Przegląd TELEKOMUNIKACYJNY
WIADOMOŚCI TELEKOMUNIKACYJNE

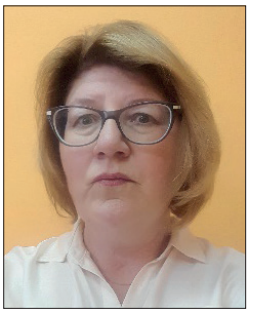
4/2022

Dwumiesięcznik Stowarzyszenia Elektryków Polskich

ISSN 1230-3496, e-ISSN 2449-7487
Cena: 52 zł (w tym 8% VAT)

WYDAWNICTWO *SIGMA-NOT*

Fot. Leszek Putkowski



Szanowni Państwo,

Z wielką przyjemnością zapraszamy do udziału w tegorocznej Multikonferencji Krajowego Środowiska Tele- i Radiokomunikacyjnego, na którą składają się dwa wydarzenia: Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji (KKRRiT) oraz Krajowe Sympozjum Telekomunikacji i Teleinformatyki (KSTiT). Multikonferencja 2022 odbędzie się na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej w dniach 7 – 9 września.

Obie konferencje mają długoletnią tradycję. Historia KSTiT sięga roku 1985, kiedy to z inicjatywy Sekcji Telekomunikacji Komitetu Elektroniki i Telekomunikacji Polskiej Akademii Nauk zorganizowano pierwsze środowiskowe sympozjum naukowe, natomiast KKRRiT odbywa się prawie nieprzerwanie i cyklicznie (za wyjątkiem 2021 r.) od roku 2001 w różnych miastach: Gdańsku, Krakowie, Łodzi, Poznaniu, Warszawie oraz Wrocławiu. Tegoroczne spotkanie jest trzecim, a zarazem ostatnim, organizowanym w formule Multikonferencji Krajowego Środowiska Tele- i Radiokomunikacyjnego. Od przyszłego roku bowiem przekształci się ona w Krajową Konferencję Radiokomunikacji i Teleinformatyki, na którą już dzisiaj serdecznie zapraszamy w imieniu przyszłych organizatorów.

Pomysł takiej integracji jest naturalną konsekwencją przenikania się dziedzin radiokomunikacji, telekomunikacji oraz informatyki. Przyczyny integracji są wielorakie, jednakże najważniejszą wydaje się być fakt, iż zarówno telekomunikacja jak i radiokomunikacja, praktycznie w pełni zaczynają bazować na rozwiązaniach software'owych (poza nielicznymi wyjątkami, jak np. sprzętowe moduły nadawczo-odbiorcze). Funkcjonalności zarządzania w systemach telekomunikacyjnych oraz radiokomunikacyjnych budowane są w postaci tzw. mikrousług uruchamianych na platformach wirtualizacyjnych. Dana mikrousługa może być wykorzystywana w różnych środowiskach (radiowym, telekomunikacyjnym czy też informatycznym), zapewniając w ten sposób wysoki stopień modularności rozwiązania, a tym samym duży poziom skalowalności całego ekosystemu ICT. Zjawisko to wyraźnie odzwierciedla tematyka Multikonferencji, w której na pierwszy plan wysuwają się tematy związane z rozwojem sieci 5G i 6G, Internetem Rzeczy oraz algorytmami sztucznej inteligencji w systemach radiowych i w sieciach telekomunikacyjnych. Ciągle aktualne kwestie stanowią zagadnienia kompatybilności elektromagnetycznej systemów i urządzeń radiowych oraz bezpieczeństwo sieci i systemów teleinformatycznych, w szczególności kryptografia i mechanizmy cyberbezpieczeństwa. Wiele dzieje się w obszarze technik multimedialnych i obrazowania trójwymiarowego, gdzie na szczególną uwagę zasługuje tematyka wirtualnej rzeczywistości. Wszystkie wspomniane zagadnienia są przedmiotem referatów plenarnych prezentowanych na Multikonferencji, których teksty znajdą Państwo w niniejszym numerze Przeglądu.

Bogaty i różnorodny program Multikonferencji obejmuje ponadto inne aktualne i ważne, z punktu widzenia zastosowań praktycznych, zagadnienia. Przewidziano 6 referatów plenarnych, 21 sesji tematycznych, 3 sesje specjalne URSI (Międzynarodowa Unia Nauk Radiowych), 2 sesje firmowe oraz panel dyskusyjny poświęcony sieciom mobilnym i multimediom. Poza referatami plenarnymi, łącznie na Multikonferencji 2022 prezentowane będzie 110 referatów, w tym 6 autorstwa finalistów Konkursu Prac Doktorskich organizowanego przez Fundację Wspierania Rozwoju Radiokomunikacji i Technik Multimedialnych.

Obecność na konferencji czołowych przedstawicieli polskiej nauki, inżynierów i ludzi biznesu, stwarza doskonałą okazję do wymiany doświadczeń i opinii w zakresie szeroko pojętej radiokomunikacji i telekomunikacji. Konferencje KKRRiT oraz KSTiT od zawsze przyciągały młodych pracowników nauki, w szczególności doktorantów i studentów – dla tych ostatnich to często pierwszy krok na drodze do kariery naukowej.

* * *

Na zakończenie pragniemy podziękować Patronom Honorowym Multikonferencji 2022 za życzliwe wsparcie inicjatywy zorganizowania tego wydarzenia i promowanie go w różnych środowiskach. Osobne wyrazy podziękowania kierujemy pod adresem Sponsorów, których finansowe wsparcie wysoko cenimy. Pragniemy podziękować członkom Komitetów Programowych KKRRiT oraz KSTiT za stworzoną atmosferę współpracy, wsparcie w niełatwym procesie recenzowania oraz cenne uwagi merytoryczne. Koleżankom i Kolegom z Komitetu Organizacyjnego oraz innym osobom zaangażowanym w organizację Multikonferencji należą się szczególne podziękowania za poświęcony czas i znakomitą współpracę. Wszystkim autorom referatów oraz uczestnikom życzymy owocnych obrad, ciekawych wrażeń ze spotkań koleżeńskich oraz udanego pobytu w Warszawie.

dr hab. inż. Kajetana Snopek, prof. uczelni.

Kajetana Snopek

Przewodnicząca KKRRiT

dr hab. inż. Jordi Mongay Batalla, prof. uczelni.

Jordi Mongay Batalla

Przewodniczący KSTiT

PRZEGLĄD TELEKOMUNIKACYJNY – WIADOMOŚCI TELEKOMUNIKACYJNE
Tele-Radio-Elektronika, Informatyka

TELECOMMUNICATION REVIEW – TELECOMMUNICATION NEWS
Tele-Radio-Electronics, Information Technology

4 ’2022

KOLEGIUM REDAKCYJNE

Redaktor naczelny: dr inż. BOGDAN ZBIERZCHOWSKI
Z-ca red. naczelnego: mgr HANNA WASIAK
Sekretarz redakcji: BOŻENA LACHOWICZ
Tel.: 604 446 028, e-mail: elektronika@red.pl.pl
Opracowanie graficzne: dr inż. PAWEŁ TOMASZEWICZ
Redaktor statystyczny: dr inż. GRZEGORZ BOROWIK
Redakcyjna strona internetowa: dr inż. MARIUSZ RAWSKI

RADA PROGRAMOWA

prof. dr hab. inż. Józef Modelski (przewodniczący), mgr inż. Krystyn Antczak, prof. dr hab. inż. Jerzy Czajkowski, prof. dr hab. inż. Andrzej Dobrogowski, dr inż. Andrzej Dulka, mgr inż. Andrzej Grześkowiak, mgr inż. Bertrand Le Guern, prof. dr hab. inż. Andrzej Jajszczyk, inż. Stefan Kamiński, inż. Zdzisław Kleszcz, mgr inż. Krzysztof Kwiecień, mgr inż. Zbigniew Lange, prof. dr hab. inż. Józef Lubacz, dr inż. Janusz Morawski, dr inż. Andrzej Wilk, prof. dr hab. inż. Tadeusz Więckowski, prof. dr hab. inż. Józef Woźniak, płk dr inż. Mieczysław Żurawski

Redakcja: ul. Ratuszowa 11 (budynek Instytutu Tele- i Radiotechnicznego), VII piętro, pokój 737, tel. 22 670-08-20 (+ poczta głosowa), tel. kom. 501 222 782. Przyjęcia interesantów w godz. 10–14.

Adres do korespondencji: ul. Ratuszowa 11, 03-450 Warszawa 1

E-mail: przeg.tel@sigma-not.pl
Internet: www.przegladtelekomunikacyjny.pl

Czasopismo dostępne wyłącznie w prenumeracie

Artykułów niezamówionych redakcja nie zwraca.
Redakcja zastrzega sobie prawo dokonywania skrótów i poprawek w nadesłanych materiałach.

Przygotowanie: Studio DTP Wydawnictwa SIGMA-NOT
Druk i oprawa: Zakład Poligrafi i Kolportażu Wydawnictwa SIGMA-NOT, www.sigma-not.pl

WYDAWNICTWO SIGMA-NOT

03-450 Warszawa, ul. Ratuszowa 11
tel.: 22 818-09-18, 22 818-98-32
Internet: http://www.sigma-not.pl
Prenumerata
e-mail: prenumerata@sigma-not.pl
Sekretariat
e-mail: sekretariat@sigma-not.pl
Dział Reklamy i Marketingu
e-mail: reklama@sigma-not.pl

Autor za publikację artykułu w czasopiśmie naukowym *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* otrzymuje 20 punktów zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 1 grudnia 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych.

Zeszyt wydany w wersji elektronicznej jako pierwotnej (referencyjnej)

Zamówienia na ogłoszenia należy kierować pod adresem Redakcji (adres jak wyżej) lub Działu Reklamy i Marketingu Wydawnictwa SIGMA-NOT, ul. Ratuszowa 11, 00-950 Warszawa, tel. 22 827-43-65, fax 22 826-80-16. Za treść i wygląd graficzny ogłoszeń Redakcja nie bierze odpowiedzialności.

Cena zeszytu: 52 zł (w tym 8% VAT)
Nakład (w tym wersja elektroniczna) do 500 egz.

TREŚĆ

Spis referatów zamieszczonych na Portalu Informacji Technicznej Sigma-NOT 95

Referaty plenarne

M. NIEMIEC
Technika kwantowa – szansa czy zagrożenie dla cyberbezpieczeństwa? ... 99
S. HAUSMAN
Techniczne aspekty oddziaływania na ludzi fal radiowych w systemach komórkowych 5G i 6G ... 105
W. BURAKOWSKI i inni
Planowane krajowe laboratorium badawcze sieci i usług 5G wraz z otoczeniem ... 110
A. DZIEMBOWSKI, D. MIELOCH
Nowe techniki kompresji wizji dla rzeczywistości wirtualnej – MPEG Immersive Video ... 116
K. STANIEC
Wydajność wąskopasmowych systemów Internetu rzeczy w trudnych warunkach propagacyjnych i zakłóceniowych ... 124

Referaty sesyjne

M. AMANOWICZ, M. KAMOL
Świadomość bezpieczeństwa cybernetycznego operatora usług kluczowych ... 135
A. DUSTOR
Zastosowanie modelowania statystycznego w weryfikacji mówcy 139
P. GARBAT i inni
System analizy obrazu dla potrzeb systemu AR ... 143
C. ZIÓŁKOWSKI, J. M. KELNER
Ukierunkowanie wiązek antenowych a przepustowość kanału radiowego w warunkach propagacji NLOS ... 147

CONTENTS

Abstracts of the papers included on the Portal Sigma-NOT 95

Plenary addres

M. NIEMIEC
Quantum technology – a cybersecurity opportunity or a threat? 99
S. HAUSMAN
Technical aspects of the impact of radio waves in 5G and 6G cellular systems on humans ... 105
W. BURAKOWSKI et al
5G national laboratory concept: 5g networks, services, measurement tools and applications ... 110
A. DZIEMBOWSKI, D. MIELOCH
New video compression techniques for virtual reality – MPEG Immersive Video ... 116
K. STANIEC
The performance of narrowband Internet of Things systems under harsh propagation and interference conditions ... 124

Selected papers

M. AMANOWICZ, M. KAMOLA
Cyber security awareness of essential services operator ... 135
A. DUSTOR
Application of statistical modelling to speaker verification ... 139
P. GARBAT et al.
Image Analysis System for Augmented Reality Application ... 143
C. ZIÓŁKOWSKI, J. M. KELNER
Orientation of antenna beams versus radio channel capacity in non-line-of-sight (NLOS) propagation conditions ... 147

LISTA RECENZENTÓW

Prof. dr hab. Adam Abramowicz
Prof. dr hab. Marek Amanowicz
Prof. dr hab. Andrzej Dobrowolski
Prof. dr hab. Piotr Gajewski
Prof. dr hab. Marcin Iwanowski
Dr hab. Jacek Izydorczyk
Prof. dr hab. Tomasz Kacprzak
Prof. dr hab. Andrzej Karwowski
Prof. dr hab. Ryszard Katulski
Prof. dr hab. Bogdan Kwolek
Prof. dr hab. Roman Kubacki
Prof. dr hab. Lidia Łukasiak
Prof. dr hab. Stanisław Osowski
Dr hab. inż. Andrzej Paszkiewicz
Prof. dr hab. Krzysztof Perlicki
Prof. dr hab. Grzegorz Różański
Prof. dr hab. Jacek Stefański
Dr hab. Zenon Szczepaniak
Prof. dr hab. Maciej Walkowiak
Prof. dr hab. Krzysztof Wesołowski
Prof. dr hab. Ryszard Zieliński

Autor za publikację artykułu w czasopiśmie naukowym *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* otrzymuje 20 punktów zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 1 grudnia 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych.

WYDAWNICTWO *SIGMA-NOT*

WARIANTY PRENUMERATY 2022:

Prenumerata papierowa – czasopismo tylko w wersji papierowej, za jego dostarczenie doliczamy opłatę roczną;

Prenumerata cyfrowa – czasopismo wyłącznie w wersji cyfrowej, dostępne na Portalu Informacji Technicznej www.sigma-not.pl, prenumerator otrzyma indywidualny link dostępu do logowania na Portalu;

PRENUMERATA PAKIET PLUS – czasopismo w wersji papierowej (bez opłaty za dostarczanie prasy) oraz cyfrowej, a także dostęp do archiwum zaprenumerowanego tytułu na Portalu Informacji TechnicznEj www.sigma-not.pl wraz z indywidualnym linkiem do logowania.

Cena jednego egzemplarza – 52 zł

CENY PRENUMERATY ROCZNEJ:

- prenumerata papierowa: 300 zł brutto + (koszty wysyłki 18 zł);
- prenumerata cyfrowa: 300 zł brutto;
- prenumerata w wersji Pakiet PLUS (papierowa + cyfrowa + archiwum) 450,- zł.

PORTAL INFORMACJI TECHNCZEJ **www.sigma-not.pl** – to największa polska internetowa baza czasopism i artykułów technicznych, wyposażona w szybką i skuteczną wyszukiwarkę tematyczną, umożliwiającą dostęp on-line do ponad 129 000 publikacji z różnych dziedzin.
Dla Państwa wygody logowanie do Portalu Informacji Technicznej stało się łatwiejsze. Portal zyskał nową, przejrzystą szatę graficzną, pojawiły się nowe funkcje ułatwiające zakup pojedynczych artykułów i całych zeszytów. Każdy prenumerator wersji cyfrowej oraz prenumeraty Pakietu PLUS otrzyma indywidulany link z kodem dostępu do zamówionego czasopisma.

Cena brutto zawiera 8% VAT na czasopisma w wersji papierowej oraz cyfrowej. W przypadku zmiany przez ustawodawcę stawki VAT na czasopisma i w konsekwencji zmiany ceny brutto, prenumerator zobowiązany jest do dopłaty różnicy.
Dla prenumeratorów zagranicznych obowiązuje cena według kursu waluty NBP (z dnia bezpośrednio poprzedzającego datę wystawienia faktury) plus koszty wysyłki.

PRENUMERATĘ ZAMAWIAMY

22 840 35 89
prenumerata@sigma-not.pl
www.sigma-not.pl

SYMBOLE TEMATYKI ARTYKUŁÓW

ZAGADNIENIA OGÓLNE	SIECI TELEKOMUNIKACYJNE	USŁUGI MULTIMEDIA	ELEMENTY UKŁADY METODY	TELETRANSMISJA	TELEFONIA	TELEINFORMATYKA	OPTOTELEKOMUNIKACJA	RADIOKOMUNIKACJA RADIOFONIA TELEWIZJA	POMIARY	EKONOMIKA PRAWO	SPONSOROWANE
GENERAL PROBLEMS	COMMUNICATION NETWORKS	SERVICES MULTIMEDIA	COMPONENTS DEVICES METHODS	TELETRANSMISSION	TELEPHONY	COMPUTER SCIENCE	OPTOCOMMUNICATIONS	RADIOCOMMUNICATION RADIO TELEVISION	MEASUREMENTS	ECONOMICS LAW	SPONSORING

Niniejszy zeszyt (4'2022) naszego miesięcznika jest związany z **Multikonferencją Krajowego Środowiska Tele- i Radiokomunikacyjnego**, na którą składają się dwa wydarzenia: **Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji** (KKRRiT) oraz **Krajowe Sympozjum Telekomunikacji i Teleinformatyki** (KSTiT). Zeszyt zawiera całość materiałów konferencyjnych. Referaty plenarne są prezentowane w formie drukowanej, zaś pełne teksty wszystkich referatów są zamieszczone na Portalu Informacji Technicznej SIGMA-NOT z nieodpłatnym dostępem: **https://sigma-not.pl/zeszyt-6934-przeglad-telekomunikacyjny-2022-4.html.** Stanowią one integralną część zeszytu. Wszystkie artykuły i referaty są recenzowane. Podstawowe informacje na temat konferencji znajdą Państwo w artykule wstępnym (II str. okładki), przygotowanym przez dr hab. inż. Kajetanę Snopek, prof. Uczelni – przewodniczącą KKRRiT i dr. hab. inż. Jordi Mongay Batalla, prof. Uczelni – przewodniczącego KSTiT.

Lista recenzentów

Marek Amanowicz, Sławomir Jerzy Ambroziak, Jordi Mongay Batalla, Andrzej Bęben, Marek Blok, Hanna Bogucka, Wojciech Burakowski, Piotr Chołda, Andrzej Chydziński, Jacek Cichocki, Tadeusz Czachórski, Adam Dąbrowski, Grzegorz Danilewicz, Marek Domański, Jerzy Domżał, Przemysław Dymarski, Piotr Gajewski, Mariusz Głąbowski, Maciej J. Grzybkowski, Sławomir Hausman, Andrzej Jajszczyk, Lucjan Janowski, Łukasz Januszkiewicz, Wojciech Kabaciński, Sylwester Kaczmarek, Ryszard Katulski, Jan Kelner, Artur Klepaczko, Adrian Kliks, Mirosław Klinkowski, Jerzy Konorski, Katarzyna Kosek-Szott, Bożena Kostek, Andrzej Krawczyk, Rafał Krenz, Andrzej Kucharski, Sławomir Kula, Krzysztof Kulpa, Mikołaj Leszczuk, Jerzy Łopatka, Wojciech Mazurczyk, Jarosław Michalak, Jacek Misiurewicz, Józef Modelski, Marek Natkaniec, Marcin Niemiec, Ewa Niewiadomska-Szynkiewicz, Andrzej R. Pach, Piotr Pacyna, Zdzisław Papir, Grzegorz Pastuszak, Krzysztof Perlicki, Zbigniew Piotrkowski, Jacek Rak, Piotr Remlein, Ryszard Romaniuk, Grzegorz Różański, Marek Różycki, Dominik Rutkowski, Jarosław Sadowski, Piotr Samczyński, Wiktor Sęga, Mirosław Siergiejczyk, Jerzy Siuzdak, Piotr Słobodzian, Stanisław Słowik, Bogusław Smólski, Kajetana Snopek, Kamil Staniec, Maciej Stasiak, Jacek Stefański, Marek Suchański, Paweł Szczepański, Krzysztof Szczypiorski, Szymon Szott, Halina Tarasiuk, Artur Tomaszewski, Krzysztof Walkowiak, Maciej Walkowiak, Krzysztof Wesołowski, Tadeusz Więckowski, Wojciech Wojtasiak, Józef Woźniak, Yevhen Yashchyshyn, Ryszard J. Zieliński, Tomasz P. Zieliński, Cezary Ziółkowski, Jerzy Żurek, Piotr Zwierzykowski.

Referaty sesji tematycznych zamieszczone na Portalu Informacji Technicznej SIGMA-NOT

REFERATY SESYJNE KSTiT 2022

ST1: Aplikacje i usługi

A. Dustor
Zastosowanie modelowania statystycznego w weryfikacji mówcy ..... (PTiWT 4 str. 139)

M.G. Stochel, P. Chołda, M.R. Wawrowski
Technical debt prioritization in telecommunication applications – why the actual refactoring deviates from the plan and how to remediate it? 151

J. Kryszyn, W. Smolik, D. Wanta, M. Midura, P. Wróblewski
Porównanie standardów openEHR oraz HL7 FHIR 156

D. Wasilewski, M. Kowalczyk
Zwiększenie bezpieczeństwa poczty elektronicznej – narzędzie wspomagające 160

P. Łubkowski
Projekt architektury systemu wspomagania decyzji o ewakuacji medycznej 164

ST2: Bezpieczeństwo sieci

M. Amanowicz, M. Kamola
Świadomość bezpieczeństwa cybernetycznego operatora usług kluczowych ..... (PTiWT4 str. 135)

S. Wnęk, R.E. Kościej, M. Niemiec
Heurystyczna detekcja ataków sieciowych z wykorzystaniem środowiska Snort 3 168

Z. Kotulski, W. Niewolski, T. Nowak, M. Sepczuk, R.-H. Hsu, S.-M. Cheng, S.-C. Chang, T.-Y. Yu, C.-F. Hung
ACMEC – Access control in 5G MEC. Joint research project of NTUST (Taiwan) and WUT (Poland) 172

M. Biernacki, R. Giermakowski, M. Rawski
Algorytmy lightweight dla systemów IoT implementowanych w FPGA 177

B. Zabołotny, M. Chmiel, M. Korona, F. Kozioł, M. Rawski
IoTrust – framework realizujący mechanizmy bezpieczeństwa end-to-end dla IoT 181

ST3: Przewarzanie danych

K. Wnęk, P. Boryło
System przetwarzania i dystrybucji danych oparty na Apache NiFi 185

P. Pyda, J. Śliwa, T. Dalecki, M. Przywuski
Efektywność replikacji danych w centrach danych 190

D. Mirowska, J. Domżał
Automatyzacja zarządzania infrastrukturą pulpitów wirtualnych w środowisku VMware Horizon 194

S. Zarębski, A. Kuźmich, S. Sitko, K. Rusek, P. Chołda
Siamese neural networks on the trail of similarity in bugs in 5G mobile network base stations 198

K. Haustein, M. Blok
Aplikacja demonstrująca działanie kodów fontannowych ... 202

ST4: Jakość usług

P. Falkowski-Gilski, T. Uhl
Badanie jakości transmisji w sieciach komórkowych na przestrzeni lat 2019-2021: wpływ pandemii COVID-19 na poziom świadczonych usług 206

B. Nowak, P. Zwierzykowski
Symulacyjne badania wpływu mechanizmu priorytetyzacji ruchu na efektywność ruchową wielousługowych węzłów sieciowych 210

R. Polak, J. Wojtuń, P. Łubkowski
Analiza jakości szyfrowanej komunikacji VoIP w granicznych warunkach czułości odbiornika radiowego 215

M. Zamłyńska, P. Falkowski-Gilski
Analiza jakości transmisji treści audio-wideo w symulowanym łączu telekomunikacyjnym z wykorzystaniem techniki OFDM 219

M. Siergiejczyk, J. Paś, A. Rosiński
Analiza niezawodnościowo-eksploatacyjna inteligentnych systemów transportowych z uwzględnieniem zakłóceń elektromagnetycznych 224

ST5: Ocena bezpieczeństwa

M. Borowski
Ocena bezpieczeństwa algorytmów zapewniających poufność oraz uwierzytelnienie danych w sieci 5g 229
D. Topyła, K. Murat, K. Zdulski, D. Paczesny, J. Bieniasz
Analiza bezpieczeństwa protokołów w wybranych interfejsach bezprzewodowych urządzeń Internetu Rzeczy 234
K. Zdulski, D. Topyła, K. Murat, D. Paczesny, J. Bieniasz
Modyfikacja oprogramowania urządzeń Internetu Rzeczy: analiza możliwości i wpływu na bezpieczeństwo 238
A. Kraśniewski, K. Szczypiorski
Kształcenie w zakresie Cyberbezpieczeństwa na Politechnice Warszawskiej 242
M. Nowak, M. Walkowski, S. Sujecki
Wsparcie procesu zarzadzania podatnościami przy użyciu konwersji oceny bazowej CVSS 2.0 do 3.x 246

ST6: Zarządzanie sieciami

K.I. Kurczab, K. Kosek-Szott
Analiza porównawcza środowisk emulacyjnych Kathará i Netkit 250
J. Mongay Batalla, S. Sujecki, T. Wichary,
Technologia dystrybucji kluczy kwantowych w sieciach 5G opartych na wielu dostawcach sprzętu 254
M. Stypiński
Analiza zależności pomiędzy parametrami sieci TPM a czasem jej synchronizacji 259
S. Kulig, A. Matiolański
Porównanie architektur aplikacji sieciowych: monolityczna i mikroserwisy 263
M. Golański, P. Tomaszewicz, M. Rawski
Środowisko testowe funkcji core baseline dla frameworku IoTrust 267

ST7: Routing sieci telekomunikacyjnych

K. Bachorski, S. Kaczmarek, M. Sac
Zastosowanie koncepcji SDN w warstwie transportowej sieci IMS/NGN 271
J. Martyna
Heurystyczny algorytm rozmieszczenia łańcucha funkcji obsługi w programowej sieci 5G 278
E. Biernacka, J. Domżał
Mechanizm zestawiania ścieżek optycznych typu bypass w sieciach IP/EON 282
P. Augustyniak, J. Skóra, P. Zwierzykowski
Przegląd otwarto-źródłowych narzędzi do zarządzania bezprzewodowymi sieciami lokalnymi 286
B. Musznicki, M. Piechowiak, P. Zwierzykowski
Problemy badawcze i modelowanie sieci sensorowych w środowiskach miejskich 291

ST8: Protokoły przekazania kluczy i algorytmy bezpieczeństwa

M. Murat, K. Zdulski, D. Topyła, M. Marzęcki, J. Bieniasz
Analiza bezpieczeństwa protokołów komunikacyjnych urządzeń Inteligentnego Domu opartych na mikrokontrolerze ESP 295
A. Dustor
System weryfikacji tożsamości na bazie Kaldi 299
M. Sepczuk
Aplikacyjna zapora sieciowa z elementami obrony mimicznej 303
M. Płachta, A. Janicki
Detekcja steganografii obrazowej z użyciem prostych sieci neuronowych 307
R. Giermakowski, M. Biernacki, M. Rawski
Implementacja w FPGA modułu PUF opartego na oscylatorach pierścieniowych 311

ST9: Sesja projektu SyMEC

A. Bęben, W. Burakowski, M. Sosnowski, D. Więcek, P. Olender, B. Belter, B. Krakowiak, M. Hoeft, M. Miszewski, A. Chomicki
Implementacja architektury SyMEC 315
M. Sosnowski, A. Bęben, W. Jóźwiak, M. Miszewski
Serwer i platforma MEC 320
B. Krakowiak, B. Belter, A. Bęben, W.Burakowski, W. Jóźwiak, M. Sosnowski,P. Olender
System sterowania i zarządzania SyMEC 325
M. Hoeft, K. Gierłowski, J. Woźniak, W. Gumiński, K. Nowicki, J. Rak
Wdrożenie systemu MEC w sieciach WLAN: Moduł współpracy serwera MEC z siecią dostępową WLAN 330
D. Więcek, T. Anuszczyk, I. Michalski, K. Rzeźniczak
Uruchomienie systemu MEC w sieciach 3GPP: Moduł współpracy serwera MEC z sieciami 4G i 5G 335
T. Łysiak, J. Kabat
Aplikacje i usługi systemu SyMEC 339

REFERATY SESYJNE KKRRiT 2022

SR1: Technika antenowa

M. Michalska, A. Kliks
Automatyczny dobór kąta pochylenia anten w systemach komórkowych 343
A. Piwowar, P. Bajurko
Rozkłady pól elektromagnetycznych w układzie anteny tubowej diagonalnej z soczewką 347
T.A. Miś
Badanie pola elektrycznego dojrzałej chmury burzowej przy pomocy długich anten napowietrznych 351
K. Bechta
Optymalizacja skutecznej charakterystyki antenowej i jej znaczenie dla oceny ekspozycji na promieniowanie elektromagnetyczne 355
W. Karcz
Metody wyznaczania rozkładu gęstości mocy w polu bliskim i dalekim wokół anten parabolicznych linii radiowych 361

SR2: Systemy radiowe do zastosowań specjalnych

K. Wesołowski
Wybór modulacji cyfrowej w nanosatelitarnym systemie transmisyjnym 367
R. Krenz, M. Sybis, I. Zainutdinov
Wykorzystanie pasma C do łączności z nano-satelitami CUBESAT 371
A. Samorzewski, A. Kliks
Modele energetyczne dla sieci bezprzewodowych z mobilnymi węzłami dostępowymi 376
S. Janji, A. Samorzewski, A. Kliks
Procedura rozmieszczenia mobilnych przekaźników dosyłowych FSO zasilanych z OZE 380

J. Dułowicz, P.K. Skokowski
Autonomiczny BSP do rozpoznania i zakłócania łączności sieci komórkowych 384

SR3: Techniki multimedialne

B. Szydełko, D. Klóska, A. Dziembowski
Parametryzacja geometrii sceny dla estymacji głębi w dekoderze wizji wszechogarniającej 388
D. Klóska, B. Szydełko, D. Mieloch
Wykorzystanie wejściowych map głębi w estymacji głębi po stronie dekodera 392
Ł. Popek, R. Perz, G. Galiński
Porównanie różnych metod detekcji i rozpoznawania zwierząt na obrazach z kamery termowizyjnej 396
P. Garbat, G. Galiński, M. Żakowski, M. Bieniek, M. Lasocki
System autonomicznej analizy obrazu dla potrzeb systemu AR ..... (PTiWT str. 143)
K. Gabor-Siatkowska, A. Janicki
Czy kalibracja okulografu w interfejsie człowiek-komputer zawsze jest niezbędna? 400

SR4: Metody sztucznej inteligencji w systemach radiowych

C. Adamczyk, A. Kliks
Wykrywanie konfliktów decyzji w inteligentnych kontrolerach radiowych w otwartych sieciach dostępu radiowego O-RAN 404
A. Brytańczyk, K. Kosek-Szott
Zastosowanie uczenia maszynowego do kontroli szybkości transmisji w sieciach standardu IEEE 802.11 408
M. Blok, K. Gmiński, P. Miklaszewski
Klasyfikator SVM w zastosowaniu do synchronizacji sygnału OFDM zniekształconego przez kanał wielodrogowy 412
M. Hoffmann
Dynamiczne wygaszanie punktów transmisyjnych w wieloantenowej sieci piątej generacji z użyciem głębokiego uczenia ze wzmocnieniem 416
Ł. Kułacz
Porównanie metod detekcji zajętości widma radiowego z wykorzystaniem uczenia federacyjnego z użyciem oraz bez użycia węzła centralnego 421

SR5: Zagadnienia warstwy fizycznej w nowoczesnych sieciach radiokomunikacyjnych (URSI – Komisja C+F)

R. Burczyk, J. Sadowski
Algorytm detekcji wąskopasmowych interferencji międzysystemowych ukrytych w paśmie pracy systemu LTE 425
J. Magiera, P. Rajchowski, K. Cwalina
Estymacja parametrów kanału NB-IoT w łączu w dół 429
O. Błaszkiewicz, P. Rajchowski, J. Sadowski
Analiza metod synchronizacji w sieci NB-IoT 433
P. Rajchowski, K.K. Cwalina, J. Sadowski
Lokalizowanie terminala użytkownika z użyciem interfejsu NB-IoT 437
A. Olejniczak, K.K. Cwalina, J. Sadowski, J. Stefański
Analiza efektywności metod głębokiego uczenia w odbiorze sygnałów GMSK 441

SR6: Projektowanie urządzeń radiokomunikacyjnych

D.W. Rosołowski, P. Korpas, M. Kajczuk, D. Gryglewski, W. Wojtasiak,
Projektowanie stacji bazowej LTE-A na pasmo 450 MHz przy użyciu uniwersalnej platformy SDR 445
D. Gryglewski, W. Wojtasiak
Projektowanie wzmacniacza Doherty'iego w praktyce 449
D.W. Rosołowski
Wybór danych uczących dla algorytmów linearyzacji wzmacniaczy mocy 453
A. Pacewicz
Wybrane metody charakteryzacji mikrofalowej materiałów magnetycznie żyrotropowych 457
J. Sadowski, J. Stefański
Zdalna realizacja zajęć w laboratoriach radiokomunikacyjnych 461

SR7: Aspekty radiowe sieci 5G i 6G

M. Suchański, J. Śliwa, K. Kosmowski
Systemy 5G w zastosowaniach militarnych 465
Z. Zakrzewski
Wąskopasmowa kompensacja DIPP w mobilnym łączu fronthaul DSB-RFoF 470
M. Wachowiak, P. Kryszkiewicz
Wpływ nieliniowości na jakość sygnału odbieranego w systemie M-MIMO OFDM 476
M. Wasilewska, H. Bogucka
Inteligentna predykcja zajętości widma przy jednoczesnej ochronie użytkownika pierwotnego 481

SR8: Transmisja radiowa w systemach Internetu Rzeczy

K.K. Cwalina, O. Błaszkiewicz, A. Olejniczak, P. Rajchowski, J. Sadowski
Estymacja i analiza stanu kanału radiowego interfejsu LTE na potrzeby realizacji heterogenicznego algorytmu multilink z użyciem głębokiego uczenia 486
C. Ziółkowski, J.M. Kelner
Ukierunkowanie wiązek antenowych a przepustowość kanału radiowego w warunkach propagacji NLOS (PTiWT str. 147)
M. Zamłyńska
Modelowanie interferencji w systemach radiokomunikacyjnych z transmisją OFDM za pomocą narzędzi statystycznych 492
J. Szewczyk, M. Nowak, P. Remlein, A. Głowacka
Platformy do wdrażania komunikacji w standardzie LORAWAN 497
V. Djaja-Jośko
Analiza wpływu bezprzewodowej synchronizacji węzłów oraz liczby lokalizowanych etykiet na skuteczność odbioru pakietów w ultraszerokopasmowym systemie lokalizacyjnym 501

SR9: Propagacja fal radiowych

S. J. Ambroziak, F.D. Cardoso, P. Kosz, M.M. Ferreira, L.M. Correia
Analiza zaników szybkozmiennych w radiowych sieciach BAN pracujących w rewerberacyjnym środowisku propagacyjnym 505
M. Nikiforuk, K. Cichoń
Tłumienie roślinności dla częstotliwości 28 GHz 511
B. Majewski, K. Malon, M. Kryk, J.M. Kelner
Tworzenie map tłumienia propagacyjnego z wykorzystaniem metody IDW 515
M. Kryk
Ocena efektywności generacji map tłumienia propagacyjnego 519
P. Komorowski
Multipleksacja promieniowania THz z wykorzystaniem niepoosiowych optycznych struktur dyfrakcyjnych..... 523

SR10: Zarządzanie częstotliwościami i inżynieria widma

P. Gajewski, J. Łopatka, P. Łubkowski
Modelowanie wielozakresowej sieci MANET z dynamicznym zarządzaniem widmem 527

R. Mądral, R. Tyniów
Aspekty regulacyjne oraz zarządzanie częstotliwościami w przypadku małych oraz bardzo małych satelitów 531
M. Mazuro, P.K. Skokowski
Sensor monitoringu widma z wykorzystaniem radia programowalnego 535
M. Rybakowski, K. Bechta
Analiza symulacyjna rzeczywistej ekspozycji na promieniowanie elektromagnetyczne od systemów wieloantenowych 5G 539
M. Wrzesiński, P. Bajurko
Zastosowanie metod kalibracyjnych w pełnofalowej symulacji elektromagnetycznej obwodów planarnych na częstotliwościach subterahercowych 544

SR11: Sieci bezprzewodowe

M. Natkaniec, M. Kras
Poprawa wydajności pracy sieci gęstych standardu IEEE 802.11AX 548
M. Natkaniec, N. Bieryt
Analiza działania mechanizmu kolorowania w sieciach gęstych standardu IEEE 802.11AX 552
S.J. Stryczek, M. Gwiazdowicz, M. Natkaniec
BTS (Background Traffic Simulator) – narzędzie do generowania rzeczywistego ruchu dla sieci WLAN w środowisku NS-3 558
G. Teca, M. Natkaniec
Connected while disconnected paradox: an IEEE 802.11 covert channel to prevent Wi-Fi attacks in an untrusted environment 562

SR12: Systemy radionawigacyjne i radiolokalizacyjne

J. Stefański, J. Sadowski
Asynchroniczna metoda jednoczesnej estymacji położenia i orientacji obiektu za pomocą dwóch nadajników 576
M.K. Bączyk
Demonstrator radaru pasywnego z odwrotną syntetyczną aperturą 571
K. Radecki
Szybki algorytm projekcji wstecznej oparty na współrzędnych barycentrycznych do zastosowań w aktywnych i pasywnych zobrazowaniach radarowych SAR 575
A. Czapiewska, R. Nicpoń
Wpływ implementacji algorytmu Chana na jego efektywność 579

SR13: Modelowanie i badanie wpływu zaburzeń WIPEM na systemy i urządzenia elektroniczne (URSI – Komisja E)

P. Słobodzian
Modelowanie oddziaływania impulsów HPM na obiekty IK 583
R. Królikowski, P. Słobodzian
Modelowanie i badanie impulsów HPM 587
P. Bieńkowski, P. Słobodzian, P. Sobkiewicz, B. Zubrzak
Modelowanie i pomiary pola oświetlającego obiekty IK w paśmie mikrofalowym 591
L. Nowosielski, B. Dudziński, R. Przesmycki, J. Kelner
Metodyka badania wnikania impulsów HPM do obiektów ekranowanych 595
J. Małecki, A. Żak, R. Studański
Problematyka oddziaływania impulsów HPM na wybrane układy elektroniczne 599

SR14: Technika dźwiękowa i systemy radiofoniczne

M. Blok, J. Banaś, M. Pietrołaj
Strategie treningu neuronowego estymatora częstotliwości tonu krtaniowego z użyciem generatora syntetycznych samogłosek 604
P. Falkowski-Gilski
Subiektywny pomiar jakości programów radiowych strumieniowanych w sieci metodą crowdsourcingu 608
R.J. Zieliński
Analiza porównawcza zmierzonych i obliczonych natężeń pól elektrycznych od nadajników SFN DAB+ 612
R. Helon, P. Korpas
Wielokanałowy rejestrator emisji radiofonicznych UKF-FM wykonany w technice SDR 616

SR15: Skutki biologiczne oraz zastosowania medyczne pól elektromagnetycznych (URSI – Komisja K)

J. Kawecki, Ł. Januszkiewicz, P. Di Barba
Zastosowanie automatycznej optymalizacji oraz uproszczonego modelu ciała do projektowania wąskopasmowej struktury ekranującej pole EM 620
P. Oleksy, Ł. Januszkiewicz
Lokalizacja bezprzewodowych endoskopów kapsułkowych z wykorzystaniem algorytmu wyboru odbiorników 624
G. Kanclerz, J. Popielak, U. Sołdaj, S. Lefek, A. Kacprzyk, T. Rok, R. Pawlak, G. Tatoń
Poszukiwanie obiektywnych ilościowych parametrów opisujących wpływ pola elektromagnetycznego na organizm człowieka 628
G. Tatoń, A. Kacprzyk, G. Kanclerz, T. Rok, R. Pawlak, E. Rokita
Najnowsze badania dotyczące częstości występowania nadwrażliwości na promieniowanie elektromagnetyczne w Polsce 631

DWUMIESIĘCZNIK STOWARZYSZENIA ELEKTRYKÓW POLSKICH WYDAWANY PRZY WSPÓŁPRACY KOMITETU ELEKTRONIKI I TELEKOMUNIKACJI POLSKIEJ AKADEMII NAUK

Przegląd TELEKOMUNIKACYJNY
TELE–RADIO–ELEKTRONIKA–INFORMATYKA

4/2022

ROK ZAŁOŻENIA 1928 ● ROCZNIK XCIV ● ISSN 1230-3496

Referaty plenarne

Marcin NIEMIEC*

Technika kwantowa – szansa czy zagrożenie dla cyberbezpieczeństwa?

Quantum technology – a cybersecurity opportunity or a threat?

Obecny stan rozwoju techniki kwantowej umożliwia przesyłanie informacji kwantowej na odległość. Systemy telekomunikacyjne korzystające z tego typu rozwiązań mogą zapewnić poufność na niespotykanym dotychczas poziomie. Jednak praktyczne zastosowanie mechaniki kwantowej ma również drugi aspekt – jest nim ryzyko złamania obecnie stosowanych algorytmów szyfrowania danych. W artykule podsumowano obecny stan techniki kwantowej, skupiając się na tych technologiach, które mają bezpośredni wpływ na cyberbezpieczeństwo.
Słowa kluczowe: cyberbezpieczeństwo, korekcja błędów, kwantowa dystrybucja kluczy, technika kwantowa

The current state of quantum technology already enables the transfer of quantum information over a distance. Telecommunications systems using this type of solutions can ensure confidentiality at an unprecedented level. However, the practical application of quantum mechanics also has a flip side of the coin – the risk of breaking currently using data encryption algorithms. This paper summarizes the current state of quantum technology by considering those solutions that have a direct impact on cybersecurity.
Key words: cybersecurity, error correction, quantum key distribution, quantum technology

DOI: 10.15199/59.2022.4.1

Patrząc wstecz na ostatnie trzy dekady rozwoju teleinformatyki, można dostrzec jak bardzo zmieniły nas nowoczesne technologie. Cyfryzacja społeczeństwa, gwałtownie przyspieszona przez pandemię w latach 2020–2022, doprowadziła do jeszcze większej zależności od systemów teleinformatycznych, świadczących usługi związane praktycznie z każdym aspektem funkcjonowania człowieka w społeczeństwie. Wiele osób twierdzi, że kolejna rewolucja technologiczna czeka na nas tuż tuż i będzie związana z praktycznym wykorzystaniem mechaniki kwantowej do przetwarzania i przesyłania informacji na odległość. Być może to tylko kolejna burza medialno-marketingowa, która przejdzie niezauważenie obok nas, ale jedno jest pewne: technika kwantowa ma potencjał, aby gruntownie przebudować podstawy ochrony danych przyjęte powszechnie za standard w cyberprzestrzeni.

Gdyby Einstein, Schrödinger czy Heisenberg obudzili się dziś, prawdopodobnie mniej zdziwieni byliby światem cyfrowym opartym na przesyłaniu danych za pomocą dwóch stanów binarnych, lecz znacznie bardziej zaciekawiłyby ich działające urządzenia wykorzystujące w praktyce zasady mechaniki kwantowej. Nawet dla nas zaskakujący jest fakt, iż doświadczenia myślowe, oparte na zjawiskach występujących w świecie atomów czy cząstek elementarnych, są realizowane w urządzeniach dostępnych nie tylko w laboratoriach naukowych, ale i w ofercie firm komercyjnych. Wśród tych rozwiązań wyróżniają się te, które bezpośrednio dotyczą cyberbezpieczeństwa.

Artykuł wprowadza czytelnika w podstawy informatyki kwantowej oraz podsumowuje obecny stan techniki kwantowej, przedstawiając te rozwiązania, które już teraz dostępne są dla użytkowników. Ze względu na rozległość tematyki, sku-

* AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, e-mail: marcin.niemiec@agh.edu.pl

piono się na technologiach, które mają bezpośredni wpływ na cyberbezpieczeństwo – zarówno ten pozytywny, jak i negatywny. Szczególną uwagę poświęcono kryptografii kwantowej – omówiony został najczęściej implementowany protokół kwantowej dystrybucji klucza oraz metody umożliwiające korekcję błędów powstałych w wyniku niedoskonałości współcześnie budowanych kanałów kwantowych. Wśród nich znajduje się metoda oparta na synchronizacji sztucznych sieci neuronowych, zaproponowana przez autora. W końcowej części artykułu nakreślono przewidywane kierunki rozwoju techniki kwantowej.

PODSTAWY (TELE-) INFORMATYKI KWANTOWEJ

Informatyka kwantowa, korzystając z cech mechaniki kwantowej, umożliwia kodowanie i przetwarzanie informacji. Choć obecnie termin ten obejmuje także rozwiązania umożliwiające przesyłanie informacji kwantowej na odległość, być może w przyszłości teleinformatyka kwantowa rozwinie się na tyle, że zaczniemy mówić o niej jako o osobnej dziedzinie. Tak czy inaczej – podstawy kodowania informacji najprawdopodobniej pozostaną jednakowe. Przede wszystkim wspólnym pojęciem będzie bit kwantowy, będący nośnikiem informacji.

Podstawową jednostką informacji cyfrowej jest bit, który może przybierać jedną z dwóch możliwych wartości: 0 lub 1. Bit kwantowy (zwany także kubitem) różni się od swojego klasycznego odpowiednika tym, że oprócz tych dwóch stanów może przyjmować każdy ze stanów pośrednich [1]. Jest to możliwe dzięki temu, że stan kwantowy (np. polaryzacja pojedynczego fotonu) może tworzyć liniowe superpozycje, będące złożeniem stanu, w którym kubit przyjmuje wartość 0 oraz stanu, w którym przyjmuje wartość 1. Stosując popularną wśród fizyków notację Diraca, tak zdefiniowany bit kwantowy można opisać jako:

$$\alpha|0\rangle + \beta|1\rangle,$$

gdzie liczby α oraz β spełniają warunek $|\alpha|^2+|\beta|^2 = 1$, dzięki czemu wskazują na prawdopodobieństwo każdego ze stanów. Innymi słowy, rozpatrywany bit kwantowy znajduje się w stanie 0 z prawdopodobieństwem $|\alpha|^2$ oraz jednocześnie w stanie 1 z prawdopodobieństwem $|\beta|^2$. Fakt, iż kubit jest w stanie kombinacji liniowej obu stanów podstawowych (0 oraz 1), prowadzi do zaskakującego zjawiska związanego z pomiarem.

Jeśli założymy, że bit kwantowy jest nośnikiem informacji, jesteśmy zmuszeni do odczytania jego stanu w odbiorniku. Załóżmy, że nośnikiem informacji jest pojedynczy foton o polaryzacji, która jest kombinacją liniową dwóch stanów ortogonalnych: dla uproszczenia nazwijmy jeden z tych stanów *polaryzacją horyzontalną* (ten stan oznaczmy jako 0), a drugi *polaryzacją wertykalną* (ten oznaczmy jako stan 1). Jeśli tylko liczby $|\alpha|^2$ oraz $|\beta|^2$ nie są zerowe, polaryzacja fotonu będzie kombinacją obu stanów podstawowych i nieformalnie będziemy o niej mogli powiedzieć, że jest to *polaryzacja ukośna*. Chcąc wykonać pomiar polaryzacji takiego fotonu – np. za pomocą filtru polaryzacyjnego lub kryształu dwójłomnego – możemy ustawić detektor *horyzontalnie* (wzdłuż osi określającej stan ortogonalny 0) lub *wertykalnie* (wzdłuż osi określającej stan ortogonalnv 1). Wynik takiego pomiaru będzie losowy, a liczby $|\alpha|^2$ oraz $|\beta|^2$ określą, z jakim prawdopodobieństwem zarejestrujemy foton w detektorze *horyzontalnym* lub *wertykalnym*. Dlatego mówimy, że pomiar zaburza stan kwantowy, choć tak naprawdę zmierzyliśmy polaryzację wzdłuż osi określającej jeden ze stanów ortogonalnych (*horyzontalną* lub *wertykalną*), a te przecież nie istniały wcześniej. Doprowadza to nas do zdumiewającego wniosku – zmierzona wielkość nie istniała przed dokonaniem pomiaru.

Kolejnym zaskakującym zjawiskiem w świecie kwantowym jest splątanie. Określa ono pewną *relację* pomiędzy stanami kwantowymi obiektów. Przykładowo – możemy utworzyć dwa fotony, których polaryzacje będą od siebie ściśle zależne. Zależność ta może być następująca: jeśli w wyniku pomiaru okaże się, że jeden z fotonów jest w stanie 0, to i drugi z nich jest również w stanie 0, ale równie prawdopodobnym wynikiem pomiaru pierwszego z fotonów jest stan 1, a wtedy również i drugi foton będzie w stanie 1:

$$\frac{1}{\sqrt{2}}|0\rangle|0\rangle + \frac{1}{\sqrt{2}}|1\rangle|1\rangle.$$

Co ciekawe, stany obu fotonów będą od siebie zależne nawet jeśli oddalimy obie cząstki od siebie na dużą odległość. Ustalenie stanu drugiej cząstki jest natychmiastowe w momencie wykonania pomiaru pierwszej z nich.

Bezpośredni wpływ na bezpieczeństwo techniki kwantowej ma jeszcze jedna zasada – zakaz klonowania nieznanego stanu kwantowego [2]. Brak możliwości sklonowania nieznanego stanu kwantowego chroni nas przed scenariuszem, w którym atakujący mógłby tworzyć wierne kopie bitu kwantowego i w ten sposób przeprowadzać wielokrotne pomiary, które doprowadziłyby do odczytania przesyłanej informacji. Ta *unikalność* kubitów, niosących informację użyteczną, wraz z nieuchronnością zmian stanów kwantowych w wyniku pomiaru, prowadzi nas w kierunku praktycznych zastosowań techniki kwantowej w cyberbezpieczeństwie.

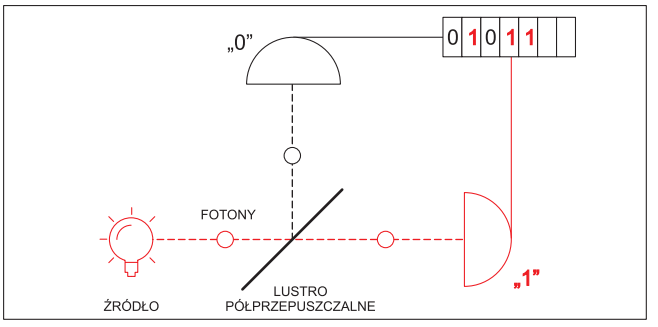
WYBRANE ROZWIĄZANIA TECHNIKI KWANTOWEJ

Technika kwantowa umożliwia budowanie praktycznych rozwiązań przez zastosowanie specyficznych cech mechaniki kwantowej. Oczywiście każde powszechnie wykorzystywane narzędzie, od smartfonu do zwykłej zaciskarki do przewodów RJ45 – jeśli będziemy je analizowali dostatecznie szczegółowo – jest obiektem kwantowym i funkcjonuje zgodnie z zasadami mechaniki kwantowej. Nie są one jednak rozwiązaniami techniki kwantowej. Obecnie potrafimy konstruować niewiele takich rozwiązań, których działanie bezpośrednio opiera się na zasadach przedstawionych w poprzednim rozdziale. Wśród nich warto wymieć generatory liczb prawdziwie losowych oraz komputery kwantowe.

Generatory liczb losowych

Liczby losowe mają zastosowanie w wielu dziedzinach, a od ich jakości nierzadko zależy bezpieczeństwo danych wrażliwych i krytycznych. Przykładem jest konieczność generowania losowych ciągów binarnych, które będą użyte jako klucze podczas szyfrowania danych przy użyciu szyfrów symetrycznych. Jeżeli klucze te nie będą ciągami losowymi, poufne dane będą narażone na ujawnienie. Stąd ograniczone zaufanie do popularnych obecnie generatorów liczb pseudolosowych dostępnych w bibliotekach programistycznych. Znacznie bezpieczniej jest zastosować generator liczb prawdziwie losowych, opierając się na efektach kwantowych.

Przykładowe rozwiązanie tego typu zostało przedstawione na rys. 1. Jeśli założymy, że lustro półprzepuszczalne z takim samym prawdopodobieństwem odbija oraz przepuszcza każdy z fotonów, jesteśmy w stanie zbudować generator losowych



■ **Rys. 1. Schemat generatora liczb losowych [3]**

ciągów binarnych. Wystarczy ustawić detektory pojedynczych fotonów, które wykryją, czy dana cząstka odbiła się od lustra, czy została przepuszczona. Jedno ze zdarzeń wygeneruje binarne zero, a drugie binarną jedynkę. Zatem rozpatrzmy bity kwantowe w stanie:

$$\frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle,$$

czyli takie, które znajdują się z takim samym prawdopodobieństwem w obu stanach podstawowych (0 oraz 1). Wystarczy wykonywać pomiary takich fotonów wzdłuż osi określającej stan ortogonalny i zapisywać, który z dwóch możliwych wyników został zaobserwowany, a każdej z możliwości przypisać binarne zero lub jedynkę. W ten sposób powstanie generator liczb prawdziwie losowych, oparty na efektach kwantowych.

Tego typu rozwiązania są już dostępne na rynku. Ich ceny zaczynają się od kilku tysięcy złotych. Przykładowe generatory liczb losowych, których zasada działania opiera się na zasadach mechaniki kwantowej, można znaleźć choćby na stronie internetowej ID Quantique [3].

Komputery kwantowe

Innym przykładem są komputery kwantowe – rozwiązanie bardzo intensywnie rozwijane w ostatnich latach. Z technicznego puntu widzenia są to urządzenia przetwarzające bity kwantowe zgodnie z wybranym algorytmem. Dwa z takich algorytmów kwantowych mają bezpośredni i bardzo istotny wpływ na współcześnie stosowane zabezpieczenia w cyberprzestrzeni [4].

W roku 1994 Peter Shor przedstawił kwantowy algorytm faktoryzacji o wielomianowej złożoności czasowej [5]. Faktoryzacja, czyli rozkładanie liczby całkowitej na iloczyn liczb pierwszych, odgrywa ważną rolę w kryptografii. Bezpieczeństwo popularnego szyfru RSA opiera się właśnie na trudności faktoryzacji dużych liczb. W skład klucza publicznego RSA wchodzi liczba, będąca iloczynem dwóch liczb pierwszych, a jeśli tylko kryptoanalityk będzie w stanie wykonać faktoryzację tej liczby – w prosty sposób pozna klucz prywatny, za pomocą którego będzie mógł deszyfrować poufne wiadomości bądź wykonać podpis cyfrowy. Dlaczego więc pomimo prawie trzydziestu lat od opracowania algorytmu Shora kryptografia asymetryczna – w tym szyfr RSA – wciąż jest stosowana w sieci Internet? Odpowiedź jest prosta: ponieważ kwantowy algorytm faktoryzacji zadziała jedynie na komputerze kwantowym. Pomimo że pierwsze komputery kwantowe zostały już skonstruowane, ich możliwości obliczeniowe są niewystarczające, aby zagrozić współcześnie stosowanym kluczom RSA (obecnie o długości co najmniej 2048 bitów).

Inny algorytm kwantowy – algorytm Grovera [6] – służy do przeszukiwania nieuporządkowanych baz danych. Aby wyszukać element w nieuporządkowanym zbiorze o M elementach, musimy wykonać średnio $M/2$ prób, sprawdzając jeden element po drugim. Jednak kwantowy algorytm Grovera potrafi rozwiązać ten problem jedynie w $\sim\sqrt{M}$ krokach. Tak znaczne przyspieszenie powoduje łatwiejsze przeprowadzanie kryptoanalizy. Zagrożenie jest realne również dla współczesnych szyfrów symetrycznych.

Czy jednak mamy powody do obaw w związku z prowadzonymi pracami nad rozwojem komputerów kwantowych? Zdania są wyraźnie podzielone. Z jednej strony grupy badaczy twierdzą, że już wkrótce opracowany zostanie komputer kwantowy, operujący na kilku tysiącach kubitów, który zagrozi bezpieczeństwu współcześnie używanych kluczy asymetrycznych [7]. Inni wskazują na zbyt poważne problemy z dekoherencją, aby w najbliższym czasie tego typu urządzenia mogły powstać i efektywnie wykonywać skomplikowane obliczenia. Główna trudność polega na wykonywaniu operacji na stanach kwantowych, które muszą być izolowane od wpływu otoczenia. Znacznie łatwiej jest przesyłać informację kwantową na odległość. Dlatego kolejny rozdział został poświęcony rozwiązaniu, które już teraz stosowane jest w praktyce i może stać się alternatywą dla obecnie stosowanej kryptografii asymetrycznej. Mowa o kwantowej dystrybucji kluczy kryptograficznych.

KRYPTOGRAFIA KWANTOWA

Kryptografia kwantowa jest połączeniem kryptografii symetrycznej z kwantową dystrybucją kluczy. Poufne dane mogą być szyfrowane za pomocą typowych szyfrów symetrycznych, np. **AES** (*Advanced Encryption Standard*). Z kolei symetryczne klucze kryptograficzne, których używają te algorytmy, są przesyłane przy użyciu stanów kwantowych.

Obecnie stosowane algorytmy do wymiany lub ustalania kluczy kryptograficznych, takie jak np. algorytm Diffie-Hellmana, mają wspólną wadę: są podatne na różnego rodzaju ataki sieciowe. A jeśli tylko intruz zdobędzie klucz kryptograficzny, wszystkie dane zaszyfrowane za jego pomocą mogą zostać odczytane. Jednak jeśli zastosujemy technikę kwantową, będziemy mogli bezpiecznie ustalić tajny klucz kryptograficzny przy użyciu bitów kwantowych, chroniąc się przed podsłuchem [8]. Zapewnią to przedstawione już zasady mechaniki kwantowej: pomiar zaburza stan kwantowy, a nieznanego stanu przed pomiarem nie można sklonować.

Kwantowa dystrybucja kluczy

Obecnie istnieje wiele algorytmów umożliwiających kwantową dystrybucję kluczy kryptograficznych. Algorytmy te można podzielić na dwie grupy, różniące się sposobem przesyłania informacji kwantowej.

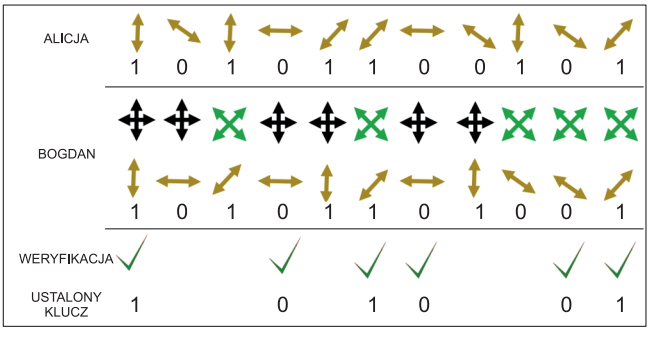
- Pierwsza grupa to rozwiązania, w których do przesłania informacji używa się pojedynczych stanów kwantowych, np. odpowiednio spolaryzowanych fotonów.
- Druga grupa – rzadziej stosowana w praktyce – to algorytmy oparte na stanach splątanych.

Obecnie protokół BB84 [9], najstarszy protokół kwantowej dystrybucji kluczy, jest jednocześnie najpopularniejszym rozwiązaniem stosowanym w urządzeniach realizujących usługę kryptografii kwantowej. Zgodnie z nim, kolejne bity informacji przesyłanej między użytkownikami są kodowane za pomocą jednej z czterech możliwych polaryzacji. Z kolei odbiorca, aby odczytać taką informację, stosuje jeden z dwóch typów detektorów.

Jeśli dwóch użytkowników (nazwijmy ich: Alicja i Bogdan), chce ustalić między sobą tajny klucz kryptograficzny za pomocą

protokołu BB84, oboje muszą zdefiniować jednoznacznie, które z polaryzacji fotonów będą odpowiadały poszczególnym bitom. Załóżmy, że foton o polaryzacji 0° lub foton o polaryzacji -45° będzie reprezentował binarne zero, natomiast fotony o polaryzacjach 90° i 45° będą nośnikami binarnych jedynek. Za pomocą takiego kodowania Alicja może wysłać do Bogdana pewien ciąg bitów. Bogdan, odbierając fotony od Alicji, oczywiście nie wie, w jakiej polaryzacji jest konkretny foton, więc w sposób losowy będzie je odbierał albo za pomocą bazy prostej, która jest w stanie odczytać jednoznacznie jedynie fotony o polaryzacji 0° i 90°, albo za pomocą bazy ukośnej, która z kolei jest w stanie odczytać poprawnie jedynie fotony o polaryzacjach: –45° i 45°.

Z tak zarejestrowanych fotonów Bogdan otrzyma pewien ciąg bitów, jednak nie będzie on identyczny z ciągiem bitów, który został wygenerowany przez Alicję. Oba ciągi będą zgadzały się ze sobą w miejscach, w których Bogdan poprawnie wybrał bazę. Pomiary takie dają jednoznaczne wyniki tylko wtedy, gdy mierzymy polaryzację wzdłuż osi określającej dany stan ortogonalny. Pozostałe pomiary dają wyniki losowe. Zatem wystarczy w sposób zupełnie jawny poinformować Alicję, jaka baza została wybrana do odczytania polaryzacji każdego z fotonów. Alicja wskazuje miejsca, w których baza została wybrana poprawnie i tym samym określa, z których bitów zbudowany zostanie klucz kryptograficzny. Przykład użycia protokołu BB84 przedstawiono na rys.2. W podanym scenariuszu pierwszy bit został odebrany poprawnie: polaryzacja fotonu 90° zostanie jednoznacznie odczytana jedynie za pomocą bazy prostej. Jednak drugi z bitów może zostać jednoznacznie poprawnie odczytany jedynie za

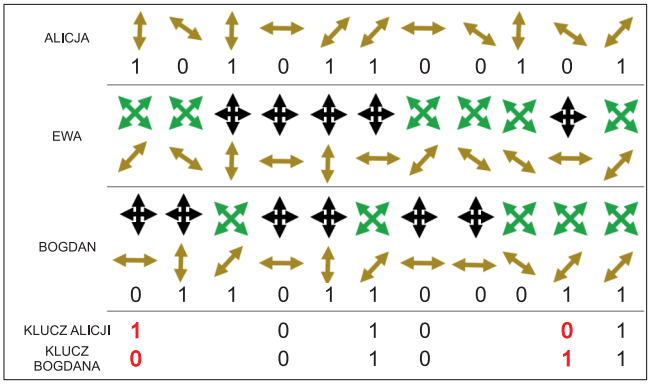


■ **Rys. 2. Przykład użycia protokołu BB84 w celu ustalenia klucza. Symbol ✚ oznacza pomiar w bazie prostej, a symbol ✕ oznacza pomiar w bazie ukośnej. Strzałki reprezentują polaryzację poszczególnych fotonów.**

pomocą bazy ukośnej, ale Bogdan wybrał w tym przypadku bazę prostą. Zatem w sposób losowy może otrzymać albo foton o polaryzacji 0° (czyli bitowe 0), albo foton o polaryzacji 90° (czyli bitowe 1). Ten bit należy więc odrzucić — nie może wchodzić w skład klucza kryptograficznego.

Dlaczego jednak protokół BB84 jest tak bezpieczny? Załóżmy, że w trakcie ustalania klucza kryptograficznego osoba niepowołana (nazwijmy atakującego Ewą) próbuje podsłuchać komunikację pomiędzy Alicją i Bogdanem. Taka sytuacja przedstawiona została na rys. 3. Aby poznać polaryzację fotonów określającą dany bit, Ewa musi użyć bazy prostej lub ukośnej (podobnie jak Bogdan). Oczywiście również wybiera bazę w sposób losowy, gdyż nie wie, która będzie odpowiednia w danej chwili.

Jeżeli Ewa wybierze nieodpowiednią bazę, zmieni w ten sposób polaryzację fotonu – w myśl zasady, że pomiar zaburza stan kwantowy. Taka sytuacja jest przedstawiona na rys. 3 z pierwszym fotonem. Po dokonaniu podsłuchu foton dociera



■ **Rys. 3. Próba podsłuchiwania klucza kryptograficznego w protokole BB84**

do Bogdana, który – nawet jeśli wybierze bazę poprawnie (biorąc pod uwagę pierwotną polaryzację fotonu wysłanego przez Alicję) – i tak może uzyskać inny bit, niż rzeczywiście został wysłany. Wystarczy więc porównać część ustalonego ciągu bitów, aby jednoznacznie przekonać się, że w trakcie ustalania klucza Ewa podsłuchiwała w kanale kwantowym, więc klucz kryptograficzny nie jest poufny.

Ewa nie jest w stanie odczytywać polaryzacji fotonów nie narażając się na zmianę ich wartości. Nie jest również w stanie sklonować nieznanego stanu kwantowego. Dzięki temu nie uda się jej podsłuchać przesyłanych danych w sposób niezauważony. Jednak kwantowa dystrybucja klucza to tylko pierwszy z etapów ustalania ostatecznego klucza kryptograficznego. Kolejne są ściśle związane z błędami w kluczu.

Błędy w ustalonym kluczu

Komunikację w kanale kwantowym cechuje wysokie prawdopodobieństwo przekłamania przesyłanych bitów. Jeśli do tego dodamy trudności z precyzyjnym kształtowaniem stanów kwantowych w nadajniku oraz niedoskonałości detektorów w odbiorniku, otrzymamy komunikację, w której błąd transmisji będzie pojawiał się znacznie częściej niż w tradycyjnych systemach teleinformatycznych. Bitowa stopa błędów w systemach z komunikacją kwantową – nazywana **QBER** (*Quantum Bit Error Rate*) – obliczana jest jako stosunek liczby błędów do całkowitej liczby bitów i zwykle wynosi kilka procent (czyli średnio kilka procent bitów klucza jest przekłamanych). Ponieważ QBER zależy od parametrów fizycznych nadajnika, kanału kwantowego oraz odbiornika, można przyjąć, że ma stałą wartość dla danego systemu. Jeśli więc poziom QBER gwałtownie wzrasta, oznacza to, że dodatkowe błędy generuje intruz podsłuchujący w kanale kwantowym. Porównując bieżącą liczbę błędów do liczby błędów, jaka zwykle powinna być w kanale transmisyjnym bez podsłuchu, użytkownicy decydują, czy przesłany klucz jest bezpieczny, czy też komunikacja była podsłuchiwana.

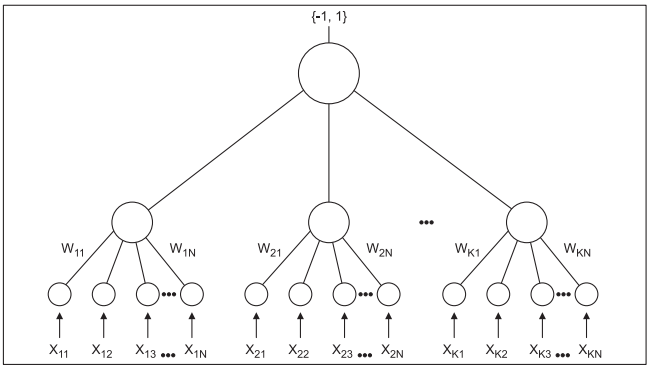
Do estymacji bitowej stopy błędów oczywiście wybiera się jedynie część ustalanego klucza. Ponieważ użyty w tym celu fragment musi zostać ujawniony, nie może już wejść w skład ostatecznego klucza. Dlatego bity użyte do estymacji QBER są kasowane. Niezwykle ważną kwestią, z punktu widzenia bezpieczeństwa i wydajności, jest odpowiedni wybór liczby bitów, na podstawie których estymowany będzie parametr QBER [10].

Jeśli poziom QBER wskazuje na to, że w kanale kwantowym bity nie zostały podsłuchane, rozpoczyna się etap korekcji błędów. Obecnie najpopularniejszym sposobem stosowanym

w praktyce jest kontrola parzystości poszczególnych bloków klucza. Nadawca i odbiorca sprawdzają parzystość wybranych fragmentów klucza, a wynik porównują, korzystając z kanału publicznego. Jeśli parzystość danego bloku się nie zgadza, blok ten dzielony jest na mniejsze części i cały proces powtarza się, aż błąd zostanie znaleziony. Aby w ten sposób nie zdradzać podsłuchującemu informacji poufnych, część bitów jest kasowana. Klucz jest skracany od razu, albo później w procesie zwanym *privacy amplification.* Etap ten również powoduje zmniejszenie długości ustalanego klucza. Z tego względu odpowiedni wybór poszczególnych parametrów na każdym z etapów ustalania klucza jest bardzo ważny z punktu widzenia bezpieczeństwa, jak i wydajności całego systemu. Dotyczy to także wyboru samego algorytmu korekcji błędów. Z tego względu warto stosować takie, które zdradzają atakującemu najmniej informacji o kluczu, jednocześnie nie powodując drastycznego obniżenia wydajności.

KOREKCJA BŁĘDÓW PRZY UŻYCIU SZTUCZNYCH SIECI NEURONOWYCH

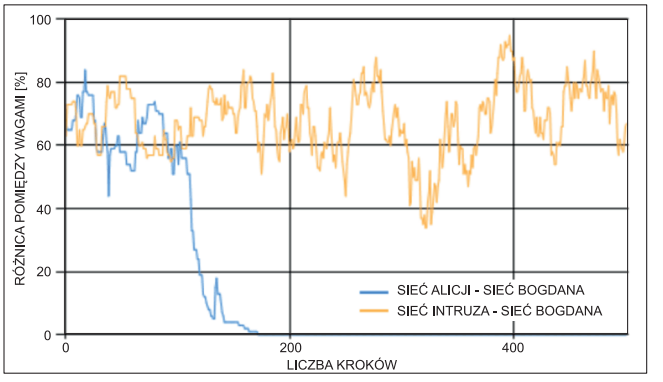
Metoda korekcji błędów oparta na synchronizacji sztucznych sieci neuronowych została zaproponowana w roku 2019 [11]. Założono w niej, że zarówno Alicja, jak i Bogdan generują własną sztuczną sieć neuronową o strukturze **TPM** (*Tree Parity Machine*). Schemat typowej sieci TPM przedstawiono na rys. 4.



■ **Rys. 4. Sztuczna sieć neuronowa o strukturze TPM**

Architektura obu sieci – zarówno u Alicji, jak i Bogdana – musi być podobna. Użytkownicy ustalają liczbę neuronów w warstwie ukrytej (K) oraz liczbę wejść każdego z neuronów w tej warstwie (N), a także zakres możliwych wag, charakteryzujących każde z połączeń między warstwą wejściową a warstwą ukrytą (L). Jedyną różnicą są wartości poszczególnych wag ($w_{11}, \dots, w_{KN}$). Alicja powinna uzależnić je od swojej wersji klucza, a Bogdan od własnej, czyli tej, która zawiera błędy związane z transmisją w kanale kwantowym. Tym sposobem obie sieci będą bardzo podobne, lecz pozostaną niewielkie różnice w miejscach występowania błędów. Aby poprawić te błędy, użytkownicy powinni użyć wybranego algorytmu uczenia maszynowego, wśród których największą skutecznością odznacza się algorytm Hebbian. Wzmacnia on te wagi, które są zgodne z wartością wygenerowaną przez całą sieć (wartość 1 lub –1), podczas stymulacji losowo wygenerowanym wektorem binarnym ($X_{11}, \dots, X_{KN}$).

Proces synchronizacji obu sieci przy użyciu algorytmu uczenia maszynowego odbywa się jedynie wtedy, gdy wartość wyjściowa sieci Alicji jest taka sama, jak wartość wyjściowa sieci Bogdana (przy założeniu, że są one stymulowane takim samem wektorem binarnym na wejściu). Zasada ta ma ogromne znaczenie również dla bezpieczeństwa, ponieważ intruz, który chciałby synchronizować własną sieć, musi czekać na moment, w którym aż trzy sieci generują jednakową wartość wyjściową. Dzięki temu sieć atakującego synchronizuje się znacznie wolniej, niż sieci Alicji i Bogdana. Przykładowy scenariusz przedstawiono na rys. 5. Sieci TPM Alicji i Bogdana zostały w pełni zsynchronizowane po niespełna 200 krokach, a intruz nie był w stanie tego zrobić nawet po 500 próbach.



■ **Rys. 5. Przykładowy proces synchronizacji sieci TPM**

Alicja i Bogdan byliby w stanie wykonać pełną synchronizację swoich sieci neuronowych, nawet jeśli początkowe wartości wag byłyby wybrane losowo. Jednak w przypadku, gdy wagi będą odzwierciedlały ciągi binarne użytkowników otrzymane w procesie kwantowej dystrybucji kluczy, cały proces będzie przebiegał znacznie szybciej. Wagi w sieciach Alicji i Bogdana będą bowiem już bardzo podobne. Proces uczenia sprowadzi się jedynie do zsynchronizowania tych kilku procent bitów, które były wynikiem zakłóceń w kanale kwantowym. To z kolei jeszcze bardziej zwiększy poziom bezpieczeństwa, gdyż intruz będzie miał mniejszą szansę na zsynchronizowanie własnej sieci. Liczne symulacje potwierdzają, że synchronizację sieci TPM w celu korekcji błędów cechuje wysoki poziom bezpieczeństwa i jednocześnie nie wymaga ona od użytkowników tak istotnego skracania klucza, jak w przypadku kontroli parzystości (choć etap *privacy amplification* wciąż jest rekomendowany).

KIERUNKI ROZWOJU

Kwantowa dystrybucja kluczy potwierdza, że technika kwantowa jest źródłem nowych rozwiązań w zakresie cyberbezpieczeństwa, które nie są jedynie ciekawymi doświadczeniami myślowymi, ale mogą być stosowane w praktyce. Już od lat 90. ubiegłego wieku zaczęły pojawiać się pierwsze systemy realizujące kwantową dystrybucję kluczy. Wszystkie te rozwiązania miały architekturę typu punkt-punkt, gdzie dwa urządzenia – nadające i odbierające informację kwantową – komunikowały się ze sobą za pomocą dedykowanego łącza. Jednak już na początku XXI wieku zaczęto tworzyć rozwiązania umożliwiające zastosowanie kryptografii w środowisku sieciowym czy w komunikacji satelitarnej [12]. Wciąż jednak poważnym wyzwaniem jest wrażliwość stanów kwantowych na dekoherencję. Problemem jest nie tylko przechowywanie i przetwarzanie kubitów w środowisku odizolowanym od wpływów zewnętrznych, ale także zakłócenia w kanale kwantowym. Ograniczają

one maksymalny dystans pomiędzy Alicją i Bogdanem, powodując wzrost liczby błędów w ustalanym kluczu. Z tego powodu prowadzone są intensywne badania w kierunku opracowania urządzeń, które będą w stanie wzmocnić przesyłany sygnał bez konieczności jego odczytywania. Obecnie największe nadzieje budzą rozwiązania oparte na teleportacji kwantowej przy użyciu splątanych fotonów.

Innym ciekawym rozwiązaniem, choć jak na razie niestosowanym w praktyce, jest kodowanie supergęste. Umożliwia ono przesłanie informacji zawartej w dwóch bitach klasycznych przy użyciu pojedynczego bitu kwantowego. Używając tego protokołu, Alicja i Bogdan mogą znacznie zwiększyć wydajność kanału telekomunikacyjnego. I to rozwiązanie opiera się na efekcie splątania, zastosowaniu wybranych operacji kwantowych i wykonaniu pomiaru kubitów.

Emocje wzbudza także pojęcie *internetu kwantowego*. Pomimo tego, że technika kwantowa nie jest jeszcze na takim poziomie rozwoju, aby już teraz zbudować działającą sieć kwantową na wzór sieci Internet, naukowcy starają się definiować i rozwiązywać problemy związane z kolejnymi etapami rozwoju takiej sieci [13]. Wśród wyzwań warto wymienić pamięć kwantową, w której będzie można przechowywać bity kwantowe i zapisywać wyniki działań poszczególnych operacji kwantowych.

* * *

Kodowanie informacji przy użyciu stanów kwantowych pojedynczych cząstek elementarnych – choć brzmi nieprawdopodobnie – jest dziś możliwe. Tego typu rozwiązania działają nie tylko w specjalistycznych laboratoriach, ale są dostępne komercyjnie dla klientów. Ich największe zalety widać przede wszystkim w zastosowaniach związanych z cyberbezpieczeństwem. Pora więc odpowiedzieć na pytanie zadane w tytule artykułu.

Komputery kwantowe z całą pewnością są zagrożeniem dla cyberbezpieczeństwa w ujęciu globalnym. Dziś trudno sobie wyobrazić działanie usług elektronicznych bez szyfrów asymetrycznych, dzięki którym protokół **TLS** (*Transport Layer Security*) zapewnia poufność w sieci Internet. Jeśli kwantowy algorytm szybkiej faktoryzacji umożliwi efektywne łamanie obecnie stosowanych kluczy asymetrycznych, skutki będą niezwykle dotkliwe. Czy zespołom pracującym nad komputerami kwantowymi uda się pokonać trudności – przede wszystkim dekoherencję bitów kwantowych – czas pokaże.

Z drugiej strony technika kwantowa wspiera nas w zapewnianiu poufności przez możliwość ustalania kluczy symetrycznych. Ogromną zaletą jest możliwość wykrycia podsłuchów pasywnych. Dzięki zasadom mechaniki kwantowej intruz nie jest w stanie bezkarnie odczytać informacji kwantowej. Poza tym, dzięki technice kwantowej, możemy korzystać z generatorów liczb prawdziwie losowych. Możliwe, że w niedalekiej przyszłości pojawią się kolejne praktyczne zastosowania. Wydajna komunikacja korzystająca z kodowania supergęstego czy teleportacja kwantowa to jedynie wybrani kandydaci. Zatem trudno polemizować z odpowiedzią, że technika kwantowa to zarówno zagrożenie, jak i ogromna szansa dla cyberbezpieczeństwa.

LITERATURA

[1] Le Bellac M. 2011. *Wstęp do informatyki kwantowej*. Wydawnictwo Naukowe PWN.

[2] Wootters W.K., W. H. Zurek. 1982. *A single quantum cannot be cloned*. Nature, 299:802–803.

[3] ID Quantique: https://www.idquantique.com [Dostęp: 24.06.2022].

[4] Hirvensalo M.. 2004. *Algorytmy kwantowe*. Wydawnictwo Szkolne i Pedagogiczne.

[5] Shor. P.W. 1994. *Algorithms for quantum computation: discrete logarithms and factoring*. 35th Annual Symposium on Foundations of Computer Science, DC, USA.

[6] Grover L.K. 1996. *A fast quantum mechanical algorithm for database search*. 28th Annual ACM Symposium on Theory of Computing, NY, USA.

[7] IBM: Expanding the IBM Quantum roadmap to anticipate the future of quantum-centric supercomputing: https://research.ibm.com/blog/ibm-quantum-roadmap-2025 [Dostęp: 24.06.2022].

[8] Niemiec M. 2009. „Kryptografia kwantowa – bezpieczeństwo na najwyższym poziomie". *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, 82, 6.

[9] Bennett C.H., G. Brassard. 1984. *Public key distribution and coin tossing*. IEEE International Conference on Computers, Systems, and Signal Processing, Indie.

[10] Niemiec M., A. R. Pach. 2013. „Management of security in quantum cryptography". *IEEE Communications Magazine*, 51, 8.

[11] Niemiec M. 2019. *Error correction in quantum cryptography based on artificial neural networks.* Quantum Information Processing, 18, 174.

[12] Mehic M., M. Niemiec, S. Rass, J. Ma, M. Peev, A. Aguado, V. Martin, S. Schauer, A. Poppe, C. Pacher, M. Voznak. 2020. *Quantum Key Distribution: A Networking Perspective*. ACM Computing Surveys, 53, 5.

[13] Wehner S., D. Elkouss, R. Hanson. *Quantum internet: A vision for the road ahead*. Science, 362, 6412.

Sławomir HAUSMAN*

Techniczne aspekty oddziaływania na ludzi fal radiowych w systemach komórkowych 5G i 6G

Technical aspects of the impact of radio waves in 5G and 6G cellular systems on humans

Przewiduje się, że w najbliższych latach systemy piątej generacji (5G) będą szybko rozwijającą się i dominującą technologią sieci komórkowych. Nieuniknione jest przy tym zagęszczenie sieci stacji bazowych, aby sprostać oczekiwaniom w zakresie jakości usług. Pojawiają się obawy społeczne dotyczące nowych zakresów częstotliwości oraz wypadkowego poziomu promieniowania pola elektromagnetycznego, pochodzącego od wszystkich źródeł tego promieniowania łącznie. Jednocześnie trwają już prace nad systemami 6G, które jeszcze bardziej będą się różnić od obecnych systemów pod względem wykorzystywanych zakresów częstotliwości (w tym subterahercowych) oraz szybkozmiennych charakterystyk promieniowania anten mMIMO. Będzie je cechować również inne zarządzanie widmem częstotliwości i łączami radiowymi, nieprzypominające systemów 1G/2G/3G/4G/5G, np. terminale będą się komunikować jednocześnie z wieloma punktami nadawczo-odbiorczymi (*Transmit/Receive Points* – **TRP**). Prawdopodobnie przestanie być używany termin *stacja bazowa* i *komórka* – systemy 6G bywają określane jako bezkomórkowe (*Cell-Free*). Użycie pasm subterahercowych i ich własności propagacyjne będą implikowały konieczność stosowania inteligentnych powierzchni odbijających (*Intelligent Reflecting Surface* – **IRS**) zarówno na zewnątrz, jak i wewnątrz budynków. To znacznie zmieni scenariusze propagacyjne i wpłynie na analizę ekspozycji. W artykule przedstawiono niektóre techniczne uwarunkowania w systemach 5G i 6G, które mogą wpływać na ocenę ekspozycji na promieniowanie pola elektromagnetycznego. Ocena wpływu fal radiowych na organizmy żywe jest domeną medycyny i nie będzie tu bezpośrednio omawiana.
Słowa kluczowe: systemy komórkowe, 5G, 6G, oddziaływanie fal radiowych na środowisko, zagrożenie dla zdrowia ludzi, modelowanie propagacji fal radiowych

Fifth-generation (5G) systems are expected to be a rapidly growing and dominant mobile network technology in the coming years. It is inevitable that the network of base stations will become denser to meet expectations in terms of quality of service. As a result, there are public concerns about the new frequency ranges and the total level of electromagnetic field radiation from all sources of this radiation combined. At the same time, work is already underway on 6G systems that will differ even more from current systems in terms of the frequency ranges used (including sub-Terahertz) and the rapidly varying radiation characteristics of mMIMO antennas. They will also feature a different management of frequency spectrum and radio links that will not resemble 1G/2G/3G/4G/5G systems, e.g. terminals will be able to communicate simultaneously with multiple Transmit/Receive Points (TRPs). It is likely that the terms "base station" and "cell" will no longer be used - 6G systems are sometimes referred to as Cell-Free. The use of sub-terahertz bands and their propagation properties will imply the need for Intelligent Reflecting Surfaces (IRS) both outdoors and indoors. This will significantly change propagation scenarios and affect exposure analysis. This article presents some of the technical considerations in 5G and 6G systems that may affect the assessment of exposure to electromagnetic field radiation. Assessment of the effects of radio waves on living organisms and is the domain of medicine and will not be directly discussed here.
Key words: cellular systems, 5G, 6G, environmental impact of radio waves, human health risks, radio wave propagation modeling

DOI: 10.15199/59.2022.4.2

Większość badań, dotyczących wpływu fal radiowych wytwarzanych przez technologie mobilne na zdrowie ludzi, dotyczyła technologii 2G i 3G. Obecnie weszliśmy już w erę implementacji systemów 5G, nie bez problemów, choćby z zarządzaniem widmem częstotliwości. Od dawna dostępne są liczne publikacje dotyczące wyników badań wpływu fal radiowych, wytwarzanych przez systemy 2G i 3G, na zdrowie ludzi i środowisko. Nie ma jednak konsensusu co do tego, jak interpretować dotychczasowe wyniki i jakie są ich implikacje dla systemów 5G, a w przyszłości 6G, nad którym już trwają prace rozwojowe i standaryzacyjne [1, 2, 3, 4].

Dla nowych i przyszłych generacji systemów istnieje jeszcze niewiele wyników badań naukowych nad wpływami biologicznymi, a jednocześnie (a może właśnie dlatego) obawy społeczne są silne. Do jednej z bardzo obszernych prac w zakresie metaanalizy oddziaływania systemów 5G należy [5]. Przeanalizowano tam (powołując się na 280 pozycji literatury) aktualne miary/metryki, przepisy dla ekspozycji 5G, biorąc pod uwagę ocenę najnowszych wytycznych takich organizacji, jak *Institute of Electrical and Electronics Engineers* (**IEEE**), *International Commission on Non-Ionizing Radiation Protection* (**ICNIRP**), *International Telecommunication Union* (**ITU**), *International Electrotechnical Commission* (**IEC**) oraz *United States Federal Communications Commission* (**FCC**), a także przepisy obowiązujące w ponad 220 krajach. Rozważono zarówno wpływ stacji bazowych 5G (*Node-B Base Stations gNBs),* jak i urządzeń końcowych **UE** (*User Equipment*). Porównano dopuszczalne poziomy gęstości mocy, natężenia pola i swoistego tempa pochłaniania energii **SAR** (*Specific Absorption Rate*), które są zdefiniowane przez organizacje międzynarodowe (np. IEEE, ICNIRP) i prze-

* Politechnika Łódzka, Instytut Elektroniki, e-mail: slawomir.hausman@p.lodz.pl

pisy krajowe (np. FCC), uwzględniając najnowsze wytyczne z lat 2019-2020 oraz wcześniejsze. Przyjęte wartości dopuszczalne okazują się podobne zarówno dla różnych wytycznych, jak i przepisów obowiązujących w różnych krajach. Tak więc, w skali globalnej, zróżnicowanie wartości uznanych za bezpieczne dla ludzi nie jest duże. W tym miejscu warto zauważyć, że większość rządów europejskich (a także poza Europą) polega na gremiach naukowych, które zakładają, że podstawowe i udokumentowane zagrożenie dla zdrowia ze strony systemów komórkowych, i innych źródeł promieniowania w zakresie radiowym, stanowi ogrzewanie tkanek ciała. W konsekwencji tego dopuszczalne poziomy promieniowania są ustalane tak, aby zapobiec szkodliwemu nagrzewaniu tkanek (szczególnie we wrażliwym na wzrost temperatury narządzie wzroku, narządach rozrodczych i mózgu). Jeśli te poziomy są przestrzegane, to uważa się, że nie ma zagrożenia dla zdrowia. Skutki nietermiczne kontra wyłącznie termiczne są osią sporu w zakresie oceny oddziaływań niejonizujących fal elektromagnetycznych na zdrowie ludzi i środowisko. Ten spór nie daje się obecnie jednoznacznie rozstrzygnąć. W dużym stopniu wynika to stąd, że wyniki badań dotyczących oddziaływań nietermicznych są zwykle nieistotne statystycznie, sprzeczne, opracowane na nieadekwatnych modelach komórkowych lub zwierzęcych itd. Ten obszar badań wymaga istotnego wzmocnienia.

W Polsce wartości dopuszczalne określa [6]. Rozpatrując sytuację ogólnoświatową, warto dostrzec, że wiele krajów nie określiło jeszcze wartości dopuszczalnych i ogólnie nie planuje formalnie wdrożenia systemów 5G. Niektóre państwa, np. Szwajcaria (a nawet poszczególne miasta, np. Bruksela w Belgii) poddały zaś ekspozycję ścisłym i (relatywnie) restrykcyjnym regulacjom. W tym przypadku bardziej rygorystyczne wartości dopuszczalne niż rekomendowane przez ICNIRP [7] i skuteczne wdrażanie sieci 5G to dwa sprzeczne cele. Dlatego warto się przyglądać rozwojowi systemów komórkowych w tych lokalizacjach, aby ocenić, czy bardziej restrykcyjne przepisy zapewnią uzyskanie jakości usług takiej, jak gdzie indziej i jaki będzie tego rachunek ekonomiczny.

Oddzielny problem to monitorowanie pól elektromagnetycznych. Sposób monitorowania pól, nawet jedynie w krajach Unii Europejskiej, jest zróżnicowany [8]. Na wcześniejszych etapach rozwoju sieci (1G – 4G) było ono stosunkowo proste. Większe wyzwanie stanowi pomiar uwzględniający instalacje 5G z antenami mMIMO z kształtowaniem wiązki. Praktycznym rozwiązaniem jest wymuszenie dużego (najlepiej maksymalnego) obciążenia transmisją danych i nakierowania wiązki/wiązek na urządzenie pomiarowe oraz rejestracja wyniku jako superpozycji pól pochodzących od wszystkich systemów. W Polsce odbywa się to zgodnie z [9]. Pomiary są prowadzone przez jednostki, mające akredytację *Polskiego Centrum Akredytacji* (**PCA**), co zapewnia ich wiarygodność.

Monitorowanie pól umożliwia weryfikację zgodności z obowiązującymi przepisami, ale nie stanowi rozwiązania problemu obaw społecznych. Tu potrzebne jest informowanie i komunikacja ze społeczeństwem. W Polsce został wdrożony system powszechnego dostępu obywateli do informacji o instalacjach oraz wynikach pomiarów PEM w otoczeniu stacji bazowych telefonii komórkowej o nazwie SI2PEM (System informacyjny o instalacjach wytwarzających promieniowanie elektromagnetyczne), opracowany przez Instytut Łączności – Państwowy Instytut Badawczy wspólnie z Ministerstwem Cyfryzacji (*https://si2pem.gov.pl*). Każdy może dowiedzieć się, jakie są wyniki pomiarów pól na interesującym go obszarze. Można to uznać za element kontroli społecznej nad źródłami fal radiowych.

PASMA CZĘSTOTLIWOŚCI DLA SYSTEMÓW NOWYCH GENERACJI

Do niedawna komunikacja komórkowa z powodzeniem wykorzystywała pasma do 6 GHz, czasem określane w literaturze anglosaskiej jako pasma *sub-6 GHz*. Charakteryzują się one korzystnymi własnościami propagacyjnymi w obszarach wysoko zurbanizowanych, ze względu na relatywnie małe straty propagacji w materiałach budowlanych. Część tego pasma, np. zakresy 700-800 MHz (tzw. pasmo zasięgowe), jest szczególnie ważna, wręcz niezbędna dla uzyskania ciągłości zasięgu w trudnych warunkach propagacyjnych i przy większych odległościach od stacji bazowych. Jednak transmisja danych w sieciach komórkowych rośnie wykładniczo od ponad dekady i oczekuje się, że ta tendencja utrzyma się w przyszłości. Wraz z rosnącą liczbą urządzeń typu smartfon, modemy czy bezprzewodowe urządzenia IoT w obecnych i przewidywanych zastosowaniach, takich jak logistyka (łańcuchy dostaw), opieka zdrowotna, transport i komunikacja samochodowa (w tym pojazdy autonomiczne), oczekuje się, że ten trend będzie umacniał się. W tym kontekście jedynym możliwym dziś rozwiązaniem jest wykorzystanie wyższych częstotliwości. Ilość dostępnego widma na falach milimetrowych jest bardzo duża w porównaniu z częstotliwościami sub-6 GHz (~50 – 100 razy). Umożliwiłoby to osiągnięcie o rząd lub dwa rzędy wielkości większej przepływność danych. Wdrożenia 5G są we wczesnej lub zaawansowanej fazie (zależnie od kraju). Jednocześnie już pojawiające się aplikacje, takie jak rozszerzona rzeczywistość (w tym holograficzna) czy immersyjna rzeczywistość rozszerzona, będą wymagały łączy o przepływności rzędu kilku terabitów na sekundę (Tbit/s), których nie zdołają obsłużyć systemy 5G (gdzie dostępna szerokość pasma jest mniejsza niż 10 GHz). Spowodowało to duże zainteresowanie eksploracją pasm terahercowych w celu uzupełnienia pasm poniżej 6 GHz oraz fal milimetrowych w systemach 6G i kolejnych. Mamy już pewne doświadczenie w wykorzystywaniu zakresu 6–300 GHz do różnych zastosowań niekomórkowych, takich jak komunikacja satelitarna, radioastronomia, teledetekcja, radary. Ze względu na postępy w technice antenowej i układów półprzewodnikowych, obecnie stało się możliwe wykorzystanie tego zakresu również do komunikacji mobilnej. Pasmo częstotliwości od 30 do 300 GHz (co odpowiada długości fali od 10 do 1 mm) nazywane jest pasmem fal milimetrowych i zapewnia o jeden lub nawet dwa rzędy wielkości większą szerokość pasma w porównaniu z pasmami poniżej 6 GHz.

Duża tłumienność (silna absorpcja) podczas propagacji przez obiekty budowalne (nawet lekkie ściany działowe) jest podstawową trudnością w implementacji systemów komunikacyjnych, wykorzystujących fale milimetrowe i subterahercowe. Jednocześnie jednak te same efekty pomogą zmniejszać zakłócenia wspólnokanałowe w systemach komórkowych nowych generacji, które charakteryzują się gęstym rozmieszczeniem komórek. Dodatkową korzyścią jest lepsze przestrzenne zwielokrotnianie częstotliwości i zwiększone bezpieczeństwo danych ze względu na wąskie wiązki anten, pracujących w zakresie fal milimetrowych (możliwość uzyskania dużych wartości zysku kierunkowego przy niewielkich wymiarach

anten). Częstotliwości od około 24 GHz do około 100 GHz są już przewidziane dla standardu 5G. Z myślą o przyszłych systemach 6G bada się dodatkowo zakres 0,1 – 10 THz.

Szerokie pasma dostępne w widmie fal milimetrowych czynią realną komunikację bezprzewodową z przepływnością wielu Gbit/s, otwierając możliwości wprowadzania innowacyjnych zastosowań, na przykład bezprzewodowych połączeń dosyłowych typu *backhaul* pomiędzy zewnętrznymi stacjami bazowymi. Zapewni to zmniejszenie kosztów wdrażania sieci nowych generacji (np. instalacji i utrzymania światłowodów, w tym kanalizacji kablowej), zwłaszcza w przypadku przyszłych sieci ultragęstych. Ponadto umożliwi przekształcenie obecnych przewodowych centrów danych w całkowicie bezprzewodowe centra z serwerami komunikującymi się radiowo z wykorzystaniem silnie kierunkowych tzw. „wiązek ołówkowych". Innym potencjalnym zastosowaniem jest komunikacja między pojazdami (V2V) w warunkach wysokiej mobilności, w tym w pociągach i samolotach. Ponadto widmo THz składa się z pasm o dostępnych szerokościach rzędu kilkudziesięciu GHz, które mogą obsługiwać przepływności do kilku Tbit/s. Dzięki temu będzie można obsługiwać aplikacje wymagające dużej przepustowości i niskich opóźnień, takie jak gry w wirtualnej rzeczywistości i wideokonferencje ultra-HD. Inne zastosowania, które skorzystają z dojrzałości komunikacji (sub)terahercowej, to komunikacja nanomaszyn, komunikacja w układach scalonych, internet nanorzeczy (IoNT) i bio-nanorzeczy [10]

Ze względu na zróżnicowaną charakterystykę propagacji kanału oraz tłumienie atmosferyczne zależne od częstotliwości, zidentyfikowano pewne pasma w widmie sub-THz/THz, które szczególnie sprzyjają zastosowaniom komunikacyjnym. Na światowej konferencji radiokomunikacyjnej (WRC) 2015 ITU opublikował listę proponowanych pasm częstotliwości w zakresie 24 – 86 GHz do globalnego wykorzystania w systemach 5G. Wybór tych pasm został dokonany po rozważeniu wielu czynników, takich jak charakterystyka propagacji kanału, globalne porozumienia oraz dostępność przyległego pasma. WRC-2019 koncentrowało się na warunkach przydziału fal milimetrowych przeznaczonych dla systemów 5G/6G. Łącznie zidentyfikowano 17,25 GHz widma [11]. Na potrzeby realizacji przyszłych systemów łączności THz, w ramach WRC 2019, zidentyfikowano również łącznie 160 GHz widma w pasmie THz z zakresu od 252 do 450 GHz. Prace będą kontynuowane w ramach WRC-23 i WRC-27.

W przypadku pasm subterahercowych na użytek 6G pojawia się interesująca i ważna kwestia, czy licencjonowanie częstotliwości będzie w ogóle potrzebne [12, 13].

Chociaż pasma subterahercowe i terahercowe mają ogromny potencjał wykorzystania ich w łączności, to istnieją znaczne wyzwania w zakresie ich komercyjnego zastosowania. W szczególności fale w tych pasmach podlegają większym stratom propagacyjnym w środowiskach zurbanizowanych i w konsekwencji mają krótsze zasięgi. Te wyzwania utrudniały do tej pory włączenie pasm fal subterahercowych do standardów i ich wdrożenia komercyjne. Wraz z postępem w nowoczesnych technologiach anten i urządzeń, wykorzystanie tych pasm do komunikacji staje się możliwe.

Systemy 6G będą łączyć w sobie funkcje komunikacji bezprzewodowej i detekcji (jest to m.in. jeden z kierunków badań finansowanych przez Unię Europejską w ramach programów Horyzont Europa, np. konkursu HORIZON-JU-SNS-2022), umożliwiając to na częstotliwościach subterahercowych nie tylko w celu zapewnienia komunikacji danych z przewidywaną przepływnością rzędu kilku Gbit/s i niskimi opóźnieniami (poniżej 1 ms), ale także jednoczesne monitorowanie środowiska za pomocą radarów i metod lokalizowania urządzeń użytkownika z dokładnością większą niż 1 cm we wszystkich trzech wymiarach. Miałoby to szczególne znaczenie w logistyce i sterowaniu pojazdami latającymi typu dron. Połączenie detekcji i szybkiej transmisji danych w jednym systemie umożliwi np. bardziej efektywną realizację kontroli i zarządzania wysoce zautomatyzowanymi środowiskami przemysłowymi, w tym robotami kroczącymi i cobotami (*Collaborative Robots*), komunikację M2M, wykorzystanie AR i VR, w tym w zastosowaniach sportowych do wzbogacania doznań kibiców (głównie AR), w pojazdach autonomicznych itp.

Takie systemy będą się nadawały do wdrożenia w złożonych środowiskach wewnątrzbudynkowych, np. w fabrykach, gdzie wiele punktów dostępowych typu **TPR** (*Transmit/Receive Points*) tworzy bezkomórkowy (*Cell-Free*) obszar usług, w ramach którego jest zrealizowane centralne zarządzanie przeniesieniami obsługi. Zastosowanie scentralizowanego podejścia koordynacyjnego (w porównaniu z podejściem rozproszonym) z optymalnym przetwarzaniem **MMSE** (*Minimum Mean Square Error*) okazało się nie tylko maksymalizować wydajność widmową, ale także redukować sygnalizację w porównaniu do standardowego podejścia rozproszonego [14]. Takie podejście do projektowania zapewnia bardzo dużą redundancję propagacyjną i szybkościową, a w konsekwencji bardzo dużą niezawodność potrzebną w środowisku przemysłowym.

Możliwość kształtowania bardzo wąskich wiązek anten w pasmach subterahercowych zapewnia zarówno zwiększanie zasięgu, jak i oszczędzanie energii przez kierowanie jej tylko w wymaganym kierunku (lub kilku kierunkach). Jednocześnie zwiększa to bezpieczeństwo danych przez utrudnienie podsłuchu. W środowisku z licznymi przeszkodami konieczne jest zapewnienie dodatkowych ścieżek propagacji, omijających przeszkody, co można zrealizować stosując inteligentne powierzchnie odbijające, opisane w dalszej części artykułu. Warto zauważyć, że w takich środowiskach występuje sytuacja, w której albo dana ścieżka propagacji jest, albo jej nie ma, bo natrafiła na przeszkodę. To tworzy nowy paradygmat, w którym aby zapewnić niezawodność kanału, należy skonstruować jak najwięcej ścieżek binarnych kanałów *on-off* pomiędzy wieloma przestrzennie różnymi nadajnikami a terminalem, w nadziei że przynajmniej niektóre z tych ścieżek nie są stłumione („wyłączone").

SYSTEMY 5G

Implementacja systemów 5G charakteryzuje się między innymi zwiększeniem liczby stacji bazowych (BS) wewnątrz i na zewnątrz budynków w środowisku wysoko zurbanizowanym, z wykorzystaniem różnego rodzaju lokalizacji (makro, mikro i femto). Można się też spodziewać dużej liczby sieci prywatnych/kampusowych, zwłaszcza do realizacji usług w obiektach przemysłowych i logistyce.

Badania prowadzone na świecie (także w Polsce) pokazują, że natężenia pola lub gęstości mocy są typowo znacznie mniejsze, niż wynikające z uwarunkowań prawnych. Jest tak nawet w lokalizacjach, w których współistnieje jednocześnie kilka

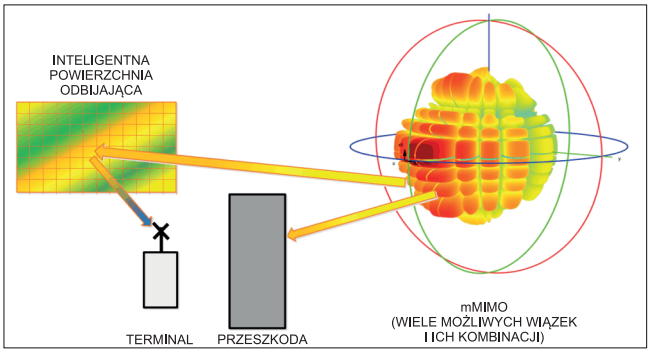
systemów różnych generacji i różnych operatorów. Metodologia oceny parametrów pól dla 5G powinna uwzględniać sterowane/przełączane charakterystyki anten mMIMO, metody dostępu do medium, zmienną w funkcji czasu liczbę i rozmieszczenie możliwych jednoczesnych wiązek. Biorąc pod uwagę bardzo dużą liczbę możliwych kombinacji tych czynników, można przypuszczać, że najbardziej odpowiednie okaże się statystyczne podejście do modelowania pól [15].

Ocena ta jest przeprowadzana drogą symulacji, w celu przewidywania poziomów odbieranego sygnału i obliczania powiązanych gęstości mocy dla typowego pojedynczego miejsca w ramach wielu konfiguracji dla każdej technologii. Badania przeprowadza się za pomocą przyrządów pomiarowych w celu walidacji i kalibracji wyników symulacji. Wyniki pokazują, że całkowity wskaźnik narażenia był bardzo niski w porównaniu z międzynarodowymi standardowymi limitami narażenia, a wyniki w [15] pokazują wkład każdej technologii dla tych szczególnych miejsc wykorzystanych w badaniu.

SYSTEMY 6G

Od pierwszych sieci komórkowych, które zostały przetestowane w latach siedemdziesiątych XX w., byliśmy świadkami kilku rewolucji w technice bezprzewodowej. Od 1G do 4G wzrost ruchu był uzyskiwany przez wykorzystywanie szerszych pasm, nowych metod modulacji, kodowania i wielodostępu oraz stopniowego zagęszczania sieci. Nawet po wprowadzeniu rewolucyjnej techniki 5G postęp się nie zatrzymał. Ma to związek choćby z przewidywanym rozwojem potrzeb w zakresie Internetu rzeczy. Międzynarodowy Związek Telekomunikacyjny (ITU) oszacował, że liczba podłączonych urządzeń IoT wzrośnie do 38,6 mld do 2025 roku i 50 mld do 2030 roku [16]. Różne zaawansowane techniki, takie jak agregacja nośnych, komunikacja wieloantenowa, a także nowe techniki modulacji, zostały już wykorzystane w kontekście obecnych generacji sieci komórkowych. Inne możliwe rozwiązania w celu spełnienia nowych oczekiwań to opracowanie lepszych technik przetwarzania sygnału dla poprawy efektywności spektralnej kanału, ekstremalne zagęszczenie sieci komórkowych. Szanse na uzyskanie znaczącej poprawy jakości usług z tych technik są znikome. Potrzebne są nowe, szersze pasma częstotliwości. Jednym z pierwszych nowych pasm dla systemów 6G może być pasmo 110 – 170 GHz.

Massive MIMO (*Multiple-Input Multiple-Output*) jest kluczową technologią 5G, która wykorzystuje wieloelementowe matryce antenowe w celu zapewnienia bardzo dużego zysku z formowania wiązki i przestrzennej multipleksacji użytkowników, a tym samym zwiększa wydajność spektralną i energetyczną. Stanowi to scentralizowane rozwiązanie o możliwościach ograniczonych przez zakłócenia międzykomórkowe nieodłącznie związane z jej „komórkowo-centryczną" konstrukcją. Przewiduje się, że systemy 6G będą miały rozproszone sterowanie mMIMO. Te cechy, w połączeniu ze skalowalnością systemu, będą odróżniać przyszłe sieci Cell-Free Massive MIMO od wcześniejszych generacji systemów bezprzewodowych. Pierwsze prace nad standaryzacją systemów 6G podjął już między innymi Międzynarodowy Związek Telekomunikacyjny (ITU). Firma Samsung opublikowała raport [17].

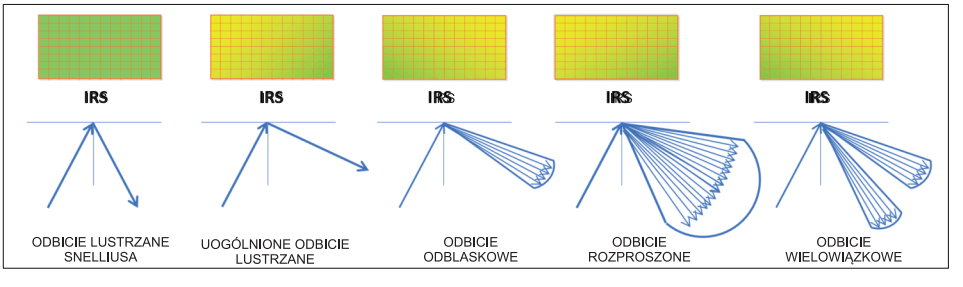


■ **Rys. 2. Ominięcie przeszkody na drodze propagacji przy zastosowaniu inteligentnej powierzchni odbijającej (IRS)**

Jak już wspomniano wcześniej, użycie pasm subterahercowych, ze względu na ich własności propagacyjne, będzie implikowało konieczność stosowania inteligentnych powierzchni odbijających **IRS** (*Intelligent Reflecting Surface*) zarówno na zewnątrz, jak i wewnątrz budynków (rys. 1 i 2).

To znacznie zmieni scenariusze propagacyjne i wpłynie na analizę ekspozycji ludzi na fale radiowe. Być może nie uda się zachować obecnych regulacji prawnych w zakresie metod pomiaru.

Warto podkreślić, że rozporządzenie Ministra Zdrowia [6], określając wartości dopuszczalne, obejmuje obecnie pasmo do 300 GHz i w tym sensie wykorzystanie pasma *D* (110 – 140 GHz) nie wymagałoby (przynajmniej formalnie) nowych uregulowań. Nie jesteśmy jednak jeszcze przygotowani do opracowania metodyki pomiarów pól przy – z założenia bardzo skomplikowanych – scenariuszach transmisji w systemach 6G. Prawdopodobnie konieczne będzie wykorzystanie metod symulacyjnych typu Monte Carlo i kalibracji oraz walidacji modeli przy użyciu metod empirycznych. Warto jednak zauważyć, że obecne metody modelowania propagacji, np. wykorzystujące śledzenie/wystrzeliwanie promieni, nie są przystosowane do analizy odbić od inteligentnych powierzchni odbijających. Tu niezbędne są nowe badania i rozwój oprogramowania użytkowego. Wydaje się również, że wartości graniczne zostały, nieco sztucznie, ekstrapolowane do 300 GHz, co może być słusznie kwestionowane. Nie ma bowiem wyczerpujących wyników badań odziaływania tego zakresu na ludzi i środowisko.



■ **Rys. 1. Różne charakterystyki odbicia przy zastosowaniu inteligentnych powierzchni odbijających (IRS)**

* * *

Badanie oddziaływania fal radiowych na ludzi i środowisko zdaje się nie nadążać za szybkim rozwojem systemów komórkowych. Nowe zakresy częstotliwości, zagęszczanie sieci stacji

bazowych, anteny mMIMO o zmiennych charakterystykach promieniowania, nowe metody modulacji i wielodostępu stwarzają już obecnie w tym zakresie nowe wyzwania. Prowadzone zgodnie z obecną metodyką badania odziaływania fal radiowych nie dadzą się wykonać w krótkim czasie. Ich przeprowadzenie będzie wymagać raczej dekad niż pojedynczych lat, zwłaszcza jeśli chce się poznać ewentualne oddziaływania długookresowe. Jednocześnie pojawiają się obawy społeczne, dotyczące nowych zakresów częstotliwości oraz wypadkowego poziomu promieniowania pola elektromagnetycznego, pochodzącego od wszystkich źródeł tego promieniowania łącznie. Chęć przyspieszenia rozwoju cywilizacyjnego i zasada ostrożności przy wdrażaniu nowych technologii nie dają się ze sobą łatwo pogodzić. Jeszcze nie wdrożono szeroko systemów 5G, a już trwają prace nad systemami 6G, które będą się jeszcze bardziej różnić od obecnych systemów, choćby pod względem wykorzystywania nowych zakresów częstotliwości (w tym subterahercowych), szybkozmiennych charakterystyk promieniowania anten mMIMO i inteligentnych powierzchni odbijających. To znacznie zmieni scenariusze propagacyjne i wpłynie na analizę ekspozycji. Wdrażanie systemów 5G, a zwłaszcza 6G (spodziewane ok. roku 2030) będzie wymagało wprowadzenia nowych metod pomiarowych i symulacyjnych, choćby właśnie w zakresie uwzględnienia inteligentnych powierzchni odbijających. Zmienna w czasie kombinacja działania takich powierzchni oraz anten typu mMIMO będzie stanowiła nowe wyzwanie w zakresie standaryzacji pomiarów pól elektromagnetycznych dla ochrony ludzi i środowiska przed ich skutkami.

LITERATURA

[1] Hoffmann M., M.Uusitalo, M.-H. Hamon, B. Richerzhagen, G. D'Aria, A. Gati, D. Lopez et al. "Expanded 6G vision, use cases and societal values – including aspects of sustainability, security and spectrum" Hexa-X Deliverable D1.2, 30/04/2021.

[2] Ericson M., H. Flinck, P. Vlacheas, S. Wänstedt et al. "Initial 6G Architectural Components and Enablers". Hexa-X Deliverable 5.1, 31/12/2021.

[3] Roth K., N.Michailow, A. Pärssinen, H. Halbauer et al. "Towards Tbps, Communications in 6G: Use Cases and Gap Analysis". Hexa-x Deliverable D2.1, 30/06/2021.

[4] Wymeersch H.et al. "6D Localisation and sensing use cases and gap analysis". Hexa-X Deliverable D3.1, 31/12/2021.

[5] Chiaraviglio L., A. Elzanaty and M. -S. Alouini, "Health Risks Associated With 5G Exposure: A View From the Communications Engineering Perspective", *IEEE Open Journal of the Communications Society*, vol. 2, pp. 2131-2179, 2021, doi: 10.1109/OJCOMS.2021.3106052.

[6] Rozporządzenie Ministra Zdrowia z dnia 17 grudnia 2019 r. w sprawie dopuszczalnych poziomów pól elektromagnetycznych w środowisku (Dz. U. z 2019 r. poz. 2448).

[7] ICNIRP, ''Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz)'', *Health Phys*., vol. 118, no. 5, pp. 483–524, 2020.

[8] Instytut Łączności – Państwowy Instytut Badawczy, *Analiza sposobów monitorowania PEM w wybranych krajach europejskich*, Warszawa 2018.

[9] Rozporządzenie Ministra Klimatu i Środowiska z dnia 15 grudnia 2020 r. w sprawie zakresu i sposobu prowadzenia okresowych badań poziomów pól elektromagnetycznych w środowisku (Dz.U. 2020 poz. 2311).

[10] Al-Zubi M.M., A. S. Mohan, P. Plapper and S. H. Ling, "Intra-Body Molecular Communication via Blood-Tissue Barrier for Internet of Bio-Nano Things," in IEEE Internet of Things Journal, 2022, doi: 10.1109/JIOT.2022.3182150.

[11] International Telecommunication Union (ITU), *Final Acts World Radiocommunication Conference 2019(WRC-19)*, ITU Publications, 2019.

[12] Standeford D. (2020). *6G raises novel spectrum and technical issues.* PolicyTracker, www. PolicyTracker.com

[13] Kurner T. and A. Hirata, "On the impact of the results of WRC 2019 on THz communications,", Third International Workshop on Mobile Terahertz System, IEEE, July 2020, str. 1–3.

[14] Bjornson E. and L. Sanguinetti, "Making Cell-Free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun*., vol. 19, no. 1, pp. 77–90, Jan. 2020.

[15] Pawlak R., P. Krawiec, J. Żurek, *On measuring Electromagnetic Fields in 5G technology*, IEEE Access 2019; 7: str. 29826 - 29835.

[16] Report ITU-R M (07/2015) IMT traffic estimates for the years 2020 to 2030.

[17] 6G Spectrum: "Expanding the Frontier", *Samsung Research*, 2020, w którym zakreśla swoją wizję systemów 6G

[18] Emilio Calvanese Strinati, George C. Alexandropoulos, Vincenzo Sciancalepore, Marco Di Renzo, Henk Wymeersch, Dinh-Thuy Phan-huy, Maurizio Crozzoli, Raffaele D'Errico, Elisabeth De Carvalho, Petar Popovski, Paolo Di Lorenzo, Luca Bastianelli, Mathieu Belouar, Julien Etienne Mascolo, Gabriele Gradon , Sendy Phang , Geoffroy Lerosey, Benoıt Denis "Wireless Environment as a Service Enabled by Reconfigurable Intelligent Surfaces: The RISE-6G Perspective" 2021 JointEuropean Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)

[19] Yuanwei Liu, Xiao Liu, Xidong Mu, Tianwei Hou, Jiaqi Xu, Marco Di Renzo, Naofal Al-Dhahir "Reconfigurable Intelligent Surfaces: Principles and Opportunities" *IEEE communications surveys & tutorials*, vol. 23, No. 3, Third quarter 2021

Wojciech BURAKOWSKI[1], Andrzej BĘBEN[1], Maciej SOSNOWSKI[1], Dariusz WIĘCEK[2], Igor MICHALSKI[2], Józef WOŹNIAK[3], Krzysztof GIERŁOWSKI[3], Michał HOEFT[3], Patryk SCHAUER[4], Arkadiusz WARZYŃSKI[4], Marek NATKANIEC[5], Piotr BORYŁO[5], Artur BINCZEWSKI[6], Bartosz BELTER[6], Maksymilian FURMANN[6]

Planowane krajowe laboratorium badawcze sieci i usług 5G wraz z otoczeniem

5G national laboratory concept: 5g networks, services, measurement tools and applications

Przedstawiono koncepcję infrastruktury PL-5G pt. *Krajowe laboratorium sieci i usług 5G wraz z otoczeniem*. Laboratorium to będzie ważną infrastrukturą badawczą w kraju, która powinna w istotny sposób przyczynić się do ukierunkowania praktycznych badań dotyczących nowych generacji dla infrastruktur sieciowo-obliczeniowych Internetu. Laboratorium to będzie zlokalizowane w czterech miastach, tj. w Gdańsku, Krakowie, Poznaniu i Wrocławiu i będzie miało wspólny system sterowania eksperymentami. Korzystanie z laboratorium będzie zdalne, co umożliwi zespołom krajowym z innych lokalizacji efektywne korzystanie z jego zasobów. Przewiduje się, że będzie również wykorzystywane w procesie dydaktycznym realizowanym przez uczelnie techniczne. Termin zakończenia budowy laboratorium przewiduje się na koniec 2023 r.
Słowa kluczowe: laboratorium sieci i usług 5G, zasady korzystania, struktura laboratorium, system zarządzania eksperymentami

The paper describes concept of PL-5G infrastructure entitled: *National laboratory of 5G networks and services jointly with measurement tools and applications*. The laboratory will be important element for improving research on new generation infrastructure for the Internet. It will be located in 4 cities, i.e. in Gdańsk, Kraków, Poznań and Wrocław. Common control system will manage all running experiments. Access to the laboratory will be remote to allow all researchers from Poland for effective usage of resources. The laboratory will be also usable for improving didactic process in technical universities. The laboratory will be open at the end of 2023.
Key words: 5G laboratory, access rules, structure of laboratory, control system for managing experiments

DOI: 10.15199/59.2022.4.3

Artykuł podsumowuje pierwszą połowę realizacji projektu POIR pt. *PL-5G: Krajowe laboratorium sieci i usług 5G wraz z otoczeniem*. Celem projektu jest udostępnienie zespołom badawczym oraz przemysłowi w pełni wyposażonego laboratorium, umożliwiającego prowadzenie badań praktycznych dotyczących techniki 5G. Technika 5G jest ważnym krokiem w kierunku unowocześnienia infrastruktury sieciowo-obliczeniowej w sieciach operatorów.

W artykule przedstawiono opracowaną w ramach realizacji projektu wizję takiego laboratorium, które powinno zacząć funkcjonować pod koniec 2023 roku. Dotychczas, tj. na koniec czerwca 2022 r., zdefiniowano jego topologię oraz wyposażenie, które będzie sukcesywnie kupowane z uwzględnieniem sformalizowanych procedur przetargowych.

Istotne pytanie, na które należało odpowiedzieć przed złożeniem wniosku, brzmiało: po co nam taki projekt w kraju? W końcu zespoły krajowe mogły uczestniczyć w projektach europejskich dotyczących techniki 5G i tam zdobyć odpowiednie doświadczenie, by wykorzystać je do rozpowszechnienia wiedzy w kraju. Również zespoły krajowe mogły korzystać ze zbudowanych już laboratoriów sieci 5G w innych krajach europejskich. Niestety, udział tych zespołów w specyfikacji techniki 5G był mniej niż skromny. Wynikało to głównie z trzech przyczyn: (1) decydującą rolę w specyfikacji zakresu tematycznego projektów europejskich miał przemysł europejski, a więc firmy takie, jak Telefonica, Nokia, Ericsson, France Telecom itd.; (2) dobór zespołów badawczych do realizacji danego projektu był ukierunkowany przez przedstawicieli przemysłu; (3) składane wnioski projektowe kierowane przez zespoły krajowe zazwyczaj nie uzyskiwały dostatecznie wystarczających ocen, czego efektem był brak finansowania. Ostatecznie w kraju nie było wystarczającej wiedzy dotyczącej techniki 5G i tym bardziej możliwości prowadzenia badań praktycznych związanych z tą techniką, w tym efektywnego włączenia tematyki techniki 5G do procesu dydaktycznego na uczelniach technicznych. Jednocześnie technika 5G jest techniką na tyle zestandaryzowaną, że wiele firm już jakiś czas temu zaczęło oferować kupno sprzętu i oprogramowania 5G, a operatorzy telekomunikacyjni zaczęli unowocześniać swoje sieci o te elementy. W efekcie obecnie istnieje wiele firm działających na terenie kraju, które oferują, głównie operatorom, swoje rozwiązania 5G.

Konsorcjantami projektu POIR *PL-5G: Krajowe laboratorium sieci i usług 5G wraz z otoczeniem* są: (1) Politechnika Warszaw-

[1] Politechnika Warszawska {wojtek, abeben, m.sosnowski}@tele.pw.edu.pl,
[2] Instytut Łączności – Państwowy Instytut Badawczy, Wrocław {d.wiecek, i.michalski}@il-pib.pl
[3] Politechnika Gdańska, jowoz@eti.pg.edu.pl, kg@kg.gda.pl, michal.hoeft@pg.edu.pl,
[4] Politechnika Wrocławska {patryk.schauer, arkadiusz.warzynski}@pwr.edu.pl,
[5] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie {natkanie, piotr.borylo}@agh.edu.pl,
[6] Poznańskie Centrum Superkomputerowo-Sieciowe {artur, bartosz.belter, mfurmann}@man.poznan.pl.

ska – lider, (2) Instytut Łączności – Państwowy Instytut Badawczy, (3) Politechnika Gdańska, (4) Politechnika Wrocławska, (5) Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie oraz (6) Instytut Chemii Bioorganicznej PAN Poznańskie Centrum Superkomputerowo-Sieciowe. Powyższe organizacje już od blisko dekady wspólnie realizują ważne projekty krajowe, m.in. ostatnio projekt PL-LAB2020. Obecny projekt jest wymieniony w mapie krajowych strategicznych infrastruktur badawczych i jest przewidziany do realizacji na lata 2021–2023. Budżet projektu to ok. 200 mln złotych.

W rozdziale następnym zostanie przedstawiona krótka charakterystyka laboratorium, obejmująca architekturę sieci międzyregionalnej, sieci w poszczególnych lokalizacjach, sieć operacyjną oraz przewidywane do wykorzystania częstotliwości. W kolejnym zostaną pokrótce opisane wstępne zasady korzystania z laboratorium, przedostatni rozdział jest poświęcony omówieniu systemu zarządzania eksperymentami, ostatni zaś jest podsumowaniem artykułu.

CHARAKTERYSTYKA LABORATORIUM

Zaprojektowana infrastruktura badawcza jest zgodna z architekturą sieci 5G i będzie się składać z trzech powiązanych laboratoriów.

1. Laboratorium sieci 5G będzie zbudowane na podstawie rozwiązań producentów technologii 5G (bezprzewodowe i przewodowe sieci dostępowe, brzegowa chmura obliczeniowa, sieć szkieletowa oraz centralna chmura obliczeniowa) wraz z otwartymi interfejsami programowymi.

2. Laboratorium symulatorów i aparatury pomiarowej 5G będzie zawierać symulatory technologii 5G oraz aparaturę pomiarową związaną z rozwojem i utrzymaniem sieci 5G.

3. Laboratorium otoczenia 5G będzie zawierać urządzenia i oprogramowanie dla tworzenia rozwiązań sieciowych, platform i aplikacji opartych na technologii 5G (między innymi integracja z technologiami Internetu Rzeczy).

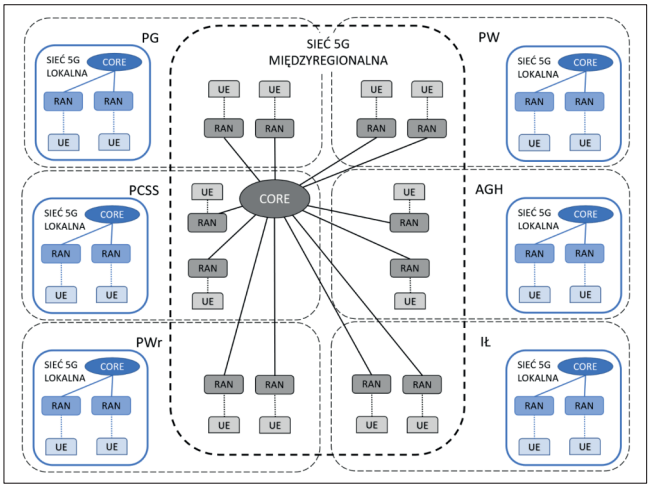
Infrastruktura PL-5G umożliwi prowadzenie badań zarówno w pojedynczym laboratorium, jak również w całym rozproszonym ekosystemie. Będzie to możliwe dzięki opracowaniu i wdrożeniu systemu zarządzania i monitorowania infrastruktury 5G w zakresie zarządzania zasobami, zarządzania eksperymentami i badaniami oraz zarządzania dostępem użytkowników. W szczególności zostanie zapewniona możliwość integracji z podobnymi laboratoriami w innych krajach UE.

Cele badawcze projektu zostały szczegółowo opisane w Agendzie Badawczej. Planowaną sieć badawczą 5G zaprojektowano w ten sposób, aby umożliwiała ona prowadzenie badań teoretyczno-eksperymentalnych dotyczących wszystkich funkcjonalności oferowanych przez sieć i usługi 5G. Zapewnienie powyższego umożliwiłoby zespołom krajowym wywieranie istotnego wpływu na udział w dalszych pracach nad rozwojem infrastruktury sieci 5G i jej właściwym wdrażaniem w sieciach operatorskich. Z infrastruktury tej powinny korzystać krajowe zespoły badawcze, producenci sprzętu i usług 5G oraz operatorzy sieci. Wiele wiodących firm działających w kraju wyraziło swoje zainteresowanie korzystaniem z tej infrastruktury.

Zakłada się, że dostęp do projektowanych laboratoriów będzie możliwy dla wszystkich krajowych zespołów badawczych oraz podmiotów przemysłowych. W tym celu zostanie opracowany regulamin dostępu do infrastruktury wraz ze wstępnym cennikiem.

Sieć międzyregionalna

W krajowym laboratorium wyróżnia się dwa rodzaje sieci 5G, którymi są sieć międzyregionalna 5G, która jest wspólna dla wszystkich lokalizacji oraz sieci lokalne 5G, umieszczone w poszczególnych lokalizacjach. Organizację sieci przedstawiono na rys. 1. Urządzenia końcowe z każdej lokalizacji są podłączone sieciami **RAN** (*Radio Access Network*) do jednej z wymienionych sieci. Intencją jest, aby sieć międzyregionalna była siecią stabilną i zbudowaną z użyciem podobnej klasy urządzeń, jak w sieciach operatorskich. Zadaniem sieci lokalnych jest przeprowadzanie eksperymentów badawczych i testowanie nowych rozwiązań, a więc sieci te powinno się budować opierając się na rozwiązaniach otwartych.

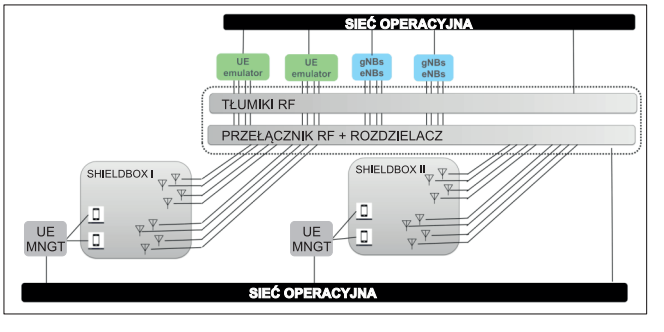


■ **Rys. 1. Sieć międzyregionalna i sieci lokalne**

Sieci w poszczególnych lokalizacjach

Laboratorium Politechniki Warszawskiej

Laboratorium ma zapewniać możliwość prowadzenia badań w środowisku złożonym z różnych implementacji rdzenia 5G, różnych sieci RAN oraz terminali i emulatorów użytkowników. Rdzeń sieci będzie uruchamiany w centrum danych, natomiast elementy wykorzystujące interfejs radiowy będą umieszczone w klatkach Faradaya (tzw. *shieldboxach*), umożliwiając pracę z częstotliwościami licencjonowanymi. Poszczególne *shieldboxy* będą zawierać terminale i głowice radiowe, które za pośrednictwem konfigurowalnego przełącznika i tłumika RF będą dołączone do stacji bazowych i emulatorów. Ich zadaniem będzie łączenie tych środowisk, zgodnie z założoną konfiguracją, bez konieczności ręcznej ingerencji w połączenia falowodowe, jak na rys. 2. W szczególności dany *shieldbox* można dołączyć do sieci międzyregionalnej.



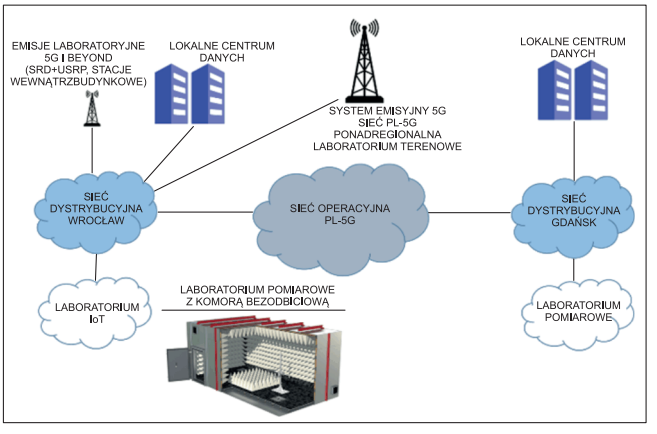
■ **Rys. 2. Elastycznie zestawiane środowisko radiowe**

Laboratorium ma umożliwić badanie technik przetwarzania na brzegu sieci (*Mobile Edge Computing* – **MEC**). W tym celu zostaną wykorzystane serwery MEC oraz otoczenie złożone z urządzeń multimedialnych i środowiska IoT, które będą źródłem danych i sygnałów do analizy. Dodatkowo laboratorium będzie miało zestaw urządzeń do badań nad językiem P4, wirtualizacją oraz zarządzaniem klastrem serwerów.

Specyfiką laboratorium PW jest jego lokalizacja w budynkach PCSS. Tym samym nie przewiduje się radiowych instalacji zewnętrznych, a duży nacisk kładziony jest na zapewnienie zdalnego dostępu do zasobów laboratorium.

Laboratorium Instytutu Łączności – PIB

Laboratorium będzie miało dwie lokalizacje – w oddziałach Instytutu we Wrocławiu i w Gdańsku. We Wrocławiu zostanie uruchomiona w pełni funkcjonalna, zintegrowana z siecią ponadregionalną, sieć emisyjna 5G, pracująca w zakresie FR1 na zewnątrz budynków, a także w formie wewnątrzbudynkowej sieć 5G w zakresach FR1, FR2 i w pasmach nielicencjonowanych, jak to przedstawiono na rys. 3.



■ **Rys. 3. Laboratoria oraz systemy sieci 5G uruchamiane w Instytucie Łączności – PIB**

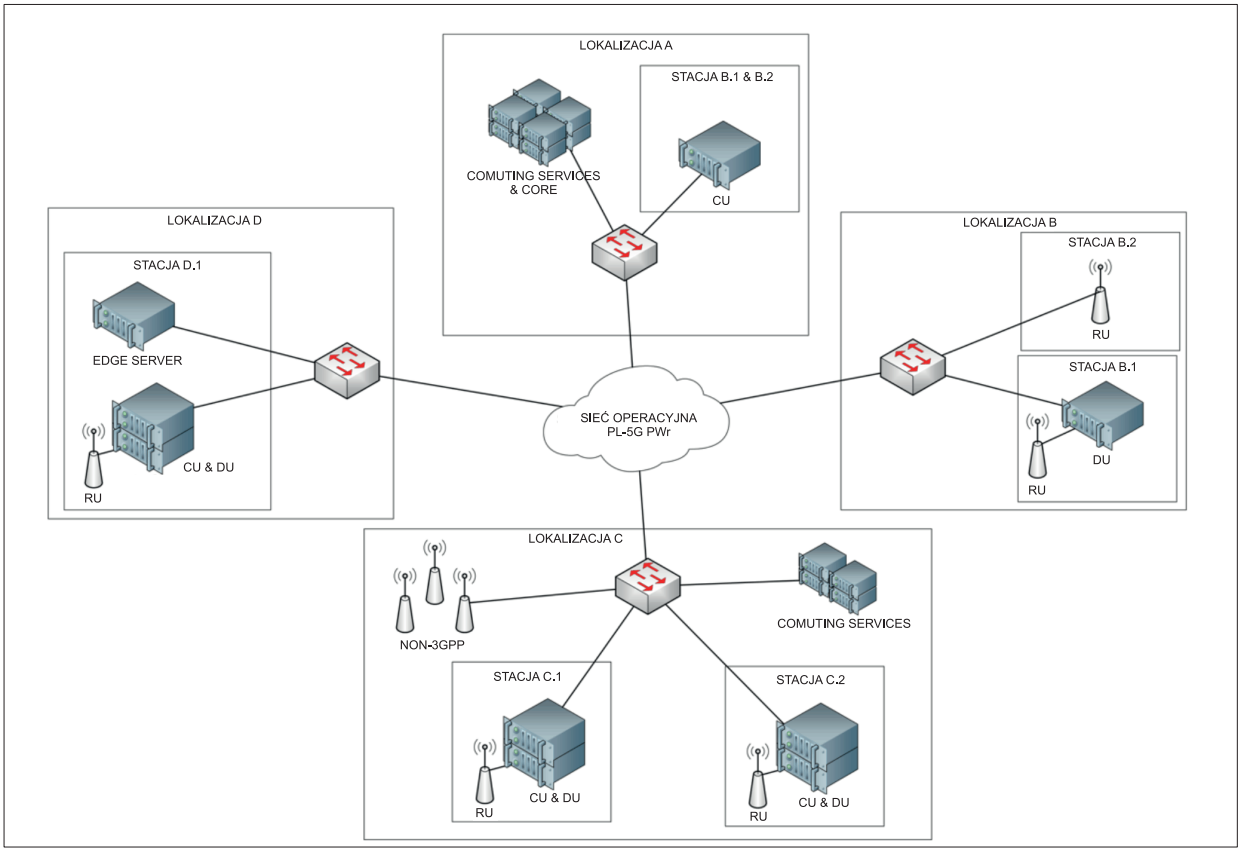
Jednym z głównych i specyficznych elementów laboratorium będzie budowana we Wrocławiu komora bezodbiciowa hybrydowa 10 m, przeznaczona do wykonywania badań kompatybilności elektromagnetycznej i badań systemów antenowych (SAC i FAR). W komorze będą mogły być wykonywane pomiary kompatybilności elektromagnetycznej, zakłóceń międzysystemowych, odporności i emisyjności urządzeń, a także badania charakterystyk systemów antenowych 5G czy parametrów urządzeń 5G (np. TRP, EIRP czy pomiary czułości urządzeń, badania współczynników ochronnych czy selektywności).

We Wrocławiu jest planowana również instalacja sieci zewnętrznej 5G w paśmie FR1, pokrywająca zasięgiem okoliczną część miasta, a także przenośne instalacje wewnętrzne w pasmach FR1, FR2 i nielicencjonowanych do celów badań w komorze bezodbiciowej i badań *beyond 5G* z wykorzystaniem np. elementów radia programowalnego SDR. Schemat koncepcyjny sieci PL-5G w IŁ-PIB przedstawiono na rys. 3.

Laboratorium Politechniki Wrocławskiej

Laboratorium umożliwi realizację rozbudowanych testów aplikacyjnych. Infrastruktura ma zapewnić zarówno demonstrację systemów usługowych w środowisku sieci 5G, jak również przeprowadzenie badań wydajności oraz możliwości integracji usług.

Ze względu na ograniczony obszar działania planowanej sieci, niezbędne jest udostępnienie odpowiednio zróżnicowanego środowiska przy ograniczonym obszarowo przemieszczaniu terminali. Kampus Politechniki Wrocławskiej oraz jego okolica umożliwia realizację wielu scenariuszy w terenie zarówno w przypadku



■ **Rys. 4. Laboratoria oraz systemy sieci 5G uruchamiane na Politechnice Wrocławskiej**

terminali użytkowników przemieszczających się powoli (piesi), jak i z większymi prędkościami (rowerzyści itp.). Ponadto będzie możliwe wydzielenie na potrzeby testów przestrzeni dla pojazdów zdalnie sterowanych, w tym autonomicznych, zarówno jeżdżących, jak i latających.

Elementy sieci 5G na Politechnice Wrocławskiej zostaną rozmieszczone w 4 lokalizacjach, jak pokazano na rys. 4. Lokalizacja A jest przeznaczona dla centrum obliczeniowego. Będą w niej znajdować się również elementy sieci dostępowej wspierające działanie stacji bazowych umieszczonych w lokalizacji B. W lokalizacji B będą umieszczone elementy stacji bazowych, w szczególności głowice radiowe oraz anteny. Stacje w lokalizacji B planowane są na pasmo FR1 oraz FR2.

Lokalizacja C będzie łączyła w sobie brzegowe centrum danych oraz stacje bazowe 5G. Stacje w tej lokalizacji planowane są na pasmo FR2. Ponadto znajdować się w niej będą urządzenia dostępowe typu non-3GPP, wykorzystujące protokoły z rodziny 802.11.

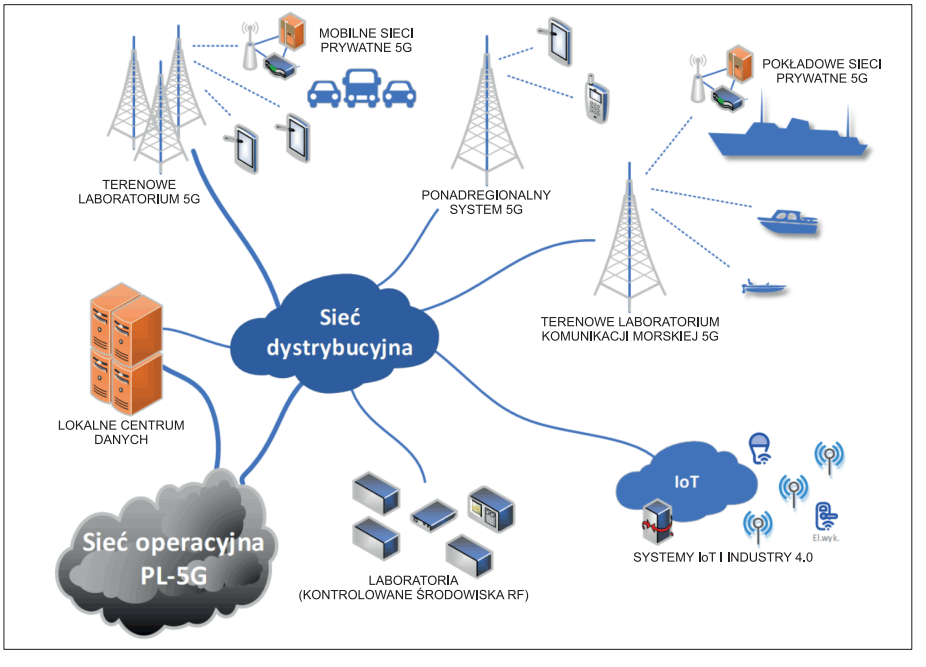
W lokalizacji D, oddalonej od głównego kampusu o ok. 3 km, zostanie umieszczona stacja bazowa na pasmo FR1. W niej dostępny będzie również serwer dla obliczeń brzegowych niewielkiej mocy.

Laboratorium Politechniki Gdańskiej

Laboratorium obejmuje zarówno węzeł ogólnokrajowej sieci ponadregionalnej, jak i systemy przeznaczone do realizacji zadań badawczych dotyczących specyficznych kierunków badań, tak jak to przedstawiono na rys. 5.

Część laboratoryjna budowanej infrastruktury obejmuje lokalne centrum danych oraz urządzenia radiowe umieszczone w kontrolowanych środowiskach propagacyjnych. W połączeniu z emulatorami urządzeń klienckich oraz urządzeniami pomiarowymi takie rozwiązanie umożliwia realizację bardzo szerokiego spektrum prac badawczych dotyczących mechanizmów i protokołów systemu 5G.

Politechnika Gdańska planuje też wdrożenie systemów terenowych, w tym sieci dostępowej 5G NR o wielokrotnym pokryciu oraz systemu przeznaczonego do badań dotyczących zastosowania rozwiązań 5G w środowisku morskim.



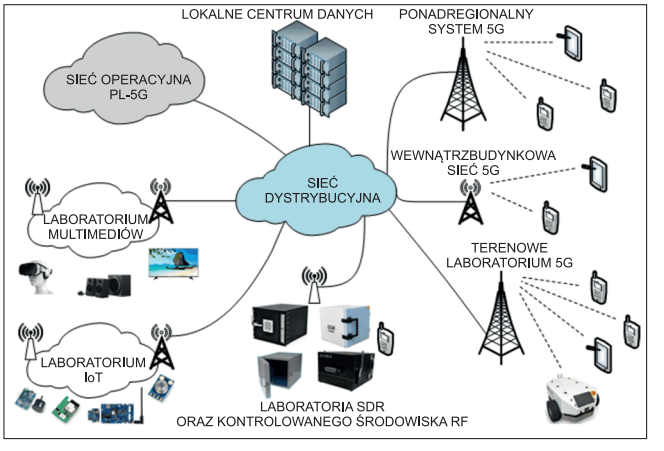
■ **Rys. 5. Laboratoria oraz systemy sieci 5G uruchamiane na Politechnice Gdańskiej**

Laboratorium Akademii Górniczo-Hutniczej

Na rys. 6 przedstawiono schemat laboratorium, który obejmuje ogólnokrajową sieć ponadregionalną, sieć wewnątrzbudynkową, jak i laboratoria specjalistyczne umożliwiające prowadzenie wszechstronnych badań naukowych związanych z sieciami i usługami 5G.

Systemy dostępowych sieci 5G będą złożone z zestawów stacji bazowych gNB i platform sprzętowych radia programowalnego SDR/OpenAirInterface, pracujących z użyciem różnych zakresów częstotliwości wraz z osprzętem koniecznym do pracy zarówno w środowisku zewnętrznym, jak i wewnątrzbudynkowym. Laboratoria sieci dostępowych będą wyposażone również w punkty dostępowe sieci Wi-Fi najnowszych generacji wraz z terminalami 5G pracującymi zgodnie ze standardami 3GPP oraz non-3GPP.

Laboratoria otoczenia sieci 5G zapewnią możliwość prowadzenia badań w zakresie nowoczesnych usług oferowanych za pośrednictwem sieci 5G oraz subiektywnego i obiektywnego postrzegania jakości tych usług przez użytkowników. Katalog usług dostępnych w tworzonym laboratorium obejmuje m.in.



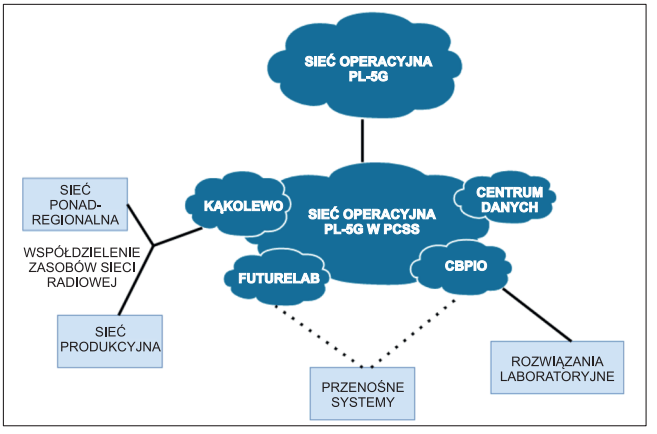
■ **Rys. 6. Laboratoria oraz systemy sieci 5G uruchamiane w AGH w Krakowie**

usługi wirtualnej i rozszerzonej rzeczywistości, hologramy, kamery 360 stopni oraz transmisje multimedialne najwyższej jakości. Dedykowane laboratorium IoT umożliwi prowadzenie kompleksowych badań, dotyczących zagadnień związanych z Internetem Rzeczy, poczynając od aspektów związanych z funkcjonowaniem infrastruktury sieciowej rdzenia 5G, przez zastosowanie zróżnicowanych technik dostępu, aż po uruchamianie i obsługę elementów aplikacyjnych związanych z usługami IoT, w szczególności z wykorzystaniem usług *fog* i *edge computing*.

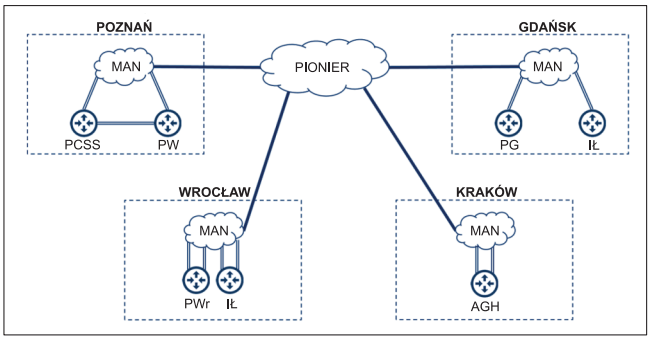
Laboratorium Poznańskiego Centrum Superkomputerowo-Sieciowego

Blokowy schemat topologii laboratorium przedstawiono na rys. 7.

Pierwszą z lokalizacji jest Centrum Badawcze Polskiego Internetu Optycz-



■ **Rys. 7. Topologia sieci 5G w węźle PCSS**



■ **Rys. 8. Topologia sieci operacyjnej**

nego, będące główną siedzibą PCSS oraz centrum przetwarzania danych. W tej lokalizacji będą skoncentrowane główne zasoby obliczeniowe laboratorium. Następną z lokalizacji jest FutureLab, gdzie przewidziane jest wykorzystanie przenośnych rozwiązań sieci 5G wraz z usługami. Obie lokalizacje mają fizyczne połączenie między sobą, co umożliwi zestawianie eksperymentów z wykorzystaniem zasobów obu lokalizacji. Ostatnią z lokalizacji jest Lotnisko EPPG, znajdujące się w Kąkolewie, które umożliwia prowadzenie badań dotyczących nowych rozwiązań i aplikacji dla bezzałogowych pojazdów powietrznych czy naziemnych lub otoczenia sieci 5G, wykorzystujących technikę przetwarzania danych na brzegu sieci.

W laboratorium PCSS planuje się zastosowanie rozwiązań zarówno laboratoryjnych, jak i produkcyjnych, zapewniających zasięgi umożliwiające przeprowadzanie eksperymentów w środowisku zbliżonym do komercyjnych wdrożeń sieci 5G. Podejście to umożliwi prowadzenie badań z wykorzystaniem urządzeń laboratoriów otoczenia związanych z multimediami, lotnictwem i komunikacją lądową czy przemysłem 4.0.

Sieć operacyjna

Sieć operacyjna krajowego laboratorium ma za zadanie połączenie laboratoriów umiejscowionych w różnych lokalizacjach, jak to pokazano na rys. 8. W tym celu będzie wykorzystywana ogólnopolska sieć światłowodowa PIONIER. Będzie ona łączyła 4 ośrodki w kraju i umożliwiała podłączanie lokalnych sieci regionalnych przez miejskie sieci akademickie.

Węzły sieci operacyjnej komunikują się przez sieć PIONIER, wykorzystując ruter operacyjny oraz dedykowane łącza światłowodowe. Przełączniki zapewniają komunikację pomiędzy wirtualnymi zasobami obliczeniowymi i dostęp do pamięci masowej, tak jak pokazano to na rys. 9. Wszystkie zasoby laboratorium badawczego będą podłączane przez przełącznik laboratoryjny (LAB), a jego konfiguracja będzie zautomatyzowana, aby dostarczać na żądanie izolowane wirtualne sieci lokalne (VLAN).

Częstotliwości pracy sieci PL 5G

W celu realizacji emisji testowych w sieci PL-5G niezbędne jest wykorzystanie częstotliwości radiowych służących do pracy w sieciach 5G NR. Zasadniczo rozróżnia się dwa podstawowe zakresy częstotliwości pracy sieci 5G NR: **FR1** (*Frequency Range 1*) – częstotliwości poniżej 7125 MHz oraz **FR2** (*Frequency Range 2*) – częstotliwości powyżej 24,25 GHz. Zakres FR1 jest najbardziej odpowiedni do pracy na rozległych obszarach (duże zasięgi stacji), podczas gdy częstotliwości zakresu FR2 wykorzystywane są do emisji lokalnych (małe zasięgi), gdy wymagana jest bardzo duża przepustowość (bardzo szerokie kanały radiowe). Możliwe są przy tym dwie podstawowe metody dostępu do widma radiowego: licencjonowany – w pasmach przeznaczonych dla systemów łączności ruchomej i nielicencjonowany – w pasmach ISM na zasadach współużytkowania widma obowiązujących w tych zakresach częstotliwości.

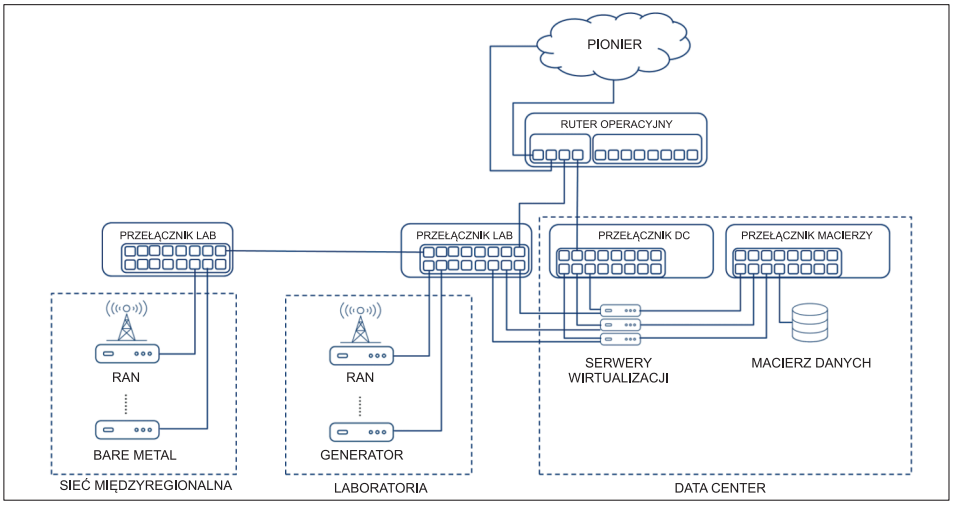
W projekcie PL-5G będą wykorzystywane zarówno częstotliwości zakresu FR1, jak i FR2, a także emisje w pasmach nielicencjonowanych ISM (tj. np. 2,4 GHz, 5 GHz). W przypadku zakresu FR1 dla każdej lokalizacji emisji będą uzgadniane z UKE odpowiednie kanały częstotliwości pasma C (zakres 3,4 – 4,4 GHz), w którym głównie pracują na świecie stacje 5G NR, a w przypadku zakresu FR2 odpowiednie kanały pasm milimetrowych (24,25 – 29,5 GHz). Dodatkowo pozyskane zostaną lokalnie niewielkie zasoby niskich częstotliwości (< 3 GHz) do celów realizacji emisji w standardach 3GPP (CAT1, NB IoT) dla systemów IoT na duże odległości.

Konkretne kanały i zakresy częstotliwości wykorzystywane w projekcie PL-5G będą zależeć od lokalizacji stacji i dostępnych w danym miejscu wolnych częstotliwości radiowych.

ZASADY KORZYSTANIA Z LABORATORIUM

Laboratoria projektu będą dostępne dla szerokiego grona użytkowników końcowych. Z infrastruktury badawczej będą mogli korzystać zarówno naukowcy, reprezentanci przedsiębiorstw (w tym przedstawiciele małych i średnich przedsiębiorstw, MŚP), jak również użytkownicy indywidualni. Zasady dostępu do infrastruktury badawczej PL-5G, a zatem prawa i obowiązki w zakresie korzystania z infrastruktury badawczej przy prowadzeniu badań naukowych lub prac rozwojowych, określone są w Regulaminie Dostępu. Podmiotem udostępniającym infrastrukturę badawczą jest członek Konsorcjum PL-5G, umożliwiający dostęp do laboratoriów. Korzystanie z infrastruktury badawczej laboratoriów w ramach projektu będzie odbywać się zgodnie z powszechnie obowiązującymi przepisami prawa, Regulaminem Dostępu oraz postanowieniami umowy zawartej z użytkownikiem końcowym.

W projekcie dostęp dla użytkowników końcowych do zasobów infrastruktury badawczej do prowadzenia prac badawczych lub rozwojowych przewiduje się w dwóch trybach. Pierwszy z trybów dostępu to tryb niegospodarczy. Użytkownicy uprawnieni do prowadzenia badań w trybie niegospodarczym to przede wszystkim pracownicy jednostek naukowych, prowadzący prace badawcze związane ze statusem swojej jednostki macierzystej. Drugi tryb dostępu do infrastruktury badawczej to tryb gospodarczy. Jest on wykorzystywany w przypadku udostępnienia zasobów i usług infrastruktury na rzecz przed-



■ **Rys. 9. Schemat węzła sieci operacyjnej**

siębiorców i MŚP oraz pracowników jednostek naukowych prowadzących badania na rzecz przedsiębiorców.

SYSTEM ZARZĄDZANIA EKSPERYMENTAMI

System zarządzania w laboratorium PL-5G kontroluje procesy realizacji eksperymentów prowadzonych przez użytkowników w rozproszonej infrastrukturze oraz monitoruje elementy tej infrastruktury.

Z technicznego punktu widzenia korzystanie z laboratoriów będzie wymagało elektronicznej rejestracji w systemie zarządzania za pomocą stworzonej do tego celu aplikacji internetowej dostępnej z poziomu przeglądarki internetowej. Rejestracja w systemie będzie obowiązkowa dla wszystkich użytkowników, którym będzie udostępniana infrastruktura.

System zostanie zbudowany z wykorzystaniem instrumentów, umożliwiających monitorowanie użycia zasobów i ich poprawnego działania. Zapewni zarządzanie dostępem do poszczególnych laboratoriów. Umożliwi tworzenie raportów o wykorzystaniu zasobów z podziałem na gospodarczy i niegospodarczy charakter ich wykorzystania.

Wszyscy użytkownicy zalogowani w systemie zarządzania PL-5G będę mieli możliwość wglądu w ofertę poszczególnych laboratoriów badawczych budowanych w projekcie, a następnie, wykorzystując dedykowany formularz, aplikowania o rezerwację i przygotowanie zasobów do przeprowadzenia swoich prac badawczych.

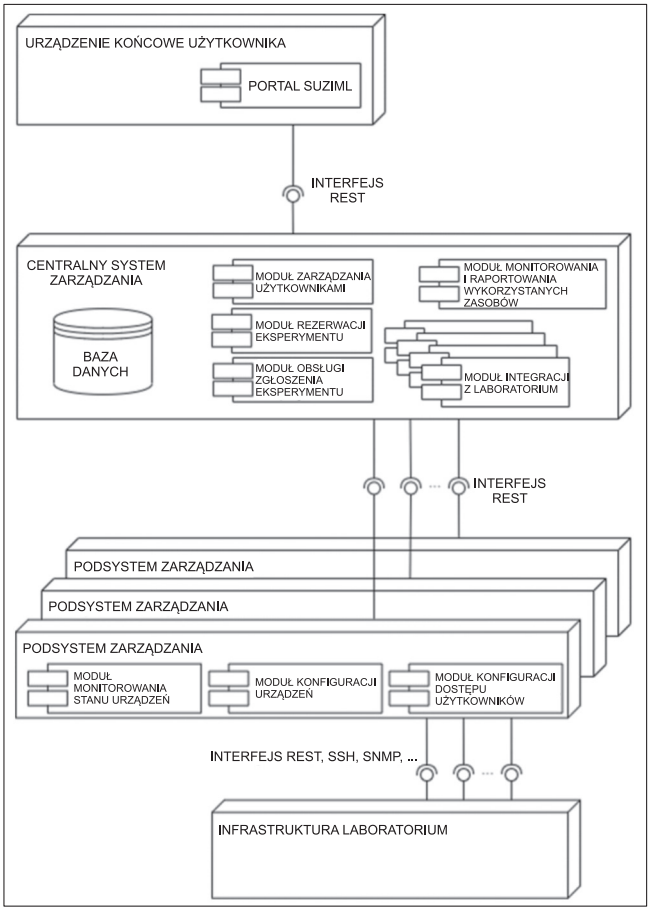
Użytkownicy będą mieli możliwość wybrania, gdzie mają znajdować się zasoby, z których chcą skorzystać, oraz wprowadzenia szczegółowej specyfikacji zasobów (np. parametrów maszyn wirtualnych uruchamianych na żądanie) oraz wskazania topologii połączeń między zasobami. System udostępni również narzędzie, zapewniające wskazanie uczestników prac (innych użytkowników zarejestrowanych w systemie), którzy mają uzyskać prawa dostępu do zasobów (dedykowanego dla potrzeb indywidualnego użytkownika laboratorium) w celu wspólnej realizacji prac badawczych.

Oprogramowanie systemu zarządzania zostanie oparte na architekturze rozproszonej. Architekturę oprogramowania przedstawiono na rys. 10. Generyczny rdzeń systemu powinien mieć możliwość współpracy z wieloma heterogenicznymi infrastrukturami laboratoriów w celu realizacji następujących procesów:

- konfiguracji urządzeń zgodnie ze specyfikacją rezerwacji wprowadzonej przez użytkownika,
- konfiguracji dostępu użytkowników (w głównej mierze zdalnego) do zarezerwowanych zasobów,
- pobrania i prezentacji informacji o dostępności zasobów.

System centralny będzie komunikował się z podsystemami zarządzania przez dedykowane moduły integracji.

W przypadku, gdy powyżej wspomniane procesy nie mogą być realizowane automatycznie (m.in. z uwagi na charakter lub ograniczenia rezerwowanych zasobów), system będzie wspierał operatorów infrastruktury PL-5G w procesie ręcznego przygotowania zasobów.



■ **Rys. 10. Architektura systemu zarządzania PL-5G**

* * *

W artykule pokrótce przedstawiono zarys planowanej sieci PL-5G, pn. *Krajowe laboratorium sieci i usług 5G wraz z otoczeniem*. Budowa tej sieci powinna być zakończona pod koniec 2023 r. Oczekuje się, iż dostęp do tego laboratorium istotnie pomoże w realizacji ważnych projektów krajowych, jak i międzynarodowych. Ponadto laboratorium powinno w sposób znaczący uatrakcyjnić proces dydaktyczny na uczelniach technicznych.

Adrian DZIEMBOWSKI*, Dawid MIELOCH*

Nowe techniki kompresji wizji dla rzeczywistości wirtualnej – MPEG Immersive Video

New video compression techniques for virtual reality – MPEG Immersive Video

Artykuł towarzyszy referatowi plenarnemu na temat najnowszej normy kodowania wizji – ISO/IEC 23090 część 12: MPEG Immersive Video (MIV). Zastosowaniem tej normy jest kompresja danych wizyjnych z wielu kamer do wykorzystania w rzeczywistości wirtualnej. Norma kodowania MIV jest niezależna od kodeków wideo, tj. obejmuje przetwarzanie wstępne i końcowe wokół istniejących kodeków, takich jak AVC, HEVC i VVC. Autorzy artykułu – wieloletni członkowie grupy eksperckiej ISO/IEC JTC1/SC29 WG11 MPEG, aktualnie WG04 MPEG Video Coding – wyjaśniają zasadę działania kodera i dekodera MIV, opisując najistotniejsze narzędzia umożliwiające efektywną kompresję wizji zarejestrowanej systemem wielokamerowym. Ponadto autorzy skupiają się na wymaganiach praktycznych systemów rzeczywistości wirtualnej, prezentując zalety i możliwości wykorzystania dwóch podstawowych profili MIV: profilu głównego (MIV Main) oraz profilu zakładającego estymację głębi po stronie odbiorczej (MIV Geometry Absent).
Słowa kluczowe: wizja wszechogarniająca, kodowanie wizji, rzeczywistość wirtualna, przetwarzanie wizji wielowidokowej

The paper accompanies the plenary lecture which provides an overview of the latest video coding standard – ISO / IEC 23090 Part 12: MPEG Immersive Video (MIV). An application of this standard is the compression of video from multiple cameras for virtual reality applications. The MIV encoding standard is independent of the video codecs, i.e., it consists of pre-processing and post-processing merged with using existing codecs such as AVC, HEVC, and VVC. The authors of the paper – long-term experts of the ISO/IEC JTC1/SC29 WG11 MPEG, currently WG04 MPEG Video Coding – explain the principles of the MIV encoder and decoder, describing the most important tools enabling effective compression of video recorded with a multi-camera system. In addition, the authors focus on the practical requirements of virtual reality systems, presenting the advantages and possibilities of using two basic MIV profiles: the main profile (MIV Main) and the profile assuming estimation of depth on the decoder side (MIV Geometry Absent).
Key word: immersive video, video coding, virtual reality, multiview video processing

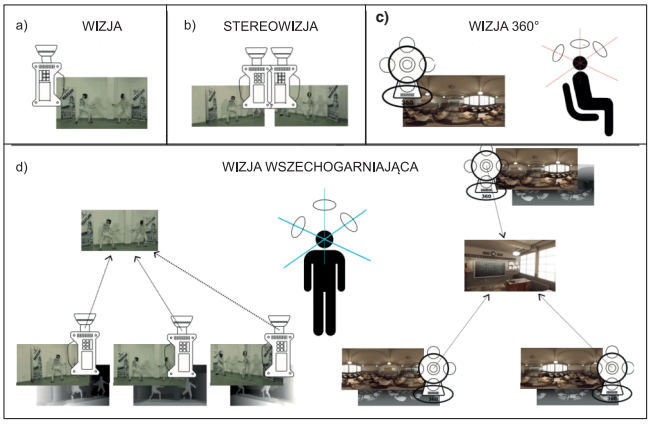
DOI: 10.15199/59.2022.4.4

WIZJA WSZECHOGARNIAJĄCA

System wizji wszechogarniającej, zwanej immersyjną, jest systemem umożliwiającym użytkownikowi (widzowi) wirtualne poruszanie się w scenie zarejestrowanej z użyciem wielu kamer. Taka nawigacja może być realizowana z wykorzystaniem wyświetlaczy nagłownych **HMD** (*Head-Mounted Displays*) [1], tradycyjnych monitorów z dołączonym urządzeniem sterującym, lub tabletów przez wykorzystanie ich dotykowych wyświetlaczy [2], [3].

W odróżnieniu od wizji 360°, gdzie widz może wyłącznie zmieniać kierunek obserwacji sceny (rys. 1b) i systemów stereoskopowych, które dodatkowo zapewniają wrażenie głębi (rys. 1c), w systemie wizji wszechogarniającej (rys. 1d) swoboda ruchów jest dużo większa, a widz może zarówno obracać się, jak i przemieszczać w scenie [4].

Pozornie może się wydawać, iż masowe wykorzystanie systemów wizji wszechogarniającej jest mrzonką lub – myśląc optymistyczniej – kwestią dalekiej przyszłości, kojarzoną raczej z futurystycznymi filmami niż z dzisiejszymi transmisjami telewizyjnymi czy internetowymi. Myślenie takie jest jednak nieuzasadnione, a kolejne przykłady wykorzystania wirtualnej nawigacji w scenie regularnie pojawiają się w transmisjach zawodów sportowych z całego świata, np. transmisji koszykówki [5] czy też igrzysk olimpijskich, zarówno letnich [6], jak i zimowych [7].

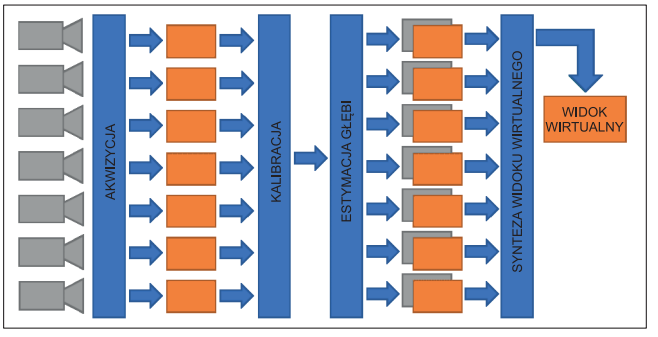


Rys. 1. Rodzaje systemów wizyjnych

Jeszcze częstsze są transmisje z uproszczonym modelem swobodnej nawigacji, w którym treść zarejestrowanej sceny jest zastępowana trójwymiarowym modelem boiska, pokazywanym przy rozstrzyganiu spornych sytuacji z użyciem technologii *Goal Line* [8] czy tenisowej *HawkEye* [9].

Oczywiście obecne systemy nie wykorzystują pełni potencjału wizji wszechogarniającej, czyli umożliwienia użytkownikowi pełnej swobody ruchu w zarejestrowanej scenie. Niemniej jednak w wymienionych systemach wykorzystywane są te same

* Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska,
e-mail: adrian.dziembowski@put.poznan.pl, dawid.mieloch@put.poznan.pl



■ **Rys. 2. Schemat przetwarzania wizji wszechogarniającej; kolorem niebieskim oznaczono kolejne kroki przetwarzania, pomarańczowym prostokątem – sekwencje wizyjne (widoki), a szarym – mapy głębi**

techniki przetwarzania wizji wielowidokowej co w systemie wizji wszechogarniającej, a więc estymacja geometrii sceny i układu kamer oraz synteza (renderowanie) kolejnych obrazów na trajektorii położeń wirtualnej kamery. Kroki te zostały opisane w następnym rozdziale.

Jedyną – choć zasadniczą – różnicą jest wybór wspomnianej trajektorii, za którą obecnie odpowiada realizator transmisji, a nie końcowy użytkownik systemu, czyli widz. Różnica ta powoduje poważne konsekwencje pod względem szerokości pasma wymaganego do przesłania treści do widza. W przypadku wysyłania jednej wybranej trajektorii do wszystkich widzów transmitowany jest tradycyjny strumień wizyjny, a więc jedna sekwencja (na przykład FullHD zakodowana koderem HEVC [10]).

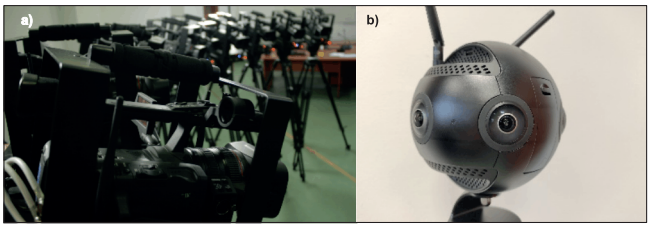
Przeniesienie możliwości wyboru trajektorii na użytkownika końcowego wiąże się natomiast z tym, iż każdy widz sam decyduje o tym, na co spogląda. Tym samym, aby umożliwić jednoczesną transmisję wizji wszechogarniającej do wielu użytkowników, nadawca nie może wybrać jednego, konkretnego strumienia wizyjnego. Zamiast tego konieczne jest wysłanie obrazów z wielu kamer.

W przypadku tak znaczącego zwiększenia ilości przesyłanych danych tradycyjne kodery wizyjne okazują się niewystarczająco efektywne, przez co stworzenie praktycznego systemu wizji wszechogarniającej bez wykorzystania dedykowanych algorytmów kompresji jest bardzo trudne, co przedstawiono w kolejnym rodziale. Próbę rozwiązania tego problemu stanowi najnowsza norma kompresji – ISO/IEC 23090-12: MPEG Immersive Video [11], którą opisano w rozdziale *MPEG Immersive Video.*

PRZETWARZANIE WIZJI WSZECHOGARNIAJĄCEJ

W celu zapewnienia użytkownikowi możliwości wirtualnej nawigacji pośród zarejestrowanej trójwymiarowej sceny należy uzyskać jej odpowiednią reprezentację. Najczęściej używana jest reprezentacja **MVD** (*Multiview Video plus Depth*) [12], w której zawartość oraz geometria sceny są przechowywane w postaci wielu widoków zarejestrowanych za pomocą kamer oraz tak zwanych map głębi (opisanych w rozdziale *Estymacja głębi*). Inne spotykane reprezentacje oparte są na chmurach trójwymiarowych punktów [13], które są częściej wykorzystywane dla danych uzyskanych za pomocą sprzętowych metod akwizycji geometrii sceny (np. systemów LiDAR [14]).

Odpowiednia reprezentacja sceny, wraz z parametrami i pozycjami użytych kamer, wykorzystywana jest do stworzenia dla użytkownika wybranego widoku w procesie syntezy widoku wirtualnego. Uproszczony schemat przetwarzania przedstawiono na rys. 2, a dalsze części tego rozdziału opisują poszczególne etapy przetwarzania wraz z przykładowymi metodami używanymi w rzeczywistych systemach wizji wszechogarniającej.



■ **Rys. 3. System wielokamerowy wyposażony w tradycyjne, perspektywiczne kamery (a); pojedyncza kamera wszechkierunkowa (b)**

Akwizycja

W systemach wizji wszechogarniającej sposób rozmieszczenia kamer oraz ich typ nie są ściśle ograniczone i w głównej mierze zależą od rozmiaru rejestrowanej sceny i końcowego przeznaczenia systemu. W przypadku użycia typowych, standardowych kamer perspektywicznych (rys. 3a) ich liczba może sięgać setek [2], ale w praktycznych systemach jest takich kamer kilka/kilkanaście w zależności od użytego rozstawienia [3].

Używając kamer wszechkierunkowych [1] (rys. 3b), można zredukować liczbę urządzeń, gdyż każde z nich rejestruje dużo większy obszar sceny niż kamera perspektywiczna. Z drugiej strony ich użycie jest często niemożliwe w przypadku systemu realizowanego na potrzeby np. transmisji sportowych, w których kamery nie mogą być umieszczone wewnątrz rejestrowanej sceny, lecz dookoła niej.

Sekwencje używane w wizji wszechogarniającej mogą być również oparte na wykorzystaniu grafiki komputerowej, zarówno przez generowanie odpowiednich scen w całości, jak przez łączenie takich scen z rzeczywistymi obiektami zarejestrowanymi przez kamery (np. przez umieszczenie naturalnych postaci w komputerowej scenerii).

Kalibracja

W celu uzyskania parametrów kamer, konieczna jest ich kalibracja. Kalibracja dla parametrów wewnętrznych kamer, która m.in. obejmuje estymację ich ogniskowej czy zniekształceń wprowadzanych przez soczewki kamer, jest łatwa do wykonania przy użyciu zdefiniowanego wzorca, np. klasycznej szachownicy. Zazwyczaj dla każdej kamery wymagane jest zarejestrowanie dziesięciu różnych widoków wzorca. Wiele metod dostępnych w bibliotekach *open source* zapewnia wystarczającą dokładność do zastosowań wizji wszechogarniającej (np. OpenCV [15]).

Kalibracja parametrów zewnętrznych kamer, tj. opisujących położenie kamer względem rejestrowanej sceny, powinna być powtarzana każdorazowo po zmianie pozycji kamer. Zazwyczaj jest wykonywana przez optymalizację funkcji opisującej zależności między punktami charakterystycznymi zarejestrowanymi jednocześnie przez kamery systemu (minimalizacja odległości od punktu do linii epipolarnej [16]). Automatyczne wyszukiwanie takich punktów odniesienia jest preferowane w praktycznych systemach wielokamerowych i jest uzyskiwane przez automatyczną ekstrakcję cech lub używanie pewnych własności obiektu kalibracyjnego, np. najjaśniejszego punktu w scenie w przypadku użycia źródła światła.

W przypadku scen zarejestrowanych za pomocą rzeczywistych kamer dobrą praktyką jest przeprowadzenie również tzw. korekcji koloru [17], która ma zminimalizować różnice w charakterystyce barwnej, występujące pomiędzy kamerami systemu, jak i zmiany charakterystyki w czasie (wynikające np. z użycia automatycznego balansu bieli). Jednoczesne zmniejszenie międzywidokowej i czasowej fluktuacji charakterystyk barwnych ma pozytywny wpływ na jakość estymowanych map głębi i wydajność kompresji takich sekwencji [18].

Estymacja głębi

Mapy głębi, najczęściej reprezentowane jako obrazy w skali szarości, przedstawiają odległość punktów zarejestrowanej sceny od płaszczyzny opisanej na przetworniku kamery (rys. 4). Mapy głębi najczęściej estymowane są algorytmicznie na podstawie widoków zarejestrowanej sceny. Głębia punktu jest wyliczana przez poszukiwanie punktu odpowiadającego, czyli punktu, który reprezentuje tę samą część sceny w innym widoku [16].

Biorąc pod uwagę dużą złożoność obliczeniową tego etapu przetwarzania sekwencji wielowidokowych, mapy głębi często są również uzyskiwane za pomocą tzw. kamer głębi [14] lub za pomocą metod opartych na zastosowaniu głębokiego uczenia [19]. W obu tych przypadkach często wymagane jest zastosowanie dodatkowego przetwarzania końcowego, mającego na celu międzywidokowe uwspólnienie map głębi, które są estymowane niezależnie dla każdego z widoków [20].

Synteza widoków wirtualnych

Synteza widoków wirtualnych stanowi ostatni etap przetwarzania danych w systemie wizji wszechogarniającej. W syntezie wielowidokowej [21], [22] zmarginalizowany jest wpływ przysłonięć obiektów. Kiedy scena rejestrowana jest przez większą liczbę kamer, obszarów całkowicie przysłoniętych we wszystkich widokach rzeczywistych jest niewiele. Metoda syntezy powinna być jednak dostosowana do systemu, w którym będzie wykorzystywana, np. przez umożliwienie syntezy z obrazów wszechkierunkowych [23], [24] czy syntezę dokonywaną w czasie rzeczywistym [25], [26], niezbędną w przypadku użycia wyświetlaczy nagłownych do obserwowania sceny.

KOMPRESJA WIZJI WSZECHOGARNIAJĄCEJ

Powszechnie wiadomo, że niezależna kompresja wielu widoków i głębi skutkuje wysokimi przepływnościami. Co więcej, nowym krytycznym ograniczeniem w wizji wszechogarniającej jest tak zwana prędkość próbkowa (*pixel rate*): odzwierciedla całkowitą liczbę próbek obrazu, które mają być zdekodowane na sekundę w celu zainicjowania syntezy nowego wirtualnego widoku. Jest to zwykle uwarunkowane praktycznymi zasobami dekoderów sprzętowych. Na przykład, dla normy HEVC, profil Main 10 na poziomie 5.2 [10], który jest kompatybilny z obecnymi sprzętowymi dekoderami (np. [27], [28]), zakłada, że dla sekwencji o rozdzielczości 4096×2048 dla 30 klatek na sekundę liczba wystąpień dekodera nie może być większa niż cztery. W przypadku wizji wszechogarniającej oznacza to, że skompresowany strumień nie może zawierać więcej niż dwa widoki i odpowiadające im mapy głębi. Co więcej, podczas gdy nowoczesne uniwersalne kodery wideo zapewniają dobrą wydajność kodowania typowych treści zarejestrowanych przez kamery, kodowanie mapy głębi jest zdecydowanie bardziej wydajne, gdy stosowane są dedykowane rozwiązania. Z tego powodu używanie tego samego kodera do obu celów jest niewłaściwe.

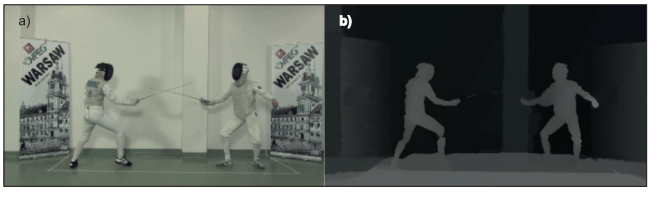
Jak opisywano wcześniej, wielowidokowa sekwencja jest rejestrowana przez kilka kamer (np. wszechkierunkowych), w których znacząca część sceny jest widoczna we wszystkich z nich. Nadmiarowość wynikająca z przestrzennego nakładania się widoków wejściowych może być wykorzystana do zmniejszenia rozmiaru danych wymaganych do pełnego odwzorowania całej trójwymiarowej sceny. Pierwsze kodery wizyjne, mające w wydajny sposób kompresować obrazy wielowidokowe przez identyfikację takich wspólnych treści, czyli MVC [29], MV-HEVC i 3D-HEVC [30], które są rozszerzeniami szeroko stosowanych norm kompresji typowych jednowidokowych sekwencji, nie są niestety dopasowane do charakterystyki wizji wszechogarniającej. Zarówno MVC, jak MV-HEVC, nie zapewniają usprawnień dla kodowania głębi (która może stanowić ponad 50% strumienia bitowego). 3D-HEVC, mimo że jego implementacje są nadal aktywnie ulepszane przez niektórych badaczy, nie jest kompatybilny z nieliniowymi układami kamer i kamerami wszechkierunkowymi, znacznie ograniczając jego użyteczność. Z tego powodu nie zaplanowano takiego typu rozszerzeń dla najnowszej normy *Versatile Video Coding* (**VVC**), lecz zdecydowano się na nowe podejście do kodowania wizji wszechogarniającej, czyli *MPEG Immersive Video* (**MIV**).

MPEG IMMERSIVE VIDEO

Obecna, najnowocześniejsza technologia kompresji dla wizji wszechogarniającej została opracowana przez grupę MPEG ISO/IEC pod nazwą *MPEG Immersive Video* [31]. Główną ideą w tej normie jest założenie, że pewna mała liczba widoków bazowych, gromadzących większość informacji o scenie, powinna być zakodowana w całości, a informacje uzupełniające (widoczne z pozostałych widoków) mogą być przekazywane w formie małych fragmentów („łatek"), tworzących „atlas", bądź pominięte, zwiększając liczbę widoków bazowych.

Z formalnego punktu widzenia, norma ISO/IEC 23090:12 MIV [11] stanowi rozszerzenie normy ISO/IEC 23090:5 V3C (*Visual Volumetric Video-based Coding*) [13] standaryzującej kodowanie scen trójwymiarowych reprezentowanych w postaci chmur punktów (*point clouds*). Kodowanie MIV wykorzystuje składnię i semantykę normy V3C, dodając techniki i narzędzia niezbędne do poprawnego kodowania i dekodowania treści w formacie MVD [12].

W odróżnieniu od wcześniejszych koderów wizji wielowidokowej (MVC, MV-HEVC i 3D-HEVC), które definiowały cały proces kodowania, począwszy od wczytania widoków wejściowych aż do tworzenia jednego wspólnego strumienia bitowego, koder MIV stanowi swoiste przetwarzanie wstępne wizji wielowidokowej, usuwając redundancję międzywidokową i zmieniając sposób reprezentacji danych wejściowych. Wyjście z kodera MIV stanowią atlasy – sekwencje wizyjne zawierające informację



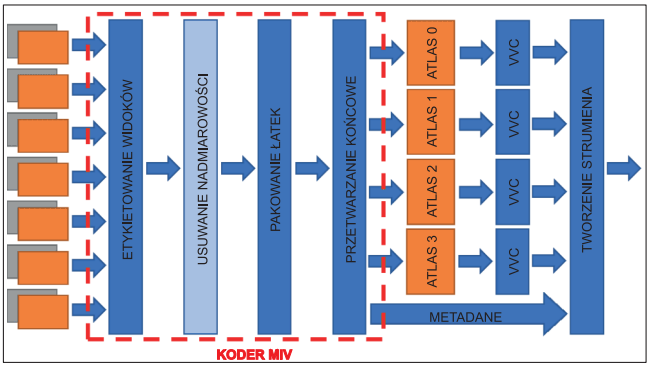
■ **Rys. 4. Widok zarejestrowany przez kamerę (a) i odpowiadająca mu mapa głębi (b), sekwencja *Fencing***

z wielu widoków wejściowych. Atlasy te są następnie kodowane przy użyciu dowolnego kodera wizyjnego, czyniąc MIV techniką całkowicie niezależną od używanej techniki kompresji (*codec--agnostic*) [32]. Tym samym, wraz z MIV, można użyć efektywnej, choć czasochłonnej kompresji, z użyciem najnowszego kodera VVC [33], jak również mniej efektywnych, ale znacznie szybszych koderów AVC [34] czy HEVC [10] z dowolnie ustawionymi parametrami [35] czy włączonymi dodatkowymi narzędziami [36]. Pełna dowolność w odniesieniu do używanego kodera wizyjnego stanowi niepodważalną zaletę MIV.

Norma MPEG Immersive Video definiuje 3 profile: główny (MIV Main), rozszerzony (MIV Extended) oraz profil z estymacją głębi po stronie dekodera (MIV Geometry Absent, GA). Poszczególne profile odpowiadają różnym wymaganiom rynku i potrzebom użytkowników. W niniejszym artykule skupiono się na dwóch z nich – MIV Main oraz MIV GA.

KODOWANIE MIV

Jak wspomniano, koder MIV stanowi przetwarzanie wstępne sekwencji wielowidokowej, umieszczając informację z widoków wejściowych w czterech atlasach. Następnie każdy z atlasów jest niezależnie kodowany z użyciem typowego kodera wizyjnego (np. VVC [33]), a skompresowane strumienie wizyjne, wraz ze strumieniem metadanych, są multipleksowane w celu stworzenia jednego strumienia bitowego transmitowanego do użytkownika. Schemat kodowania wizji z użyciem kodera MIV przedstawiono na rys. 5.



■ **Rys. 5. Uproszczony schemat kodowania MIV, czerwoną ramką zaznaczono koder MIV, kolorem niebieskim oznaczono kolejne kroki kodowania (jasnoniebieskim – kroki opcjonalne), pomarańczowym – sekwencje wizyjne przed i po kodowaniu koderem MIV**

Wyodrębniono cztery podstawowe etapy przetwarzania wizji w koderze MIV (czerwona ramka): etykietowanie widoków, usuwanie nadmiarowości, pakowanie łatek do atlasów i przetwarzanie końcowe. Kroki te zostały opisane w kolejnych podrozdziałach.

Szczegółowy opis kodera MIV można znaleźć w artykule [31] oraz dokumencie opisującym działanie modelu testowego MIV [37].

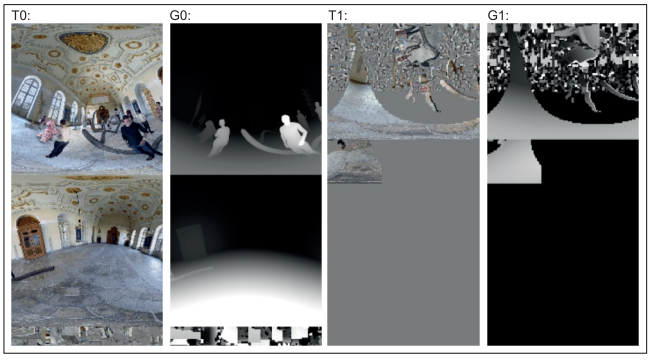
MIV Main

Profil główny normy MIV (MIV Main) opisuje podstawowy sposób kodowania wizji wszechogarniającej, przeznaczony dla prostych, praktycznych systemów swobodnej nawigacji. W takich systemach kluczowe jest szybkie dekodowanie treści z użyciem prostych i tanich urządzeń końcowych, np. telefonów komórkowych czy tabletów.

Koder MIV, działający w trybie MIV Main, realizuje wszystkie cztery główne kroki przetwarzania wizji wielowidokowej, generując dwa atlasy tekstury i dwa atlasy głębi (rys. 6). Atlasy te zawierają informację ze wszystkich widoków wejściowych.

Etykietowanie widoków

W pierwszym kroku kodowania koder MIV decyduje, które z widoków wejściowych niosą najwięcej istotnych informacji. Widoki te są określane widokami bazowymi i będą umieszczo-



■ **Rys. 6. Cztery atlasy wygenerowane przez koder MIV działający w trybie MIV Main; sekwencja** *Museum*

ne w atlasach w całości (2 widoki w atlasach T0 i G0 na rys. 6). Istotna (nienadmiarowa) informacja z pozostałych widoków wejściowych (określanych widokami dodatkowymi) również jest umieszczana w atlasach, aczkolwiek widoki te są wcześniej przetwarzane w celu usunięcia redundancji międzywidokowej.

Liczba widoków bazowych zależy od rozdzielczości sekwencji. W przypadku sekwencji wysokiej rozdzielczości (4K) wysłanie większej liczby widoków bazowych niż jeden spowodowałoby, iż w atlasach zabrakłoby miejsca na informację z pozostałych widoków. Dla sekwencji o rozdzielczości mniejszej (np. FullHD) widoków bazowych jest więcej.

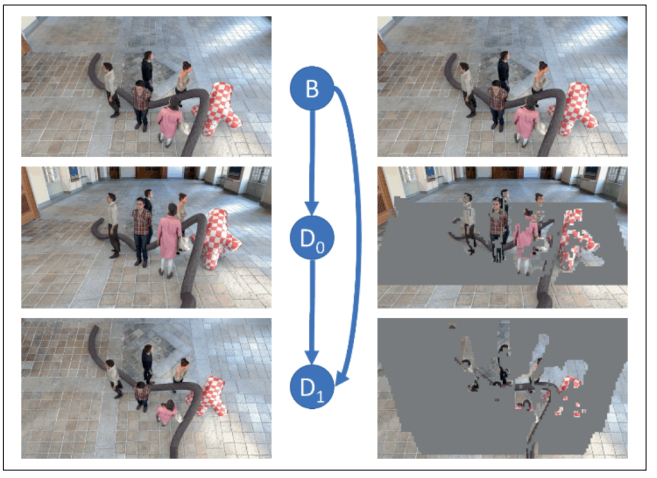
Wybór widoków bazowych dokonywany jest w sposób automatyczny na podstawie analizy rozmieszczenia kamer w systemie. Jeżeli jest tylko jeden widok bazowy, wybierany jest widok z najbardziej centralnej kamery systemu. W przypadku większej liczby (k) widoków bazowych, wybierane są widoki z k kamer najbardziej oddalonych od siebie. Widoki bazowe wybierane są przy użyciu algorytmu **PAM** (*Partitioning Around Medoids*) [38].

Usuwanie nadmiarowości

W celu usunięcia międzywidokowej nadmiarowości widoków dodatkowych, stosuje się przerzutowanie punktów (pikseli) pomiędzy poszczególnymi widokami.

W pierwszym kroku wszystkie punkty widoku bazowego (widoków bazowych, węzeł B na rys. 7) są rzutowane do każdego widoku dodatkowego. W przypadku, gdy kolor i głębia punktu (x, y) w widoku dodatkowym są takie same jak kolor i głębia przerzutowane do punktu (x, y) z widoku bazowego, punkt ten jest uznawany za nadmiarowy i usuwany (szare obszary na rys. 7) [39].

Następnie wybierany jest ten spośród widoków dodatkowych (węzeł D_0 na rys. 7), w którym usunięto najmniej nadmiarowych punktów, a cały algorytm jest iteracyjnie powtarzany przez przerzutowanie nienadmiarowych punktów z tego widoku do pozostałych widoków dodatkowych. Algorytm kończy się w momencie, gdy usunięto nadmiarowe punkty ze wszystkich widoków wejściowych.



■ **Rys. 7. Proces usuwania nadmiarowości międzywidokowej. Po lewej – widoki wejściowe: B (bazowy), D_0 i D_1 (widoki dodatkowe), po prawej – widoki po usunięciu informacji nadmiarowej (szare obszary). Strzałkami oznaczono kierunek rzutowania informacji pomiędzy widokami**

Pakowanie łatek do atlasów

Po usunięciu nadmiarowości widoki dodatkowe zawierają wyłącznie informację nienadmiarową. Informacja ta, wraz z pełnymi widokami bazowymi, jest umieszczana w atlasach w postaci mozaiki łatek, gdzie łatka oznacza prostokątny obszar widoku wejściowego obejmujący wzajemnie połączone nieusunięte punkty. W przypadku widoku bazowego łatka obejmuje cały obszar widoku.

Poszczególne łatki są sekwencyjnie pakowane do atlasów w kolejności malejącej powierzchni, z użyciem algorytmu MaxRect [40].

Oczywiście, w celu zapewnienia dekodowalności, algorytm pakowania musi być odwracalny, a więc informacja o rozmiarze i oryginalnym położeniu każdej łatki w widoku wejściowym musi zostać przesłana. W koderze MIV informacja ta jest umieszczana w strumieniu metadanych (rys. 5).

Przetwarzanie końcowe

Atlasy stworzone w procesie pakowania zawierają znacznie mniej informacji niż widoki wejściowe. Niemniej jednak charakterystyka atlasu (wiele małych obiektów, krawędzi i wysokich częstotliwości) czyni je treścią bardzo trudną do efektywnego zakodowania z użyciem tradycyjnych koderów wizyjnych.

Dlatego też przed zakodowaniem atlasy są dodatkowo przetwarzane. W przypadku głębi rozdzielczość atlasów jest dwukrotnie zmniejszana w obu kierunkach, a zakres dynamiczny głębi jest normalizowany. Poszczególne łatki w atlasach tekstury są natomiast modyfikowane przez usunięcie składowej stałej lumy i obu chrom, zmniejszając amplitudę i liczbę krawędzi pomiędzy łatkami.

Ponadto zarówno w głębi, jak i teksturze, łatki nie zawierają wyłącznie obszarów nieusuniętych (o nieregularnych krawędziach), ale również fragmenty ich otoczenia [41], dostosowując się do działania kodera wizyjnego, który operuje na strukturze blokowej jednostek CU [10].

MIV Geometry Absent

Inne podejście do kodowania wizji wszechogarniającej polega na zastąpieniu transmisji map głębi ich estymacją ze zdekodowanych widoków w dekoderze lub w chmurze obliczeniowej [42]. Takie rozwiązanie umożliwia uwzględnienie w strumieniu bitów dwukrotnie większej liczby atlasów tekstury niż w typowej kompresji widoków i map głębi za pomocą MIV Main (por. rys. 6 i 8), przy tej samej prędkości próbkowej.

Profil MIV GA umożliwia strumieniom bitowym MIV niezawieranie map głębi, zachowując tylko informacje wymagane do ich rekonstrukcji. Szczegóły dotyczące wszystkich rozwiązań znajdujących zastosowanie w tym profilu można znaleźć w [43].

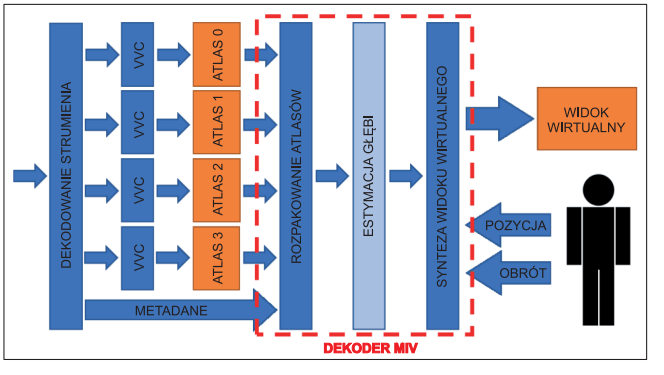
■ **Rys. 8. Cztery atlasy wygenerowane przez koder MIV działający w trybie MIV GA; sekwencja** ***Museum***

Etykietowanie i pakowanie widoków

W przypadku estymacji głębi po stronie dekodera nie jest przeprowadzane usuwanie nadmiarowości, gdyż jest ono dokonywane na podstawie map głębi widoków wejściowych, które nie muszą być dostępne w koderze. Zatem podczas etykietowania widoków wybierane są tylko widoki podstawowe, jednak w większej liczbie niż w przypadku podstawowego profilu, gdyż nie wysyłając głębi, możliwe jest umieszczenie dodatkowych pełnych widoków w atlasach bez zmiany prędkości próbkowej. W związku z używaniem jedynie pełnych widoków, ich pakowanie sprowadza się do umieszczenia ich kolejno w atlasie, znacząco upraszczając ten krok i zmniejszając złożoność kodera.

Przetwarzanie końcowe

Jedną z wad profilu MIV GA jest brak możliwości skorzystania z wysokiej jakości map głębi, które mogą znajdować się po stronie kodera. Te mapy mogą być dostarczane wraz sekwencjami wygenerowanymi komputerowo lub estymowane



■ **Rys. 9. Uproszczony schemat dekodowania MIV, czerwoną ramką zaznaczono dekoder MIV, kolorem niebieskim oznaczono kolejne kroki dekodowania (jasnoniebieskim – kroki opcjonalne), pomarańczowym – sekwencje wizyjne przed i po dekodowaniu dekoderem MIV**

za pomocą ręcznych lub automatycznych metod poprawy ich jakości w przypadku treści naturalnych.

Aby możliwe było wykorzystanie tych wartościowych danych, są one przesyłane w strukturze Geometry Assistance SEI, umożliwiającej transmisję pewnych informacji pobocznych, które pomagają estymatorowi głębi po stronie dekodera w uzyskaniu map głębi o wyższej jakości przy jednoczesnym zmniejszeniu złożoności estymacji [44]. Jedną z przykładowych przesyłanych informacji jest flaga wskazująca dla poszczególnych bloków mapy głębi, czy można użyć dla danego obszaru głębię z poprzedniej ramki w związku z brakiem ruchomych obiektów.

Kodowanie atlasów

Jak opisano wcześniej, do zakodowania atlasów stworzonych przez koder MIV można wykorzystać dowolną technikę kodowania wizji. W związku z tym każdy atlas kodowany jest niezależnie, umożliwiając traktowanie każdego z nich w taki sam sposób jak tradycyjnej sekwencji wizyjnej. Na koniec powstałe w koderze MIV metadane są multipleksowane ze strumieniami wideo, tworząc jeden wspólny strumień bitowy.

DEKODOWANIE MIV

Tak jak koder MIV stanowi przetwarzanie wstępne sekwencji wielowidokowej, tak dekoder MIV jest przetwarzaniem końcowym po zdekodowaniu przesłanych strumieni wizyjnych (rys. 9).

Po zdemultipleksowaniu strumienia otrzymywane są 4 strumienie wizyjne oraz strumień metadanych. Każdy ze strumieni wizyjnych jest dekodowany z użyciem typowego dekodera wizyjnego, a zdekodowane sekwencje (atlasy) wraz z metadanymi są następnie przetwarzane przez dekoder MIV w celu wygenerowania widoku oglądanego przez końcowego użytkownika systemu wizji wszechogarniającej.

MIV Main

Jak podkreślono wcześniej, podstawowym zastosowaniem profilu głównego są proste, praktyczne systemy, w których kluczowa jest szybkość dekodowania. W związku z tym dekoder MIV, działający w trybie MIV Main (rys. 9) jest uproszczony i składa się z dwóch podstawowych etapów: rozpakowywania atlasów i syntezy widoków wirtualnych.

Rozpakowanie atlasów

Pierwszym etapem przetwarzania sekwencji wielowidokowej w dekoderze MIV jest odtworzenie widoków wejściowych, w których umieszczane są wszystkie łatki przesłane w ramach atlasów. Oczywiście nie wszystkie widoki są odtwarzane w całości, ponieważ część powierzchni widoków dodatkowych została bezpowrotnie usunięta w koderze, gdyż była nadmiarowa.

Synteza widoku wirtualnego

W drugim i ostatnim etapie dekodowania dokonywana jest synteza widoku wirtualnego prezentowanego widzowi. Wykorzystywane są do niej odtworzone widoki wejściowe, parametry kamer przesłane w strumieniu metadanych oraz informacja o położeniu i kierunku obserwacji sceny przez użytkownika. Sam proces syntezy nie jest zdefiniowany w normie MIV, a więc do wygenerowania widoku wirtualnego może zostać użyty – w zależności od potrzeb – dowolny syntezer. W szczególności, w praktycznym systemie wizji wszechogarniającej, w którym dekoder powinien działać w czasie rzeczywistym i generować widok wirtualny zaraz po aktualizacji położenia widza, powinien zostać użyty syntezer czasu rzeczywistego, np. [26], [24].

W oprogramowaniu modelowym MIV [37] zastosowany został algorytm VWS [25], działający wolniej, lecz umożliwiający uzyskanie wirtualnych widoków o wysokiej jakości.

Dekoder MIV Geometry Absent

W związku z koniecznością przeprowadzenia czasochłonnego procesu estymacji map głębi, dekoder w profilu MIV GA jest zdecydowanie bardziej wymagający pod względem złożoności obliczeniowej niż dekoder dla MIV Main [42]. Chociaż złożoność obliczeniowa estymacji głębi może być postrzegana jako wada używania tego profilu, możliwe jest znaczne ograniczenie jej przez estymację tylko dla części widoków, które są niezbędne do syntezy żądanego widoku wirtualnego. Bez dodatkowego kanału sprzężenia zwrotnego jest to niemożliwe w MIV Main, dla którego estymacja głębi po stronie kodera dla wszystkich widoków jest obciążeniem dla takich aplikacji, jak strumieniowanie wydarzeń na żywo.

Obecnie w badaniach dotyczących normy MIV, do celów estymacji głębi (zarówno po stronie kodera, jak i dekodera), wykorzystywane jest oprogramowanie **IVDE** (*Immersive Video Depth Estimation*) [45], które stanowi rozwinięcie metody estymacji głębi opartej na segmentacji obrazu [46]. Metoda ta jest rozwijana równolegle z modelem testowym MIV i zapewnia ulepszenia takie, jak zwiększenie jakości map głębi dla skompresowanych widoków wejściowych [47], zwiększające jej użyteczność w badaniach nad MIV GA.

BADANIA EKSPERYMENTALNE

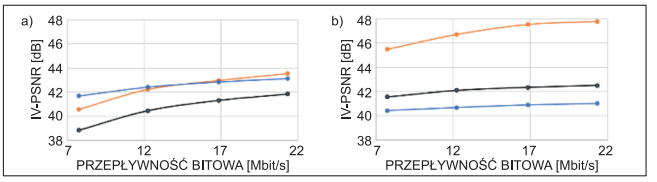
Przedstawiono tu porównanie obu profili normy MIV, zestawiając je z istniejącym podejściem, w którym widoki wejściowe i towarzyszące im głębie kodowane są w sposób niezależny, bez jakiejkolwiek redukcji redundancji międzywidokowej (*multiview simulcast*).

Metodologia badań

Eksperymenty zostały przeprowadzone, zgodnie z warunkami ustalonymi przez grupę ISO/IEC MPEG Video Coding (VC) [48], na zbiorze testowym zawierającym 15 zróżnicowanych sekwencji testowych, wliczając w to sekwencje naturalne zarejestrowane rzeczywistymi systemami wielokamerowymi i wygenerowane komputerowo, zarówno tradycyjne (perspektywiczne), jak i wszechkierunkowe (360°).

Do eksperymentów użyto oprogramowania modelowego MIV [37], a stworzone atlasy kodowano przy użyciu szybkiej implementacji kodera VVC [33]: VVenC [49]. W przypadku profilu MIV GA dodatkowo użyto oprogramowania IVDE [45] do estymacji głębi po stronie dekodera.

Każdą sekwencję testową zakodowano czterokrotnie, z różnymi wartościami parametru kwantyzacji QP [10], pokazując efektywność kompresji kodowania MIV dla różnych przepływności bitowych. Jakość syntezowanych widoków wirtualnych mierzono z użyciem dwóch obiektywnych metryk jakości: WS-PSNR [50], będącej rozszerzeniem miary PSNR dla sekwencji 360°, a także używanej w pracach grupy MPEG VC nad MIV metryki dostosowanej do oceny zniekształceń typowych dla wizji wszechogarniającej – IV-PSNR [51].



■ **Rys. 10. Krzywe RD dla sekwencji perspektywicznych (a) i 360° (b); zależność przedstawiono jako IV-PSNR [dB] w funkcji przepływności bitowej [Mbit/s]**

Dla obu testowanych profili MIV zastosowano praktyczne ograniczenia liczby i rozmiaru atlasów (4 atlasy, każdy z nich o rozdzielczości do 8 Mpix). W przypadku podejścia *multiview simulcast*, w którym są kodowane widoki wejściowe (o różnej rozdzielczości), zastosowano jedno ograniczenie – łączna liczba przesyłanych pikseli była mniejsza niż 4× 8 MPix.

Porównanie profili MIV

Porównanie trzech testowanych podejść ilustruje rys. 10 w formie krzywych **RD** (*Rate-Distortion*), obrazujących zależność jakości sekwencji prezentowanej widzowi od przepływności bitowej zakodowanego strumienia. Przedstawione krzywe zostały uśrednione po wszystkich sekwencjach testowych.

Jak pokazano, w przypadku sekwencji perspektywicznych efektywność kompresji za pomocą kodera MIV (obu testowanych profili) jest dużo wyższa, niż dla niezależnego kodowania widoków wejściowych i odpowiadających im map głębi.

Dla systemów o większej przepływności bardziej efektywne jest użycie profilu MIV Main. MIV GA sprawdza się lepiej dla systemów o mocno ograniczonej przepływności. W takim przypadku atlasy głębi stworzone przez koder MIV, działający w trybie MIV Main, są na tyle mocno kompresowane, że dokładność informacji przestrzennej staje się niezadowalająca. Dlatego wydajniej wyestymować ją w odbiorniku.

W przypadku sekwencji 360° profil MIV Main umożliwia diametralną poprawę jakości w porównaniu do podejścia *multiview simulcast*. Profil MIV GA wydaje się mniej efektywny, aczkolwiek nie jest to problem samego profilu, lecz ograniczeń istniejących metod estymacji map głębi [46], które aktualnie nie są w stanie wyestymować dobrej jakości map dla takiego rodzaju sekwencji.

■ **Tabela 1. Porównanie efektywności kodowania** *multiview simulcast* **i dwóch profili MIV przedstawione w postaci delt Bjontegaarda; ujemna wartość delty Bjontegaarda oznacza, iż MIV jest efektywniejszy, niż** *multiview simulcast*.

Typ		Nazwa sekwencji	MIV Main a *multiview simulcast*		MIV GA a *multiview simulcast*	
			WS-PSNR	IV-PSNR	WS-PSNR	IV-PSNR
Stworzona komputerowo	360°	Chess	lepiej	lepiej	681.4%	-32.6%
		ChessPieces	lepiej	lepiej	gorzej	-8.6%
		ClassroomVideo	-24.6%	-16.2%	129.4%	25.8%
		Hijack	-54.1%	-59.5%	gorzej	gorzej
		Museum	-18.8%	-25.8%	133.6%	42.3%
	Perspektywiczna	Cadillac	-4.3%	-24.8%	-74.6%	-72.6%
		Fan	-32.5%	-44.4%	-91.8%	-83.7%
		Kitchen	-34.6%	-54.9%	-39.0%	-24.9%
		Mirror	-38.5%	-45.9%	-67.9%	-67.6%
Naturalna		Carpark	-47.6%	-50.5%	-64.0%	-61.9%
		Fencing	-32.2%	-33.1%	-38.8%	-54.3%
		Frog	-6.1%	-24.4%	-61.5%	-61.7%
		Hall	-77.6%	-68.3%	-88.2%	-63.7%
		Painter	-16.2%	-29.3%	-68.2%	-58.8%
		Street	-18.8%	-40.7%	-51.1%	-56.9%

Wyniki dla poszczególnych sekwencji przedstawiono w tabeli 1. Dla każdej sekwencji zaprezentowano wartość delty Bjoentegaarda [52], określającej procentową zmianę przepływności przy zachowaniu takiej samej jakości, wyliczoną na podstawie odległości pomiędzy dwiema krzywymi RD. Deltę Bjontegaarda wyznaczono dla obu metryk używanych w pracach grupy MPEG VC: WS-PSNR [50] i IV-PSNR [51].

Brak wartości w tabeli 1 (np. dla sekwencji Chess) oznacza, iż krzywe są na tyle odległe od siebie, że niemożliwe jest wyznaczenie dla nich delty Bjontegaarda. W takim przypadku lepszą lub gorszą efektywność kompresji kodera MIV opisano słownie.

DALSZE PRACE

Norma MIV została ukończona w tym roku, a więc wydawać by się mogło, iż prace nad efektywnym kodowaniem wizji wszechogarniającej zostały zakończone z sukcesem.

Jednakże pomimo że MIV umożliwia skuteczną kompresję sekwencji wielowidokowych i ma spory potencjał do bycia normą powszechnie używaną, nie rozwiązuje wszystkich potencjalnych problemów i wyzwań towarzyszących przetwarzaniu wizji wszechogarniającej. Stąd też obecnie trwają prace nad drugą edycją normy MIV [53], która ma rozwijać techniki kompresji zastosowane w MIV, dodatkowo zwiększając jej wszechstronność.

Wśród głównych wyzwań stojących aktualnie przed ekspertami grupy MPEG VC zajmującymi się tworzeniem normy MIV można wymienić :

- efektywną kompresję scen nielambertowskich (duża liczba odbić i przezroczystości),
- możliwość użycia systemów z dynamicznie zmieniającymi się parametrami (ruchome kamery, możliwość dynamicznej zmiany ogniskowej poszczególnych kamer),
- możliwość użycia dodatkowych sensorów głębi umożliwiających szybszą lub lepszą estymację głębi dla pozostałych kamer systemu,
- rozszerzenia w kierunku wirtualnej i rozszerzonej rzeczywistości oraz możliwość wstawiania do sceny zarejestrowanej systemem wielokamerowym obiektów reprezentowanych przez chmury punktów.

Praca finansowana ze środków przyznanych przez Ministerstwo Edukacji i Nauki.

LITERATURA

[1] Jeong J.B. i in. 2019. „Towards 3DoF+ 360 video streaming system for immersive media". *IEEE Access* 7: 136399-136408.

[2] Tanimoto M. i in. 2012. „FTV for 3-D Spatial Communication". *Proc. of the IEEE* 100 (4): 905-917.

[3] Stankiewicz O. i in. 2018. „A free-viewpoint television system for horizontal virtual navigation". *IEEE Tr. on Multimedia* 20 (8): 2182-2195.

[4] Wien M. i in. 2019. „Standardization status of immersive video coding". *IEEE J. on Emerging and Selected Topics in Circuits and Systems* 9 (9).

[5] Oyman O. i in. 2019. „Virtual Reality Industry Forum's View on State of the Immersive Media Industry". *SMPTE Motion Imaging Journal* 128 (8): 91-96.

[6] Hayes C. 2021. „Olympic games broadcasting: Immerse yourself in the olympics this summer". *Engineering & Technology* 16 (6): 2225.

[7] Wu F. 2021. *Construction of Digital Dynamic Sports System Platform Based on VR Technology.* CIPAE Conference 2021.

[8] *Alhadad S.A. 2018. „Enhancing Sma*rt Sport Management based on Information Technology". *IOSR-JSPE* 5 (5): 19-26.

[9] Xu S. i C.Tang 2021. *Application of AI technology in sports competitions.* AIVRV 2021*, Sanya, Chiny.*

[10] Sullivan G. i in. 2012. „Overview of the High Efficiency Video Coding (HEVC) standard". *IEEE Tr. on Circuits and Systems for Video Technology* 22 (12): 1649-1668.

[11] ISO/IEC. 2022. Standard ISO/IEC DIS 23090-12, Information technology – Coded Representation of Immersive Media – Part 12: MPEG immersive video.

[12] Müller K. i in. 2011. „3-D Video Representation Using Depth Maps". *Proc. of the IEEE* 99 (4): 643-656.

[13] ISO/IEC. 2021. Standard ISO/IEC 23090-5, Information technology – Coded Representation of Immersive Media – Part 5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC).

[14] ang E.S. i in. 2019. „Video-Based Point-Cloud-Compression Standard in MPEG: From Evidence Collection to Committee Draft [Standards in a Nutshell]". *IEEE Signal Processing Magazine* 36 (3): 118-123.

[15] https://learnopencv.com/camera-calibration-using-opencv/

[16] Szeliski R. 2022. *Computer Vision: Algorithms and Applications*. The University of Washington.

[17] Dziembowski A., M. Domański 2018. *Adaptive color correction in virtual view synthesis.* 3DTV-CON, Helsinki, Finlandia.

[18] Dziembowski A. i in. „Color correction for immersive video applications". *IEEE Access* 9: 75626-75640.

[19] Laga H. i in. 2022. „Survey on Deep Learning Techniques for Stereo-Based Depth Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 1738-1764.

[20] Mieloch D. i in. 2021. „Depth Map Refinement for Immersive Video". *IEEE Access* 9: 10778-10788.

[21] Fachada S. i in. 2018. *Depth image based view synthesis with multiple reference views for virtual reality*. 3DTV-CON, Helsinki, Finlandia.

[22] ISO/IEC. 2019. *Versatile view synthesizer (VVS) 2.0 manual*. Dok. ISO/IEC JTC 1/SC29/WG11 MPEG N18172.

[23] Dziembowski A. i in. 2019. *Virtual view synthesis for 3DoF+ video*. PCS 2019, Ningbo, Chiny.

[24] Stankowski J., A. Dziembowski 2022. *Real-time CPU-based view synthesis for omnidirectional video*. WSCG 2022, Pilzno, Czechy.

[25] Fleureau J. i in. 2020. *An Immersive Video Experience with Real-Time View Synthesis Leveraging the Upcoming MIV Distribution Standard*. ICMEW 2020.

[26] Stankowski J., A. Dziembowski . 2020. *Fast view synthesis for immersive video systems.* WSCG 20, Pilzno, Czechy.

[27] „2021 TV Video Specifications". https://developer.samsung.com/smarttv/develop/specifications/media-specifications/2021-tv-video-specifications.html.

[28] „H.264/H.265 Video Codec Unit" https://www.xilinx.com/products/intellectual-property/v-vcu.html#overview.

[29] Merkle P. i in. 2006. *Efficient Compression of Multi-View Video Exploiting Inter-View Dependencies Based on H.264/MPEG4-AVC*. ICME 2006, Toronto, Kanada.

[30] Tech G. i in. 2016. „Overview of the Multiview and 3D Extensions of High Efficiency Video Coding". *IEEE Tr. on Circuits and Systems for Video Technology* 26 (1): 35-49.

[31] Boyce J. i in. 2021. „MPEG Immersive Video Coding Standard". *Proc. of the IEEE* 109 (9): 1521-1536.

[32] Grzelka A. i in. 2022. *The study of the video encoder efficiency in decoder-side* depth *estimation applications*. WSCG 2022, Pilzno, Czechy.

[33] Bross B. i in. 2021. „Overview of the Versatile Video Coding (VVC) standard and its applications". *IEEE Tr. on Circuits and Systems for Video Technology* 31 (10): 3736-3764.

[34] Sullivan G., T. Wiegand 2005. „Video compression – from concepts to the H.264/AVC standard". *Proc. of the IEEE* 93.

[35] Szekiełda J. i in. 2021. *The influence of coding tools on immersive video coding.* WSCG 2021, Pilzno, Czechy.

[36] Samelak J. i in. 2021. *Efficient immersive video compression using Screen Content Coding.* WSCG 2021, Pilzno, Czechy.

[37] ISO/IEC. 2021. *Test Model 11 for MPEG Immersive Video.* Dok. ISO/IEC JTC1/SC29/WG04 MPEG VC N0142, Online.

[38] Kaufman L., P.J. Rousseeuw 1990. *Partitioning Around Medoids (Program PAM).* Wiley Series in Probability and Statistics, Hoboken, USA: 68–125.

[39] Mieloch D. i in. 2022. *Color-dependent pruning in immersive video coding.* WSCG 2022, Pilzno, Czechy.

[40] Jylänki J. 2010. *A thousand ways to pack the bin—A practical approach to two-dimensional rectangle bin packing*. Tech. Rep., Online.

[41] Dziembowski A. i in. 2022. *Spatiotemporal redundancy removal in immersive video coding*. WSCG 2022, Pilzno, Czechy.

[42] Garus P. i in. 2022. "Immersive Video Coding: Should Geometry Information Be Transmitted as Depth Maps?". *IEEE Tr. on Circuits and Systems for Video Technology* 32 (5)*:* 3250-3264*.*

[43] Mieloch D. i in. 2022. „Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video". *IEEE Tr. on Circuits and Systems for Video Technology, Early Access.*

[44] Szydełko B. i in. 2022. „Recursive block splitting in feature-driven decoder-side depth estimation". *ETRI Journal* 44 (1): 38-50.

[45] ISO/IEC. 2021. *Manual of IVDE 3.0*. Dok. ISO/IEC JTC1/SC29/WG4 MPEG VC/N0058.

[46] Mieloch D. i in. 2020. „Depth Map Estimation for Free-Viewpoint Television and Virtual Navigation". *IEEE Access* 8: 5760-5776.

[47] Mieloch D. i in. 2021. *Point-to-Block Matching in Depth Estimation*. WSCG 2021, Pilzno, Czechy.

[48] ISO/IEC. 2022. *Common Test Conditions for MPEG Immersive Video*. Dok. ISO/IEC JTC 1/SC 29/WG04 N0169.

[49] Wieckowski A. i in. 2021. *VVenC: an open optimized VVC encoder in versatile application scenarios*. *Proc. SPIE* 11842, App. of Digital Image Proc. XLIV.

[50] Sun Y. i in. 2017. „Weighted-to-spherically-uniform quality evaluation for omnidirectional video". *IEEE Signal Processing Letters* 24 (9): 1408-1412.

[51] Dziembowski A. i in. 2022. „IV-PSNR – the objective quality metric for immersive video applications". *IEEE Tr. on Circuits and Systems for Video Technology*, Early Access.

[52] Bjoentegaard G. 2001. *Calculation of average PSNR differences between RD-Curve*s. ITU-T VCEG Meeting, Austin, USA.

[53] ISO/IEC. 2022. *Use cases and requirements for MIV – edition-2 (final*). Dok. ISO/IEC JTC 1/SC 29/WG02 N0157.

Kamil STANIEC*

Wydajność wąskopasmowych systemów Internetu rzeczy w trudnych warunkach propagacyjnych i zakłóceniowych

The performance of narrowband Internet of Things systems under harsh propagation and interference conditions

Internet rzeczy jest zjawiskiem wywodzącym się od sieci sensorowych, te zaś są wynikiem zrozumienia korzyści wynikających z „dania rzeczom głosu", czyli umożliwienia miliardom czujników, mierników, liczników, sond, sensorów komunikowania swoich odczytów bez ingerencji człowieka. Z uwagi na niskie wymagania dotyczące przepustowości, systemy te wykorzystują wąskie kanały częstotliwościowe, ze względów propagacyjnych najczęściej ulokowane w pasmach poniżej 1 GHz, co sprzyja dużym zasięgom albo dobrej penetracji sygnału do wnętrz budynków. Omówiono kilka wybranych systemów transmisyjnych dla podkreślenia różnorodności rozwiązań w warstwie fizycznej. Odrębny rozdział poświęcono zagadnieniu zagłuszania, które wydaje się jedną z najskuteczniejszych – mimo iż mało wyrafinowanych – metod zaburzania czy blokowania transmisji w wąskopasmowych systemach IoT. Przegląd literaturowy wskazuje na liczne próby zapobiegania skutkom zagłuszania lub – alternatywnie – wykorzystania zagłuszania do ochrony transmisji przed podsłuchem. Z powodu wysokich czułości systemów IoT, pomiar ich wydajności w warunkach zakłóceniowych bądź wielodrogowych wymaga odpowiedniej adaptacji zaplecza laboratoryjnego, umożliwia jednak uzyskanie miarodajnych wyników dotyczących rekomendowanych nastaw w warunkach granicznie wysokich tłumień czy niskich wartości nośnej do szumu i zakłóceń oraz transmisji wielodrogowej.
Słowa kluczowe: sieci niskich przepustowości, ultrawąskopasmowe, LoRa, Weightless, SigFox, komora bezechowa, rewerberacyjna

The Internet of Things is a phenomenon derived from sensor networks, and these are the result of understanding the benefits of "giving things a voice", i.e., enabling billions of sensors, meters, gauges, probes, sensors to communicate their readings without human intervention. Due to low bandwidth requirements, these systems use narrow frequency channels, most often located in bands below 1 GHz for propagation reasons, which favors long ranges or good signal penetration into building interiors. Several selected transmission systems were discussed to emphasize the diversity of solutions in the physical layer, devoting a separate chapter to the jamming issue, which seems to be one of the most effective – although not very sophisticated – methods of disrupting or blocking transmissions in narrowband IoT systems. The literature review shows numerous attempts to prevent the effects of jamming or, alternatively, to use jamming for protecting transmissions from eavesdropping. Due to the high sensitivity of IoT systems, measuring their performance in disturbed or multi-path conditions requires appropriate adaptation of the laboratory facilities, however, it allows obtaining reliable results regarding the recommended settings in conditions of extremely high attenuation or low carrier to noise and interference values as well as multi-path transmission.
Key words: low throughput networks, ultranarrowband, LoRa, Weightless, SigFox, anechoic chamber, reverberation

DOI: 10.15199/59.2022.4.5

GENEZA INTERNETU RZECZY

W dobie dojrzałości Internetu dostrzeżono, iż sensory, mierniki i detektory także tworzą swoistą „społeczność" sieciową, uczestniczącą w ruchu sieciowym i generującą dane o nie mniejszej ważności, niż te wytwarzane przez człowieka, a często strategiczne dla podtrzymania jakości życia, kontroli procesów przemysłowych czy monitorowania stanu środowiska naturalnego. Urządzenia te, czy ogólniej – obiekty, mogą komunikować się albo z człowiekiem (operatorem), albo z innymi urządzeniami. W tym drugim przypadku określane są też czasem w literaturze akronimem **MTD** (*Machine-Type Device*), czyli urządzeniami typu maszynowego, które do wymiany danych między sobą nie wymagają ludzkiej ingerencji. Sam proces komunikacji między nimi określany jest jako **M2M** (*Machine-to-Machine*) lub **MTC** (*Machine-Type Communications*). Stanowi on jeden z paradygmatów tzw. Internetu rzeczy **IoT** (*Internet of Things*), czyli takiego odłamu ogólnoświatowej sieci, w którym urządzenia MTD są partnerami równorzędnymi względem ich ludzkich odpowiedników, jeśli chodzi o ważność generowanego przez nie ruchu telekomunikacyjnego, coraz bardziej niezbędnego do podtrzymania sprawnego funkcjonowania naszej cywilizacji technicznej.

Termin *Internet rzeczy* został po razy pierwszy użyty w trakcie jednej z prezentacji firmowych dla Procter & Gamble przez Kevina Ashtona w 1999 r. [1], chociaż sam autor obecnie promuje nazwę *Internetu dla rzeczy* [2]. Początkowo przewidywano, iż technologią, służącą do realizacji założeń IoT, będzie identyfikacja radiowa **RFID** (*Radio Frequency Identification*), która miała być narzędziem wykonawczym rozmaitych zadań zlecanych przez komputery obiektom (rzeczom), takim jak aktuatory, liczniki, sterowniki itp. [3]-[7]. Za nieformalną datę pojawienia się Internetu rzeczy, jako faktu, można natomiast przyjąć przełom lat 2008 – 2009 r. Wówczas to (według badań własnych Cisco opisanych w [8]) iloraz komunikujących się w sieci „obiektów" do ludzi przekroczył wartość 1, wynosząc już w 2010 r. 1,84, podczas gdy jeszcze w 2003 r. był szacowany na 0,08.

Ideę tę zaprezentowano na rys. 1, umożliwiając jednocześnie dostrzeżenie podziału systemów Internetu rzeczy na rozwiąza-

* Katedra Telekomunikacji i Teleinformatyki, Wydział Informatyki i Telekomunikacji. Politechnika Wrocławska, e-mail: kamil.staniec@pwr.edu.pl

nia komórkowe **CIoT** (*Cellular IoT*) oraz pozostałe, znane jako **LTN** (*Low Throughput Network*). Wspólnym obszarem napływu danych, czy to pochodzących od tradycyjnych systemów (komórkowych, szerokopasmowego dostępu do Internetu itp.), czy od systemów, w których informację generują czujniki, liczniki i mierniki, jest chmura, w której realizowane są zarówno usługi sieciowe, jak i bazodanowe. Nadejście ery Internetu rzeczy spowodowało jednak modyfikacje także w tym zakresie, m.in. doprowadzając do powstania koncepcji *obliczeń we mgle* (*fog computing*), efektywniejszej z energetycznego i transmisyjnego punktu widzenia oraz ze względów cyberbezpieczeństwa niż obliczenia w chmurze, a także innych koncepcji, takich jak *obliczenia brzegowe* (*edge computing*), *inteligencja tła* (*ambient intelligence*) itp.

Z usługowo-technologicznego punktu widzenia struktura Internetu rzeczy składa się z tzw. domen pionowych i poziomych. Pionowe można utożsamić z konkretnymi obszarami aplikacyjnymi i powiązanymi z nimi firmami oraz procesami standaryzacyjnymi (choć na różną skalę), odbywającymi się w obrębie każdego z tych obszarów. Domeny poziome z kolei wiążą się z systemami telekomunikacyjnymi, służącymi do realizacji założeń i zadań określonych w domenach pionowych, w zakresie tworzenia rzeczywistych rozwiązań komunikacyjnych i tym samym stanowiącymi dla nich wspólną bazę.

Ze względu na typ zastosowania technik Internetu rzeczy, wydaje się iż z użytkowego punktu widzenia daje się on podzielić na dwa główne nurty: masowy i krytyczny. Linię podziału stanowią tutaj parametry jakościowe transmisji wymagane w każdym z nich. Masowy IoT ([9], [10]), któremu dedykowane są systemy opisywane w niniejszym artykule, związany z przenoszeniem ruchu na potrzeby inteligentnych budynków, logistyki, śledzenia i zarządzania flotą, monitorowania upraw, hodowli oraz w szerokim zakresie – inteligentne pomiary mediów wymagają spełnienia niżej wymienionych warunków, przy tym detalicznego opisu charakterystyki ruchu MTC, wyrażonego w postaci tzw. modelu mobilnego autonomicznego raportowania **MAR** (*Mobile Autonomous Reporting*) [11]:

- niskich kosztów i bardzo małego poboru energii elektrycznej, zapewniającego bezobsługową pracę urządzenia przez lata;
- małych objętości danych;
- zdolności obsługi wielkiej liczby urządzeń, liczonej nawet w dziesiątkach tysięcy na punkt dostępowy (stację bazową);
- dużej czułości (od 20 dB do 30 dB większej niż standardowe systemy bezprzewodowe, np. **GPRS** – *General Packet Radio Service*).

Sieci IoT z grupy LTN mogą być realizowane przy użyciu jednej z dwóch podstawowych technik: z rozpraszaniem bezpośrednim **SS** (*Spread Spectrum*) oraz z transmisją ultrawąskopasmową **UNB** (*Ultra-Narrowband*). Niezależnie od wyboru, najważniejszym celem jest uzyskanie określonego zysku przetwarzania G_p na tyle dużego, aby zapewnić możliwie największy zakres tolerancji dla tłumienia propagacyjnego, określanego akronimem **MCL** (*Maximum Coupling Loss*). MCL z kolei ma w zamierzeniu przełożyć się na jeden z dwóch efektów:

- zdolność pracy systemu przy realnych zasięgach rzędu kilkunastu, a nawet kilkudziesięciu kilometrów;
- dobrą penetrację sygnału radiowego do miejsc trudno dostępnych, np. przez warstwy ziemi i betonu, oddzielających licznik/miernik wody czy gazu, które w tradycyjnych systemach często stanowią zaporę trudną do pokonania; ponieważ tego typu scenariusz jest typowy dla zastosowań IoT, techniki transmisji stosowane w systemach Internetu rzeczy powinny dysponować odpowiednimi mechanizmami, umożliwiającymi skompensowanie wpływu tłumienia dodatkowego wywołanego wspomnianymi przeszkodami propagacyjnymi, najczęściej mieszczącego się w zakresie 20–30 dB.

PASMA CZĘSTOTLIWOŚCIOWE PRACY SYSTEMÓW INTERNETU RZECZY

Niniejszy rozdział jest poświęcony zarówno aspektom widmowym, jak i pewnym normom regulującym, a przez to ułatwiającym jednoczesną pracę wielu urządzeń końcowych, przez redukowanie ryzyka powstania wzajemnych zakłóceń. Jak wynika z zaleceń **ERC** (*European Research Council*) i **ETSI** (*European Telecommunications Standards Institute*), zestawionych w tabeli 1, dla większości podpasm zakresu 863-870 MHz określono dość rygorystyczne limity dla cyklu pracy **DC** (*Duty Cycle*), oscylujące zazwyczaj wokół 1%, nigdy natomiast nieprzekraczające 10%. Z zestawienia tego wynika jednak także, iż dopuszcza się, aby limity cyklu pracy były zamiennie stosowane z jedną z technik tzw. uprzejmego dostępu do medium **PMA** (*Polite Medium Access*). Ich głównym celem jest wyposażenie urządzeń IoT (czyli rzeczy) w pewien stopień „świadomości" własnego otoczenia elektromagnetycznego (EM), aby unikać kolizji z transmisjami od innych urządzeń końcowych. Mechanizmami służącymi do realizacji tego celu jest technika szacowania zajętości kanału **CCA** (*Clear Channel Assessment*) [12]. Definiuje ona dwie metody uprzejmego dostępu do medium: zwinną adaptację częstotliwości **AFA** (*Adaptive Frequency Agility*) [12] oraz nasłuch przed nadawaniem **LBT** (*Listen Before Talk*) [13], opracowane właśnie z myślą o współdzieleniu wspólnego, relatywnie wąskiego pasma przez dużą liczbę użytkowników.

Dla stworzenia korzystnych warunków propagacyjnych, powodujących stosunkowo niskie tłumienia sygnału, dla systemów IoT zakłada się pracę w pasmach poniżej 1 GHz. Wymaganie zaś ogólnej dostępności i niekosztownego użytkowania ograniczyło te pasma do grupy nielicencjonowanych pasm **ISM** (*Industrial, Scientific, Medical*), przedstawionych w tabeli 1. Bazowymi dokumentami definiującymi te pasma są: specyfikacja ETSI EN 300 220-2 [12], rekomendacja ERC Rec. 70-03 [15] oraz Europejska Tablica Przydziału Częstotliwości [16], oryginalnie dedykowane m.in. dla urządzeń krótkozasięgowych **SRD** (*Short-Range Devices*), zaadaptowanych następnie także dla potrzeb dalekosiężnych wąskopasmowych systemów IoT. Jak widać, pasma te z definicji umożliwiają transmisję wąskopasmową, dodatkowo z istotnymi ograniczeniami głównie w zakresie maksymalnej skutecznej mocy promieniowanej **ERP** (*Equivalent Radiated Power*) oraz cyklu pracy *DC*. Ograniczenia te określono w artykule 8(2) Dyrektywy 2014/53/EU [17], definiującym specyfikację interfejsów radiowych oraz przydzielanie im klas urządzeń radiowych, oraz w [14], gdzie podano wartości minimalne dla gęstości mocy oraz inne wymagania (dotyczące DC i LBT) dla systemów pracujących w paśmie poniżej 1 GHz, stosujących rozpraszanie widma metodą skakania po częstotliwościach **FHSS** (*Frequency Hopping Spread Spectrum*) lub rozpraszania bezpośredniego **DSSS** (*Direct Sequence Spread Spectrum*).

METODY ROZPRASZANIA WIDMA STOSOWANE W SYSTEMACH INTERNETU RZECZY

Jedną z najpopularniejszych odmian techniki rozpraszania widma jest kluczowanie bezpośrednie DSSS, w którym efekt rozproszenia uzyskuje się na drodze operacji alternatywy wyklu-

■ **Tabela 1. Główne pasma częstotliwości ISM (według [12]-[16]), przeznaczone dla wąskopasmowych systemów IoT**

Symbol pasma według [11]	Zakres częstotliwości	Maks. ERP	Ograniczenie dostępu do łącza	Maks. BW	Potencjalne zastosowanie
D	169,400-169,475 MHz	500 mW	DC ≤ 10%	50 kHz	Bezprzewodowe aparaty słuchowe **ALD** (*Assistive Listening Device*), urządzenia krótkozasięgowe
E	169,400-169,4875 MHz	10 mW	DC ≤ 0,1%	Całe pasmo	
F	169,4875-169,5875 MHz	10 mW	DC ≤ 0,001% DC ≤ 0,1% w godz. 00:00-06.00 czasu lokalnego	Całe pasmo	
G	169,5875-169,8125	10 mW	DC ≤ 0,1%	Całe pasmo	
H	433,05-434,79 MHz	10 mW	DC =10%	Całe pasmo	Niekreślone urządzenia krótkozasięgowe (*non-specific SRD*)
I	433,05-434,79 MHz	1mW [1]	—	Całe pasmo	
J	434,04-434,79 MHz	10 mW	—	25 kHz	
K	863-865 MHz	10 mW	—	Całe pasmo[3]	Mikrofony, bezprzewodowe systemy audio, np. do monitoringu dzieci, słuchawek
		25 mW	DC ≤ 0,1% lub PMA[2]		Nieokreślone urządzenia krótkozasięgowe (*non-specific SRD*)
L	865-868 MHz	25 mW [4]	DC ≤ 1% lub PMA[2]	Całe pasmo[3]	Czytniki i etykiety (tagi) RFID
M	**868-868,60 MHz**	**25 mW**	**DC ≤ 1% lub PMA[2]**	Całe pasmo[3]	IEEE 802.15.4 (ZigBee), LoRa, SigFox, Wireless-M, Z-Wave
N	868,70-869,200 MHz	25 mW	DC ≤ 0,1% lub PMA[2]	Całe pasmo[3]	Przemysłowe, Wireless-M
O	869,40-869,65 MHz	25 mW	DC ≤ 0,1% lub PMA[2]	Całe pasmo	Przemysłowe i komunikacyjne (Wireless-M, LoRa, SigFox)
P	**869,40-869,65 MHz**	**500 mW**	**DC ≤ 10% lub PMA[2]**	**Całe pasmo**	
Q	869,70-870,00 MHz	5 mW	DC Bez ograniczeń	Całe pasmo	Systemy alarmów przeciwpożarowych i antywłamaniowych
R	869,70-870,00 MHz	25 mW	DC ≤ 1% lub PMA[2]	Całe pasmo	
X (lub 'g' wg [15])	863,00-870,00 MHz	25 mW	DC ≤ 0,1% lub PMA[2]	Całe pasmo[3]	Wąsko- i szerokopasmowe systemy inne niż krótkozasięgowe [15]

[1] dodatkowo, dla BW>250 kHz, widmowa gęstość mocy **PSD** (*Power Spectral Density*) powinna wynosić -13 dBm/10 kHz
[2] realizowanego na drodze stosowania jednej z technik: LBT lub AFA (jak podano jawnie w analogicznej tabeli w [13])
[3] z wyłączeniem zastosowań o krótkim zasięgu SRD ograniczonych do BW=300 kHz
[4] PSD=-4,5 dBm/100 kHz, możliwe także PSD=+6,2 dBm/100 kHz przy ograniczeniu pasma pracy jedynie do zakresu 865-868 MHz

czającej XOR (*exclusive OR*) strumienia bitów informacyjnych (w bitach/s) z szybkozmienną sekwencją pseudolosową. W niektórych systemach (np. UMTS) sekwencje te często są ciągami wzajemnie ortogonalnymi, co zapewnia możliwości rozróżniania sygnałów pochodzących od różnych nadajników. Różnica gęstości mocy $S(f=0)$ pomiędzy obiema wersjami sygnału, tj. skupionym, zajmującym szerokość BW_{inf} (*bandwidth*) i rozproszonym o szerokości BW_{SS}, umożliwia sformułowanie – uniwersalnej dla wszystkich technik rozpraszania widma – zasady definiującej G_p, zgodnie ze wzorem (1). Dla podkreślenia ogólności tej zasady, właściwej także dla innych metod SS, takich jak **CSS** (*Chirp Spread Spectrum*), rozpraszania przez skakanie po czasie **THSS** (*Time Hopping SS*) czy skakania po częstotliwości FHSS:

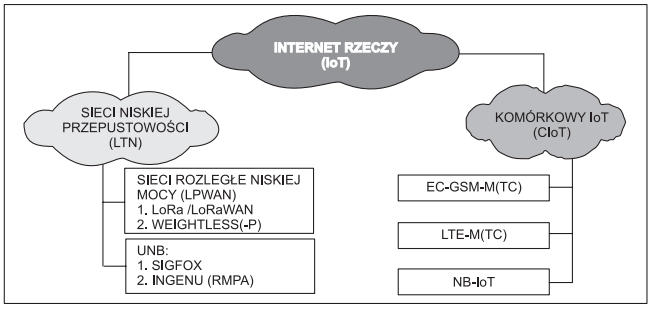
$$G_p[dB] = 10\log\left({BW_{SS}}/{BW_{inf}}\right) \tag{1}$$

Innym sposobem rozpraszania widma jest generowanie sygnałów typu *chirp* [18]-[21], zwanych także świergotowymi, ze względu na analogię do dźwięków wydawanych przez nietoperze i delfiny. Podstawy teoretyczne tej metody zostaną tu omówione szczegółowo, ponieważ jej odmiana – umożliwiająca stosowanie modulacji wielowartościowych – jest stosowana w jednym z kluczowych systemów podgrupy LPWAN, czyli LoRaWAN (a konkretnie w jego warstwie fizycznej – LoRa). W technice tej (CSS) rozproszenie uzyskuje się przez ciągłą zmianę częstotliwości w czasie trwania symbolu T_s (świergotu) w zakresie BW [Hz], czyli od częstotliwości początkowej (niższej) f_0 do końcowej (wyższej) f_1, zdefiniowanych względem nośnej częstotliwości środkowej f_c, będącej częstotliwością chwilową w czasie t=0. W rezultacie częstotliwość chwilowa $f_i(t)$ zmienia się, zgodnie ze wzorem (2), w zakresie BW [Hz], w przedziale czasowym $T_s/2 \le t \le T_s/2$, przy czym zmiana ta może mieć charakter liniowy bądź wykładniczy. Parametr μ w tym wzorze przyjmuje wartość +1 (*downchirp*) lub –1 (*upchirp*), co odpowiada impulsowi z narastającą lub opadającą częstotliwością. W podstawowej implementacji CSS wartość μ determinuje wartość konkretnego bitu. Przy tym zazwyczaj μ =+1 odpowiada logicznej jedynce, μ = –1 zaś logicznemu zeru, choć pozostaje to kwestią konwencji przyjętej w danym systemie telekomunikacyjnym. Parametr α z kolei wyznacza tempo zmian częstotliwości i zadany jest wzorem (3). Wynika z niego, że w liniowych sygnałach świergotowych ich szerokość kanału BW jest różnicą częstotliwości f_1 i f_0, zaś w wykładniczych – ich ilorazem.

$$f_i(t) = \begin{cases} f_c + \mu\alpha t = f_c + \mu\frac{BW}{T_s}t & \textit{dla liniowego} \\ f_c \cdot \alpha^{\mu t} = f_c \cdot \left[\left(\frac{f_1}{f_0}\right)^{\frac{1}{T_s}}\right]^{\mu t} & \textit{dla wykładniczego} \end{cases} \tag{2}$$

$$\alpha = \frac{1}{2\pi}\frac{d^2\varphi(t)}{t^2} = \begin{cases} \frac{BW}{T_s} = \frac{(f_1 - f_0)}{T_s} & \textit{dla liniowego} \\ BW^{\frac{1}{T_s}} = \left(\frac{f_1}{f_0}\right)^{\frac{1}{T_s}} & \textit{dla wykładniczego} \end{cases} \tag{3}$$

Nazwa *systemy ultrawąskopasmowe* (UNB) odnosi się do faktu, że kanały radiowe używane do dwukierunkowej komuni-

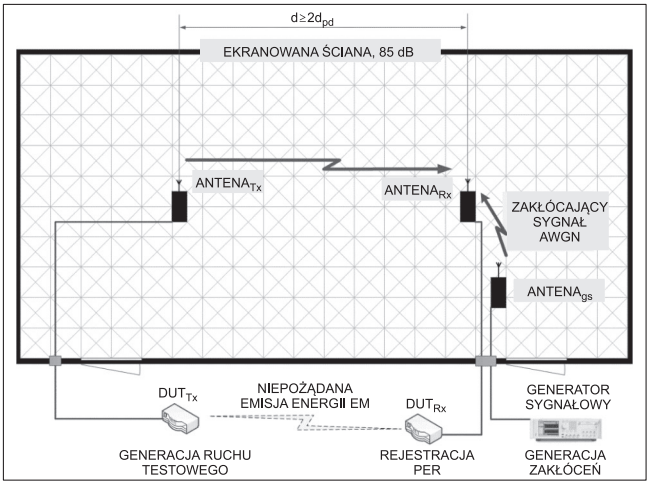


■ **Rys. 1. Podział systemów Internetu rzeczy**

kacji są wyraźnie węższe w porównaniu do całkowitej dostępnej szerokości pasma (zazwyczaj w proporcji ok. 1:100) [22]. Podczas gdy typowe wąskopasmowe systemy radiowe LTN czy CIoT stosują kanały o BW rzędu od kilkudziesięciu do kilkuset kHz [23], sygnały UNB wymagają BW w zakresie od stu do kilkuset Hz. Transmisja tego typu implikuje zatem przynajmniej dwie oczywiste korzyści.

● Można założyć, że charakter zaników wielodrogowych będzie typu płaskiego, gdyż dla kanałów UNB nietrudno jest spełnić założenie, aby szerokość kanału była (dużo) mniejsza niż szerokość pasma koherencji BW_{coh} zadanego wzorem (4), w którym τ_{RMS} (*root mean square delay spread*) oznacza rozmycie opóźnienia [24, 25].

● Istnieje możliwość jednoczesnej, bezzakłóceniowej transmisji od wielu terminali jednocześnie z powodu dostępności dużej częstotliwościowo liczby kanałów. Przykładowo, system SigFox, pracujący w zakresie o łącznej szerokości 192 kHz, dysponuje w tym paśmie liczbą 1920 rozłącznych kanałów „w górę" o indywidualnej szerokości 100 Hz (choć w rzeczywistości dobierane są losowo bez zachowania jakiegokolwiek rastru kanałowego). Ten fakt zaś implikuje z kolei możliwość znaczącego uproszczenia protokołu wielodostępu, a zatem zredukowania wiążącego się z nim narzutu w postaci dodatkowego opóźnienia, zmniejszenia szybkości transmisji, wyeliminowania strat energii na potrzeby obsługi procedur CCA czy zminimalizowania niezbędnej sygnalizacji. Typową metodą realizacji wielodostępu w systemach ultrawąskopasmowych jest wielodostęp z losowym podziałem czasu i częstotliwości **RFTDMA** (*Random Frequency Time Division Multiple Access*), opisany np. w [26] (rozdz. 4.3.3). W metodzie tej najważniejszymi cechami są: asynchroniczny dostęp węzła (urządzenia końcowego) do medium, losowość przydziału zarówno w dziedzinie czasu, jak i częstotliwości oraz brak protokołu rywalizacyjnego.



■ **Rys. 2. Schemat zestawu pomiarowego do badań odporności/wydajności systemów TLN na zakłócenia EM z wykorzystaniem komory bezechowej**

$$BW_{coh} = \frac{\sqrt{3}}{2\pi \cdot \tau_{rms}} \tag{4}$$

PREZENTACJA WARSTWY FIZYCZNEJ W PRZYKŁADOWYCH SYSTEMACH IOT

Poniższy przegląd nie jest wyczerpujący, lecz prezentuje trzy przykłady systemów dominujących w grupie LTN, umożliwiając jednocześnie dostrzeżenie rozległości różnic w realizacji warstwy fizycznej. To – jak się okaże w dalszej części artykułu – ma decydujący wpływ na osiągi tych systemów w ekstremalnych warunkach propagacyjnych.

Weightless

Weightless jest otwartym standardem komunikacyjnym przeznaczonym do obsługi dalekosiężnych łączy w warunkach silnego tłumienia propagacyjnego, w obecności zakłóceń, opracowanym i wspieranym przez Weightless SIG. System zaprojektowano do realizacji transmisji dwukierunkowej, w pełni synchronicznej, na potrzeby sieci dalekiego zasięgu Internetu rzeczy o ograniczonej przepustowości i zredukowanych wymaganiach dotyczących opóźnień. Został on przystosowany do pracy w pasmach częstotliwości ISM poniżej 1 GHz [27]. Ze względu na prawne ograniczenia obowiązujące w Europie, główna uwaga skoncentruje się jednak na pasmie X band (863–870 MHz), a w szczególności na podpasmach M oraz P, zgodnie z tabelą 1, dla których zdefiniowano następujące ograniczenia na ERP oraz DC:

- podpasmo M: 868,00–868,60 MHz, ERP=25 mW, $DC \leq 1\%$;
- podpasmo P: 869,40-869,65 MHz, ERP=500 mW, $DC \leq 10\%$.

System ten jest – obok LoRaWAN – jednym z czołowych „graczy" w przestrzeni systemów LPWAN.

Podobnie jak w przypadku LoRa, w systemie Weightless również stosuje się transmisję z widmem rozproszonym dla polepszenia odporności na zakłócenia. W tym przypadku jednak widmo jest rozpraszane bezpośrednio, z wykorzystaniem technik przeplotu bitowego i randomizacji danych (*whitening*). Rozpraszanie to jednak stosowane jest jedynie trybie z modulacją OQPSK. W drugim trybie, wykorzystującym **GMSK** (*Gaussian Minimum Shift Keying*), nie stosuje się rozpraszania widma. W pierwszym z trybów dostępne są dwie wartości SF: 4 oraz 8, przy czym w SF=8 zdefiniowano dwa zestawy sekwencji rozpraszających, tj. $(8,1)_0$ i $(8,1)_1$, stosowane odpowiednio na bitach parzystych i nieparzystych. W słowie synchronizacyjnym używa się natomiast jedynie sekwencji $(8,1)_0$.

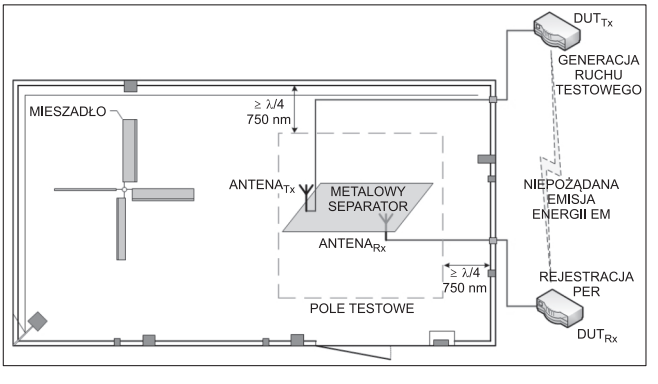
LoRaWAN

LoRaWAN (*Long Range Widea Area Network*) jest dalekosiężnym otwartym systemem telekomunikacyjnym, wspieranym obecnie przez LoRa Alliance. Pod nazwą LoRa należy natomiast rozumieć zamknięte, własnościowe rozwiązanie dla warstwy fizycznej systemu [28], [29], opracowane przez francuską firmę Cycleo (obecnie Semtech), opierające się na odmianie transmisji z widmem rozproszonym przy użyciu wielowartościowej metody CSS. Oznacza to, że w odróżnieniu od tradycyjnej, binarnej

techniki CSS, w której pojedynczy symbol *chirp* kodował tylko jeden bit, wersja tej metody zastosowana w LoRa umożliwia zakodowanie od 7 do 12 bitów w symbolu. Z założenia LoRa charakteryzuje się długim czasem bezobsługowej pracy na baterii (przez ok. 10 lat przy sprzyjających warunkach), niskimi szybkościami transmisji (do 22 kbit/s dla SF=7 i BW=500 kHz oraz 50 kbit/s w warstwie fizycznej z użyciem modulacji FSK) oraz dużymi zasięgami (2-5 km w obszarach zurbanizowanych, do 15 km na terenach podmiejskich). System może pracować, zależnie od regionu świata, w różnie zdefiniowanych pasmach ISM, w których obowiązują odmienne zasady dotyczące m.in. ERP, DC, liczby i szerokości obowiązkowych kanałów, stosowania mechanizmów PMA, długości preambuły itp. Specyfikacja [30] dla warstwy fizycznej LoRa określa wartości tych parametrów indywidualnie dla następujących pasm używanych w różnych częściach świata:

- EU 433 MHz ISM;
- CN 470–510 MHz;
- Chiny 779–787 MHz ISM;
- EU 863–870 MHz ISM w podpasmach (wg tab. 12):
  - M: 868-868,60 MHz (łącznie 600 kHz), ERP=25 mW, DC≤1%;
  - P: 869,40-869,65 MHz (łącznie 250 kHz), ERP=500 mW, DC≤10%;
- US 902–928 MHz ISM;
- Australia 915–928 MHz ISM;
- Południowa Korea 920–923 MHz ISM.

Aby w pojedynczym sygnale *chirp* typu LoRa można było zakodować do SF=12 bitów, konieczne jest zdefiniowanie unikatowej trajektorii częstotliwości w obrębie pojedynczego sygnału *chirp* dla każdej z 2^{SF} próbek, z których każda trwa $T=BW^{-1}$ [s], zaś cały sygnał $T_s=T\cdot 2^{SF}$ [s]. Efekt modulacji uzyskuje się przez przypisanie każdemu symbolowi unikatowego przesunięcia (offsetu), wynoszącego *k* czipów (próbek) względem częstotliwości początkowej f_0 kanału, kodując tym samym jego wartość (tj. wartość danego stanu modulacyjnego). Tym samym, dla utworzenia dyskretnej przestrzeni dla *m*-stanowej modulacji, cały kanał BW zostaje podzielony na 2^m interwałów, każdy o szerokości $\Delta f=BW\cdot 2^{-m}$, zlokalizowanych na dyskretnych częstotliwościach: $\{f_0+0\cdot\Delta f, f_0+1\cdot\Delta f, f_0+2\cdot\Delta f, \ldots, f_0+(2^m-1)\cdot\Delta f\}$. Każda z nich stanowi częstotliwość początkową, reprezentującą *k*-tą ze zbioru 2^m *m*-bitowych sekwencji binarnych. Innymi słowy, dana *k*-ta sekwencja moduluje świergot przez inicjalizację (w czasie T_0) jego narostu, począwszy od częstotliwości $f_0+k\cdot\Delta f$ aż do osiągnięcia (trajektorią liniową albo wykładniczą) częstotliwości granicznej sygnału (f_F), a następnie skokowo redukując częstotliwość chwilową do f_0 i kontynuując jej narost przez pozostałą część trwania sygnału T_s.



■ **Rys. 3. Schemat stanowiska pomiarowego wpływu ekstremalnej wielodrogowości na pracę systemu IoT w komorze rewerberacyjnej LKE Politechniki Wrocławskiej**

Zmodyfikowany względem oryginalnej postaci sygnał świergotowy $x_k^{(SF)}$ o częstotliwości chwilowej, będącej funkcją kwadratową czasu, przedstawia wzór (5) [20]. We wzorze tym *k* oznacza liczbę przesuniętych czipów (offset), zaś oryginalny sygnał *chirp* jest cyklicznie przesunięty o *k* czipów/próbek. Po stronie odbiorczej odebrany sygnał x_k przemnażany jest przez nieprzesunięty sprzężony sygnał *chirp* o $x_{k=0}$. Demodulacja preambuły i słowa synchronizacyjnego SyncWord, zapewniająca uzyskanie referencji czasowej niezbędnej do demodulacji dalszej części sygnału, możliwa jest dzięki cesze ortogonalności sygnałów bazowych LoRa o tym samym współczynniku rozproszenia *SF*. Stanowi ona, iż własności korelacji wzajemnej pomiędzy 2^{SF} możliwymi sygnałami bazowymi LoRa o różnych offsetach (*m*, *k*) są zadane relacją (6). Wynikowy sygnał $x_{dem}^{(SF)}$ (zadany wzorem (7)) poddawany jest następnie szybkiej transformacie Fouriera FFT. Rezultatem tego jest widmo z maksymalnym prążkiem w próbce o indeksie odpowiadającym wartości poszukiwanego offsetu *k*, co kończy proces demodulacji danego świergotu (jak prześledzono dokładnie w [26], rozdz. 5.1.5.1). Dla kompletności opisu należy zauważyć, że ortogonalność świergotów LoRa jest zachowana także dla sygnałów o różnej wartości *SF*, niezależnie od ich offsetu (także dla identycznych), o czym traktuje zależność (8). Cecha ta stanowi jednocześnie podstawę podejścia planistycznego systemu LoRaWAN, polegającego na tworzeniu do sześciu logicznych sieci (różniących się wartościami *SF*), obsługiwanych przez pojedynczą stację bazową, albo na agregowaniu szybkości transmisji do danego użytkownika, dzięki przypisaniu mu kilku różnych *SF*.

$$x_k^{(SF)}(nT) = \sqrt{\frac{1}{2^{SF}}}\, exp\left[j2\pi\cdot\frac{\left((n+k)mod2^{SF}\right)^2}{2^{SF+1}}\right] \tag{5}$$

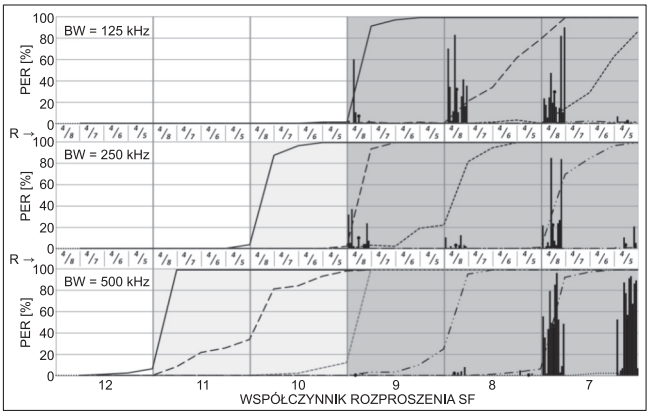
$$\sum_{n=0}^{2^{SF}-1} x_m^{(SF)}(nT)\cdot x_k^{*(SF')}(nT) = \begin{cases}1 & dla\ m=k\\ 0 & dla\ m\neq k\end{cases} \tag{6}$$

$$x_{dem}^{(SF)}(nT) = x_k^{(SF)}(nT)\cdot x_0^{*(SF)}(nT) \tag{7}$$

$$\sum_{n=0}^{2^{SF}-1} x_k^{(SF)}(nT)\cdot x_k^{*(SF')}(nT) \approx 0,\ \ \forall\, SF\neq SF' \tag{8}$$

SigFox

Od urządzeń systemu SigFox, jak od każdego systemu UNB (np. mniej znanych: Ingenu, Weightless-N), wymaga się ekstremalnej energooszczędności. Wydatnie podkreśla to zastosowany tu protokół wielodostępu RFTDMA, w którym kanał transmisji jest wybierany losowo w dziedzinie czasu i częstotliwości, bez inicjowania jakichkolwiek mechanizmów rywalizacyjnych. Najbliższym odpowiednikiem RFTDMA jest protokół ALOHA bez wstępnego nasłuchiwania zajętości kanału. W odróżnieniu jednak od klasycznego ALOHA, częstotliwości nośne w systemach UNB są dobierane z zakresu ciągłego (tu o łącznej szerokości 192 kHz, jak opisano dalej) zamiast z predefiniowanego dyskretnego zbioru kanałów. Cecha ta umożliwia jednocześnie zastosowanie w terminalach nadajników słabszej jakości, których niepewność generowanej częstotliwości jest większa od szerokości samego kanału BW, wynoszącego w łączu „w górę", **UL** (*uplink*), zaledwie 100 Hz. Stanowi to kolejną zaletę takich systemów w postaci obniżonych kosztów urządzeń nadawczo-odbiorczych, choć negatywnie wpływającą na wartości współczynnika szumowego **NF** (*Noise Factor*) czy marginesu implementacyjnego **IM** (*Implementation Margin*), co z kolei odbywa się ze szkodą dla budżetu energetycznego łącza. Każdy pakiet danych powtarzany jest w łączu UL trzykrotnie, każdorazowo



■ **Rys. 4. Zestawienie obszarów odporności i podatności na zakłócenia i wielodrogowość dla warstwy fizycznej LoRa**

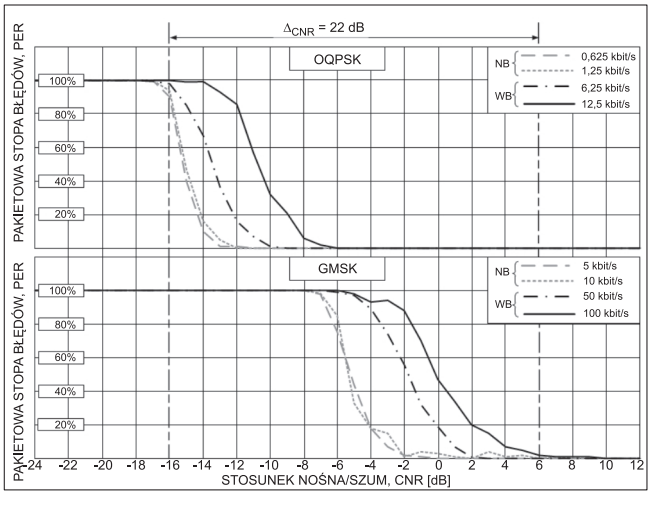
na innym, losowo wybranym kanale. Tym samym transmisja każdej pojedynczej wiadomości trwa 6,24 s, dając w efekcie zysk przetwarzania G_p wynikający z tych repetycji równy ok. 4,8 dB. Do najważniejszych cech systemu SigFox należą [31]:

- praca we fragmencie podpasma M (według tabeli 1), obejmującym łączną szerokość 192 kHz w zakresie 868,034 MHz–868,226 MHz;
- niesymetryczna transmisja dwukierunkowa UL/DL (*downlink*, czyli łącze „w dół”);
- szerokość kanału *BW* w transmisji UL: 100 Hz;
- szerokość kanału *BW* w transmisji DL: 600 Hz;
- topologia sieciowa: gwiazda;
- maksymalna dopuszczalna ERP: 14 dBm (25 mW);
- stosowane modulacje: DBPSK (w łączu UL), GFSK (w łączu DL);
- szybkość transmisji R_b w łączu UL: 100 bit/s;
- szybkość transmisji R_b w łączu DL: 600 bit/s;
- czułość P_{min}, przy *NF*=5 dB w DL lub 3 dB w UL, SNR=7 dB oraz zysku przetwarzania G_p=4,8 dB, wynosząca ok. -144 dBm w UL i 134 dBm w DL;
- dynamika tłumienia MCL=158 dB w UL i 161 dB w DL;
- całkowita długość ramki transmisyjnej: 26 B (w tym 12 B pakietu danych użytkownika, *payload*);
- czas transmisji ramki UL: 2,08 s;
- liczba transmisji pojedynczego pakietu: 3;
- systemowe ograniczenie maksymalnej liczby wysłanych wiadomości: 140 pakietów/dobę; wynika ono z konieczności spełnienia warunku *DC*=1%, co oznacza że wziąwszy po uwagę ponad 6-sekundowy czas transmitowania pojedynczej wiadomości (trzykrotnie powtórzonej), w ciągu godziny można transmitować łącznie przez 36 sekund, to jest przez czas wystarczający do nadania statystycznie niecałych sześciu wiadomości; w skali dobowej przekłada się to na maksymalnie ok. 140 wiadomości (pakietów);
- zasięg roboczy: do 50 km w terenach otwartych, do 5 km w terenach zabudowanych.

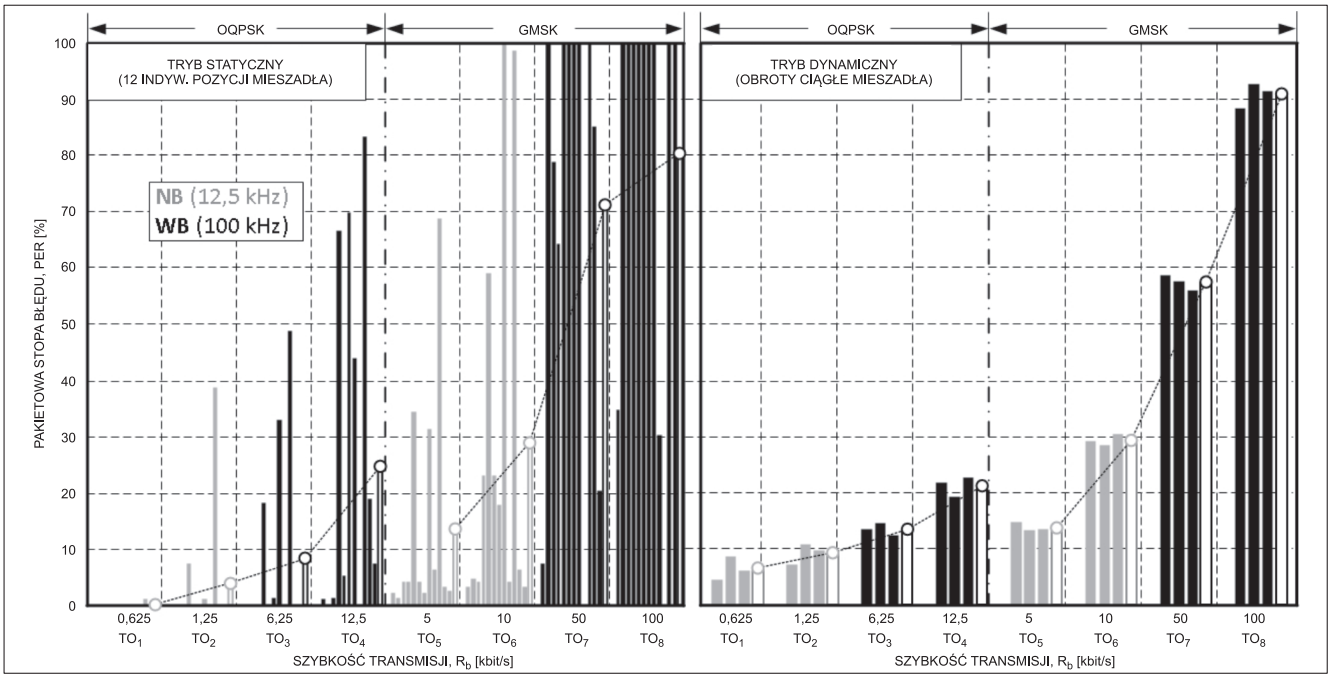
ZAKŁÓCENIA ELEKTROMAGNETYCZNE W SIECIACH INTERNETU RZECZY

Problem możliwych zakłóceń elektromagnetycznych w sieciach IoT, dosyłających sygnał w segmencie dostępowym pomiędzy warstwą czujników/mierników a stacjami bazowymi LAP, został już zauważony i opisany w ETSI TR 102 691 [32]. Dokumenty traktują temat ogólnie, przedstawiając potencjalne cele i konsekwencje zakłóceń/zagłuszeń w działaniu systemu M2M oraz podstawowe mechanizmy zapobiegania im. Zidentyfikowano dwie główne grupy zakłóceń: intencjonalne i nieintencjonalne. Celem ataków intencjonalnych jest zazwyczaj chęć uzyskania dostępu do przesyłanych danych telemetrycznych bądź ich modyfikacja. Zakłócenia nieintencjonalne z kolei, utożsamiane z negatywnymi oddziaływaniami elektromagnetycznymi, mogą prowadzić do utraty łączności pomiędzy urządzeniami IoT a stacjami bazowymi. Odpowiednio zaplanowane rozmieszczenie stacji bazowych na danym obszarze powinno w dużym stopniu zredukować prawdopodobieństwo takiego zdarzenia. Znacznie trudniejszymi do przewidzenia i kontrolowania są zakłócenia pochodzące ze źródeł innych niż nadajniki badanego systemu, współdzielących to samo pasmo częstotliwości (co wystąpiło w 2021 r. we Wrocławiu, wskutek awarii zabezpieczenia przeciwprzepięciowego w jednym z nadajników radia cyfrowego DAB+, generującego harmoniczną w paśmie 434 MHz). Niezależnie od źródła, zagłuszenia (*jamming*) stanowią obecnie jedną z najprostszych i najefektywniejszych metod degradacji funkcjonalności sieci IoT, a nawet ich całkowitego sparaliżowania.

Praca systemów IoT w nielicencjonowanych pasmach ISM z jednej strony wiąże się z brakiem opłat licencyjnych, z drugiej strony sprawia, że współużytkowanie wspólnych pasm powoduje nieuniknione podnoszenie się szumu tła, wraz z uruchamianiem kolejnych nadajników, ale też bezprzewodowych mikrofonów, zdalnie sterowanych zabawek itp. Problemem podnoszenia się tego szumu zajęto się m.in. w [33], wykazując, że w porównaniu do wcześniej wykonanych (tj. w 2004 r.) pomiarów szerokopasmowych w paśmie 200 MHz–3000 MHz dla warunków miejskich, opisanych w [34], pomiary przeprowadzone w 2016 r. wskazują na pojawienie się w ciągu dekady wielu nowych źródeł promieniowania oraz znacznego zagęszczenia istniejących, o nieprzewidywalnej charakterystyce czasowej i niedających się zakwalifikować jako biały addytywny szum gaussowski (AWGN). Jako źródła oczekiwanych transmisji zidentyfikowano systemy komórkowe (GSM, UMTS, LTE) oraz DVB-T. Przypadkowy i trudny do opisania za pomocą modeli statystycznych charakter mają zdarzenia skutkujące emisją promieniowania, takie jak: przejazd samochodu emitującego fale EM z niedostatecznie ekranowanej elektroniki pokładowej, sygnały transmitowane przez piloty zdalnego otwierania drzwi, bram i systemów automatyki domowej itp. We wnętrzach budynków



■ **Rys. 5. Wyniki pomiarów PER dla systemu Weightless uzyskane w komorze bezechowej**

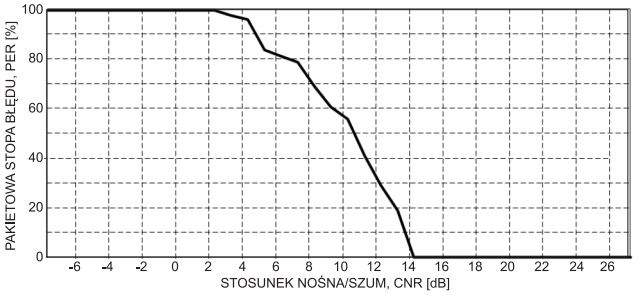


■ **Rys. 6. Wyniki pomiarów PER dla systemu Weightless uzyskane w komorze rewerberacyjnej PWr LKE**

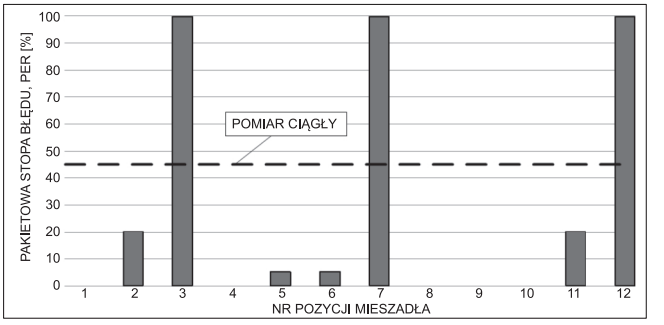
źródłami zakłóceń jest natomiast sprzęt komputerowy, w tym szczególnie serwerownie, bezprzewodowe sieci lokalne WLAN czy peryferia komunikujące się ze sobą za pomocą Bluetooth. Podniesiony szum tła we wnętrzach budynków może okazać się szczególnie problematyczny dla systemów IoT, które z definicji dedykowane są głównie do akwizycji danych z tych właśnie lokalizacji od modemów umieszczonych bezpośrednio przy/na urządzeniu pomiarowym.

Analogiczne badania zamieszczone w [35] i [36] dotyczyły jedynie pasma ISM 868,0-868,6 MHz (pasmo M w tabeli 1), mierzonego w mieście Aalborg w Danii. Zaobserwowano tam, że 22% spośród zebranych próbek należy zakwalifikować do poziomu wysokiego (tj. powyżej 105 dBm), wśród nich także liczne próbki o mocy 65 dBm. Za najbardziej odpowiedni model dla takiego profilu gęstości prawdopodobieństwa przyjęto uogólniony rozkład ekstremalnych wartości **GEV** (*Generalized Extreme Value Distribution*), stosowany zazwyczaj do opisu zjawisk skrajnych, najczęściej używany w meteorologii i hydrologii. Symulacje wpływu tych zakłóceń na pracę systemów LoRaWAN i SigFox wykazały, że powodują one spadek skutecznego zasięgu, odpowiednio do poziomu 78% oraz poniżej 50%, względem zasięgu osiągalnego przy braku zakłóceń. Oznacza to, że nawet pomimo trzykrotnego powtarzania każdej wiadomości w systemie SigFox na innym kanale częstotliwościowym, system SigFox okazuje się bardziej podatny na wpływ interferencji niż system Lora, w którym za dobrą odporność na zakłócenia odpowiada dość wysoki poziom zysku przetwarzania w zakresie 11–23 dB.

Pomiary wpływu zakłóceń na pracę urządzeń LoRaWAN, przeprowadzone dla wpływu sygnału harmonicznego i udokumentowane w [37], umożliwiają stwierdzenie, że sygnał ten nie stanowi zauważalnego źródła zakłóceń, dopóki nie przekracza mocy sygnału LoRa o 5 dB dla *SF*=7 oraz 19,5 dB dla *SF*=12, przy sprawności kodowania *R*=0,67. Z prac późniejszych można wymienić [38], gdzie wykazano, iż dzięki predykcji modelu generowania danych w jednym urządzeniu sieci LoRaWAN, jedno urządzenie jest w stanie skutecznie zablokować inne, poznawszy jego cykl pracy, wielkość pakietów i harmonogram transmisji (czyli na drodze analizy statystycznej ruchu wychodzącego). Skutecznym odzyskiwaniem celowo zakłóconego sygnału, tym razem od urządzeń systemu ZigBee (opartego na specyfikacji IEEE 802.15.4), zajęto się w [39], stosując trójfazową metodę regeneracji sygnału, przez: 1. bank filtrów, 2. fakt korzystania z rozpraszania widma przez urządzenia IEEE 802.15.4 oraz 3. kodowanie korekcyjne **FEC** (*Forward Error Correction*). Statystyk utraty pakietów, zbieranych z węzłów przekaźnikowych (*relay*), służących do wykrywania zagłuszeń, użyto także w [40], na potrzeby klasyfikowania węzłów w sieciach wieloskokowych na: niezakłócające, podejrzane i zakłócające, prowadząc w rezultacie do zmiany kanału roboczego w sieci. W [41] z kolei udowodniono, iż sygnał zakłócający, większy o min. 13 dB od mocy kanału **NPBCH** (*NB-IoT Physical Broadcast Channel*), może uniemożliwić prawidłową detekcję danych zawartych w bloku **MIB** (*Master Information Block*), zawierającym parametry stacji bazowej i odczytanie niewłaściwej wartości. Podobne wnioski sformułowano w [42], gdzie zademonstrowano (na podstawie danych o zmierzonych czułościach granicznych zaprezentowanych szczegółowo w [26]) realną możliwość zagłuszenia stacji bazowej systemów Weightless oraz SigFox sygnałem ciągłym generowanym z użyciem mocy wypromieniowanej zgodnej z obowiązującym limitem dla pasma ISM 868 MHz. W [43], wykorzystując teorię gier, opracowano metodę rozpraszania sygnału IoT pomiędzy podnośnymi **OFDM** (*Orthogonal Frequency Division Multiplexing*), opartą na algorytmie ewolucyjnym. Autorzy [44] zaproponowali modułową strukturę łańcucha bloków *blockchain* dla sieci WSN, w której w razie drastycznej redukcji wartości **CNIR** (*Carrier-Noise And Interference*) w jednym ze zbioru predefiniowanych dla sieci kanałów komunikacyjnych jej koordynator rozsyła na innym kanale z tego zbioru agenta **IDS** (*Intrusion Detection System*) do wszystkich węzłów. Informuje on o zmianie kanału roboczego,



■ **Rys. 7. Wyniki pomiaru PER dla systemu SigFox w warunkach kontrolowanych zakłóceń**



■ **Rys. 8. Wyniki pomiarów PER dla systemu SigFox w komorze rewerberacyjnej**

jednocześnie rozpoczynając emisję ciągłą sygnału na skorumpowanym kanale, celem odwrócenia uwagi węzła zagłuszającego. Opis zdecentralizowanej metody (tzw. IoTligent) adaptacyjnej zmiany kanału w urządzeniach LoRa, z wykorzystaniem algorytmów uczących, podano w [45], wykazując iż w ciągu dziesiątek sekund uczenia się urządzenia są w stanie poprawnie zidentyfikować kanały zakłócane i używać alternatywnych, z puli 8 kanałów dostępnych w podpaśmie X z tabeli 1. Jeszcze inna propozycja proaktywnego tym razem zwalczania źródeł zagłuszania została przedstawiona w [46]. Wykorzystano w niej transmisję z widmem rozproszonym, przydzielając unikatową, losowo wygenerowaną sekwencję rozpraszającą (o kompromisowo dobranej długości) dla każdej pary węzłów, do przeciwdziałania zarówno zakłóceniom pochodzącym od szerokopasmowych źródeł szumu **BBN** (*Broadband Noise*), jak i od skompromitowanych węzłów należących do atakowanej sieci. Podobny pomysł podano w [47], postulując zagłuszanie urządzenia zagłuszającego przez częste emisje flar radiowych, emitowanych przez węzły sieci sensorowej w postaci pustych pakietów, tym samym zmuszając to urządzenie do ciągłej pracy reaktywnej aż do wyczerpania jego zasobów energetycznych. W [48] użyto algorytmu równowagi Nasha do detekcji węzła zakłócającego, obserwując CNIR w sieci oraz zmiany w prawdopodobieństwie pomyślnego dostarczenia pakietów pomiędzy węzłami. Podobne podejście prezentuje [49], gdzie legalne węzły (tzw. liderów), jak i sam zakłócacz (reaktywny, tzw. naśladowca), zamodelowano jako adwersarzy w hierarchicznej grze, których interesy (choć przeciwne) zmierzają do osiągnięcia równowagi: liderzy dążą do kompromisu między CNIR a mocą nadawczą (którą jednocześnie zdradzają swoje transmisje), naśladowca zaś optymalizuje próg mocy detekcji transmisji i dopasowuje moc zakłóceń, w celu jej optymalnego wykorzystania i uniknięcia zbyt szybkiego wyczerpania własnej baterii. Pomysł ciągłej referencyjnej emisji kontrolowanej został także wykorzystany w [50], gdzie zaproponowano, aby jeden z węzłów o dobrze poznanej charakterystyce sygnału radiowego (w tym CNIR oraz odpowiedzi impulsowej, zmierzonych przed uruchomieniem sieci), służył za źródło sygnału odniesienia, potrzebne zarówno do detekcji zagłuszeń, jak i korekty zagłuszanych transmisji. Metodę opartą na proaktywnym zestawianiu optymalnej ścieżki trasowania pakietów, przy jednoczesnym doborze najlepszego kanału radiowego, zademonstrowano w [51]. Jest ona szczególnie przydatna w wieloskokowych sieciach kognitywnych IoT, z natury przystosowanych do zmian kanału roboczego.

Zupełnie przeciwną koncepcją względem ochrony przed zagłuszeniem jest celowa emisja sygnałów zagłuszających, zmierzająca do zapobieżenia podsłuchowi czy zachowania poufności wymiany informacji. W [52] zajęto się koncepcją kooperacyjnego zakłócania urządzenia zagłuszającego, w której źródłami emisji owych intencjonalnych zakłóceń są dwa sąsiadujące z węzłem źródłowym węzły sieciowe, z których jeden jest zamierzonym odbiorcą wiadomości. Jak wykazano, metoda sprawdza się szczególnie dobrze w sieciach wieloskokowych, zaś główną korzyścią (i zamierzeniem autorów) jest zachowanie poufności komunikacji przez drastyczne obniżenie poziomu CNIR w urządzeniu podsłuchującym. Podobnym zagadnieniom poświęcono też np. prace [53] i [54]. Odchodząc nieco od kontekstu WSN czy IoT, warto dodać dla kompletności przeglądu, iż niepożądane emisje elektromagnetyczne, generowane podczas normalnej pracy procesorów w urządzeniach wbudowanych, spędzają także sen z powiek producentom takich układów. Już we wczesnych pracach bowiem [55], [56] wykazano, iż taki ulot elektromagnetyczny stanowi źródło cennej informacji dla strony atakującej, umożliwiając skuteczną ekstrakcję klucza kodu kryptograficznego używanego w algorytmach szyfrujących urządzeń, takich jak smartfony, komputery, bankomaty czy inne „rzeczy" IoT, powszechnie stosujących algorytmy kryptograficzne. Oprócz samego nasłuchu, strona atakująca może zastosować wstrzykiwanie zaburzeń do układu, przez zakłócanie zegara czy sygnałów napięciowych. Z racji dwojakich zagrożeń wynikających z przechwytywania emisji EM, ryzyku temu nadano akronim **EM-SCA/FI** (*Electromagnetic Side-Channel Analysis And Fault Injection*). Najpopularniejszymi rozwiązaniami w zakresie ochrony przed nim są [57]-[62]: 1. ciągłe wykrywanie prób modyfikacji/zaburzania pracy zegara czy napięcia w układzie (wspierane obecnie algorytmami uczenia maszynowego, np. w [62]), 2. wbudowanie do chipu sondy pola magnetycznego dla wykrycia prób zbliżeniowego monitorowania emisji, 3. emitowanie szumu EM dla zmylenia odczytów zbliżeniowych.

POMIARY WYDAJNOŚCI SYSTEMÓW IOT W TRUDNYCH WARUNKACH PROPAGACYJNYCH

Przyrostowa tendencja wdrażania różnych systemów LTN, współdzielących te same pasma częstotliwości, czyni niezbędnym opracowanie ujednoliconych procedur badania ich wydajności. Problem jest tym bardziej ważki, że dostęp do zasobów częstotliwościowych odbywa się tu (co wynika z definicji pasm nielicencjonowanych) w sposób niepodlegający koordynacji przez jakąkolwiek instytucję nadzorującą. Ciężar realizacji bezkolizyjnego wielodostępu, przy użyciu technik typu LBT czy AFA, został tym samym przeniesiony na systemy. Mimo tego, w przypadku dużego zagęszczenia urządzeń nadawczych,

rzędu dziesiątek/setek tysięcy na miasto, a takie jest założenie sformułowane dla masowego Internetu rzeczy (np. w [9]), do jednoczesnych zakłóceń niewątpliwie będzie dochodzić. Z punktu widzenia urządzeń klienckich IoT, zakłócenia te będą miały charakter systematycznie podnoszącego się tła szumowego, prowadzącego do stopniowego zmniejszania się CNIR. Kolejne podrozdziały będą poświęcone prezentacji zarysów autorskiej metodyki pomiaru wydajności transmisyjnej, rekomendowanej dla systemów wąskopasmowych IoT z grupy LTN. Następnie zaś zostaną przedstawione wyniki badań pomiarowych przeprowadzonych w Laboratorium Kompatybilności Elektromagnetycznej Politechniki Wrocławskiej (LKE PWr) dla systemów omówionych poprzednio. Pomiary wydajności/odporności proponuje się wykonywać dla dwóch wariantów, przyjmując w obu przypadkach pakietową stopę błędów **PER** (*Packet Error Rate*) jako miarę jakości odbioru, a tym samym odporności:

- w obecności szumów i zakłóceń elektromagnetycznych o różnym poziomie mocy i stałej charakterystyce czasowo-widmowej;
- w warunkach propagacji wielodrogowej (tutaj: ekstremalnej, oferowanej w nieobciążonej komorze rewerberacyjnej).

Metodyka pomiaru wydajności (odporności) na szumy i zakłócenia

Ze względu na możliwy wpływ podnoszącego się stopniowo tła elektromagnetycznego w miarę coraz głębszej penetracji rynku telemetrycznego przez radiowe systemy IoT, należy w pomieszczeniach ekranowanych zapewnić maksymalną możliwą:

- izolację elektromagnetyczną od źródeł zewnętrznych;
- eliminację efektu propagacji wielodrogowej, będącej rezultatem licznych odbić fali radiowej od ścian pomieszczenia czy znajdujących się w nim obiektów.

Oba warunki są możliwe do spełnienia w komorze bezechowej, której pierwotnym i bazowym przeznaczeniem jest badanie charakterystyki promieniowania różnych obiektów emitujących fale EM (w tym głównie anten). Komora bezechowa, stanowiąca infrastruktury LKE PWr, gwarantuje skuteczność ekranowania na poziomie 85 dB. Proponowane do badań odpornościowych i wydajnościowych stanowisko pomiarowe przedstawiono schematycznie na planie tejże komory na rys. 2 (gdzie d_{pd} oznacza odległość pola dalekiego).

Metodyka pomiaru wydajności (odporności) w warunkach ekstremalnie wielodrogowych

Do pomiarów wpływu wielodrogowości na działanie systemów IoT rekomenduje się stosowanie komory rewerberacyjnej (rys. 3). Wykorzystano wyniki wcześniej przeprowadzonych pomiarów rozmycia opóźnienia τ_{RMS} dla różnego stopnia obciążenia komory, wyrażonego liczbą absorberów obniżających jej odbiciowość. W ich rezultacie uzyskano profil $\tau_{RMS,}$ będący jej odpowiedzią na stopniowe dokładanie paneli absorpcyjnych [63], począwszy od stanu komory nieobciążonej (z N=0 panelami), a kończąc na stanie komory przeciążonej (z N=12 panelami), dla której dalszej obciążanie nie powodowało znaczących zmian τ_{RMS}. Warunki skrajnie wielodrogowe odpowiadają τ_{RMS}≈1,55 μs, czyli wartości leżącej w zakresie typowym dla makrokomórek miejskich, w których τ_{RMS}=1 – 3 μs [24].

WYNIKI POMIARÓW WYDAJNOŚCI WYBRANYCH SYSTEMÓW IOT W TRUDNYCH WARUNKACH PROPAGACYJNYCH

W niniejszym rozdziale zaprezentowano wyniki pomiarów wybranych systemów IoT (z grupy LTN), uzyskane na podstawie pomiarów przeprowadzonych według procedur opisanych w poprzednich rozdziałach (rys. 4 – 8). Za miarę jakości transmisji przyjęto tu pakietową stopę błędów, każdorazowo prezentując profile PER uzyskane dla różnych wartości CNIR (w przypadku badań podatności na zakłócenia) oraz różnych nastaw (dotyczących mieszadła) w komorze rewerberacyjnej. Dokładniejsze ich omówienie można znaleźć w publikacjach: [26], [42], [63]. Przytoczenie ich w tym miejscu ma na celu zademonstrowanie zachowania systemów w obszarach granicznych dla ich działania oraz zróżnicowania tej odpowiedzi, w zależności od konkretnych nastaw, tj. wyboru **MCS** (*Modulation-Coding Scheme*), wartości SF (w przypadku metody LoRa), szerokości kanału BW i innych parametrów, charakterystycznych dla konkretnych systemów. Daje to planiście sieci, realizowanej przy użyciu danego systemu, dość szeroki zakres możliwości, zależnie od oczekiwanego poziomu odporności systemu.

* * *

Intencją autora było przybliżenie tematyki związanej z problemem zagłuszania, aspektu raczej nieczęsto poruszanego, lecz – jak się okazuje – stanowiącego istotne i realne zagrożenie dla efektywnego działania dalekosiężnych systemów Internetu rzeczy. Szczególnie narażoną grupą systemów są te należące do rodziny LTN, ze względu na współdzielenie relatywnie wąskich pasm ISM ulokowanych poniżej 1 GHz. Tym samym istotne stają się dwa czynniki, mogące być źródłem zagłuszania sygnałów pożądanych: 1) emisje pochodzące ze źródeł celowo zakłócających (jako forma cyberataku) oraz 2) podnoszący się poziom tła elektromagnetycznego (zjawisko udokumentowane), w miarę wzrostu liczby sieci IoT działających na wspólnym terenie. Jeszcze innym efektem, tym razem propagacyjnym, mogącym ujemnie oddziaływać na systemy IoT, jest propagacja wielodrogowa, szczególnie istotna w przypadku instalacji urządzeń końcowych w halach czy terenach odsłoniętych, z licznymi powierzchniami odbijającymi. W celu zbadania poziomu odporności tych systemów na wspomniane zjawiska, zaproponowano modyfikację zaplecza laboratoryjnego, służącego do wykonywania pomiarów z zakresu kompatybilności elektromagnetycznej. Zademonstrowano także wyniki pomiarów, mogące służyć jako wyniki odniesienia dla projektowania sieci tych systemów z użyciem nastaw redukujących ryzyko zagłuszenia czy zniekształcenia (wskutek wielodrogowości) transmitancji kanału radiowego.

LITERATURA

[1] Ashton K., „That 'Internet of Things' Thing. In the real world, things matter more than ideas", *RFiD Journal*, na podstawie prezentacji dla Procter & Gamble (P&G) w 1999, June 22, 2009.

[2] *Peter Day's World of Business*, BBC World Service (BBC), plik audio dostępny online pod adresem http://downloads.bbc.co.uk/podcasts/radio/worldbiz/worldbiz_20150319-0730a.mp3 (ostatni dostęp 25.06.2022).

[3] Huvio E., J.Grönvall., K.Främling, *Tracking and tracing parcels using a distributed computing approach*, SOLEM, Olav (ed.) Proceedings of the 14th Annual Conference for Nordic Researchers in Logistics (NOFOMA'2002), Trondheim, Norway, 12–14 June 2002, pp. 29-43.

[4] Magrassi P. *Why a Universal RFID Infrastructure Would Be a Good Thing*, Gartner research report G00106518, 2nd May 2002.

[5] Magrassi P., T.Berg, *A World of Smart Objects*, Gartner research report R-17-2243, 12th August 2002.

[6] *Internet of Things — An action plan for Europe*, COM(2009) 278 final, Commission of the European Communities, 18th June 2009.

[7] Wood A., "The internet of things is revolutionizing our lives, but standards are a must", *The Guardian*, 31st March 2015.

[8] Evans D., *The Internet of Things: How the Next Evolution of the Internet Is Changing Everything*, Cisco White Paper, April 2011.

[9] 3GPP TR 22.861 V14.1.0 (2016-09), *3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers for Massive Internet of Things*; Stage 1 (Release 14).

[10] ITU-R M.2083 (09/2015), *IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*, November 2014.

[11] 3GPP TR 36.888, *Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE*, v.12.0.0, June 2013.

[12] ETSI EN 300 220-2 V3.1.1 (2016-11), *Short Range Devices (SRD) operating in the frequency range 25 MHz to 1 000 MHz; Part 2: Harmonized Standard covering the essential requirements of article 3.2 of Directive 2014/53/EU for non specific radio equipment.*

[13] ETSI TR 102 313 V1.1.1 (2004-07), *Electromagnetic compatibility and Radio spectrum Matters (ERM); Frequency-agile Generic Short Range Devices using Listen-Before-Transmit (LBT); Technical Report.*

[14] ETSI EN 300 220-1 V2.1.1 (2005-04), *Electromagnetic compatibility and Radio spectrum Matters (ERM); Short Range Devices (SRD); Radio equipment to be used in the 25 MHz to 1 000 MHz frequency range with power levels ranging up to 500 mW; Part 1: Technical characteristics and test methods.*

[15] ERC Recommendation 70-03, *Relating to the use of Short Range Devices (SRD*), Tromsø, 19 May 2017.

[16] *The European table of frequency allocations and applications in the frequency range 8.3 kHz to 3000 GHz*, ERC REPORT 25, zatwierdzona w marcu 2019.

[17] Dyrektywa Parlamentu Europejskiego i Rady 2014/53/UE z dnia 16 kwietnia 2014 r. w sprawie harmonizacji ustawodawstw państw członkowskich dotyczących udostępniania na rynku urządzeń radiowych i uchylająca dyrektywę 1999/5/WE.

[18] Flandrin P., *Time-frequency/time-scale analysis*, ISBN 0-12-259670-9", Academic Press, San Diego, 1999.

[19] Springer A. et al., *Spread spectrum communications using chirp signals*, IEEE Proc. Eurocomm., pp. 166-170, 2000.

[20] Goursaud Z., J.-M.Gorce., *Dedicated networks for IoT: PHY/MAC state of the art and challenges*, EAI endorsed Transactions on Internet of Things, HAL Id: hal-01231221, 2015.

[21] IEEE 802.15.4-2011: *IEEE Standard for Local and metropolitan area networks – Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs).*

[22] Minh-Tien Do, Z.Goursaud, J.-M.Gorce., *Interference modelling and analysis of random FDMA scheme in Ultra Narrowband Networks*, Intern. Conf. on Application of Information and Comm. Technologies (AICT), July 2014, Paris, France.

[23] Win M.Z., P.C.Pinto., L.A.Shepp, *A mathematical theory of networks interference and its applications*, Proceedings of the IEEE, vol. 97, no. 2, pp. 205-230, Feb. 2009.

[24] Saunders S.R., A. Aragón-Zavala, *Antennas and propagation for wireless communications systems. 2nd edition*, John Wiley & Sons Ltd., New York, 2007.

[25] ITU, ITU-R P1407-1,"Multipath propagation and parameterization of its characteristics"

[26] Staniec K., *Radio Interfaces in the Internet of Things Systems: Performance studies*, Springer; 1st ed. 2020.

[27] *Weightless(-P). System specification*, version 1.03, 7th November 2017.

[28] *LoRa Modulation Basics*, Semtech, Application Note AN1200.22.

[29] Sornin N., M.Luis, T.Eirich, T.Kramp, O.Hersent, *LoRaWAN Specification*, Version:V1.0.2, July 2016.

[30] LoRa Alliance Technical committee, *LoRaWAN Regional Parameters*, Version: V1.0, July 2016.

[31] *SigFox. Technical overview*, SigFox, May 2017. Dokument elektroniczny dostępny pod adresem https://www.disk91.com/wp-content/uploads/2017/05/4967675830228422064.pdf (ostatni dostęp dn. 06.12.2018).

[32] *Machine-to-Machine communications (M2M); Smart Metering Use Cases*, ETSI TR 102 691 V1.1.1 (2010-05).

[33] Palaios A., V.Miteva, J. Riihijärvi, P.Mähönen, *When the whispers become noise: a contemporary look at radio noise levels*, 2016 IEEE Wireless Communications and Networking Conference (WCNC), 3-6 April 2016. DOI: 10.1109/WCNC.2016.7564891.

[34] Ming-Hui Chang, K.-H. Lin, *A comparative investigation on urban radio noise at several specific measured areas and its applications for communications*, IEEE Transactions on Broadcasting, vol. 50, issue: 3, Sept. 2004, pp. 233-243.

[35] Vejlgaard B., M. Lauridsen, H.C.Nguyen, I.Kovács, P.E.Mogensen & M. Sørensen (2017). *Interference Impact on Coverage and Capacity for Low Power Wide Area IoT Networks*. In IEEE Wireless Communications and Networking Conference 2017.

[36] Lauridsen M., B. Vejlgaard, I. Z. Kovacs, H. Nguyen and P. Mogensen, *Interference Measurements in the European 868 MHz ISM Band with Focus on LoRa and SigFox*, 2017 IEEE Wireless Communications and Networking Conference (WCNC), 2017, pp. 1-6, doi: 10.1109/WCNC.2017.7925650.

[37] *SX 1272/3/6/7/8: LoRa Modem*, Semtech, Designer's Guide, AN1200.13, rev 1, July 2013.

[38] Max Ingham, Jims Marchang, Deepayan Bhowmik, *IoT security vulnerabilities and predictive signal jamming attack analysis in LoRaWAN*, The Institution of Engineering and Technology, Special Section: Security on Mobile and IoT Devices, 2020, doi: 10.1049/iet-ifs.2019.0447.

[39] Zicheng Chi, Yan Li, Xin Liu, Wei Wang, Yao Yao, Ting Zhu, and Yanchao Zhang. 2020. *Countering cross-technology jamming attack*. In Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '20). Association for Computing Machinery, New York, NY, USA, 99–110. DOI:https://doi.org/10.1145/3395351.3399367.

[40] Jagdeep Singh, Isaac Woungang, Sanjay Kumar Dhurandher, Khuram Khalid, *A jamming attack detection technique for opportunistic networks*, Internet of Things, Volume 17, 2022, 100464, ISSN 2542-6605, https://doi.org/10.1016/j.iot.2021.100464.

[41] Morillo, G. and U.Roedig (2021) *Jamming of NB-IoT synchronisation signals*, 26th European Symposium on Research in Computer Security (ESORICS) 2021, Virtual Event, 04-08 October.

[42] Staniec K., M. Kowal, *On Vulnerability of Selected IoT Systems to Radio Jamming – A Proposal of Deployment Practices*. Sensors 2020, 20, 6152. https://doi.org/10.3390/s20216152.

[43] Namvar N., W. Saad, N. Bahadori and B. Kelley, *Jamming in the Internet of Things: A Game-Theoretic Perspective*, 2016 IEEE Global Communications Conference (GLOBECOM), 2016, pp. 1-6, doi: 10.1109/GLOCOM.2016.7841922.

[44] Mbarek B., Ge, M. & Pitner, T. An adaptive anti-jamming system in HyperLedger-based wireless sensor networks. Wireless Netw 28, 691–703 (2022). https://doi.org/10.1007/s11276-022-02886-1.

[45] Moy C., "Artificial Intelligence for Jamming Mitigation in IoT Networks: LoRaWAN Field Measurements Using IoTligent," 2021 XXXIVth General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS), 2021, pp. 1-4, doi: 10.23919/URSIGASS51995.2021.9560289.

[46] Renzo E. E.R. Navas, F. Cuppens, N. Boulahia Cuppens, L. Toutain, G.Z. Papadopoulos, Physical resilience to insider attacks in IoT networks: Independent cryptographically secure sequences for DSSS anti-jamming, *Computer Networks,* Volume 187, 2021, 107751, ISSN 1389-1286, doi: https://doi.org/10.1016/j.comnet.2020.107751.

[47] Sciancalepore, S.; Oligeri, G.; Di Pietro, R. Strength of Crowd (SOC)–Defeating a Reactive Jammer in IoT with Decoy Messages. Sensors 2018, 18, 3492. https://doi.org/10.3390/s18103492.

[48] Manikandan C., and Alamelumangai, V., Performance Analysis of Nash Algorithm to Detect the Jamming Attack in IoT Network (November 27, 2020). Compliance Engineering Journal Volume 11, Issue 3, 2020 ISSN NO: 0898-3577; Performance Analysis of Nash Algorithm to Detect the Jamming Attack in IoT Network Page no:7-13, Available at SSRN: https://ssrn.com/abstract=3972569.

[49] Tang, X. P. Ren and Z. Han, "Jamming Mitigation via Hierarchical Security Game for IoT Communications," in IEEE Access, vol. 6, pp. 5766-5779, 2018, doi: 10.1109/ACCESS.2018.2793280.

[50] Zhang P. and S. Sun, "One Node to Guard All: Jamming-Resistant and Low-Latency Communication for IoT," 2018 IEEE Global Communications Conference (GLOBECOM), 2018, pp. 1-6, doi: 10.1109/GLOCOM.2018.8647322.

[51] Haythem Bany Salameh, Rawan Derbas, Moayad Aloqaily, and Azzedine Boukerche. 2019. Secure Routing in Multi-hop IoT-based Cognitive Radio Networks under Jamming Attacks. In Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWIM '19). Association for Computing Machinery, New York, NY, USA, 323–327. DOI:https://doi.org/10.1145/3345768.3355944.

[52] Abdullah Z., G. Chen, M. A. M. Abdullah and J. A. Chambers, "Enhanced Secrecy Performance of Multihop IoT Networks With Cooperative Hybrid-Duplex Jamming," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 161-172, 2021, doi: 10.1109/TIFS.2020.3005336.

[53] Fan X. and Y. Huo, "Security Analysis of Cooperative Jamming in Internet of Things with Multiple Eavesdroppers," 2019 IEEE Global Communications Conference (GLOBECOM), 2019, pp. 1-6, doi: 10.1109/GLOBECOM38437.2019.9014015.

[54] Lin H., Hong W., Bin W., Fei P., Run-Fa L, H. Song, Jie T., Xiumin W., "Cooperative Jamming for Physical Layer Security Enhancement in Internet of Things," in IEEE Internet of Things Journal, vol. 5, no. 1, pp. 219-228, Feb. 2018, doi: 10.1109/JIOT.2017.2778185.

[55] Agrawal D. , B. Archambeault, J.R. Rao, P. Rohatgi, "The EM Side--Channel(s)", Cryptographic Hardware and Embedded Systems – CHES 2002, 4th International Workshop, Redwood Shores, CA, USA, August 13-15, 2002. DOI: 10.1007/3-540-36400-5_4.

[56] Chih-Hsu Yen and Bing-Fei Wu, "Simple error detection methods for hardware implementation of Advanced Encryption Standard," in *IEEE Transactions on Computers*, vol. 55, no. 6, pp. 720-731, June 2006, doi: 10.1109/TC.2006.90.

[57] Fox-IT. TEMPEST attacks against AES. Technical report.

[58] Danial J. et. al., "EM-X-DL: Efficient cross-device deep learning side-channel attack with noisy em signatures", *ACM Journal on Emerging Technologies in Computing Systems* (JETC), 18(1), pp. 1-17, Sep. 2021.

[59] Das D. et al. EM and Power SCA-Resilient AES-256 in 65nm CMOS Through >350x Current-Domain Signature Attenuation. In IEEE ISSCC, 2020.

[60] Ghosh A. , D. Das, J. Danial, V. De, S. Ghosh and S. Sen, "Syn-STELLAR: An EM/Power SCA-Resilient AES-256 With Synthesis-Friendly Signature Attenuation," in *IEEE Journal of Solid-State Circuits*, vol. 57, no. 1, pp. 167-181, Jan. 2022, doi: 10.1109/JSSC.2021.3113335.

[61] Ghosh A. , D. Das, J. Danial, V. De, S. Ghosh and S. Sen, "36.2 An EM/Power SCA-Resilient AES-256 with Synthesizable Signature Attenuation Using Digital-Friendly Current Source and RO-Bleed-Based Integrated Local Feedback and Global Switched-Mode Control," 2021 IEEE International Solid- State Circuits Conference (ISSCC), 2021, pp. 499-501, doi: 10.1109/ISSCC42613.2021.9365978.

[62] Ghosh A. , D. Das, S. Ghosh and S. Sen, "EM SCA & FI Self-Awareness and Resilience with Single On-chip Loop & ML Classifiers," 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2022, pp. 592-595, doi: 10.23919/DATE54114.2022.9774588.

[63] Staniec K., P. M. Kowal, LoRa performance under variable interference and heavy-multipath conditions. *Wireless Communications and Mobile Computing.* 2018, vol. 2018, art. 6931083, s. 1-9, doi: 10.1155/2018/6931083.

Marek AMANOWICZ*, Mariusz KAMOLA*

Świadomość bezpieczeństwa cybernetycznego operatora usług kluczowych

Cyber security awareness of essential services operator

Przedstawiono koncepcję modelowania cyberprzestrzeni operatora usług kluczowych, umożliwiającą tworzenie świadomości bezpieczeństwa cybernetycznego dla świadczonych usług, głównych procesów biznesowych oraz systemów, a także jej wykorzystanie w procesie zarządzania cyberbezpieczeństwem.
Słowa kluczowe: cyberbezpieczeństwo, modelowanie cyberprzestrzeni, operator usług kluczowych, świadomość sytuacyjna

The paper presents the concept of modelling essential service provider cyberspace to create cybersecurity awareness for services, core business processes and systems and its use in the cybersecurity management process.
Key words: cybersecurity, cyberspace modelling, essential services operator, situational awareness

DOI: 10.15199/59.2022.4.6

Ochrona przed rosnącą liczbą coraz bardziej wyrafinowanych i złożonych ataków cybernetycznych wymaga osiągnięcia świadomości bezpieczeństwa w cyberprzestrzeni[1)], zapewniającej pozyskiwanie w czasie rzeczywistym bieżącej informacji o zidentyfikowanych zagrożeniach i potencjalnych skutkach, w tym o wpływie podatności systemów teleinformatycznych (**IT** – *Information Technology*) i sterowania przemysłowego (**OT** – *Operational Technology*) na powiązane z nimi procesy i usługi. Zmasowane ataki cybernetyczne na ważne obiekty państwowe i przemysłowe poprzedzające inwazję Rosji na Ukrainę dobitnie podkreślają wagę, jaką należy przywiązywać do wdrażania mechanizmów zapewniających szybkie wykrywanie prób takich działań i uruchamiania mechanizmów ich tłumienia, zanim spowodują realne zagrożenia.

Sytuację komplikuje silne powiązanie usług, procesów i współdzielonych elementów infrastruktury IT/OT [5,8]. Te wzajemne powiązania są różnej i złożonej natury [4]. Ich precyzyjne rozeznanie jest niezbędne do identyfikacji zagrożeń w cyberprzestrzeni oraz oceny wpływu na bezpieczeństwo zarówno w skali lokalnej (np. pojedynczego procesu czy usługi), jak i globalnej (np. firmy, sektora, państwa). Wykrycie takich zależności prowadzi do dokładniejszej oceny poziomu krytyczności pojedynczego elementu infrastruktury, a nawet całego sektora.

Przeprowadzanie pełnej i wiarygodnej oceny zagrożeń świadczonych usług wymaga precyzyjnej identyfikacji oraz opisu wszystkich istotnych elementów cyberprzestrzeni wraz z ich wzajemnymi oddziaływaniami. Wskutek propagacji zagrożenia, nawet niewielka zmiana stanu pojedynczego elementu może powodować istotne zagrożenie bezpieczeństwa usług czy kluczowych procesów operatora. Identyfikacja powiązań elementów cyberprzestrzeni operatora umożliwia wykrycie całych łańcuchów zależności wewnętrznych, ale także ustalenie najbardziej krytycznych elementów i wdrożenie skutecznych mechanizmów poprawy bezpieczeństwa.

W artykule przedstawiono koncepcję modelowania cyberprzestrzeni operatora usług kluczowych, umożliwiającą tworzenie świadomości bezpieczeństwa cybernetycznego dla świadczonych usług, głównych procesów biznesowych oraz systemów, a także jej wykorzystanie w procesie zarządzania cyberbezpieczeństwem. W następnym rozdziale zaprezentowano model cyberprzestrzeni oraz sposób wyznaczania statycznych parametrów obiektów. Modelowanie dynamicznych stanów elementów cyberprzestrzeni, wynikających ze zmaterializowanych zagrożeń bezpieczeństwa, jest opisane w kolejnym rozdziale. Całość podsumowano wnioskami dotyczącymi dalszego rozwoju zaproponowanych rozwiązań i ich wykorzystania do zarządzania świadomością bezpieczeństwa cybernetycznego przez operatorów usług kluczowych.

MODELOWANIE STATYCZNYCH ATRYBUTÓW OBIEKTÓW

Model cyberprzestrzeni

W celu uproszczenia formalnego opisu, poszczególne elementy cyberprzestrzeni operatora będą w dalszej części określane jednym ogólniejszym pojęciem *obiekt biznesowy* (w skrócie

[1)] Świadomość bezpieczeństwa w cyberprzestrzeni to postrzeganie zagrożeń dla elementów cyberprzestrzeni oraz zdarzeń w odniesieniu do czasu i przestrzeni, rozumienie ich znaczenia oraz przewidywanie przyszłego stanu tych elementów.

* NASK-PIB, Warszawa, e-mail: marek.amanowicz@nask.pl, mariusz.kamola@nask.pl

obiekt lub zamiennie *węzeł*), który reprezentuje, w zależności od rozpatrywanego kontekstu, usługę, proces, system lub jego element.

Sieć powiązań jest definiowana jako graf skierowany z atrybutami opisanymi na jego łukach i wierzchołkach. Funkcje opisane na łukach grafu charakteryzują wpływ zagrożeń, wynikających z naruszenia atrybutów bezpieczeństwa powiązanych obiektów, tj. poufności (*c – confidentiality*), integralności (*i – integrity*) oraz dostępności (*a – availability*) oraz czas propagacji zagrożenia. Funkcje opisane na wierzchołkach grafu, charakteryzujące miedzy innymi rodzaj obiektu, jego lokalizację w cyberprzestrzeni oraz rolę/znaczenie biznesowe, są określane przez operatora. Drugą grupę tworzą funkcje wyliczane w procedurach określania krytyczności obiektów oraz obrazowania bieżącego stanu ich bezpieczeństwa.

Należy zauważyć, że niektóre obiekty mogą tworzyć grupy obiektów redundantnych, tj. alternatywnych względem pozostałych obiektów grupy[2]. Ponadto, dla niektórych obiektów, zwanych dalej obiektami nadrzędnymi, możliwe jest ustalenie ich istotności biznesowej określającej ich znaczenie z punktu widzenia prowadzonego przez operatora biznesu, wyrażone osiąganym zyskiem, ponoszoną stratą lub inną miarą definiowaną przez operatora.

Przyjęto następujące oznaczenia węzłów sieci powiązań:
V_n – węzeł sieci powiązań reprezentujący *n-ty* obiekt biznesowy, gdzie: n=1,2, ..., N ; N – liczba obiektów,
V_b^g – węzeł sieci powiązań reprezentujący *b-ty* nadrzędny obiekt biznesowy, gdzie: b=1,2,...,B, B – liczba obiektów nadrzędnych,
I_b^g – istotność *b-tego* nadrzędnego obiektu biznesowego, tj. znaczenie obiektu V_b^g z punktu widzenia prowadzonego przez podmiot biznesu.

Krytyczności obiektów

Proces budowy sieci powiązań obejmuje dwa zadania, tj. identyfikację i opis obiektów biznesowych oraz ich hierarchizację. Opiera się on na metodzie **AHP** (*Analytic Hierarchy Process*) zaproponowanej w [6].

Proces identyfikacji ma na celu stworzenie bazy danych o obiektach i ich atrybutach, w tym między innymi o ich lokalizacji, identyfikacji komórek organizacyjnych odpowiedzialnych za ich utrzymanie oraz uwarunkowań ich funkcjonowania w przypadku wystąpienia zdarzeń wpływających na bezpieczeństwo, w szczególności podatności na zagrożenia bezpieczeństwa. Proces hierarchizacji ma na celu ustalenie wpływu poszczególnych obiektów wraz z ich charakterystyką oraz istotnością poszczególnych obiektów ze względu na konsekwencje wynikające z naruszenia ich bezpieczeństwa.

Każdy obiekt w cyberprzestrzeni operatora usługi jest opisany co najmniej dwiema wielkościami statycznymi, tj.:

- istotnością odziaływania na powiązane z nim obiekty, nazywaną dalej względną krytycznością,
- istotnością wpływu na osiągnięcie przez operatora zakładanych celów biznesowych, nazywaną dalej krytycznością w kontekście organizacji.

Wpływ naruszenia dowolnego atrybutu bezpieczeństwa obiektu V_n na atrybuty bezpieczeństwa powiązanego z nim obiektu V_j jest określony wartościami macierzy X_{nj} odziaływania tych obiektów, przedstawionej w tabeli 1.

[2] w danej chwili operacyjnie aktywny jest tylko jeden obiekt

Przyjęto, że jeżeli oddziaływanie obiektów ma być określone ze względu na wszystkie atrybuty bezpieczeństwa, będzie ono wyznaczane według zależności (1).

$$s_{nj} = \sum_{(k,l)} x_{nj}^{kl}; \quad k \in (a,i,c);\ l \in (a,i,c) \tag{1}$$

Tabela 1. Macierz x_{nj} powiązań atrybutów bezpieczeństwa

		V_j		
		a	*i*	*c*
V_n	*a*	x_{nj}^{aa}	x_{nj}^{ai}	x_{nj}^{ac}
	i	x_{nj}^{ia}	x_{nj}^{ii}	x_{nj}^{ic}
	c	x_{nj}^{ca}	x_{nj}^{ci}	x_{nj}^{cc}

Do oceny wpływu bezpieczeństwa obiektu V_n na osiąganie zakładanych celów biznesowych wykorzystuje się wskaźnik υ_n wyznaczany z zależności (2).

$$\upsilon_n = \sum_{z=1}^{Z} s_{nz} \cdot \upsilon_z \tag{2}$$

tj. jako ważoną sumę krytyczności s_{nz} obiektu V_n względem wszystkich Z obiektów, na które on oddziałuje, gdzie wagami są wyznaczone wcześniej krytyczności υ_z w kontekście całej organizacji tych Z obiektów.

Znajomość krytyczności obiektów umożliwia operatorowi opracowanie działań zapobiegawczych i ich ujęcie między innymi w planach utrzymania ciągłości działania, w celu zmniejszenia skutków oddziaływania zagrożeń na realizacje jego kluczowych celów strategicznych. Zaproponowana procedura określania wartości tych wskaźników umożliwia prowadzenie analiz typu *what-if* ukierunkowanych na poszukiwanie takich rozwiązań technicznych i proceduralnych, które będą prowadzić do zmniejszenia ryzyka związanego z propagacją zagrożeń między obiektami sieci powiązań.

MODELOWANIE STANÓW DYNAMICZNYCH

Dynamika propagacji zagrożeń, wieloaspektowość oddziaływań międzywęzłowych oraz możliwość występowania cykli w sieci stanowią czynniki determinujące architekturę opracowanego modelu symulacyjnego sieci obiektów biznesowych. Każdy z powyższych czynników ma charakter komplikujący zadanie modelowania propagacji zagrożenia w sieci. Uwzględnienie cykli w sieci, zasadniczo niepożądanych w kontekście analizy, jest niezbędne w modelowaniu układów automatyki i stanowi ich immanentną cechę. Prowadzi ono do wdrożenia w modelu nieliniowych zależności międzywęzłowych, aby zapewnić stabilność numeryczną w symulacji. Z kolei rozpiętość skali czasowej możliwych oddziaływań, w połączeniu z nieliniowościami, wyklucza zastosowanie klasycznych, synchronicznych modeli dynamiki, np. Friedkina [2], kierując konstrukcję symulatora w stronę architektury zdarzeniowej. Wreszcie wieloaspektowość oddziaływań nakazuje rozważenie możliwych sposobów ich agregacji w przypadku (najczęstszym) wpływu na działanie i bezpieczeństwo jednego obiektu przez wiele innych.

Oprócz powyższych uwarunkowań, model powinien spełniać pewne postulaty natury pragmatycznej, wynikające ze specyfiki pozyskiwania danych początkowych i aktualizujących

i ostatecznie wpływające na jego użyteczność dla operatora. Uniwersalność interpretacji pojęcia obiekt biznesowy nie stanowi tutaj przeszkody, a jest raczej ułatwieniem w ekstrakcji cech wspólnych takich obiektów w kontekście modelowania zagrożeń, umożliwiając ostatecznie określenie klasy obiektów-węzłów, które będą jednorodne i o ograniczonej, uniwersalnej funkcjonalności. Natomiast wszelkie bardziej zaawansowane zjawiska i zależności w systemie są modelowane nie przez polimorfizm klas modeli dla poszczególnych węzłów, ale przez rozbudowanie struktury ich połączeń o węzły sztuczne, dodatkowe. Takie podejście zapewnia stopniową ewolucję złożoności modelu przez złożoność struktury, a bez ingerencji w logikę kodu źródłowego. Jest ono pewną analogią do niekwestionowanego paradygmatu modelowania dowolnie złożonych odwzorowań przez głębokie sztuczne sieci neuronowe o stosunkowo prostej funkcji aktywacji pojedynczego węzła-neuronu.

Pojęciem kluczowym prezentowanego modelu jest obserwowany stan $\boldsymbol{q}_j$ obiektu biznesowego V_j. Jest to wielkość wektorowa o składowych odpowiadających stopniowi zaspokojenia wymagań funkcjonowania obiektu w aspekcie dostępności, integralności i poufności, $\boldsymbol{q}_j=[q_j^a, q_j^i, q_j^c]$. Każda z tych składowych może przyjmować wartości od 0 (wymaganie zupełnie niezaspokojone) do 1 (wymaganie w pełni zaspokojone).

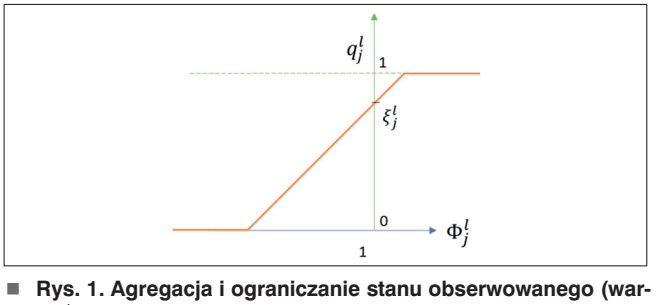
Zasadniczo wymaga się, aby każdy z obiektów sieci spełniał wymagania w wyżej wymienionych aspektach. Obserwowany stan obiektu jest sumą tzw. stanu autonomicznego ξ_j oraz wpływu obserwowanych stanów węzłów nań wpływających, ograniczoną do zakresu <0;1>. Na rys. 1 przedstawiono ideę tego sumowania; przy czym łączny wpływ obcych stanów reprezentuje Φ_j^l. Taka logika dotyczy każdego z aspektów z osobna, tj. $l \in (a, i, c)$.

Łączny wpływ obcych stanów może być sumą minimum lub maksimum wartości, z jakimi te stany oddziałują na aspekt k funkcjonowania węzła . W przypadku sumy otrzymuje się wzór:

$$\Phi_j^l = \sum_{n\in\mathcal{E}(j),k\in(a,i,c)} x_{nj}^{kl}\,(q_n^k - \xi_n^k) \qquad (3)$$

gdzie n indeksuje elementy zbioru $\varepsilon(j)$ węzłów oddziałujących na V_j. Wartości oddziaływania obiektów mają zatem w modelu dynamicznym interpretację wzmocnienia, z jakim uszczerbek w funkcjonowaniu obiektu V_n przenosi się na V_j.

Wpływy uszczerbków w funkcjonowaniu obiektów, stanowiące składniki sumy w (3), propagują się w sieci z właściwą sobie dynamiką odzwierciedlającą naturę oddziaływania obiektów. Mogą mieć charakter np. proporcjonalny, inercyjny, prostego opóźnienia – podobnie jak typowe człony układów automatyki – albo kompozycji tychże członów. W szczególnych zastosowaniach [7], gdy znany jest charakter i parametry tej propagacji, uzasadnione jest zastosowanie konkretnych i precyzyjnych formuł. Natomiast w przypadku ogólnym najistotniejszą rolę gra opóźnienie liczone od zaobserwowania uszczerbku w funkcjonowaniu obiektu V_n do zaobserwowania pierwszych symptomów wpływu tego uszczerbku na uzależniony od niego stan obiektu V_j – i w ten właśnie sposób odbywa się symulacja w dziedzinie czasu modelu dynamicznego. Oznacza to w szczególności, że wyznaczane stany obiektów są stanami ustalonymi, lecz charakter przejścia pomiędzy wyznaczonymi stanami nie jest uwzględniony. Uwzględniona została natomiast asymetria w dynamice oddziaływania w przypadku uszczerbku względem poprawy wartości stanu obiektu. Wynika ona ze spostrzeżeń praktycznych, np. oddziaływanie niedostępności systemu sterowania na zdolność świadczenia wspieranej przez niego usługi realizuje się zdecydowanie szybciej, niż przywrócenie jego dostępności. Stąd w modelu występują oba czasy propagacji, przy czym dopuszczalne są czasy zerowe.



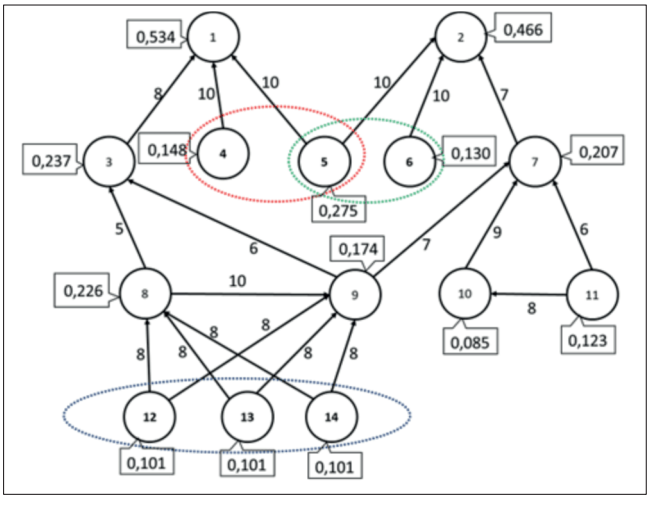
■ **Rys. 1. Agregacja i ograniczanie stanu obserwowanego (wartość ξ_j^l<1 przyjęto na potrzeby zilustrowania metody**

PRZYKŁAD

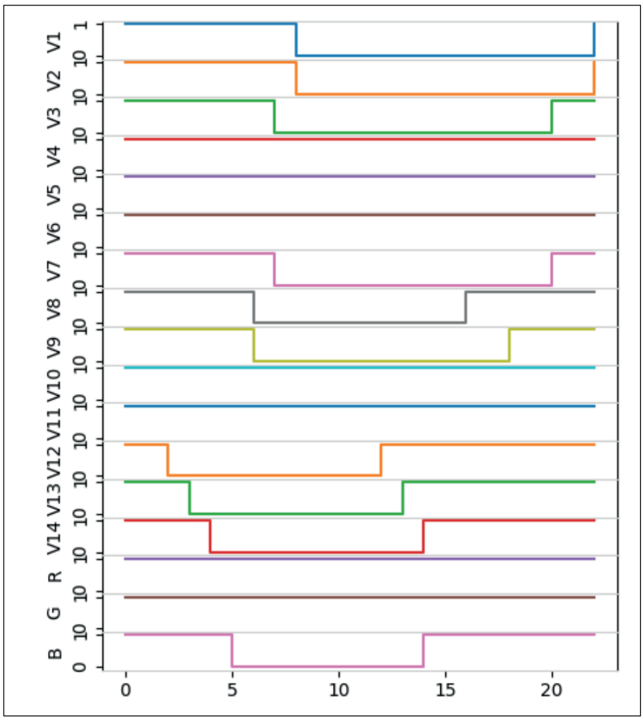
Załóżmy, że w procesie identyfikacji oraz hierarchizacji obiektów została zbudowana sieć powiązań, jak na rys. 2. Obiekty ponumerowano kolejnymi liczbami naturalnymi. Obiekty o numerach 4, 5, 6, 12, 13 i 14 tworzą trzy grupy redundantne (oznaczone na rys. 2 liniami przerywanymi). Obiekty o numerach 1 i 2 są nadrzędnymi obiektami biznesowymi (np. kluczowymi usługami świadczonymi przez operatora, stanowiącymi źródło jego dochodów), których istotność biznesową operator określił na l_1^g = 800 oraz l_2^g = 700 jednostek umownych.

Przyjmijmy także, że w procesie hierarchizacji zostały określone krytyczności powiązań obiektów s_{nj} ze względu na ich dostępność, jak oznaczono na rys. 2. Zamieszczono na nim także wyniki obliczeń krytyczności obiektów v_n w kontekście całej organizacji.

Dla ilustracji modelowania stanów dynamicznych sieci przedstawionej na rys. 2 przyjmijmy, że w chwili początkowej wszystkie obiekty znajdują się w stanie pełnej dostępności. Niech w chwilach 2, 3, 4 obiekty 12, 13 i 14 kolejno stają się niedostępne, natomiast przywrócenie ich dostępności następuje w chwilach 12, 13 i 14. Podobnie jak w przypadku określania krytyczności obiektów, model dynamiczny umożliwia uwzględnienie odziaływania obiektów redundantnych. Efekt propagacji niedostępności grupy obiektów redundantnych na pozostałe



■ **Rys. 2. Przykładowa sieć powiązań obiektów**



■ **Rys. 3. Stany węzłów w funkcji czasu w scenariuszu zaniku, a następnie wznowienia działania usług redundantnych. Propagacja ubytków z opóźnieniem równym 1, propagacja poprawy stanu z opóźnieniem równym 2**

obiekty sieci zilustrowano na rys. 3. Niedostępność ostatniego obiektu redundantnego grupy powoduje niedostępność całej grupy, co w sieci przykładowej propaguje się aż do węzłów nadrzędnych.

Wprowadzenie kolejnych powiązań między węzłami redundantnymi umożliwia modelowanie bardziej złożonych zależności logicznych, m.in. sekwencję uruchamiania kolejnych tzw. zimnych rezerw oraz ograniczenia zasobowe (*de facto*, czasowe) ich funkcjonowania. Model można rozbudowywać o kolejne zależności logiczne bez potrzeby modyfikowania podstawowego algorytmu symulacji – zgodnie z założeniami przedstawionymi poprzednio.

Zaimplementowany algorytm symulacji jest typowym algorytmem zdarzeniowym, przy czym przez zdarzenie rozumie się zmianę (zaplanowaną) stanu autonomicznego węzła oraz zmianę (wynikową) wartości oddziaływania między węzłami. Prototyp symulatora wykonano w języku Python 3, wykorzystując standardową bibliotekę obliczeń numerycznych *numpy*, natomiast do celów testów wydajności posłużono się dodatkową popularną biblioteką modelowania grafów *NetworkX*. Wydajność symulatora oszacowano empirycznie, symulując awarię pojedynczego węzła w sieci losowej (model Erdősa-Rényiego [1]) o 1000 węzłach, przy czym każdy węzeł oddziałuje średnio na dwóch sąsiadów. Na komputerze osobistym typu i5 z zegarem 1,8 GHz średni czas symulacji wyniósł ok. 40 s. Istnieje możliwość wykonania testów na sieciach bardziej realistycznych względem przedstawionej na rys. 2, tj. utworzonych na podstawie zależności modułów oprogramowania [3].

* * *

Przedstawiona metoda modelowania wpływu wzajemnych powiązań usług, procesów i infrastruktury IT/OT na świadczenie usług i realizację kluczowych procesów biznesowych stanowi element projektu realizowanego przez NASK-PIB we współpracy z operatorem usług kluczowych. Dotychczas zaproponowane rozwiązania, uzupełnione o nowe, umożliwiające między innymi dynamiczne szacowanie ryzyka i jego propagację w sieci powiązań, zarządzanie podatnościami, integrację z wewnętrznymi systemami bezpieczeństwa, a także z systemem teleinformatycznym Krajowego Systemu Cyberbezpieczeństwa, zapewnią opracowanie i wdrożenie kompleksowego systemu zarządzania świadomością sytuacyjną operatorów usług kluczowych.

LITERATURA

[1] Erdős P., A.Rényi. 1959. "On Random Graphs". *Publicationes Mathematicae*. 6:290–297.

[2] Friedkin N., E.Johnsen. 1990. "Social influence and opinions." *Journal of Mathematical Sociology*, 15(3-4) : 193-206.

[3] Kamola M., 2019. "Sensitivity of Importance Metrics for Critical Digital Services Graph to Service Operators' Self-Assessment Errors." *Security and Communication Networks*. Available online: https://doi.org/10.1155/2019/7510809.

[4] Petit F., Lewis Lawrence P. 2016. "Incorporating Logical Dependencies and Interdependencies into Infrastructure Analyses". *Goerge Mason University,* Available online: https://cip.gmu.edu/2016/02/17/incorporating-logical-dependencies-and-interdependencies-into-infrastructure-analyses.

[5] Rinaldi S. et al. 2001. "Identifying, understanding, and analysing critical infrastructure interdependencies". *IEEE Control Systems*, 21 (6) : 11 – 25.

[6] Saaty Th. L. 2008. „Relative measurement and its generalisation in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process". *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 102 : 251–318.

[7] Rucco P., et al. 2012. "Dynamic functional modelling of vulnerability and interoperability of Critical Infrastructures." *Reliability Engineering & System Safety*, 105 : 51 – 63.

[8] Zimmerman R. 2004. "Decision-making and the vulnerability of interdependent critical infrastructu-re". *IEEE International Conference on Systems, Man and Cybernetics,* 5 : 4059 – 4063.

Adam DUSTOR*

Zastosowanie modelowania statystycznego w weryfikacji mówcy

Application of statistical modelling to speaker verification

Przedstawiono zastosowanie modelowania statystycznego, opartego na mieszaninach rozkładów Gaussa w głosowym systemie biometrycznym. Przedstawiono przykładową realizację systemu weryfikacji mówcy z wykorzystaniem otwartego oprogramowania. Uzyskane wyniki umożliwiają zastosowanie takiego systemu weryfikacji tożsamości jako dodatkowego zabezpieczenia zwiększającego poziom bezpieczeństwa systemu teleinformatycznego.
Słowa kluczowe: bezpieczeństwo, biometria, modelowanie statystyczne, weryfikacja mówcy

The paper presents the application of statistical modelling based on Gaussian mixture models to a voice biometric system. The implementation of a speaker verification system based on open-source software is presented. Obtained results enable the application of such an identity verification system as an additional security layer which increases the security level of a teleinformation system.
Key words: biometrics, security, speaker verification, statistical modelling

DOI: 10.15199/59.2021.4.8

Bezpieczeństwo i związane z nim zagadnienia są niezwykle ważne w nowoczesnych systemach informatycznych. W związku z tym istnieje ogromne zapotrzebowanie na skuteczne oraz niezawodne metody identyfikacji i weryfikacji tożsamości. Dotychczasowe metody weryfikacji oparte na określonej wiedzy (hasło, numer PIN) lub posiadanym sprzęcie (karta bankowa, klucz) są wysoce zawodne. Hasło wymaga zapamiętania i jest podatne na podsłuch lub odgadnięcie. Klucz wymaga natomiast odpowiedniego przechowywania i można go zgubić, zniszczyć, skopiować lub też ukraść. Najnowocześniejszą metodą weryfikacji tożsamości jest oparcie się na cechach biometrycznych ludzkiego ciała. Wśród stosowanych obecnie można wyróżnić cechy fizyczne i behawioralne. Do najważniejszych cech fizycznych należą: odciski palców, wzór tęczówki oka, obraz twarzy, kształt dłoni. Cechy behawioralne to przede wszystkim głos i podpis odręczny. Nie istnieje cecha biometryczna, która zapewniałaby bezbłędną weryfikację tożsamości, dlatego metody biometryczne są zwykle uzupełnieniem metod tradycyjnych.

Systemy zabezpieczeń, oparte na cechach fizycznych, są zazwyczaj bardziej niezawodne i stabilne w czasie. Wzór tęczówki oka lub odcisk palca nie zmienia się w czasie, podczas gdy biometria behawioralna, taka jak głos ludzki, jest w znacznym stopniu uzależniona od stanu emocjonalnego lub fizycznego mówiącego. W związku z tym dokładność systemów biometrycznych opartych na głosie jest znacznie niższa, niż systemów opartych na tęczówce oka.

W najbliższej przyszłości systemy biometryczne będą bardzo powszechne niemal we wszystkich systemach informatycznych i bezpieczeństwa. W niektórych krajach, na przykład w Chinach, takie systemy są niemal wszędzie. Rozpoznawanie twarzy stosowane w większości kamer ulicznych służy do identyfikacji poszukiwanych przez policję przestępców.

BIOMETRIA GŁOSU – PODSTAWY

Automatyczne rozpoznawanie mówcy umożliwia, na podstawie mierzalnych cech sygnału mowy, jego zidentyfikowanie bądź zweryfikowanie. Potrzeba rozpoznawania osób na podstawie głosu pojawia się tam, gdzie nie ma możliwości wykorzystania innych metod biometrycznych (np. rozmowy telefoniczne, łączność radiowa, komunikacja z komputerem). Inne przykłady zastosowania rozpoznawania mówcy to systemy ochrony danych przed nieuprawnionym dostępem, sterowanie urządzeniami za pomocą głosu przez osoby uprawnione, zakupy telefoniczne, dostęp do komputerów, serwisów informacyjnych lub baz danych. Rozpoznawanie mówiącego znajduje również szerokie zastosowanie w kryminalistyce i sądownictwie.

W biometrii głosu wyróżnia się dwa podzagadnienia: identyfikację i weryfikację. Identyfikacja w najprostszym przypadku to proces, w którym wypowiedź nieznanej osoby jest przypisywana do jednej z M klas, gdzie M to liczba mówców mających w systemie rozpoznawania swój model głosu. Weryfikacja mówcy natomiast jest procesem potwierdzenia lub odrzucenia deklarowanej tożsamości na podstawie wypowiedzi weryfikowanego mówcy.

W systemach rozpoznawania mówców wyróżnia się systemy zależne i niezależne od tekstu. W zależnych treść językowa wypowiedzi testowej powinna być taka sama, jak treść zarejestrowana podczas treningu. W systemach niezależnych natomiast wypowiedzi uczące i testowe mogą zawierać całkowicie różną treść lingwistyczną. Konsekwencją zależności od tekstu jest konieczność wypowiadania określonych haseł w trakcie procedury rozpoznawania oraz ich częste powtarzanie w przypadku trudności z rozpoznaniem użytkownika. Systemy takie, mimo że są bardziej niezawodne i prostsze od systemów niezależnych od

* Politechnika Śląska, Gliwice, e-mail: adam.dustor@polsl.pl

tekstu, nie są odporne na odtworzenie przez oszusta wcześniej zapisanego hasła mówcy, co stanowi ich bardzo poważną wadę.

W procesie identyfikacji sekwencja wielowymiarowych wektorów parametrów, uzyskiwanych z kolejnych ramek sygnału mowy, jest porównywana z zapisanymi w systemie rozpoznawania modelami głosu poszczególnych mówców. Maksymalne podobieństwo wskazuje na tożsamość mówcy. Niestety, wraz ze wzrostem liczby modeli głosu rośnie prawdopodobieństwo błędnej identyfikacji. Jest to jedna z przyczyn znacznie większego praktycznego znaczenia weryfikacji głosu, która to jest niewrażliwa na liczbę zarejestrowanych użytkowników.

W procesie weryfikacji sekwencja parametrów porównywana jest zawsze tylko z jednym modelem głosu, modelem mówcy, którego tożsamość jest weryfikowana. Jeżeli podobieństwo jest odpowiednio duże, to tożsamość mówcy jest pozytywnie zweryfikowana, gdy natomiast zbyt małe, to mówca jest traktowany jako oszust i odrzucany przez system biometryczny. W związku z tym istnieje możliwość popełnienia błędów dwojakiego rodzaju. Błąd fałszywej akceptacji **FA** (*False Acceptance*), gdy oszust jest błędnie uznany za osobę, której tożsamość jest weryfikowana. Błąd fałszywego odrzucenia **FR** (*False Rejection*), gdy uprawniony mówca o tożsamości zgodnej z deklarowaną jest błędnie uznawany za oszusta i odrzucany przez system rozpoznawania.

Ekstrakcja parametrów

W biometrii głosu niezwykle istotnym zagadnieniem jest odpowiedni wybór parametrów wyliczanych z poszczególnych segmentów sygnału mowy. Powinny to być parametry o niskiej zmienności wewnątrzosobniczej, aby stan fizyczny i psychiczny mówiącego oraz upływ czasu miały niewielki wpływ na poziom błędów weryfikacji, a zarazem parametry te powinny cechować się dużą zmiennością międzyosobniczą. Skuteczność systemu rozpoznawania w dużej mierze zależy od tego, w jakim stopniu parametry fizyczne sygnału mowy oddają cechy osobnicze mówcy.

W rozpoznawaniu mówcy wykorzystuje się parametry uzyskiwane z przebiegu czasowego, jak rozkład interwałów czasowych, względne długości czasu wypowiedzi poszczególnych elementów fonetycznych, obwiednia czasowa amplitudy, parametry analizy przejść przez zero sygnału mowy. Alternatywnie wykorzystywane są parametry uzyskiwane z widma sygnału mowy, jak uśrednione widmo amplitudowe, widmo mocy, częstotliwość tonu krtaniowego, parametry formantowe, widmo krótkoterminowe, momenty widmowe. W procesie ekstrakcji parametrów wykorzystywana jest również analiza predykcyjna LPC i uzyskiwane z niej współczynniki predykcji, odbicia i pochodzące od nich parametry cepstralne LPCC.

Najniższe poziomy błędów uzyskiwane są dla parametrów cepstralnych LPCC oraz uwzględniających nieliniową charakterystykę ludzkiego słuchu parametrów MFCC [10].

Model statystyczny głosu

Drugim kluczowym problemem w rozpoznawaniu głosu, oprócz ekstrakcji parametrów, jest zagadnienie modelowania mówiącego. Do stworzenia modeli wykorzystuje się bardzo wiele algorytmów, z których najważniejsze to algorytmy grupowania danych (*k*-means i pochodne), modelowanie głosu oparte na mieszaninach rozkładów Gaussa **GMM** (*Gaussian Mixture Models*) [4, 11], klasyfikatory oparte na maszynie wektorów nośnych **SVM** (*Support Vector Machine*) [13], zbiory rozmyte [2], klasyfikatory kernelowe [3], analiza czynnikowa JFA [6] oraz *i*-wektory [1, 5], a także uczenie głębokie [9, 12, 14].

Złotym standardem w biometrii głosu są mieszaniny GMM, które stanowią zwykle punkt odniesienia do innych metod modelowania. W modelu GMM mówca reprezentowany jest przez kombinację liniową M rozkładów normalnych, z których każdy jest opisany wektorem wartości średnich $\boldsymbol{\mu}_i$, macierzą kowariancji $\mathbf{C}_i$ oraz pewnym współczynnikiem wagowym p_i. Prawdopodobieństwo, że wektor parametrów $\mathbf{x}$ pochodzi od mówcy opisanego modelem λ, jest równe:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}) \tag{1}$$

gdzie $b_i(\mathbf{x})$ jest D-wymiarowym rozkładem normalnym o parametrach $\boldsymbol{\mu}_i$ oraz $\mathbf{C}_i$ przy założeniu, że suma wszystkich wag p_i jest równa 1. W rezultacie mówca w systemie rozpoznawania jest reprezentowany przez model λ, będący zbiorem parametrów $\lambda = \{p_i, \boldsymbol{\mu}_i, \boldsymbol{C}_i\}$ dla $i = 1,2, \dots M$. Zakładając niezależność wektorów parametrów uzyskanych z wypowiedzi rozpoznawanej $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_N]$, prawdopodobieństwo, że wypowiedź należy do mówcy o modelu λ, wynosi:

$$p(\mathbf{X}|\lambda) = \prod_{j=1}^{N} p(\mathbf{x}_j|\lambda) \tag{2}$$

W procesie testowania systemu weryfikacji podstawą podjęcia decyzji o akceptacji mówcy jest znormalizowana wartość prawdopodobieństwa $S(\mathbf{X})$:

$$S(\mathbf{X}) = \frac{1}{N}\sum_{j=1}^{N} \log p\left(\mathbf{x}_j \middle| \lambda\right). \tag{3}$$

W procesie uczenia wykorzystywany jest algorytm **EM** (*Expectation-Maximization*). Istotną wadą modelu GMM jest duża liczba parametrów, a zatem trudności z ich właściwą estymacją podczas fazy uczenia. Trudności z wytrenowaniem modeli można częściowo zniwelować, stosując tzw. model odniesienia GMM-UBM [11]. Ze wszystkich wypowiedzi pewnej grupy mówców (różnych od mówców testowanych) tworzy się jeden duży model GMM, zwany modelem odniesienia i z niego tworzy się następnie, na podstawie procedury adaptacyjnej, modele GMM poszczególnych mówców.

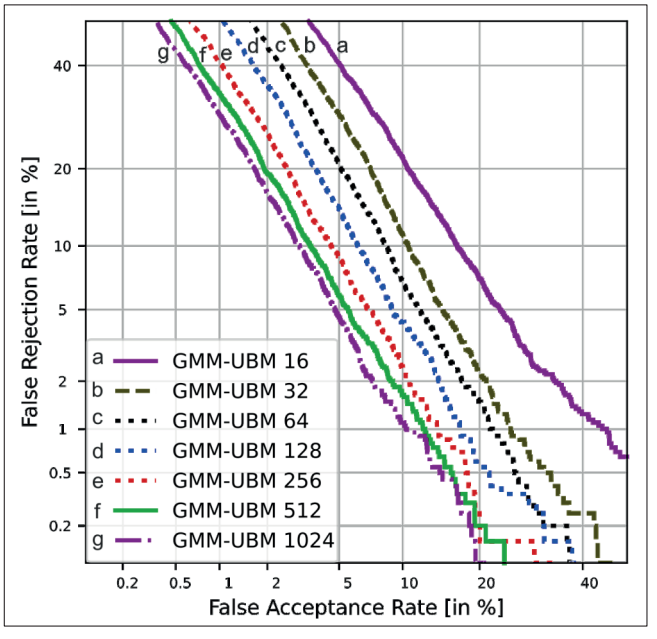
IMPLEMENTACJA

Przykładowy system weryfikacji mówców zrealizowano na komputerze typu PC z wykorzystaniem oprogramowania Anaconda oraz oprogramowania *open-source*, w którym zaimplementowano wszystkie bloki funkcjonalne systemu weryfikacji mówców. Całość kodu napisano w języku Python z wykorzystaniem podstawowych pakietów, jak NumPy, SciPy, matplotlib oraz SIDEKIT [7], zainstalowanych w osobnym środowisku utworzonym za pomocą oprogramowania Anaconda. W badaniach wykorzystano cały zasób mowy TIMIT, stanowiący zbiór 6300 wypowiedzi 630 mówców z USA (10 wypowiedzi na mówcę). Zasób ten, udostępniany za opłatą przez **LDC** (*Linguistic Data Consortium*), jest jednym z podstawowych zasobów wykorzystywanych w badaniach nad rozpoznawaniem mowy i mówcy.

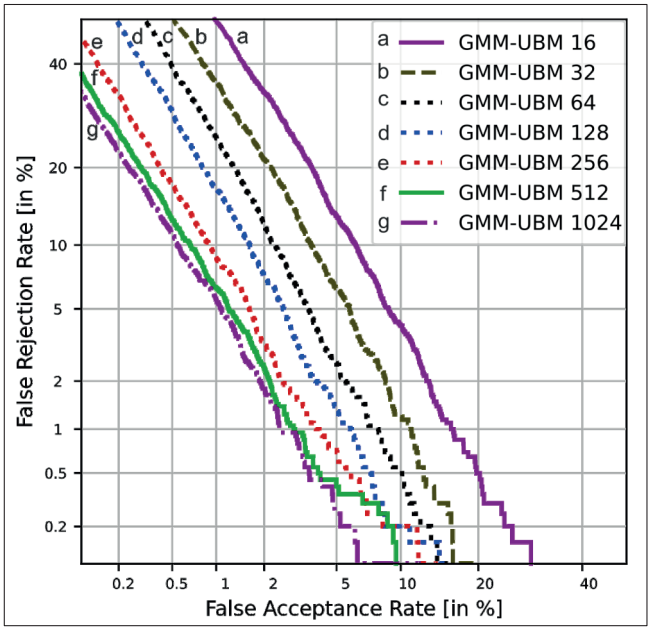
Do modelowania głosu wykorzystano zmodyfikowaną wersję mieszanin GMM, a mianowicie model GMM-UBM. Ze wszystkich wypowiedzi 130 mówców utworzono za pomocą algorytmu EM jeden model odniesienia z diagonalnymi macierzami kowariancji. Z modelu odniesienia utworzono na podstawie procedury MAP modele 500 mówców. W tym celu wykorzystano po 6 wypowiedzi uczących na model, pozostałe 4 wypowiedzi każdego z 500 mówców były wykorzystane do testowania. Łącz-

na liczba testów, w których jest możliwy błąd FR, wyniosła 2000 (4*500), a liczba testów z możliwym błędem fałszywej akceptacji FA wyniosła 998 000 (4*499*500) dla każdej kombinacji wymiarowości wektora cech i rzędu modelu. W trakcie treningu systemu modyfikowano tylko wektory wartości średnich, macierze kowariancji były takie same jak w modelu odniesienia.

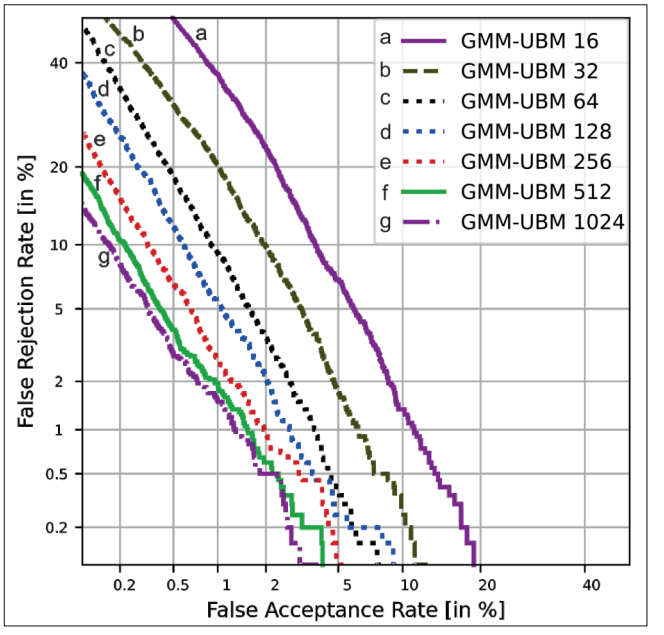
Podczas testowania systemu weryfikacji sprawdzono wpływ złożoności modelu (liczba składników mieszaniny GMM) oraz wymiarowości wektora cech obliczanych z kolejnych segmentów sygnału mowy na wartości błędów FR i FA. Wartość rzędu modelu zmieniano w zakresie 16 do 1024. Wektor cech stanowiły parametry cepstralne MFCC o liczbie współczynników odpowiednio 5, 10, 15 i 20. Do parametrów MFCC dołączono parametry ΔMFCC. Sygnał poddano segmentacji na ramki o długości 25 ms z wstępną preemfazą. Ze względu na niewielkie zaszumienie, nie zastosowano detekcji i wycięcia ciszy.



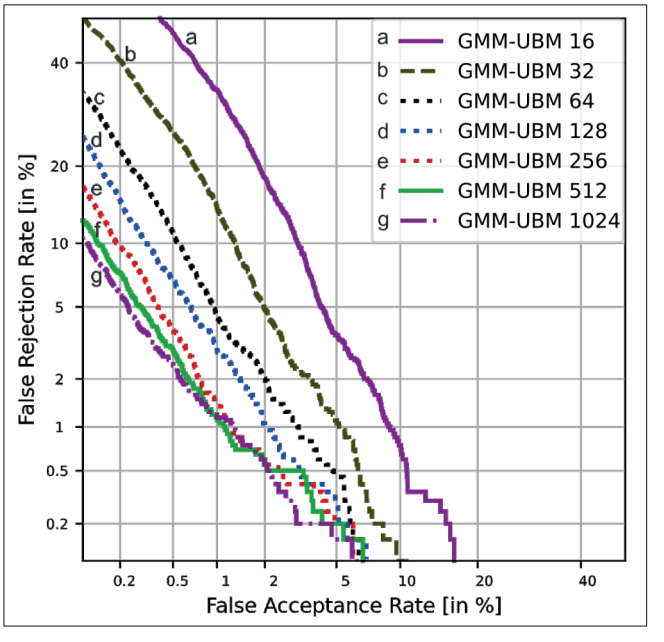
■ **Rys. 1. Wpływ złożoności modelu GMM-UBM na błędy weryfikacji (5 współczynników MFCC na ramkę)**



■ **Rys. 2. Wpływ złożoności modelu GMM-UBM na błędy weryfikacji (10 współczynników MFCC na ramkę)**



■ **Rys. 3. Wpływ złożoności modelu GMM-UBM na błędy weryfikacji (15 współczynników MFCC na ramkę)**



■ **Rys. 4. Wpływ złożoności modelu GMM-UBM na błędy weryfikacji (20 współczynników MFCC na ramkę)**

Na rys. 1, 2, 3 i 4 przedstawiono zachowanie systemu weryfikacji mówcy w zależności od liczby składników mieszanin GMM na model mówcy. Zależność pomiędzy błędami FR i FA przedstawiono opierając się na krzywych DET [6, 8].

Można zauważyć, że na poziom błędów systemu weryfikacji ma wpływ zarówno liczba składników mieszanin GMM, jak i wymiarowość wektora cech. Jest to szczególnie widoczne dla modeli o niższej złożoności. Model o 512 składnikach zapewnia praktycznie takie same osiągi, jak model o 1024 składowych. Jest to wskazówką, że nie ma sensu dalsze komplikowanie modelu, gdyż powoduje to problemy ze znacznie wydłużonym czasem

fazy treningu systemu. Najniższe poziomy błędów uzyskano dla 512 mieszanin i wymiarowości wektora 40 (20 MFCC i 20 ΔMFCC). Wartość błędu EER zmniejsza się wówczas do poziomu około 1% (rys. 4).

Warto się zastanowić, czy taki poziom błędów jest satysfakcjonujący z punktu widzenia bezpieczeństwa systemu teleinformatycznego. Dobranie takiej wartości progu decyzyjnego, przy której wartości FR i FA są sobie równe i wynoszą około 1%, oznaczałoby, że średnio co setny uprawniony weryfikowany użytkownik jest błędnie odrzucany lub też co setny oszust jest fałszywie akceptowany. Poziom błędów rzędu 1% jest stosunkowo wysoki dla samodzielnego systemu bezpieczeństwa, jednak taki system biometryczny może stanowić uzupełnienie tradycyjnego systemu weryfikacji tożsamości opartego na wiedzy (numer PIN) lub sprzęcie (karta bankowa).

W niektórych zastosowaniach komercyjnych (bankowość elektroniczna) wartość progu dobiera się w taki sposób, aby zminimalizować prawdopodobieństwo fałszywej akceptacji oszusta, nawet kosztem częstszego odrzucania uprawnionych użytkowników. Na przykład dla systemu z rys. 4 (model GMM-UBM 1024) możliwe jest dobranie takiego progu, aby system pracował przy parametrach $FR \approx 10\%$ i $FA = 0{,}1\%$ lub $FR \approx 5\%$ i $FA = 0{,}2\%$ co zapewnia zmniejszenie ryzyka akceptacji oszusta odpowiednio 10 lub 5 razy.

* * *

W artykule omówiono najważniejsze bloki funkcjonalne głosowego systemu weryfikacji tożsamości wraz z jego implementacją praktyczną w języku Python. Wykazano, że przy odpowiednim doborze wymiarowości wektora cech i liczby składowych w modelu GMM można uzyskać wartości błędu EER rzędu 1%. Dobierając odpowiednio wartość progu decyzyjnego, można zmniejszyć prawdopodobieństwo fałszywej akceptacji nieuprawnionych użytkowników do wartości 10^{-3}. Uzupełnienie tradycyjnego systemu weryfikacji tożsamości, opartego na wiedzy lub posiadaniu, o moduł biometryczny zapewni istotne zmniejszenie prawdopodobieństwa nieautoryzowanego dostępu do systemu teleinformatycznego.

LITERATURA

[1] Dehak N., P.J.Kenny, R.Dehak, P.Dumouchel, P.Ouellet, 2011. „Frontend factor analysis for speaker verification". *IEEE Transactions on Audio, Speech, and Language Processing* 19(4).

[2] Dustor A., P.Kłosowski, 2013. „Biometric Voice Identification Based on Fuzzy Kernel Classifier". *Communications In Computer and Information Science* 370: 456-465. Springer-Verlag.

[3] Dustor A., P.Kłosowski, J.Izydorczyk, 2014. „Speaker Recognition System with Good Generalization Properties". *International Conference on Multimedia Computing and Systems (ICMCS)*, 206-210.

[4] Dustor A. 2020. „Speaker verification with TIMIT corpus – some remarks on classical methods". *Proceedings of the 24th International Conference on Signal Processing Algorithms, Architectures, Arrangements and Applications (SPA)*.

[5] Jiang Y., F.H.F.Leung, 2021. „Vector-Based Feature Representations for Speech Signals: From Supervector to Latent Vector". *IEEE Transactions on Multimedia* 23: 2641-2655.

[6] Kabir M. M., M.F.Mridha, J.Shin, I.Jahan, A.Q.Ohi, 2021. „A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities". *IEEE Access* 9: 79236-79263.

[7] Larcher A., K.A.Lee, S.Meignier, 2016. „An extensible speaker identification SIDEKIT in Python". *International Conference on Audio Speech and Signal Processing ICASSP*.

[8] Martin A., G.Doddington, T.Kamm, M.Ordowski, M.Przybocki, 1997. „The DET curve in assesment of detection task performance". *Proceedings of the EUROSPEECH*: 1895–1898.

[9] Ohi Q., M.F.Mridha, M.A.Hamid, M.Monowa, 2021. „Deep Speaker Recognition: Process, Progress, and Challenges". *IEEE Access* 9: 89619-89643.

[10] Rabiner L. R., B.H.Juang , 1993. *Fundamentals of speech recognition*. Prentice Hall.

[11] Reynolds D. A., T.F.Quatieri, 2000. „Speaker Verification Using Adapted Gaussian Mixture Models". *Digital Signal Processing* 10: 19–41.

[12] Snyder D., D.Garcia-Romero, G.Sell, D.Povey, S.Khudanpur, 2018. „Xvectors: Robust dnn embeddings for speaker recognition". *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: 5329–5333.

[13] Vapnik V., 1999. *The nature of statistical learning theory*. Springer.

[14] Zhou T., Y.Zhao, J.Wu, 2021. „ResNeXt and Res2Net Structures for Speaker Verification". *IEEE Spoken Language Technology Workshop (SLT)*: 301-307.

Piotr GARBAT*, Grzegorz GALIŃSKI**, Mateusz BIENIEK*, Michał ŻAKOWSKI*, Maciej LASOCKI***

System analizy obrazu dla potrzeb systemu AR

Image Analysis System for Augmented Reality Application

Przedstawiono ideę systemu wizyjnego umożliwiającego analizę obrazów z wielu kamer dla potrzeb systemów rzeczywistości wspomaganej. Przedstawione rozwiązanie zapewnia efektywne wyznaczanie pozycji i ścieżek ruchu poruszających się osób w przestrzeni miejskiej oraz ich aktywności. W artykule przedstawiono kluczowe moduły analizy obrazu z wielu widoków, w tym: podsystem wykrywania i śledzenia pieszych (MOT) oraz podsystem wykrywania aktywności użytkowników. Omawiane podsystemy zostały przebadane w celu określenia efektywności zaproponowanych algorytmów określania ścieżki ruchu i rozpoznawania aktywności.
Słowa kluczowe: rozpoznawanie aktywności, analiza wielowidokowa, uczenie głębokie

This article presents the idea of a vision system that allows for the analysis of images from multiple cameras for the purposes of augmented reality systems. The proposed solution effectively determines the position and paths of movement of moving people in the urban space and their activities. The article presents the critical modules of the image analysis from multiple views, including the pedestrian detection and tracking (MOT) subsystem and the user activity detection subsystem. The subsystems were stud to determine the effectiveness of the proposed algorithms for determining the traffic path and recognizing activity.
Key words: action recognition, multiview analysis, deep learning

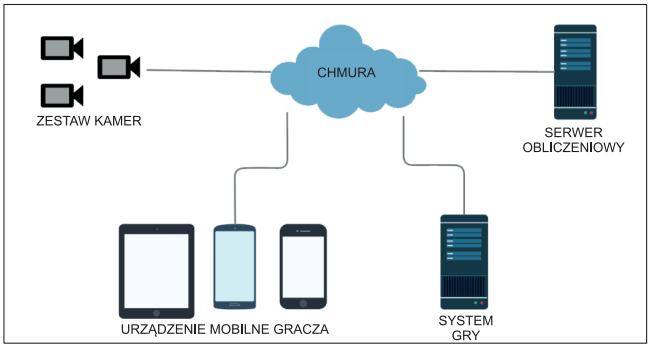
DOI: 10.15199/59.2022.4.7

W ostatnich latach obserwuje się niezwykle dynamiczny rozwój technologii wzbogaconej rzeczywistości **AR** (*Augmented Reality*), powodowany rosnącym zainteresowaniem wielu branż, w tym branży rozrywkowej. Wraz z rozwojem technologicznym, oczekiwania wobec wrażeń dostarczanych przez tego typu systemy są coraz wyższe. Na rynku z sukcesem pojawiają się coraz nowsze rozwiązania umożliwiające w coraz większym stopniu zaangażowanie zmysłów użytkownika w świat wirtualny. Jednym z przykładów tego typu innowacji są gry w AR. Łączenie obrazu rzeczywistego z wirtualną sceną gry umożliwia przeniesienie rozgrywki do realnych miejsc. Możliwość rozgrywki na urządzeniach mobilnych, takich jak tablety, smartfony i przenośne systemy gier, otwiera nowe perspektywy dla twórców tego typu gier. Jedną z głównych zalet takiego rozwiązania jest praktycznie nieograniczony obszar do odtworzenia immersyjnego środowiska, ponieważ dodawane są wyraźne cechy gry w środowisku użytkownika w czasie rzeczywistym. Przykładem tego typu gier są cieszące się popularnością gry terenowe typu: „Pokemon GO" oparte na systemach lokalizacji gracza. Niestety, obecne rozwiązania pozostawiają wiele do życzenia zarówno w zakresie precyzji pozycjonowania gracza, jak i ograniczonych możliwości bezpośredniej interakcji między graczami podczas rozgrywki w czasie rzeczywistym. Proponowane rozwiązanie ma za zadanie zapewnić możliwość precyzyjnego określenia pozycji gracza w rzeczywistej scenie oraz możliwość jego oddziaływania na rozgrywkę pozostałych graczy znajdujących się w danym miejscu.

SYSTEM ANALIZY OBRAZÓW

Ogólna koncepcja

Proponowane rozwiązanie oparte jest na systemie wizyjnym, wykorzystującym zewnętrzne kamery umieszczone w przestrzeni, w której prowadzona jest gra (rys. 1). Jest on jednym z komponentów zaprojektowanej architektury. Jego zadaniem jest zapewnienie śledzenia ruchu gracza, rozpoznawanie wybranych aktywności i przesyłanie informacji o jego położeniu w czasie rzeczywistym, tak aby rozgrywka była płynna. System gry, otrzymując na bieżąco informacje o położeniu gracza, umożliwia śledzenie postępów, jak również wprowadza nowe perspektywy interakcji w rozgrywce wieloosobowej.



■ **Rys. 1. Ogólny schemat zaprojektowanego rozwiązania**

* Instytut Mikroelektroniki i Optoelektroniki PW, Warszawa, e-mail: piotr.garbat@pw.edu.pl,
** Instytut Radioelektroniki i Technik Multimedialnych PW, Warszawa, e-mail: grzegorz.galiński@pw.edu.pl,
*** Wydział Architektury PW, Warszawa, e-mail: maciej.lasocki@pw.edu.pl

Opisywany system składa się z trzech podstawowych elementów: zestawu kamer umożliwiających obserwację z wielu kierunków oraz w wielu miejscach, modułu detekcji i śledzenia ruchu wielu obiektów jednocześnie – **MV-MOT** (*Multi-View Multiple Object Tracking*) oraz modułu rozpoznawania aktywności – **MV-VAR** (*Multi-View Video Action Recognition*).

Moduł wykrywania i śledzenia graczy

Pierwszy z modułów wykorzystuje metodę FairMOT [5]. Rozwiązanie to wykorzystuje sztuczną sieć neuronową typu koder – dekoder do przetwarzania obrazu na potrzeby detekcji ruchomych obiektów oraz ich klasyfikacji. W efekcie możliwe jest śledzenie ruchu dowolnego obiektu oraz powiązanie ze sobą jego wystąpień w różnych położeniach i momentach czasu (rys. 2). W zaprojektowanej architekturze przewidziano obsługę wielu kamer, które zostaną uruchomione wraz z rozwojem projektu. Zapewnią one obserwację dużego obszaru, stanowiącego pole gry. Celem związanym z systemem detekcji jest uzyskanie jak najwyższej dokładności śledzenia i rozpoznawania gracza, co umożliwi precyzyjne określenie jego położenia na mapie gry. Nie bez znaczenia jest tutaj fakt występowania wielu kamer, który niesie za sobą konieczność przebudowy architektury algorytmu FairMOT, aby był on zdolny do powiązania ze sobą wystąpień obiektów pomiędzy osobnymi widokami.

■ **Rys. 2. Przykładowa ramka sekwencji z wykrytymi przechodniami**

Moduł rozpoznawania aktywności graczy

Kolejnym z komponentów systemu jest moduł rozpoznawania aktywności gracza [1]. W trakcie rozgrywki gracz otrzymuje zadanie wykonania określonej akcji. Obecnie scenariusz definiuje sześć aktywności: korzystanie z telefonu, robienie zdjęcia, uniesienie jednej ręki, uniesienie dwóch rąk, siadanie, kucanie. Poprawne wykonanie i rozpoznanie akcji umożliwi systemowi przypisanie gracza do odpowiedniej postaci w systemie śledzenia postaci. W danej chwili jedna aktywność może być użyta do synchronizacji maksymalnie jednego gracza w obrębie jednej lokalizacji.

Dane wejściowe systemu stanowią strumienie wideo z kamer miejskich. Dostępnych jest kilka lokalizacji, a w każdej z nich umieszczone są dwie kamery rejestrujące obraz z różnych kątów widzenia. Dane są wysyłane do chmury, gdzie są synchronizowane i poddawane analizie w komponencie rozpoznawania aktywności. System udostępnia API umożliwiające odpytanie go o aktywności wykonywane przez wszystkie lub konkretne osoby z danej lokalizacji w ciągu ostatnich *N* sekund.

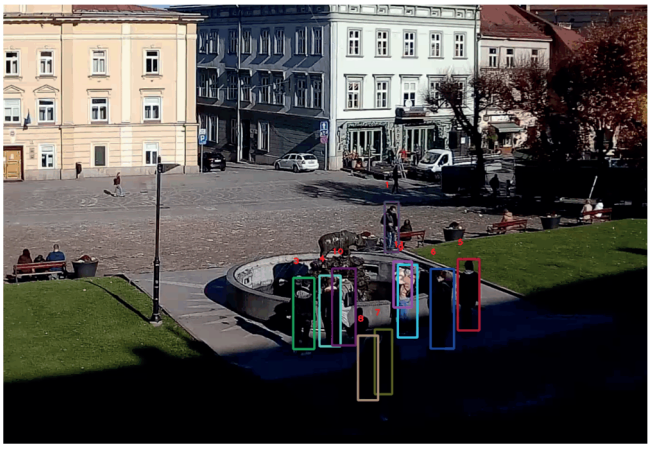
Do rozpoznawania aktywności wykorzystano algorytmy uczenia maszynowego i głębokie sieci neuronowe. Poza samymi algorytmami istotną rolę pełnią dane treningowe. Składają się one z jednogodzinnych sekwencji wideo nagranych przez cztery kamery w dwóch lokalizacjach oraz ręcznie dokonanych oznaczeń pozycji osób i wykonywanych aktywności.

Sieć neuronowa przyjmuje na wejściu ok. 2-sekundową sekwencję wideo i zwraca prawdopodobieństwa obecności akcji w tej sekwencji. Taka definicja wejścia-wyjścia określa sposób przygotowania danych. W jednej próbce treningowej powinna być widoczna dokładnie jedna akcja. Użyte zostały w tym celu oznaczenia osób w postaci ramek. Dla każdej oznaczonej akcji można wyznaczyć obszar, w którym poruszała się ramka postaci w czasie trwania akcji. Ten obszar definiuje najmniejszy prostokąt, który zawiera ramki we wszystkich położeniach i posłuży do wycięcia sekwencji wideo opisującej daną akcję. W przypadku aktywności, trwających dłużej oraz obejmujących znaczny obszar znaczny obszar ze względu na przemieszczanie się osoby, tak uzyskane sekwencje nie stanowią jeszcze dobrych danych do uczenia, gdyż znaczna część obszaru nie zawiera danej akcji przez większość czasu. Ponadto stosunek szerokości do wysokości takich sekwencji znacznie odbiega od 1, co nie jest pożądane w algorytmach wykorzystujących sieci neuronowe. Dlatego też dłuższe akcje zostały pocięte na krótsze według następującego algorytmu: jeśli aktywność trwa dłużej niż 3 sekundy, jest dzielona na trzysekundowe odcinki z przesunięciem równym 2,4 sekundy. Wyznaczone obszary wycięcia bez podziału oraz z podziałem na krótsze akcje ilustruje rys. 3.

■ **Rys. 3. Wyznaczone obszary wycięcia dla przykładowej ramki sekwencji testowej**

Dane pochodzą z kamer umieszczonych na dość dużej wysokości, a więc widoczność sceny jest zadowalająca i umożliwia śledzenie wielu postaci. Występują jednak pewne cechy utrudniające rozpoznawanie, jak np. występowanie cienia. Osoby przebywające w cieniu są częściowo lub całkowicie niewidoczne na sekwencjach wideo, co wyklucza możliwość wykonywania rozpoznawania (rys.4).

Ponadto dane zawierają oznaczenia osób poruszających się w grupach w niedalekiej odległości od siebie. Takie osoby mogą przesłaniać się wzajemnie, wprowadzając błąd w procesie trenowania. Przesłanianie może być też spowodowane obiektami



■ **Rys. 4. Przykładowy obraz ilustrujący trudne warunki oświetleniowe**

architektury, jak np. fontanna. Powiązanym problemem jest ustawienie osoby tyłem do kamery, przez co część informacji jest tracona. Na uwagę zasługuje też problem wyznaczenia granicy dwóch akcji. Przykładowo akcje korzystania z telefonu i robienia zdjęcia pokrywają się. W większości przypadków ta druga zawiera się w pierwszej. Granica między tymi akcjami jest dość umowna.

Jednym ze sposobów zwiększenia ilości danych i zmniejszenia błędu generalizacji jest tzw. osadzanie danych. Podczas trenowania wykonywane są przekształcenia danych, takie jak: odbicie lustrzane, zmiana kolorystyki, kadrowanie. Ponadto, w celu zwiększenia efektywności wykrywania ruchu, w odrębnej sieci obliczany jest przepływ optyczny (rys. 5), który następnie przekazywany jest obok standardowych danych do wejścia sieci rozpoznającej aktywności.

EKSPERYMENTY

Do oceny działania systemu stosowana jest metodologia podobna jak w konkursie *Multiple Object Tracking* (**MOT**) Challenge [3]. W konkursie tym zdefiniowany został zestaw miar oceniających zarówno jakość detekcji obiektów w pojedynczych ramkach sekwencji wideo, jak również jakość śledzenia obiektów w kolejnych ramkach i poprawność wyznaczania ich ścieżki ruchu. Podstawowymi miarami do oceny jakości detekcji w pojedynczym obrazie są: liczba poprawnie wykrytych obiektów **TP** (*True Positives*), liczba niewykrytych obiektów **FN** (*False Negatives*) oraz liczba fałszywych detekcji **FP** (*False Positives*).

W kontekście śledzenia obiektu w kolejnych ramkach sekwencji wideo istotne jest, aby ten sam obiekt miał zawsze nadany taki sam identyfikator. Oznacza to, że system rozpoznaje go jako „ten sam" obiekt w kolejnych ramkach. Jeśli tak nie jest, tzn. w kolejnych ramkach temu samemu rzeczywistemu obiektowi zostały nadane różne identyfikatory, jest to traktowane jako błąd. Do oceny systemu pod tym względem, w każdej ramce sekwencji, oprócz wyżej wymienionych TP, FN i FP, wyliczana jest również liczba zmiany identyfikatora obiektu **IDSW** (*Identity Switch*).

Do oceny poprawności identyfikacji obiektów obliczane są ponadto miary IDP, IDR i IDF1 [4]. IDP jest miarą precyzji identyfikacji (*identification precision*) i określa, jaka część dokonanych detekcji i przypisania identyfikatorów obiektu jest poprawna. Z kolei **IDR** (*Identification Recall*) określa, jaka część spośród wszystkich obiektów referencyjnych (*ground truth*) została poprawnie zidentyfikowana. Natomiast IDF1, obliczana jako stosunek poprawnych identyfikacji do średniej liczby wszystkich detekcji i obiektów referencyjnych, łączy obydwie powyższe miary i umożliwia ocenę systemu za pomocą pojedynczej wartości:

$$IDP = \frac{ID_{TP}}{(ID_{TP}+ID_{FP})} \tag{1}$$

$$IDR = \frac{ID_{TP}}{(ID_{TP}+ID_{FN})} \tag{2}$$

$$IDF1 = \frac{2*ID_{TP}}{2*ID_{TP}+ID_{FP}+ID_{FN}} \tag{3}$$

gdzie: ID_{TP} jest liczbą poprawnych identyfikacji ID (*true positive*), ID_{FP} jest liczbą błędnych identyfikacji FP (*False Positive*), a ID_{FN} jest liczbą niezidentyfikowanych obiektów referencyjnych FN (*False Negative*).

Do całościowej oceny jakości systemu detekcji wykorzystywane są dwie miary, uwzględniające wyniki uzyskiwane dla poszczególnych ramek. Są to: dokładność śledzenia obiektów **MOTA** (*Multiple Object Tracking Accuracy*) oraz precyzja lokalizacji obiektów **MOTP** (*Multiple Object Tracking Precision*).

MOTA jest jedną z najpowszechniej wykorzystywanych miar do oceny jakości detekcji i śledzenia obiektów ze względu na to, że uwzględnia różne źródła błędów wymienione powyżej (niewykrycie obiektów, fałszywe detekcje oraz zmiana identyfikatora):

$$\mathrm{MOTA} = 1 - \frac{\sum_t (FN_t+FP_t+IDSW_t)}{\sum_t GT_t} \tag{4}$$

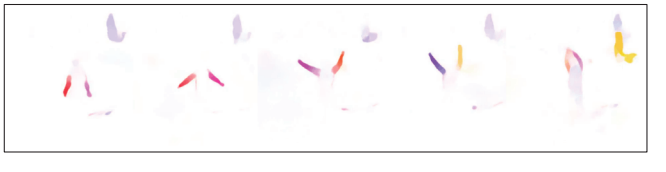
gdzie: t jest numerem ramki, GT jest liczbą obiektów referencyjnych (*Ground Truth*), a *FN*, *FP* i *IDSW* – jak opisano powyżej.

Natomiast MOTP jest miarą określającą dokładność przestrzennej lokalizacji obiektów w obrazie, gdyż uwzględnia wzajemne pokrycie obszarów określonych przez ramki (*bounding box*) wyznaczające obiekt:

$$\mathrm{MOTP} = \frac{\sum_{t,i} d_{t,i}}{\sum_t TP_t} \tag{5}$$

gdzie $d_{t,i}$ określa stopień pokrycia ramki otaczającej obiekt i z odpowiadającym mu obiektem referencyjnym.

Ponadto do oceny jakości śledzenia obiektów (wyznaczania ścieżek) każdy ze śledzonych obiektów kwalifikowany jest do kategorii: przeważnie śledzony (*mostly tracked*), częściowo śledzony (*partially tracked*) i przeważnie gubiony (*mostly lost*). Decyzja podejmowana jest na podstawie poprawności przypisania identyfikatora obiektu w kolejnych ramkach, a zatem zgodności ścieżki obiektu śledzonego ze ścieżką rzeczywistą. Obiekt kwalifikowany jest jako „przeważnie śledzony", jeśli jego ścieżka jest poprawnie wyznaczona na co najmniej 80% jej długości, a jako „przeważnie gubiony", jeśli ścieżka jest poprawnie wyznaczona na mniej niż 20% jej długości.



■ **Rys. 5. Przykładowa sekwencja obrazów przepływu optycznego**

Powyższe miary służą do oceny jakości detekcji i śledzenia obiektów w pojedynczej sekwencji wideo.

Jednym z zadań systemu jest śledzenie ruchu wielu postaci znajdujących się w obrębie pola gry w czasie rzeczywistym. W tym celu przygotowano zestaw danych, w postaci nagrań zrealizowanych przez zespół projektowy na rynku Przemyśla. Dane te, wraz z ogólnodostępnym zbiorem MOT17, wykorzystano do nauki modelu sieci neuronowej. Uzyskane wyniki zweryfikowano natomiast, wykorzystując publiczny zbiór MOT16 oraz nagrania przygotowane przez zespół. Uzyskane wyniki przedstawiono w tabeli 1.

Tabela 1. Wyniki uzyskane na wybranych nagraniach

Parametr	Jednostka	MOT16-13	Przemyśl
IDF1	[%]	80,17	89,68
IDP	[%]	85,15	92,64
IDR	[%]	75,74	86,91
MOTA	[%]	77,26	87,47
MOTP	[%]	22,33	15,00

Przygotowany model okazał się skuteczny w warunkach miejskich, co zaowocowało m.in. możliwością wykorzystania wyników śledzenia w pracach nad drugim komponentem zaprojektowanej architektury – systemem rozpoznawania akcji.

Do wyznaczenia pozycji XYZ w przestrzeni 3D konieczna jest znajomość parametrów wewnętrznych kamer oraz kalibracja układu kamer, w celu wyznaczenia parametrów zewnętrznych (obrót i przesunięcie kamer względem siebie). Do wykonania kalibracji układu kamer potrzebny jest zestaw punktów, dla których znana jest rzeczywista pozycja XYZ w przestrzeni 3D, jak również ich położenie w obrazach pobranych z poszczególnych kamer. Procedurę tę należy wykonać oddzielnie dla każdej lokalizacji, w której ma działać system detekcji i śledzenia. Po skalibrowaniu układu kamer, na podstawie wyników detekcji w różnych sekwencjach (widokach), można wyliczyć pozycję XYZ śledzonych obiektów, stosując procedurę triangulacji.

Do oceny systemu pod względem dokładności pozycjonowania w przestrzeni 3D obliczana jest odległość euklidesowa pomiędzy rzeczywistą pozycją obiektu a tą wyznaczoną przez system. Obliczana jest odległość dla każdego obiektu, jak również wartość uśredniona dla wszystkich obiektów.

System realizowany w ramach projektu *Opracowanie i implementacja systemu sztucznej inteligencji wirtualnych postaci pozwalającego na symulację ich realistycznych zachowań i interakcji z graczem na podstawie autonomicznej analizy obrazu uczestników gry w czasie rzeczywistym*" POIR.01.02.00-00-0204/20.

LITERATURA

[1] Carreira J., A.Zisserman. 2018. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". arXiv:1705.07750 (dostępne pod adresem: https://arxiv.org/abs/1705.07750)

[2] Feichtenhofer Ch., Fan Haoqi, J.Malik, H.Kaiming. 2019. "SlowFast Networks for Video Recognition". arXiv:1812.03982v3 (dostępne pod adresem: https://arxiv.org/abs/1812.03982).

[3] Milan A., L.Leal-Taixe, I.Reid, S.Roth, K.Schindler. 2016. "MOT16: A Benchmark for Multi-Object Tracking", arXiv 1603.00831, 2016 (dostępne pod adresem: https://arxiv.org/abs/1603.00831)

[4] Ristani E., F.Solera, R.Zou, R.Cucchiara, C.Tomasi. 2016. „Performance Measures and a Data Set for Multi-target, Multi-camera Tracking". In: Hua Gang, Jégou, Hervé (eds), „Computer Vision – ECCV 2016 Workshops", ECCV 2016. *Lecture Notes in Computer Science*, vol 9914. Springer

[5] Zhang Yifu, Wang Chunyu, Wang Xinggang, Zeng Wenjun, Liu Wenyu. 2021. "FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking". *IEEE Transactions on Communications Systems*, 11 (4), 360–393.

Cezary ZIÓŁKOWSKI*, Jan M. KELNER*

Ukierunkowanie wiązek antenowych a przepustowość kanału radiowego w warunkach propagacji NLOS

Orientation of antenna beams versus radio channel capacity in non-line-of-sight (NLOS) propagation conditions

Wpływ dokładności ukierunkowania wiązek antenowych na przepustowość kanału radiowego ma istotne znaczenie w praktycznym wykorzystaniu systemów massive-MIMO. W artykule przedstawiono metodykę oceny wpływu niewspółosiowości wiązek systemów antenowych na poziom mocy odbieranego sygnału, a w konsekwencji przepustowość kanału. Ocenę ilościową tego zagadnienia przedstawiono na podstawie badań symulacyjnych przeprowadzonych dla przykładowego scenariusza z uwzględnieniem propagacji w zakresie 28 GHz w warunkach NLOS.
Słowa kluczowe: propagacja wielodrogowa, przepustowość kanału, wieloelipsoidalny model propagacji

The influence of the orientation accuracy of the antenna beams on the radio channel capacity is important in the practical use of massive-**MIMO** (*Multiple-Input Multiple-Output*) systems. This paper presents a methodology for assessing the impact of misalignment of antenna systems bundles on the received signal power level and, consequently, the channel capacity. The quantitative assessment of this issue was presented based on simulation tests carried out for an example scenario, taking into the propagation in the range of 28 GHz for non-line-of-sight (NLOS) conditions account.
Key words: channel capacity, multi-ellipsoidal propagation model, multi-path propagation

DOI: 10.15199/59.2022.4.9

Efektywny wzrost pojemności sieci bezprzewodowych w stosunku do sieci **LTE** (*Long Term Evolution*) jest jednym z głównych celów rozwoju systemów piątej generacji (5G). Ten cel wynika z dynamicznego wzrostu liczby użytkowników i rosnącego zapotrzebowania na zakres realizowanych usług telekomunikacyjnych. W celu zapewnienia większych szerokości pasm poszczególnych łączy radiowych, systemy 5G – oprócz dolnych zakresów częstotliwości mikrofal (fal decymetrowych i centymetrowych) [1] – funkcjonują również w pasmach fal milimetrowych. Jednakże zasadniczym mankamentem wykorzystania tych zakresów pasma jest istotny wzrost tłumienia wzdłuż ścieżki propagacji pomiędzy nadajnikiem (TX) a odbiornikiem (RX). Podstawowym sposobem kompensacji tego wzrostu tłumienia jest zwiększanie zysku energetycznego systemu antenowego, który jest odwrotnie proporcjonalny do rozwartości charakterystyk promieniowania. Z tego powodu, w systemach 5G, zwłaszcza dla zakresów fal milimetrowych, stosowane są anteny kierunkowe lub systemy wieloantenowe, na podstawie których realizowana jest technologia massive-**MIMO** (*Multiple-Input Multiple-Output*) [1]. Tego typu systemy antenowe umożliwiają formowanie wiązek [1], co zapewnia realizację techniki multipleksacji przestrzennej łączy radiowych.

Głównym celem artykułu jest przedstawienie metodyki oceny wpływu parametrów i doboru orientacji systemu anten na przepustowość kanału radiowego w warunkach **NLOS** (*Non-Line-Of-Sight)*. Minimalizacja strat mocy wynikających z niedopasowania wiązek antenowych i optymalny wybór ich ukierunkowania jest podstawą analizy jakości łącza radiowego zdefiniowanego przez **SNR** (*Signal-to-Noise Ratio*). Otrzymane wyniki mają charakter statystyczny i są podstawą do oceny stopnia degradacji przepustowości kanału w wielodrogowym środowisku propagacji z uwzględnieniem warunków NLOS. Prezentowana metodyka umożliwia ocenę skuteczności procedur wyznaczania optymalnej orientacji wiązek antenowych, która zapewnia maksymalną przepustowość w określonych warunkach propagacji. Do oceny zmian poziomu odbieranego sygnału dla różnych kierunków wiązek antenowych stosuje się wieloeliptyczny (lub wieloelipsoidalny) model propagacyjny **MPM** (*Multi-elliptical/ellipsoidal Propagation Model*) [4], oparty na wieloelipsoidalnej geometrii położenia elementów rozpraszających falę elektromagnetyczną. Wyniki uzyskiwane na podstawie tego modelu są porównywalne z rezultatami, jakie daje procedura zalecana przez standard 3GPP [6], co pokazano w publikacji [3]. Innowacyjnym elementem opracowanej metodyki jest możliwość niezależnej oceny wpływu zarówno środowiska, jak i parametrów systemów antenowych na możliwości transmisyjne kierunkowych łączy w warunkach NLOS.

PRZEPUSTOWOŚĆ A TYP ŚRODOWISKA I SYSTEMU ANTENOWEGO

Twierdzenie Shannona-Hartleya wprowadza fundamentalną zależność, która opisuje względną przepustowość kanału transmisyjnego jako funkcję SNR:

$$C_f[\text{bit/s/Hz}] = \log_2(1 + SNR) \tag{1}$$

gdzie: $SNR = P_f/P_n$, P_f to moc sygnału użytecznego, a P_n reprezentuje moc sygnału zakłóceń.

* Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Wydział Elektroniki, Instytut Systemów Łączności, Warszawa, e-mail: cezary.ziolkowski@wat.edu.pl, jan.kelner@wat.edu.pl

W przeprowadzonej analizie wyrażenie (1) definiuje wartość referencyjną przepustowości kanału radiowego z uwzględnieniem dookólnych charakterystyk promieniowania systemu antenowego i propagacji w warunkach wolnej przestrzeni. W tym przypadku tłumienie propagacji jako funkcję odległości D opisuje zależność:

$$PL_f(D)[\mathrm{dB}] = 20\log(4\pi D/\lambda) \tag{2}$$

gdzie $\lambda = c/f_0$, c i f_0 są odpowiednio prędkością propagacji i częstotliwością fali elektromagnetycznej.

W środowisku propagacji wielodrogowej, zwłaszcza w warunkach NLOS, poziom sygnału użytecznego ulega redukcji, zatem w tych warunkach zmniejsza się *SNR*, a w konsekwencji przepustowość kanału. Przedstawiona w artykule analiza przepustowości w środowisku wielodrogowym opiera się na modelu tłumienia typu **CI** (*Close-In free space reference distance path loss model*)[2]. Model ten opisuje zależność:

$$PL_m(D)[\mathrm{dB}] = PL_m(D_0) + 10\,PLE \cdot \log(D/D_0) \tag{3}$$

gdzie PLE jest współczynnikiem tłumienia propagacji zależnym od typu środowiska, a D_0 to odległość referencyjna.

W celu uwzględnienia wpływu warunków propagacyjnych wprowadzone zostało pojęcie współczynnika środowiskowego $K_e(D)$, który wyraża relację pomiędzy mocami odbieranych sygnałów w środowisku wielodrogowym $P_m(D)$ oraz w wolnej przestrzeni $P_f(D)$ dla systemów anten dookólnych. Uwzględniając odwrotną proporcjonalność pomiędzy mocą odbieranego sygnału a tłumienia środowiska, współczynnik ten można wyrazić jako:

$$K_e(D) = \frac{P_m(D)}{P_f(D)} = \frac{PL_f(D)}{PL_m(D)} \tag{4}$$

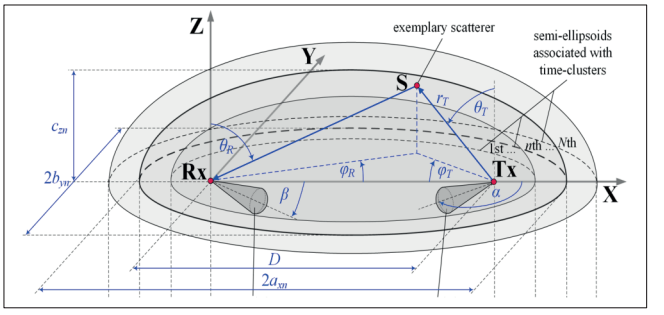
Aby uwzględnić wpływ systemu antenowego na przepustowość kanału, wprowadza się współczynnik $K_a(D)$ definiowany zależnością:

$$K_a(D) = \frac{P_s(D)}{P_m(D)} \tag{5}$$

Wprowadzenie współczynników $K_e(D)$ i $K_a(D)$ umożliwia wyznaczenie zależności funkcjonalnej pomiędzy mocami sygnałów $P_s(D)$ i $P_f(D)$ odbieranych w łączach z antenami o wąskich i dookólnych wiązkach w środowiskach propagacji wielodrogowej i w wolnej przestrzeni. Stąd względną przepustowość C kanału z uwzględnieniem środowiska propagacji wielodrogowej i systemu antenowego o wąskich wiązkach można wyrazić wzorem:

$$C = \log_2(1 + K_e(D)K_a(D)SNR) \tag{6}$$

Z równania (4) wynika, że podstawą do wyznaczenia współczynnika środowiskowego $K_e(D)$ są modele tłumienia dla wolnej przestrzeni i dla warunków propagacji wielodrogowej. Natomiast ocena współczynnika systemu antenowego $K_a(D)$ opiera się na wykorzystaniu geometrycznego modelu propagacji MPM. Dzięki temu możliwe jest odwzorowanie wpływu zarówno parametrów anteny, jak i właściwości transmisyjnych środowiska na przepustowość kanału.



■ **Rys. 1. Struktura geometryczna modelu MPM**

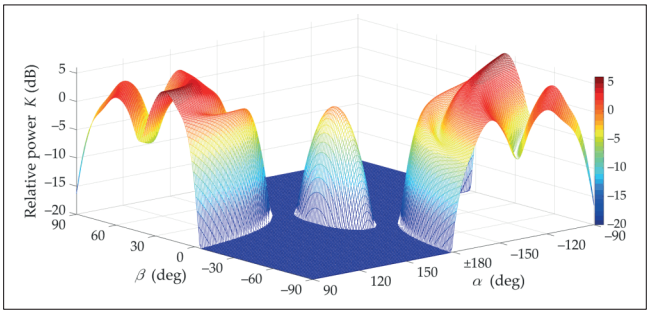
ESTYMACJA KĄTOWEGO ROZKŁADU MOCY

Wieloelipsoidalny model propagacji MPM jest podstawą dla estymacji kątowego rozkładu mocy **PAS** (*Power Angular Spectrum*), $p(\varphi_R)$, sygnału docierającego do odbiornika, gdzie $\varphi_R = (\theta_R, \varphi_R)$ reprezentuje kąty odbioru sygnału **AOA** (*Angle Of Arrival*) w płaszczyźnie odpowiednio elewacji i azymutu. Na rys. 1 przedstawiona jest geometria MPM, czyli obszarów potencjalnego występowania elementów rozpraszających, które reprezentowane są przez współogniskowe półelipsoidy [4]. Taka struktura geometryczna wynika z profilu opóźnienia mocy **PDP** (*Power Delay Profile*), która jest jedną z podstawowych charakterystyk opisujących właściwości transmisyjne kanału.

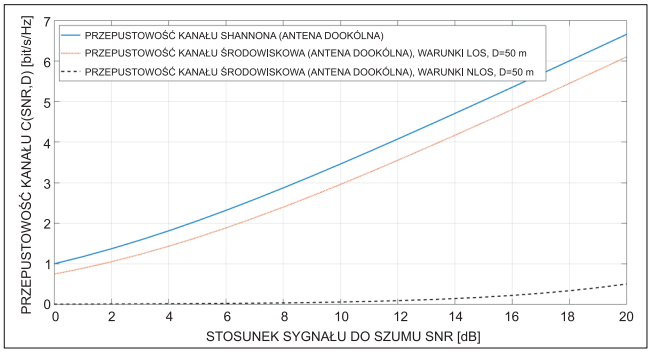
Parametry struktury geometrycznej MPM są ściśle powiązane z właściwościami transmisyjnymi środowiska propagacji opisanego przez PDP. W przypadku środowisk z propagacją wielodrogową można zaobserwować obecność kilku lub kilkunastu ekstremów lokalnych funkcji PDP. Oznacza to, że składowe pochodzące od pojedynczych rozproszeń i docierające z różnymi opóźnieniami decydują o kątowym rozkładzie poziomu odbieranego sygnału. To jest podstawą wykorzystania półelipsoid do odwzorowania najbardziej prawdopodobnych lokalizacji elementów rozpraszających. Liczba elipsoid jest równa liczbie ekstremów PDP, które reprezentują klastry czasowe składowych docierających do odbiornika o jednakowych opóźnieniach. Estymacja $p(\varphi_R)$ polega na wyznaczeniu trajektorii ścieżek propagacji pochodzących z TX i docierających do RX. Ścieżki te uwzględniają wieloelipsoidalną geometrię położeń elementów rozpraszających. Ale struktura geometryczna MPM odwzorowuje lokalizacje elementów rozpraszających. Zatem przecięcie kierunku drogi propagacji z poszczególnymi półelipsami wskazuje położenie elementów rozpraszających. Na podstawie kąta wypromieniowania **AOD** (*Angle Of Depature*), $p(\varphi_T) = p(\theta_T, \varphi_T)$ gdzie θ_T i φ_T są kątami wypromieniowania odpowiednio w płaszczyźnie elewacji i azymutu, można wyznaczyć współrzędną radialną rozpraszacza w układzie kulistym z początkiem w TX. Dokonując transformacji współrzędnych względem początku układu w punkcie RX, wyznacza się kierunki odbioru sygnału. Szczegółowy opis procedury estymacji kątowego rozkładu mocy $p(\theta_R, \varphi_R)$ na podstawie MPM przedstawiono w [3]. Wyznaczony w procedurze symulacyjnej PAS jest podstawą do oceny mocy $P_s(D)$ odbieranego sygnału zgodnie z zależnością:

$$P_s(D) = \int_{-180^\circ}^{180^\circ}\int_{0^\circ}^{90^\circ} p(\theta_R, \varphi_R)|g(\theta_R, \varphi_R)|^2\, d\theta_R d\varphi_R \tag{7}$$

gdzie: $|g(\theta_R, \varphi_R)|^2$ jest unormowaną charakterystyką odbioru mocy anteny odbiorczej.



■ **Rys. 2. Względny współczynnik mocy *K* (α, β, D) jako funkcja kierunków wiązek antenowych α i β**



■ **Rys. 3. Przepustowość jako funkcja SNR dla dookólnych charakterystyk systemu antenowego i zróżnicowanych warunków propagacyjnych**

Z równań (5) i (7) wynika, że obliczenie współczynnika $K_a(D)$ sprowadza się do estymacji $p(\theta_R, \varphi_R)$.

UKIERUNKOWANIE WIĄZEK ANTENOWYCH A MOC SYGNAŁU

Model MPM nie zapewnia oceny tłumienia w sposób bezpośredni, lecz jedynie daje możliwość oceny PAS jako funkcji znormalizowanej. W praktyce do oceny utraty ścieżki można wykorzystać wiele modeli pochodzących ze statystycznie uśrednionych danych pomiarowych. Modele te są jednak zdefiniowane dla wiązek skierowanych na siebie i d dla wybranych szerokości wiązek **HPBW** (*Half-Power BeamWidth*) systemu antenowego, np. [2]. Przedstawiona w [3] metodyka oparta na MPM zapewnia modyfikację tłumienia i bilansu mocy dla różnych HPBW i orientacji wiązek antenowych względem modeli empirycznych. Podstawą tej modyfikacji jest względny współczynnik mocy $K(\alpha, \beta, D)$, który reprezentuje względną moc dla zróżnicowanych kierunków wiązek antenowych

$$K(\alpha, \beta, D)[dB] = 10\log\frac{P_S(\alpha,\beta,D)}{P_S(\alpha=180°,\beta=0°,D)} \qquad (8)$$

gdzie $P_s(\alpha, \beta, D) = P_s(D)$ jest mocą odebranego sygnału dla kierunków i wiązek anten odpowiednio nadawczej i odbiorczej jak na rys. 1.

Badania symulacyjne są podstawą do oceny strat mocy wynikających z niedopasowania wiązek antenowych w łączu kierunkowym oraz umożliwiają optymalny dobór ich orientacji w warunkach NLOS. W przeprowadzonych badaniach wykorzystano scenariusz przestrzenny jak na rys. 1 oraz przyjęto następujące założenia:

- częstotliwość nośna jest równa f_0 = 28 GHz;
- PDP oparty jest na modelu linii opóźniającej **TDL** (*Tapped-Delay Line)* ze standardu 3GPP TR 38.901 [6], tj. TDL-B, dla analizowanej częstotliwości parametry TDL są przyjęte na podstawie scenariusza typu **UMa** (*Urban Macro*), dla którego rms rozmycia opóźnienia **DS** (*Delay Spread*) wynosi σ_r = 266 ns;
- odległość pomiędzy TX i RX jest równa D = 50 m;
- w charakterystykach promieniowania mocy wiązek antenowych uwzględnione są tylko listki główne;
- charakterystyki wiązek antenowych są odwzorowane za pomocą modelu Gaussa [5];
- wartości HPBW anten nadawczej i odbiorczej są takie same zarówno w płaszczyźnie azymutu, jak i elewacji i wynoszą $HPBW_{T\theta,R\theta} = HPBW_{T\varphi,R\varphi} = 10°$;
- zyski energetyczne anten nadawczej i odbiorczej są równe i wynoszą $G_T = G_R$ = 24,6 dBi;
- wysokości anten nadawczej (7 m) i odbiorczej (1,5 m) przyjęto na podstawie scenariuszy pomiarowych [2];
- liczba ścieżek generowanych dla każdego klastra czasowego (półelipsoidy) w badaniach symulacyjnych metodą Monte-Carlo wynosi L = 10.

Na rys. 2 przedstawione są wyniki badań obrazujące zależność wartości współczynnika mocy $K(\alpha, \beta, D)$ jako funkcji kątów skierowania wiązek antenowych TX (α) i RX (β) dla D = 50 m.

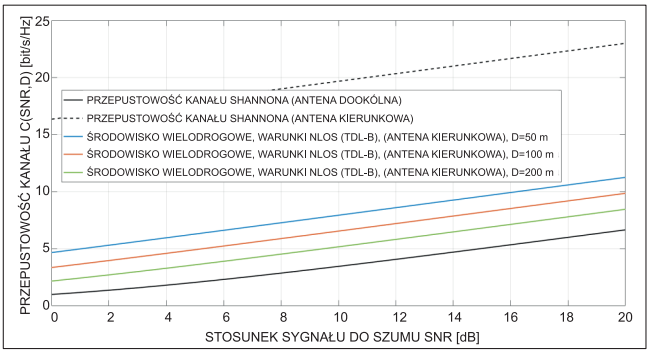
Z wykresu przedstawionego na rys. 2 wynika, że w warunkach NLOS, stosując optymalny dobór kątów wiązek TX i RX przy ograniczonym zakresie zmian $|\alpha|$ <90° oraz przy założonych parametrach anten i środowiska propagacji, można uzyskać wzrost odbieranej mocy o około 6 dB.

Oznacza to, że w tych warunkach propagacji wzajemna współosiowość wiązek nie zapewnia najwyższego poziomu odbieranego sygnału. Dlatego w systemach wykorzystujących technologie *massive*-MIMO i formowania wiązki pożądane jest stosowanie dodatkowych procedur przetwarzania odbieranych sygnałów, które zapewnią optymalny dobór kierunków wiązek antenowych.

UKIERUNKOWANIE WIĄZEK ANTENOWYCH A PRZEPUSTOWOŚĆ

Przedstawione wyniki pokazują, że w warunkach propagacji NLOS uzyskanie maksymalnego poziomu mocy odbieranego sygnału w łączu radiowym z antenami kierunkowymi wymaga wyznaczenia optymalnego ukierunkowania wiązek antenowych. W tej części artykułu przedstawiona jest ocena wpływu zarówno właściwości transmisyjnych środowiska, jak i ukierunkowania anten na przepustowość w warunkach propagacji NLOS. Jako dane referencyjne przyjęto przepustowość łącza radiowego z dookólnym systemem antenowym w warunkach propagacji wolnoprzestrzennej. Ocena poziomu odbieranego sygnału, który w sposób bezpośredni decyduje o przepustowości łącza radiowego, opiera się na wykorzystaniu modelu MPM w procedurze badań symulacyjnych. Badania te przeprowadzono dla założeń opisanych wcześniej z uwzględnieniem zmian odległości pomiędzy Tx i Rx. Właściwości transmisyjne środowiska propagacji odwzorowuje model propagacji CI opisany zależnością (3). Na wstępie prezentowane są skutki degradacyjnego wpływu zarówno parametrów transmisyjnych środowiska, jak i warunków propagacji na przepustowość. Ten wpływ odwzorowuje zróżnicowanie przepustowości łączy radiowych z dookólnymi

systemami antenowymi jako funkcji referencyjnych wartości SNR, które odpowiadają warunkom propagacji wolnoprzestrzennej. Dla propagacji wolnoprzestrzennej z warunkami LOS, wielodrogowej z warunkami LOS i wielodrogowej z warunkami NLOS, wykresy przepustowości jako funkcje wartości SNR przedstawiono na rys. 3. Dla propagacji wielodrogowej warunki LOS i NLOS są zdeterminowane przez wartości parametru **PLE** (*Path Loss Exponent*) równe odpowiednio *PLE* = 2,1 i *PLE* = 3,4. Uzyskane wykresy pokazują, że przepustowość kanału radiowego w warunkach **LOS** (*Line-Of-Sight*) w odniesieniu do wolnej przestrzeni jest mniejsza jedynie o 0,5 bit/s/Hz, natomiast dla warunków NLOS to zróżnicowanie jest znacznie większe.



■ **Rys. 4. Przepustowość jako funkcja SNR dla kierunkowego systemu antenowego i warunków NLOS**

Zastosowanie systemów antenowych, których charakterystyki promieniowania/odbioru mają wąskie wiązki, jest jednym ze sposobów minimalizacji negatywnych skutków wielodrogowości propagacji w warunkach NLOS. Efekty wykorzystania przestrzennie selektywnych systemów antenowych pokazano na rys. 4.

Wspólny kierunek maksymalnego promieniowania i odbioru anten zapewnia statystycznie kilkunastokrotny wzrost przepustowości kanału radiowego. Dla analizowanego łącza przepustowość jest większa o (1,5÷3,5)bit/s/Hz w odniesieniu do wolnoprzestrzennego kanału radiowego z dookólnym systemem antenowym. Wyniki analizy prezentowane na rys. 4 pokazują, że przepustowość kanału radiowego zależy również od odległości pomiędzy Tx i Rx. Dwukrotne zmniejszenie odległości powoduje wzrost przepustowości kanału radiowego o około 1,5 bit/s/Hz w całym zakresie zmienności SNR. W warunkach NLOS ukierunkowanie bezpośrednie anteny nadawczej i odbiorczej nie zapewnia uzyskania maksymalnego poziomu odbieranego sygnału. Dlatego w tych warunkach celowe jest zastosowanie systemu sterowania, który zapewni wyznaczanie kierunku maksymalnego poziomu odbieranego sygnału. Z rys. 4 wynika, że wykorzystanie kierunku maksymalnego poziomu sygnału zapewnia dodatkowy przyrost przepustowości o (2÷3)bit/s/Hz. Przyrost ten jest tym większy, im odległość pomiędzy Tx i Rx jest większa. W przypadku odległości *D* = 50 m wybór kierunku maksymalnego poziomu zapewnia statystyczny przyrost przepustowości kanału o około 2bit/s/Hz, natomiast dla odległości *D* = 200 m przyrost ten wzrasta do wartości 3 bit/s/Hz. Bardziej szczegółową analizę przepustowości zawarto w [7].

* * *

Przedstawiona w artykule problematyka poświęcona jest ocenie wpływu parametrów systemu antenowego, ze szczególnym uwzględnieniem jego ukierunkowania, na przepustowość kanału radiowego w warunkach NLOS. Potrzeba podjęcia takiej tematyki związana jest z wdrażaniem nowych technologii w systemach antenowych ze sterowaniem parametrami wiązek, których przykładem jest *massive*-MIMO. Przeprowadzona ocena ma charakter statystyczny i opiera się na badaniach symulacyjnych, w których do odwzorowania skutków zjawisk propagacyjnych wykorzystany został wieloelipsoidalny model propagacji MPM. Uzyskane wyniki pokazują, że w warunkach NLOS zastosowanie systemów antenowych o wąskich wiązkach zapewnia nawet kilkunastokrotny wzrost przepustowości kanału radiowego w porównaniu z systemem dookólnym. Dodatkowo zastosowanie systemu sterowania wiązką i wyboru kierunku maksymalnego poziomu odbieranego sygnału prawie dwukrotnie zmniejsza degradację przepustowości kanału. Powyższe wnioski wydają się oczywiste. Jednakże zastosowana procedura zapewnia ilościową ocenę efektywności wykorzystania systemów antenowych ze sterowanymi wiązkami w warunkach NLOS. Wykorzystanie w procedurze symulacyjnej geometrycznego modelu propagacji MPM daje możliwość uwzględnienia nie tylko parametrów i charakterystyk systemu antenowego, ale również typu środowiska propagacji. To decyduje o oryginalności metody analizy przepustowości kanału radiowego opisanej w tym artykule.

LITERATURA

[1] Busari S.A., K.M.S. Huq, S. Mumtaz, L.Dai, J. Rodriguez 2018. "Millimeter-wave massive MIMO communication for future wireless systems: A survey". *IEEE Communications Surveys & Tutorials*, 20(2): 836–869.

[2] Rappaport T.S., G.R.MacCartney, M.K.Samimi, S. Sun, 2015. "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design". *IEEE Transactions on Communications*, 63(9): 3029–3056.

[3] Bechta K., J.M. Kelner, C.Ziółkowski, L.Nowosielski, 2021. "Inter-beam co-channel downlink and uplink interference for 5G New Radio in mm-wave bands". *Sensors*, 21(3): 793.

[4] Ziółkowski C., J.M.Kelner, 2018. "Statistical evaluation of the azimuth and elevation angles seen at the output of the receiving antenna". *IEEE Transactions on Antennas and Propagation,* 66(4): 2165–2169.

[5] Vaughan R., J. Bach Andersen, 2003. *Channels, propagation and antennas for mobile communications*. Institution of Engineering and Technology: London, UK, 2003.

[6] 3GPP. 2019. *Study on channel model for frequencies from 0.5 to 100 GHz (3GPP TR 38.901 Version 16.0.0 Release 16)*. 3rd Generation Partnership Project (3GPP), Technical Specification Group Radio Access Network: Valbonne, France, 2019.

[7] Ziółkowski C., J.M.Kelner, J.Krygier, A.Chandra, A.Prokeš , 2021. "Radio channel capacity with directivity control of antenna beams in multipath propagation environment". *Sensors*, 21(24): 8296.

TECHNICAL DEBT PRIORITIZATION IN TELECOMMUNICATION APPLICATIONS – WHY THE ACTUAL REFACTORING DEVIATES FROM THE PLAN AND HOW TO REMEDIATE IT?

PRIORYTETYZACJA DŁUGU TECHNICZNEGO W APLIKACJACH TELEKOMUNIACYJNYCH – DLACZEGO RZECZYWISTY WYSIŁEK REDUKCJI ODBIEGA OD PLANU I JAK TEMU ZARADZIĆ?

Marek G. Stochel [1,2]
marek.stochel@motorolasolutions.com

Piotr Chołda [2]
piotr.cholda@agh.edu.pl

Mariusz R. Wawrowski [1]
mariusz.wawrowski@motorolasolutions.com

[1]Motorola Solutions, Kraków, Poland
[2]AGH University of Science and Technology, Institute of Telecommunications, Kraków, Poland

DOI: 10.15199/59.2022.4.10

Abstract: **This paper focuses on application of technical debt prioritization technique to the telecommunication software managing the fleet of devices for a video surveillance system. Technical debt for this application was gathered, categorized and prioritized according to the approach named CoDVA. However, the implementation of technical debt reduction did not follow exactly the expected guidelines. The three main causes of this phenomenon were discovered: continuous refactoring approach, sizing of technical debt items, and the broadened scope of refactoring activities. Therefore, we propose to adopt a specific definition of technical debt and follow a few rules for defining its scope and granularity.**

Słowa kluczowe: **dług techniczny, inżynieria oprogramowania, priorytetyzacja długu technicznego, rozwój i ewolucja oprogramowania, zarządzanie długiem technicznym.**

Keywords: **Software Development; Software Engineering; Software Engineering and Debt Metaphors; Software Maintenance and Evolution; Technical Debt; Technical Debt Management, Technical Debt Prioritization.**

1. INTRODUCTION

The evolution of telecommunication software in use, in ever-changing landscape of technologies, tools, and business models causes massive creation of technical debt [1]. However, its understanding evolves in time, and a widely used definition is as follows: *"In software-intensive systems, **technical debt** consists of design or implementation constructs that are expedient in the short term, but set up a technical context that can make a future change more costly or impossible. Technical debt is a contingent liability whose impact is limited to internal system qualities—primarily, but not only, maintainability and evolvability"* [2]. In our previous paper, we claim that this definition should be expanded to embrace all software artefacts responsible for delivering a product to a customer [3], as the business perspective shifts towards service model, rather than one-time sale. Moreover, even with one-time sale the seller is responsible for providing agreed-upon service (bound by Service Level Agreement) for a predefined time. Hence, all software artefacts constituting stable development environment, ensuring testability (automation), code analysis tools, continuous integration and delivery, etc., may be subject to technical debt dynamics. For this reason, they should be considered a part of the telecommunication software solution, which becomes actually a service offered to the customers.

This paper presents the initial results from over 1-year-long experiment, and describes the application of the technical debt prioritization technique to a product managing a fleet of connected devices for a video surveillance system. We decided to use Continuous Debt Valuation Approach (CoDVA) technique for technical debt valuation and prioritization, which we developed and presented earlier [4]. We pose a research question: *Is prioritized technical debt reduction based on CoDVA effective (executed as per plan, bringing tangible benefits)?* During the experiment, a few interesting observations were made. The implementation of technical debt reduction did not follow exactly the expected guidelines. We have discovered three major reasons for that: implementation of continuous refactoring approach, the size of desired technical debt reduction effort (planned refactorings), and the additional required activities which were not originally considered a part of technical debt reduction. Now, after considering the results, we propose to adopt a specific definition of technical debt and follow a few rules for defining the scope of technical debt reduction, prioritizing it, and executing refactoring effort to improve overall software quality and maintainability. The initial results of this experiment proved that quality of the product improved over time, as downward trend of stabilization issues was observed.

Paper structure. The remainder of the paper covers the following topics: Section 2 discusses the background and related work. Section 3 presents methodology, including data collection and analysis processes. The results are shown and discussed in Section 4. Finally, conclusions and future work are presented in Section 5.

2. RELATED WORK

The technical debt term is broadly used; however, due to its metaphorical origin it gradually became ambiguous [5], therefore the operational definition used in our work was quoted earlier in Section I, accompanied by additional explanations. Moreover, the **technical debt** term means also a set of all technical debt items associated

with the system [2]. An atomic **technical debt item (TDI)** points to an immature software development artefact indicating the difference between its current and desired state [4]. The act of reducing technical debt is called **refactoring** [6], or remediation [2]. The evolution of software development and delivery techniques, namely the adoption of the DevOps paradigm and its Continuous Integration and Continuous Delivery practices (CI/CD) requires technical debt to follow. Hence, a broader pool of software development artefacts will be considered a part of telecommunication system or service [3], as more software artefacts undergo technical debt dynamics. We should also avoid a trap of understanding technical debt only according to the scope presented by the tools, e.g. SonarQube[1]. Our understanding should revolve around the difference between the current and the desired state of the artefacts constituting a system or service. Moreover, a change of surrounding solutions (evolving technologies and their adoption by the industry) may be perceived as an act of incurring technical debt by our product (as it results in a change of its desired state).

As technical debt management should be the goal, not the relentless refactoring, many approaches to technical debt prioritization emerged [9][10]. Some papers are mostly conceptual; others discuss frameworks standardizing the process of prioritization of technical debt [10]. In our previous work, we introduced the CoDVA technique for prioritizing technical debt [4]. This approach enables valuation of technical debt to prioritize its reduction according to the business perspective, i.e. indicating changes with the potentially highest positive impact on the future evolution of the product. The evolution of the product itself should be not only driven by adjusting the codebase to enable future changes (which is a risk in agile development approach), but should ensure an explicit allocation of effort to introduce bigger (e.g. architectural) changes [11][12]. Hence continuous refactoring approach [8], driven within the context of ongoing development, may only be a partial solution. Moreover, the perception of the desired future state to optimize towards is provided by the business in the form of a product roadmap (a pipeline of future functionalities expected to be added to the system). In summary, this paper drives further the experiments in adopting the CoDVA approach for technical debt prioritization and provides insight into further evolution of this original approach.

3. METHODOLOGY

This exploratory case study was conducted in a large, international company, which creates, maintains and operates the mission-critical telecommunications systems. The studied application was a software managing a fleet of devices constituting a video surveillance system. During the period of one year several versions of the product were released. The CoDVA technique for technical debt prioritization was applied to drive refactoring effort in parallel with adding new functionality. One agile scrum team, cooperating with many stakeholders from North America and Europe (e.g. Product Managers, Sales Representatives, project teams) was responsible for development and maintenance of this product, and certain budget for refactoring was agreed upon with the Product Manager.

Firstly, technical debt was collected during a brainstorm session within the development team. The results of this session were discussed with stakeholders responsible for business return of engineering investments (Product Managers). Such refined technical debt was stored as separate records in the development database, one record for each atomic TDI. A single record described in detail: scope of work, estimated effort, product context pointing to an area requiring update, and potential consequences associated (justification for a change).

Secondly, the prioritized business roadmap was discussed with the Product Managers. The minor difference between the CoDVA approach presented in our previous work [4] and the currently used one was the valuation process. Instead of using pure monetary value of benefits (deriving it from expected sales), a broader perspective was adopted, taking into account parameters like investment size, alignment with overall product strategy, time-to-market expectations, and many more. As a result, the business value is calculated by the Product Manager for each feature, called feature score (**f_score**). The normalized f_score value (in the range 0..1) is called relative feature priority **f_priority**. This constitutes a basis for prioritization of product roadmap.

Thirdly, a mapping of TDIs on the feature roadmap was prepared (see a sample excerpt in Tab. 1 below). Such a matrix helped to understand the **TDI_impact** of the refactoring effort on the product (S, M, L, XL), explicitly listing anticipated benefits.

			Features (Id and Size)								
Technical Debt Items			F10	F11	F12	F13	F14	F15	F16	F17	F18
Id	Type	Size	XL	XL	XL	XL	L	S	M	S	XL
TDI-1	testability	M	XL	XL	XL	-	S	S	M	XL	S
TDI-2	testability	S	XL	L	L	-	S	S	M	XL	S
TDI-6	CI/CD	XL	XL	XL	XL	-	-	S	-	M	-
TDI-3	CI/CD	M	-	XL	XL	-	L	L	L	L	L
TDI-4	maintainability	S	XL	XL	S	-	-	L	S	-	-
TDI-8	CI/CD, Perfomance	S	XL	XL	XL	-	-	-	-	-	-
TDI-15	testability	S	-	XL	XL	-	M	M	M	M	M
TDI-17	CI/CD	S	-	XL	XL	-	S	S	S	S	S
TDI-18	CI/CD	S	-	XL	XL	-	S	S	S	S	S
TDI-7	Development Environment	M	XL	-	XL	-	-	-	-	XL	-
TDI-10	CI/CD	S	L	-	XL	-	M	M	M	M	M
TDI-14	Development Environment	S	-	-	XL	S	-	-	-	S	-
TDI-9	maintainability	S	L	-	S	-	-	-	-	-	-
TDI-19	CI/CD, testability	M	M	XL	XL	-	M	M	M	M	M

Tab. 1 Impact of TDIs on feature roadmap.

Next **CoDVA_INDEX** for each refactoring was calculated, based on: feature scores, feature sizes, TDI impacts, TDI costs, and a **release_cadence** factor. The more features were positively impacted by a given TDI refactoring, the higher CoDVA_INDEX was associated with it. Additionally, feature scores were modified based on their assignment to future releases of the product (release_cadence). The later in the release chain the feature was planned, the lesser was its value taken into account for calculating CoDVA_INDEX, as the scope

[1] https://www.sonarqube.org/

volatility for future product releases had to be taken into account. Expecting no changes in the scope of future releases, we may keep the values intact for the complete roadmap (release_cadence→1), expecting ~100% change every release there is no point for assessing anything beyond the features planned for the next release (release_cadence→0). As a result, a prioritized list of TDIs was created, and refactoring effort was planned accordingly.

3.1. Data collection process

During the experiment, the development of the prioritized sequence of these improvements and their impact on the product was tracked to assess the effectiveness of the CoDVA approach, i.e. whether the improvements were executed as per plan and produce tangible benefits, as it was stated in our research question. As the **dynamics of a feature roadmap** played an important role in prioritizing technical debt items, we were monitoring the volatility of relative feature priorities for both: features already delivered and features still planned in the roadmap. Next, based on priorities on the feature roadmap we were recalculating the CoDVA_INDEX and prioritized the list of technical debt items accordingly. For each of the TDIs we monitored the status of implementation, and if refactoring could not be conducted or was delivered prior to expectations, we collected data to understand the causes.

Additionally, to observe the impact on the product, we were monitoring a number of stabilization cycles required to release a new product version. Each stabilization cycle embraced execution of a test plan and correction of all issues preventing the release of a new product version. Once all critical issues were addressed and the test cycle did not expose any new blockers, the new product version could be released.

3.2. Data analysis process

Agile development is conducted in phases, called sprints. The result of each sprint should be a potentially releasable product version. Whether the version is actually released, it is a decision of the Product Manager, who assesses its value from the customer perspective. In each sprint, the development team meets for an event called refinement. Its goal is to decompose the planned functionality in a way it is understood and ready for development in the next one or further sprints. This is also an opportunity to decompose technical debt items the same way. In this moment the team assesses the potential value of technical debt items, decomposes the top ones and, just before the next sprint starts, selects ones to be refactored (plans technical debt reduction). After the sprint ends, the potentially shippable product increment is demonstrated to stakeholders, including the results of technical debt reduction efforts. Additionally, an assessment of prioritized technical debt was made to understand whether it is following the plan. Every release (usually after a few sprints) the assessment of the feature roadmap was conducted, feature priorities were reviewed and the technical debt roadmap was updated (e.g. assessing if new elements should be added).

Additionally, an ad-hoc update of the technical debt roadmap, or the feature roadmap was called if either engineering team or business team were getting new data which affected the plans. The engineering team could run into unexpected complexity or wanted to consult their ideas with the Product Manager, and on the other hand, the Product Manager had new updates reflecting the customer feedback and prioritization. In extreme cases, a sprint or a release was cancelled to prioritize a new urgent business need.

4. DISCUSSION OF RESULTS

Technical debt prioritization is a dynamic process, and should reflect business perspective to provide the highest return on investment. Observing high dynamics of the business roadmap, it is evident that we should not base our judgment of benefits on engineering assessment only. Pure engineering perspective may result in creating waste by optimizing codebase which will not be changed in the future or, even worse, which will be retired. In the latter case, whatever is the engineering perception of the technical debt associated with code to be retired, the potential value of refactoring efforts is zero from the customer or business perspective.

First, we monitored the trend of relative value of features from the business perspective that evidently demonstrated dynamics of the product roadmap. For clarity, we present only 13 of them that exceeded (at least once) the arbitrarily chosen threshold of relative feature priority (0.35). They are presented here separately in two figures: Fig.1 shows six features not yet delivered and Fig. 2 presents seven features that were already delivered to customers as a part of subsequently released new product versions.

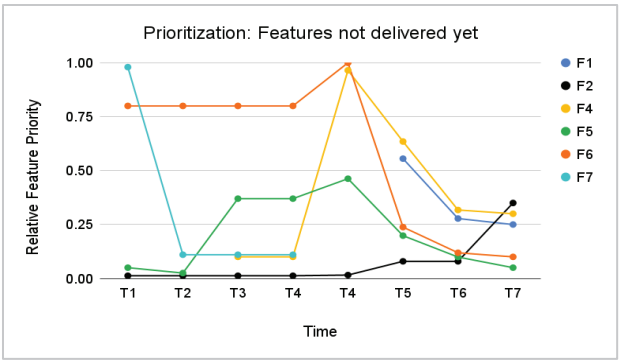


Fig. 1. Product roadmap: planned features.

Moreover, in Fig. 1 and Fig. 2 only the subset of features is presented for legibility reasons, as the number of features was a few times higher. As we can see, even though the prioritization was dynamically updated we could find a situation that two (or more) features were having the same relative priority (FN7 in Fig. 2 and F2 in Fig 1). In such a situation, the engineering team consulted with the Product Manager which feature should be selected for implementation. In this case feature F2 was chosen.

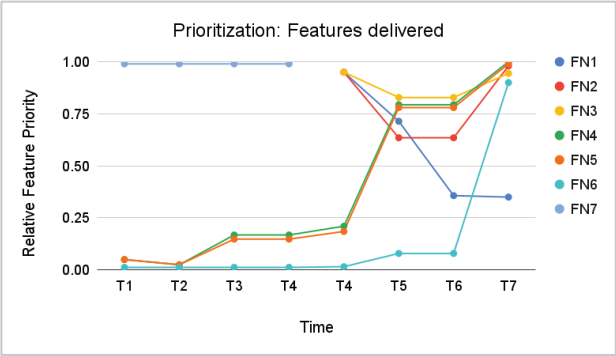


Fig. 2. Product roadmap: implemented features.

An additional noteworthy observation on the feature roadmap dynamics is the fact that features prioritized on top of the list once at certain point of time may be deprioritized, and not implemented at all. This is the result of the dynamics in the sales process; e.g. the window of opportunity for the sale may disappear (see features F4, F6, F7 in Fig. 1).

Tab. 2. Technical Debt Reduction – Prioritized TDIs

TDI	*TDI Type*	*Status*	*Comment*
TDI-1	testability	done	
TDI-2	testability	done	
TDI-3	CI/CD	done	
TDI-4	maintainability	done	
TDI-5	maintainability	done	
TDI-6	CI/CD	too big	
TDI-7	Development Environment	blocked	
TDI-8	CI/CD, Perfomance	done	
TDI-9	maintainability	done	
TDI-10	CI/CD		
TDI-11	security		
TDI-12	maintainability		
TDI-13	maintainability		
TDI-14	Development Environment	done	context
TDI-15	testability		
TDI-16	testability		
TDI-17	CI/CD	done	context
TDI-18	CI/CD	done	context
TDI-19	CI/CD, testability		
TDI-20	logging/telemetry		
TDI-21	maintainability		
TDI-22	security		
TDI-23	maintainability		
TDI-24	testability		
TDI-25	maintainability		
TDI-26	maintainability	done	context
TDI-27	CI/CD	done	context
TDI-28	maintainability	done	context
TDI-29	maintainability	done	context
TDI-30	logging/telemetry, performance		
TDI-31	logging/telemetry, performance		
TDI-32	maintainability		
TDI-33	maintainability		
TDI-34	maintainability		
TDI-35	maintainability	too big	
TDI-36	maintainability		

Furthermore, interesting observations were made while working on the improvement roadmap (prioritization of TDIs). Even though the top most items were addressed (TDI1-TDI5, TDI7-TDI8), there were a few TDIs addressed which were prioritized significantly lower on the list. The reason for that was continuous refactoring effort, a practice that is well aligned with agile development pace. Once a development context is open, some TDIs may be addressed with minimal cost while adding new functionality. However, bigger changes (e.g. some significant architectural improvements) could not be addressed this way, as the major focus of the team, within a sprint lasting two weeks, is to deliver explicitly planned scope (sprint goals).

Another observation is the fact that some of technical debt items will not be addressed at all, as their sheer size is too big to justify investments over new roadmap features awaiting implementation. Usually the larger size goes in pair with unclear scope of work. An example here could be refactoring of all regression tests to use a new simulator or be aligned with a different test framework. In such a situation, the solution is to decompose it into smaller TDIs and propose alignment of such effort with ongoing product development (new features).

Additional early observation (already included in Tab. 2) was the broadened scope of technical debt reduction activities to embrace also CI/CD perspective. If these improvements were not included in the scope, the engineering team would have (or at least should have) tried to prioritize them over the ones explicitly listed anyway. This is because the faster feedback from development and greater knowledge on the product usage (for instance, due to telemetry) cause the better codebase quality, and finally the product.

Tab. 3. Stabilization cycles required to release a new version

Product Version	*Stabilization cycles*
v1	6
v2	2
v3	2
v4	2
v5	1
v6	1

Finally, we observe that the investments made for this product started paying off and resulted in less error-prone code and faster time-to-market (Tab. 3).

5. CONCLUSIONS AND FUTURE WORK

Techniques of prioritizing technical debt as well as understanding the dynamics of its reduction processes allow for rational spending of limited funds and reasonable balancing of efforts between the software product development and optimization of the current codebase. Adopting the CoDVA approach for technical debt prioritization resulted in tangible benefits. The engineering team improved the product in a way that adding a new functionality is easier and less error-prone,

hence achieving faster time-to-market (less stabilization cycles were required to release a new product version).

Technical debt refactoring was following the plan in general (top-most improvements were applied); however, a few lower priority TDIs were refactored as well. Considering overall cost of investments this is a desired situation, as the engineering team uses the opportunity of ongoing development not only to add new functionality but also to improve the exact part of the code they are currently working on. The cost of adding a refactoring change in the same context as new development is usually significantly lower than opening a new context and introducing the change there, which makes the continuous refactoring approach useful.

Another factor which should be considered is sizing technical debt items. They cannot be significantly bigger than the effort required to develop a new feature, as they will be constantly postponed. The biggest size of a single a TDI that was implemented during our experiment was not surpassing the cost of a middle-size feature. Moreover, such TDI should have been explicitly planned and agreed upon with the Product Manager. Finally yet importantly, we also observed some minor dynamics of TDI priorities presented in Tab. 2 caused by the changing feature roadmap priorities. However, as the fluctuation was negligible, we refrained from explicitly showing this dynamics.

In order to achieve better predictability of the improvement roadmap we claim that the following rules should be applied: proper sizing of TDIs small enough to be considered for implementation, shifting focus towards more costly TDIs, limiting the number of explicitly allocated TDIs and embracing continuous refactoring approach for smaller TDIs. Finally, the CI/CD related refactorings should be treated with the same focus as sheer product codebase (they should be considered a part of a product by our definition), and prioritized accordingly.

As this is the second application of the CoDVA technique, we assume that this approach can be replicated, and the results may be generalized for telecommunication software. The future work will embrace finalization of this experiment, further analysis of TDI prioritization dynamics on more diverse set of TDIs, as in this case mostly generic improvements (affecting a higher amount of features) were implemented. Finally, the results of applying the CoDVA approach should be compared with other prioritization techniques.

REFERENCES

[1] Kruchten, Philippe. "The end of agile as we know it." In 2019 IEEE/ACM International Conference on Software and System Processes (ICSSP), pp. 104-104. IEEE, 2019.

[2] Kruchten, Philippe, Robert Nord, and Ipek Ozkaya. Managing technical debt: Reducing friction in software development. Addison-Wesley Professional, 2019.

[3] Stochel, Marek G., Piotr Chołda, and Mariusz R. Wawrowski. "Adopting DevOps Paradigm in Technical Debt Prioritization and Mitigation," submitted to SEAA2022.

[4] Stochel, Marek G., Piotr Chołda, and Mariusz R. Wawrowski. "Continuous Debt Valuation Approach (CoDVA) for Technical Debt Prioritization." In 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 362-366. IEEE, 2020.

[5] Stochel, Marek G., Piotr Chołda, and Mariusz R. Wawrowski. "On Coherence in Technical Debt Research: Awareness of the Risks Stemming from the Metaphorical Origin and Relevant Remediation Strategies." In 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 367-375. IEEE, 2020.

[6] Avgeriou, Paris, Philippe Kruchten, Ipek Ozkaya, and Carolyn Seaman. "Managing technical debt in software engineering (dagstuhl seminar 16162)." In Dagstuhl Reports, vol. 6, no. 4. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[7] Chen, Lianping, and Muhammad Ali Babar. "Towards an evidence-based understanding of emergence of architecture through continuous refactoring in agile software development." In 2014 IEEE/IFIP Conference on Software Architecture, pp. 195-204. IEEE, 2014.

[8] Vassallo, Carmine, Fabio Palomba, and Harald C. Gall. "Continuous refactoring in CI: A preliminary study on the perceived advantages and barriers." In 2018 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 564-568. IEEE, 2018.

[9] Lenarduzzi, Valentina, Terese Besker, Davide Taibi, Antonio Martini, and Francesca Arcelli Fontana. "Technical debt prioritization: State of the art. A systematic literature review." arXiv preprint arXiv:1904.12538 (2019).

[10] Pina, Diogo, Alfredo Goldman, and Graziela Tonin. "Technical Debt Prioritization: Taxonomy, Methods Results, and Practical Characteristics." In 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 206-213. IEEE, 2021.

[11] Kruchten, Philippe. "Voyage in the Agile Memeplex: In the world of agile development, context is key." Queue 5, no. 5 (2007): 38-44.

[12] Stochel, Marek G., Mariusz R. Wawrowski, and James J. Waskiel. "Adaptive agile performance modeling and testing." In 2012 IEEE 36th Annual Computer Software and Applications Conference Workshops, pp. 446-451. IEEE, 2012.

PORÓWNANIE STANDARDÓW OPENEHR ORAZ HL7 FHIR

COMPARISON OF OPENEHR AND HL7 FHIR STANDARDS

Jacek Kryszyn[1]; Waldemar T. Smolik[1]; Damian Wanta[1]; Mateusz Midura[1], Przemysław Wróblewski[1]

[1] Politechnika Warszawska, Instytut Radioelektroniki i Technik Multimedialnych, Warszawa, jacek.kryszyn@pw.edu.pl (JK), waldemar.smolik@pw.edu.pl (WTS), damian.wanta@pw.edu.pl (DW), mateusz.midura.dokt@pw.edu.pl (MM), przemyslaw.wroblewski@pwe.du.pl (PW)

DOI: 10.15199/59.2022.4.11

Streszczenie: **Systemy informatyczne w medycynie charakteryzują się potrzebą bezpiecznego przechowywania i przetwarzania dużych ilości wrażliwych danych medycznych przy jednoczesnym zapewnieniu możliwości współpracy z innymi systemami. Istnieją dwa konkurencyjne standardy medyczne, które na to pozwalają: openEHR oraz HL7 FHIR. W niniejszej pracy omówiono oba te standardy oraz porównano je ze sobą. Przedstawiona została architektura obu systemów, podobieństwa i różnice, sposoby modelowania danych oraz zapewnienia interoperacyjności.**

Abstract: **Information systems in medicine are characterized by the need to securely store and process large amounts of sensitive medical data while ensuring interoperability with other systems. There are two competing medical standards that allow this: openEHR and HL7 FHIR. In this paper, both of these standards are discussed and compared with each other. The architecture of both systems, the similarities and differences, methods of data modeling and ensuring interoperability were presented.**

Słowa kluczowe: **systemy informatyczne w medycynie, informatyzacja ochrony zdrowia, openEHR, HL7 FHIR**

Keywords: **medical computer systems, healthcare information technology, openEHRm HL7 FHIR**

1. WSTĘP

W ochronie zdrowia stosowanych jest wiele standardów informatycznych, w tym standardów dedykowanych dla medycyny. Niektóre z nich definiują ogólne koncepcje systemów medycznych, a niektóre koncentrują się na wybranym aspekcie. Ogólną koncepcję zawierają standardy takie jak ISO/TS 18308 opisujący wymagania architektury elektronicznego rekordu pacjenta (electronic healthcare record – EHR), CEN/TC EN 13940-1 opisujący koncepcję systemu ochrony zdrowia, czy też standard HL7 RIM, który opisuje referencyjny model informacji. LOINC, SNOMED CT i ICD-10 systematyzują terminologię medyczną. DICOM definiuje format obrazowych danych medycznych i komunikację obrazowych systemów medycznych i bazodanowych. Obiektowy, otwarty i rozszerzalny standard DICOM jest przykładem wzorcowym, chociaż jego wadą jest koncentracja na danych obrazowych. Wiele standardów informatyki medycznej jest niekompletnych w tym sensie, że opisuje tylko jeden aspekt, na przykład format przekazywanych danych, bez definiowania protokołu ich przekazywania. Przykładem takiego standard jest HL7 Messaging, zarówno w wersji v2.x jak i v3. W przypadku tego standardu struktura obiektu informacyjnego jest zaczerpywana z innych standardów: HL7 RIM (Reference Information Model) i HL7 CDA (Clinical Document Architecture). Co więcej, HL7 v3 nie narzuca kodowania danych, lecz proponuje domyślne kodowanie w XML-u. HL7 EHR-S FM opisuje model funkcjonalny systemów medycznych. CEN/ISO EN 13606 opisuje komunikację systemów w kontekście zachowania oryginalnego znaczenia klinicznego, jak i zachowania poufności danych szczególnie wrażliwych. Większość standardów wykorzystywanych w informatyce medycznej, takich jak amerykański CCR (Continuity of Care Record), HL7 CDA czy ISO 13606, przyjmuje podejście skoncentrowane na dokumentach (np. raport, dokumentacja pacjenta, wypis, itp.). Modelowana jest złożona struktura dokumentów klinicznych z danymi medycznymi i opisywany jest format, w jakim dane są wymieniane (np. XML). Standardy te nie opisują protokołu komunikacyjnego ani nie koncentrują się na przekazywaniu atomowych jednostek danych dotyczących zdrowia (takich jak: tętno, ciśnienia krwi, glukoza, itp.). Chociaż HL7 CDA jest przeznaczony do budowania złożonych dokumentów klinicznych, może być używany do przekazywania poszczególnych elementów danych, ale z założenia standard ten jest bardzo złożony i sam w sobie, bez określenia protokołu komunikacyjnego, nie zapewnia współdziałania systemów informatycznych.

Standard HL7 FHIR i standard openEHR, które są przedmiotem niniejszej pracy, są stosunkowo nowe, a ich historia zaczyna się odpowiednio w 2012 i w 2006 roku. Akronimy standardów czyta się odpowiednio jak angielskie wyrazy „fire" i „open air". Standardy te budzą duże zainteresowanie i cieszą się popularnością, co wynika z większej kompletności specyfikacji. W obu przypadkach standard obejmuje zarówno format jak i protokół komunikacji. Oba standardy wykorzystują istniejące i powszechnie wykorzystywane standardy informatyczne. Protokół warstwy aplikacji opiera się na warstwie protokołu HTTP. Można zamiennie stosować kodowanie danych w formacie XML i JSON. Zakłada się bezstanową architekturę oprogramowania, która jest obecnie bardzo popularna w biznesowych rozwiązaniach internetowych. Obecnie dostępna jest duża liczba implementacji otwartoźródłowych, które umożliwiają testowanie lokalne i zdalne. Istnieje również duża społeczność wspierających się wzajemnie programistów.

Standardy FHIR i openEHR, choć zaproponowane przez różne organizacje użyteczności publicznej, są bardzo podobne w koncepcji i wielu aspektach, ale można w

nich znaleźć zasadnicze różnice. Celem niniejszej pracy jest analiza porównawcza obu standardów. W porównaniu uwzględnione będą takie aspekty jak ogólna koncepcja i zapewnienie kompatybilności systemów, model danych i protokół komunikacyjny poziomu aplikacji. Omówiona zostanie elastyczność, obejmująca między innymi rodzaj obsługiwanych systemów bazodanowych, składnię i złożoność zapytań oraz obsługiwane formaty kodowania danych. Oceniona zostanie możliwość rozszerzania standardu, obejmująca dodawanie nowych elementów danych i terminologii. Przedstawione zostaną istniejące biblioteki i oprogramowanie w wersji otwartoźródłowej.

2. OGÓLNA CHARAKTERYSTYKA STANDARDÓW

2.1. HL7 FHIR

Standard HL7 FHIR (Fast Healthcare Interoperability Resources) określa format i protokół wymiany informacji medycznych między różnymi systemami komputerowymi, niezależnie od sposobu ich przechowywania w tych systemach. Przez format należy rozumieć wiele formatów, które są dopuszczone przez organizację opracowującą ten standard, czyli HL7 (Health Level Seven). FHIR opiera się na standardach internetowych takich jak HTTP, XML, JSON i OAuth, co niewątpliwie ułatwia implementację oprogramowania. FHIR zakłada, że komunikacja za pomocą protokołu HTTP będzie odbywała się według założeń stylu architektury oprogramowania REST (REpresentational State Transfer), w którym protokół komunikacyjny jest bezstanowy [1]. W takiej architekturze odbiorca nie zachowuje stanu sesji. Nadawca wysyła za każdym razem kompletne dane, które muszą być zrozumiałe w izolacji od poprzednich lub następnych komunikatów. Architektura bezstanowa ułatwia skalowalność i rozwój oprogramowania, jednak może wpływać na wydajność systemu ze względu na każdorazowe przesyłanie kompletnego stanu sesji. W standardzie FHIR przesyłane są obiekty informacji, nazywane zasobami (Resources), które bazują na innych standardach HL7, takich jak HL7 CDA, co pozwala na łatwiejszą integrację z istniejącymi systemami używającymi struktury danych zdefiniowanej w CDA (Clinical Document Architecture). Zasób opisywany jest przez model informacji, który definiuje elementy danych, ograniczenia i relacje pomiędzy obiektami, takimi jak pacjent, procedura, spotkanie, obserwacja itp. Obecnie istnieje 151 zdefiniowanych zasobów, a ponad 50 kolejnych jest „proponowanych". Zasoby mają pięć cech: adres URL, który identyfikuje zasób (na przykład url/patient/2345786), metadane identyfikacyjne, czytelne dla człowieka podsumowanie XHTML, zestaw zdefiniowanych elementów danych – inny zestaw dla każdego typu zasobu oraz rozszerzalny blok do przyszłego zdefiniowania. Każdy typ zasobów zdefiniowany w FHIR ma menedżera, do którego odwołujemy się poprzez URL [Serwer FHIR]/[typ], gdzie [typ] jest nazwą typu zasobu. Elementy danych zasobów zawierają relacje do innych elementów i zasobów. Interfejs programowania aplikacji (ang. Application programming interface, API) opisuje protokół FHIR jako zestaw operacji, znanych jako „Interactions", na zasobach. Interakcje są wykonywane bezpośrednio na zasobach serwera przy użyciu żądań i odpowiedzi protokołu HTTP. Podstawowe interakcje zebrane są w Tabeli 1. Bardziej zaawansowane interakcje to Batch/Transaction, Operation, VRead, Patch i Capabilities. Interakcje są podzielone na typy: Instancja, Typ i System. API opisuje również końcówki systemowe (ang. endpoints) służące do walidacji i dokumentacji.

Tab. 1. Interakcje FHIR

Nazwa	
Create	utwórz nowy zasób z przypisanym identyfikatorem serwera = POST url/{resourceType}
Read	odczytanie aktualnego stanu zasobu = GET url/{resourceType}/{id}
Update	zaktualizuj istniejący zasób według jego identyfikatora (lub utwórz go, jeśli nowy) = PUT url/{resourceType}/{id}
Delete	usuń zasób = DELETE url/{resourceType}/{id}
Search	wyszukaj zasoby według kryteriów filtrowania = GET url/{resourceType}?search parameters
History	pobierz historię zmian dla określonego zasobu = GET url/{resourceType}/{id}/history

W zależności od serwera, dostępne mogą być różne zestawy interakcji, jak również różne zestawy obsługiwanych zasobów. Serwery muszą obsługiwać zasób oświadczenia o możliwościach (Capability Statement), który opisuje, które interakcje i zasoby są dostępne. Specyfikacja standardu przewiduje rozszerzanie możliwości zasobów poprzez rozszerzenia i profilowanie (Extensions i Profiles). Zasoby mogą być przekazywane wymiennie w formacie XML lub JSON. Standard uwzględnia autoryzację dostępu i szyfrowanie danych.

2.2. openEHR

Standard openEHR, opracowany przez openEHR Foundation, opisuje zarządzanie, przechowywanie, wyszukiwanie i wymianę danych w systemach elektronicznej dokumentacji medycznej (EHR) [2][6]. openEHR posługuje się dwupoziomowym modelem danych, w którym informacja (poziom 1) oddzielona jest od wiedzy (poziom 2). Pierwszy poziom, definiowany za pomocą języka UML, zawiera model referencyjny z niewielkim zestawem klas, które opisują sposób reprezentacji danych w rekordzie pacjenta. Zawiera typy danych, struktury danych, identyfikatory i wzorce projektowe. Ogólna struktura karty zdrowia pacjenta opisywana jest przez klasy: obserwacja, ocena, instrukcja, akcja, wpis administratora. Inne klasy to na przykład: kontekst medyczno-prawny i szczegóły audytu. Poziom 1 obejmuje również takie elementy, jak kompozycja (Composition), sekcja (Sections), obserwacja (Entry) i różne typy danych, takie jak ilość (Quantity) i zakodowany tekst (Coded text). Drugi poziom, definiowany za pomocą języka ADL (Archetype Definition Language), dostarcza formalnych definicji pojęć klinicznych w postaci archetypów i szablonów. Archetypy zawierają metadane, zestaw nazw, reguł i ograniczeń opisujących, jak używać bloków modelu referencyjnego do tworzenia struktur danych przypominających drzewa. Archetypy to maksymalne zbiory danych zawierające wszystkie elementy danych wymagane przez wszystkie

przypadki użycia. Archetypy określają, które punkty danych są obowiązkowe, a które opcjonalne, ich typ i kryteria walidacji. Archetypy są projektowane przez specjalistów dziedzinowych przy użyciu Modelu Archetypów, zwykle w międzynarodowych środowiskach współpracy (w tym na platformach dyskusyjnych/przeglądowych). Szablony służą do łączenia kilku archetypów w większą strukturę, na przykład w kompozycję, przeznaczoną do konkretnego przypadku użycia, na przykład jako podstawa wypisu ze szpitala. Szablon może również ograniczać, ukrywać lub ustawiać wartości domyślne w archetypach i modelu odniesienia, na którym się opiera. Szablony nie „dodają" nowych koncepcji klinicznych; używają i ograniczają koncepcje zdefiniowane przez istniejące archetypy. openEHR definiuje komunikację w warstwie aplikacji z wykorzystaniem protokołu HTTP. Zakłada się bezstanową architekturę komunikacji REST, jednak możliwe jest zastosowanie innych rozwiązań takich jak SOAP lub Google Protocol Buffers. openEHR posługuje się pojęciem zasobu (Resource), który może być identyfikowany, adresowany, obsługiwany lub zarządzany poprzez adres URI/URL. Zasoby muszą być kodowane w formacie XML lub JSON. API podzielone jest na funkcje dotyczące rekordu pacjenta (EHR), zawierające zapytania kierowane do serwera (Query) i wprowadzające nowe definicje i modele na serwerze (Definitions). Przykładowe polecenia API zebrane są w Tabeli 2.

Tab. 2. Funkcje API

Polecenie	*Znaczenie*
POST EHR	Utwórz rekord EHR POST url/ehr
POST EHR	Utwórz rekord z zewnętrznym ID
GET EHR	Odczytaj rekord o numerze ID GET url/ehr/{EHR ID}
GET EHR	Rekord o zewnętrznym ID
POST Template	Zarejestruj szablon
GET Template	Odczytaj pojedynczy szablon GET url/definition/template/adl1.4/
GET Template	Odczytaj listę szablonów
POST Composition	Zapisz kompozycję POST url/ehr/{EHR ID}/composition
PUT Composition	Zapisz kompozycję
GET Composition	Odczytaj kompozycję
GET Composition	Odczytaj wersję kompozycji
DELETE Composition	Logiczne usunięcie kompozycji
Execute Query	SELECT c/name/value FROM COMPOSITION c WHERE c/uid/value = $uid GET {baseUrl}/query/myQuery?uid=2340-sd424fdv87::url::1

3. ANALIZA PORÓWNAWCZA

3.1. Ogólna koncepcja i kompatybilność systemów

Oba standardy wspierają wymianę, a także trwałość danych dotyczących opieki zdrowotnej. W przypadku obu standardów głównym hasłem jest interoperacyjność systemów (interoperability). Demistyfikując ten nowy w języku angielskim termin, należy wyjaśnić, że jest on propagowany przez przedstawicieli średniego i małego biznesu. Ma on oznaczać kompatybilność systemu z powszechnym standardem, w przeciwieństwie do kompatybilności z własnościowym formatem narzuconym przez dużą firmę. Zarówno HL7 FHIR jak i openEHR, do pewnego stopnia, umożliwiają realizację kompatybilnych w sensie standardu systemów obsługujących elektroniczny rekord pacjenta. Wydaje się, że openEHR lepiej opisuje problem konwersji na inne standardy. openEHR jest ściśle hierarchiczny i wdraża dwupoziomowe podejście do modelowania (model referencyjny i model archetypowy) dla swoich struktur danych i koncepcji. FHIR przyjmuje bardziej płaską reprezentację pojęć, chociaż odnosi się do hierarchicznego modelu zawartego w HL7 CDA i pozwala na tworzenie pakietów (Bundle) odpowiadających kompozycjom w openEHR (Composition). Sposób kodowania i komunikacji jest podobny, choć openEHR zawiera rozszerzenia. openEHR w przeciwieństwie do FHIR wykorzystuje dziedziczenie do realizacji specjalizacji danych i kompozycji danych złożonych.

3.2. Elastyczność i rozszerzalność standardu

Niezależnie od standardu możliwa jest obsługa dowolnego systemu bazodanowego, relacyjnego jak i nieSQL-owego. Obie specyfikacje obsługują różne typy danych prostych i złożonych. Zapytania AQL (Archetype Query Language) openEHR bardzo przypominają składnią SQL, ale nie jest łatwo zrealizować nimi zaawansowanego przetwarzania danych. To samo zapytanie zapisane za pomocą REST openEHR jest dość długie, złożone i nieczytelne. W standardzie FHIR zapytania są krótsze i prostsze, nawet w przypadku szczegółowych zapytań. Zarówno FHIR jak i openEHR używają do kodowania danych formatu XML i JSON, przy czym openEHR pozwala dodatkowo na stosowanie własnego języka (ADL/AOM). Kodowanie w formacie JSON wygląda podobnie dla obu standardów, jednak w przypadku openEHR kodowanie za pomocą XML-a prowadzi do bardzo obszernych plików tekstowych. Wynika to z faktu, że w openEHR atrybuty elementów danych są kodowane jak kolejne elementy danych. W FHIR wartości są przenoszone w atrybutach, a elementy danych nie zawierają wartości. Archetypy openEHR i zasoby FHIR są zbudowane z różnych komponentów, które mogą być użyte w innych kompozycjach i pakietach. Repozytoria archetypów i zasobów są systematycznie rozbudowywane (300 archetypów dla openEHR i 151 zasobów dla FHIR). W openEHR kompozycje, rozszerzenia i specjalizacje archetypów oparte są na mechanizmach bezpośredniego odniesienia, zagnieżdżania potomnych archetypów oraz wzorców. W FHIR złożone struktury są budowane za pomocą referencji do zasobów, zasobów zawartych i pakietów. Mechanizmy te pozwalają na rozbudowywanie standardu, a z drugiej strony zmniejszają koszty rozwoju oprogramowania.

W przypadku obu standardów wewnętrzny system terminologiczny nie jest wyczerpujący. Oba standardy obsługują szereg zewnętrznych systemów terminologicznych do kodowania nazw elementów, a czasem wartości danych. Dostępne są narzędzia programistyczne do dodawania terminologii przy budowaniu nowych archetypów i zasobów, na przykład Archetype Editor lub clinFHIR. Oba standardy pozwalają na zapis znaczenia danych klinicznych i wykorzystanie znaczenia w zapytaniach.

3.3. Istniejące narzędzia i oprogramowanie

Nie sposób w tak krótkim tekście omówić istniejące narzędzia i oprogramowanie, nawet te najważniejsze z nich. Podanych zostanie tylko kilka reprezentatywnych przykładów. W przypadku FHIR warto wspomnieć o bibliotekach HAPI i Firely .Net SDK. Biblioteka HAPI jest implementacją w języku Java o otwartym kodzie źródłowym (University of Toronto). Biblioteka ta definiuje klasy dla każdego zasobu FHIR. Biblioteka HAPI udostępnia przykład implementacji zarówno serwera zbudowanego na serwerze WWW jak i klienta FHIR. Biblioteka dostarcza również narzędzie do walidacji modelowanych zasobów FHIR w celu zapewnienia zgodności zasobów ze specyfikacją. Firely .Net SDK jest biblioteką dla języka C#. Zawiera między innymi: klasy opisujące modele danych, klasy pomocnicze do pracy z metadanymi, analizator składniowy języka XML i JSON, przykład klienta usług FHIR, narzędzie do walidacji danych na podstawie profili oraz narzędzie do weryfikacji zapytań. Firely, jako firma komercyjna, udostępnia bezpłatną edycję serwera FHIR do celów testowych i edukacyjnych. Udostępnia również publiczną końcówkę do testów dostępną w Internecie. Serwery FHIR, w tym otwartoźródłowe, udostępniają duże firmy takie jak IBM czy Microsoft. Podstawową biblioteką dla openEHR jest implementacja standardu w języku Java udostępniana przez openEHR Foundation. Po przekazaniu tej biblioteki fundacji przez zespół ze Szwecji, stała się otwartoźródłową biblioteką referencyjną [7]. Jest podzielona na kilka komponentów oprogramowania. openehr-rm implementuje wszystkie specyfikacje modelu informacji referencyjnych i zapytania. rm-builder służy do konstruowania obiektów za pomocą archetypów. Komponenty openehr-aom i openehr-ap zapewniają reprezentację archetypów, tworzenie obiektów i obsługę walidacji. adl-parser implementuje język ADL i przekształca archetypy w formacie tekstowym na formę AOM (Archetype Object Model). Inne implementacje w języku Java to archie i EHRbase. Istnieją również implementacje dla innych języków: opEHR.NET dla języka C# [3], pyEHR dla języka Python, skoba dla języka Ruby. W testowaniu oprogramowania openEHR można wykorzystać otwartoźródłowe serwery, takie jak EHRServer, EHRbase czy OpenEyes [5]. Przy testowaniu oprogramowania można skorzystać z publicznie dostępnego zbioru danych do badania poprawności i wydajności serwerów ORBDA [4].

4. WNIOSKI I PODSUMOWANIE

Celem niniejszej pracy było przedstawienie i porównanie dwóch standardów przeznaczonych do wymiany danych w systemach ochrony zdrowia: HL7 FHIR i openEHR. Oba standardy stawiają sobie za cel umożliwienie wymiany danych medycznych między systemami różnych producentów. Choć każdy standard z osobna zapewnia realizację tego celu, to jednak patrząc z punktu użytkownika, kompatybilność systemów, jaką dostarczają, jest ograniczona istnieniem drugiego standardu. Z przeprowadzonej analizy wynika, że standardy choć różnią się modelem danych, to są bardzo podobne, a przede wszystkim mają pełnić tę samą funkcję. Nie było przedmiotem tej pracy ocena, który standard wyprze konkurencyjny. Wydaje się, że najlepszym wyjściem byłoby scalenie obu specyfikacji w jedną. Aby obiektywnie stwierdzić, czy któryś z tych dwóch standardów jest lepszy, można byłoby przeprowadzić testy wydajnościowe serwerów implementujących oba rozwiązania w celu stwierdzenia, czy nie ma między nimi znaczącej różnicy. Pozwoliłoby to ustalić, czy któraś z dwóch architektur ma przewagę. Jak pokazano w [5], zastosowanie openEHR w szpitalnym systemie informatycznym wiąże się z zauważalnym zmniejszeniem wydajności i przepustowości w stosunku do rozwiązania opartego o własnościowy model danych. Brak jest podobnych testów dla innych implementacji openEHR oraz HL7 FHIR. Większość prac naukowych skupia się na zastosowaniach obu standardów i nie rozstrzyga tych kwestii.

LITERATURA

[1] Sundvall Erik, Nyström Mikael, Karlsson Daniel, Eneling Martin, Chen Rong, Örman Håkan. 2013. „Applying representational state transfer (REST) architecture to archetype-based electronic health record systems". *BMC Medical Informatics and Decision Making*, 13.

[2] Min Lingtong, Tian Qi, Lu Xudong, Duan Huilong. 2018. „Modeling EHR with the openEHR approach: An exploratory study in China Philip Payne". *BMC Medical Informatics and Decision Making*, 18.

[3] Demski Hans, Garde Sebastian, Hildebrand Claudia. 2016. „Open data models for smart health interconnected applications: The example of openEHR Standards, technology, machine learning, and modeling". *BMC Medical Informatics and Decision Making*, 16.

[4] Teodoro Douglas, Sundvall Erik, Junior Mario João, Ruch Patrick, Freire Sergio Miranda. 2018. „ORBDA: An openEHR benchmark dataset for performance assessment of electronic health record servers". *PLOS One*, 13.

[5] Kryszyn Jacek, Cywoniuk Kamil, Smolik Waldemar T., Wanta Damian, Wróblewski Przemysław, Midura Mateusz. 2022. „Performance of an openEHR based hospital information system". *International Journal of Medical Informatics*, 162.

[6] Pahl Christina, Zare Mojtaba, Nilashi Mehrbakhsh, Borges Marco Aurélio de Faria, Weingaertner Daniel, Detschew Vesselin, Supriyanto Eko, Ibrahim Othman. 2015. „Role of OpenEHR as an open source solution for the regional modelling of patient data in obstetrics". *Journal of Biomedical Informatics*, 55: 174–187.

[7] Chen Rong, Klein Gunnar. 2007. „The openEHR java reference implementation project". *Studies in Health Technology and Informatics*, 129: 58–62.

ZWIĘKSZENIE BEZPIECZEŃSTWA POCZTY ELEKTRONICZNEJ – NARZĘDZIE WSPOMAGAJĄCE

ELECTRONIC MAIL SECURITY INCREASING - A SUPPORTING TOOL

Damian Wasilewski[1]; Marcin Kowalczyk[2]

[1] Instytut Telekomunikacji, Politechnika Warszawska, Warszawa, Daniel.Wasilewski01@pw.edu.pl
[2] Instytut Telekomunikacji, Politechnika Warszawska, Warszawa, Marcin.Kowalczyk@pw.edu.pl

DOI: 10.15199/59.2022.4.12

Streszczenie: **Artkuł prezentuje wyniki projektu mającego na celu zwiększenia bezpieczeństwa korzystania z poczty elektronicznej e-mail. Efektem końcowym prac jest narzędzie w formie zewnętrznego serwisu dokonujące automatycznej detekcji niebezpiecznych wiadomości w oparciu o zawartość nagłówka oraz odnośniki i powiązane z nimi strony internetowe, zawarte w treści wiadomości. Integralną częścią rozwiązania jest klasyfikator decyzyjny oparty o uczenie maszynowe oraz sztuczną inteligencję. Narzędzie integruje się z popularnymi klientami pocztowymi takimi jak MS Outlook za pomocą dedykowanych wtyczek.**

Abstract:**The paper presents the results of the project aimed at increasing the security of electronic mail. The final result of the work is a tool in a form of an external service that automatically detects dangerous e-mail messages based on the content of the headers as well as links and related with them websites included in the message content. An integral part of the solution is a decision classifier based on ma-chine learning and artificial intelligence approach. The developed tool integrates with popular e-mail clients such as MS Outlook using dedicated plugins.**

Słowa kluczowe: **detekcja zagrożeń, klasyfikacja poczty elektronicznej, uczenie maszynowe**

Keywords: **detection of threats, mail classification, machine learning**

1. WSTĘP

Usługa poczty elektronicznej e-mail ze względu na jej powszechność jest jedną z najczęściej wykorzystywanych usług w cyfrowym świecie. Szacuje się, że z usługi tej na chwilę obecną korzysta ponad 4 mld użytkowników, a liczba ta ciągle rośnie [1]. Usługa ta jest też ze względu na swój charakter jedną z najbardziej podatnych na różnego rodzaju cyberataki. Mowa tu zarówno o działaniach, których celem jest wyłudzenie (wykradzenie) poufnych informacji, czy też zaszkodzeniu użytkownikowi, np. na skutek doprowadzenia do instalacji niebezpiecznego oprogramowania, jak też rozsyłce „niechcianej" poczty w formie tzw. spamu, np. maili zawierających treść reklamowe, itp. Zgodnie z badaniami Delloite z roku 2020 aż 91% wszystkich cyberataków rozpoczyna się od prób wyłudzeń w postaci ataku phising, dokonywanych z udziałem poczty elektronicznej e-mail [2]. Czyni to też tę usługę nader pożądaną (trudno wyobrazić sobie współczesny świat bez niej), ale też wysoce niebezpieczną. Dlatego też w zasadzie od początków jej istnienia prowadzi się ciągłe prace mające podnieść poziom jej bezpieczeństwa. W tym m.in. nakierowane na możliwość detekcji i blokowania niebezpiecznej korespondencji przed jej otwarciem przez końcowego odbiorcę i odpowiednio wczesne, prewencyjne powiadamianie go o potencjalnym niebezpieczeństwie. W tym względzie mamy swoistego rodzaju nieustanny „wyścig zbrojeń". Atakujący bo-wiem również systematycznie udoskonalają techniki i metody ataku [3-6]. Powoduje to, że nowe rozwiązania w zakresie ochrony poczty e-mail muszą być coraz bardziej zaawansowane. Wiele z nich wykorzystuje możliwości adaptacyjnego uczenia, które są oferowane przez uczenie maszynowe ML (ang. *Machine Learning*) oraz sztuczną inteligencję AI (ang. *Artficial Intelligence*).

W oparciu o to podejście powstało również opracowane przez Autorów narzędzie w formie zewnętrznej usługi osadzonej w chmurze MS Azure. Narzędzie to wg. wiedzy autorów jest jednym z pierwszych, które w ocenie czy wiadomość jest bezpieczna, czy też niesie za sobą ryzyko zagrożeń, typu wyłudzenie danych, bazuje na treści wiadomości, w szczególności obecności w nich linków URL (ang. *Unified Resource Locators*) do zewnętrznych stron i ich zawartości HTML.

Narzędzie może integrować się z popularnymi klientami pocztowymi za pomocą dedykowanych wtyczek.

W aktualnej wersji wtyczka taka został przygotowana dla klienta pocztowego MS Outlook. Integralnym elementem opracowanego narzędzia jest klasyfikator wiadomości pocztowych, działający w oparciu o sieć neuronową głębokiego uczenia.

Dalsza część artykułu oprócz prezentacji samego narzędzia (rozdział 3) zawiera wyniki przeprowadzonych badań w obszarze wyboru odpowiednio skutecznego algorytmu ML i AI (rozdział 4), jak też wprowadzenie w obszar projektu, obejmujące przybliżenie zagrożeń i metod przeciwdziałania im w odniesieniu do korzystania z korespondencji elektronicznej (rozdział 2).

2. POCZTA E-MAIL – ZAGROŻENIA, METODY ZAPOBIEGANIA

Historia poczty elektronicznej e-mail, zapewniającej wymianę wiadomości tekstowych między użytkownikami, rozpoczyna się w latach 60-tych ubiegłego stulecia

(pierwsze studia konceptualne). Pierwsza fizyczna implementacja miała miejsce w roku 1971. Od tego czasu usługa ta przeszła długą drogę rozwoju, zyskując ogromną popularność.

2.1. Poczta e-mail - zagrożenia

Wciąż pomimo dużej liczby rozwiązań wdrażanych w obszarze ochrony poczty elektronicznej wielu jej użytkowników staje się ofiarą różnego rodzaju ataków, czy też nadużyć, dokonywanych przy jej udziale. Mowa tu zarówno o uciążliwych, ale też w większości przypadków nieszkodliwej, dystrybucji i otrzymywania różnego rodzaju niechcianych wiadomości w formie tzw. spamu, np. treści reklamowych, jak też niosących za sobą po-ważne zagrożenia ataków typu phising. Te ostatnie nakierowane są na wyłudzanie poufnych wiadomości, czy też zmuszenia użytkowników do instalacji niebezpiecznego oprogramowania. Przykładem czego mogą być np. rejestratory (ang. keylogers), służące do pozyskiwania np. danych logowania poprzez zapamiętywanie ciągów znaków wprowadzanych z klawiatury. Nasz komputer może też np. zostać przekształcony w koparkę kryptowalut lub też stać się jednym z wielu uczestniczących w prowadzaniu masowych ataków sieciowych. Wszystko to oczywiście bez naszej wiedzy i zgody. Mimo ciągle wzrastającej w tym względzie wśród użytkowników poczty elektronicznej e-mail świadomości możliwych zagrożeń, ostateczna liczba popełnianych przestępstw na tym polu nie maleje, przy jednocześnie wzrastającej liczbie tego typu prób. Podczas ataków wyłudzających typu phising atakujący często wykorzystują takie metody jak: fałszywe linki (adresy URL, zaszyte w treści wiadomości) prowadzące do niebezpiecznych/ nieautoryzowanych stron czy serwisów www, zainfekowane załączniki lub też bazują na meto-dach socjotechnicznych. Przykłady podejmowanych i możliwych prób ataków w oparciu o te mechanizmy można mnożyć. Dlatego też tak ważne jest ciągłe opracowywanie nowych rozwiązań, jak też udoskonalanie obecnych, w zakresie podnoszenia bezpieczeństwa poczty elektronicznej e-mail.

2.2. Bezpieczeństwo korespondencji elektronicznej – metody zapobiegania

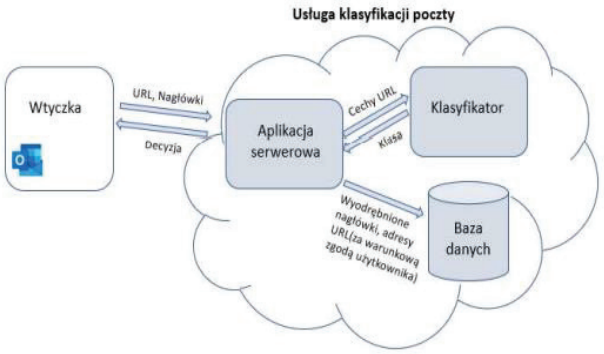
W czasach powstawania pierwszych standardów dla przesyłanie korespondencji elektronicznej niewiele myślano o jej bezpieczeństwie. Wzrost jej popularności oraz duża podatność na różnego rodzaju ataki podejmowane z jej udziałem wymusiły zmiany w tym zakresie. Na przestrzeni lat powstało wiele metod i standardów rozszerzających podstawową implementację rozwiązań dla poczty elektronicznej e-mail. Dotyczyły one wielu problemów, takich jak weryfikacja tożsamości użytkowników, czy też filtrowanie niechcianych lub też niebezpiecznych z punktu widzenia odbiorcy wiadomości. Na chwilę obecną powszechnie stosowane są takie protokoły rozszerzające jak SFP (ang. *Sender Policy Framework*), DKIM (ang. *DomainKeys Identified Mail*), oraz DMARC (ang. *Domain-based Message Authentication, Reporting and Conformance*) [7]. Dostawcy usług pocztowych instalują także na swojej infrastrukturze wiele dodatkowych rozwiązań służących wykrywaniu i filtrowaniu wiadomości szkodliwych typu *spam*, oraz *phising*.

3. OPRACOWANE NARZĘDZIE

Filtrowanie, a następnie blokowanie niebezpiecznej korespondencji wyłącznie na podstawie danych zawartych w różnego rodzaju nagłówkach analizowanej wiadomościu bez odniesienia do rzeczywistej ich treści bywa nieskuteczne. Z tego też powodu w podjętych pracach, które miały na celu opracowanie narzędzia zwiększającego poziom bezpieczeństwa korzystania z poczty elektronicznej e-mail, skupiono uwagę na możliwości zapewnienia wczesnego wykrywania niebezpiecznej korespondencji w oparciu o jej treść, ze szczególnym uwzględnieniem obecnych w niej adresów URL (hiperłączy), oraz kryjących się za nimi stronami/serwisami www (analiza ich struktury).

3.1 Założenia funkcjonalne – architektura narzędzia

Zbudowane narzędzie ma postać usługi zewnętrznej, która jest dostępna z poziomu chmury MS Azure. Nie ma jednak przeszkód, aby rozwiązanie to było dostępne z poziomu dowolnie innej, lub też funkcjonowało jako rozwiązanie typu on-promise. Funcjonalny schemat blokowy narzędzia został przedstawiony na Rys. 1.



Rys. 1. Architektura funkcjonalna narzędzia

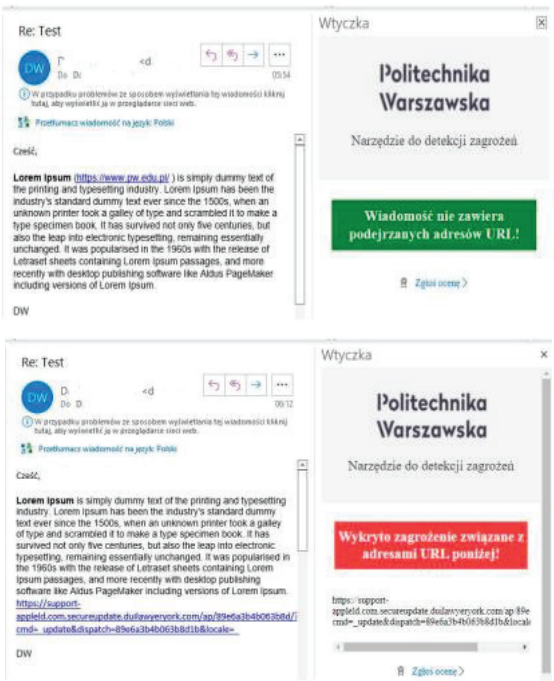
Podstawą rozwiązania jest **aplikacja serwerowa**. Zapewnia ona m.in. możliwość obsługi zgłoszeń użytkowników z prośbą o weryfikację wiadomości przekazywanych z poziomu końcowego klienta pocztowego. Aplikacja ta dokonuje analizuje treść wiadomości mail ze szczególnym uwzględnieniem obecności hiperłączy URL do zewnętrznych stron (serwisów), które potencjalnie mogą być niebezpieczne. Dla tych ostatnich określany jest wektor metadanych w postaci zestawu blisko **80 cech** w podziale na trzy grupy (zestawy), tj.: (1) *zestaw cech wyodrębnionych z każdego adresu URL zawartego w wiadomosci e-mail*, (2) *cechy drzewa DOM z dokumentów HTML związanych z wyodrębionymi adresami URL*, (3) *zestaw cech oceny domeny będącej elementem wyodrębnionego adresu URL*. Wyodrębniony w ten sposób wektor cech jest przekazywany jako informacja wejściowa do **klasyfiktora**. Na jego podstawie podejmuje on decyzję o tym czy w jego ocenie wiadomość niesie za sobą ryzyko, czy też jest bezpieczna.

Klasyfikator pełniący rolę decyzyjną w opracowanym narzedziu został oparty w aktualnej wersji

o sztuczną sieć neuronową głębokiego uczenia, wytypowaną w oparciu o przeprowadzone badania (rozdział 4-ty).

Ostatnim elementem po stronie zewnetrznego serwisu jest baza danych, w której za zgodą użytkownika końcowego istnieje możliwość zapamiętywania wiadomości i powiązanych z nią metadanych wyodrębnionych przez aplikację serwerową. Dane te w powiązaniu z informacją zwrotną udzielaną przez użytkowników o poprawności dokonanej oceny mogą posłużyc w przyszłości do ponownego wytrenowania modelu klasyfikatora.

Integralną częścią rozwiązania jest też **wtyczka** dla klienta pocztowego. W aktualnej wersji jest ona dostępna dla programu MS Outlook. Zapewnia możliwość korzystania ze zbudowanego narzędzia. Pozwala m.in. na: pozyskiwanie oceny korespondencji elektronicznej, prezentacje wyników oceny bezpieczeństwa, zgłaszanie błędnej oceny poprawności dokonanej klasyfikacji, oraz przesyłanie logu zdarzeń na serwer w przypadku wystąpienia problemów z oprogramowaniem. Rozpoczęcie pracy z wtyczką polega na zaimportowaniu przygotowanej konfiguracji w postraci pliku *xml*. Wtyczka pozwala na zweryfikowanie zagrożenia związanego z daną korespondencją poprzez wykonanie akcji *Zweryfikuj wiadomość*. W efekcie czego kopia wiadomości jest przesyłana do usługi, gdzie odbywa się jej weryfikacja. W zależności od wyniku użytkownik dostaje komunikat pozytywnej, bądź też negatywnej oceny, Rys. 2. W przypadku tej ostatniej wskazywane są również podejrzane adresy URL.



Rys. 2. Przykłady pozytywnej, jak też negatywnej oceny wiadomości e-mail pod względem zagrożeń

Użytkownikowi przysługuje możliwość zakwestionowania dokonanej przez narzędzie oceny, w przypadku gdy jego zdaniem odnośnik jest bezpieczny.

4. OPRACOWANIE KLASYFIKATORA – BADANIA

Ze względu na potrzebę realizacji wysoce generycznego narzędzia zdecydowano się wykorzystać przy opracowaniu klasyfikatora podejście oparte o uczenie maszynowe i sztuczną inteligencję. Wymagało to przeprowadzenia dodatkowych badań mających na celu wskazanie metody zapewniającej odpowiednio dużą skuteczność. W tym celu badaniom na wspólnym zbiorze uczącym poddano modele klasyfikatorów, które zostały opracowane na bazie różnych metod ML i AI. Na tym etapie wykorzystany został zbiór *Web page phishing detection* z serwisu Mendeley Data [8]. Zawierał on adresy URL oraz struktury drzewa DOM dokumentów HTML, do których zgromadzone URL prowadziły lub nadal prowadzą. Zbiór ten zawierał dane źródłowe, ale również wyodrębnione na ich podstawie cechy (metadane). W przeprowadzonych testach wykorzystano zbiór danych liczący 11430 rekordów. Badaniom zostały poddane opracowane modele klasyfikatorów na bazie takich metod ML i AI jak: regresja logiczna, k-najbliższych sąsiadów (kNN), naiwny klasyfikator Bayesa, maszyna wektorów nośnych SVM (ang. *Suport Vector Machine*), las decyzyjny, oraz głęboka sieć neuronowa. Topologia tej ostatniej została zobrazowana na Rys. 3.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 70)]	0
dense (Dense)	(None, 512)	36352
dense_1 (Dense)	(None, 512)	262656
dropout (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 128)	65664
dense_3 (Dense)	(None, 128)	16512
dropout_1 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 64)	8256
dense_5 (Dense)	(None, 1)	65

Total params: 389,505
Trainable params: 389,505
Non-trainable params: 0

Rys. 3. Topologia zaprojektowanej sieci neuronowej

Badania zostały przeprowadzone dla zbalansowenego (wyniki w Tabela 1), jak też niezbalansowanego (Tabela 2) zbioru uczącego. Wykonanie tych ostatnich było szczególnie istotne ze względu na domniemanie, że w rzeczywistości liczebność odbieranych wiadomości niebezpiecznych w stosunku do wszystkich jest mimo wszystko stosunkowo mała. Do oceny jakości (wiarygodności) poszczególnych modeli klasyfikatorów zastosowano metryki zdefiniowane w oparciu o tzw. macierz pomyłek, znanej też jako macierz konfuzji (ang. *confusion matrix*). Szczegółowy opis ich znaczenia zamieszczono w Tabeli 3.

Tab. 1. Metryki oceny rozpatrywanych algorytmów dla zbioru zbalansowanego

Metoda	*Dokładność*	*Precyzja*	*Czułość*	*F1-score*
Regresja logiczna	*0,9366*	*0,9186*	*0,9589*	*0,9383*
kNN	*0,9370*	*0,9472*	*0,9261*	*0,9366*
Naiwny klasyfikator Bayesa	*0,8827*	*0,8929*	*0,8711*	*0,8819*
SVM	*0,9594*	*0,9564*	*0,9630*	*0,9597*
Las decyzyjny	*0,9408*	*0,9181*	*0,9686*	*0,9427*
Głęboka sieć neuronowa	*0,9583*	*0,9507*	*0,9672*	*0,9589*

Tab. 2. Metryki oceny rozpatrywanych algorytmów dla zbioru niezbalansowanego

Metoda	*Dokładność*	*Precyzja*	*Czułość*	*F1-score*	*G-mean*
Regresja logiczna	*0,9043*	*0,6475*	*0,9658*	*0,7753*	*0,9280*
kNN	*0,9521*	*0,8307*	*0,9044*	*0,8660*	*0,9327*
Naiwny klasyfikator Bayesa	*0,8909*	*0,6305*	*0,8737*	*0,7324*	*0,8840*
SVM	*0,9294*	*0,7139*	*0,9795*	*0,8258*	*0,9488*
Las decyzyjny	*0,8565*	*0,5447*	*0,9761*	*0,6992*	*0,9011*
Głęboka sieć neuronowa	*0,9368*	*0,7400*	*0,9653*	*0,8378*	*0,9840*

Tab. 3. Metryki – opis znaczenia

Metryka	*Opis*
Dokładność	*mówi o tym, ile elementów z całego zbioru zostało poprawnie rozpoznanych*
Precyzja	*parametr definiujący precyzję rozpoznawania klasy mniejszościowej*
Czułość	*parametr mówiący o tym jak dobrze algorytm rozpoznaje obiekty klasy większościowej*
F1-score	*średnia harmoniczna dla precyzji i czułości*
G-mean	*średnia geometryczna metryk czułości oraz precyzji, czym większa jest wartość parametru tym lepsza skuteczność algorytmu.*

Uzyskane wyniki wyrażone wartością poszczególnych metryk wskazują na największą użyteczność rozwiązań opartych o kNN, jak też głębokiej sieci neuronowej. W szczególności gdy mowa o zbiorze niezbalansowanym. Zadawalające metryki uzyskano również dla modelu klasyfikatora SVM. O ostatecznym wyborze kasyfikatora na bazie sieci neuronowej głębokiego uczenia zadecydowały w szczególności potencjalne korzyści mareketingowe. Nie oznacza to, że w kolejnych odsłonach narzędzia nie będzie ono rozwijane w kierunku łączenia kilku metod ML i AI jako podstawy dla wykorzystywanego klasyfikatora. Dzięki czemu spodziewana jest dalsza poprawa jego wiarygodności.

5. PODSUMOWANIE

Popularność usługi poczty elektronicznej e-mail powoduje, że jest ona celem wielu ataków, które przybierają różne formy, w tym również ataków typu *phising*. Systematycznie udoskonalane techniki ich prowadzenia wymuszają z jednej strony nieustanną pracę nad poprawą świadomości zagrożeń i technik ich unikania wśród użytkowników (najsłabsze ogniwo), z drugiej zaś opracowywanie coraz skuteczniejszych narzędzi przeciwdziałania im. Mowa tu m.in. o skutecznym filtrowaniu i blokowaniu wiadomości niebezpiecznych zanim dotrą one do odbiorcy końcowego. W opracowywaniu tych ostatnich szczególnie użytecznym może być uczenie maszynowe oraz sztuczna inteligencja. Narzędzia na ich bazie są w stanie adaptacyjnie uczyć się nowych wzorców nadużyć w oparciu o gromadzone na bieżąco zbiory danych (wiedze), bez konieczności wprowadzania rozległych zmian programistycznych.

Zaproponowane narzędzie w postaci zewnętrznej usługi dostępnej w chmurze MS Azure, integrujące się z popularnymi klientami pocztowymi, takimi jak MS Outlook, jest tego dobrym przykładem. Przy jego implementacji wykorzystane zostały wnioski z przeprowadzonych badań w zakresie przydatności różnego rodzaju algorytmów ML i AI dla skutecznej klasyfikacji poczty. Przeprowadzone badania wykazały szczególnie dużą przydatność na tym polu metody kNN oraz sztucznej sieci neuronowej głębokiego uczenia. Wybór tej ostatniej na etapie końcowej implementacji rozwiązania miał w dużej mierze charakter uznaniowy. Zadecydował jej większy potencjał marketingowy. Autorzy zaznaczają jednak, że nie ma żadnych przeszkód technicznych, aby na dalszych etapach rozwoju projektu wprowadzić również dodatkowe warstwy oceny poczty przychodzącej w oparciu też o inne algorytmy, co powinno podnieść w ostatecznym rozrachunku wiarygodność zaproponowanego rozwiązania.

W ocenie Autorów wykorzystanie ML i AI w połączeniu z możliwościami chmur obliczeniowych, takich jak MS Azure, jest drogą do opracowywania i udostępniania na szeroką skalę coraz bardziej zaawansowanych, ale też stosunkowo tanich, narzędzi w obszarze ochrony korespondencji elektronicznej e-mail, czego przykładem jest prezentowane w niniejszym artykule narzędzie.

LITERATURA

[1] https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/

[2] https://www2.deloitte.com/my/en/pages/risk/articles/91-percent-of-all-cyber-attacks-begin-with-a-phishing-email-to-an-unexpected-victim.html

[3] Awad W.A., Elseuofi S.M. 2012. "Machine Learning Methods for Spam E-mail Classification", *Int. Jour. of Computer Science and Information Techn.*, 3(1), 173-184

[4] Rot A, Olszewski B. 2017. „Advanced Persistent Threats Attacks in Cyberspace. Threats, Vulnerabilities, Methods of Protection", *Comp. Science and Information Systems*, 113–117.

[5] Valecha R., Mandakar P., Rado H.R. 2022. „Phishing Email Detection Using Persuasion Cues". IEEE Transactions on Dependable and Secure Computing, (19), 747-756

[6] Salloum S., Gaber T., Vadera S., Shaalan K. 2021. „Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey", *Procedia Computer Science,* (189), 19-28

[7] Nightingale S. 2017, „Email Authentication Mechanisms: DMARC, SPF and DKIM", *National Institute of Standards and Technology*

[8] Hannousse A., Yahiouche S. 2021, "Web page phishing detection", Mendeley Data, V3, doi: 10.17632/c2gw7fy2j4.3

PROJEKT ARCHITEKTURY SYSTEMU WSPOMAGANIA DECYZJI O EWAKUACJI MEDYCZNEJ*

THE ARCHITECTURE CONCEPT FOR A MEDICAL EVACUATION DECISION SUPPORT SYSTEM

Piotr Łubkowski[1]

[1] Wojskowa Akademia Techniczna, Warszawa, piotr.lubkowski@wat.edu.pl

DOI: 10.15199/59.2022.4.13

Streszczenie: **Współczesne rozwiązania wykorzystywane w środowisku wojskowym oparte są o model architektury korporacyjnej rozumianej jako zbiór właściwości danej organizacji. Ramy architektury korporacyjnej proponowanej do zastosowania w omawianym środowisku wyznacza architektura korporacyjna NATO Architecture Framework w wersji 4. Celem NAFv4 jest zapewnienie standardu opracowywania i opisywania architektury korporacyjnej na potrzeby wojskowe. W artykule zaprezentowano praktyczną implementację NAFv4 w opracowaniu architektury systemu wspomagania decyzji o ewakuacji medycznej, którego ideą jest monitorowaniu parametrów życiowych żołnierza oraz wykorzystanie rozszerzonej świadomości sytuacyjnej do wsparcia procesu decyzyjnego w zakresie skutecznej ewakuacji poszkodowanych nie tylko na polu walki.**

Abstract: **Contemporary solutions used in military environment are based on Enterprise Architecture model understood as a set of properties of a given organization. The framework of enterprise architecture proposed for use in this environment is specified by NATO Architecture Framework version 4. The purpose of NAFv4 is to provide a standard for developing and describing enterprise architecture for military purposes. The paper presents practical implementation of NAFv4 in development of decision support system architecture for medical evacuation, which idea is to monitor vital signs of a soldier and use extended situational awareness to support decision making process in effective evacuation of casualties not only on the battlefield.**

Słowa kluczowe: **architektura korporacyjna, ewakuacja medyczna, wspomaganie decyzji, taksonomia zdolności.**

Keywords: **enterprise architecture, decision support, medical evacuation, taxonomy of capabilities.**

1. WSTĘP

1.1. Geneza problemu

Opracowanie i wykonanie teleinformatycznego systemu wspomagania decyzji ewakuacji medycznej (*SWDEwMED*), opartego na integracji monitoringu parametrów życiowych żołnierza z systemem zabezpieczenia medycznego oraz z systemem zarządzania walką szczebla taktycznego jest zasadniczym celem Projektu realizowanego w ramach realizacji umowy nr DOB-SZAFIR/09/B/006/01/2021 o wykonanie i finansowanie projektu realizowanego na rzecz obronności i bezpieczeństwa państwa w ramach konkursu 1/SZAFIR/2020 pt. "System wspomagania decyzji ewakuacji medycznej oparty na monitorowaniu parametrów życiowych żołnierza i rozszerzonej świadomości sytuacyjnej". Projekt realizowany jest wspólnie przez naukowców z WAT, inżynierów z firmy Teldat Sp. z o.o. sp.k. oraz przedstawicieli Wojskowego Instytutu Medycznego.

Realizacja projektu wynika z potrzeb Sił Zbrojnych RP w zakresie osiągnięcia priorytetowych zdolności w obszarze technologii zabezpieczenia medycznego, w tym zdolności zdefiniowanych w *FMN Spiral Specification*.

Opracowany na VI poziomie gotowości technologii demonstrator wykorzystywać będzie w warstwie sprzętowej układ sensorów, w warstwie analitycznej struktury przechowywania danych oraz zaawansowane rozwiązania z zakresu teorii ewidencji, zaś w warstwie zobrazowania rozszerzoną świadomość sytuacyjną.

W ramach realizowanego Projektu zostaną opracowane rozwiązania w zakresie diagnostyki pola walki i ewakuacji medycznej nie tylko dla potrzeb Sił Zbrojnych, ale także innych jednostek wykorzystujących elementy zabezpieczenia medycznego [2].

1.2. Zabezpieczenie i ewakuacja medyczna

Współczesne konflikty zbrojne charakteryzują się dużą dynamiką działań oraz wykorzystaniem zaawansowanego potencjału technicznego, który wspiera w jak najszerszym zakresie zasoby ludzkie funkcjonujące w otoczeniu określonej organizacji o zdefiniowanej strukturze. Ich wynikiem może być mniejszy lub większy stopień zagrożenia życia lub zdrowia czynnika ludzkiego, którego wyrazem są straty lub czasowe wykluczenie z możliwości użycia indywidulanego żołnierza lub grup żołnierzy. Celem minimalizacji strat własnych od szeregu lat stosowane są na szeroką skalę działania zmierzające do zapewnienia skutecznej pomocy medycznej na polu walki oraz ewakuacji medycznej celem zapewnienia leczenia stacjonarnego. Efektem tych działań są opracowane w skali lokalnej oraz globalnej procedury zabezpieczenia medycznego. Mają one różnoraki charakter uzależniony od rodzaju instytucji czy organizacji oraz typu misji czy operacji [1].

Dobrze zorganizowany system ewakuacji medycznej jest jednym z elementów zdefiniowanych w „Strategii rozwoju systemu bezpieczeństwa narodowego RP 2022". Zdolność do zabezpieczenia medycznego pola walki została również zidentyfikowana jako zdolność prioryte-

towa w osiąganiu przewagi na przyszłym polu walki. Celem zabezpieczenia medycznego jest opracowanie rozwiązań zapewniających szybkie i pewne działanie w zabezpieczeniu medycznym w tym między innymi szybka ewakuacja medyczna (*Priorytetowe kierunki badań w resorcie obrony narodowej na lata 2017-2026*).

Zbliżone wymagania nakładają na SZ RP zobowiązania sojusznicze wynikające z przynależności do struktur NATO. Wykaz zdolności operacyjnych oraz technicznych zawarty został w *FMN Spiral Specification Roadmap 2020* (na lata 2021-2030) i obejmuje między innymi osiągnięcie zdolności w ramach następujących obszarów:

Spirala 3 – dowodzenie i kierowanie, zarządzanie i śledzenie procesu ewakuacji medycznej;

Spirala 4 – przygotowanie i przyjmowanie raportów o sytuacji medycznej *SITREP*;

Spirala 5 – usługa monitorowania lokalizacji pacjenta, raportowanie medyczne – *MEDISTREP*.

Segregacja żołnierzy na polu walki (tzw. *triage*), ewakuacja rannych i chorych z pola walki do dedykowanych punktów medycznych, udzielanie im pierwszej pomocy medycznej oraz kwalifikowanej pomocy o profilu chirurgicznym i internistycznym w pełnym lub ograniczonym zakresie a następnie ewakuacja do miejsc leczenia ambulatoryjnego należą do podstawowych przedsięwzięć realizowanych w procesie ewakuacji taktycznej w ramach standardu TCCC [6]. Wymieniony standard wykorzystywany jest w SZ RP oraz w innych formacjach mundurowych, w tym w Policji, Straży Pożarnej czy w jednostkach Ratownictwa Medycznego.

2. ARCHITEKTURA RAMOWA SYSTEMU

Opracowanie architektury projektowanego systemu wspomagania decyzji o ewakuacji medycznej jest jednym z zasadniczych przedsięwzięć realizowanego Projektu.

Architektura systemu określa jego komponenty oraz definiuje funkcje tych komponentów i występujące między nimi relacje. Składowe elementy architektury oddziałują ze sobą wzajemnie za pośrednictwem interfejsów. Interfejs wskazuje jakie usługi oferuje dany element.

Architektura może być także rozumiana jako formalny opis systemu lub szczegółowy plan systemu będący podstawą jego realizacji. W odniesieniu do pojęcia architektury systemu słuszną jest także definicja mówiąca, że jest to struktura elementów (komponentów), ich relacje oraz zasady i wytyczne do jego projektowania i rozwoju.

W literaturze występuje szereg modeli opisujących proces tworzenia architektury systemu, które zasadniczo uwzględniają obszary takie jak aspekt operacyjny, informacyjny i techniczny.

W procesie tworzenia architektury systemu i jego elementów niezbędne jest poprawne zdefiniowanie architektury ramowej systemu, która stanowi punkt wyjścia do wyboru architektury referencyjne a wreszcie docelowej opisywanego systemu.

Architektura ramowa (*Framework Architecture*) określa standardowy zestaw kluczowych informacji biznesowych i technicznych służących opracowaniu architektury systemu. W przypadku gdy systemy są tworzone na bazie tej samej architektury ramowej, możliwe jest osiągnięcie zdolności do współdziałania systemów (interoperacyjności) oraz zapewnienie spójności całej architektury korporacyjnej (*Enterprise Architecture*). Do podstawowych i najczęściej wykorzystywanych wzorców architektury ramowej zaliczyć należy: TOGAF (*The Open Group Architecture Framework*) [5], IAF (*Integrated Architecture Framework*), DoDAF (*The Department of Defense Architecture Framework*) [4] czy NAF (*NATO Architecture Framework*) [3].

TOGAF określa ramy architektoniczne, które po dostosowaniu do potrzeb konkretnej organizacji umożliwiają stworzenie jej architektury docelowej. Opracowany i rozwijany przez konsorcjum *The Open Group* składa się z siedmiu głównych części obejmujących ogólny przegląd koncepcji architektury aż do modeli referencyjnych. Zasadniczą wadą rozwiązania jest brak możliwości zastosowania do budowy systemów informatycznych. Z kolei IAF stanowi ramową architekturę korporacyjną, która obejmuje infrastrukturę biznesową, informacyjną, systemową i technologiczną. IAF jest rozwijana przez globalną społeczność architektów od 1996 r i zakłada wykorzystanie czterech poziomów abstrakcji: kontekstowego, koncepcyjnego, logicznego i fizycznego. Ramy architektury korporacyjnej zdefiniowane w DoDAF zapewniają wizualizację dla określonych problemów interesariuszy poprzez punkty widzenia zorganizowane według różnych widoków. Przedstawione rozwiązanie dedykowane jest do systemów złożonych związanych z problematyką integracji i interoperacyjności. Wzorzec architektury ramowej NAF wykorzystywany jest natomiast do opracowywania i opisywania architektury zarówno systemów wojskowych jak i biznesowych. NAF zapewnia możliwość identyfikacji wykorzystania dostępnych środków do wsparcia strategii całej organizacji oraz realizacji założonych celów i zadań w oparciu o zasady, wytyczne i opisy wykorzystywane w procesie projektowania , prezentacji i komunikacji dotyczących architektury korporacyjnej.

Z uwagi na charakter projektu oraz proponowany obszar zastosowania wskazanym jest zastosowanie do opracowania projektu architektury systemu wsparcia decyzji ewakuacji medycznej architektury ramowej NAF. Takie podejście stworzy możliwość integracji opracowywanego rozwiązania z pozostałymi elementami wchodzącymi w skład rozwiązań dedykowanych dla SZ RP.

3. PROJEKT ARCHITEKTURY SYSTEMU SWD-EWMED

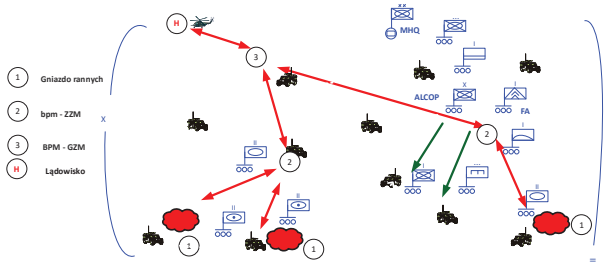
Ramy architektury korporacyjnej proponowanej do zastosowania w odniesieniu do realizowanego Projektu wyznacza architektura korporacyjna NAF w wersji 4. NAFv4 spełnia wymagania niezbędne do opracowania architektury systemu SWD-EwMED ze względu, na fakt, że system ten łączy w sobie elementy organizacyjne, funkcjonalne i teleinformatyczne stosowane w środowisku łączności i informatyki sił zbrojnych.

Projekt architektury systemu wspomagania decyzji o ewakuacji medycznej

Projekt architektury systemu SWD-EwMED stanowiący praktyczną implementację NAFv4 [3], [4] składa się z następujących elementów:

1. Koncepcja systemu:
   a. Taksonomia zdolności – widok C1;
   b. Wizja systemu – widok C2;
   c. Zależności pomiędzy zdolnościami – widok C3.
2. Specyfikacja usług systemu SWD-EwMED:
   a. Taksonomia usług – widok S1;
   b. Interfejsy usług – widok S3;
   c. Interakcje usług – widok S7.
3. Logiczna struktura systemu SWD-EwMED:
   a. Typy węzłów – widok L1;
   b. Interakcje między węzłami – widok L3;
   c. Logiczne powiązania między węzłami – widok L6;
   d. Logiczny model danych – widok L7.

Operacyjny widok proponowanej architektury z uwzględnieniem interfejsów i przepływów przedstawiony został na Rys. 1.



Rys. 1. Schemat widoku operacyjnego dla systemu ewakuacji medycznej

Przedstawiony widok składa się z określonej liczby węzłów logicznych, które realizują zadania związane z monitorowaniem parametrów życiowych żołnierza, ich analizą oraz umożliwiają wnioskowanie o stanie jego zdrowia wraz ze zobrazowaniem zebranej i zinterpretowanej informacji w postaci graficznej.

Zdolność do realizowania określonych działań związanych z procesem wsparcia decyzji o ewakuacji medycznej przedstawiona jest w oparciu o taksonomię zdolności reprezentowaną za pomocą widoku C1 zgodnie ze specyfikacją NAFv4 (Rys. 2).

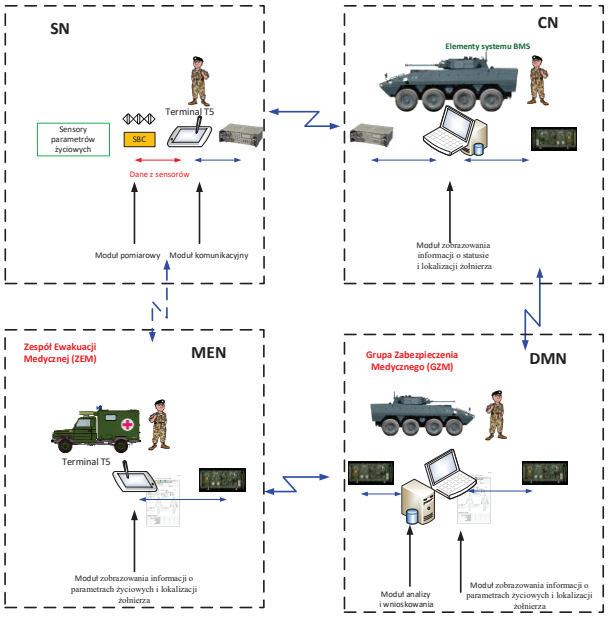
Rys. 2. Taksonomia zdolności systemu SWD-EwMED – widok C1

Zaprezentowane zdolności obejmują monitorowanie sygnałów życiowych, przesyłanie parametrów życiowych, analizę i wnioskowanie oraz zobrazowanie informacji o parametrach życiowych i lokalizacji żołnierza. Wymienione zdolności wynikają z wymagań na systemy ewakuacji medycznej opisane powyżej.

W oparciu o przyjęte założenia opracowana została logiczna struktura systemu SWD-EwMED, która obejmuje następujące powiązane ze sobą węzły (obiekty operacyjne) realizujące określone funkcje:

SN – Soldier Node;
CN – Commander Node;
DMN – Decision Making Node;
MEN – Medical Evacuation Node.

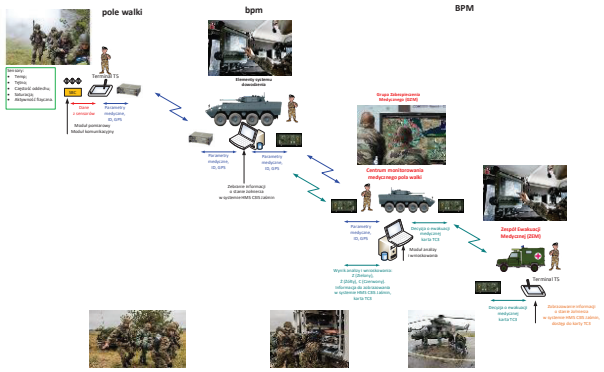
W zaproponowanej strukturze można dodatkowo wyróżnić elementy w postaci modułów i komponentów (Rys. 3). Węzeł (node) jest w omawianej architekturze rozumiany jako element najwyższego poziom opisu architektury, o unikalnej budowie, grupujący elementy niższego rzędu w celu ułatwienia prac projektowych m.in. poprzez odwzorowanie logicznych zależności pomiędzy elementami na ujęcie architektoniczne wyższego poziomu abstrakcji.



Rys. 3. Struktura logiczna systemu SWD-EwMED

Moduł uważany jest za artefakt architektoniczny pośredniego poziomu, o unikalnej budowie, grupujący elementy niższego rzędu w celu ułatwienia prac projektowych m.in. poprzez odwzorowanie fizycznych zależności pomiędzy elementami na ujęcie architektoniczne wyższego poziomu abstrakcji. Komponent to element najniższego poziomu opisu architektury, o unikalnej budowie, powtarzalny w ramach różnych podsystemów.

Niezmiernie istotnym elementem projektowanej architektury SWD-EwMED jest wizja systemu, która umożliwia zidentyfikowanie warstw odpowiedzialnych za realizację zidentyfikowanych zdolności. W zaprezentowanym widoku zidentyfikowane zostały: warstwa funkcjonalna, warstwa informacyjna i warstwa fizyczna (Rys. 4).



Rys. 4. Wizja systemu SWD-EwMED

Warstwa fizyczna odpowiedzialna jest za diagnozowanie obiektów, tj. żołnierzy, za pomocą rejestracji sygnałów, które mogą wskazywać na szeroko pojęte zagrożenia zdrowia i życia żołnierzy. Podstawowym elementem warstwy fizycznej systemu jest układ sensorów rozmieszczonych na węźle (SN) objętym diagnozowaniem. Dedykowane sensory, prowadzą ciągłą i niewymagającą interakcji z operatorem, akwizycję oraz wstępną analizą danych.

W warstwie funkcjonalnej przewidywana jest analiza danych. Analiza wstępna przebiega w oparciu o osobisty moduł analizy zlokalizowany na węźle (możliwość identyfikacji dwóch stanów: zagrożenie, brak zagrożenia). Z kolei analiza szczegółowa realizowana jest przez moduł analizy i wnioskowania zlokalizowany na stanowisku centrum monitorowania medycznego pola walki. W warstwie tej realizowane są zadania bezpośredniej interakcji z operatorem systemu – automatyczna detekcja zagrożeń, wyszukiwanie informacji, prezentacja oraz generacja meldunków.

Warstwa informacyjna rozwinięta jest w oparciu o właściwości zbliżone do warstwy informacyjnej Węzła teleinformatycznego WTi dla której opracowywane są dedykowane interfejsy zapewniające transfer danych. Transfer danych z poziomu węzła SN (moduł pomiarowy – moduł komunikacyjny) do modułu analizy i wnioskowania (węzeł DMN) jest zasadniczą funkcjonalnością warstwy informacyjnej. Pozyskiwane dane (informacje) przekazywane są między elementami systemu za pomocą opracowanego zbioru protokołów sieciowych, umożliwiających niezawodny transfer danych dzięki systemom łączności wykorzystywanym na polu walki w warunkach ograniczonej dostępności łączy, rozspójnienia sieci lub czasowego braku komunikacji.

Kolejną funkcjonalnością przewidywaną w warstwie informacyjnej jest przekazywanie decyzji wypracowanej w module analizy i wnioskowania do modułu zobrazowania informacji o parametrach życiowych i lokalizacji żołnierza (węzły DMN, CN oraz MEN). Dane uzyskane w drodze wypracowania decyzji są prezentowane operatorowi systemu w postaci zwięzłych meldunków zawierających informacje o parametrach życiowych żołnierza, jego lokalizacji wraz ze wskazaniem optymalnej trasy dla grupy ewakuacji medycznej z wykorzystaniem dynamicznych map. Taka forma prezentacji umożliwi operatorowi szybką ocenę zaistniałej sytuacji lub potencjalnego zagrożenia oraz na bezzwłoczne podjęcie decyzji o użyciu grup ewakuacyjnych. Prezentowane informacje zostaną udostępnione w oparciu o moduł wizualizacji interoperacyjny z systemem zarządzania polem walki. Z tego względu w warstwie informacyjnej opracowywany jest interfejs interoperacyjności pomiędzy modułem analizy i wnioskowania a modułem zobrazowania informacji o parametrach życiowych i lokalizacji żołnierza.

4. WNIOSKI

Przedstawiony projekt architektury systemu wspomagania decyzji o ewakuacji medycznej zrealizowany został jako praktyczna implementacja architektury ramowej zgodnej z NAFv4. Wykorzystanie architektury NAFv4 podyktowane zostało charakterem realizowanego Projektu, który jest bezpośrednio związany z obszarem łączności i informatyki. Zastosowanie zaproponowanego rozwiązania umożliwiło opracowanie elementów składowych architektury systemu SWD-EwMED w postaci węzłów, modułów i komponentów. Określone zostały także funkcjonalności oraz rozmieszczenie elementów w systemie. Zaproponowano powiązania oraz relacje pomiędzy elementami systemu, co gwarantuje właściwe jego funkcjonowanie w złożonym środowisku eksploatacji, jakim niewątpliwie jest współczesne pole walki.

Projekt architektury systemu SWD-EwMED podlega aktualnie implementacji praktycznej do modelu laboratoryjnego, który po weryfikacji stanie się podstawą opracowania demonstratora technologii na VI poziomie gotowości technologii.

LITERATURA

[1] Ciuka-Witrylak Małgorzata, Jarosławska-Kolman Karina, Grosset Ryszard.2018. „Nowoczesne rozwiązania w zakresie segregacji medycznej na przykładzie systemu EvaCopNet". *Bezpieczeństwo i Technika Pożarnicza*, Vol. 50, iss. 2, 74-89.

[2] Łubkowski Piotr, Opracowanie finalnej wersji koncepcji architektury systemu SWD-EwMED, Raport Etap 1 Zadanie 1, 2022.

[3] NATO Architecture Framework (NAF) version 4, styczeń 2018, AC/322-D(2015)0009, https://www.nato.int/nato_static_fl2014/assets/pdf/pdf_2018_08/20180801_180801-ac322-d_2018_0002_naf_final.pdf

[4] The DoDAF Architecture Framework Version 2.02, https://dodcio.defense.gov/Library/DoD-Architecture-Framework/. Dostęp: 05.05.2022.

[5] The TOGAF® Standard, 10th Edition. https://www.opengroup.org/togaf. Dostęp: 05.05.2022.

[6] Trzos Arkadiusz. 2016. „Teleinformatyka w ratownictwie medycznym i medycynie ratunkowej". *Ratownictwo medyczne wobec współczesnych zagrożeń*, 133-150.

* *Artykuł powstał w ramach projektu nr DOB-SZAFIR/09/B/006/01/2021 finansowanego przez NCBiR.*

HEURYSTYCZNA DETEKCJA ATAKÓW SIECIOWYCH Z WYKORZYSTANIEM ŚRODOWISKA SNORT 3

Heuristic detection of network attacks using the Snort 3 environment

Sebastian Wnęk[1]; Rafał Kościej[2]; Marcin Niemiec[3]
[1,2,3]AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, al. Mickiewicza 30, 30-059 Kraków
[1]swnek@student.agh.edu.pl, [2]kosciej@student.agh.edu.pl, [3]niemiec@agh.edu.pl

DOI: 10.15199/59.2022.4.14

Streszczenie: **Artykuł przedstawia algorytm detekcji ataków sieciowych, bazujący na podejściu heurystycznym. Szczegółowo opisany został sposób w jaki opracowane rozwiązanie klasyfikuje ruch sieciowy i podejmuje decyzję o zidentyfikowanym zdarzeniu. Algorytm został zaimplementowany w otwartoźródłowym środowisku Snort 3. Dodatkowo zaprezentowano model wspomagający detekcję, oparty na algorytmach uczenia maszynowego.**

Abstract: **The paper describes a heuristic algorithm for network attack detection. The principle of operation of this solution and attack classification were described in detail. The mechanizm was implemented in the Snort 3 – the well-known open source environment for intrusion detection. Additionally, a model supporting attack detection based on machine learning was presented.**

Słowa kluczowe: **bezpieczeństwo, uczenie maszynowe, Snort 3, algorytm heurystyczny, IDS**

Keywords: **security, machine learning, Snort 3, heuristic algorithm, IDS**

1. WSTĘP

Każdego dnia korzystamy z sieci komputerowej logując się do poczty elektronicznej, systemu bankowego banku czy przesyłając ważne wiadomości. Wraz ze wzrostem poziomu cyfryzacji otaczającego nas świata, przenosimy coraz więcej danych do świata wirtualnego. Niestety – skutkuje to również tym, że wrażliwe informacje mogą zostać przechwycone i wykorzystane przez osoby niepowołane. Przykładem takich danych są poufne hasła dostępu. Celem ataków sieciowych stają się także serwisy internetowe, które dostarczają usługi użytkownikom końcowym. Również ataki na dostępność usług, takie jak DoS czy DDoS są już codziennością [1]. W ostatnich latach obserwuje się ciągły wzrost kampanii z wykorzystaniem złośliwego oprogramowania (*ang. malware*). Według raportu firmy Purplesec [2] w 2009 roku wystąpiło 12,4 mln infekcji złośliwym oprogramowaniem, natomiast w 2018 roku było to już 812,67 mln. Daje to 65-krotny wzrost liczby infekcji w ciągu 10 lat.

Użytkownicy cyberprzestrzenii są ciągle narażeni na nowe zagrożenia. Jednak podobnie jak w rzeczywistym świecie, gdzie możemy spotkać różne instytucje i służby dbające o bezpieczeństwo obywateli, tak i w świecie wirtualnym eksperci od cyberbezpieczeństwa opracowują nowe rozwiązania zwiększające poziom ochrony. Jednakże ciągle pojawiają się nowe rodzaje ataków, które są coraz bardziej wyspecjalizowane i trudniej wykrywalne. Dlatego aby zwiększać poziom bezpieczeństwa warto rozwijać nowe metody detekcji ataków sieciowych.

W niniejszy artykuł opisuje nowy algorytm heurystyczny oparty na flagach, entropii oraz współdzielonej wiedzy, która jest gromadzona w zależności od środowiska w jakim będzie wykorzystywane to rozwiązanie. Rozwiązanie to zostało zaprezentowane po raz pierwszy w publikacji [3]. Niniejszy artykuł rozszerza to rozwiązanie o praktyczną implementację w środowisku Snort 3 oraz dodatkowe możliwości detekcji i budowania wiedzy na temat podejrzanych adresów IP. W artykule zaprezentowano koncepcję i wstępne wyniki badań, w których algorytm heurystyczny usprawniono poprzez wykorzystanie modelu klasyfikacji ruchu w oparciu o uczenie maszynowe.

2. SYSTEMY DETEKCJI ZAGROŻEŃ

Systemy typu IDS [4] (*Intrusion Detection System*) są to rozwiązania używane do analizy ruchu sieciowego i wykrywania nietypowych zachowań w sieci, które mogą być potencjalnym zagrożeniem dla lokalnej infrastruktury sieciowej. Systemy tego typu mają za zadanie wykrywanie, a dzięki temu alarmowanie administratorów o możliwym ataku, jak również neutralizację prawdopodobnego zagrożenia, jeśli tylko system wyposażony jest w taką funkcjonalność [5]. Systemy umożliwiające prewencję przed atakami nazywane są IPS (*Intrusion Prevention System*). Rozwiązania takie można podzielić na dwa główne typy:

- rozwiązania programowe (np. Snort),
- rozwiązania całościowe, tzn. składające się z urządzenia wraz z oprogramowaniem).

Każdy z głównych typów IPS można także podzielić w zależności od tego, gdzie system jest stosowany. Spotykanymi rozwiązaniami są: *Host IPS* (HIPS) oraz *Network IPS* (NIPS). HIPS to rozwiązanie działające na pojedynczym hoście i blokujące ataki wymierzone w komputer danego użytkownika. Drugie rozwiązanie, NIPS, stosowane jest do analizy ruchu sieciowego dla całej infrastruktury sieciowej. Ma ono

za zadanie wstępną analizę ruchu i wykrywanie oraz blokowanie zdarzeń niepożądanych.

3. ALGORYTM HEURYSTYCZNY

Heurystyczny algorytm detekcji ataków sieciowych opiera się przede wszystkim na flagach charakteryzujących dane zagrożenie oraz obliczonej entropii danego pakietu. Oryginalna wersja algorytmu została zaproponowana w artykule [3]. Jego składowe zostały opisane w kolejnych podrozdziałach.

3.1. Flagi

Aby sklasyfikować dane zagrożenie zdecydowano się na zdefiniowanie pięciu rodzajów flag. Wartości tych flag mogą być zmieniane przez użytkownika czy administratora, tak aby dostosować działanie algorytmu do przyjętej polityki bezpieczeństwa.

Pierwsza z pięciu to flaga *dangerous*, określająca jaki poziom zagrożenia jest związany z danym adresem IP. W ramach tej flagi, można skorzystać z trzech wartości: High, Medium oraz Low (ich domyślne wartości to odpowiednio: 6, 5 i 2). Kolejna flaga, *attack,* określa jaki typ ataku był powiązany z adresem IP. Może to być np: Malware, Phishing, Ransomware lub DDoS. Flaga *range* określa wpływ ataku na inne komponenty sieci, np. serwer lub router. Możliwe wartości tej flagi to: Single (oznacza, że ucierpieć może tylko zaatakowana stacja), Partial (zaatakowane mogą zostać najbliższe komponenty zainfekowanej stacji) oraz Complete (oznacza, że atak może rozprzestrzenić się na całą sieć). Flaga *access* określa czy do danego ataku potrzebna jest interakcja użytkownika wewnątrz sieci. Flaga ta została podzielna na dwa typy: None oraz User (oznacza to, że atak wymaga interakcji użytkownika). Ostatnim rodzajem flagi jest flaga *availability*. Określa ona, jakie szkody może wyrządzić atak: None, Partial (atak powoduje obniżenie wydajności oraz utratę wybranych funkcjonalności) oraz Complete (całkowita utrata kontroli nad jednostką).

Zaproponowane flagi określają cechy danego zagrożenia, które jednak mogą być od siebie zależne. Z tego powodu wprowadzono współczynniki określające wpływ wybranych flag na końcową wartość pakietu, decydującą o podjętej decyzji (detekcja ataku lub brak zagrożenia). I tak wartości flag *dangerous* oraz *attack* mnożone są przez współczynniki α oraz β, a ich suma powinna wynosić 1. Jest to zabieg mający na celu dostosowanie flag do wybranego modelu ryzyka i zdecydowanie, co ma większy priorytet w kontekście detekcji ataku.

3.2. Wartość entropii

Entropia jest pojęciem szeroko stosowanym w teorii informacji. Została zaproponowana przez Claude'a Shannona w roku 1949 [6]. Jest średnią wartością informacji przenoszoną przez pojedynczą wiadomość. Na podstawie entropii można stwierdzić czy wydarzenie jest rzadkie czy częste. Im wyższa entropia, tym wydarzenie jest mniej prawdopodobne. Wzór pozwalający określić entropię w sposób ilościowy przyjmuje postać:

$$H(s) = \sum_{i=1}^{n} p(x_i)\, log\frac{1}{p(x)} = -\sum_{i=1}^{n} p(x)\, log(p(x_i))$$

Formuła ta jest wykorzystywana do obliczenia wartości entropii dla każdego z analizowanych pakietów. Otrzymana w ten sposób wartość pomaga w detekcji, gdyż na jej podstawie możemy stwierdzić czy pakiet z danego adresu IP jest zdarzeniem rzadkim czy częstym w danej sieci. To z kolei pomaga w wykrywaniu niektórych typów ataków sieciowych – w szczególności ataków typu DoS i DDoS. Warto wspomnieć, że wartość entropii jest mnożona przez współczynnik γ, aby nie zdominowała ostatecznej decyzji dotyczącej pakietu.

3.3. Parametry detekcji

W tabeli 1 przedstawiono wartości domyślne dla podstawowych parametrów związanych z detekcją zagrożenia. Zostały one dobrane tak aby korespondowały z domyślnymi wartościami poszczególnych flag. Model został opracowany w taki sposób aby detekcja była jak najbardziej uniwersalna, a administrator miał możliwość dostosowania algorytmu do własnej infrastruktury sieciowej. Dlatego wszystkie domyślne wartości są jedynie przykładowe i były stosowane podczas weryfikacji funkcjonalnej.

Tab. 1. Wartości domyślne parametrów detekcji

Parametr	Wartość domyślna
sensitivity	20
entropy	6
packet_value	15
filename_malicious	" "

Parametr *sensitivity* określa wartość progu, powyżej którego dany pakiet zostaje uznany za zagrożenie. Z kolei *packet_value* określa początkową wartość każdego pakietu. Dodatkowo określona została maksymalna wartość entropii (parametr *entropy*), powyżej którego pakiet automatycznie zostaje uznany za zagrożenie. Ostatnim z przedstawionych parametrów jest *filename_malicious*, który wskazuje na repozytorium z listą podejrzanych adresów IP. Pozyskanie takich adresów może odbywać się na zasadzie wymiany informacji między firmami w ramach utworzonej federacji lub poprzez wykonanie kompleksowej analizy zagrożeń (np. przez *Threat Intelligence Group*). Podejrzane adresy IP, które brały udział w innych atakach, wraz z odpowiednio określonymi flagami, znajdują się w pliku, którego ścieżkę podaje się w parametrze *filename_malicious*. Algorytm podczas analizy każdego pakietu sprawdza czy informacja o źródłowym adresie IP znajduje się w pliku. Jeżeli tak, pobiera flagi oraz oblicza końcową wartość pakietu. Przyjętym formatem jest Comma-Separated Value (CSV), a adres IP oraz flagi oddzielone są od siebie przecinkami.

3.4. Wartość końcowa pakietu

Decyzja dotycząca analizowanego pakietu jest podejmowana na podstawie tzw. końcowej wartości pakietu. Sposób jej obliczania przedstawia wzór:

$$PV_f = PV_i + (\alpha \cdot dangerous) + (\beta \cdot attack) + range + access + availability + (\gamma \cdot entropy)$$

gdzie:

- α jest współczynnikiem zagrożenia (jego domyślna wartość to 0,65);
- β to współczynnik ataku o domyślnej wartości 0,35;
- γ to współczynnik entropii o domyślnej wartości 0,5;
- PV_f jest końcową wartością pakietu;
- PV_i to początkowa wartość pakietu.

Obliczona w ten sposób wartość określa w sposób ilościowy ryzyko związane z danym pakietem. Następnie wartość końcowa pakietu jest porównywana z progiem detekcji (parametr *sensitivity*), a jeżeli jest wyższa – pakiet klasyfikowany jest jako zagrożenie. Wartości poszczególnych flag mogą zostać dobrane indywidualnie np: poprzez administratora sieci. Podobnie jak współczynniki α, β oraz γ określające wpływ danej flagi i wartości entropii na końcową decyzję. Dzięki temu użytkownik/administrator może swobodnie dobierać wpływ każdego z czynników na końcową wartość pakietu.

4. DETEKCJA ATAKU

Funkcjonalność detekcji ataków sieciowych za pomocą algorytmu heurystycznego zaimplementowano w środowisku Snort [7]. Jest to popularny system detekcji zagrożeń, który jest rozwiązaniem programowym o otwartym kodzie źródłowym. Od roku 2013 środowisko Snort jest zarządzane i rozwijane przez firmę Cisco. Obecnie dostępna jest nowa wersja tego środowiska, nazwana: Snort 3.

Zgodnie z zaproponowanym algorytmem detekcji, do danego adresu IP zostają przypisane odpowiednie flagi. Ich wartości są domyślne lub mogą zostać podane przez użytkownika. Kiedy pakiet z danym adresem zostaje odebrany, przypisywana jest mu wartość początkowa *packet_value*. Następnie od tej wartości zostają dodane wartości poszczególnych flag, przypisanych do danego adresu IP, pomnożone przez odpowiednie współczynniki. Jeżeli wartość pakietu przekroczy wartość *sensitivity* wówczas zostaje on uznany za podejrzany, a następnie jest przekazany do dalszej analizy. Oprócz wartości pakietu, algorytm bierze pod uwagę również entropię danego pakietu. Jeżeli przekracza ona wartość *entropy* wówczas bez względu na wartość *packet_value*, pakiet zostanie uznany za podejrzany i przesłany do dalszej analizy.

Jak wspomniano, informacja na temat zagrożeń jest przechowywana w repozytorium. Przykładowy pojedynczy wpis ma następującą postać:

```
192.168.0.1,H,D,P,N,N,0,0.0
```

Na samym początku znajduje się podejrzany adres IP, następnie występują flagi w następującej kolejności: *dangerous, attack, range, access, availability.* Na końcu znajduje się liczba pakietów odebranych z tego adresu IP oraz przypisana wartość entropii. Dwie ostatnie liczby są na bieżąco aktualizowane – na koniec działania programu nowe wartości są zapisywane.

5. ANALIZA RUCHU SIECIOWEGO

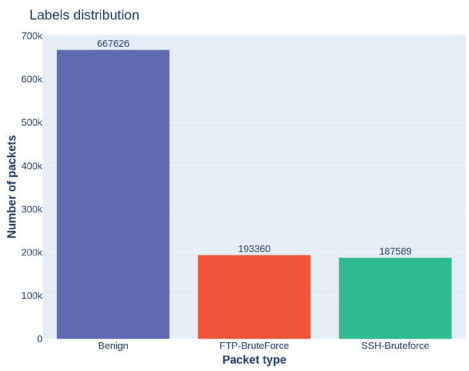
Do poprawnego działania heurystycznego algorytmu detekcji wykorzystywany jest zbiór podejrzanych adresów IP. Może być on na bieżąco aktualizowany przez zaprojektowany w tym celu model klasyfikacji ruchu sieciowego oparty na uczeniu maszynowym. Przykładowo, gdy w ciągu ostatnich kilku minut model znajduje adresy, które z dużym prawdopodobieństwem mogą być atakiem, może dodać je do listy podejrzanych adresów IP. Wprawdzie model powstały w wyniku uczenia maszynowego nie będzie miał 100% skuteczności, ale wytypowane adresy znalazłyby się jedynie w zbiorze podejrzanych adresów IP. Oznacza to że nie byłby jeszcze jednoznacznie sklasyfikowane jako ostrzeżenie o przeprowadzonym ataku, ale algorytm heurystyczny zajmowałby się detekcją ataków przy uwzględnieniu nowych adresów. Dzięki takiej korelacji, zaproponowany system w krótkim czasie może dostosować się do aktualnej sytuacji w sieci i aktywnie reagować na zmiany. Takie podejście podnosi poziom oferowanego bezpieczeństwa.

5.1. Zbiór danych

Do wytrenowania modelu potrzebne są dane, na podstawie których będzie mogła odbywać się klasyfikacja ruchu sieciowego. Autorzy artykułu użyli w tym celu zbiór danych, opracowany przez Communications Security Establishment (CSE) i Canadian Institute for Cybersecurity (CIC) [8], podzielony na kilka części. Aktualnie została wykorzystana część z atakami typu brute force.

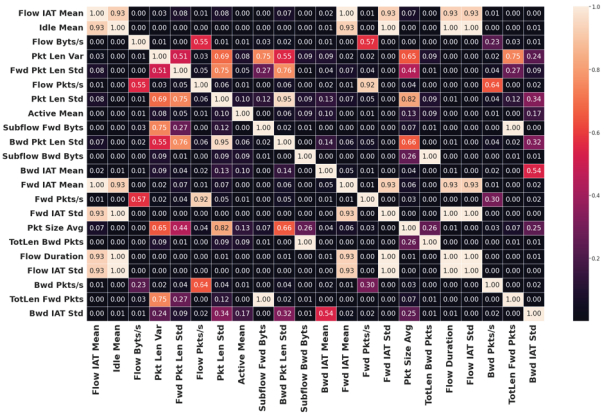
5.2. Analiza

Ataki typu brute force są bardzo powszechne w sieciach komputerowych. Mogą doprowadzić do włamania na konto takiego użytkownika, który stosuje słabe hasła dostępu. Wybrany scenariusz ze zbioru danych został opracowany z myślą o atakach słownikowych na konta SSH i MySQL.



Rys. 1. Rozkład pakietów

Na rysunku 1 przedstawiono rozkład poszczególnych typów pakietów. Realistyczny ruch sieciowy został oznaczony jako “Benign”. Zebrane dane składają się z 80 klas i ponad miliona wierszy. Jednym z problemów jest niezbalansowanie danych [9]. Aby wyeliminować ten problem zostały wybrane kluczowe klasy i usunięte duplikaty. Usunięto również podział na rodzaje ataku FTP i SSH tak, żeby model przewidywał jedynie czy dany pakiet to atak czy nie. Na rysunku 2 przedstawiono tzw. strukturę *heat map*, która dostarcza informacji o tym, w jaki sposób poszczególne klasy są od siebie zależne.



Rys. 2. Mapa współczynnika korelacji wybranych klas

5.3. Rezultaty

Do klasyfikacji danych został wybrany *Random Forest Classifier*. Wybór ten był spowodowany cechami tej metody – przede wszystkim dlatego, że szybko obsługuje duże zbiory danych i może automatycznie równoważyć analizowane zbiory. Uzyskane wyniki zostały przedstawione w tabeli 2.

Tab. 2. Miary wydajności modelu

Metryka	Wartość
Accuracy	0,910
Precision	0,999
Recall	0,754
F1 Score	0,859

Wybrane miary modeli klasyfikacyjnych oparte na algorytmach uczenia maszynowego służą do oceny tego, jak dobrze modele sprawdzają się w danym kontekście.

- Dokładność (*accuracy*) to metryka wydajności modelu. Informuje o tym, jak często model będzie prawidłowo przewidywał wynik, biorąc pod uwagę całkowitą liczbę przypadków.
- Precyzja (*precision*) modelu mierzy odsetek etykiet przewidywanych jako pozytywne, które są rzeczywiście poprawne.
- Miara *recall* reprezentuje zdolność do poprawnego przewidywania wyników pozytywnych spośród faktycznie pozytywnych. Jest to istotna różnica w stosunku do precyzji. Im wyższy wynik recall, tym model lepiej radzi sobie z identyfikacją zarówno pozytywnych, jak i negatywnych przykładów.
- Wynik *F1* dostarcza ogólnych informacji o jakości wyników modelu i jest używany jako metryka w scenariuszach, w których należy wybrać jeden z wyników *precision* lub *recall*.

Na podstawie metryki *accuracy*, model ma pozornie wysoką skuteczność ale biorąc pod uwagę wartości *recall* i *F1*, model w znacznym stopniu błędnie oznacza normalny ruch w sieci jako atak. Jednak dzięki działaniu z algorytmem heurystycznym, błędnie wytypowane adresy nie wywołają fałszywego alarmu o ataku. Uzyskane wyniki są obiecujące, a dalsze prace będą skupione na opracowaniu modeli charakteryzujących się jeszcze lepszymi wynikami detekcji. Przełoży się to na wyższą jakość całego systemu detekcji ataków, który wykorzystuje nowy algorytm heurystyczny, ponieważ repozytorium podejrzanych adresów IP będzie zawierało aktualne dane o zdarzeniach występujących w monitorowanej sieci.

LITERATURA

[1] Warburton David. 2022. “2022 Application Protection Report: DDoS Attack Trends”, https://www.f5.com/labs/articles/threat-intelligence/2022-application-protection-report-ddos-attack-trends [Dostęp: 09.06.2022]

[2] “Cyber Security Statistics The Ultimate List Of Stats, 2021, Data & Trends”, https://purplesec.us/resources/cyber-security-statistics [Dostęp: 09.05.2022]

[3] Niemiec, M.; Kościej, R.; Gdowski, B. 2021. “Multivariable Heuristic Approach to Intrusion Detection in Network Environments.” Entropy, 23, 776. doi:10.3390/e23060776

[4] Stallings, W. 2011. “Cryptography and Network Security: Principles and Practice, 5th ed.” Prentice Hall: Hoboken, NJ, USA.

[5] Oliveira, N.; Praça, I.; Maia, E.; Sousa, O. 2021. “Intelligent Cyber Attack Detection and Classification for Network-Based Intrusion Detection Systems” Appl. Sci., 11, 1674. doi:10.3390/app11041674

[6] Shannon, C. E., & Weaver, W. 1949. “A Mathematical theory of Communication”

[7] Snort Team. “Snort 3 User Manual” https://www.snort.org/snort3 [Dostęp: 09.06.2022]

[8] Communications Security Establishment (CSE) & the Canadian Institute for Cybersecurity (CIC), 2018. “CSE-CIC-IDS2018” https://www.unb.ca/cic/datasets/ids-2018.html [Dostęp: 09.06.2022]

[9] Guo Haixiang, Li Yijing, i in. 2017. “Learning from class-imbalanced data: Review of methods and applications”. Expert Systems with Applications, 2017, 73, 220-239 doi:10.1016/j.eswa.2016.12.035

ACMEC – KOTROLA DOSTĘPU W SIECI 5G MEC. WSPÓLNY PROJEKT BADAWCZY NUST (TAJWAN) I PW (POLSKA)

ACMEC - ACCESS CONTROL IN 5G MEC.
JOINT RESEARCH PROJECT OF NTUST (TAIWAN) AND WUT (POLAND)

Zbigniew Kotulski[1]; Wojciech Niewolski[2]; Tomasz Nowak[3], Mariusz Sepczuk[4], Ruei-Hau Hsu [5] ,Shin-Ming Cheng[6]; Shou-Chung Chang[7]; Tse-Yuan Yu[8]; Cheng-Feng Hung[9]

[1] Warsaw University of Technology, Warsaw, zbigniew.kotulski@pw.edu.pl
[2] Warsaw University of Technology, Warsaw, wojciech.niewolski@wp.pl
[3] Warsaw University of Technology, Warsaw, tomasz.nowak@gmail.com
[4]Warsaw University of Technology, Warsaw, mariusz.sepczuk@pw.edu.pl
[5]National Sun Yat-sen University, Taiwan, rhhsu@mail.cse.nsysu.edu.tw
[6] National Taiwan University of Science and Technology, Taiwan, smcheng@mail.ntust.edu.tw
[7] National Sun Yat-sen University, Taiwan, shouchungchang@gmail.com
[8] National Sun Yat-sen University, Taiwan, jas27690@gmail.com
[9] National Sun Yat-sen University, Taiwan, gyrgey18912@gmail.com

DOI: 10.15199/59.2022.4.15

Streszczenie: **Artykuł przedstawia wyniki prac w ramach projekty polsko-tajwańskiego zrealizowane przez zespoły z National Taiwan University of Science and Technology w Taipei i Instytutu Telekomunikacji Politechniki Warszawskiej. Celem projektu jest propozycja nowej architektury bezpieczeństwa dla sieci 5G ze szczególnym uwzględnieniem systemu kontroli dostępu do usług realizowanych w technologii MEC. Artykuł przedstawia kolejne etapy budowy i badania architektury oraz jej nowe bloki funkcjonalne które tworzą zaawansowany system nadzoru i udostępniania serwisów w środowiskach brzegowych – MEC Enabler oraz modularny schemat uwierzytelnienia MECAC.**

Abstract: **The paper presents the results of the Polish-Taiwanese project by teams from the National Taiwan University of Science and Technology in Taipei and the Institute of Telecommunications of the Warsaw University of Technology. The project proposes a new security architecture for 5G networks, emphasizing the access control system to services implemented in the MEC technology. The paper presents the successive stages of building and testing the architecture and its new functional blocks, which create an advanced system of supervision and protection of applications hosted in the edge environments - MEC Enabler and modular authentication framework MECAC.**

Słowa kluczowe: **kontrola dostępu, sieci mobilne piątej generacji (5G), system udostępniania MEC, modularny system udostępniania MECAC, wielodostępowe przetwarzanie brzegowe (MEC).**

Keywords: **5G mobile networks, access control, Multi-access Edge Computing (MEC), MEC Enabler, MECAC.**

1. INTRODUCTION

The 5G network is a modern approach to mobile networking and services that assume extremely high parameters of its performance. The MEC (**Multi-access Edge Computing)** technology can be integrated with 5G (fifth generation) networking to satisfy these requirements and give additional network functionalities. In such a compound structure, stakeholders are engaged: the network provider, MEC-hosted applications providers, and the network and the applications end-users. They all must access the network, the servers, and applications and cooperate honestly without reducing the system's security. This project proposes a transparent and reliable 5G MEC security architecture with an adaptive access control system.

The project results are several detailed solutions that complement the standard 5G network security architecture, allowing to increase the level of its security in radio access to the network and control of access to services while ensuring a high level of quality parameters of the network operation. In the following sections, we present the completed tasks and results from the project. Section 2 describes the project objectives in detail. Section 3 discusses the results obtained in six Workpackages constituting the project, implemented in WUT (WP 1 - WP 3) and NTUST (WP 4 - WP 6). Section 4 summarizes the work and outlines the directions for the future development of the designed security architecture.

2. OUTLINE OF THE PROJECT

The project proposes a new security architecture for 5G MEC networks, focusing on the access control system services and adaptation of the architecture to actual environmental constraints. The aim of the research is an access control system that will enable access to the service in a time that meets the quality requirements of the 5G network [5]. Moreover, in conjunction with MEC technology [1], the constructed scheme should enable the service to be performed with high data transfer speed and low delays. Another expectation concerning the constructed access control system is enhanced resistance to attacks from the radio network. In addition, the system should adapt to the changing conditions of the network operation while

maintaining high-quality parameters. It should also react to network attacks ensuring a high level of security.

To obtain the expected results of the project, we proposed and solved the following tasks:

- Precise identification of components, interfaces, parties, objects and subjects, relations, and implicit dependencies;
- Identification of threats, vulnerabilities, and risks and proposition of countermeasures;
- Design of access control security architecture for 5G MEC, including the creation of a new access control instance called the MEC Enabler;
- Proposals of new security algorithms and protocols and new methods of investigations and testing. A new JWT has been proposed, called JSON MEC Access Token (JMAT), for the distribution of MEC-Enabler-based authorization credentials;
- Development of experimental platform of 5G MEC using *software-defined radio* (SDR) and open-source software;
- Design of practical rouge base station (BS) attack for 5G MEC;
- Testing, verifying, and analyzing created parts of the solutions mentioned above using a proposed testbed, experimental platform, and practical attacks;
- Optimization of the solution to create an adaptable security architecture;
- Designing the MEC inspector for the operational 5G network by analyzing the responses resulting from crafted inputs.

In Section 3, we present the main work carried out so far in designing a new access control system for the 5G MEC network and compare the proposed solutions with the designs of other authors.

3. RESEARCH TASKS AND RESULTS

3.1. Access control security architecture for 5G MEC

ETSI in [2] has specified 5G security architecture. It defines six security areas for protecting the access procedures and components' domains during the network operation:

I. **Network access security** enables user equipment (UE) to securely authenticate and access services via the network.
II. **Network domain security** enables network nodes to securely exchange signaling and user plane data.
III. **User domain security** protects the user's access to mobile equipment.
IV. **Application domain security** enables applications in the user and provider domains to securely exchange messages.
V. **Service-based architecture (SBA) domain security** enables network functions to communicate securely within the serving network domain and other network domains.
VI. **Visibility and configurability of security** enable the user to be informed whether a security feature is in operation or not.

Designing the security architecture for the 5G MEC network, we expect that it will enable:

- Granting access to the service in the MEC as required by the service provider,
- Performing grant access procedure on time consistent with 5G requirements,
- Enabling the service to be provided with the parameters needed by 5G,
- Adapting the parameters of granting access and providing the service to the environmental conditions (context).

Our security architecture (see [6], [9]) extends the ETSI model to ten security domains and, at the same time, adds protection to the MEC services and network slices. It also includes the new MEC Enabler component, which defines and manages security credentials for access control in MEC-based services.

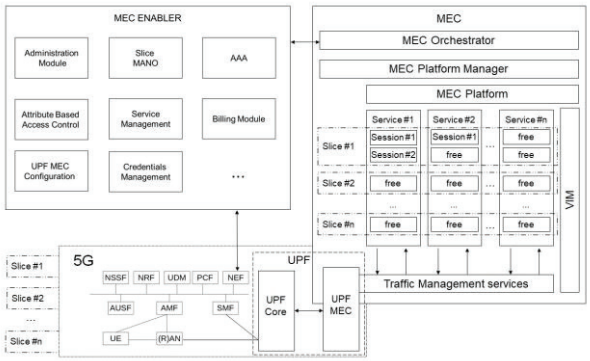


Fig. 1. MEC Enabler in the 5G MEC network

The MEC Enabler (see Fig. 1) is crucial in providing elastic and secure access control services in 5G MEC networks. It consists of several logical blocks (working in a microservices architecture), including:

- **Administration Module** supports all management operations and configures other modules.
- **Slice MANagement and Orchestration (MANO)** manages the slice life cycle in the MEC area, reserving resources for specific slices and marking them according to service type, network identification, or any other rule.
- **Attribute-Based Access Control** stores all policy data concerning services, slices, available connections, and other MEC information.
- **Service Management** checks available services, their utilization, and the number of resources dedicated to them.
- **Credentials Management** creates tokens, which can be used in MEC resources' authorization process. The created token is monitored by this module and refreshed or deleted according to service needs.
- **Billing Module** is used for billing and storing information about MEC usage, which can be used in various business models.
- **User Plane Function (UPF) MEC Configuration** matches and properly configures UPF MEC and links proper network slices with dedicated MEC resources.
- **AAA** is responsible for authentication, authorization, and accounting of all requests to MEC services. After positive verification, the AAA module triggers the

whole MEC access procedure starting from the UPF MEC Configuration Module, which prepares a new network configuration that allows creating a connection with the chosen service and finalizing with the Credentials Management, which generates a token for service authorization.

The proposed security architecture has been implemented and tested in a virtualized environment.

3.2. Testing 5G MEC access control security architecture

The designed architecture must fulfill requirements associated with the network operation, security, usability, etc. Thus, their functionalities should be tested by simulations. A basic version of the MEC Enabler in a testbed was deployed (see Fig.2). Moreover, a concept of access control was created at this stage of the project. The designed access control approach assumes the usage of JMAT tokens to use the service. The performance and security tests have shown that JMAT meets the quality and security requirements for MEC 5G networks [9]. At a later phase of the project, the testing environment was expanded with the elements necessary to simulate the service provision while ensuring a proper level of security met by the access control mechanism. Simultaneously, the added components were tested for correct operation and integration. The next step was the choice of 5G MEC services that can be deployed in the created testbed, simulated or emulated, and adequately tested. The services were selected from vertical industries that can use the potential of the 5G network, in particular the 5G MEC network [10]. Finally, the achieved results are compared with theoretical assumptions, and the architecture is optimized.

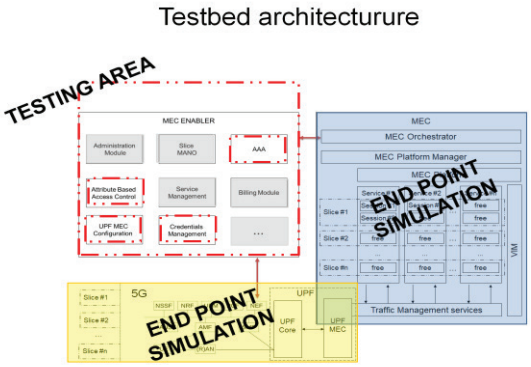


Fig. 2. The testbed architecture

This task includes constructing the testbed with its emulators and simulators, making tests and analyses, and generating experimental datasets for adaptive optimization of the system.

3.3. Optimization of the architecture and access control system

This task aims to design, implement, and apply an iterative method supporting the improvement of the access control system in 5G MEC. Its purpose is to optimize the security architecture, select the best algorithms and protocols, and improve the new proposed solutions. Artificial Intelligence is an excellent candidate for solving those problems. An experimental dataset (a training set) must be generated to use the method. Finally, the access control system will be optimized according to several criteria, e.g., the applied resources and operations costs, its impact on QoS / QoE, expected risks and security level (Quality of Protection), etc.

In the scope of this section, we have archived the following goals and milestones. First, we optimized the JMAT token generation process [9]. The main performance gain was obtained from changes in the data structures that hold policies and by removing out-of-the-box JSON parsing supported by Python 3. The second milestone is the high-level improvement method for implementing improvements in our platform. The improvement method has the following assumptions:

- We use improvement scenarios that are independent of each other or have a small impact on each other.
- Those scenarios should refer to the entities identified as subjects of improvement. The list of improvement scenarios can be extended during further maintenance.
- Scenarios could be applied individually to improve the final solution with each procedure.
- The scenario itself can also be iterative by design. It could be archived by:
  - changing the entity n-times to obtain better and better results in following iterations. This rule does not have to be absolute to prevent the algorithm falls into a local extremum. It cannot explore the rest of the solution space,
  - updating the training dataset with more data to improve the automated algorithm (machine learning algorithm, AI algorithm).

The third milestone was to define example improvement scenarios. One of them is the improvement of pyABAC [7] rules evaluation that is used to determine if the service request should be accepted or denied due to security policy constraints. Incoming requests might be rejected or allowed with pyABAC policies evaluation. The logical OR or AND operation is performed on those policies (depending on the evaluation algorithm). If we guess which policy will return the allowed result, the allowed result might be calculated faster for the OR operation. The request can be classified as a request allowed by the n-th policy with a classification algorithm. Then, this policy can be executed in parallel to get an instant decision. Analogically, the same applies to the AND operation and the denied result. The expected benefit is a statistical reduction of the time needed for checking ABAC policies. The risks are policy modification will cancel the calculations, the policies might have too low success rate, and the total time spent handling requests will be higher than in the naïve method.

3.4. Designing access control for MEC-based applications

Realizing ultra-reliable and low latency communications (URLLC) by multi-access edge computing framework enables application providers to deploy services on near-user edge sides for lower transmission delay and high bandwidth in the fifth-generation mobile networks. Authentication and key exchange are of utmost to ensure the essential security, i.e., user identification and communication confidentiality, for the access control of MEC. However, the design of MEC-based access control in the specification by the standard telecommunication body,

i.e., 3GPP, integrates the control of access to MEC-based applications with the control of communications by the infrastructure of mobile networks. The design may be adverse to providing MEC-based applications by the commercial off-the-shelf telecommunication infrastructures. Thus, our work introduces a new method of MEC-based access control, as shown in Fig. 3, which allows a standardized authentication framework, i.e., OpenID Connect (OIDC), to be integrated for the access control of MEC-based applications with minimized changes to the infrastructures. User authentication [13] for access control [12] is necessary to guarantee the correctness of the access right. In addition, to improve ease of use for service providers, the integration of user authentication and MEC platform access control is indispensable. Thus, a proposed modular authentication framework named MECAC box is independent of 5G and MEC architecture. MECAC supports access control to the network traffic from user equipment to a MEC platform and the access control to MEC-based applications for a minimal dependency on each other. By using the MECAC box, application providers can employ MEC-based access control functions for authentication and secure communications by MECAC.

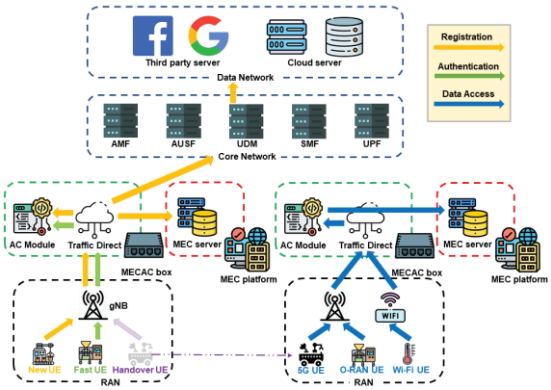


Fig. 3. A system model of secure multi-access edge computing-based access control

This study proposes an access control framework for the practice of MECAC, by which MEC-based application servers with an access control module can authenticate user identity, protect user privacy against identity analysis, and support partial distributed access control. Our security design supports handover functionality and real-time active revocation with high scalability. The contributions of this scheme are shown as follows.

- **Isolation between communication and application access control:** MECAC introduces a two-tier access control for the flexibility to manage MEC platforms' access control and MEC application servers' data upon them by TSPs and MEC service providers, respectively. The management of the access control logic to the services by MEC-based applications is separated from that of the traffic access control to the MEC platform with MEC-based applications.
- **Compatible standards authentication:** In this work, the authentication framework provided by MECAC for access control to the MEC-based application can interface with the standard authentication framework, i.e., OIDC [10].
- **Efficient deployment:** Service providers do not need to modify their original service to deploy MECAC between the base station and MEC platform. Service providers only need to change the configuration file on MECAC for the access traffic control to the corresponding MEC platform for specific applications

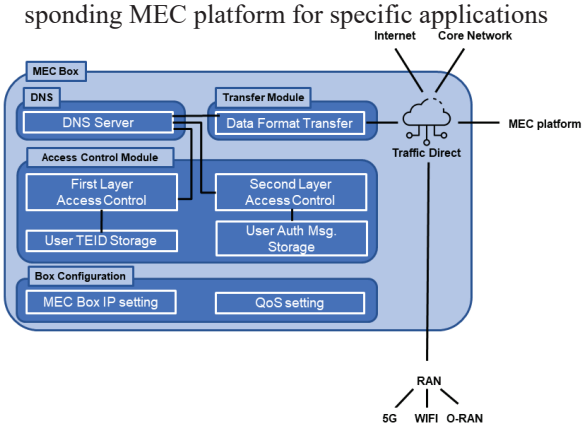


Fig. 4. The system architecture of the MECAC Box

Fig. 4 shows the system architecture of the MECAC. The MECAC system enables the traffic direction for each UE to a MEC platform. It also allows the authentication of each access request to a MEC-based app server. Moreover, the MECAC system can configure the mapping of the domain name of a service to the specific IP of the corresponding app server on the MEC platform domain.

- **Traffic direct:** This entity can redirect the traffic from UE to the Internet or UE to the MEC platform and access MEC-based applications based on the TEID assigned by the UE. Whenever a UE wants to use services on the MEC platform, the traffic direct is responsible for the user's MEC service usability.
- **Access control module:** The access control module is the core of the MECAC system, composed of a modularized access control API. Service providers do not need to modify any of their applications deployed on the MEC platform to achieve access control functions through this module. The access control API can be divided into two parts, one is the server-end, and the other is the UE-end. When the UE wants to access the application service data on the MEC application server, it will first send a request from its access control API and then send it to the access control API of the server to respond. In addition, we also divide the access control module into two tiers.

We compare the security features of papers related to edge computing and the MEC platform, as shown in Tab. 1. All research has achieved the access control property since it is an essential feature in the system for a service provider which can avoid an out-of-authority entity accessing the data. However, almost no research considers the two layers of access control simultaneously, whether for the access control of the communication layer or the application layer. Considering the mobile characteristic of the user, the support of user mobility is regarded as an essential feature in the mobile network design, which can make the system more complete and give users a more convenient experience. Efficiency means the degree of ease deploys access control functionality on the MEC application

server. Besides, compatibility allows the service provider to interface with the system without modifying the original system directly.

Tab. 1. Comparison of security properties

	[8]	*[7]*	*[4]*	*[3]*	*Our Scheme*
Access control	One-tier	One-tier	One-tier	One-tier	Two-tier
Mobility	✓	-	-	✓	✓
Support OIDC Standard	✓	-	-	-	✓
Efficiency	Medium	Medium	Low	Low	High
Compatibility	Low	Low	Low	Medium	High

4. SUMMARY

MEC is regarded as an essential framework in the 5G environment, which can provide users with a low-latency service experience for the application requirements of URLLC. However, the decentralized data storage feature hinders access control to MEC-based applications. This work analyzes the security issues for designing a MEC-based access control framework and proposes the corresponding solution to support user registration and authentication with traffic redirection for MEC-based access control. The proposed MEC-Enabler and MECAC are practical solutions with minimized changes and gaps to the existing infrastructure and service providers for developing the MEC-based applications. Also, for mobility, our proposed MECAC supports handover authentication. In addition, the authentication framework is compatible with the existing authentication framework standard, i.e., OIDC, for high flexibility in implementing user authentication for MEC-based applications.

BIBLIOGRAPHY

[1] ETSI GS MEC 003 V2.1.1. 2019-01. *Multi-access Edge Computing (MEC); Framework and Reference Architecture.*

[2] ETSI TS 133 501 V16.3.0. 2020-08. *5G; Security architecture and procedures for 5G System*.

[3] He, Quanwen, Hui Lin; Jia Hu; Xiaoding Wang, 2021. "A Novel Cross-domain Access Control Protocol in Mobile Edge Computing," 2021 IEEE Global Communications Conference (GLOBECOM), pp. 1-6, doi: 10.1109/GLOBECOM46510.2021.9685650.

[4] HewaTharaka, An Braeken, Mika Ylianttila, and Madhusanka Liyanage. 2020, "Multi-Access Edge Computing and Blockchain-based Secure Telehealth System Connected with 5G and IoT," GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020, pp. 1-6, doi: 10.1109/GLOBECOM42002.2020.9348125.

[5] ITU-R M.2150-0. 2021-02. *Detailed specifications of the radio interfaces of IMT-2020.*

[6] Kotulski Zbigniew, Wojciech Niewolski, Tomasz W. Nowak, and Mariusz Sepczuk. 2021. "New Security Architecture of Access Control in 5G MEC". In: Thampi S.M., et al. (eds) *Security in Computing and Communications. SSCC 2020. Communications in Computer and Information Science*, vol 1364, pp. 77-91. Springer, Singapore. doi: 10.1007/978-981-16-0422-5_6

[7] Ksentini Adlen and Pantelis A. Frangoudis, 2020. "Toward Slicing-Enabled Multi-Access Edge Computing in 5G," *IEEE Network* **34**(2): 99-105. doi: 10.1109/MNET.001.1900261.

[8] Li Chi-Yu, Ying-Dar Lin, Yuan-Cheng Lai, Hsu-Tung Chien, Yu-Sheng Huang, Po-Hao Huang, and Hsueh-Yang Liu, 2020. "Transparent AAA Security Design for Low-Latency MEC-Integrated Cellular Networks," *IEEE Transactions on Vehicular Technology*, 69 (3): 3231-3243. doi: 10.1109/TVT.2020.2964596.

[9] Niewolski Wojciech, Tomasz W. Nowak, Mariusz Sepczuk, and Zbigniew Kotulski. 2021. "Token-Based Authentication Framework for 5G MEC Mobile Networks" *Electronics* 10, no. 14: 1724. https://doi.org/10.3390/electronics10141724

[10] Nowak Tomasz W., Mariusz Sepczuk, Zbigniew Kotulski, Wojciech Niewolski, Rafał Artych, Krzysztof Bocianiak, Tomasz Ośko, and Jean-Philippe Wary. 2021. "Verticals in 5G MEC-use cases and security challenges". *IEEE Access*, **9**: 87251-87298. doi: 10.1109/ACCESS.2021.3088374

[11] *Py-ABAC's documentation*. 2022-04.07. https://py-abac.readthedocs.io/en/latest/

[12] Sandhu Ravi and Pierangela Samarati. 1994, "Access control: principle and practice," *IEEE Communications Magazine*, **32** (9) : 40-48. doi: 10.1109/35.312842.

[13] Wang Peng, Chien-Ming Chen, Saru Kumari, Mohammad Shojafar, Rahim Tafazolli, and Yi-Ning Liu. "HDMA: Hybrid D2D Message Authentication Scheme for 5G-Enabled VANETs," 2021. *IEEE Transactions on Intelligent Transportation Systems*, **22** (8) : 5071-5080. doi: 10.1109/TITS.2020.3013928.

Algorytmy lightweight dla systemów IoT implementowanych w FPGA

Lightweight Algorithms for IoT systems implemented in FPGA

Mateusz Biernacki[1]; Radosław Giermakowski[2]; Mariusz Rawski[3]

[1] Politechnika Warszawska, Warszawa, m.biernacki@tele.pw.edu.pl
[2] Politechnika Warszawska, Warszawa, r.giermakowski@tele.pw.edu.pl
[3] Politechnika Warszawska, Warszawa, m.rawski@tele.pw.edu.pl

DOI: 10.15199/59.2022.4.16

Streszczenie: **Szybki rozwój urządzeń Internetu Rzeczy przekłada się na ogromny wzrost liczby produkowanych urządzeń. Tak gwałtowny rozwój tej dziedziny pociąga za sobą konieczność rozwijania aspektów towarzyszących takich jak kryptografia. Obecnie nie ma standardu, który określałby w jaki sposób zabezpieczać urządzenia IoT dlatego w tej pracy zostały poddane analizie główne prymitywy kryptograficzne. Następnie poruszone były aspekty różnych rozwiązań lightweight na podstawie których wybrane zostały takie, które mogą być używane w systemach Internetu Rzeczy.**

Abstract: **The rapid development of Internet of Things devices translates into a huge increase in the number of manufactured devices. Such rapid development of this field implies the need to develop accompanying aspects such as cryptography. Currently, there is no standard that defines how to secure IoT devices therefore in this work the main cryptographic primitives have been analyzed. Then the aspects of different lightweight solutions were addressed based on which the ones that can be used in Internet of Things systems were selected.**

Słowa kluczowe: **Kryptografia, zabezpieczenie, Internet Rzeczy, systemy wbudowane.**

Keywords: **Cryptography, security, Internet of Things, embedded systems.**

1. WSTĘP

Rozwój urządzeń i systemów, które zaliczają się do kategorii IoT w ostatnich latach bardzo przyspieszył co zaowocowało nowymi urządzeniami, które coraz bardziej wchodzą w życie codzienne. Tak znaczący rozwój tej dziedziny spowodował wzrost zainteresowania aspektami bezpieczeństwa. Coraz więcej osób zaczęło sobie zadawać pytanie na ile urządzenia takiego rodzaju są bezpieczne. Konwencjonalne systemy, gdzie do dyspozycji są znacznie większe zasoby w porównaniu z urządzeniami sieci IoT pozwalają na stosowanie wielu metod zabezpieczania [7]. Natomiast sposób zabezpieczania urządzeń oraz rozwiązań Internetu Rzeczy wymaga skupienia się na innych cechach i aspektach, ponieważ tam zużycie energii i zasobów bywa kluczowe.

Różne rozwiązania Internetu Rzeczy ukierunkowane są na spełnianie różnych celów. Innego podejścia będzie wymagała realizacja architektury Smart City, a innego zaś Smart Home [7], pomimo, że spełniać będą podobną funkcję wysokiego poziomu, czyli poprawę jakości życia na danym obszarze. Trzeba też pamiętać, że każde z takich rozwiązań nie może być realizowane dokładnie tak samo [2]. Różne realizacje będą skupiały się na zupełnie innym aspekcie, czasem będzie to poziom bezpieczeństwa, czasem liczba przetwarzanych danych, a czasem czas życia urządzenia na zasilaniu akumulatorowym.

Zapewnienie bezpieczeństwa urządzeń o ograniczonych zasobach jest wymagające z kilku powodów. Pierwszym nadrzędnym problemem, który napotykany jest bardzo często w urządzeniach IoT to zasilanie bateryjne, które wymusza stosowanie algorytmów, które są energetycznie efektywne [6]. Kolejnym aspektem jest sam problem złożoności obliczeniowej. Konwencjonalne algorytmy mogą być zbyt rozbudowane, żeby były możliwe do uruchomienia na bardzo małych procesorach lub układach ASIC, które mogą znajdować się na przykład w kartach RFID. Kolejnym aspektem jest efekt skali. Rosnąca liczba urządzeń powoduje, że sposób ich zabezpieczania, ale i transmisji danych powinien być odporny i dostosowany do milionowych serii danego rozwiązania.

Projekt IoTrust realizowany w Zakładzie Cyberbezpieczeństwa WEiTI PW umożliwia realizację rekonfigurowalnego frameworku, który realizuje wszystkie najważniejsze funkcje bezpieczeństwa [3]. Pozwala na modułowe dołączanie odpowiednio wybranych i przygotowanych algorytmów kryptograficznych, które mogą cechować się wysokim poziomem bezpieczeństwa, dużą przepustowością lub bardzo ograniczonym zapotrzebowaniem na zasoby. Jest to o tyle istotne, że wcześniej wspomniane zróżnicowane podejście w typowych rozwiązaniach wymagałoby projektowania systemu od początku, zaś w przypadku IoTrust możliwa jest zamiana tylko tych modułów, które nie spełniają założeń projektowych.

2. ALGORYTMY LIGHTWEIGHT

Dziedzina lekkich algorytmów kryptograficznych jest relatywnie młoda. Jednak wzbudza zainteresowania wśród wielu osób z różnych dziedzin. Objawia się to znaczną liczbą algorytmów, które powstają i są publikowane. Autorzy takich algorytmów muszą zmierzyć się z bardzo trudnym zadaniem jakim jest kompromis

pomiędzy trzema cechami, które są najbardziej istotne w IoT [1] tak jak na Rys.1.

Biorąc pod uwagę szyfry blokowe, gdzie zwiększanie poziomu bezpieczeństwa przez wydłużanie długości klucza powoduje znaczące zwiększenie zajętości zasobów. Podobnie jest w przypadku optymalizacji skupionej na szybkości działania, gdzie bardzo często odbywa się to kosztem ograniczenia poziomu bezpieczeństwa algorytmu. Zazwyczaj udaje się spełnić przynajmniej dwa z trzech wymogów, jednak spełnienie wszystkich trzech jednocześnie wymogów jest wyjątkowo trudne.

Rys. 1 Kompromis w podejściu do projektowania algorytmu

Projektowanie algorytmów lekkich może być zrealizowane stosując przynajmniej trzy podejścia. Pierwszym z nich jest optymalizacja już istniejących dobrze znanych i ustandaryzowanych algorytmów kryptograficznych w taki sposób, żeby znacząco poprawić ich wydajność programową lub sprzętową przy jednoczesnym braku zmian w samym sposobie wyznaczania szyfrogramu. Kolejnym podejściem jest niewielka zmiana istniejącego algorytmu, w taki sposób, żeby uzyskać lepsze wyniki. Taka niewielka zmiana powinna być przeprowadzona w taki sposób, żeby nie zmieniać znacząco samego sposobu szyfrowania. Ostatni sposób jest zdecydowanie najtrudniejszy i wymaga zaprojektowania algorytmu od podstaw z nakierowaniem na uzyskanie najlepszych wyników we wszystkich trzech kategoriach.

Trzeba mieć jednocześnie na uwadze, że współczesne szyfry, które nie są orientowane w stronę lightweight nie koniecznie muszą posiadać cechy, które pozwolą na wydajną implementację sprzętową. Jest to związane bezpośrednio z ich przeznaczeniem, czyli wykorzystaniem ich w rozwiązaniach programowych. Dlatego pierwsza metoda, która ma na celu optymalizację może nie przynieść zamierzonych rezultatów dla ustandaryzowanych algorytmów. Jest to dość oczywiste, ponieważ większość takich rozwiązań będzie działała lub działa na urządzeniach klasy PC lub urządzeniach wbudowanych, gdzie zamiast dedykowanego rozwiązania sprzętowego znajduje się procesor lub mikrokontroler.

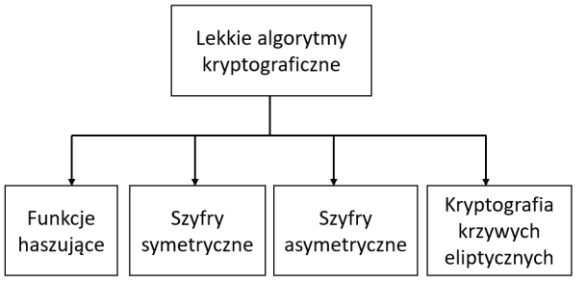
Projektowanie szyfrów lightweight jest wyjątkowo trudne, ponieważ jest to bardzo obszerna dziedzina, która ze względu na wybrany kierunek może wymagać zupełnie innego podejścia. Gdy weźmiemy pod uwagę koncepcję Smart City, w której rozpatrywane są zarówno urządzenia zasilane bateryjnie jak i podłączone do stałego źródła prądu oraz fakt, że takich urządzeń mogą być setki tysięcy to okazuje się, że jest to kolejny problem wymagający pewnego kompromisu.

Skupiając się głównie na rozwiązanych sprzętowych w ramach tego dokumentu można zauważyć, że rozwiązania PUF (*Phisically Unclonable Function*) mogą być implementowane w układach FPGA bez większych problemów [5]. Jest to znaczący krok w kierunku prototypowania systemów z wykorzystaniem układów FPGA, ponieważ pozwala to na generowanie prywatnych kluczy, które są zależne tylko od tego w jaki sposób został wyprodukowany chip. Klucze prywatne są bardzo ważne i mogą być wykorzystywane do celów związanych z zabezpieczeniem samych danych w urządzeniu lub w celach kryptografii asymetrycznej. W przypadku kryptografii symetrycznej nie jest to aż na tyle ważne, ponieważ w takiej sytuacji i tak musi dojść do wymiany klucza z np. serwerem lub odbiornikiem co samo w sobie już jest pewnego rodzaju zagrożeniem, dlatego w takich sytuacjach lepiej wykorzystać inne metody.

3. KLASYFIKACJA LEKKICH ALGORYTMÓW KRYPTOGRAFICZNYCH

Można wyróżnić cztery główne grupy lekkich algorytmów kryptograficznych. Zgodnie z Rys.2. Kryptografia krzywych eliptycznych jest asymetryczna, jednak oparcie tej dziedziny o matematykę krzywych eliptycznych realizuje nieco odmienne podejście od pozostałych co pozwala odseparować taki dział od pozostałych.

Tematyka związana z prymitywami kryptograficznymi, mającymi zastosowanie w kryptosystemach Internetu Rzeczy działającymi na urządzeniach o ograniczonych zasobach jest bardzo obszerna. Wymaga to znacznie szerszego podejścia i bardziej rozbudowanej analizy, dlatego w ramach artykułu podjęte zostaną tylko funkcje skrótu oraz szyfry symetryczne, czyli najbardziej popularne prymitywy lekkiej kryptografii.

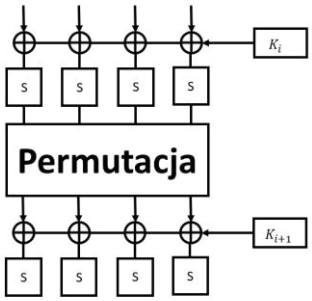


Rys. 2 Podział algorytmów kryptograficznych

3.1. Szyfry symetryczne

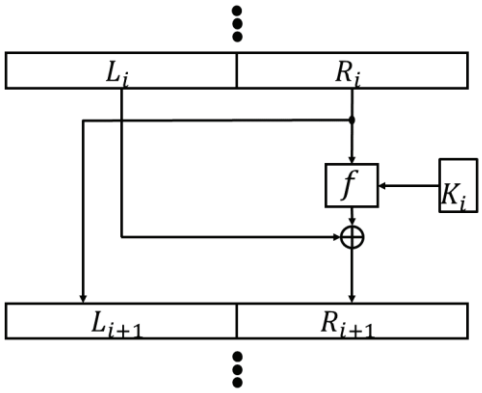
Szyfry blokowe mają charakterystyczną cechę, która polega na tym, że pojedynczy blok danych wejściowych jest przetwarzany jednocześnie. Szyfry blokowe, pomimo że tworzą jedną z czterech grup algorytmów

mogą zostać podzielone na dwa kolejne zbiory. Jednym z nich są algorytmy, które generują szyfrogram na podstawie sieci permutacji-podstawienia Rys.3. Polega to na tym, że pojedyncza runda algorytmu składa się z operacji, które realizują podstawienie danych przez moduł podstawienia (*S-Box*) oraz permutacji (*P*-Box) zależnie od algorytmu. Takie algorytmy przede wszystkim wykorzystują dobrze implementowane operacje takie jak XOR oraz przesunięcia bitowe. Klucz w większości sytuacji generowany jest dla każdej pojedynczej rundy co niestety ma negatywny wpływ na opóźnienia. Istnieją rozwiązania, które nie korzystają z osobnego klucza dla pojedynczej rundy, jednak wtedy taki algorytm jest podatny na ataki.



Rys. 3 Przetwarzanie w sieci permutacji-podstawienia

Drugą grupą są algorytmy oparte na sieci Feistel. Oparte są na zależności podziału bloku danych na dwie części, z których jedna poddawana jest operacji przetwarzania danych.



Rys. 4 Przetwarzanie w ramach sieci Feistel

Ważnym jest dokonanie wyboru odpowiedniego algorytmu, który może być wykorzystywany jako szyfr blokowy w realizacjach Internetu Reczy. Istnieje wiele różnych algorytmów lekkich, które mogą zostać poddane analizie pod kątem wyboru, jednak pod uwagę warto wziąć przede wszystkim *AES,* czyli algorytm zaproponowany przez organizację NIST. Jest to aktualnie najbardziej popularny szyfr blokowy, który nie jest wprost kojarzony z algorytmami lightweight jednak jego popularność oraz zoptymalizowane implementacje sprawiają, że konieczne jest jego porównanie z innymi algorytmami. Kolejnym algorytmem, na który warto zwrócić uwagę to PRESENT, czyli jeden z pierwszych algorytmów, który jest ukierunkowany na bardzo ograniczone zapotrzebowanie na zasoby. Niewątpliwie jest to jeden z lepszych algorytmów lekkich, ponieważ został uznany w certyfikacie ISO/IEC 29192-2. Kolejne dwa algorytmy należą do jednej rodziny i zostały zoptymalizowane w dwóch kierunkach, czyli realizacji programowej – SPECK oraz realizacji sprzętowej – SIMON. Szyfry blokowe należy rozpatrywać w kategoriach zajętości komórek logicznych w przypadku układów FPGA oraz przepustowości pamiętając o długości klucza i bloku danych implementowanych jako dedykowane rdzenie sprzętowe IPCore.

Tab. 1. Porównanie algorytmów lightweight [2]

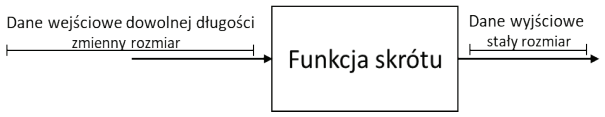
Algorytm	Klucz [b]	Blok [b]	Zajętość [kGE[3]]	Przepustowość [kb/s]
AES[1]	128	128	2,4	56,64
PRESENT[1]	128	64	1,3	12,12
PRESENT[1]	80	64	1,0	12,4
SIMON[2]	128	128	1,3	22,9
SPECK[2]	128	128	1,4	12,9

Tak jak przedstawiono w Tab.[1] można zaobserwować, że AES nie jest algorytmem lekkim i charakteryzuje się znacznie wyższym zapotrzebowaniem na zasoby GE, jednak przy wyższym zapotrzebowaniu jest w stanie zapewnić znacznie wyższą przepustowość i jest to prawie pięciokrotnie wyższa przepustowość przy mniej niż dwukrotnie większym zapotrzebowaniu na komórki logiczne. Algorytm SIMON i SPECK charakteryzują się podobnym zapotrzebowaniem na komórki logiczne względem PRESENT jednak nie są ujęte w ISO/IEC 29192-2.Algortym AES jest ustandaryzowany i powszechnie wykorzystywanym w konwencjonalnych systemach ponadto jego wyższa przepustowość względem proponowanych rozwiązań lekkich jest dodatkową zaletą. Kolejnym algorytmem wartym rozważenia jest PRESENT, który jest szeroko znany. Można więc przyjąć, że AES oraz PRESENT właśnie są jednymi z lepszych wyborów, jeśli chodzi o wykorzystanie szyfrów blokowych w krypto-systemach IoT.

3.2. Funkcje haszujące

Funkcje haszujące nazywane czasem funkcjami skrótu posiadają cechą, która jest bardzo ważna w każdym systemie. Dla dowolnej długości danych wejściowych generują skrót o znanym i ustalonym rozmiarze tak jak pokazano to na Rys.5.

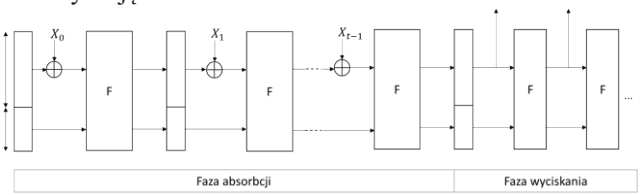
[1] Sieć permutacji podstawienia
[2] Sieć Feistel
[3] GE ozn. ang. *Gate Equivalents*



Rys. 5 Schemat działania funkcji skrótu

Funkcje skrótu opierają się głównie na trzech architekturach: Merkle-Damgård, gąbki oraz funkcji kompresji Davies–Meyer'a. Z punktu widzenia algorytmów lightweight najbardziej istotna i obszerna grupa algorytmów zbudowana jest na bazie gąbki. Sposób w jaki działa algorytm oparty na architekturze gąbki przedstawia Rys. 6.

Aktualnie najbardziej popularną na świecie funkcją skrótu jest SHA-3. Jest to odpowiednia wersja algorytmu Keccak, który jest jednym z ważniejszych kandydatów do analizy funkcji skrótu. Algorytm Keccak jest parametryzowany i przy odpowiedniej konfiguracji może realizować funkcję SHA-3. Kolejnymi kandydatami są funkcje spongent oraz PHOTON, które podobnie umożliwiają parametryzację.



Rys. 6 Schemat gąbki w funkcjach skrótu

Przedstawione algorytmy będą rozpatrywane w wersjach skrótu o szerokości 128 oraz 256 bitów.

Tab. 2. Porównanie funkcji skrótu[4]

Algorytm	Skrót [bitów]	Cykle zegara na blok	Struktura	Zajętość [GE]
PHOTON	128	996	*Gąbka*	1122
Keccak	128	1000	*Gąbka*	509
spongent	128	2380	*Gąbka*	1060
PHOTON	256	996	*Gąbka*	2177
Keccak	256	1200	*Gąbka*	2079
spongent	256	9520	*Gąbka*	1950

Wyniki przedstawione w Tab. 2 opracowane na bazie dedykowanych rdzeni IPCore pokazują, że dla wersji wyniku wynoszącego 128 bitów algorytm Keccak charakteryzuje się najmniejszą zajętością i niemal najlepszym czasem odpowiedzi. W przypadku wersji 256 bitów wyniki nie są aż tak znacząco różne od siebie jednak należy pamiętać, że w tej wersji algorytm Keccak może być utożsamiany ze standardem SHA-3 co niewątpliwie pokazuje, że jest to najlepszy algorytm, który można zaimplementować jako funkcję skrótu w rozwiązaniach IoT.

4. PODSUMOWANIE

W ramach tego artykułu zostały omówione rozwiązania kryptografii symetrycznej oraz funkcji skrótu. Są to jedne z ważniejszych kategorii, które należy rozpatrywać w ramach urządzeń IoT.

Realizując zgrubną analizę dostępnych algorytmów kryptograficznych udało się ustalić takie, które będą najbardziej dostosowane do wymogów i oczekiwań w systemach Internetu Rzeczy. Decyzja została podjęta w głównej mierze na podstawie wyników implementacji sprzętowych oraz rozpoznawalności danego algorytmu. Rozpoznawalność algorytmu nie jest wymierną miarą jednak w przypadku zabezpieczenia systemów całkiem istotną, ponieważ ustandaryzowane lub dobrze zbadane rozwiązania pozwalają na nieco większe zaufanie. W ramach tej pracy udało się ustalić, że dobrze rokujące algorytmy, które można wykorzystać w IoT to dla funkcji skrótu Keccak, zaś w przypadku szyfrów blokowych będzie to AES oraz PRESENT. Konieczna jest dalsza analiza wydajności, ponieważ każdy z algorytmów pozwala na dalsze optymalizacje architektury, które mogą poprawić wyniki. Będzie to przedmiotem dalszych prac badawczych polegających na opracowaniu dedykowanych sprzętowych rdzeni IPCore i analizie ich złożoności sprzętowej i efektywności.

FINANSOWANIE

Praca współfinansowana przez Narodowe Centrum Badań i Rozwoju w ramach projektu Nr CYBERSECIDENT/456446/III/NCBR/2020.

LITERATURA

[1] Berkel, A., Singh, Prince. (2018). An Information Security Architecture for Smart Cities. 10.1007/978-3-319-94214-8_11.

[2] Bhardwaj, Isha, Ajay Kumar and Manu Bansal. "A review on lightweight cryptography algorithms for data security and authentication in IoTs." 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC) (2017): 504-509.

[3] Chmiel, M., Korona, M., Kozioł, F., Szczypiorski, K., Rawski, M. Discussion on IoT Security Recommendations against the State-of-the-Art Solutions. Electronics,10(15), 2021.

[4] Hammad, B. , Jamil, N. , Rusli, M., Z`aba, M., "A survey of Lightweight Cryptographic Hash Function", International Journal of Scientific & Engineering Research Volume 8, Issue 7, 2017

[5] Majzoobi, M., Koushanfar F., Potkonjak, M. "Lightweight secure PUFs," 2008 IEEE/ACM International Conference on Computer-Aided Design, 2008, pp. 670-673, doi: 10.1109/ICCAD.2008.4681648.

[6] McKay, K. , Bassham, L. , Sonmez Turan, M. and Mouha, N. (2017), Report on Lightweight Cryptography, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=922743 (Accessed April 27, 2022).

[7] Xu, T., Wendt, J. B., Potkonjak, M. "Security of IoT systems: Design challenges and opportunities," 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2014, pp. 417-423, doi: 10.1109/ICCAD.2014.7001385.

IoTrust – framework realizujący mechanizmy bezpieczeństwa end-to-end dla IoT

IoTrust – framework implementing end-to-end security mechanisms for IoT

Bartosz Zabołotny[1]; Marta Chmiel[1]; Mateusz Korona[1]; Fryderyk Kozioł[1]; Mariusz Rawski[1]

[1]Instytut Telekomunikacji, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska
(b.zabolotny, m.chmiel, m.korona, f.koziol, m.rawski)@tele.pw.edu.pl

DOI: 10.15199/59.2022.4.17

Streszczenie: **Zabezpieczenie urządzeń Internet of Things (IoT) przed obecnymi zagrożeniami cybernetycznymi, ale także przed zagrożeniami, które mogą się pojawić w trakcie życia tych urządzeń, jest krytycznym wyzwaniem dla twórców systemów wbudowanych. Niskobudżetowe urządzenia IoT wymagają niestandardowego rozwiązania zaprojektowanego w celu sprostania wymaganiom środowisk o ograniczonych zasobach uruchomieniowych. Konieczne jest opracowanie rozwiązań, które zapewnią projektantom IoT narzędzia potrzebne do tworzenia efektywnych mechanizmów zabezpieczenia ich urządzeń. W artykule przedstawiono koncepcję frameworku IoTrust, będącego platformą dla mechanizmów cyberbezpieczeństwa IoT w postaci System on Chip realizującego konfigurowalne środowisko Trusted Execution Environment.**

Abstract: **Securing Internet of Things (IoT) devices from current cyber threats, but also from threats that may arise during the life of these devices, is a critical challenge for embedded system developers. Low-budget IoT devices require a custom solution designed to meet the demands of environments with constrained resources. There is a need to develop solutions that provide IoT designers with the tools they need to create effective security mechanisms for their devices. This paper presents the concept of the IoTrust framework, which is a platform for IoT cybersecurity mechanisms in the form of an System on Chip implementing a configurable Trusted Execution Environment.**

Słowa kluczowe: **IoT, cyberbezpieczeństwo, SoC, TEE, HWRoT.**

Keywords: **IoT, cybersecurity, SoC, TEE, HWRoT.**

1. WSTĘP

Internet rzeczy IoT (*Internet of Things*) to rewolucyjna koncepcja obejmująca wielki ekosystem, w którym połączone ze sobą urządzenia i usługi gromadzą, wymieniają i przetwarzają dane w celu dynamicznego dostosowania się do kontekstu. Szacuje się, że jest nawet 12 miliardów aktywnych urządzeń IoT, a liczba ta może podwoić się w nadchodzących 5 latach za sprawą nowych sensorów, aparatów *smart home* czy też kamer monitoringu wizyjnego [1][2].

Niestety wiele urządzeń IoT jest oddanych do użytku bez odpowiednich środków bezpieczeństwa i są one narażone na różnego rodzaju ataki. Jaskrawym przykładem takiego zdarzenia jest zdalne przejęcie kontroli nad samochodem Jeep na autostradzie w St. Louis przez zespół badaczy [3]. Znaczące skutki mogą odnosić ataki *Distributed Denial of Service* (DDoS), których źródłem są sieci *botnet* składające się z tysięcy zainfekowanych urządzeń IoT, jak w przypadku wirusa *Mirai*, gdzie atakowani byli dostawcy usług internetowych w USA i Francji [4][5].

Według Agencji Unii Europejskiej ds. Cyberbezpieczeństwa ENISA stawienie czoła wyzwaniom związanym z zabezpieczeniem produktów i usług IoT jest dzisiaj fundamentalnym zagadnieniem. Agencja przygotowała zestaw rekomendacji dla konstruowania urządzeń IoT, które mają być wykorzystane w infrastrukturze krytycznej [6]. Również amerykański Narodowy Instytut Standaryzacji i Technologii (NIST) zdefiniował podstawowy zestaw funkcji z zakresu bezpieczeństwa cybernetycznego (*Core Cybersecurity Feature Baseline*) [7], który jest zbiorem właściwości oraz możliwości technicznych implementowanych w urządzeniu niezbędnych do obsługi typowych mechanizmów cyberbezpieczeństwa, które chronią to urządzenie, jego dane oraz całe systemy i ekosystemy IoT.

W [8] projektanci IoTrust przedstawili syntezę z analizy literatury i materiałów dotyczących urządzeń IoT w kontekście cyberbezpieczeństwa oraz istniejących rozwiązań do zabezpieczenia IoT. Bazując na wynikach tej analizy opracowano koncepcję architektury sprzętowo-programowego *frameworku* IoTrust, który zastosowany w procesie projektowania urządzeń IoT zapewni ich bezpieczeństwo informatyczne. Koncepcja zakłada wykorzystanie rozwiązań z obszaru *Hardware Root of Trust* (HWRoT) oraz *Trusted Execution Environment* (TEE). Proponowane rozwiązanie wykorzystuje również elementy *Trusted Computing*. Zaproponowana architektura frameworku zakłada połączenie wspomnianych mechanizmów i opracowanie nowych, aby w rezultacie stworzyć konfigurowalne i kompleksowe rozwiązanie zapewniające bezpieczeństwo węzłów IoT.

W dalszej części artykułu przedstawiono ogólny zarys frameworku IoTrust, opis jego głównych modułów oraz szczegóły implementacji TEE.

2. Framework IoTrust

Framework IoTrust ma stanowić platformę dla mechanizmów cyberbezpieczeństwa IoT w postaci *System on Chip* (SoC) realizującego konfigurowalne środowisko TEE. Środowisko to budowane jest z rdzeni IPCore opisanych w języku opisu sprzętu HDL realizujących mechanizmy HWRoT oraz biblioteki programowej umożliwiającej implementację funkcji TEE z użyciem opracowywa-

nych komponentów sprzętowych. Proponowane rozwiązanie może być dostosowane do implementacji IoT zarówno w technologii FPGA (*Field Programmable Gate Array*), jak i ASIC (*Application Specific Integrated Circuit*). Częścią frameworku są biblioteki udostępniające zaufane Application Programming Interface (API) umożliwiające realizację typowych mechanizmów cyberbezpieczeństwa z użyciem zaimplementowanego zestawu funkcji *core baseline*. Opracowane rozwiązanie będzie stanowić kompleksowy, konfigurowalny, programowo/sprzętowy framework do budowy zaufanego środowiska uruchomieniowego TEE dla rozwiązań IoT umożliwiającego realizację mechanizmów bezpieczeństwa *end-to-end* dopasowanych do wymagań użytkownika.

Realizacja frameworku IoTrust wymaga opracowania koncepcji i implementacji zbioru konfigurowalnych modułów systemu SoC, które funkcjonalnie rozwiązują problemy bezpieczeństwa wyszczególnione przez autorów publikacji [8]. W tym celu konieczne jest użycie zarówno komponentów sprzętowych, jak i programowych oraz mechanizmów komunikacji między nimi. Na *Rys. 1* przedstawiono architekturę przykładowego systemu SoC dla urządzenia IoT opartego na standardowej szynie danych i typowych blokach pamięci czy urządzeniach wejścia/wyjścia wraz ze specjalizowanymi komponentami frameworku IoTrust.

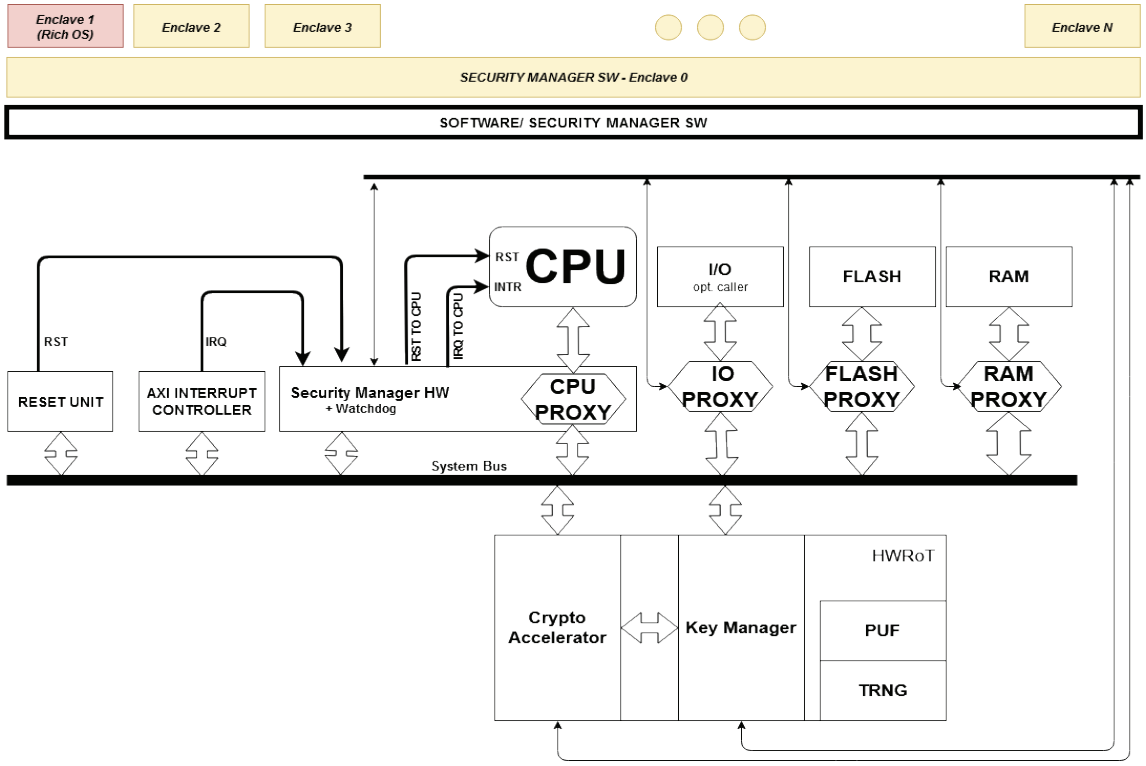
Kluczowymi elementami proponowanego rozwiązania są bloki:

- **Security Manager Hardware** oraz **Security Manager Software**, które umożliwiają realizację koncepcji *Trusted Computing* poprzez tworzenie enklaw, w których wykonywany jest kod i zapewnienie ich fizycznej izolacji (każda enklawa ma dostęp do wydzielonych zasobów sprzętowych), co z odpowiednim użyciem modułu HWRoT pozwala na realizację środowiska TEE,
- **Hardware Root of Trust**, który pełni rolę *źródła zaufania* w systemie i udostępnia bezpieczne sprzętowe implementacje niezbędnych algorytmów kryptograficznych,
- **RAM Proxy**, który realizuje obsługę dostępu do zewnętrznej pamięci RAM z zabezpieczaniem danych.

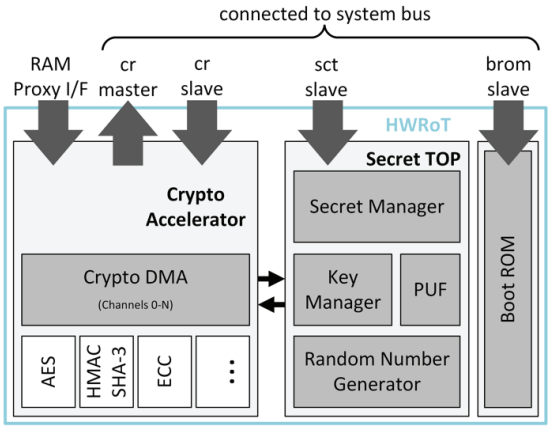
2.1. Implementacja Hardware Root of Trust

Hardware Root of Trust (HWRoT, także *Trust Anchor*) to komponent, który składa się z modułów obliczeniowych, kodu wykonywalnego oraz danych (np. kluczy kryptograficznych), udostępniający funkcje bezpieczeństwa niezbędne do zapewnienia zaufania w ramach platformy (np. poufność, integralność, weryfikacja, autoryzacja, bezpieczne przechowywanie oraz aktualizacja). Jego bezwzględnymi cechami są niezmienność (*immutability*) oraz przewidywalność – w określonych warunkach zawsze zachowuje się w ten sam sposób. Sprzętowa implementacja umożliwia spełnienie powyższych warunków.

HWRoT przez zagwarantowaną niezmienność (np. kod wykonywalny ładowany z pamięci tylko do odczytu) jest traktowany jako zaufany element systemu. Dzięki usługom bezpieczeństwa, które oferuje, może weryfikować poprawność kolejno uruchamianych modułów programistycznych. W ten sposób zaufanie może być propagowane w ramach platformy i formowany jest tzw. łańcuch zaufania (*Chain of Trust*), którego pierwszym elementem jest HWRoT, a ostatnim system operacyjny lub uruchomiona w nim aplikacja. Bezpieczne uruchomienie systemu z utworzeniem łańcucha zaufania nazywane jest procesem *secure boot*.



Rys. 1 Schemat blokowy systemu SoC wykorzystującego framework IoTrust



Rys. 2 Schemat blokowy modułu HWRoT w IoTrust

Układ Hardware Root of Trust platformy IoTrust (Rys. 2) składa się z trzech głównych komponentów. Pierwszym z nich jest **pamięć tylko do odczytu** dla pierwszego *bootloadera*, który przeprowadza wstępną inicjalizację systemu oraz odpowiednio steruje blokami HWRoT w procedurze *secure boot*. Kolejnym elementem jest blok ***crypto accelerator***, który implementuje niezbędne prymitywy kryptograficzne (również tzw. algorytmy lekkie, odpowiednie dla urządzeń o ograniczeniach sprzętowych) wraz z układem DMA umożliwiającym wydajne operacje na danych. Jest to moduł, który również pośrednio umożliwia szyfrowanie pamięci RAM dla poszczególnych procesów. Ostatnim jest blok ***secret top***, który zarządza kluczami kryptograficznymi i innymi sekretami (np. fizyczny oraz logiczny identyfikator urządzenia, tożsamość właściciela itp.). Odpowiada również za generację bezpiecznych kryptograficznie ciągów losowych (wykorzystywanych np. do generacji kluczy kryptograficznych w samym HWRoT) oraz zawiera blok ***Physical Unclonable Function***, który realizuje unikalny *fingerprint* dla każdego układu scalonego.

2.2. Realizacja Trusted Execution Environment

Środowisko *Trusted Execution Environment* to wydzielony bezpieczny obszar jednostki obliczeniowej, gdzie zapewniono izolację od innych uruchomionych procesów. Ten obszar gwarantuje autentyczność wykonywanego kodu, integralność zasobów (rejestrów *Central Processing Unit* (CPU), pamięci czy urządzeń wejścia/wyjścia) oraz poufność kodu, danych i stanów pamięci stałej. Często zapewniane są również mechanizmy zdalnej atestacji wykonywanego kodu. TEE to sposób implementacji ogólnego pojęcia *Trusted Computing* łączącego wiele technologii, aby zapewnić gwarancje dotyczące oprogramowania uruchomionego na urządzeniu.

2.2.1. Enklawa

Enklawę zdefiniowano jako zarządzane przez komponent **Security Manager** bezpieczne środowisko uruchomieniowe (*Runtime*), wraz z opisującymi je metadanymi (zbiór uprawnień dostępu do sprzętu, sumy kontrolne czy cyfrowy podpis kodu) oraz izolowaną przestrzenią adresową, w którym uruchamiane są instancje zweryfikowanych aplikacji. Każda enklawa uruchomiona na procesorze pracuje w osobnej wirtualnej przestrzeni adresowej.

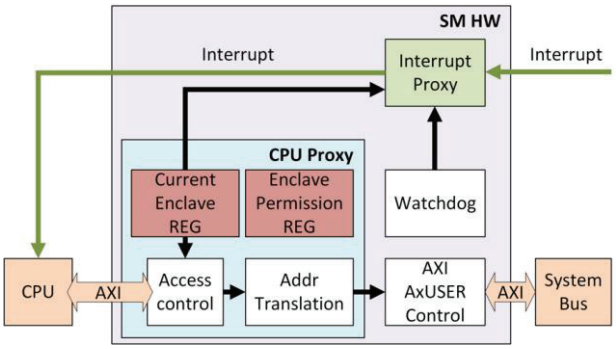
2.2.2. Security Manager Software

Security Manager Software (SM-SW) to programowy moduł odpowiedzialny za zarządzanie enklawami i ich cyklem życia – programowy hipernadzorca (*hypervisor*). Przydziela czas procesora enklawom oraz kolejkuje enklawy oczekujące. Wykorzystując emulowane przerwania programowe udostępnia enklawom możliwość interakcji z resztą systemu w sposób bezpieczny za pomocą API SM-SW. Umożliwia wymianę danych między enklawami.

W skład modułu SM-SW wchodzi również obsługa przerwania generowanego przez moduł Security Manager Hardware (SM-HW), związanego z przełączeniem aktywnej enklawy. Jest to procedura, która umożliwia zagwarantowanie bezpiecznego i pewnego przełączenia między niekoniecznie zaufaną enklawą, a kodem hipernadzorcy posiadającego prawie pełny dostęp do wszystkich komponentów systemu.

2.2.3. Security Manager Hardware

Moduł **Security Manager Hardware** (SM-HW, Rys. 3) jest podłączony bezpośrednio do jednostki CPU zapewniając mechanizmy związane z *Trusted Computing*. W bloku **CPU Proxy** zaimplementowano sprzętowe tłumaczenie adresów transakcji na szynie systemowej z przestrzeni wirtualnej widzianej przez daną enklawę do przestrzeni fizycznej szyny systemowej. Transakcje naruszające uprawnienia dostępu zostają odrzucone. Dodatkowo informacja o aktualnie aktywnej enklawie jest przekazywana do innych bloków frameworku IoTrust, aby zapewnić izolację procesów od urządzeń systemu oraz umożliwić dostępy do pamięci (szyfrowanie/deszyfrowanie) związane z numerem wykonywanej enklawy.



Rys. 3 Schemat blokowy modułu SM-HW IoTrust

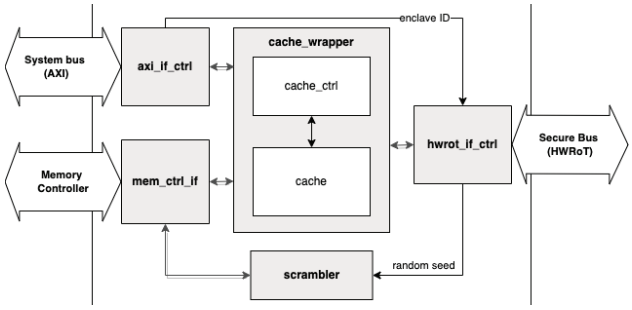
Przerwania systemowe są przechwytywane i przekazywane do jednostki obliczeniowej tylko w przypadku, gdy wykonywany jest kod enklawy uprawnionej do jego obsługi. Mechanizm rejestru *Current Enclave* i przerwania *Watchdog* umożliwia kontrolę bezpieczeństwa systemu. W przypadku braku obsługi przerwania przez wyznaczoną enklawę w odpowiednim czasie, moduł *Watchodog* sygnalizuje nieprawidłowe działanie systemu na skutek błędu lub w wyniku cyberataku. Następuje wtedy zgłoszenie zdarzenia w logu oraz reset systemu.

Zmiany konfiguracyjne oraz naruszenia dostępowe są zgłaszane za pomocą modułu logowania zdarzeń cy-

berbezpieczeństwa i zostają *zapieczętowane*, aby zapewnić ich poufność oraz integralność przy odczycie przez uprawniony podmiot.

2.1. Implementacja RAM Proxy

Zadaniem modułu **RAM Proxy** jest zabezpieczanie i obsługa dostępów do zewnętrznej pamięci RAM. Jest to niezbędny element, który pośredniczy między SM-HW oraz kontrolerem pamięci przy wymianie danych. Ponadto ściśle współpracuje z modułem HWRoT w celu enkrypcji i dekrypcji danych. Komponent ten (Rys. 4) ma trzy główne interfejsy – systemowy, do kontrolera pamięci oraz do HWRoT (*secure bus*).



Rys. 4 Schemat blokowy modułu RAM Proxy IoTrust

RAM Proxy po otrzymaniu z systemu żądania operacji z użyciem pamięci, określa rodzaj operacji i identyfikuje zgłaszającą ją enklawę. Głównym elementem układu jest bufor *Cache* przechowywujący ostatnio odczytane bloki pamięci. Jego zastosowanie zwiększa wydajność układu, dzięki ograniczeniu liczby interakcji z pamięcią oraz HWRoT. Przetwarzanie każdego żądania rozpoczyna się od sprawdzenia zawartości *Cache*. Gdy przechowywane tam bloki nie zawierają żądanego fragmentu danych, inicjowany jest odczyt z pamięci.

Każdy odczytany blok przekazywany jest do HWRoT do dekrypcji wraz z identyfikatorem enklawy, aby możliwe było wykorzystanie właściwego klucza kryptograficznego. Jeśli operacją jest odczyt, odszyfrowany blok danych przekazywane jest w odpowiedzi na szynie systemowej. W przypadku zapisu – odpowiedni fragment jest podmieniany, blok danych jest ponownie szyfrowany w HWRoT i przez *Cache* przekazywany do kontrolera pamięci. Możliwe jest zastosowanie dodatkowego zabezpieczenia w postaci *scramblingu* adresów pamięci z wykorzystaniem źródła losowości z HWRoT.

3. Podsumowanie

Na podstawie przeglądu aktualnych publikacji, obserwujemy rosnącą świadomość zagrożeń związanych z cyberbezpieczeństwem, na które narażone są urządzenia działające w ramach sieci IoT. Coraz więcej instytucji wydaje zalecenia dotyczące zabezpieczenia elementów takich sieci. Framework projektowany i realizowany w ramach projektu IoTrust jest propozycją, która ma na celu zapewnienie cyberbezpieczeństwa urządzeń IoT z wykorzystaniem koncepcji HWRoT. Docelowo ma być to rozwiązanie konfigurowalne (np. w zakresie wykorzystywanych algorytmów kryptograficznych), kompatybilne z różnymi platformami sprzętowymi, np. procesorami różnych producentów.

Wykorzystanie koncepcji TEE i SoC jest zgodne z najnowszymi trendami w dziedzinie zapewnienia bezpieczeństwa systemów wbudowanych. Badania prezentowane w literaturze pokazują, że TEE ma duże praktyczne znaczenie i możne być wykorzystywane do konstruowania złożonych systemów. Jednak brak formalnej definicji TEE sprawia, że istniejące rozwiązania znacząco się różnią. Dla potrzeb realizacji Projektu są wykorzystywane najlepsze wzorce zaproponowane przez czołowych producentów rozwiązań typu SoC, jak Xilinx, czy ARM.

FINANSOWANIE

Praca współfinansowana przez Narodowe Centrum Badań i Rozwoju w ramach projektu Nr CYBERSECIDENT/456446/III/NCBR/2020.

LITERATURA

[1] Sinha, Satyajit, State of the IoT: Number of connected IoT devices growing 9% to 12.3 billion globally, cellular IoT now surpassing 2 billion, 22.09.2021, https://iot-analytics.com/number-connected-iot-devices/ (Dostęp 06.05.2022)

[2] Ericsson, IoT connections outlook, https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/iot-connections-outlook (Dostęp 06.05.2022)

[3] Greenberg, Andy, Hackers Remotely Kill a Jeep on the Highway – With Me in It, 21.07.2021, https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/ (Dostęp 06.05.2022)

[4] Graff, Garret M. How a Dorm Room Minecraft Scam Brought Down the Internet, 13.12.2017, https://www.wired.com/story/mirai-botnet-minecraft-scam-brought-down-the-internet/ (Dostęp 06.05.2022)

[5] Green, Andy, The Mirai Botnet Attack and Revenge of Internet of Things, 27.10.2016, https://www.varonis.com/blog/the-mirai-botnet-attack-and-revenge-of-the-internet-of-things, (Dostęp 06.05.2022)

[6] European Union. Agency for Network and Information Security, Baseline security recommendations for IoT in the context of critical information infrastructures, 11.2017, doi:10.2824/03228 (Dostęp 06.05.2022)

[7] Fagan, M.; Megas, K.N.; Scarfone, K.; Smith, M. IoT device cybersecurity capability core baseline. Technical report, National Institute of Standards and Technology, Gaithersburg, MD, 2020. doi:10.6028/NIST.IR.8259a. (Dostęp 06.05.2022)

[8] Chmiel, M.; Korona, M.; Kozioł, F.; Szczypiorski, K.; Rawski, M. Discussion on IoT Security Recommendations against the State-of-the-Art Solutions. Electronics 2021, 10, 1814. https://doi.org/10.3390/electronics10151814 (Dostęp 06.05.2022)

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

SYSTEM PRZETWARZANIA I DYSTRYBUCJI DANYCH OPARTY NA APACHE NIFI

DATA PROCESSING AND DISTRIBUTION SYSTEM BASED ON APACHE NIFI

Karol Wnęk[1]; Piotr Boryło[2]

[1] AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, karolwnek@student.agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, piotr.borylo@agh.edu.pl

DOI: 10.15199/59.2022.4.18

Streszczenie: **Głównym celem niniejszego artykułu jest zaprezentowanie wpływu autorskich poprawek domyślnej konfiguracji na szybkość przetwarzania danych przez Apache NiFi. Dodatkowo zbadano jak skaluje się wydajność wraz ze wzrostem liczby węzłów w klastrze obliczeniowym. Uzyskane wyniki szczegółowo przeanalizowano pod kątem wydajności oraz wartości kluczowych wskaźników.**

Abstract: **The main purpose of this article is to present the impact of authors' tweaks to the default configuration on the processing speed of Apache NiFi. Additionally, how the performance scales with the increasing number of nodes in a computing cluster has been examined. Achieved results were thoroughly analyzed in terms of performance and key indicators.**

Słowa kluczowe: **Apache NiFi, Big Data, ETL, przetwarzanie danych**

Keywords: **Apache NiFi, Big Data, data processing, ETL**

1. WSTĘP

Nieustannie postępująca komputeryzacja oraz rozwój i popularyzacja usług świadczonych za pośrednictwem sieci Internet przekładają się na generowanie i konsumpcję coraz większej ilości danych cyfrowych. Zbadaniem całego rynku, a w szczególności ilości informacji tworzonych, modyfikowanych oraz konsumowanych w ujęciu rocznym, zajęła się organizacja Statista. W opracowanym raporcie można przeczytać, że w 2010 roku były to 2 zetabajty informacji w skali globalnej, 2015 już 15,5 ZB, a w 2020 roczna wartość wzrosła o kolejne 50 ZB [10]. Według prognoz, w roku 2025 przekroczona zostanie granica 180 ZB. Taka dynamika wzrostu w połączeniu z wolumenem danych, stanowią motywację do powstawania nowych gałęzi przemysłu, jak np. Big Data zajmująca się analizą zbiorów danych, szukaniem korelacji między nimi i formułowaniem na tej podstawie wniosków. Szacunkowa wartość tego sektora w roku 2023 może wynieść aż 103 miliardy dolarów [8]. Konieczne są więc narzędzia pozwalające na sprawne manipulowanie danymi, ponieważ ręczne analizowanie zbiorów wielkości nawet kilku gigabajtów może być zbyt czasochłonne. Jednym z dostępnych rozwiązań jest, rozwijany przez fundację Apache, narzędzie NiFi służące do automatyzacji przetwarzania oraz dystrybucji danych.

Apache NiFi można zdefiniować jako oprogramowanie do pobierania, przetwarzania oraz ładowania danych (ang. ETL - *Extract, Transform and Load*). Typowym scenariuszem użycia jest pobranie surowych danych, a następnie przystosowanie ich do formatu obsługiwanego przez odbiorcę. Przykładem może być odpytanie czujnika temperatury o wynik pomiaru i konwersja uzyskanej odpowiedzi do zapytania bazodanowego. Poszczególne etapy realizowane są w sposób przepływowy. Oznacza to, że gotowa konfiguracja przyjmuje postać ścieżki przetwarzającej składającej się z wyspecjalizowanych modułów odpowiedzialnych za wprowadzanie zmian. Moduły są ze sobą połączone kolejkami umożliwiającymi przechowanie kwantów danych. Cały system działa w sposób asynchroniczny, a wykonywane w modułach procesy są od siebie niezależne i mogą być wykonywane równolegle. Do dyspozycji użytkownika oddano ponad 250 gotowych modułów przetwarzających, a dodatkowo istnieje możliwość zdefiniowania nowych, realizujących niestandardowe zadania. Dużą zaletą Apache NiFi jest możliwość połączenia kilku węzłów obliczeniowych w jeden klaster, co pozwala na zwiększenie wydajności przetwarzania danych bez konieczności wprowadzania zmian w ścieżce przetwarzającej. Poniżej przedstawiono podstawową terminologię i parametry odnoszące się do Apache NiFi:

FlowFile, reprezentuje pojedynczy kwant danych składający się z dwóch elementów: zawartości oraz atrybutów (metadanych w postać par klucz - wartość).

FlowFile Processor, moduł wykonujący określone zadanie na danych wejściowych, wynik przesyła na jedno lub więcej wyjść. Ma on pełny dostęp do atrybutów oraz treści *FlowFile*. W danej chwili może obsługiwać pojedyncze pliki lub ich grupy. Przykłady predefiniowanych modułów to *LogMessage* służący do zapisywania komunikatów do dziennika zdarzeń, *UpdateAttribute* modyfikujący atrybuty oraz *ReplaceText* służący do zmiany zawartości *FlowFile*.

Connection, połączenie pomiędzy modułami przetwarzającymi, które pełni też rolę kolejki pozwalając łączyć bloki pracujące z różną szybkością. Oprócz tego wspiera obsługę *FlowFile* z różnym priorytetem oraz w różnych węzłach klastra w celu równoważenia obciążenia.

Concurrent Tasks, liczba wątków przypisana danemu modułowi przetwarzającemu w ramach jednego węzła, domyślnie 1.

Run Duration, określa jak długo moduł przetwarzający może obsługiwać kolejne *FlowFile* bez potrzeby zapisywania zmian. Większe wartości skutkują wzrostem przepustowości ale kosztem większych opóźnień. Domyślną wartością jest 0 ms, która oznacza, że moduł przetwarzający zapisuje wprowadzone zmiany dla każdego *FlowFile* osobno.

2. STATE-OF-THE-ART

Apache NiFi jest narzędziem dojrzałym, które stanowiło podstawę wielu projektów. Począwszy od projektu mającego na celu zbudowanie stabilnego systemu ładowania danych [4], a skończywszy na bardziej wymagających, jak np. analiza danych przestrzennych [5] czy monitorowanie bezpieczeństwa pieszych [3]. Ponadto Apache NiFi jest często stosowany na potrzeby Internetu rzeczy (ang. *Internet of Things*, IoT) [2], [1]. Interesujące są również zastosowania w analizie sentymentu [7] oraz jako jeden z elementów składowych projektu MONSOON, mającego na celu ułatwienie procesu optymalizacji produkcji przemysłowej [9]. Alternatywne do Apache NiFi rozwiązania, takie jak Azure Synapse oraz Azure Data Factory, opisano w artykule [6].

3. BADANIA I UZYSKANE WYNIKI

W ramach tej pracy zaproponowano oryginalną modyfikację konfiguracji Apache NiFi w celu poprawy szybkość przetwarzania danych. Wszystkie eksperymenty przeprowadzono na zbiorze danych CORD-19[1]. Dodatkowo w pracy zbadano wpływ liczby węzłów w klastrze na wydajność Apache NiFi. Scenariusze testowe obejmowały konfiguracje składające się ze zmiennej liczby hostów. Szczegółowa dokumentacja oraz pliki konfiguracyjne są dostępne w publicznym repozytorium git[2].

Zadaniem realizowanym przez badaną instancję Apache NiFi było wyodrębnienie unikatowej listy prac naukowych cytowanych w artykułach wchodzących w skład CORD-19. Pozyskiwane były między innymi, tytuł, autorzy oraz data publikacji. Dane wejściowe były w formacie json, a wyniki zapisywano jako pliki csv. Opracowaną ścieżkę przetwarzającą podzielono na trzy części. Pierwsza odpowiadała za załadowanie danych, druga dokonywała właściwego przetwarzania, a ostatnia zapisywała wyniki. Z uwagi na czytelność i ograniczone miejsce, szczegółowe rysunki przedstawiające konfiguracje poszczególnych modułów przetwarzających oraz połączenia między nimi umieszczono w repozytorium git[2].

W celu osiągnięcia szybszego przetwarzania, wprowadzono stosowne zmiany w ustawieniach modułów przetwarzających. Zmodyfikowano parametry *Concurrent Tasks* oraz *Run Duration*. Poszczególne wartości umieszczono w Tabeli 1 (domyślne ustawienia to odpowiednio 1 oraz 0 ms). Kolejność wierszy odpowiada miejscu w ścieżce przetwarzającej. Dodatkowo w obrębie grupy modułów odpowiedzialnej za ładowanie danych zmniejszono maksymalny rozmiar kolejek z 1 GB do 200 MB w celu oszczędzania pamięci RAM. Natomiast w grupie odpowiedzialnej za przetwarzanie danych podniesiono limit buforowanych *FlowFile* z domyślnych 10 000 do 30 000. Zaproponowane usprawnienia są oryginalnym wkładem niniejszej publikacji. Poszczególne ustawienia dobrano eksperymentalnie i iteracyjnie, tak aby uzyskać bezbłędną pracę systemu przy jak największej wydajności.

Tabela 1. Zaproponowane wartości parametrów dla poszczególnych modułów przetwarzających.

Moduł przetwarzający	*Concurrent Tasks*	*Run Duration [ms]*
GetFile	10	N/A
NiFi Flow	10	N/A
EvaluateJsonPath	10	500
UpdateAttrinute	10	500
JoltTransformJSON	7	1000
EvaluateJsonPath	2	1000
SplitJson	3	500
EvaluateJsonPath	10	2000
ExtractAuthors	5	2000
UpdateAttribute	5	2000
ReplaceText	10	2000
HashContent	5	2000
DetectDuplicate	12	2000
ReplaceText	10	2000
MergeContent	5	N/A
ReplaceText	5	50
UpdateAttribute	1	0
PutFile	5	25

Wszystkie badania wykonano z wykorzystaniem maszyn wirtualnych. Każda z nich miała do dyspozycji dwa rdzenie procesora AMD Ryzen 9 5900X o średnim taktowaniu 4,4 GHz, 4256 MB pamięci operacyjnej o taktowaniu 3600 MHz oraz 60 GB przestrzeni dyskowej (dysk NVMe PCIe 3.0). Zbiór testowy oraz wynikowy przechowywany był we wspólnym dla wszystkich hostów folderze udostępnionym (dysk NVMe PCIe 4.0). Użyto systemu operacyjnego Ubuntu Linux 20.04 LTS.

Monitorowanymi parametrami były *Load Average* (średnie obciążenie systemu Linux), procentowe wykorzystanie procesora, zajętość pamięci operacyjnej

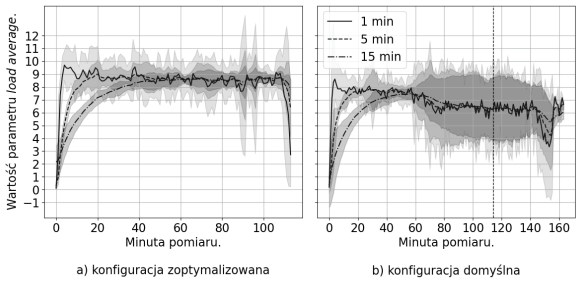
[1] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[2] https://github.com/karolwn/Apache_NiFi

oraz czas przetwarzania pojedynczego pliku i całego zbioru danych. Badano scenariusze dla klastrów złożonych z liczby węzłów zmieniającej się od 1 do 5 (z krokiem 1). Szczegółowe wyniki zaprezentowano jedynie dla scenariuszy skrajnych (1 i 5 węzłów w klastrze). Przedstawione wyniki powstały z uśrednienia ośmiu powtórzeń. Na wykresach zaznaczono 95% przedziały ufności, a w Tabeli 2, ze względu na dużą zmienność parametrów obciążenia systemu widoczną szczególnie w przypadku 5 węzłów, 68%.

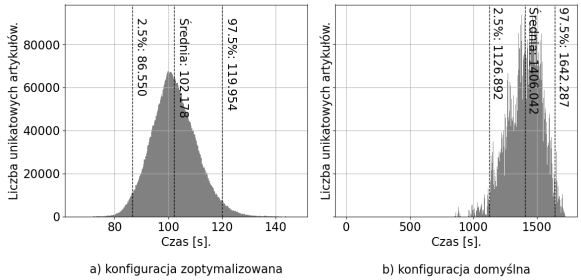
3.1. Jeden węzeł

Rysunek 1 przedstawia zmiany wartości parametru *Load Average* w czasie przetwarzania danych. Poszczególnymi stylami linii oznaczono jego wyniki za okres jednej, pięciu oraz piętnastu minut. Pionowa przerywana linia bez etykiety na wszystkich wykresach oznacza moment zakończenia przetwarzania zbioru wejściowego przez zoptymalizowany węzeł na tle wyników dla węzła domyślnego. Wartości *Load Average* węzła zoptymalizowanego oscylują wokół wartości 8,5, a w przypadku domyślnym jest to wartość 7. Różnica ta wynika z wyższego średniego zużycia procesora węzła dla konfiguracji zaproponowanej w tym artykule. Wahania jednominutowego *Load Average* są, z dużym prawdopodobieństwem, spowodowane ustawieniami planistów poszczególnych modułów przetwarzających, które w rezultacie nie pracują w sposób ciągły. Dzięki przedziałom ufności przedstawionym na Rysunku 1 możliwe jest zobrazowanie niestabilnej pracy infrastruktury w przypadku scenariusza bazowego. Dla domyślnej konfiguracji wartości *Load Average* za okres 15 minut pomiędzy 60-tą, a 160-tą minutą przetwarzania zawierają się w przedziale od 4 do 9. Taka rozbieżność spowodowana była całkowitym zapełnieniem się kolejek, co skutkowało przerwaniami pracy modułów przetwarzających. Dzięki zaproponowanej optymalizacji udało się wyeliminować te problemu i doprowadzić do poprawnej i stabilnej pracy systemu.



Rys. 1. Parametr Load Average w funkcji czasu przetwarzania zbioru wejściowego (1 węzeł).

Poszczególne czasy przebywania *FlowFile* w systemie, tzn. od momentu załadowania z dysku do przekazania do zapisu, przedstawiono na Rysunku 2. Autorskie poprawki ścieżki przetwarzającej pozwoliły na zredukowanie średnich czasów przetwarzań o jeden rząd wielkości z 1 406 do 102 sekund.



Rys. 2. Rozkład czasu potrzebnego do wyodrębnienia pojedynczego, cytowanego artykułu (1 węzeł).

Instancja Apache NiFi poddana optymalizacji wykorzystywała zasoby procesora serwera o około 30 punktów procentowych bardziej efektywnie w przypadku użycia jednego węzła. Natomiast wykorzystanie pamięci operacyjnej pozostało na zbliżonym poziomie. Należy to tłumaczyć tym, że maszyny wirtualne Java (ang. *JVM*) miały do dyspozycji taką samą ilość pamięci RAM. Szczegółowe wyniki umieszczono w Tabeli 2. Wysokie zużycie CPU przez procesy systemowe w przypadku konfiguracji domyślnej jest związane z licznymi operacjami wejścia/wyjścia, które wywołane były przez przeciążone kolejki. Wynika to z faktu, że Apache NiFi starając się oszczędzać pamięć operacyjną dokonuje zapisu *FlowFile* do pliku na dysku. Przepełnienie kolejek nie występowało w konfiguracji zoptymalizowanej.

Tabela 2. Wyniki numeryczne.

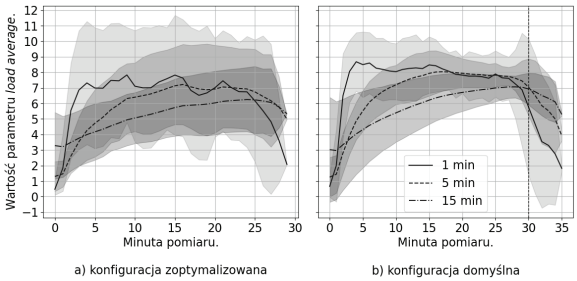
Nr węzła	*CPU - użytkownik [%]*	*CPU - system [%]*	*Wykorzystany RAM [MB]*
jeden węzeł - konfiguracja zoptymalizowana			
1	81.36 ± 11.65	13.01 ± 1.54	3005.48 ± 125.98
jeden węzeł - konfiguracja domyślna			
1	49.25 ± 6.37	30.1 ± 6.14	3006.04 ± 105.79
pięć węzłów - konfiguracja zoptymalizowana			
1	58.18 ± 18.88	13.36 ± 4.05	2939.28 ± 225.73
2	57.75 ± 19.17	12.55 ± 3.35	2725.8 ± 247.69
3	60.96 ± 17.13	12.85 ± 3.85	2734.32 ± 268.28
4	59.38 ± 19.96	13.23 ± 4.38	2729.3 ± 261.57
5	59.0 ± 18.81	13.79 ± 4.29	2736.59 ± 264.59
średnia	59.05 ± 18.79	13.15 ± 3.99	2773.06 ± 253.57
pięć węzłów - konfiguracja domyślna			
1	41.58 ± 21.67	22.83 ± 13.6	2990.34 ± 183.84
2	51.2 ± 10.69	22.58 ± 13.75	2666.11 ± 226.08
3	47.91 ± 11.57	21.97 ± 13.17	2628.47 ± 233.14
4	38.34 ± 21.47	22.24 ± 13.48	2628.78 ± 234.11
5	49.63 ± 10.74	22.96 ± 12.59	2641.24 ± 227.66
średnia	45.73 ± 15.23	22.52 ± 13.32	2710.99 ± 220.97

3.2. Pięć węzłów

Uruchomienie klastra złożonego z pięciu węzłów przyniosło spadek średniego zużycia procesora przez procesy użytkownika o około 20 punktów procentowych w przypadku konfiguracji

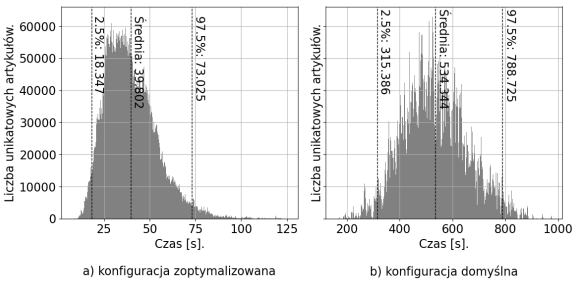
zoptymalizowanej, natomiast dla klastera domyślnego nie odnotowano istotnych zmian. Należy zauważyć, że zmienność wykorzystania zasobów jest tak duża, iż zatarły się statystyczne różnice w przypadku wykorzystania CPU przez użytkownika pomiędzy konfiguracjami. Szczegółowe wyniki również umieszczono w Tabeli 2. Wykorzystanie pamięci operacyjnej w obydwu konfiguracjach spadło o około 300 MB względem scenariusza z jednym węzłem.

Klaster składający się z pięciu zoptymalizowanych węzłów przetwarzających charakteryzował się szerszymi przedziałami ufności parametru *Load Average* niż ten z ustawieniami domyślnymi, który dodatkowo posiadał wyższe średnie wartości (Rysunek 3). Wynika to z wolnych mocy przerobowych systemu dopracowanego, co w połączeniu z krótszym całkowitym czasem przetwarzania pozwala równocześnie stwierdzić, że klaster bazowy był przeciążony.



Rys. 3. Parametr Load Average w funkcji czasu przetwarzania zbioru wejściowego (5 węzłów).

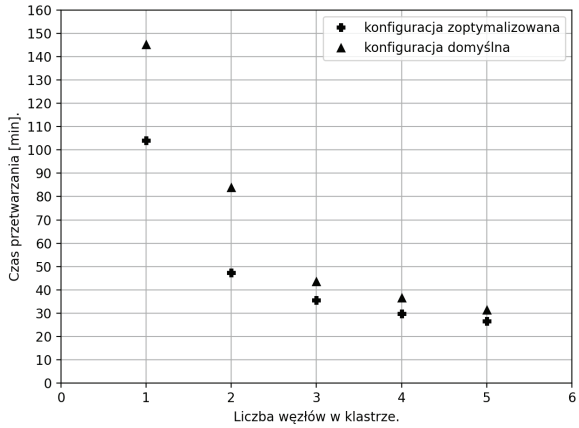
Uruchomienie klastra składającego się z pięciu węzłów przetwarzających pozwoliło na znaczne obniżenie wartości średniego czasu niezbędnego do uzyskania jednego wpisu wynikowego (Rysunek 4). W przypadku zoptymalizowanym była to redukcja względem instancji jednowęzłowej o około minutę, a w domyślnym aż o 14 minut. Mimo to nadal autorska konfiguracja ma istotną przewagę. Klaster dopracowany potrzebował średnio około 40 sekund, a domyślny aż 534 sekund na wyodrębnienie jednego wpisu wynikowego.



Rys. 4. Rozkład czasu potrzebnego do wyodrębnienia pojedynczego, cytowanego artykułu (5 węzłów).

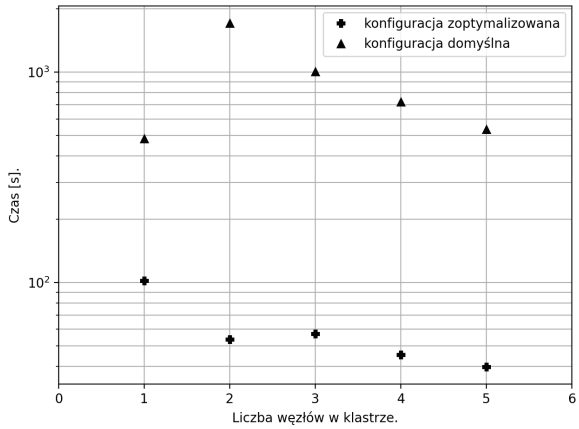
3.3. Wyniki zbiorcze

Rysunek 5 przedstawia całkowity czas przetwarzania zbioru wejściowego w zależności od liczby węzłów w klastrze. Zarówno w konfiguracji domyślnej jak i zoptymalizowanej największy zysk dało włączenie do klastra drugiego węzła. Pozwoliło to na około dwukrotne przyspieszenie przetwarzania. Dodawanie kolejnych instancji nie niosło za sobą już tak dużego zysku. Wynika to najprawdopodobniej ze zwiększonego narzutu wymaganego do zarządzania klastrem i równoważenia obciążenia. Co ciekawe, autorska konfiguracja pozwoliła na przetworzenie zbioru wejściowego w czasie zbliżonym do tego, jaki potrzebował klaster złożony z *n+1* węzłów w konfiguracji domyślnej. Interesujące jest także zmniejszenie różnicy pomiędzy konfiguracją zoptymalizowaną i domyślną po włączeniu trzeciego węzła. Świadczy to o tym, że przy mniejszej liczbie hostów klaster domyślny był przeciążony.



Rys. 5. Sumaryczny czas przetwarzania całego zbioru.

Pomimo stosunkowo niedużych różnic w całkowitych czasach przetwarzania zbioru wejściowego. Czas wymagany do obsłużenia pojedynczego *FlowFile* wykazał istotnie większą rozbieżność, na poziomie rzędu wielkości (Rysunek 6). Jest to rezultat długiego oczekiwania *FlowFile* w kolejkach przeciążonego systemu w konfiguracji domyślnej.



Rys. 6. Średnie czasy przetwarzania pojedynczych plików.

4. PODSUMOWANIE

Przedstawione w pracy wyniki pokazują istotną poprawę wydajności ścieżki przetwarzania danych na skutek zaproponowanych modyfikacji parametrów. Dzięki autorskiej konfiguracji udało się uzyskać krótsze czasy przetwarzania pojedynczej jednostki danych i całego zbioru, a także znacznie bardziej efektywne wykorzystanie dostępnych zasobów. Dodatkowo wykazano, że czas przetwarzania danych przez Apache NiFi maleje nieliniowo wraz ze wzrostem liczby węzłów w klastrze, a ta zmiana jest zależna od konfiguracji parametrów.

Analiza wykorzystania procesora przez proces użytkownika odpowiedzialny za przetwarzanie danych może służyć na potrzeby oceny systemu jako przeciążony (konfiguracja domyślna) lub dający potencjał do dalszej poprawy osiągów (konfiguracja autorska).

Potencjalne kierunki dalszych prac to badanie udziału poszczególnych modułów w wykorzystaniu zasobów oraz analiza zależności od użytej platformy sprzętowej i zbioru wejściowego. Dodatkowo można dokonać implementacji automatycznego skalowania klastra w zależności od obciążenia. Wymagałoby to użycia dodatkowych narzędzi takich jak Ansible, czy Terraform, ponieważ NiFi nie zapewnia takiej funkcjonalności. Przy czym należy zaznaczyć, że jakakolwiek zmiana w klastrze wymaga jego pełnego restartu wraz z podmianą odpowiednich plików konfiguracyjnych, co potencjalnie może utrudnić skalowanie zasobów z czasie rzeczywistym.

LITERATURA

[1] Somphop Chanthakit, Phongsak Keeratiwintakorn, i Choopan Rattanapoka. 2019. „An iot system design with real-time stream processing and data flow integration". *Research, Invention, and Innovation Congress (RI2C)*: 1–5, 2019. doi: 10.1109/RI2C48728.2019.8999968.

[2] Rustem Dautov and Salvatore Distefano. „Stream processing on clustered edge devices". *IEEE Transactions on Cloud Computing*: 1–1, 2020. doi: 10.1109/TCC.2020.2983402.

[3] Daniel Gonzalez, Gerardo Granados, Juan Battini, Russell Carter, Tom Nguyen, Jungsoo Lim, and Russell Abbott. 2019. „Case study: Environmental safety monitoring system for sidewalk". *8th Mediterranean Conference on Embedded Computing (MECO)*: 1–4. doi: 10.1109/MECO.2019.8760037.

[4] H. Isah and F. Zulkernine. 2018. „A Scalable and Robust Framework for Data Stream Ingestion". *2018 IEEE International Conference on Big Data (Big Data)*: 2900-2905, doi: 10.1109/BigData.2018.8622360.

[5] Sang-Su Kim, Wang-Ro Lee, and Jun-Hui Go. 2019. „A study on utilization of spatial information in heterogeneous system based on Apache NiFi". *International Conference on Information and Communication Technology Convergence (ICTC)*: 1117–1119, 2019. doi: 10.1109/ICTC46691.2019.8939734.

[6] Vitalii Mayuk, Ivan Falchuk, and Piotr Muryjas. 2021. „The comparative analysis of modern etl tools". *Journal of Computer Sciences Institute*. 19: 126–131, doi: 10.35784/jcsi.2631.

[7] Abhinay Pandya, Panos Kostakos, Hassan Mehmood, Marta Cortes, Ekaterina Gilman, Mourad Oussalah, and Susanna Pirttikangas. „Privacy preserving sentiment analysis on multiple edge data streams with Apache NiFi". 2019. *European Intelligence and Security Informatics Conference (EISIC)*: 130–133. doi: 10.1109/EISIC49498.2019.9108851.

[8] Christo Petrov. 25+ impressive big data statistics for 2021. URL https://techjury.net/blog/big-data-statistics/. data dostępu: 2021-12-11.

[9] M. Sarnovsky, P. Bednar, and M. Smatana. 2017. „Data integration in scalable data analytics platform for process industries". *IEEE 21st International Conference on Intelligent Engineering Systems (INES)*: 187–192. doi: 10.1109/INES.2017.8118553.

[10] Statista. „Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 (in zettabytes)". URL https://www.statista.com/statistics/871513/worldwide-data-created/. data dostępu: 2021-12-11.

EFEKTYWNOŚĆ REPLIKACJI DANYCH W CENTRACH DANYCH
EFFICIENCY OF DATA REPLICATION IN DATA CENTERS

Piotr Pyda, Joanna Śliwa, Tomasz Dalecki, Michał Przywuski

Wojskowy Instytut Łączności - Państwowy Instytut Badawczy, Zegrze
{p.pyda; j.sliwa; t.dalecki; m.przywuski}@wil.waw.pl

DOI: 10.15199/59.2022.4.19

Streszczenie: **Centra danych przechowują dane dla potrzeb działania usług, aplikacji z których korzystają użytkownicy. W celu zwiększenia niezawodności w dostępie do danych oraz w obawie przed uszkodzeniem, zniszczeniami, atakami cybernetycznymi dane z serwerów replikuje się na inne centra, serwery, instancje. W artykule przedstawiono przykładowe wyniki z rzeczywistego systemu, w którym zbadano wpływ przepustowości łączy pomiędzy centrami danych na czasy replikacji.**

Abstract: **Data centers store the necessary data for the operation of services and applications used by users. In order to increase the reliability of data access and to protect the system from damage or destruction, data from servers are replicated to other centers, servers and instances. The article presents examples of results from a real system in which the influence of bandwidth between data centers on replication times was examined.**

Słowa kluczowe: **centra danych, maszyny wirtualne, replikacja danych.**

Keywords: **data centers, data replication, virtual machines.**

1. WSTĘP

Centra danych są obecnie ważnym elementem systemów korporacyjnych. Projekt centrum danych opiera się na sieci zasobów obliczeniowych i pamięci masowej, które umożliwiają dostarczanie współdzielonych aplikacji i danych. Nowoczesny projekt centrum danych obejmuje serwery, systemy pamięci masowej, routery, przełączniki, zapory sieciowe i kontrolery dostarczania aplikacji. Razem zapewniają:

- infrastrukturę sieciową, która łączy serwery (fizyczne i zwirtualizowane), usługi centrum danych, pamięć masową i łączność zewnętrzną z lokalizacjami użytkowników końcowych;
- infrastrukturę pamięci masowej do wydajnego przechowywania danych;
- zasoby obliczeniowe. Serwery zapewniają przetwarzanie, zasoby dyskowe, pamięć lokalną i łączność sieciową, które są wykorzystywane przez aplikacje.

Komponenty te zwykle przechowują i zarządzają danymi i aplikacjami o znaczeniu krytycznym dla firmy, dlatego bezpieczeństwo centrum danych i odtwarzanie po awarii mają kluczowe znaczenie w projektowaniu centrum danych. Aby zapewnić niezawodność i odporność na uszkodzenia, w wielu praktycznych zastosowaniach dane są zazwyczaj replikowane w wielu kopiach i przechowywane na więcej niż jednym serwerze. Chociaż idealnie byłoby replikować dane w jak największej liczbie kopii, ograniczenie pojemności pamięci masowej serwerów jest wąskim gardłem.

Aby zapewnić lepsze wrażenia użytkownika, centrum danych powinno starać się obsługiwać maksymalną liczbę jednoczesnych żądań danych, biorąc pod uwagę popularność danych dla użytkowników oraz pojemność serwera. Chodzi o to, aby dać użytkownikom dokładnie te zasoby i usługi, których potrzebują. Dlatego dzielenie pamięci masowej i wydajności obliczeniowej odbywa się bardzo łatwo dzięki maszynom wirtualnym, które mogą migrować z określonej maszyny na inną i mogą być również replikowane. Operacje te są wykonywane za każdym razem, gdy występuje konserwacja serwera, uszkodzenia sprzętu lub oprogramowania, zniszczenia podstawowego centrum lub ataku cybernetycznego. Aby osiągnąć odporność na uszkodzenia, sugeruje się redundancję sprzętu i oprogramowania. W przypadku maszyny wirtualnej nadmiarowość oznacza replikowanie oryginalnej maszyny wirtualnej na innej maszynie fizycznej. Replikacja dodatkowo ma jeszcze jedną zaletę może powodować równoważenie obciążenia żądań serwerów [1]. Tematyka centrów danych jest szeroko badana. Między innymi w [2] autorzy zbadali, jak umieścić maszyny wirtualne na serwerach, aby spełnić wymagania centrum danych. W [3] autorzy rozważali topologie obciążeń intensywnie wykorzystujących dane i zaproponowali architekturę alokacji zasobów w chmurze. W [4] przeprowadzono badanie efektywności energetycznej centrum danych. Autorzy zaproponowali również zasady architektury, zasady alokacji zasobów i algorytmy planowania maszyn wirtualnych. Prace badawcze w [5] [6] omawiały schematy efektywnego wykorzystania zasobów centrum danych w celu oszczędzania energii.

W artykule skupiono się na badaniu wpływu pojemności łącza na czas replikacji pomiędzy centrami danych. Założono, że działa jedno podstawowe centrum danych i kolejne - zapasowe jest przygotowywane do przejęcia jego roli na wypadek awarii. Autorzy spróbowali odpowiedzieć na pytanie jakie przepustowości między centrami podstawowym i zapasowym są konieczne by warto było przeprowadzać replikację przez sieć, a także ile czasu potrzeba do odtworzenia przykładowych serwerów dla różnych przepustowości łączy międzywęzłowych. W szczególności porównano replikację różnicową wykonaną metodą Site Recovery oraz replikację na poziomie aplikacji. Jako aplikację wykorzystano MS

SQL wraz z dostępnymi mechanizmami replikacyjnymi. Wyniki mogą być wykorzystane do zabezpieczenia odpowiednich łączy przy replikacjach wojskowych centrów danych.

2. BADANY SYSTEM

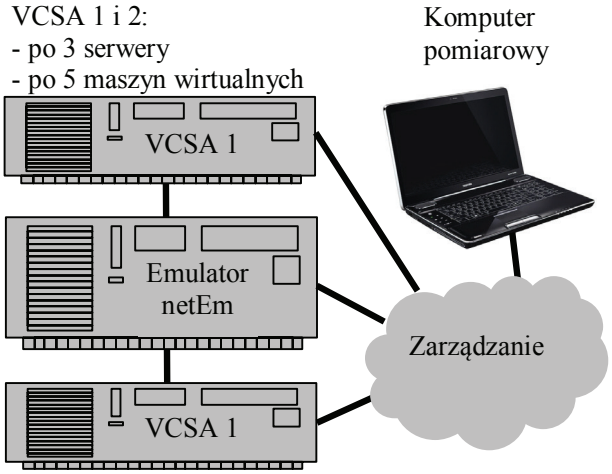
Konfiguracja testowa przedstawiona na rysunku 1 składa się z emulatora łącza netEm oraz z dwóch VCSA (ang. vCenter Server Appliance), których skład jest jednakowy, tj. pojedynczy VCSA zawiera 3 serwery fizyczne dla 5 maszyn wirtualnych, 1 serwer fizyczny jako zasób dyskowy. Zakłada się, że replikacja jest prowadzona z VCSA1 do VCSA 2. Przedstawione w tabeli 1 proponowane zestawy parametrów zawierają ustawienia jakie należy ustawić w emulatorze sieci tak, aby odwzorować warunki rzeczywistej sieci łączności. Tabela zawierają minimalne parametry jakich należy się spodziewać w rzeczywistej sieci przewodowej o zasięgu około 50km oraz radioliniowej o zasięgu około 25km. Opóźnienie w tabeli jest sumą opóźnienia propagacji zależne od długości łącza oraz opóźnienia transmisji zależnego od szybkości bitowej i długości ramki.
Jako emulator łącza danych: Centos 8 pracujący z jądrem zawierającym emulator netEm. System skonfigurowano do pracy jako router, który przekazuje pakiety pomiędzy interfejsami fizycznymi wymuszając parametry odpowiadające poszczególnym łączom transmisyjnym (patrz Tab.1).
Przyjęto następujące założenia:

- konfiguracja VCSA składa się z równocześnie pracujących 5 maszyn wirtualnych, każda z nich posiada własną bazę danych,
- średni zapis do bazy danych: 100 rekordów co 1 sek., średnia wielkość jednego rekordu: 1 kB surowych danych (wielkość rekordu w bazie danych może być większa, ze względu na wprowadzanie przez bazę danych swojego narzutu),
- czas przerwy w połączeniu pomiędzy replikowanymi maszynami wirtualnymi na VCSA 1 i VCSA 2 około 1 godz.

Tab. 1 Zestawy parametrów do emulatora netEm

Zestaw parametrów	*Szybkość bitowa [Mbit/s]*	*Opóźnienie [ms]*	*Jitter [ms]*	*Straty [%]*
Kabel 0	10000	0,25	0,001	10^{-12}
Kabel 1	1000	0,25	0,001	10^{-12}
Kabel 2	100	0,26	0,02	10^{-12}
Kabel 3	10	0,59	0,4	10^{-12}
Radiolinia 1	34	0,48	0,02	10^{-6}
Radiolinia 2	8	1,58	0,4	10^{-6}
Radiolinia 3	2	5,87	1,0	10^{-6}

Badania zostały wykonane z użyciem replikacji aplikacji MS SQL oraz przyrostowej replikacji maszyn wirtualnych VMware, Zbadano czas potrzebny na replikację oraz porównano, który mechanizm replikacji jest szybszy. Dla VMware wykorzystano mechanizmu VMware Site Recovery
Badanie przeprowadzono dla różnych łączy pomiędzy replikowanymi urządzeniami zgodnie z tabelą 1.



Rys 1. Konfiguracja testowa do zbadania czasu replikacji

Do zarządzania testami wykorzystano laptop z systemem Windows 10 Professional z zainstalowanym następującym oprogramowaniem:

- Hyper-V manager – konsola MMC (ang. Microsoft Management Console) do zarządzania wirtualizatorem Hyper-V, który umożliwia m.in. tworzenie maszyn wirtualnych, tworzenie wirtualnych przełączników oraz konfigurację mechanizmu replikacji.
- Zdalny pulpit umożliwiający podłączenie do serwerów Hyper-V, gdzie udostępniono konsolę Powershell.
- Powershell – umożliwia odczyt logów serwerów Hyper-V.
- Klient SSH – do pracy zdalnej na systemach Centos 8.
- Zabbix – oprogramowanie do monitorowania infrastruktury teleinformatycznej. Wykorzystywane do odczytu liczników przesłanych danych. Odczyt wykonywany był z przełączników VDX co 5s poprzez protokół SNMP. Odczyty były wizualizowane na wykresach wykorzystania łącza w czasie.

Dla potrzeb badań opracowano oprogramowanie pomocnicze do generacji i weryfikacji danych testowych oraz oprogramowanie umożliwiające konfigurację symulatora łącza danych:

- GenFiles.sh – skrypt bash do generacji plików z danymi pseudolosowymi. Parametry: ścieżka, liczba plików, wielkość pojedynczego pliku. Wynikiem jest utworzenie plików z danymi pseudolosowymi oraz pliku z listą skrótów SHA256 utworzonych plików. Lista pozwala zweryfikować poprawność replikacji. Dla ułatwienia badań plik z listą skrótów jest umieszczany na partycji przeznaczonej do replikacji, tak aby można było łatwo zweryfikować poprawność replikacji bez potrzeby dodatkowego kopiowania pliku.
- kabel0.sh, kabel1.sh, kabel2.sh, kabel3.sh oraz radio1.sh, radio2.sh i radio3.sh – skrypty bash służące do ustawiania parametrów symulatora łącza danych zgodnie z profilami opisanymi w Tab. 1.

3. WYNIKI

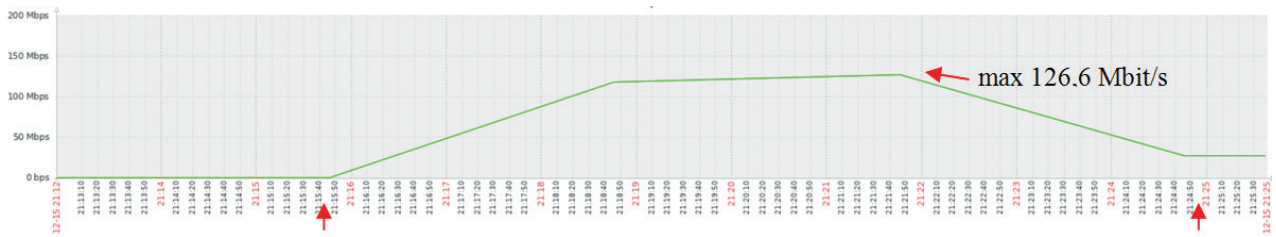
Przebieg badania obejmował konfigurację klastrów VCSA oraz ich synchronizację. Następnie klastry zostały odłączone od siebie i następowała generacja rekordów do bazy danych w każdej maszynie wirtualnej. Po wygenerowaniu w ciągu godziny około 360 tys. rekordów, każdy o średniej wielkości 1 kB, następowało podłączenie drugiego klastra za pomocą łącza o parametrach jak w tabeli 1 oraz włączeniu odpowiedniego trybu replikacji Site Recovery lub replikacji MS SQL. Następnie zmierzono czas, po jakim maszyny z VCSA 1 były zsynchronizowane na VCSA 2. Po zsynchronizowaniu maszyny były gotowe do przełączenia

Wyniki pomiarów odnoszą się do środowiska badawczego, w którym wykorzystuje się serwery firmy Dell R430, wirtualizatory Vmware ESXI v.7 oraz maszyny wirtualne z systemami operacyjnymi Windows Server 2019 z interfejsem graficznym i aplikacjami MS SQL. Jako oprogramowanie do replikacji przyrostowej zostało użyte Site Recovery v.8. Jako replikację aplikacyjną MS SQL użyto replikacji transakcyjnej.

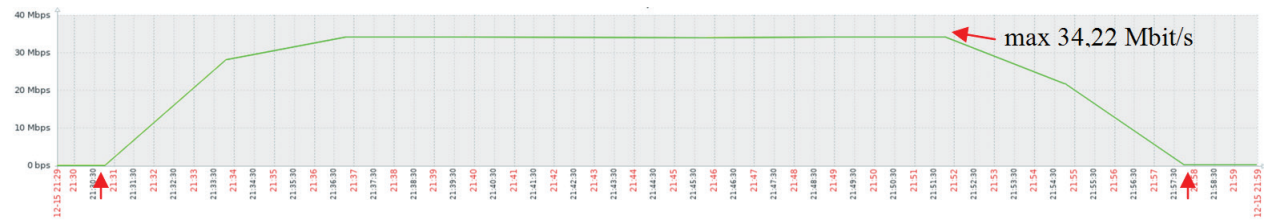
Otrzymane wyniki dla obu wariantów pomiarów zostały zamieszczone w tabeli 2. Dodatkowo, na rysunkach od Rys. 2 do Rys. 5 zostały przedstawione przykładowe wykresy zajętości łącza dla replikacji Site Recovery oraz dla replikacji aplikacyjnej MS SQL, które otrzymano przy wykorzystaniu oprogramowania Zabbix.

Tab. 2 Wyniki pomiarów czasu replikacji aplikacji MS SQL i 5 maszyn wirtualnych VMware

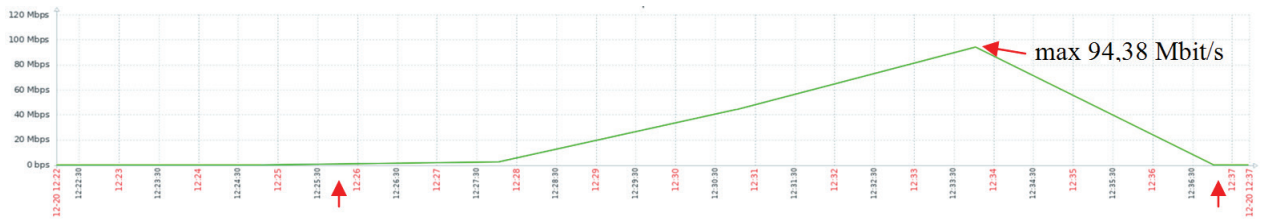
Zestaw parametrów łącza transmisyjnego	*Czas replikacji Site Recovery [hh:mm:ss]*	*Czas replikacji na poziomie aplikacji MS SQL [hh:mm:ss]*
Kabel 0	0:08:12	0:11:45
Kabel 1	0:08:17	0:11:55
Kabel 2	0:09:04	0:15:00
Kabel 3	1:19:13	1:03:00
Radiolinia 1	0:22:29	0:24:00
Radiolinia 2	1:37:12	1:12:00
Radiolinia 3	Powyżej 8 godz.	3:42:00



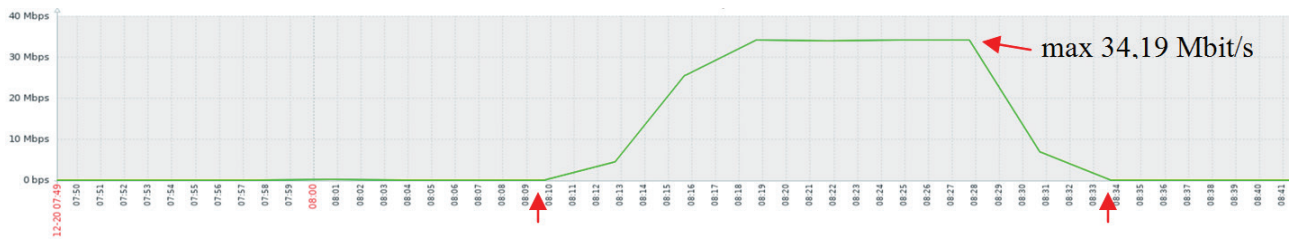
Rys. 2. Wykres zajętości łącza dla replikacji Site Recovery, typ łącza: Kabel 0



Rys. 3. Wykres zajętości łącza dla replikacji Site Recovery, typ łącza: Radiolinia 1



Rys. 4. Wykres zajętości łącza dla replikacji aplikacyjnej MS SQL, typ łącza: Kabel 0



Rys. 5. Wykres zajętości łącza dla replikacji aplikacyjnej MS SQL, typ łącza: Radiolinia 1

4. WNIOSKI

Dla pomiarów Site Recovery zaobserwowano zbliżone czasy replikacji dla Kabel 0, Kabel 1 i Kabel 2 ze względu na to, że ruch nie był w stanie wykorzystać pełnego dostępnego łącza. Z wykresów przedstawionych na rysunkach 2 i 3 widać, iż maksymalna szybkość transmisji dla Kabel 0 wynosiła 126 Mbit/s, natomiast ograniczenie dla Kabel 2 spowodowało zmniejszenie tej wartości do około 100Mbit/s. Średnia szybkość transmisji wynosiła dla Kabel 0 około 68 Mbit/s, natomiast dla Kabel 2 około 52 Mbit/s. Szybkości dla Kabel 1 są niemal identyczne jak dla Kabel 0, również ruch nie jest w stanie wykorzystać dostępnego łącza. Spośród łączy przewodowych najdłuższy czas replikacji osiągnięto dla Kabel 3, co było do przewidzenia.

W przypadku łączy radioliniowych zebrano wyniki dla dwóch łączy, dla łącza 2 Mbit/s czas replikacji przekraczał 8 godz. w związku z tym, przerwano pomiar jako nieużyteczny w rzeczywistych warunkach. Dla łączy Radiolinia 1 i radiolinia 2 otrzymano odpowiednio 22:29 min. oraz 1:37:12, co jest wynikiem przewidywalnym wynikającym z przepustowości łącza i charakterystyk opóźnieniowych.

Jeżeli weźmiemy pod uwagę wielkość łącza niezbędnego do replikacji, należy odrzucić łącza, które potrzebują więcej niż 1 godz. na replikację, gdyż przy korzystaniu z takiego łącza ciągle mamy nieaktualne dane pomiędzy serwerami - bazami danych. A należy pamiętać, że w rzeczywistym przypadku na wykorzystywanym łączu może pojawić się ruch wynikający z działania innych aplikacji i usług w centrum danych. Zatem łącza Kabel 3, Radiolinia 2 i Radiolinia 3 są niewskazane do wykorzystania w systemie.

Zaobserwowano, że dla szybkich łączy replikacja Site Recovery jest przebiega szybciej od replikacji aplikacyjnej, natomiast w łączach wolniejszych (Kabel 3, Radiolinia 2 i Radiolinia 3) lepiej radzi sobie replikacja aplikacyjna. Jednak jak wspomniano powyżej czas replikacji dla badanego przypadku jest nieakceptowany.

W przypadku replikacji aplikacyjnej podobnie jak w replikacji Site Recovery nie udało się wykorzystać całego dostępnego łącza. Dlatego wyniki dla łączy Kabel 0, Kabel 1 i Kabel 2 są zbliżone i nieco dłuższe niż dla przypadku pierwszego. Dłuższy czas wynika z faktu, iż aplikacja po przerwie w dostępie do sieci i do głównej instancji, najpierw przesyła informacje związane z uzyskaniem adresu IP (DHCP) oraz wymienia wiadomości o aktualnym statusie baz danych, a następnie przesyła brakujące rekordy. Podobnie jak w poprzednim przypadku, przepustowości łącza, dla których uzyskano czas replikacji powyżej 1 godz. nie powinny być brane pod uwagę przy projektowaniu systemu, bowiem przy założonym napływie rekordów replikowana baza ciągle byłaby niezsynchronizowana i wielkość danych do replikacji zwiększała by się w czasie. Zatem należy rozważać łącza o przepustowościach 34 Mbit/s i wyższych.

Czasy odnotowane w tabeli 2 podczas replikacji z wykorzystaniem funkcji Site Recovery odzwierciedlają czas działania mechanizmu wykonywania replikacji. Do pełnego przywrócenia funkcjonalności maszyny wirtualnej należy ją następnie odtworzyć przy pomocy panelu Site Recovery. Operacja zajmuje 14 sekund i nie stanowi obciążenia dla sieci klastra. Maszyna zostanie odtworzona z wykorzystaniem danych z ostatniej udanej synchronizacji. Dopiero po takiej operacji możliwe jest przełączenie pomiędzy bazami MS SQL.

5. PODSUMOWANIE

W artykule przedstawiono wyniki badań replikacji maszyn wirtualnych będących częścią centrów danych. Zbadano czas wykonania replikacji różnicowej dla różnych parametrów łącza pomiędzy replikowanymi centrami. Porównano czas replikacji różnicowej z czasem synchronizacji aplikacji na przykładzie bazy danych MS SQL.

Przyjęte konfiguracje są w przybliżeniu takie jakich można się spodziewać w rzeczywistych warunkach pracy. Badania przeprowadzono dla różnych typów łączy pomiędzy obiektami, które odwzorowują zarówno szybkie łącza światłowodowe bez zakłóceń, jak również łącza bezprzewodowe z wprowadzanymi opóźnieniami i stratami. Łącza bezprzewodowe mogą być odzwierciedleniem, wykorzystania w siłach zbrojnych połączeń radioliniowych np. dla potrzeb wojskowych centrów danych. Jednakże w badaniach nie były wykorzystywane urządzenia utajniające i szyfrujące.

LITERATURA

[1] O. Raluca i P. Florin, „Energy-Efficient Virtual Machine Replication for Data Centers", 17th International Symposium on Parallel and Distributed Computing (ISPDC), 2018, s. 126-132, doi: 10.1109/ISPDC2018.2018.00026

[2] M. Wang, X. Meng and L. Zhang, "Consolidating virtual machines with dynamic bandwidth demand in data centers", INFOCOM 2011 Proceedings IEEE, pp. 71-75, 2011.

[3] G. Lee, N. Tolia, P. Ranganathan and R. H. Katz, "Topology-aware resource allocation for data-intensive workloads", SIGCOMM Comput. Commun. Rev., vol. 41, no. 1, pp. 120-124, Jan 2011.

[4] A. Beloglazov, J. Abawajy and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing", Future Generation Computer Systems, vol. 28, no. 5, pp. 755-768, 2012.

[5] K. Patel, N. Patel, and H. Patel, "Efficient resource allocation strategy to improve energy consumption in cloud data centers," in Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. ACM, 2016, p. 76

[6] K. Zheng, X. Wang, L. Li, and X. Wang, "Joint power optimization of data center network and servers with correlation analysis," in INFOCOM, 2014 Proceedings IEEE, 2014, pp. 2598–2606

AUTOMATYZACJA ZARZĄDZANIA INFRASTRUKTURĄ PULPITÓW WIRTUALNYCH W ŚRODOWISKU VMWARE HORIZON

AUTOMATION OF INFRASTRUCTURE MANAGEMENT OF VIRTUAL DESKTOPS IN THE VMWARE HORIZON ENVIRONMENT

Diana Mirowska[1]; Jerzy Domżał[1]

[1] AGH w Krakowie,
dmirowska@student.agh.edu.pl, jerzy.domzal@agh.edu.pl

DOI: 10.15199/59.2022.4.20

Streszczenie: **Prezentowana praca ma na celu przybliżyć pojęcie automatyzacji zarządzania usługami oferowanymi przez środowisko VMware Horizon w środowisku infrastruktury akademickiej z zastosowaniem dostarczanego API oraz zaproponować gotowy model projektu zarządzania usługami. Szczególny nacisk został położony na automatyzację tworzenia pul maszyn wirtualnych oraz opracowanie wygodnego i niezawodnego interfejsu użytkownika. Dodatkowo przedstawiona analiza pozwala użytkownikom na lepsze zrozumienie mechanizmów działania systemu.**

Abstract: **The presented work aims to introduce the concept of VMware Horizon services automation in an academic infrastructure environment using the provided API and to propose a ready-to-use project model. Particular emphasis is placed on automating the creation of virtual machine pools and preparation of a convenient and reliable user interface. In addition, the presented analysis allows novice users to better understand the mechanisms of system operation.**

Słowa kluczowe: **VMware, wirtualizacja, automatyzacja, zarządzanie infrastrukturą.**

Keywords: **VMware, virtualization, automatization, infrastructure management.**

1. WSTĘP

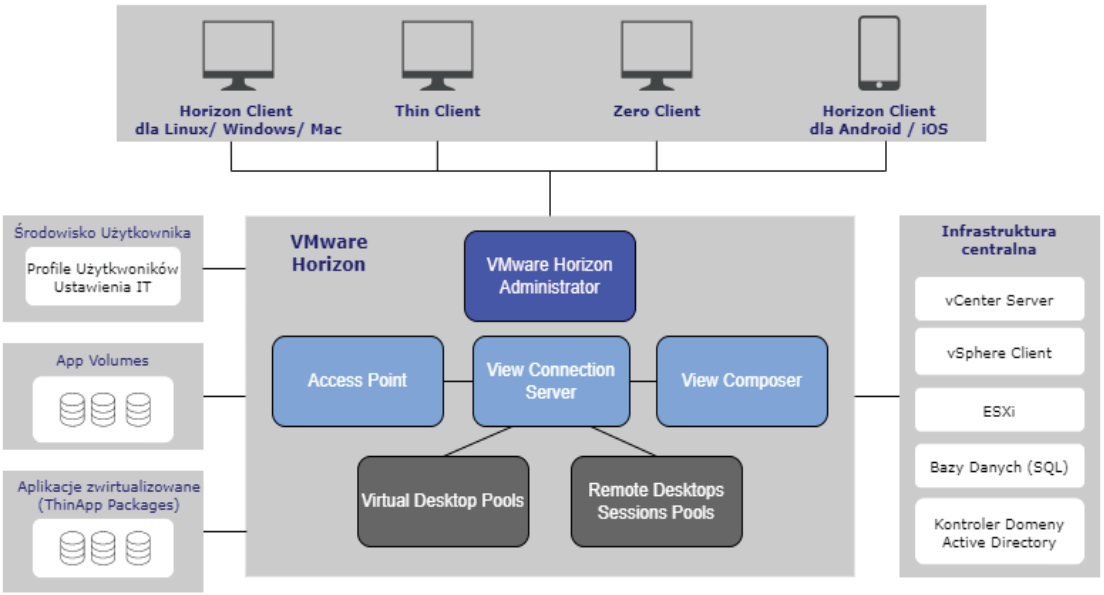
Obecnym trendem w zarządzaniu infrastrukturą sieci korporacyjnych i uczelnianych jest dążenie do maksymalnej centralizacji zasobów i sterowania urządzeniami. Użytkownicy korzystają z udostępnianych zasobów zdalnie w ramach posiadanych uprawnień, co pozwala na lepszą kontrolę ich użycia oraz wczesne wykrywanie zagrożeń i potencjalnych nadużyć. Odpowiednie przygotowanie infrastruktury do wykorzystania zdalnego i automatyzacja zarządzania pozwala na świadczenie usług w modelu Infrastracture as a Service (IaaS) o wysokim poziomie niezawodności oraz najwyższej jakości.

Artykuł ma na celu przybliżyć pojęcie automatyzacji oprogramowania VMware Horizon w środowisku akademickim z zastosowaniem dostarczanego API oraz zaproponować gotowy model infrastruktury. Przedstawione rozwiązanie odpowiada na zapotrzebowanie, jakie istnieje na rynku małych i średnich przedsiębiorstw oraz instytucji edukacyjnych, które nie mogą pozwolić sobie na komercyjne rozwiązania tego typu [2].

Wirtualizacja Pulpitów cieszy się niezmiennie rosnącą popularnością, a trend ten przybrał na sile w związku z wybuchem pandemii wirusa Sars-Cov-2 i koniecznością wdrożenia zdalnego modelu pracy. W klasycznym podejściu fizyczna maszyna wraz z całym oprogramowaniem oraz zasobami sprzętowymi jest udostępniana do użytku zdalnego przy użyciu narzędzi zdalnego pulpitu, takich jak TeamViewer, AnyDesk bądź Microsoft Remote Desktop Protocol (RDP). Jest to rozwiązanie zdecentralizowane, a więc tanie i bezproblemowe we wdrożeniu, wymaga jednak dedykowanych stacji roboczych o odpowiednich zasobach, które rzadko są w pełni wykorzystywane [1], [7].

Odmiennym podejściem cechuje się Infrastruktura Pulpitów Wirtualnych (Virtual Desktop Infrastructure - VDI), która jest środowiskiem zarządzanym centralnie. W rozwiązaniu tym wszystkie pulpity wirtualne znajdują się na serwerze zasobów [3]. Jest to odpowiedź na ciągle rosnące zapotrzebowanie na wysokiej klasy stacje robocze cechujące się dostępnością i mobilnością. Pozwala to na wykorzystanie w sposób optymalny scentralizowanych zasobów znajdujących się na serwerach centrum danych, dzięki czemu możemy uzyskać oszczędności sprzętowe, zmniejszone zużycie energii elektrycznej oraz dostępność indywidualnych pulpitów w dowolnej lokalizacji. Dodatkowo uzyskujemy możliwość wygodnego zarządzania udostępnianymi pulpitami, ich bezpieczeństwem oraz zasobami, których potrzebują użytkownicy.

W artykule omówiono proces automatyzacji zarządzania środowiskiem VDI z wykorzystaniem VMware Horizon – platformy do zarządzania, monitorowania i udostępniania Infrastruktury Pulpitów Wirtualnych. Firma VMware udostępnia konsolowe narzędzie VMware PowerCLI bazujące na języku skryptowym PowerShell oraz od wersji 8 VMware Horizon Server API zgodne ze standardami REST API. O ile powszechnie możemy spotkać gotowe rozwiązania zarówno komercyjne, jak również open source podejmujące tematykę automatyzacji narzędzi VMware takich jak VMware vSphere, na przykład dedykowany moduł Ansible czy dedykowane środowiska SDK dla języka Python o tyle temat automatyzacji pracy z VMware Horizon, szczególnie z zastosowaniem języka Python, nie jest poruszany. Nie są również dostępne gotowe narzędzia ułatwiające pracę ze środowiskiem VMware Horizon.



Rys. 1. Infrastruktura Pulpitów Wirtualnych z wykorzystaniem VMware Horizon.

2. ZAŁOŻENIA PROJEKTOWE

2.1. Struktura rozwiązań VMware

Rozwiązanie omawiane w artykule bazuje na wirtualizacji sprzętowej ESXi zarządzanej centralnie z poziomu serwera VMware vCenter z naciskiem na automatyzację tworzenia pul maszyn wirtualnych przy wykorzystaniu serwera zarządzania VMware Horizon. Wirtualizacja sprzętowa typu 1 bazuje na dostępie Hypervisora do zasobów sprzętowych z pominięciem warstwy systemu hosta [4]. Wykorzystanie oprogramowania VMware Horizon z dodatkiem VMware View Composer daje możliwość tworzenia grup pulpitów wirtualnych o współdzielonych zasobach, wspólnej konfiguracji i przeznaczeniu, które są automatycznie przyłączane do domeny Active Directory. Projekt rozwiązania został wykonany w istniejącym środowisku Instytutu Informatyki na wydziale Informatyki, Elektroniki i Telekomunikacji AGH w Krakowie. Poglądowy schemat wykorzystanej infrastruktury znajduje się na Rysunku 1.

2.2. Schemat systemu automatyzacji

Opracowany system składa się z elementów t.j.:

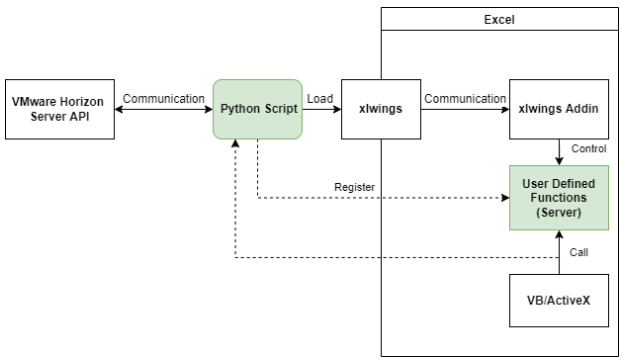
- interfejs użytkownika, który stanowią przygotowane arkusze skoroszytu programu Microsoft Excel,
- skrypty w języku Python obsługujące REST API wystawiane przez VMware Horizon,
- xlwings - dodatek do programu Microsoft Excel będący jednocześnie biblioteką Python.

System działa w następujący sposób: operator systemu zbiera dane od klientów maszyn wirtualnych w formie kwestionariusza Microsoft Forms, a następnie zebrane dane zostają wyeksportowane do arkusza Microsoft Excel w formacie xls. W normalnym trybie pracy z VMware Horizon przy dynamicznie zmieniającym się środowisku laboratoryjnym, konieczne jest żmudne przepisywanie danych z arkusza do wieloetapowego kreatora tworzenia pul. Znacząco wydłuża to proces ich przygotowania i jest obarczone ryzykiem popełnienia błędów.

Dane przechowywane w systemie dotyczące maszyn wirtualnych to przede wszystkim: nazwa puli maszyn wirtualnych, szablon nazewnictwa maszyn wirtualnych, nazwa wyświetlana, informacje o obrazie źródłowym maszyny wirtualnej, typ puli, liczba aktywnych maszyn w puli, ustawienia dotyczące zasobów, ustawienia dotyczące protokołu dostępu do maszyny, domena Active Directory, docelowe miejsce ulokowania maszyny wirtualnej. Połączenie Microsoft Excel - skrypty - serwer VMware Horizon pozwala na wygenerowanie arkuszy odpowiadających poszczególnym funkcjom systemu.

Funkcje zawarte w systemie to: podgląd i odświeżanie stanu pul maszyn wirtualnych na serwerze, aktualizacja konfiguracji pul wirtualnych, utworzenie nowej puli o wybranych parametrach, usunięcie wybranej puli.

Użytkownik ma możliwość wyboru parametrów typu: szablon maszyny wirtualnej, klaster serwerów, domena Active Directory; spośród zasobów dostępnych na serwerze VMware Horizon i pobieranych automatycznie z serwera do arkusza; oraz deklaracji własnych parametrów takich jak np. nazwa puli maszyn wirtualnych.



Rys. 2. Schemat komunikacji systemu.

Na Rysunku 2. przedstawiony został schemat komunikacji elementów systemu podczas wykonywania polecenia utworzenia nowej puli maszyn wirtualnych, a następnie aktualizacji listy pul maszyn w arkuszu Microsoft Excel. Jak widzimy, wywołanie akcji z poziomu interfejsu użytkownika - kliknięcie w arkuszu przycisku powiązanego z makrem Microsoft Excel - uruchamia proces programu xlwings VB. Proces ten uruchamia instancję maszyny wirtualnej Pythona, która wykonuje akcje powiązane z wywołanym makrem - w pierwszej kolejności wysyła zapytanie HTTP z poświadczeniem użytkownika w celu zalogowania się do serwera i uzyskania danych sesji, a następnie wstrzykuje dane sesji do kolejnego zapytania, na przykład create desktop i wysyła je do serwera. Proces Pythona czeka na potwierdzenie zakończenia akcji i kod statusu, przekazuje informacje o poprawnym bądź błędnym wykonaniu akcji i zamyka się.

3. OBSŁUGA API W JĘZYKU PYTHON

Udokumentowane REST API VMware Horizon wersji 7 umożliwia jedynie prostą konfigurację użytkowników oraz monitoring serwera [6]. Nie zezwala jednak na manipulację pulami maszyn wirtualnych, co jest celem zaproponowanego rozwiązania. Konieczna więc była analiza wsteczna (ang. reverse engineering) sposobu komunikacji między przeglądarkowym interfejsem administracyjnym a serwerową częścią VMware Horizon [5]. W artykule przeanalizowano kilka akcji wymaganych do wykonania zadań wybranych do zautomatyzowania: proces logowania, listowanie podsumowania pul maszyn, listowanie głównych (szablonowych) maszyn wirtualnych, tworzenie nowej puli maszyn wirtualnych.

3.1. Logowanie klienta do usługi

Aby wykonać jakiekolwiek operacje przy użyciu REST API na serwerze VMware Horizon wymagane jest uwierzytelnienie i autoryzowanie klienta. W tym celu trzeba wykonać zapytanie podając nazwę użytkownika, hasło oraz domenę, do której dostęp chce mieć klient. Ścieżka służąca do logowania to /view-vlsi/rest/v1/login. Serwer oczekuje zapytania HTTP POST oraz treści w formacie JSON (JavaScript Object Notation). W przypadku powodzenia, serwer zwróci odpowiedź z kodem 200 - OK, a także przekaże ciasteczka z kluczami sesji oraz tokenem zabezpieczającym przed atakami CSRF (Cross-Site Request Forgery) [8]. W naszym scenariuszu wykorzystania, nie ma ryzyka ataku CSRF, ale z uwagi na fakt podszywania się pod przeglądarkowy interfejs administracyjny, gdzie takie ryzyko istnieje, musimy dostosować się do zabezpieczeń stosowanych przez serwer. Zarówno ciasteczka, jak i token muszą zostać zachowane na potrzeby wykonywania przyszłych zapytań.

3.2. Tworzenie puli maszyn wirtualnych

Automatyzacja tworzenia nowych pul maszyn wirtualnych to jeden z głównych celów zaproponowanego rozwiązania. Zapytanie do serwera REST odpowiedzialne za tę akcję jest obszerne i wymaga podania wielu wartości. Spora część z nich może być jednak ustawiona na stałą wartość, ponieważ tylko jedna ich konfiguracja jest dopuszczalna w danym środowisku. Zapytanie to musi być wysłane metodą POST na ścieżkę API: view-vlsi/rest/v1/desktop/create. Parametry konieczne do utworzenia puli maszyn wirtualnych pobieramy z dostępnych endpointów typu GET:

- /view-vlsi/rest/v1/queryservice/DestopSummaryView
- /view-vlsi/rest/v1/BaseImageVm/list

Na Listingu 1. przedstawiony jest wycinek formatu wiadomości, którą trzeba zawrzeć w zapytaniu o utworzenie nowej puli maszyn wirtualnych.

4. IMPLEMENTACJA AUTOMATYZACJI PRZY UŻYCIU MS EXCEL I XLWINGS

Interfejs użytkownika w programie Microsoft Excel składa się ze skoroszytu arkusza kalkulacyjnego utworzonego z użyciem xlwings przy zastosowaniu programu napisanego w języku Python. Pakiet xlwings składa się z dwóch głównych elementów: z dodatku (ang. add-in) do programu MS Excel oraz z modułu dla języka Python. Dodatek do programu MS Excel zawiera kod w języku Microsoft Visual Basic odpowiedzialny za komunikację z środowiskiem uruchomieniowym języka Python [9]. Kod ten uruchamia usługę pośrednią, pozwalającą na rejestrowanie funkcji w języku Python jako funkcji środowiska MS Excel. Pozwala również na rejestrowanie funkcji jako procedur języka Visual Basic, dzięki czemu możliwe jest uruchamianie kodu w języku Python z wstawek Visual Basic. Pomaga to np. w sprzęganiu z interfejsem Visual Basic ActiveX, umożliwiając tworzenie przyjaznego dla użytkownika interfejsu graficznego.

```
{
 "type": type,
 "base":{
 "name": name,
 "displayName": display_name,
 "accessGroup":"AccessGroup/ACCESSGROUPID"
 },
 "desktopSettings":{
 /* ...Constant values... */ },
 "automatedDesktopSpec":{
 /* ...Constant values... */
 "vmNamingSpec":{
 "namingMethod":"PATTERN",
 "patternNamingSettings":{
 "namingPattern": naming_pattern,
 "maxNumberOfMachines": max_machines,
 "numberOfSpareMachines": spare_machines,
 "provisioningTime": "UP_FRONT" }, },
/* ...Constant values... */  }  }
```

Listing 1.: Fragment sparametryzowanej konfiguracji puli przekazywanej w zapytaniu HTTP.

4.1. Zarządzanie arkuszem MS Excel przy użyciu xlwings

Niektóre operacje, np. pobranie listy dostępnych pul maszyn lub listy obrazów bazowych powinny zawsze być wpisywane do tego samego podarkusza. W tym celu zaimplementowana została funkcja przedstawiona na Listingu 2, która stara się znaleźć pożądany arkusz w zbiorze już istniejących, lub tworzy nowy, gdy ten nie istnieje.

Funkcja ta nie jest rejestrowana jako procedura/makro lub funkcja arkusza, ponieważ nie ma ona być widoczna z poziomu aplikacji MS Excel lub języka Visual

Basic. Jest ona dostępna wyłącznie na użytek kodu po stronie środowiska języka Python.

```
Def find_or_create_new(book, sheet_name):
 for sh in book.sheets:
 if sh.name == sheet_name:
 return sh

 book.sheets.add(sheet_name)
 return book.sheets.active
```

Listing 2.: Funkcja szukająca istniejącego arkusza o żądanej nazwie lub tworząca go gdy ten nie istnieje.

4.2. Listowanie dostępnych pul maszyn

Bardzo przydatne jest wypisanie listy dostępnych pul maszyn oraz ich parametrów (Listing 3). Wartości te pobierane są za pomocą klasy HorizonManager odpowiedzialnej za komunikację z serwerem REST usługi VMware Horizon. Lista pobranych pul zostaje przekazana do arkusza o nazwie ”Desktop Pools”. W przypadku niepowodzenia wykonana zostaje funkcja horison_error_handler(), która informuje użytkownika o błędach.

```
@xw.sub
 def list_pools():
 mgr = HorizonManager()
 mgr.set_error_handler(horizon_error_handler)
 mgr.login(*ask_user_credentials())

 desktops = mgr.desktop_summary_view()
 wb = xw.Book.caller()
 sheet: xw.Sheet =
find_sheet_of_name_or_create_new(wb, 'Desktop
Pools')
 r = sheet.range('A1',
f'A{DesktopPool.columns_number()}')
 r.value = DesktopPool.column_names
 i = 2
 for d in desktops:
 r = sheet.range((i, 1), (i,
DesktopPool.columns_number()))
 r.value = list(d.get_as_dict().values())
 i += 1
```

Listing 3: Procedura pobierająca listę dostępnych pul maszyn wirtualnych.

5. WNIOSKI

W ramach przedstawionego rozwiązania opisano system, który został opracowany na potrzeby Instytutu Informatyki na wydziale IEiT AGH w Krakowie.

Interfejs użytkownika przygotowany w skoroszycie programu Microsoft Excel, obsługuje podstawowe funkcjonalności systemu: dodawanie, edycję oraz kasowanie pul maszyn wirtualnych. System ten znacznie ułatwia pracę ze środowiskiem VMware, a w szczególności tworzenie pul maszyn wirtualnych w sposób zautomatyzowany. Dodatkowo eliminuje ryzyko przypadkowej edycji bądź błędnego wyboru kluczowych elementów infrastruktury takich jak zasoby dyskowe czy operacyjne maszyny wirtualnej.

Kolejną zaletą systemu jest jego elastyczność i rozszerzalność. Każdy użytkownik systemu zaznajomiony z językiem programowania Python będzie w stanie dodać do niego kolejne funkcjonalności związane z serwerem VMware Horizon bądź dowolnym innym zarządzanym narzędziem z rodziny VMware, które obsługuje REST API.

System ten został przetestowany podczas tworzenia nowych obrazów wzorcowych dla maszyn wirtualnych, gdzie koniecznym było wielokrotne tworzenie różnych typów pul maszyn wirtualnych z pojedynczymi zmianami w konfiguracji. W takim scenariuszu interfejs w programie Microsoft Excel sprawdził się bardzo dobrze, ponieważ pozwalał na szybką zmianę dowolnych parametrów, bez konieczności każdorazowego uzupełniania pełnej konfiguracji tworzonej. Przedstawione rozwiązanie omawia rzadko poruszany temat automatyzacji pracy ze środowiskiem VMware Horizon i stanowi unikatową propozycję gotowego rozwiązania problemu.

LITERATURA

[1] Agarwal Shalabh, Biswas Rana, Nath Asoke. 2014. „Virtual Desktop Infrastructure in Higher Education Institution: Energy Efficiency as an Application of Green Computing". *2014 Fourth International Conference on Communication Systems and Network Technologies* : 601-605

[2] Alagappan Annamalai, Venkataraman Sivakumar, Sivakumar Subitha. 2016. „Virtual desktop infrastructure for rendering education technology in multifaceted learning platforms — A case study at Botho University." *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)* : 1717-1720.

[3] Casanova Louis, Marcel, Kristianto Edy. 2017. „Comparing RDP and PcoIP protocols for desktop virtualization inVMware enviroment." *2017 5th International Conference on Cyber and IT Service Management (CITSM)* : 1-4.

[4] Chrobak Paweł. 2014. „Implementation of Virtual Desktop Infrastructure in academic laboratories”. *Proceedings of the Federated Conference on Computer Science and Information Systems*: 1139–1146.

[5] Laitkorpi Markku, Koskinen Johannes, Systa Tarja. 2006. „A uml-based approach for abstracting application interfaces to rest-like services.” *13th Working Conference on Reverse Engineering* : 134–146.

[6] Liu Fangtao, Yang Zhiyong. 2018. „Design of vmware vsphere automatic operation and maintenance system based on python.”. *In 2018 International Conference on Advanced Mechatronic Systems (ICAMechS)* : 283–286.

[7] Nakhai Pedram Hossein, Anuar Nor Badrul. 2017. "Performance evaluation of virtual desktop operating systems in virtual desktop infrastructure". *IEEE Conference on Application, Information and Network Security (AINS)*, 2017: 105-110.

[8] Sangsanit Krisada, Kurutach Werasak, Phoomvuthisarn Suronapee. 2018. „Rest web service composition: A survey of automation and techniques.” *In 2018 International Conference on Information Networking (ICOIN)* : 116–121.

[9] Zumstein Felix. 2021 „Python for ExcelA Modern Environment for Automation and Data Analysis.” *O'Reilly Media, Inc., 005 Gravenstein Highway North, Sebastopol, CA 95472*.

SYJAMSKIE SIECI NEURONOWE NA TROPIE PODOBIEŃSTWA BŁĘDÓW W STACJACH BAZOWYCH SIECI KOMÓRKOWYCH 5G

SIAMESE NEURAL NETWORKS ON THE TRAIL OF SIMILARITY IN BUGS IN 5G MOBILE NETWORK BASE STATIONS

Sebastian Zarębski[1,2], Aliaksandr Kuzmich[1], Sebastian Sitko[1], Krzysztof Rusek[2], Piotr Chołda[2]

[1] NOKIA, Bobrzyńskiego 46, 30-348 Kraków, Poland
[2] AGH University of Science and Technology, Institute of Telecommunications, Al. Mickiewicza 30, 30-059 Kraków, Poland

DOI: 10.15199/59.2022.4.21

Abstract: **To improve the R&D process, by reducing duplicated bug tickets, we used an idea of composing BERT encoder as Siamese network to create a system for finding similar existing tickets. We proposed several different methods of generating artificial ticket pairs, to augment the training set. Two phases of training were conducted. The first showed that only and approximate 9% pairs were correctly identified as certainly similar. Only 48% of the test samples are found to be pairs of similar tickets. With the fine-tuning we improved that result up to 81%, proving the concept to be viable for further improvements.**

Keywords: **fault detection, machine learning, natural language processing**

1. INTRODUCTION

R&D process is inherently complex and require significant financial and human resources. Employees of telecommunication companies, who are involved in the implementation of new technologies, should be aware of the current technical problems that other engineers are facing. To improve the process, it is necessary to design and create a system to search and compare newly discovered defects with ones which are already fixed.

However, bug descriptions for 5G mobile network base stations are composed of many complementary fragments. In NOKIA's process, a title and a description are used to describe the symptom. It includes a list of the steps performed that led to the observation and a characteristic of the expected and observed test results. Noticed symptoms are also described by a set of system logs, recorded automatically during the device operation. Bug descriptions (tickets) are composed from text fragments that are characterized by a very high accumulation of phrases and words that are not included in general English dictionary. Thus, it is futile to expect that the application of neural networks, which were evaluated on open and general data sources, will produce similar results when used in the technological process. The specific corporate dataset is characterized by a high content of abbreviations and specific company's and telecommunication jargon. It makes tickets hard to understand without having a proper knowledge – see Figure 1.

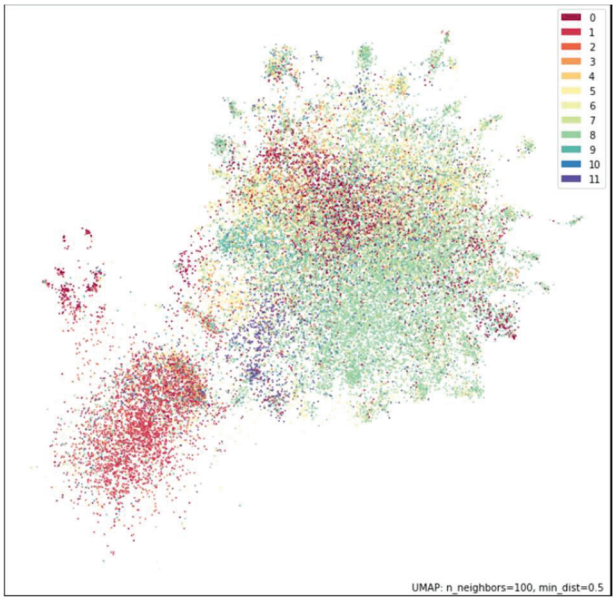


Fig. 1. Visualisation of encoded bugs descriptions across 12 functional areas. Tickets form a big and rather uniform cluster, showing how similar they are.

In our work, we focused on using an existing neural network architecture (transformer) to create a system that finds semantically similar error descriptions. This was accomplished by using a combination of identical networks into a so-called Siamese network. By applying additional filters in the form of categorical data (product name, software version, hardware type used), we improved the results by rejecting bug reports that were deceptively similar to each other. By creating a corpus of important vocabulary from 3GPP documentation [1], we were able to improve the efficiency of the overall tool and improve the processing of a text.

We give the following structure to this paper. In Section Two we present the exact problem statement. In Section Three we point to literature that inspired our idea. In Section Four we describe the methods we used. In Section Five we conclude our research with results and findings.

2. PROBLEM STATEMENT

In Nokia, as soon as bug report is written, the fault coordinator reviews it to see if problem is already known or similar one has been already submitted. If so, the ticket may be closed and declared invalid. In other cases, there begins a so-called "investigation" phase, where a selected team of programmers and/or specification architects looks for the source of the problem. Then a fix is created and provided to a tester. If the patch is sufficient and the problem no longer exists, the ticket can be closed. This process is described in Figure 2.

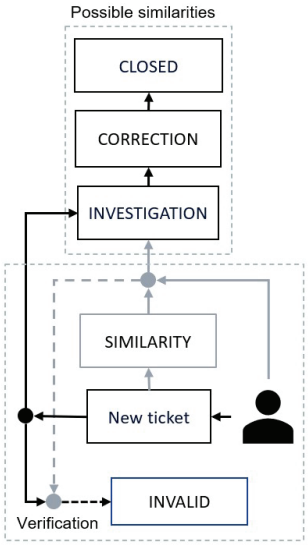


Fig. 2. Typical bug report status diagram. The thick black line describes the present process, the thick grey line the proposed improvement of the process.

There is a risk that two or more coordinators - especially from different offices and time zones - will not notice that another department saw the same or similar error. Such situation creates a strong need to enable such checking already in the phase of creating a ticket. It would improve the process by reducing an effort of coordinators and others, who can focus on different (perhaps more important) tasks.

3. LITERATURE REVIEW

There are already well-known approaches of an efficient text encoding. FastText algorithm introduced use of n-grams into a skip gram model, where cosine similarity was used to evaluate an effectiveness of a solution [2]. Approaches that do not use machine learning were developed, focusing on indexing textual data [3]. However, our scenario requires to use a method, that can be adapted to very inclusive meanings in a dataset. Therefore, powerful neural network architecture must be utilised – to properly find and understand complex meanings. Lately, the encoding model called the BERT was presented, claiming first places in many evaluation metrics [4]. Another research proved that it might be used in a semantic comparison of sentences. For this an architecture called a Siamese network [5] must be used. Several researchers proved that using a BERT encoder in such scenario produces high results in multiple tasks [6][7]. Noticeably, it required to run the training in two phases – normal, in which encoder weights are frozen (non-trainable) and the second, called the fine-tuning [8], in which we unblock chosen number of encoder layers. To use the model, after encoding text into its dense representation, such vector must be checked against database of existing bugs. For this, the Annoy library that extends the idea of the ANN [9], may be utilised. It provides an optimal way of computing distances between pairs of vectors (Euclidean, cosine etc).

4. METHODS

Problem description is composed of several textual parts. These are: title, expected observation outcome, observed situation, hardware details and set of logs pointing to the possible root cause of a symptom. To represent it as a data sample, we merge all parts into single sentence. Then the text is tokenized. Each token, which is rarely a whole word, is called a word piece, as mentioned in [4]. In our case a final vector has usually between three to five hundred tokens.

We have found that only 2% of the entire history of software bugs are known as very similar items (duplicates). Such are defined by the description of symptoms indicating the same bug, if detected in the same product with the same software version, on the same hardware configuration. The resulting subsets usually contain 2 to 5 reports linked to each other. One ticket is defined as a parent and the rest are attached to it. The relatively small number of similar tickets is a positive fact for the company - but not for our problem. The irony is that by proposing this solution, we reduce the small corporate dataset even further. For this reason, the final dataset must be specially prepared to compensate a deficiency of a training samples. Therefore, we need to artificially generate a larger number of pairs, which will allow the neural network to learn the relationships between selected text fragments. To achieve this goal, we defined six types of tickets pairs, which we then extracted from the dataset and combined into separate datasets. Rules of creating them are as follow:

- Set A - tickets are combined only from the internal subgroup (related to the parent ticket, but without it),
- Set B – as in set B, but along with a parent ticket,
- Set C - tickets are combined between different subgroups (having different parent tickets, without them),
- Set D – as in set C, but having the same categorical values,
- Set E - tickets are combined randomly between parent ticket and those that did not have similar tickets,
- Set F - tickets are combined between different subgroups (having different parent tickets).

Sets A and B are composed of pairs with a positive similarity relation, that is, pairs of tickets which are simi-

lar to each other. Sets C through F are sets of pairs of tickets that are dissimilar to each other. The test set is the same for all subsets and contains unique pairs of similar tickets (created analogously to set B). Thus, we can verify which approach produce better results - whether the network should specialize in determining the similarity or dissimilarity of the given inputs.

In the result each version of the dataset was approximately three to four times bigger than the original one. Next step is to split them on training, validation and testing subsets. By doing it naively (taking randomly samples up to a defined amount) we could eventually lead to data leak (pairs generated from the same problem could be split between training and testing set). To avoid such situation, we propose following short procedure:

1) dataset must be divided on smaller clusters (grouped by parent problem) - clusters must be treated as "whole samples",
2) clusters must be sorted by parent's date of correction (newest problems on the top to put most of them in testing or validation subset – to evaluate model against the current company's technological backlog),
3) then clusters may be separated on desired subsets.

Having prepared training and testing sets, we can determine the architecture of the neural network. Based on the research cited in the Section Three, we focused on using two BERT encoders connected in a Siamese network as presented in Figure 3. Because both encoders share parameters (neurons weights), this configuration can be simplified to the use of just one.

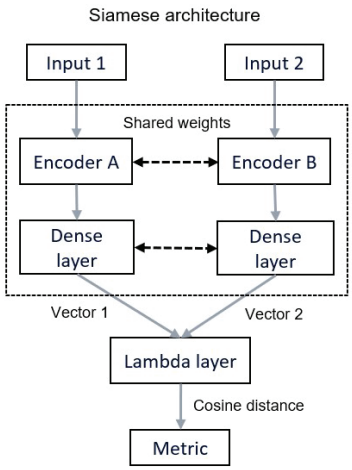


Fig. 3. Generic Siamese network architecture with shared encoder and dense layer's weights

To compare a pair of tickets, two input layers are used. The encoder is terminated by a trainable deep layer, which learns the representation of vectors after they have pushed through it. The whole network is finished with a Lambda layer, whose purpose is to compute the cosine distance between obtained encodings. In this process, we use a loss function called the contrastive loss. The intuition here is that we want our similar vectors to be as close to 1 as possible since log(1) = 0 (thus loss equal to zero). However, as the network is computing cosine distance, not cosine similarity, the final labels are determined in such way that d=0 means two vectors are ideally similar and d=1 means that they are completely different. In production scenario, Lambda layer is omitted, and dense vectors are fed to the ANN, to cluster the most similar ones. These are the final predictions.

5. RESULTS AND CONCLUSION

Let us define our expectations for the results. Tables 1 and 2 contain the metrics of testing each set, for the process with frozen encoder weights and unfrozen (fine-tuning). The first two columns show the percentage of pairs classified as similar when the neural network was very confident (distance d= [0; 0.25]) and confident (d= (0.25; 0.5]). The third column is the sum of these results, which can be compared to the fourth column, which contains the rest of the prediction results. Distance d= (0.5; 1] means that the network considered those pairs to be rather dissimilar or completely different from each other.

The first phase of training, which assumes that BERT encoder layers are frozen, prove that internal vocabulary related to the problem is difficult to use and the network is unable to catch much of important context. Results are presented in Table 1. It is noticeable that the confidence of the answers about the similarity of the samples is quite low. At best 9% strong confidence is obtained and just 48% of testing pairs are marked as somehow similar. These results are observed for the set E, which represented pairs that were dissimilar to each other - and that are selected almost randomly. An additional problem arises here because the pairs are combined almost independently of the process. Due to the randomness, there is control over exactly which pairs would be formed. Thus, a different draw might change this result. Therefore, it is worth looking at the second-best result, that is, for the set F, in which we combine dissimilar pairs by pairing tickets belonging to other groups (having a different parent tickets).

Tab. 1. A demonstration of the effectiveness of the neural network depending on the approach taken to create the training set – frozen encoder.

SET	Metric boundaries			
	[0; 0.25]	(0.25; 0.5]	[0; 0.5]	(0.5; 1]
A	6.70%	29.82%	36.52%	63.48%
B	5.56%	35.84%	41.40%	58.60%
C	3.56%	30.36%	33.92%	66.08%
D	4.32%	35.02%	39.34%	60.66%
E	**8.76%**	**48.02%**	**56.78%**	**43.22%**
F	6.5%	41.94%	48.44%	51.56%

The huge difference to these results can be observed in Table 2. This proves how crucial the phase of fine-tuning the network with BERT encoder is. Our re-search has shown that in the case of smaller sets - such as in this case - it is enough to unlock only one encoder layer, that is, the last one, to obtain good results and save some memory on GPU card. After such training, the neural network was able to identify as similar more than 81% of the given test samples, of which it identified only half a percent as probably similar. Only 17% of the samples could not be classified as related and similar.

Tab. 2. A demonstration of the effectiveness of the neural network depending on the approach taken to create the training set – fine tuning.

SET	Metric boundaries			
	[0; 0.25]	(0.25; 0.5]	[0; 0.5]	(0.5; 1]
A	**81.12%**	0.56%	**82.68%**	**17.32%**
B	80.34%	0.50%	80.84%	19.16%
C	79.92%	0.82%	80.74%	19.26%
D	76.82%	**0.84%**	77.66%	22.34%
E	75.40%	0.82%	76.22%	23.78%
F	79.54%	0.54%	80.08%	19.92%

It is worth to note that in the process of creating 5G network base stations, Nokia prepared about 500 of its hardware versions (motherboards and radio heads). Based on the document prepared by the 3GPP organization containing all the abbreviations and phrases, we were able to identify about 3,500 technical abbreviations and a dictionary containing about 30,000 words (including 40% not found in the common English dictionary).

Additionally, the procedure of generating a testing set is still to be improved. It is still naïve from business perspective, as the algorithm pays no attention if samples of each product (e.g., radio head types) are represented equally in the training and testing sets. This can be improved.

Eventually, we have obtained a highly automated system that can be used to support the daily work of testers in the company. However, it should be noted that this study is based on working with pairs that we know that are similar. To get a complete picture, it is also necessary to find how many of such pairs, determined by the model, were falsely indicated as related (so-called false positives). We can decrease some of these wrong predictions by applying filters in the form of comparisons of categorical values associated with reports.

An important aspect not addressed in this work is the fact of response uncertainty. For data with such specific and content, it would be useful to be able to estimate the uncertainty of assessing the similarity of two tickets. One should then consider applying the problem in the context of variational inference. This will allow the solution to correctly estimate two types of uncertainty - epistemic (related to lack of knowledge) and aleatoric (related to probability calculus itself). If we want to suggest potentially similar error descriptions to the user, we must also be able to answer how confident we are in giving such answer.

LITERATURE

[1] 3GPP. 2021. version 17.1.0. “Vocabulary for 3GPP Specifications”. 3rd Generation Partnership Project (3GPP), *Technical Specification (TS) 21.905*. [Online]. Available: http://www.3gpp.org/DynaReport/21905.htm

[2] Bojanowski Piotr et al., 2017. “Enriching Word Vectors with Subword Information”. *Transactions of the Association for Computational Linguistics.* (5): 135-146

[3] Elastic. 2022. “Elasticsearch: The Official Distributed Search & Analytics Engine”. [Online]. Available: https://www.elastic.co/elasticsearch/

[4] Devlin Jacob et al. June 2019. “BERT: Pre-training of deep bidirectional transformers for language understanding”. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (1): 4171–4186. Association for Computational Linguistics.

[5] Reimers Nils, Gurevych Iryna. 2019. “Sentence-bert: Sentence embeddings using siamese bert-networks”. *CoRR*.

[6] Lavi Dor, Medentsiy Volodymyr, Graus David. 2021. “conSultantBERT: Fine-tuned Siamese Sentence-BERT for Matching Jobs and Job Seekers.” *ArXiv* abs/2109.06501

[7] Chicco Davide. 2021. “Siamese Neural Networks: An Overview”. *Artificial Neural Networks. Methods in Molecular Biology*. (2190): 73–94.

[8] Merchant Amil et al. 2020. “What happens to BERT embeddings during fine-tuning?”, *Proceedings of the Third Blackbox NLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. 33–44. [Online]. Available: https: //aclanthology.org/2020.blackboxnlp-1.4

[9] Li Wen et al. 2020. "Approximate Nearest Neighbor Search on High Dimensional Data — Experiments, Analyses, and Improvement". *IEEE Transactions on Knowledge and Data Engineering*. 32 (8): 1475-1488.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

APLIKACJA DEMONSTRUJĄCA DZIAŁANIE KODÓW FONTANNOWYCH
THE APPLICATION DEMONSTRATING THE FOUNTAIN CODES

Kamila Haustein[1]; Marek Blok[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, hausteinkamila@gmail.com
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, marek.blok@pg.edu.pl

DOI: 10.15199/59.2022.4.22

Streszczenie: **Istotną cechą kodów fontannowych jest możliwość odtwarzania nadawanych danych niezależnie od jakości kanału. Wynika to z tego, że kody te nie mają z góry założonej zawartości informacyjnej, a kolejne symbole nadmiarowe są generowane, dopóki jest taka potrzeba. Jest to szczególnie przydatne w transmisjach broadcastowych, bo każdy z odbiorców może zdekodować dane tak wcześnie, na ile pozwala jakość łącza. Aby przybliżyć funkcjonowanie tych kodów studentom, powstała zaprezentowana w pracy aplikacja demonstrująca ich kodowanie i dekodowanie.**

Abstract: **An important feature of the fountain codes is the ability to reconstruct the transmitted data regardless of the conditions in the channel. They are rateless codes, so they do not have any predetermined code rate, and successive redundant symbols are generated as long as it is needed. This is especially useful in broadcast transmissions, where each recipient can restore the data as early as possible depending on link quality. In order to familiarize students with the functioning of these codes, an application was created that demonstrates their coding and decoding.**

Słowa kluczowe: **nadmiarowe kodowanie korekcyjne, kanał z wymazywaniem, kody fontannowe.**

Keywords: **erasure channel, error correction coding, fountain codes.**

1. WSTĘP

W 1998 roku Michael Luby zapoczątkował cyfrowe kody fontannowe LT (*Luby Transform Codes*) [6], które umożliwiają rozszerzanie kodu w trakcie transmisji o kolejne pozycje kontrolne zależnie od potrzeb. Struktura kodów nie jest z góry ustalona, stąd zawsze mamy możliwość wygenerowania kolejnego symbolu, który w zestawieniu z wybranymi poprzednimi symbolami może pozwolić na odtworzenie nadawanej wiadomości. Jednocześnie symbole nadawane można odtworzyć na wiele sposobów na podstawie różnych zestawów symboli odebranych. Wraz z dodawaniem kolejnych symboli kodowych liczba takich możliwości rośnie, a co za tym idzie rośnie odporność kodu na wymazywanie jego symboli.

Jeżeli przyjąć, że kodowane symbole to kolejne pakiety do nadania, to kody fontannowe pozwalają na odtworzenie wszystkich pakietów nadawanej wiadomości, w sytuacji, gdy kanał transmisyjny wymazuje w sposób losowy przesyłane pakiety kodowe. Ponieważ z założenia liczba zakodowanych pakietów, które może wygenerować źródło wiadomości jest nieograniczona, to niezależnie od prawdopodobieństwa wymazania pakietów w kanale, odbiorca prędzej czy później odbierze zestaw pakietów wystarczający do odtworzenia całej nadawanej wiadomości.

W przypadku kodów fontannowych liczba pakietów nadawanych może być określona w trakcie kodowania zależnie od warunków panujących w kanale, bez konieczności zakładania z góry ustalonej zawartości informacyjnej kodu (*code rate*) [7]. Ta cecha kodów *rateless* jest szczególnie przydatna w przypadku transmisji rozgłoszeniowych, gdy straty pakietów dla każdego odbiorcy są inne. To odbiorca, a nie nadawca, określa, ile pakietów jest mu potrzebnych do prawidłowego odebrania wiadomości. W efekcie kod się „dopasowuje" do pojemności informacyjnej kanału. Odbiorca dysponujący kanałem wysokiej jakości odtworzy wiadomość nadawaną bez konieczności odbierania pakietów nadmiarowych lub z niewielką ich liczbą. Z kolei odbiorca z kanałem niskiej jakości będzie wymagał wielu pakietów nadmiarowych, których liczba jest powiązana z wymaganą zawartością informacyjną kodu zależną od faktycznej pojemności informacyjnej kanału, której nie musimy znać z góry. Dzięki tej właściwości kody te znalazły swoje miejsce w nowoczesnych rozwiązaniach i standardach.

W celu przybliżenia funkcjonowanie tych ważnych kodów opracowana została zaprezentowana w pracy aplikacja, zrealizowana w Matlabie, demonstrująca działanie algorytmu kodowania i dekodowania kodów fontannowych [2]. Zaimplementowane narzędzie umożliwia m.in. pracę krokową, czy też obserwowanie równoczesnej transmisji do kilku odbiorców przy niezależnych stratach pakietów. Program ten umożliwia zweryfikowanie warunków koniecznych do prawidłowego odebrania nadanej wiadomości w zależności od właściwości macierzy generującej przy pakietach losowo usuwanych przez kanał podczas transmisji.

2. RODZAJE KODÓW FONTANNOWYCH

Ogólną ideą kodów fontannowych jest rozpraszanie informacji o pakietach do nadania w wielu pakietach przesyłanych przez kanał w taki sposób, żeby możliwe było odtworzenie pakietów źródłowych na podstawie wielu różnych podzbiorów pakietów odebranych. Uodparnia to przesyłane dane na losowe straty pakietów. Koncepcję przypadkowych kodów fontannowych przedstawiono w p. 2.1. W ramach tej koncepcji powstały różne rozwiązania praktyczne takie jak kody Tornado [5], LT [7] oraz

Raptor [3]. Kody te i ich właściwości pokrótce przedstawiono w p. 2.2-2.4.

2.1. Przypadkowy kod fontannowy

Pakiet wyjściowy kodera kodu fontannowego jest sumą modulo 2 losowych pakietów źródłowych $\boldsymbol{x}$, gdzie x_l dla $l = 1,2, \ldots, N$ to l-ty pakiet, który chcemy przesłać do odbiorcy. Oznacza to, że pakiet nadawany y_j można zapisać w postaci mnożenia macierzowego realizowanego w arytmetyce modulo 2

$$y_j = \boldsymbol{g}_j \boldsymbol{x} \tag{1}$$

gdzie $\boldsymbol{g}_j$ jest wektorem kolumnowym o elementach przyjmujących wartość 0 albo 1, wskazujących na pakiety źródłowe tworzące pakiet nadawany y_j. Kolejne kolumny $\boldsymbol{g}_j$ tworzą macierz generującą nadajnika $\boldsymbol{G}_{K,N}$, gdzie K jest liczbą nadanych pakietów. Dla przypadkowego kodu fontannowego kolejne kolumny macierzy generującej $\boldsymbol{G}_{K,N}$ są generowane losowo.

Jak można zauważyć $\boldsymbol{g}_j[l] = 1$ oznacza, że w pakiecie y_j jest zawarta informacja o pakiecie źródłowym x_l. Jednak w celu odtworzenia tego pakietu z odebranego pakietu y_j należy wyeliminować wpływ pozostałych pakietów źródłowych

$$x_l = y_j \oplus \sum_{k=1}^{N} \boldsymbol{g}_j[k] x_k \tag{2}$$

Żeby odtworzyć wszystkie pakiety źródłowe x_i, odbiornik odbiera kolejne pakiety nadawane y_j tak długo, aż skompletuje N niezależnych kolumn macierzy $\boldsymbol{G}_{K,N}$, które tworzą macierz kwadratową $\boldsymbol{G}$. Problem sprowadza się wtedy do rozwiązania równania macierzowego

$$\boldsymbol{x} = \boldsymbol{G}^{-1} \boldsymbol{y} \tag{3}$$

Łatwo zauważyć, że aby przeciwdziałać stratom (wymazaniom) pakietów w kanale, wystarczy w macierzy $\boldsymbol{G}_{K,N}$ umieścić nadmiarowe kolumny. W takiej sytuacji zadaniem odbiornika jest wytypowanie spośród odebranych pakietów tych, które odpowiadają niezależnym kolumnom macierzy generującej.

2.2. Kody Tornado

Kody Tornado [5] w przeciwieństwie do ogólnych założeń kodów fontannowych nie należą do kodów rateless. Głównym atutem tych kodów jest możliwość przesyłania dużej liczby pakietów do wielu użytkowników zachowując przy tym małą liczbę pakietów nadmiarowych [4]. Dodatkowo cechują się one niską złożonością algorytmu dekodowania, co wynika to z tego, że pakiety nadmiarowe są zależne od niewielkiej liczby pakietów informacyjnych. Stąd kody te należą do grupy kodów LDPC (*low density parity check*) [4].

W kodzie Tornado na podstawie pakietów informacyjnych tworzona jest założona z góry liczba nadmiarowych pakietów kontrolnych. Następnie na ich podstawie tworzone są kolejne pakiety nadmiarowe chroniące pakiety kontrolne niższego poziomu. Kod taki może zawierać kilka kolejnych poziomów zabezpieczeń pozycji nadmiarowych, gdzie ostatni poziom jest zabezpieczany klasycznym kodem blokowym, który jest bardziej odporny na straty, np. kodem Reeda-Solomona [5].

Kody Tornado są oparte na rzadkim grafie dwudzielnym. Wszystkie węzły tego grafu, reprezentujące pakiety nadawane stanowią razem słowo kodowe. Wśród tych węzłów, węzły lewe reprezentują symbole informacyjne, a węzły prawe symbole kontrolne. Wielopoziomowa struktura kodów Tornado sprawia, że nie są one odpowiednie dla dużych systemów przesyłania danych ze względu na złożoność algorytmu ich kodowania oraz dekodowania proporcjonalną do długości kodu [6].

2.3. Kody LT

Kody LT (*Luby transform codes*) [7] są pierwszymi kodami fontannowymi o znaczeniu praktycznym. Charakteryzują się one tym, że liczba przesyłanych zakodowanych pakietów jest potencjalnie nieograniczona i jest określona w trakcie transmisji [7]. Zakłada się, że nadajnik przesyła kolejne zakodowane pakiety tak długo, aż dekoder będzie w stanie odtworzyć dane źródłowe.

W efekcie rozszerzania kodu na bieżąco w trakcie transmisji, dane źródłowe można odtworzyć niezależnie od liczby wymazanych pakietów. Kolejne pakiety są tworzone zgodnie z zależnością (1), przy czym wektor $\boldsymbol{g_j}$ jest tworzony w koderze losowo. Najpierw losowany jest stopień bieżącego węzła kontrolnego $\boldsymbol{d}$, a następnie w wektorze $\boldsymbol{g_j}$ umieszczane jest losowo $\boldsymbol{d}$ jedynek. Parametr $\boldsymbol{d}$ określa liczbę losowo wybranych pakietów źródłowych, które tworzą zakodowany pakiet.

Właściwościach kodu LT są ściśle związane z rozkładem prawdopodobieństwa stopni węzłów $\boldsymbol{d}$. W praktyce wykorzystuje się solidny rozkład solitonowy (*robust soliton distribution*) [5], który gwarantuje, że większość pakietów jest niskiego stopnia. Przekłada się to na niską złożoność kodowania i dekodowania oraz dużą wydajność dla dużej liczby kodowanych pakietów [8]. Ponadto kody LT są uniwersalne ze względu na to, że są one prawie optymalne dla każdego kanału z wymazywaniem.

2.4. Kody Raptor

Wadą kodów LT jest to, że wraz z kolejnymi odtworzonymi pakietami źródłowymi spada prawdopodobieństwo tego, że kolejny pakiet odebrany umożliwi odtworzenie nowego pakietu źródłowego. Kody Raptor [3] są rozszerzeniem kodów LT, które ograniczają ten problem i pozwalają na wcześniejsze odtworzenie wszystkich symboli wejściowych. Zasadniczo są one oparte na kodach LT, z tą różnicą, że kodowanie dzieli się na dwa etapy. Pakiety źródłowe są wstępnie kodowane standardowym kodem blokowym (*pre-coder*) zabezpieczającym przed wymazaniami. Następnie tak zabezpieczone pakiety są kodowane z użyciem klasycznego kodu LT. Dwustopniowa struktura kodu oznacza, że dekoder kodu LT nie musi odtworzyć wszystkich symboli kodu wstępnego, bo pozwala on na odtworzenie brakujących symboli [8]. Kosztem niewielkiej dodatkowej nadmiarowości wprowadzanej przez koder wstępny, dekoder LT nie musi oczekiwać na odebranie pakietu, do którego w ramach losowania trafiły ostatnie brakujące pakiety źródłowe [9].

Jednym z typu kodów Raptor są kody RaptorQ [3], które są najbardziej zaawansowanym kodem fontannowym. Charakteryzują się one znacznie lepszym prawdopodobieństwem dekodowania i lepszą wydajnością w porównaniu z użyciem samego kodu LT.

3. DEKODOWANIE KODÓW FONTANNOWYCH

W poprzednim punkcie opisano koncepcję kodowania kodów fontannowych, z kolei w niniejszym punkcie przedstawiono opis algorytmu dekodowania, który został zaimplementowany w aplikacji (rys. 2). Algorytm ten na podstawie odebranych pakietów i odpowiadających im kolumn macierzy $\boldsymbol{G}$ tworzy macierz $\widehat{\boldsymbol{G}}$ składającą się z zestawu N jej niezależnych kolumn oraz odpowiadający jej zestaw pakietów $\widehat{\boldsymbol{y}}$. Jest wystarczające do odtworzenia wektora $\boldsymbol{x}$ (3). Równocześnie, już w trakcie tej selekcji, rozwiązywany jest układ równań (3). Algorytm składa się z sześciu głównych etapów:

1. Inicjalizacja dekodera.

2. Odbiór kolejnego pakietu y_{k+1} i generacja odpowiadającej mu kolumny $\boldsymbol{g}_{k+1}$.

3. Wstępne przetworzenie pakietu y_{k+1} oraz kolumny $\boldsymbol{g}_{k+1}$ z użyciem bieżącej wersji macierzy $\widehat{\boldsymbol{G}}$ oraz wektora $\widehat{\boldsymbol{y}}$ w celu wyzerowania elementów $\boldsymbol{g}_{k+1}$ na pozycjach od 1 do k.

4. Weryfikacja czy przetworzona kolumna $\boldsymbol{g}_{k+1}$ może być dołączona do macierzy $\widehat{\boldsymbol{G}}$. (4a) Jeżeli przetworzona kolumna $\boldsymbol{g}_{k+1}$ ma jedynkę na $k+1$ pozycji, to dołączenie jej do macierzy $\widehat{\boldsymbol{G}}$ oraz dołączenie pakietu y_{k+1} do wektora $\widehat{\boldsymbol{y}}$ i aktualizacja macierzy $\widehat{\boldsymbol{G}}_{\boldsymbol{R}}$ oraz wektora $\widehat{\boldsymbol{y}}_{\boldsymbol{R}}$. (4b) Jeżeli przetworzona kolumna $\boldsymbol{g}_{k+1}$ ma zero na $k+1$ pozycji, to dołączenie jej do macierzy rezerwowej $\widehat{\boldsymbol{G}}_{\boldsymbol{R}}$ oraz dołączenie pakietu y_{k+1} do wektora $\widehat{\boldsymbol{y}}_{\boldsymbol{R}}$.

5. Sprawdzenie czy odtworzono wszystkie pakiety nadawanej wiadomości $\boldsymbol{x}$. Jeżeli $k = N + 1$ kończymy dekodowanie (7), w przeciwnym razie przechodzimy do przetwarzania kolejnego pakietu: zwiększenie k o jeden.

6. Weryfikacja czy w macierzy rezerwowej $\widehat{\boldsymbol{G}}_{\boldsymbol{R}}$ są kolumny z jedynką na k +1 pozycji. Jeżeli tak, to (7a) wydobycie pierwszej z nich jako kolumny $\boldsymbol{g}_{k+1}$ oraz z wektora $\widehat{\boldsymbol{y}}_{\boldsymbol{R}}$ i przejście do p. 4a. Jeżeli nie, to przejście do p. 2.

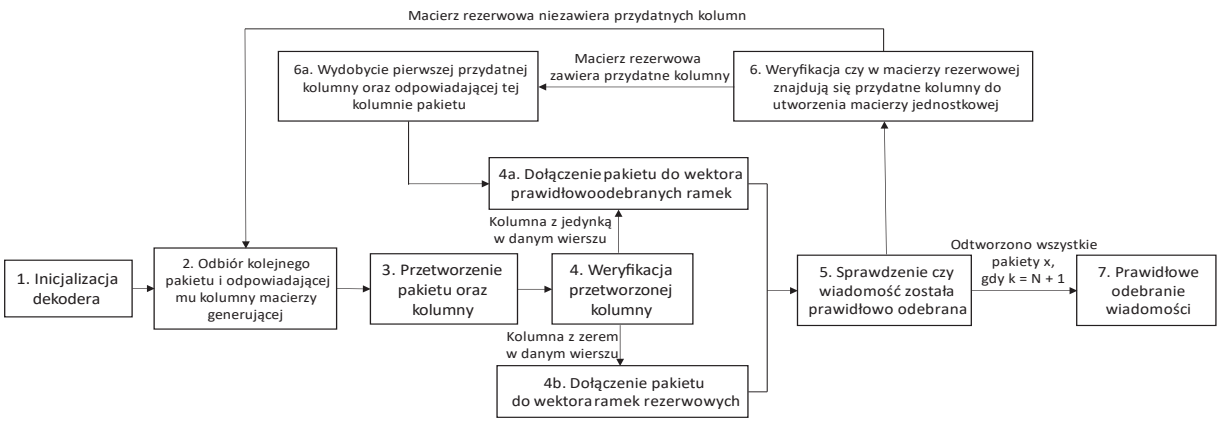
4. APLIKACJA DEMONSTRUJĄCA

Opracowana aplikacja demonstruje działanie kodów fontannowych. Została ona zorganizowana w sposób, który jest intuicyjny dla użytkownika i łatwy w obsłudze. Umożliwia ona wybór symulacji transmisji do jednego odbiorcy, kilku odbiorców oraz testowanie kodów fontannowych. Użytkownik samodzielnie może dokonać wyboru wartości parametrów takich jak: typ danych nadawanych, ich długość, liczbę pakietów, typ macierzy generującej, prawdopodobieństwo wymazywania lub numery usuniętych pakietów. W aplikacji zostały przedstawione również efekty transmisji w postaci parametrów liczbowych, takich jak liczba iteracji, liczba odebranych pakietów, liczba wymazanych pakietów, łączny czas dekodowania, najdłuższy czas dekodowania jednego pakietu itd. Dodatkowo wyniki zostały zaprezentowane w postaci wykresów, które ułatwiają zrozumienie algorytmu kodowania i dekodowania.

4.1. Przesyłanie danych do jednego odbiorcy

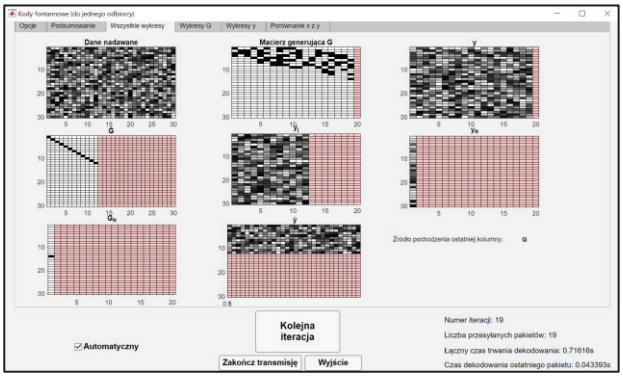
Po wyborze tej opcji użytkownik ma możliwość wyboru typu danych nadawanych: losowe lub wskazany obraz. Jeżeli wskazany zostanie obraz na ekranie będzie wyświetlona informacja odnośnie jego nazwy oraz rozmiaru. Kolejnym wyborem jest wybór sposobu generowania macierzy generującej. Zależnie od tego wyboru możliwe jest obserwowanie różnic w przebiegu procesu dekodowania dla różnych strategii generowania macierzy generującej. Mamy tu sześć możliwości.

Losowy: Wykorzystywane kolumny macierzy generującej są generowane w sposób losowy, a odpowiadające im pakiety odbierane, aż do momentu prawidłowego odebrania wiadomości. Scenariusz 1: Macierz generująca jest macierzą jednostkową, która ma wymiary równe liczbie pakietów danych nadawanych. Scenariusz 2: Kolumny macierzy generującej są generowane losowo. Dekodowanie zatrzymuje się po odebraniu N pakietów. W przypadku nieprawidłowego odebrania wiadomości można kontynuować dekodowanie z losową macierzą generującą. Scenariusz 3: Generowana jest macierz, która uniemożliwia prawidłowe odebranie wiadomości bez skorzystania z możliwości wygenerowania dodatkowych kolumn. W celu prawidłowego odebrania wiadomości użytkownik musi kontynuować odbieranie kolejnych pakietów. Scenariusz 4: Umożliwia wczytanie macierzy generującej z pliku tekstowego. Scenariusz 5: Macierz generująca posiada jedynki, które są skupione lokalnie, pozwalające na odtworzenie pakietów na podstawie bliskich sobie pakietów odebranych. Użytkownik określa zawartość informacyjną kodu oraz wielkość okna pakietów źródłowych składających się na pakiet kodowy. Okno aplikacji



Rys. 1. Schemat algorytmu dekodowania kodu fontannowego

podczas procesu dekodowania losowych danych źródłowych po wyborze Scenariusza 5 zostało zaprezentowane na rys. 2.



Rys. 2. Okno programu po wyborze scenariusza 5

4.2. Przesyłanie danych do kilku odbiorców

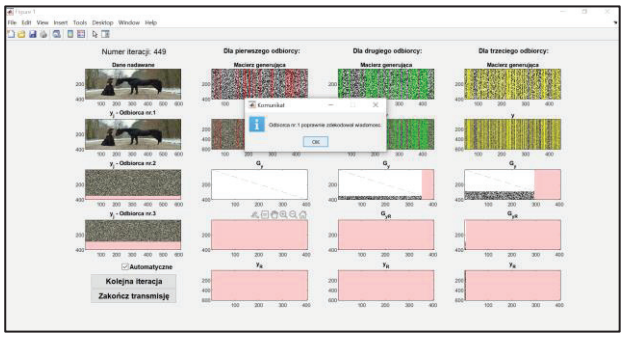
Zaimplementowane narzędzie pozwala użytkownikowi na śledzenie i przesyłanie jednakowej wiadomości do trzech różnych odbiorców i równoległe porównywanie kolejnych etapów procesu dekodowania. W celu łatwiejszego zobrazowania algorytmu działania, każdemu odbiorcy przypisano unikatowy kolor wymazywanych pakietów co możemy zauważyć na rys. 3. Po poprawnym odebraniu wiadomości przez wszystkich trzech odbiorców otwiera się zakładka Podsumowanie. W ramach tej zakładki użytkownik ma możliwość porównania wartości parametrów opisujących efektywność dekodowania przesyłanej wiadomości w zależności od zadanego prawdopodobieństwa wymazywania.

4.3. Testowanie kodów fontannowych

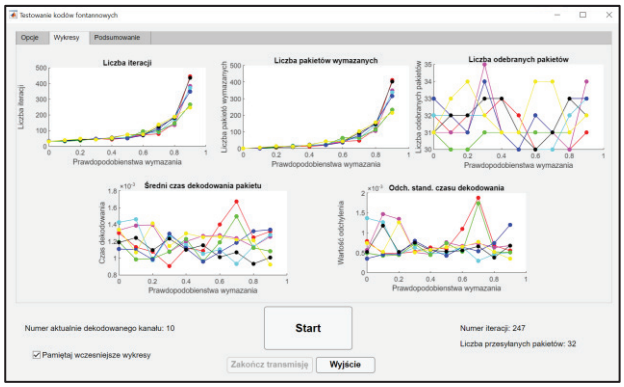
Moduł ten umożliwia użytkownikowi zaobserwowanie zmian w skuteczności procesu dekodowania w zależności od wartości prawdopodobieństwa wymazywania dla różnych macierzy generujących lub odmiennego wzorca wymazania pakietów. Użytkownik ma możliwość zachowania wcześniejszych wyników i dołożenie do nich wyników testów dekodowania, dla kolejnych parametrów kodera lub typów macierzy generującej, wykreślonych w funkcji prawdopodobieństwa wymazywania (rys. 4).

5. PODSUMOWANIE

Opracowana aplikacja umożliwia łatwiejsze zrozumienie działania kodów fontannowych. Dodatkowo ukazuje ich właściwości i zastosowania. Użytkownik ma możliwość dokładnego przeanalizowania algorytmu kodowania i dekodowania obserwując między innymi, że kody te pozwalają na prawidłowe odtworzenie przesyłanych danych niezależnie od prawdopodobieństwa wymazywania pakietów, które wpływa jedynie na to jak szybko nastąpi prawidłowe odtworzenie wiadomości. Dalsze prace nad aplikacją będą ukierunkowane na uzupełnienie jej o możliwość wyboru konkretnych typów kodów fontannowych (Tornado, LT czy Raptor), co pozwolił na porównywanie skuteczność poszczególnych rozwiązań oraz ocenić różnice w sposobie ich działania i porównanie prawdopodobieństwa poprawnego odebrania wiadomości dla określonej liczby zakodowanych pakietów przesyłanych przez kanał.



Rys. 3. Transmisja do wielu odbiorców; prawidłowe odtworzenie wiadomości przez pierwszego odbiorcę



Rys. 4. Okno aplikacji z wykresami przedstawiającymi wyniki badania wpływu prawdopodobieństwa wymazywania na poszczególne mierzone parametry symulowanych transmisji

LITERATURA

[1] Bello Philip. 1963, „Characterization of randomly time-variant channels", IEEE Transactions on Communications Systems, 11 (4): 360–393.

[2] Haustein K. 2021. "Aplikacja demonstrująca działanie kodów fontannowych", projekt dyplomowy inżynierski, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska.

[3] Ho T., 2003, „Summary of raptor codes", Scientific Commons.

[4] Joshi G. i inni, 2010, „Fountain codes", Global telecommunications conference.

[5] Khisti A., 2003, „Tornado codes and Luby transform codes", s. 1-12.

[6] Luby M. G., Mitzenmacher M., Shokrollahi M. A., Spielman D. A., Stemann V, 1997, „Practical lossresilient codes", s. 150-159.

[7] MacKay, David JC, 2003: „Information theory, inference and learning algorithms".

[8] MacKay, David JC, 2005, „Fountain codes", IEE Proceedings-Communications, s. 1062-1068.

[9] Nagasubramanian K., 2008: „Code design for erasure channels with limited or noisy feedback".

[10] Yao W. i inni, 2008, „Research on fountain codes in deep space communication", Congress on Image and Signal Processing.

BADANIE JAKOŚCI TRANSMISJI W SIECIACH KOMÓRKOWYCH NA PRZESTRZENI LAT 2019-2021: WPŁYW PANDEMII COVID-19 NA POZIOM ŚWIADCZONYCH USŁUG

STUDY ON QUALITY OF TRANSMISSION IN CELLULAR NETWORKS BETWEEN 2019-2021: IMPACT OF THE COVID-19 PANDEMIC ON THE LEVEL OF PROVIDED SERVICES

Przemysław Falkowski-Gilski[1]; Tadeus Uhl[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki,
ul. Narutowicza 11/12, 80-233 Gdańsk, przemyslaw.falkowski@eti.pg.edu.pl
[2] Akademia Morska w Szczecinie, Wydział Inżynieryjno-Ekonomiczny Transportu,
ul. Wały Chrobrego 1-2, 70-500 Szczecin, t.uhl@am.szczecin.pl

DOI: 10.15199/59.2022.4.23

Streszczenie: **Pandemia COVID-19 znacząco ograniczyła mobilność użytkowników, a w szczególności studentów. Nauka zdalna miała szczególny wpływ na sposób przydziału zasobów w sieciach komórkowych. Niniejsza praca przedstawia wyniki badań dotyczących jakości transmisji w środowisku zewnątrzbudynkowym. Kampanię pomiarową w latach 2019-2021 przeprowadzono na terenie kampusu uczelni technicznej. Każdego roku badania wykonano przy użyciu własnej autorskiej aplikacji mobilnej, zainstalowanej na 50 smartfonach działających pod kontrolą systemu Android.**

Abstract: **The COVID-19 pandemic has significantly limited user mobility, mainly students. Remote learning had a particular impact on resource allocation in case of cellular networks. This paper presents results of a quality evaluation in an outdoor environment. The measurement campaign between 2019-2021 was carried out at the campus of a technical university. Each year studies were performed using our own custom-build mobile application, installed on 50 smartphones running Android OS.**

Słowa kluczowe: **pomiary jakości, QoN (Quality of Network), QoS (Quality of Service), technologie mobilne, transmisja multimediów.**

Keywords: **mobile technologies, multimedia transmission, QoN (Quality of Network), QoS (Quality of Service), quality evaluation.**

1. WSTĘP

Pandemia COVID-19 znacząco ograniczyła mobilność użytkowników. Wiele osób z przyczyn osobistych i zdrowotnych musiało zrezygnować ze swobodnego przemieszczania się. Fakt ten dotyczył w szczególności studentów uczelni wyższych, którzy w znacznym stopniu przeszli na naukę zdalną. Z pewnością wymusiło to konieczność reorganizacji mechanizmów przydziału zasobów w naziemnych sieciach komórkowych. Niniejsza praca przedstawia wyniki pomiarów dotyczących jakości transmisji na terenie kampusu uczelni technicznej. Badania obejmują lata 2019-2021, przeprowadzone w środowisku zewnątrzbudynkowym na otwartej przestrzeni. Pomiary wykonano przy użyciu własnej autorskiej aplikacji mobilnej, działającej pod kontrolą systemu operacyjnego Android. Kampania pomiarowa w każdym roku obejmowała grupę 50 urządzeń mobilnych w postaci smartfonów, kompatybilnych ze standardem LTE (Long Term Evolution).

Tak dobrana metodyka pomiarowa, określana mianem *crowdsourcing*, czyni ją elastyczną, kontrolowalną, niezależną od systemów pomiarowych firm komercyjnych [8, 9], a także obniża przez to znacznie koszty podjętych badań. Należy podkreślić, że tego rodzaju testy dotyczą poziomu QoN (Quality of Network), a zatem są związane z pomiarami parametrów warstwy dostępowej badanej sieci. Drugim poziomem testów może być QoS (Quality of Service). Tutaj uwzględnia się warstwy wyższe logicznego modelu sieci, a zatem zaimplementowane w niej aplikacje. Ich jakość odgrywa zdecydowanie nadrzędną rolę. Tego rodzaju badania są bardzo złożone, czasochłonne i wymagają wyspecjalizowanych technik i urządzeń pomiarowych [7]. W większości przypadków są one obłożone licencjami, co czyni je kosztownymi.

Po krótkim wstępie omówione zostanie środowisko pomiarowe. Dalej nastąpi opis przeprowadzonych badań, tzn. przedstawione zostaną urządzenia i metodyka pomiarowa. Kolejny punkt poświęcony będzie rdzeniu niniejszej pracy, tj. zaprezentowaniu wyników pomiarów wraz z ich interpretacją. Pracę zakończy podsumowanie i wgląd do dalszych badań.

2. TEREN KAMPUSU

Politechnika Gdańska należy do największych uczelni technicznych w Polsce. Każdego roku kształci ok. 20 tys. studentów na studiach dziennych i zaocznych I i II stopnia. Prowadzi także zajęcia dla doktorantów w ramach Szkoły Doktorskiej. Uczelnia zatrudnia ponad 2,5 tys. pracowników, w tym ok. 1200 stanowią naukowcy. Kampus obejmuje 8 wydziałów i 77 hektarów miejskiej przestrzeni. Bez wątpienia jest to interesujące miejsce do przeprowadzenia badań naukowych dotyczących jakości usług w naziemnych sieciach komórkowych.

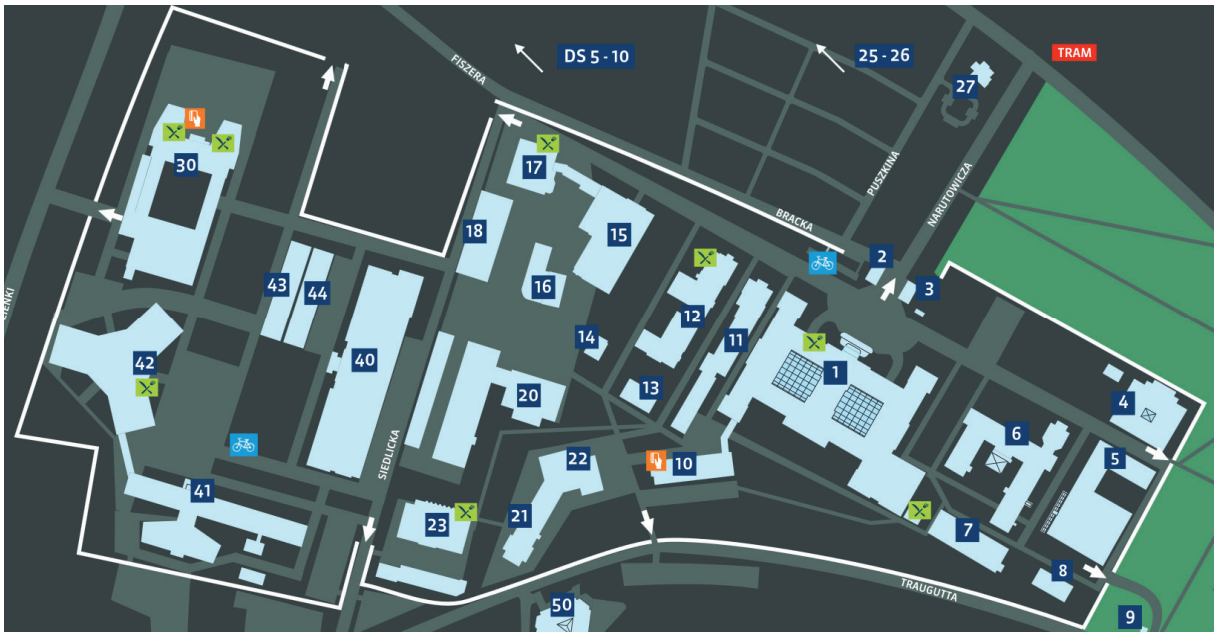
Kwestia zapewnienia odpowiedniej jakości, w tym wysokiej przepływności bitowej zarówno w łączu w dół jak i w górę, a także niskiego poziomu opóźnienia, jest kluczowe przy realizacji usług czasu rzeczywistego (np. transmisja sygnałów mowy, wideokonferencja), jak i usług strumieniowania multimediów. Jest to szczególnie

istotne biorąc pod uwagę rosnącą liczbę użytkowników, dostępność urządzeń mobilnych, wszechobecność sieci bezprzewodowych oraz idące za tym wzrastające oczekiwania względem efektywnych i niezawodnych usług [1].

Warto zaznaczyć, że współczesne urządzenia mobilne oferują znacznie więcej niż jedynie wymianę informacji tekstowych SMS (Short Message Service), multimedialnych MMS (Multimedia Messaging Service), czy transmisję sygnałów mowy. Smartfony oferują szerokie możliwości komunikacji oraz dzielenia się i konsumowania wszelkiego rodzaju treści [6]. Posiadają szereg zintegrowanych czujników oraz modułów komunikacji bezprzewodowej. Wiele informacji dotyczących ich stanu, trybu pracy oraz aktualnych wskazań można odczytać z poziomu aplikacji mobilnej zainstalowanej na urządzeniu abonenckim.

3. BADANIA

W latach 2019-2021 szereg zajęć na uczelniach wyższych prowadzonych było w trybie zdalnym lub mieszanym, tzw. hybrydowym. Same wykłady w większości przypadków odbywały się przy pomocy platform strumieniowych typu Clickmeeting, MS Teams, Zoom, itp. [4]. Wymagania dotyczące transmisji audio-wideo w czasie rzeczywistym przedstawiono w Tabeli 1. Dlatego też postanowiono zbadać jak rozdzielane są zasoby pomiędzy urządzeniami mobilnymi rozlokowanymi na wybranym terenie kampusu uczelni technicznej.



Rys. 1. Mapa Kampusu Politechniki Gdańskiej, na którym prowadzono pomiary [5]

Tab. 1. Najpopularniejsze usługi multimedialne i ich wymagania jakościowe

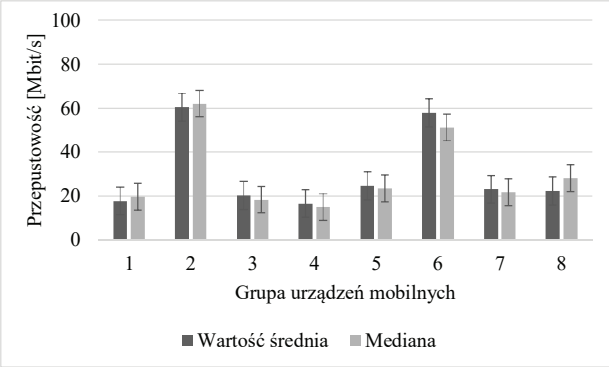
Usługa	*Przepływn. [Mbit/s]*	*Opóźnienie [ms]*
Wideokonferencja	1-4	100
Strumieniowanie w jakości SD	3-4	100 (w czasie rzeczywistym) 300 (z buforowaniem)
Strumieniowanie w jakości HD	5-8	100 (w czasie rzeczywistym) 300 (z buforowaniem)

Badania prowadzono podczas roku akademickiego na przestrzeni 3 lat od 2019 do 2021 na terenie otwartym Kampusu Politechniki Gdańskiej, przedstawionym na rys. 1. Każdego roku wykonywano testy przy użyciu 50 smartfonów, obejmujących szeroką gamę producentów, odmiennych specyfikacji technicznych oraz wersji systemu operacyjnego Android. Testy przeprowadzono przy pomocy autorskiej aplikacji mobilnej w ciągu tygodnia o różnych porach dnia. Każda osoba dokonywała pomiarów w 10 unikalnych punktach (powyżej 30 pomiarów na punkt), równomiernie rozlokowanych na terenie Kampusu PG. Pogoda była słoneczna, bez silnych wiatrów czy zachmurzenia, bez opadów atmosferycznych, zaś średnia temperatura wynosiła ok. 15-25 stopni Celsjusza.

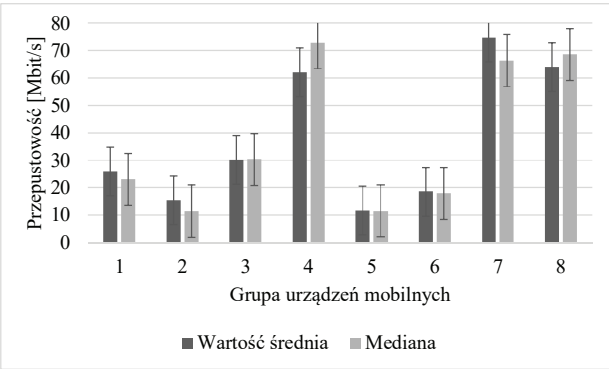
4. WYNIKI POMIARÓW

Wyniki pomiarów były gromadzone lokalnie w pamięci wewnętrznej urządzenia mobilnego, a następnie przeniesione do bazy danych do celów późniejszej obróbki i analizy. Rezultaty, w formie przepustowości (łącze w dół i w górę) oraz opóźnienia (ping), przedstawiono na rys. 2-10. Wyniki z roku 2019 opublikowane we wcześniejszej pracy [2], a także nowsze z lat 2020 oraz 2021, zaprezentowano odpowiednio na rys. 2-4, 5-7, 8-10.

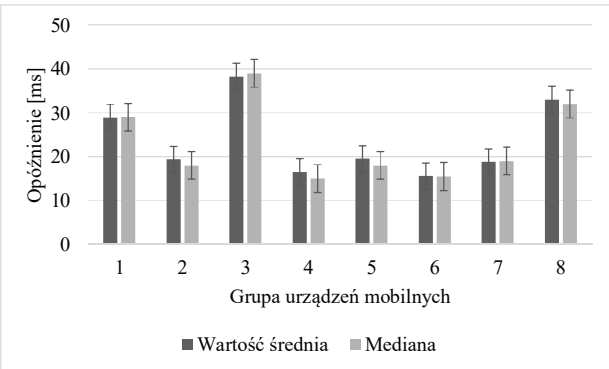
Urządzenia podzielono na 8 grup, w zależności od m.in. modelu i/lub producenta, liczby rdzeni procesora, ilości pamięci RAM, czy też wersji systemu operacyjnego Android. Warto zaznaczyć, że poziom ufności przy analizie danych wynosił 0,95. Oczywiście wszystkie urządzenia pracowały w standardzie LTE.



Rys. 2. Pobieranie danych (łącze w dół) – rok 2019



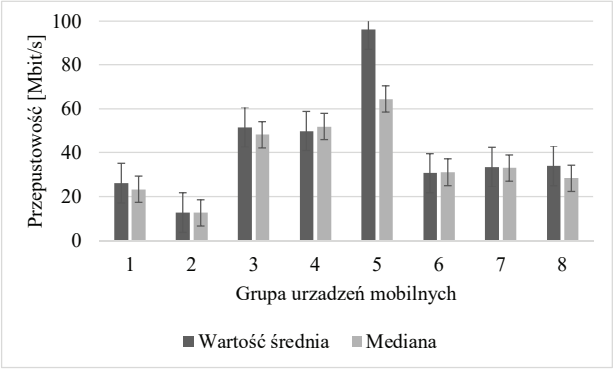
Rys. 3. Wysyłanie danych (łącze w górę) – rok 2019



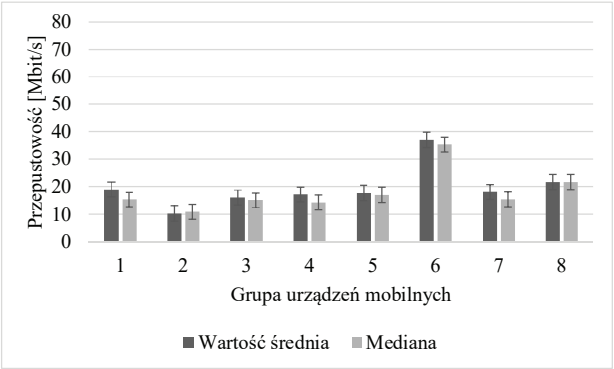
Rys. 4. Opóźnienie (ping) – rok 2019

Przedstawione wyniki pomiarów mogą być zinterpretowane w następujący sposób. W roku 2019 średnia wartość przepustowości w przypadku łącza w dół (pobieranie danych), przy uwzględnieniu 8 grup urządzeń końcowych, wynosiła ok. 30 Mbit/s, w roku 2020 blisko 40 Mbit/s, a w roku 2021 oscylowała na poziomie 50 Mbit/s. Jak zaobserwowano, im mniej użytkowników było zalogowanych do sieci komórkowej (w 2019 pełna praca stacjonarna na uczelni, od 2020 powolne przechodzenie na pracę zdalną, z kolei począwszy od 2021 pełna praca zdalna), tym większa była oferowana przepustowość na pojedynczą grupę urządzeń abonenckich. To zrozumiałe z uwagi na wykorzystywaną tam metodę współdzielenia medium, do podziału pasma w warstwie dostępowej.

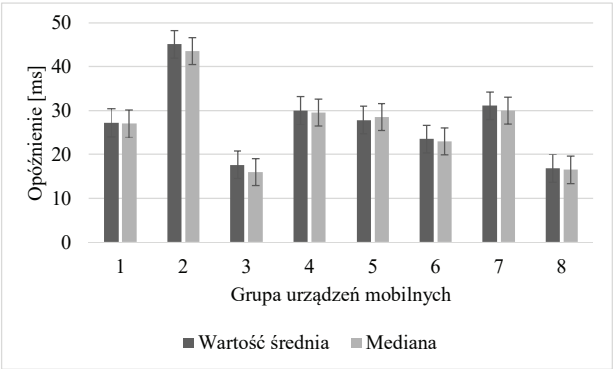
Powyższej tendencji nie można natomiast odnotować dla łącza w górę (wysyłanie danych). W roku 2019 średnia przepustowość w tym kierunku oscylowała na poziomie ok. 35 Mbit/s. Tutaj szczególnie urządzenia końcowe z grup 4, 7 i 8 podnosiły znacząco tę wartość. W latach 2020-2021 średnia przepustowość w tym kierunku zmniejszyła się do wartości ok. 20 Mbit/s. Mogło to wynikać z faktu zwiększonej liczby danych udostępnianych przez użytkowników. Przy wzmożonej pracy zdalnej wiązało się to zapewne z udostępnianiem własnego obrazu z poziomu pulpitu lub wskazanej aplikacji, komunikację głosową oraz oczywiście zamieszczanie sprawozdań na dedykowanej platformie do zdalnego nauczania.



Rys. 5. Pobieranie danych (łącze w dół) – rok 2020

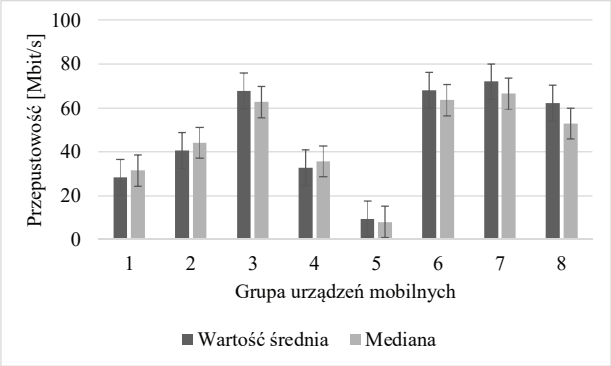


Rys. 6. Wysyłanie danych (łącze w górę) – rok 2020

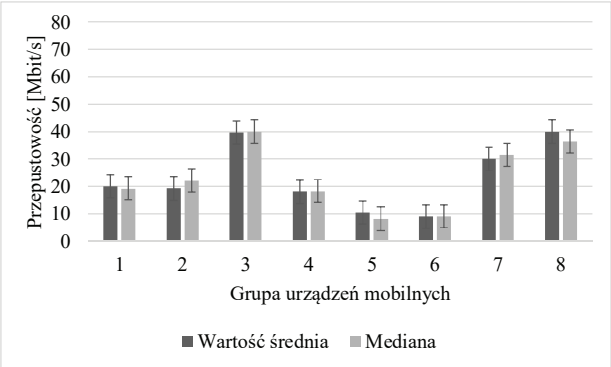


Rys. 7. Opóźnienie (ping) – rok 2020

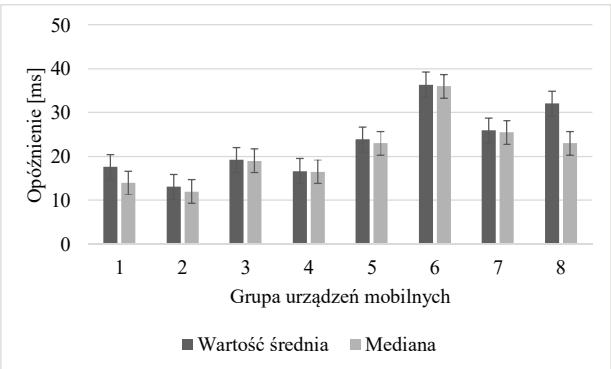
Wartości średnich opóźnień w sieci komórkowej w analizowanym okresie oscylowały na poziomie odpowiednio ok.: 24 ms (2019), 27 ms (2020) i 22 ms (2021). W tym wypadku nie można dopatrzyć się znaczącej korelacji pomiędzy wartością opóźnienia (ping) a liczbą użytkowników na terenie badanego kampusu. Wynika to zapewne z faktu, że ilość zapytań do serwerów zdalnego nauczania i pokrewnych repozytoriów dokumentów nie zmieniła się. Zmieniła się natomiast lokalizacja źródła (urządzenia), czyli miejsce generowania zapytań, związana z tzw. trybem *home office*.



Rys. 8. Pobieranie danych (łącze w dół) – rok 2021



Rys. 9. Wysyłanie danych (łącze w górę) – rok 2021



Rys. 10. Opóźnienie (ping) – rok 2021

5. PODSUMOWANIE

Wyniki analizy w sposób praktyczny opisują zmiany jakie zaszły w kwestii przydziału zasobów w przypadku naziemnych systemów telefonii komórkowej. Nauka zdalna i związana z nią ograniczona mobilność użytkowników miała zauważalny wpływ na prędkości transmisji w łączu w górę i w dół, natomiast niewielki na poziom opóźnień. Przeprowadzone badania uzupełniają analizy przedstawione przez autorów w roku 2020 [2].

Otrzymane wyniki pomiarów na przestrzeni lat 2019-2021 wykazują zdolność badanej sieci dostępowej do realizacji szeregu usług multimedialnych. Praca posiada walory praktyczne, stanowiące cenne wsparcie dla operatorów sieci komórkowych w kwestii optymalizacji zarządzania dostępnymi zasobami. Warto zaznaczyć, że proces edukacji w trybie zdalnym jest odmienny w zależności od typu zajęć, w tym: wykładów, seminariów, ćwiczeń, laboratoriów i projektów. Każdy z nich wymusza inny poziom aktywności, zarówno nauczyciela jak i uczniów. Szczególnie jeżeli bierzemy pod uwagę usługi realizowane w czasie rzeczywistym, związane m.in. z dwukierunkową komunikacją głosową, wideokonferencją, czy też udostępnianiem i dzieleniem się zasobami w formie strumienia danych multimedialnych lub wgrywanych sprawozdań. W tej kwestii kluczową rolę odgrywają zarówno łącze w dół, łącze w górę oraz oczywiście opóźnienie.

Można zauważyć, że mniejsza liczba urządzeń i aktywnych połączeń przyczyniała się do wzrostu przepustowości dla indywidualnych użytkowników (szczególnie w przypadku prędkości pobierania danych). W roku 2019 rzadko które urządzenie uzyskiwało wartości powyżej 60 Mbit/s. Z kolei w latach 2020-2021 spora ich liczba zbliżała się lub nawet przekraczała próg 80 Mbit/s. Wyraźnie spadły także wartości opóźnień. Co ciekawe w latach 2019-2020 oscylowało ono pomiędzy 20-40 ms, z kolei w roku 2021 wyraźnie spadło. Przyszłe badania mogłyby uwzględniać szersze spektrum aktywności użytkowników, czy też wybranych usług strumieniowych [3].

LITERATURA

[1] Elsayed, Mahmoud Said, Nhien-An Le-Khac, Anca Delia Jurcut. 2021. „Dealing with COVID-19 network traffic spikes [cybercrime and forensics]". *IEEE Security & Privacy* 19 (1) : 90–94.

[2] Falkowski-Gilski, Przemysław, Tadeus Uhl. 2020 „Badanie jakości sieci komórkowych w przededniu technologii 5G na wybranym przykładzie". *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne* 7-8 : 338–341.

[3] Falkowski-Gilski, Przemysław, Tadeus Uhl. 2020. „Current trends in consumption of multimedia content using online streaming platforms: A user-centric survey". *Computer Science Review* 37 : 100268.

[4] Favale Thomas, Francesca Soro, Martino Trevisan, Idilio Drago, Marco Mellia. 2020. „Campus traffic and e-Learning during COVID-19 pandemic". *Computer Networks* 176 : 107290.

[5] Kampus PG. 2022. https://campus.pg.edu.pl/ (online). dostęp: 20.04.2022.

[6] Lutu, Andra, Diego Perino, Marcelo Bagnulo, Enrique Frias-Martinez, Javad Khangosstar. 2020. „A characterization of the COVID-19 pandemic impact on a mobile network operator traffic". *Proceedings of the ACM Internet Measurement Conference* 1 : 19–33.

[7] Nowicki Krzysztof, Tadeus Uhl. 2017. QoS/QoE in the heterogeneous Internet of things (IoT). *Beyond the Internet of Things*, 165–196. Springer.

[8] Rodhe & Schwarz. 2022. https://www.rohde-schwarz.com/ch/loesungen/broadcast-and-media/overview/broadcast-und-medientechnik-uebersicht_229836.html (online). dostęp: 20.04.2022.

[9] Systemics-PAB. 2022. https://www.syspab.eu (online). dostęp: 20.04.2022.

SYMULACYJNE BADANIA WPŁYWU MECHANIZMU PRIORYTETYZACJI RUCHU NA EFEKTYWNOŚĆ RUCHOWĄ WIELOUSŁUGOWYCH WĘZŁÓW SIECIOWYCH

SIMULATION STUDIES ON THE IMPACT OF TRAFFIC PRIORITIZATION MECHANISM ON EFFICIENCY OF MULTISERVICE NETWORK NODES

Błażej Nowak; Piotr Zwierzykowski;
Politechnika Poznańska, Poznań, błażej.nowak@put.poznan.pl
Politechnika Poznańska, Poznań, piotr.zwierzykowski@put.poznan.pl

DOI: 10.15199/59.2022.4.24

Streszczenie: **Artykuł poświęcony jest badaniom wpływu mechanizmu priorytetyzacji ruchu na efektywność ruchową wielousługowych węzłów sieciowych. Przedstawiono w nim również najważniejsze algorytmy oraz mechanizmy wykorzystane do tworzenia środowiska symulacyjnego zastosowanego w badaniach. Artykuł zawiera wyniki badań które ukazują możliwości oceny wpływu mechanizmu priorytetyzacji ruchu na efektywności ruchową węzła sieciowego..**

Słowa kluczowe: **Symulacja, inżynieria ruchu, węzły sieciowe, priorytety**

Keywords: **Simulation, traffic engineering, prioritization, network nodes.**

1. WSTĘP

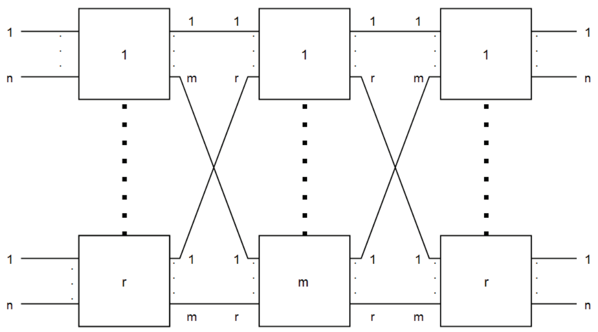
Zwiększająca się różnorodność usług sieciowych, o wzrastającym zapotrzebowaniu na przepływności, oraz rosnąca liczba urządzeń podłączonych do globalnej sieci transmisji danych, stawia nowe wymagania przed sieciami telekomunikacyjnymi. W celu sprostania tym wymaganiom niezbędne jest nie tylko zwiększenie przepływności łączy telekomunikacyjnych, ale także zapewnienie możliwości różnicowania jakości obsługi różnych klas strumieni ruchu w węzłach telekomunikacyjnych. Obszarem badań, będącym tematem tego referatu, jest określenie charakterystyk ruchowych blokowalnych pól komutacyjnych o dobrze znanych strukturach [1-4]. Od wielu lat w węzłach sieciowych o dużych pojemnościach wykorzystywane są pola komutacyjne o strukturze Closa, w tym przede wszystkim pola 3-sekcyjne pola Closa [5-8].

Celem tego artykułu jest opracowanie symulatora wielousługowych blokowalnych pól komutacyjnych o strukturze Closa, stosowanych w węzłach nowoczesnych sieci telekomunikacyjnych, umożliwiającego określenie wpływu charakterystyk ruchowych takich pól na sprawność ruchową węzłów sieci.

Dalsza część artykułu zorganizowana jest w następujący sposób. W rozdziale drugim przedstawiono budowę węzła sieciowego. Rozdział trzeci opisuje algorytmy i mechanizmy wykorzystywane podczas tworzenia środowiska symulacyjnego wykorzystywanego do modelowania węzła. W rozdziale czwartym przedstawiono wyniki przykładowych eksperymentów symulacyjnych. Artykuł kończy krótkie podsumowania zawierające najważniejsze wnioski wypływające z przedstawionych badań.

2. BUDOWA WĘZŁA SIECIOWEGO

Podstawowym elementem węzła sieciowego jest pole komutacyjne. W artykule założymy, że badane węzły mają strukturę pola Closa [5, 9]. Pojedynczy komutator jest najmniejszym elementem wyposażonym w określoną liczbę wejść i wyjść. Pole Closa składa się z nieparzystej liczby sekcji. W artykule przedstawiamy badania przeprowadzone dla 3-sekcyjnego pola Closa o strukturze pokazanej na *Rys.1.*



Rys. 1. Struktura trzysekcyjnego pola Closa

Na *Rys. 1* przyjęto następujące oznaczenia:

n - liczba wejść
r - liczba komutatorów sekcji pierwszej i trzeciej
m - liczba komutatorów sekcji środkowej

Podstawowym elementem sterowania polem jest algorytm wyboru drogi połączeniowej, który określa sposób połączenia wybranego wejścia z wybranym wyjściem pola. W artykule rozważmy pola komutacyjne z algorytmem selekcji punkt-punkt. W takich polach urządzenie sterujące określa komutator pierwszej sekcji, na którego łączu wejściowym pojawiło się zgłoszenie

klasy i. Następnie system sterujący znajduje komutator ostatniej sekcji posiadający wolne łącze wyjściowe, które dysponuje t_i wolnymi jednostkami alokacji w żądanym kierunku. W kolejnym kroku urządzenie sterujące podejmuje próbę zestawienia połączenia pomiędzy wybranym komutatorem pierwszej i ostatniej sekcji. Istnienie wolnej drogi połączeniowej umożliwia realizację połączenia. W przypadku braku możliwości zestawienia połączenia, zgłoszenie jest tracone z powodu blokady wewnętrznej. Jeżeli żaden z komutatorów ostatniej sekcji nie dysponuje wolnymi t_i wolnymi zasobami w żądanym kierunku, zgłoszenie jest tracone z powodu blokady zewnętrznej [9].

3. ŚRODOWISKO SYMULACYJNE

W środowisku symulacyjnym przyjęto następujące zasady działania systemu z priorytetami:

- wszystkie oferowane systemowi klasy zgłoszeń mają przypisany priorytet,
- obsługa zgłoszeń klas o niższym priorytecie nie ma wpływu na prawdopodobieństwo blokady zgłoszeń o wyższym priorytecie,
- w sytuacji braku wolnych zasobów pojawienie się nowego zgłoszenia o wyższym priorytecie powoduje zakończenie obsługi (wyparcie) aktualnie obsługiwanych zgłoszeń o niższym priorytecie.

Symulator trzysekcyjnego pola Closa z ruchem wielousługowym i priorytetyzacją został zaimplementowany w języku programowania C# z wykorzystaniem techniki programowania obiektowego. Symulacja opiera się na metodzie przeglądania zdarzeń Metoda ta bazuje na posortowanej liście zdarzeń względem czasu ich występowania [10]. Algorytmem selekcji drogi połączeniowej jest algorytm punkt-punkt (przedst. w r.2). Algorytm ten został rozszerzony o możliwość obsługi priorytetów zgłoszeń (*Rys.2*). Opracowany symulator jest przystosowany do obsługi mieszaniny wielu klas ruchów wielousługowych z priorytetami. Środowisko symulacyjne pozwala na badania różnych rodzajów ruchów tj. ruchów typu Erlang, Engset i Pascal, często oznaczanych BPP (Bernoulli, Poissona, Pascal). Danymi wejściowymi symulatora są:

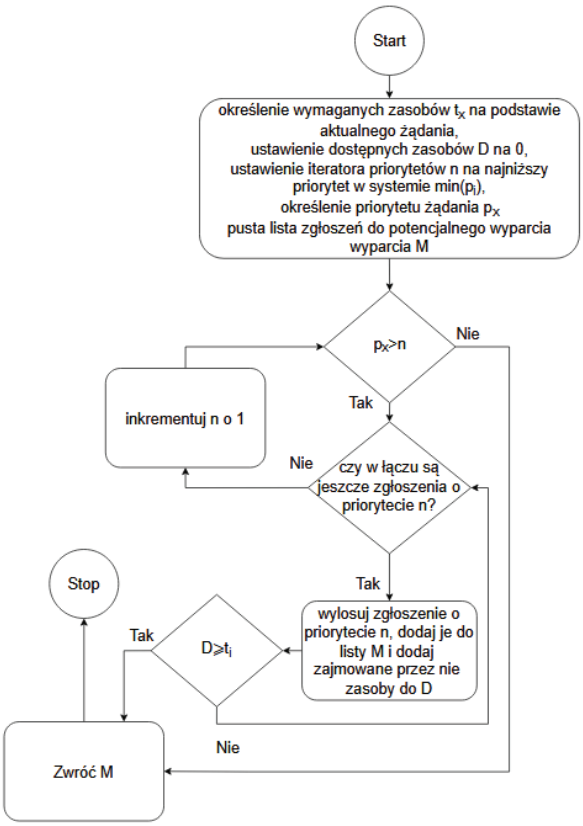
- parametry trzysekcyjnego pola Clossa (*Rys.1*):
  - m - liczba komutatorów sekcji środkowej
  - n - liczba łączy wejściowych w komutatorze pierwszej sekcji i wyjściowych w komutatorze ostatniej sekcji
  - r - liczba komutatorów pierwszej i ostatniej sekcji
- f- pojemność pojedynczego łącza
- lista klas z określonym typem ruchu (Erlang, Engset, Pascal) i ewentualną liczbą źródeł ruchu n - $C_{i,\{er,en_n,pa_n\}}$
- lista zasobów żądanych przez każdą z klas - t_i
- priorytet dla każdej z klas - p_i
- a - średnie natężenie na jednostkę alokacji

W wyniku zastosowania metody przeglądania zdarzeń zostały zdefiniowane dwa zdarzenia:

- pojawienie się żądania obsługi,
- koniec obsługi żądania.

Gdy dojdzie do zdarzenia "pojawienie się żądania" sprawdzany jest stan pola pod kątem wolnych zasobów z uwzględnieniem wybranego algorytmu selekcji. Gdy nie znaleziono dostatecznych zasobów symulator zaczyna brać pod uwagę priorytet żądania i gromadzić żądania do potencjalnego wyparcia na rzecz wyższego priorytetu. Jeżeli w zgromadzonych zgłoszeniach zostanie zapewniona dostateczna liczba zasobów to zostaną one wyparte - stracone a zgłoszenie o wyższym priorytecie wchodzi do pola komutacyjnego. Gdy po uwzględnieniu priorytetów nadal nie ma na przychodzące zgłoszenie dostatecznej liczby zasobów to zostaje ono stracone. Niezależnie od tego czy żądanie zostało przyjęte planowane jest kolejne zdarzenie typu "pojawienie się żądania" i "koniec obsługi żądania". Gdy dojdzie do zdarzenia "koniec obsługi żądania" zasoby pola komutacyjnego są zwalniane a zgłoszenie zostało obsłużone.

W przypadku kiedy dochodzi do wyparcia zgłoszeń, z łącza które będzie częścią drogi połączeniowej, z racji ich niższego priorytetu pojawia się kwestia w jaki sposób wybrać te zgłoszenia. W omawianym symulatorze zaimplementowany został najprostszy algorytm: w systemie o N priorytetach zgłoszenia do potencjalnego wyparcia przez klasę o priorytecie p_i wybierane są rozpoczynając od najniższego priorytetu $min(p_n)$, a kończąc na priorytecie p_{i-1}. Wybór jest losowy, równomierny. Gdy wyczerpią się zgłoszenia priorytetu n, a nie zapewniono jeszcze dostatecznych zasobów symulator zaczyna losować z puli zgłoszeń o priorytecie $n + 1$.



Rys. 2. Algorytm wybieranie zgłoszeń do wyparcia na rzecz wyższego priorytetu z łącza biorącego udział w drodze połączeniowej.

Proces wypierania obejmuje całą drogę połączeniową. Dlatego trzeba zwrócić uwagę, że aby doszło do realizacji połączenia w żądanym kierunku to określone łącza wejściowe i wyjściowe każdej sekcji muszą zapewnić żądaną liczbę zasobów a więc z każdego łącza na drodze połączeniowej będą wypierane zgłoszenia.

4. Wyniki przykładowych eksperymentów symulacyjnych

Opracowany program symulacyjny został wykorzystany do zbadania trzysekcyjnego pola Closa o następujących parametrach:

- liczba wejść w komutatorze pierwszej sekcji $n = 4$
- pojemność pojedynczego łącza $f = 40$
- liczba komutatorów w pierwszej i drugiej sekcji $m = r = 4$
- algorytm selekcji punkt-punkt

Eksperyment symulacyjny był powtarzany 8 razy dla różnych konfiguracji priorytetów ruchu oferowanego przy czym parametry które pozostawały niezmienne to:

- liczby żądanych jednostek alokacji każdej klasy i jej typ: $t_0 = 4$, $C_{0,er}$, $t_1 = 5$, $C_{1,pa_{300}}$, $t_2 = 8$, $C_{2,en_{300}}$, $t_3 = 10$, $C_{3,er}$
- stosunek ruchu oferowanego:

$$a_0 \cdot t_0 = a_1 \cdot t_1 = a_2 \cdot t_2 = a_3 \cdot t_3$$

Schemat priorytetów przyjęty w poszczególnych eksperymentach:

- eksperyment 1: $p_0 = p_1 = p_2 = p_3$
- eksperyment 2: $p_0 > p_1 > p_2 > p_3$
- eksperyment 3: $p_0 < p_1 < p_2 < p_3$
- eksperyment 4: $p_0 > p_1 = p_2 = p_3$
- eksperyment 5: $p_0 = p_1 = p_2 < p_3$

Wyniki każdego z eksperymentów są zaprezentowane w formie tabel i wykresów w zależności od natężenia na jednostkę alokacji (1) w systemie:

$$a = \sum_i \frac{a_i \cdot t_i}{n \cdot m \cdot f} \quad (1)$$

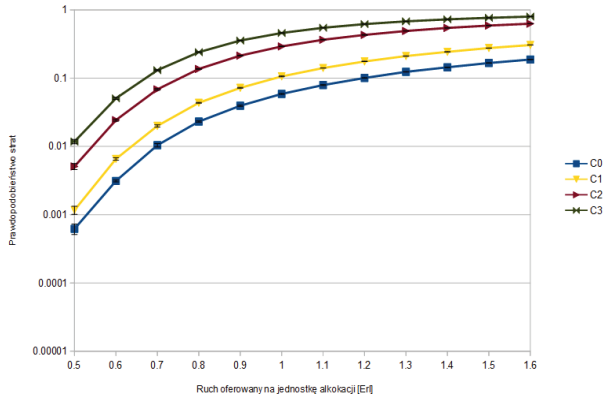
Gdzie i to indeks klasy, a_i, t_i to odpowiednio natężenie i liczba żądanych zasobów z kolei n, m, f to liczba wejść komutatora pierwszej sekcji, liczba komutatorów pierwszej sekcji i pojemność pojedynczego łącza. Każdy punkt jest opatrzony 95% przedziałem ufności dla serii 7 symulacji obliczonym na podstawie wzoru

$$(\underline{X} - t_\alpha \sqrt{\frac{\sigma^2}{d}}; \underline{X} + t_\alpha \sqrt{\frac{\sigma^2}{d}}) \quad (2)$$

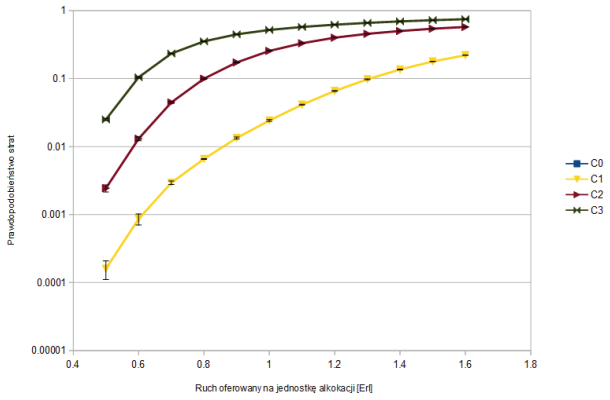
gdzie $\underline{X}$ to średnia z d prób, t_α to wartość rozkładu t-studenta dla $d - 1$ stopni swobody a σ^2 to wariancja.

Długość każdej symulacji to 100 tysięcy jednostek czasu przy czym czas stabilizacji systemu został przyjęty na 200 jednostek czasu. Czas potrzebny na wykonanie eksperymentu mieścił się w przedziale 15 - 25 minut na komputerze z systemem operacyjnym Windows 10 z procesorem Intel Core i5-4570 CPU 3.20 GHz i pamięcią RAM 16GB. Wyniki eksperymentów 1-3 zostały przedstawione na *Rys. 3, 4, 5* a wyniki eksperymentów 4 i 5 na *Rys. 6, 7*.

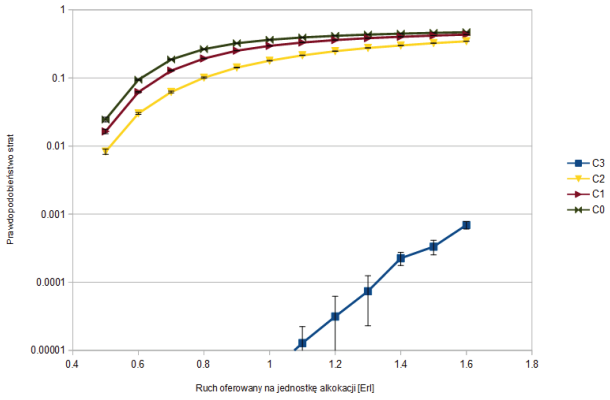
W pierwszym eksperymencie *(Rys. 3)* wyniki są typowe dla pola komutacyjnego z ruchem wielousługowym, czyli zgłoszenia żądające więcej zasobów mają jednocześnie większe prawdopodobieństwo strat. Po nałożeniu priorytetów w eksperymencie 2 *(Rys. 4)* widać, że klasa o najwyższym priorytecie ma na tyle małe prawdopodobieństwo strat, że nie występuje na wykresie. Dzieje się tak gdyż ta klasa ma całe pole na wyłączność - jeżeli miałoby dojść do zapełnienia łączy w polu i do utraty zgłoszenia dla klasy o najwyższym priorytecie to zanim to nastąpi sprawdzana jest zawartość pola pod kątem obsługiwanych zgłoszeń o niższych priorytetach (tutaj C_1, C_2, C_3) jeżeli takie zgłoszenia są aktualnie w polu to zostaną one wyparte przez nadchodzące zgłoszenie jeżeli wyparcie ich jest w stanie zapewnić dostateczne zasoby dla t_0.



Rys.3. Prawdopodobieństwo strat całkowitych dla eksperymentu 1.



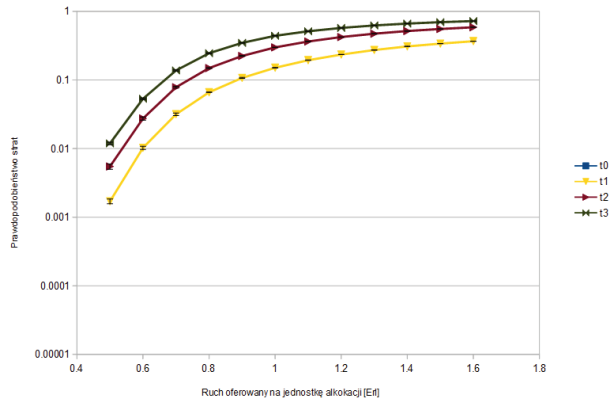
Rys.4. Prawdopodobieństwo strat całkowitych dla eksperymentu 2.



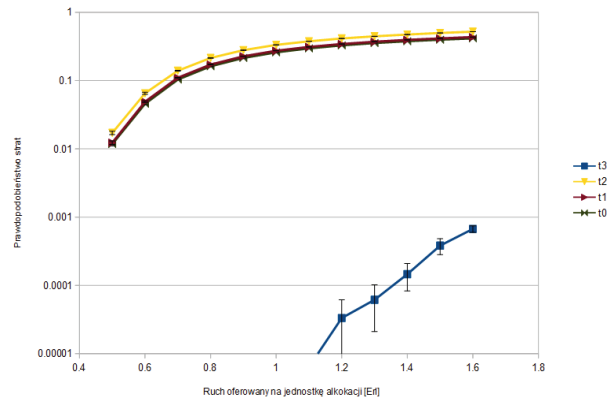
Rys.5. Prawdopodobieństwo strat całkowitych dla eksperymentu 3.

Analizując wyniki eksperymentu 2 *(Rys. 4)* w odniesieniu do eksperymentu 1 *(Rys. 3)* widać, że klasy ruchu o niższych priorytetach jak C_2, C_3 mają porównywalne prawdopodobieństwo strat. Można to wytłumaczyć wykładniczym kształtem rozkładu czasu obsługi zgłoszeń. Gdy nie mamy priorytetów a węzeł przyjmie zgłoszenie które będzie niefortunnie długo obsługiwane to będzie ono zabierało zasoby aż do końca obsługi. W przypadku gdy nałożone są priorytety to zgłoszenia które zabierają zasoby zbyt długo a mają niski priorytet zostaną wyparte przez wyższe priorytety zwalniając miejsce na zgłoszenia potencjalnie krótszym czasie obsługi co poprawia ogólną wydajność pola.

Porównując eksperyment 2 *(Rys. 4)* i 3 *(Rys. 5)* można zauważyć, że ustawienie priorytetów zgodnie z zasadą im większe żądanie tym wyższy priorytet zbliżamy do siebie prawdopodobieństwa strat klas niższych podczas gdy ustawienie priorytetów zgodnie z zasadą im mniejsze żądanie tym większy priorytet zachowuje odstępy między prawdopodobieństwami strat. W obu przypadkach prawdopodobieństwa strat dla najwyższego priorytetu są drastycznie obniżone w porównaniu do braku priorytetów.



Rys.6. Prawdopodobieństwo strat całkowitych dla eksperymentu 4.



Rys.7. Prawdopodobieństwo strat całkowitych dla eksperymentu 5.

Porównując 5 *(Rys. 7)* z 3 *(Rys. 5)* można zaobserwować, że klasy o najwyższym priorytecie mają takie same prawdopodobieństwa strat co potwierdza wcześniejsze stwierdzenie, że klasy o najwyższym priorytecie mają do dyspozycji całe pole na wyłączność. Podobna sytuacja zachodzi dla porównania eksperymentu 4 *(Rys. 6)* z 2 *(Rys. 4)*.

Odnosząc wyniki eksperymentu 4 *(Rys. 6)* do 2 *(Rys. 4)* można dojść do wniosku, że priorytetyzacja od pewnego momentu daje znikomy zysk w prawdopodobieństwie strat. W eksperymencie 2 *(Rys. 4)* klasy C_2 *i* C_3 mają zbliżone prawdopodobieństwa strat do eksperymentu 4 *(Rys. 6)* mimo tego, że klasy C_1, C_2, C_3 mają ten sam

priorytet natomiast różnica jest widoczna dla klasy C_1 która zauważalnie traci. Podobna sytuacja jest widoczna przy porównaniu eksperymentu 5 *(Rys. 7)* i 3 *(Rys. 5)*, klasa C_1 wyraźnie traci.

5. PODSUMOWANIE

W artykule przedstawiono małą część spośród przeprowadzonych przez autorów badań. Badania obejmowały różną liczbę klas ruchu, różne ich rodzaje udziały w całym ruchu z uwzględnieniem wielu scenariuszy przypisania priorytetów do określonych klas ruchu. Z uwagi na ograniczenia długości artykułu, tylko niektóre wyniki mogły być przedstawione w artykule. Analizując wyniki przeprowadzonych eksperymentów można wyciągnąć następujące wnioski:

- nadając priorytety różnym klasom ruchu możemy skutecznie manipulować prawdopodobieństwem strat części klas,
- wpływ nałożenia priorytetu maleje wraz z liczbą nałożonych priorytetów,
- dla najwyższego priorytetu wszystkie inne zgłoszenia nie istnieją i mogą być wyparte z pola w każdej chwili,
- dla klas o priorytecie innym niż najwyższy w stratach pojawia się składnik wynikający z wyparcia.

Zgodnie z najlepszą wiedzą autorów opracowany symulator jest pierwszym symulatorem wielousługowych pól komutacyjnych na poziomie zgłoszeń, który uwzględnia priorytetyzację klas ruchu. W poprzednich pracach opisane są symulatory pól komutacyjnych z mechanizmem rezerwacji zasobów dla poszczególnych klas ruchu [9]. Mechanizm ten można traktować jako mechanizm dwustopniowej priorytetyzacji. Ograniczeniem tego rozwiązania są jednak tylko dwa poziomy priorytetów (rezerwacja lub jej brak) oraz całkowite wyłączenie zasobów zarezerwowanych z ogółu zasobów sieciowych tzn. nawet gdy nie są one wykorzystywane przez klasy dla których zostały przeznaczone to nie wykorzystają je inne klasy ruchu.

Rozwiązanie przedstawione w artykule, którego działanie ilustrują wyniki przeprowadzonych eksperymentów symulacyjnych nie ma takich ograniczeń. Może zostać zastosowane dla wielu różnych mieszanin ruchu, w tym ruch multimedialnego, o zróżnicowanych żądaniach i priorytetach.

W przyszłości symulator zostanie użyty do weryfikacji modeli analitycznych, które pozwolą obliczać prawdopodobieństwa strat w polach komutacyjnych obsługujących wielousługowe ruchy z nałożonymi priorytetami. W planach rozwoju symulatora jest między innymi optymalizacja i implementacja algorytmów wypierania klas niższych na rzecz klas o wyższych priorytetach.

Zastosowanie mechanizmu priorytetyzacji może umożliwić optymalizację sieci rozumianą jako polepszenie wydajności ruchowej węzła sieciowego. Można go również wykorzystać w celach analiz ruchowych skutków działań promocyjnych operatorów, które nie tylko mogą zwiększyć obciążenie węzłów, ale również wprowadzić priorytetyzację obsługi różnych użytkowników sieci.

PODZIĘKOWANIA

Badania zostały sfinansowane ze środków Ministerstwa Edukacji i Nauki (projekt nr 0313/SBAD/1307).

LITERATURA

[1] Wojciech Kabaciński: Nonblocking Electronic and Photonic Switching Fabrics, Springer,2005

[2] M. Sobieraj, P. Zwierzykowski, E. Leitgeb: Simulation Studies of Elastic Optical Networks Nodes with Multicast Connections, Human-centric Computing and Information Sciences, 2022, vol. 12, pp. 05-1-05-13

[3] M. Stasiak, M. Sobieraj, P. Zwierzykowski: Modelling of multi-service switching networks with multicast connections, IEEE Access, 2022, vol. 10, pp. 5359-5377

[4] M. Głąbowski, H. Ivanov, E. Leitgeb, M. Sobieraj, M. Stasiak, "Simulation studies of elastic optical networks based on 3-stage Clos switching fabric," Optical Switching and Networking, vol. 36, 2020. https://doi.org/10.1016/j.osn.2020.100555

[5] C. Clos: A study of non-blocking switching networks, Bell Labs Tech. J., vol. 32, no. 2, pp. 406-424, 1953

[6] V. E. Benes: A 'thermodynamic' theory of traffic in connecting networks, Bell Labs Tech. J., vol. 42, no. 3, pp. 567-607, May 1963.

[7] H.J.S. Dorren, N. Calabretta, O. Raz: A 3-stage CLOS architecture for high-throughput optical packet switching, Proceedings of Asia Communications and Photonics Conference and Exhibition, Shanghai, China, 2009, 1-6.

[8] A. Aakanksha, M. Shailendra, G.Tarun: Routing in all-optical three stage-Clos interconnection networks,' International Journal of Computer Science and Engineering, 2012, 2(1), pp. 19–22.

[9] M. Sobieraj: Modelowanie pól komutacyjnych z mechanizmami progowymi i wielousługowymi źródłami ruchu, Poznań University of Technology, PhD dissertation, 2014.

[10] J. Tyszer: Object-Oriented Computer Simulation of Discrete-Event Systems, Springer, ISBN: 0792385063, 1999.

ANALIZA JAKOŚCI SZYFROWANEJ KOMUNIKACJI VOIP W GRANICZNYCH WARUNKACH CZUŁOŚCI ODBIORNIKA RADIOWEGO

THE ANALYSIS OF ENCRYPTED VOIP COMMUNICATION QUALITY IN LIMITING RADIO RECEIVER SENSITIVITY CONDITIONS

Rafał Polak[1]; Jarosław Wojtuń[2]; Piotr Łubkowski[3]

[1] Wojskowa Akademia Techniczna, Warszawa, rafał.polak@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, jaroslaw.wojtun@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Warszawa, piotr.lubkowski@wat.edu.pl

DOI: 10.15199/59.2022.4.25

Streszczenie: **Zasadniczym wyzwaniem związanym z realizacją współczesnych usług komunikacji multimedialnej jest zapewnienie wysokiego poziomu bezpieczeństwa, wysokiej dostępności i wysokiej jakości usług w zróżnicowanym heterogenicznym środowisku sieciowym. Implementacja dowolnego algorytmu szyfrowania może skutkować zasadniczym pogorszeniem jakości zwłaszcza w warunkach pogorszenia parametrów kanału radiowego. Zjawisko to jest szczególnie często obserwowane na granicy czułości odbiornika. W artykule przedstawiono metodykę i wyniki badania wpływu szyfrowania SCIP na jakość komunikacji VoIP w szerokopasmowym kanale radiowym w granicznych warunkach czułości odbiornika radiowego. Pomiary przeprowadzono za pomocą testera MultiDSLA oraz metod PESQ i POLQA.**

Abstract: **The main challenge associated with the implementation of modern multimedia communication services is to ensure a high level of security, high availability and high-quality services in the heterogenous network environment. Implementation of any encryption algorithm may lead to a deterioration of quality especially in the radio communication networks due to changes in the radio channel parameters. This phenomenon is particularly often observed at the sensitivity limit of the radio channel. The article presents the methodology and results of the study of the impact of SCIP encryption, on the quality of VoIP communication in in a wideband radio channel in limiting condition of channel sensitivity. The measurements were performed using the MultiDSLA tester and the PESQ and POLQA methods.**

Słowa kluczowe: **testowanie jakości VoIP, szyfrowanie SCIP, szerokopasmowa transmisja radiowa.**

Keywords: **voice quality testing, SCIP encryption, wideband radio communication.**

1. WSTĘP

Zapewnienie odpowiedniego poziomu bezpieczeństwa transmisji danych i głosu w systemach łączności jest jednym z najważniejszych aspektów we współczesnej komunikacji sojuszniczej i koalicyjnej. Protokół SCIP umożliwia wykorzystanie mechanizmów kryptograficznych do szyfrowania transmisji w sposób zestandaryzowany, zapewniając interoperacyjność istniejących i nowo tworzonych sojuszniczych i koalicyjnych systemów łączności. Ważną zaletą protokołu SCIP jest wykorzystanie klasycznego modelu usług VoIP. Wymagania dotyczące transmisji w czasie rzeczywistym przez bezpieczne kanały komunikacji (tj. komunikacji głosowej) są definiowane wokół ograniczeń funkcjonalnych i jakościowych. Wymagania dotyczące jakości w warstwie aplikacji są zdefiniowane w zaleceniu ITU-T G.1010 i obejmują parametry związane z nawiązywaniem, rozłączaniem i blokowaniem połączeń. Niezwykle ważne jest jednak zapewnienie jakości na poziomie sieci, ponieważ umożliwia ono przewidywanie wielkości wymaganego pasma oraz poziomu strat w warstwie aplikacji.

W odniesieniu do poufności, integralności i uwierzytelniania aplikacji VoIP, należy również zwrócić uwagę na zmniejszenie efektywnej szerokości pasma. Może to prowadzić do ogólnego obniżenia jakości usługi VoIP w związku ze wzrostem liczby traconych pakietów, zwiększeniem wartości opóźnienia i jittera. Problem dostępności pasma jest szczególnie widoczny w systemach łączności radiowej, tak charakterystycznych dla taktycznej łączności wojskowej. Kanał radiowy ulega degradacji ze względu na występujące pogorszenie stosunku sygnału do szumu (*SNR*), wielościeżkową propagację czy zakłócenia dopplerowskie. Negatywny wpływ na warunki propagacji mają także zaniki oraz wysoki poziom zakłóceń. Dodatkowo, w przeciwieństwie do wąskopasmowego kanału radiowego, szerokopasmowe kanały radiowe charakteryzują się mniejszą odpornością na ukierunkowane zakłócenia. Niezmiernie istotnym ograniczeniem jest także czułość odbiornika na granicy zakresu. Zjawisko powyższe może negatywnie wpływać na jakość komunikacji.

Problematyka wpływu szyfrowania na jakość komunikacji jest przedmiotem szeregu publikacji. Wyniki prac prezentowane w [6], [9] dotyczą szyfrowanych usług głosowych realizowanych w oparciu o szerokopasmowe systemy komunikacji przewodowej i bezprzewodowej. Z kolei w **Błąd! Nie można odnaleźć źródła odwołania.** przedstawione zostały uwarunkowania realizacji usługi VoIP w środowisku sieci kognitywnych. Zależność pomiędzy wpływem szyfrowania na jakość realizacji usługi głosowej dyskutowana jest w [4], [5]. Prezentowane opracowania nie odnoszą się jednak do oceny realizacji usługi szyfrowanej komunikacji VoIP w środowisku szerokopasmowych sieci radiowych w warunkach ograniczonej czułości, które to zjawisko często towarzyszy tego typu sieciom.

Zatem problemem badawczym prezentowanym w dalszej części artykułu jest ocena jakości ruchu VoIP szyfrowanego protokołem SCIP w granicznych warunkach czułości odbiornika radiowego. Wymiernym efektem przeprowadzonych badań jest opracowana metodyka badania wpływu protokołów szyfrujących na jakość komunikacji głosowej w kanałach radiowych.

2. CHARAKTERYSTYKA ŚRODOWISKA BADAWCZEGO

2.1. Protokół SCIP

SCIP – *Secure Communications Interoperability Protocol* jest zespołem norm, opracowywanym dla pojedynczych użytkowników wymagających bezpiecznej komunikacji multimedialnej w relacji „end to end" we wszelkich typach systemów. Jest to technologia normalizowane przez NATO (*ADatP34G*) i zalecana do użytku w telekomunikacyjnych systemach koalicyjnych i narodowych.

Technologia SCIP określa całe spektrum zagadnień związanych z tworzeniem terminali końcowych systemu, normalizując sposób cyfryzacji sygnałów mowy, sposób transmisji danych użytkownika i sposób utajniania. Normalizowana jest także sygnalizacja terminali z tym związana, a wynikowy strumień danych cyfrowych może być transmitowany przez dowolne medium fizyczne, zależne od typu terminala (np. radiostacja osobista, radiostacja KF/UKF, telefon VoIP itp.). Konsekwentne stosowanie tej normy zapewnia ujednolicenie struktury wyżej wymienionych terminali, w tym radiostacji, gdzie obecnie panuje duża dowolność.

Technologię SCIP charakteryzuje możliwość wykorzystania w wojskowych systemach opartych na protokołach IPv4 i IPv6, a także w obecnie wykorzystywanych systemach wojskowych i cywilnych, dzięki następującym cechom:

- wymaga niewielkich przepustowości dzięki bardzo oszczędnemu kodekowi fonicznemu (MELPe), o szybkości 2,4 kb/s, co umożliwia pracę poprzez wąskopasmowe środki radiowe KF/UKF, natomiast w przypadku wystarczających zasobów pasma jest używany kodek G.729D (7,2 kb/s) o wysokiej jakości;
- jest projektowana do pracy simpleksowej, dupleksowej oraz układów punkt – wielopunkt.

2.2. Szerokopasmowy system komunikacji radiowej

Rozwój współczesnych systemów dowodzenia i kierowania nierozerwalnie wiąże się ze stale rosnącymi potrzebami w zakresie wymagań, którym sprostać muszą systemy łączności. Od kilku lat rozwój wojskowych systemów łączności ukierunkowany jest na technologie szerokopasmowe i rozwiązania związane z wykorzystaniem protokołu IP [1], [3]. Radiostacje szerokopasmowe pozwalają na tworzenie łączy bezprzewodowych o dużej przepustowości, które umożliwiają jednoczesna transmisję głosu i danych przy zapewnieniu odpowiedniej jakości dla tych usług. Ponadto radiostacje szerokopasmowe umożliwiają tworzenie sieci o elastycznej architekturze. Radiostacje stale monitorują swoje otoczenie i w razie potrzeby automatycznie tworzą nowe trasy routingu. Z drugiej strony należy pamiętać, że radiostacje szerokopasmowe pracujące w ramach jednej sieci prócz ruchu od użytkowników muszą stale przekazywać ruch utrzymaniowy. Wielkość tego ruchu zwiększa się wraz ze wzrostem liczy węzłów (radiostacji) w sieci radiowej.

W badaniach wykorzystane zostały dwa róże rozwiązania pozwalające na zbudowanie szerokopasmowych sieci radiowych, tj. radiostacje R-450C firmy Transbit oraz radiostacje AN/PRC-117G firmy L3Harris. Radiostacje te należą do urządzeń najnowszej generacji opartych na technologii radia definiowanego programowo – SDR (*Software Defined Radio*).

Radiostacja R-450C umożliwia tworzenie bezprzewodowych sieci pakietowych IPv4 i IPv6 [8]. Pracuje w zakresie częstotliwości od 225 MHz do 40 0MHz z kanałem o szerokości od 1 MHz, 2 MHz i 4 MHz. Radiostacja umożliwia tworzenie sieci MANET o przepustowości do 8 Mb/s dla relacji i nie mniej niż 40 Mb/s dla sieci, przy zasięgach ograniczonych do 6 km w ruchu i do 15 km na postoju. Ze względu na duże przepustowości sprzęt ten przeznaczony jest do realizacji rozszerzonych funkcji BMS, w tym transmisji obrazu wideo w czasie rzeczywistym.

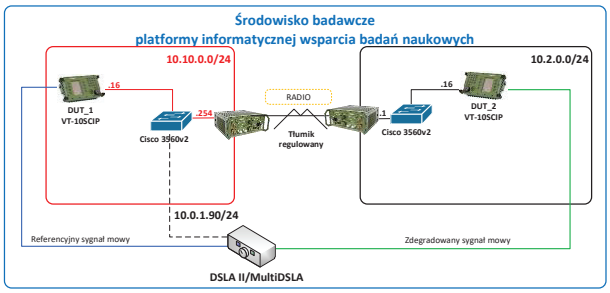
Radiostacja AN/PRC-117G pozwala na budowę sieci szerokopasmowych [7]. Umożliwia transmisję danych pakietowych z protokołem IPv4. Szerokopasmowy tryb pracy ANW2C (*Adaptive Networking Wideband Waveform Revision C*) pracuje w zakresie częstotliwości od 225 MHz do 2 GHz z kanałem o szerokości 0,5MHz, 1,2 MHz, 2,5 MHz lub 5 MHz. Maksymalna liczba radiostacji w sieci ANW2C wynosi 30. ANW2C wykorzystuje technologię dynamicznej alokacji kanałów w celu zwiększenia przepustowości sieci do 10 Mb/s. Bezpieczeństwo łączności jest zapewnione poprzez moduł szyfrujący klasy Type-1, certyfikowany przez NSA (*National Securyty Agency*), który umożliwia przesyłanie danych niejawnych o klauzuli Top Secret.

3. METODYKA BADAWCZA

Celem oceny wpływu granicznych wartości czułości kanału radiowego na jakość szyfrowanej komunikacji VoIP opracowana została metodyka badawcza bazująca na wykorzystaniu narzędzia pomiarowego w postaci dedykowanego testera MultiDSLA. Tester wyposażony jest w interfejs mikrotelefonu pozwalający na bezpośrednie podłączenie do toru rozmównego. Opcja ta została wykorzystana w przeprowadzonych badaniach. Metodologia badań została opracowana i zweryfikowana we wcześniejszych pracach prowadzonych przez autorów w związku z testowaniem usług czasu rzeczywistego w sieciach heterogenicznych [9].

Badania dotyczące wpływu protokołu SCIP na jakość połączenia VoIP zostały przeprowadzone w testowym stanowisku eksperymentalnym, które składa się z dwóch domen, zwanych odpowiednio D_1 oraz D_2 (Rys. 1). W każdej domenie znajduje się terminal głosowy VT-10SCIP (oznaczony jako DUT - Device Under Test), w którym zaimplementowano protokół SCIP oraz dwie

radiostacje. W badaniach wykorzystano dwie sieci radiowe. Połączenie VoIP jest zestawiane pomiędzy terminalami VT-10SCIP, a ruch głosowy jest przesyłany z DUT_1 do DUT_2 przez kanał radiowy z regulowanym tłumikiem. W związku z tym wszelkie zakłócenia mogą wpływać na poziom oceny jakości odbieranego sygnału mowy, który jest wyrażany za pomocą wskaźnika MOS (*Mean Opinion Score*). Wartości MOS zawierają się w przedziale od 1 do 5, przy czym wartość 5 oznacza najlepszą jakość mowy.



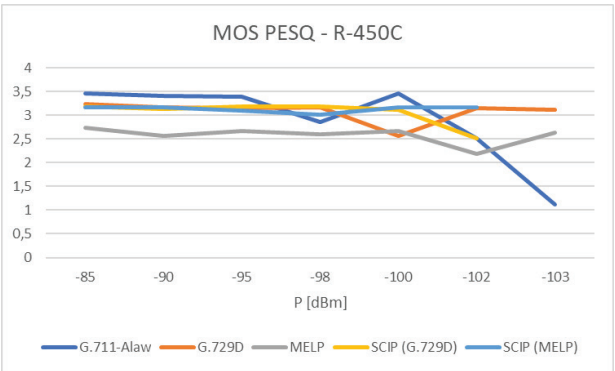
Rys. 1. Schemat stanowiska pomiarowego

Przedstawiony na schemacie tester DSLA II (*Digital Speech Level Analyzer*) wraz z aplikacją MultiDSLA stanowi profesjonalny system do pomiaru i analizy jakości głosu. Dostarcza on referencyjny sygnał głosowy dla DUT-1 i odbiera zniekształcony sygnał od DUT_2. Sygnały te stanowią podstawę do oszacowania jakości sygnału mowy za pomocą obiektywnych metod pomiarowych PESQ (*Percepcyjna Ocena Jakości Mowy, ang. Perceptual Evaluation of Speech Quality*) i POLQA (*Percepcyjna Obiektywna Ocena Jakości Słuchu, ang. Perceptual Objective Listening Quality Analysis*). Do badań wykorzystywane zostały pliki referencyjne reprezentujące próbki mowy żeńskiej i męskiej o odpowiedniej mocy sygnału zgodnie z ITU-T P.862 oraz ITU-T P.862.3.

Podczas badań wykorzystano następujące kodeki mowy obsługiwane przez DUT_1: G.711-Alaw, jako kodek odniesienia o wysokiej przepływności 64 kb/s i kodeki G.729D oraz MELP, które były badane bez i z szyfrowaniem za pomocą SCIP. Przyjmując za podstawę opracowany scenariusz obejmujący szyfrowaną komunikację VoIP w kanale radiowym o zmiennej charakterystyce, dokonano oceny jakości transmisji sygnału mowy za pomocą metod PESQ i POLQA. Poniżej przedstawiono wyniki przeprowadzonych badań z uwzględnieniem scenariuszy i przyjętych uwarunkowań.

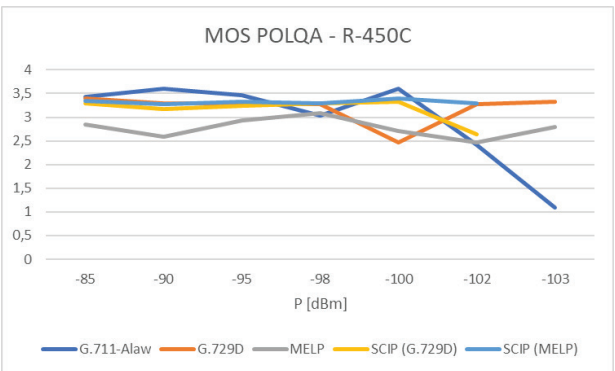
4. ANALIZA WYNIKÓW

Badanie zostały przeprowadzone dla warunków brzegowych pracy radiostacji na granicy czułości odbiornika radiowego. Podczas badań metodą empiryczną zweryfikowano, że pierwsze zmiany w transmisji głosowej pojawiają się przy mocy odbieranej na poziomie -90 dBm. W związku z tym w artykule przedstawiono wyniki badań dla mocy odbieranej radiostacji na poziomie od -85 dBm do całkowitej utraty transmisji. Na wykresie (Rys. 2) przedstawiono ocenę jakości mowy metodą PESQ z wykorzystaniem radiostacji R-450C. Transmisja z wykorzystaniem kodeków charakteryzującymi się niską przepływnością (G.729D i MELP) bez szyfrowania była realizowana do utraty łączności na podobnym poziomie, transmisja z wykorzystaniem kodeka G.711-Alaw również była realizowana do utraty łączności natomiast od mocy odbieranej na poziomie -100 dBm widać wyraźny spadek jakości mowy. Dla transmisji szyfrowanej protokołem SCIP zarówno dla kodeka G.729D, jak i MELP transmisja została przerwana na poziomie mocy odbieranej wynoszącej -102 dBm.



Rys. 2. Ocena jakości metodą PESQ dla R-450C

Wyniki badań oceny jakości mowy metodą POLQA z wykorzystaniem radiostacji R-450C (Rys. 3) są zbieżne z wynikami otrzymanymi metodą PESQ.

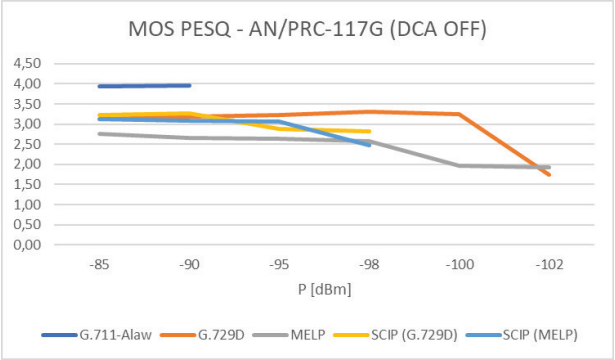


Rys. 3. Ocena jakości metodą POLQA dla R-450C

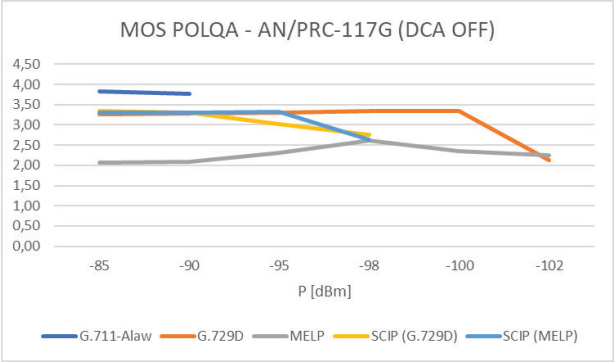
Na kolejnych wykresach (Rys. 4 i Rys. 5) przedstawiono wyniki badań oceny transmisji mowy metodami PESQ i POLQA z wykorzystaniem radiostacji AN/PRC-117G. Należy pamiętać, że transmisja danych z wykorzystaniem radiostacji AN/PRC-117G i szerokopasmowego trybu pracy ANW2C jest już na poziomie radiostacji zabezpieczona kryptograficznie poprzez moduł szyfrujący klasy Type-1. Radiostacje AN/PRC-117G poosiadają dodatkową opcję umożliwiającą dynamiczny przydział zasobów do kanału radiowego (DCA, *Dynamic Capacity Allocation*). Włączenie trybu DCA zapewnia większą szybkość transmisji danych przy jednoczesnym zwiększeniu opóźnień w sieci. Badania zostały przeprowadzone przy

wyłączonej opcji DCA. W badaniach tryb ANW2C wykorzystywał kanał radiowy o szerokości 1,2 MHz w zakresie częstotliwości od 350,2 MHz do 351.4 MHz.

Przedstawione wyniki wskazują, że dla połączeń szyfrowanych dodatkowo protokołem SCIP następuje gwałtowne zerwanie transmisji na poziomie -98 dBm, co wynika z faktu braku możliwości zestawienia sesji sygnalizacyjnej SCIP z uwagi na duży czas opóźnienia w kanale radiowym. Dla transmisji nieszyfrowanej protokołem SCIP takie uwarunkowania nie występują.



Rys. 4. Ocena jakości metodą PESQ dla AN/PRC-117G w trybie pracy z wyłączonym DCA



Rys. 5. Ocena jakości metodą POLQA dla AN/PRC-117G w trybie pracy z wyłączonym DCA

Wartości MOS uzyskane dla kodeka G.711 są większe niż w przypadku radiostacji R-450C. Wynika to z większej przepustowości sieci radiowej. Wykorzystywany jest tutaj kanał o szerokości 1,2 MHz. Radiostacja R-450C wykorzystuje kanał 1 MHz.

5. WNIOSKI

W artykule przedstawiony został wpływ warunków brzegowych pracy radiostacji na granicy czułości odbiornika radiowego na jakość szyfrowanej komunikacji VoIP, gdzie szyfrowane zostało zrealizowane z wykorzystaniem protokołu SCIP normalizowanego przez NATO i rekomendowanego do użytku w koalicyjnych i narodowych systemach łączności.

Otrzymane wyniki wskazują jednoznacznie na wpływ warunków brzegowych na jakość komunikacji VoIP, a zatem również wskazują na zależność jakości QoS od warunków realizacji połączenia.

Uzyskane wyniki wskazują również, że nawet przy bardzo zdegradowanych parametrach granicznych istnieje możliwość komunikacji głosowej o akceptowalnej jakości. Warto również zauważyć, że dla obu typów zastosowanych radiostacji uzyskano podobne wartości współczynnika MOS, przy czym graniczna wartość czułości odbiornika przy której jest jeszcze możliwa realizacja usługi jest znacznie większa w przypadku radiostacji R-450. Uzyskane wartości MOS zawierają się w przedziale 3÷2,5, co w odniesieniu do zastosowań wojskowych jest wartościami zadowalającymi [2].

Warto również zauważyć, że zastosowanie kodeka MELP pozwala na uzyskanie zadowalającej jakości połączenia głosowego przy niskim poziomie mocy sygnału na wejściu odbiornika. Najlepsze wyniki uzyskano dla kodeka G.729 D, który w tych warunkach zapewnia najwyższą jakość komunikacji głosowej. Należy jednak pamiętać, że pomiary przeprowadzono w odniesieniu do radia szerokopasmowego, które umożliwia stosowanie tego typu kodeków. W przypadku łączności z wykorzystaniem kanałów wąskopasmowych mogą wystąpić problemy z dostępnością zasobów.

LITERATURA

[1] Golan, Edward, Kraśniewski, Adam, Skarżyński, Paweł, Romanik, Janusz. 2011. „Ocena potrzeb i możliwości wykorzystania szerokopasmowych radiostacji osobistych i pokładowych na szczeblu taktycznym", *Journal of KONBiN* 2 (18) : 89–106.

[2] Hersent, O., Petit, J-P., Gurle, D. 2005. "Beyond VoIP Protocols". *Wiley* : 80.

[3] Jabłoński, Wiesław. 2020. „Rewolucja w eterze", *Wojskowy Instytut Wydawniczy*.

[4] Kolahi, S.S., Mudaliar, K., Zhang, C., Gu, Z. 2017. "Impact of IPSec security on VoIP in different environments". *9th International Conference on Ubiquitous and Future Networks ICUFN 2017*, 979–982.

[5] Kumaar, Dinesh R. i in. 2016. "Engineering Search Schema for Encrypted Data in Cloud Networks When QOE Meets QOP". *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*. Volume 23 Issue 1 : 97–103.

[6] Łubkowski, Piotr, Polak, Rafał, Sierzputowski, Robert. 2020. „The measurements of the secured voice communication quality in a broadband radio channel". *Radioelectronic Systems Conference 2019, Proceedings of SPIE: The International Society for Optical Engineering*. vol. 11442 : 1–8.

[7] Materiały dostępne na stronie producenta radiostacji AN/PRC-117G. *https://www.l3harris.com/resources/datasellspec-sheets/anprc-117g-multiband-manpack-datasheet*, dostęp 25.04.2022 r.

[8] Materiały dostępne na stronie producenta radiostacji R-450C. *https://www.transbit.pl/product/r-450c?categorySlug=0,* dostęp 25.04.2022 r.

[9] Polak, Rafał i in. 2019. „Secure voice transmission in the coalitional communication". *XII Conference on Reconnaissance and Electronic Warfare Systems*. Red. Piotr Tomasz Kaniewski. t. 11055 : 1–7.

ANALIZA JAKOŚCI TRANSMISJI TREŚCI AUDIO-WIDEO W SYMULOWANYM ŁĄCZU TELEKOMUNIKACYJNYM Z WYKORZYSTANIEM TECHNIKI OFDM

ANALYSIS OF TRANSMISSION QUALITY OF AUDIO-VIDEO CONTENT IN A SIMULATED TELECOMMUNICATION LINK WITH OFDM TECHNIQUE

Monika Zamłyńska [1]; Przemysław Falkowski-Gilski [2]

[1] Akademia Wojsk Lądowym im. Generała Tadeusza Kościuszki, Wrocław, monika.zamlynska@gmail.com
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, przemyslaw.falkowski@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.26

Streszczenie: **Wdrożenie niezawodnego systemu komunikacji audio-wideo przynosi wiele korzyści. Z uwagi na fakt, że ilość dostępnego pasma stale się kurczy, badacze koncentrują się na nowatorskich metodach transmisji. Obecnie technika OFDM (Orthogonal Frequency Division Multiplexing) jest szeroko stosowana zarówno w mediach przewodowych, jak i bezprzewodowych. W pracy przedstawiono badania jako-ści QoS (Quality of Service) symulowanego łącza transmisji danych, przy zmiennych poziomach SNR (Signal-to-Noise Ratio) oraz BER (Bit Error Rate).**

Abstract: **Application of a reliable audio-video communication system brings many advantages. As the amount of available bandwidth continues to shrink, researchers focus on novel types of transmission. Nowadays, OFDM (Orthogonal Frequency Division Multiplexing) is widely utilized both in wired and wireless media. In this paper we investigate the QoS (Quality of Service) parameters of a simulated data transmission system, including SNR (Signal-to-Noise Ratio) and BER (Bit Error Rate).**

Słowa kluczowe: **badania jakości, OFDM (Orthogonal Frequency Division Multiplexing), transmisja multimediów, QoS (Quality of Service)**

Keywords: **quality evaluation, OFDM (Orthogonal Frequency Division Multiplexing), multimedia transmission, QoS (Quality of Service)**

1. WSTĘP

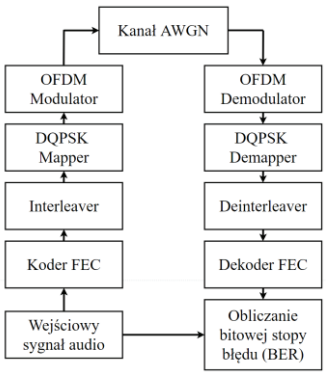
Proces projektowania, wdrożenia i utrzymania niezawodnych usług multimedialnych jest wymagającym zadaniem. Wraz z rozwojem możliwości urządzeń stacjonarnych i przenośnych, rosną oczekiwania użytkowników dotyczące jakości świadczonych usług. Każdego roku zwiększa się pokrycie i przepustowość zarówno sieci przewodowych, jak i bezprzewodowych. Coraz więcej systemów telekomunikacyjnych bazuje na technice OFDM (Orthogonal Frequency Division Multiplexing) [3, 6]. Dlatego tak ważne jest zbadanie technicznych aspektów jakości transmisji w sieciach heterogenicznych, m.in. przy zmiennych warunkach pasma, stosunku sygnał-szum SNR (Signal-to-Noise Ratio), czy też stopie błędów BER (Bit Error Rate). Warto zaznaczyć, że znaczna większość pakietów danych w sieciach obejmuje sygnały multimedialne, w tym materiały audio-wideo [8, 11].

Mając na uwadze wspomniane czynniki i aspekty, postanowiono dokonać analizy jakości parametrów technicznych, tzw. QoS (Quality of Service) [5], dla zasymulowanego łącza telekomunikacyjnego.

2. TRANSMISJA SYGNAŁÓW FONII

2.1. Opis symulacji toru fonii

W celu oceny jakości transmisji dźwięku opracowano model symulacyjny toru telekomunikacyjnego w środowisku Matlab/Simulink. Model ten umożliwiał m.in. wyznaczenie bitowej stopy błędu (BER) dla transmitowanych plików dźwiękowych. Schemat blokowy dla części nadawczej i odbiorczej przedstawiono na rys. 1.



Rys. 1. Schemat blokowy symulowanego toru komunikacyjnego

Najważniejsze parametry zamodelowanego systemu przedstawiono w Tabeli 1.

Tab. 1. Parametry zamodelowanego toru komunikacyjnego dla plików muzycznych

Parametr	*Wartość*
Przepływność bitowa	32 kbit/s
Częstotliwość próbkowania	44,1 kHz
Liczba kanałów	Pojedynczy (mono)
Kodek audio	OGG Vobis
Kodowanie protekcyjne	BCH + kod. splotowe
Modulacja	DQPSK + OFDM
Liczba podnośnych OFDM	1350
Odstęp między podnośnymi	1 kHz
Szerokość kanału	1,44 MHz
Model kanału	AWGN

Pliki dźwiękowe wykorzystane w symulacji były przetworzone z użyciem kodeka Ogg Vobis, w trybie pojedynczego kanału (mono). Z tego względu nie zastosowano multipleksacji. Użyty w modelu system był inspirowany standardem DAB (Digital Audio Broadcasting) [12, 14, 16] z parametrami kodowania zawartymi w Tabeli 2. Wstępnie sygnał kodowany był za pomocą kodu BCH (Bose-Chaudhuri-Hocquenghem) [2].

Tab. 2. Parametry kodowania dla sygnałów fonicznych

Kodowanie	*Parametry*	
BCH	Wielomian:	(511, 439, *t*=8)
Kodowanie splotowe		(133, 171, 145, 133)

Druga faza obejmowała kodowanie splotowe, które zapisuje sekwencję binarnych wektorów wejściowych w celu utworzenia binarnych wektorów wyjściowych [7]. Stosowany jest wielomian ósemkowy 133, 171, 145, 133 z długością wymuszoną 7 bitów [10]. Użycie dwóch kodowań pozwala na transmisje praktycznie wolną od błędów [2]. Następnie dane poddawane były mechanizmowi przeplotu. Zaprojektowany dedykowany blok przekształcał dane według losowej permutacji.

W kolejnym kroku dane poddawane były przekształceniu na symbole z pomocą modulatora w paśmie podstawowym, zgodnie z modulacją DQPSK (Differential Quadrature Phase Shift Keying). W przypadku kanału transmisyjnego zamodelowano wariant z białym szumem AWGN (Additive White Gaussian Noise).

Niech $X(k)$ będzie strumieniem bitów na wyjściu modulatora DQPSK. Sygnał ten był najpierw rozdzielany na N równoległych strumieni. Następnie dla każdego symbolu liczona była odwrotna transformata Fouriera IDFT (Inverse Discrete Fourier Transform):

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j\frac{2\pi}{N}nk} \quad (1)$$

$$\bigwedge_{n\epsilon N} n = 0,1,\dots,N-1 \quad (2)$$

Operację tę wykonywano w modulatorze OFDM (strona nadawcza). W przypadku demodulatora OFDM (strona odbiorcza) wykonywana była operacja odwrotna:

$$X(k) = \sum_{k=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}nk} \quad (3)$$

$$\bigwedge_{k\epsilon N} k = 0,1,\dots,N-1 \quad (4)$$

Następnym krokiem było przygotowanie i przetwarzanie wybranych sygnałów dźwiękowych.

2.1. Przetwarzanie sygnału dźwiękowego

Według zaleceń organizacji 3GPP (3rd Generation Partnership Project) [1], usługi komunikacji głosowej można podzielić na trzy grupy, o zmiennych wymaganiach dotyczących dopuszczalnego opóźnienia oraz stopy błędów, jak pokazano w Tabeli 3.

Tab. 3. Wymagania jakościowe dotyczące transmisji sygnałów fonii

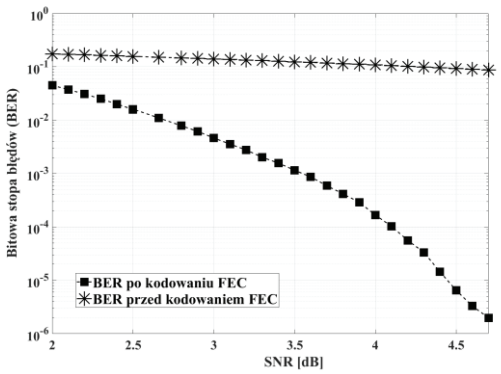
Typ usługi	*Dozwolone opóźnienie [ms]*	*Dozwolona stopa błędów*
Strumieniowanie (w czasie rzeczywistym)	100	10^{-2}
Strumieniowanie (buforowane)	300	10^{-6}
Strumieniowanie (na żywo)	100	10^{-3}

To badanie koncentruje się na komunikacji głosowej. Zgodnie z Tabelą 3, usługi te dopuszczają opóźnienia w zakresie od 100 ms (głos konwersacyjny) do 300 ms (głos niekonwersacyjny). Jeżeli chodzi o poziom błędów, to przyjęty próg waha się od 10^{-6} do 10^{-2}. W zależności od wybranego wariantu, ma to znaczący wpływ na jakość przesyłanych próbek mowy.

Próbki sygnału fonii użyte podczas badania pochodziły z zasobów ITU-T [13]. Składały się z różnych zdań wypowiedzianych przez lektorów męskich i żeńskich w różnych językach. Próbki oryginalne dostępne były w formacie WAV 16-bit PCM. Pliki przetworzone i przesłane przez zamodelowany tor komunikacyjny dostępne były w formacie OGG Vorbis. Przepływność bitową ustawiono na poziomie 32 kbit/s.

2.2. Wyniki

Na rys. 2 przedstawiono wyniki bitowej stopy błędów (BER) w odniesieniu do SNR, w wariancie z zastosowaniem kodowania korekcyjnego FEC (Forward Error Correction) i bez.



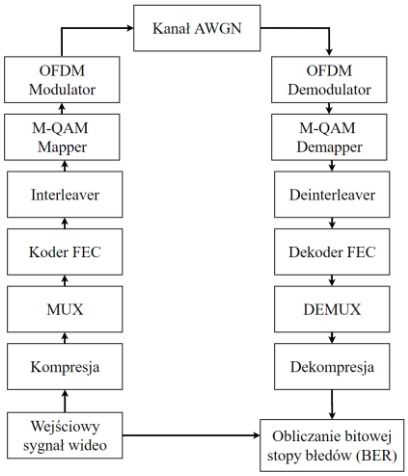
Rys. 2. Wyniki zależności bitowej stopy błędów BER do składowej SNR przed i po zastosowaniu kodowania FEC

W sumie test obejmował 24 pliki audio o zmiennym SNR z zakresu od 2 do 4,7 dB. Poziom BER był monitorowany zarówno przed kodowaniem korekcyjnym FEC, jak i po kodowaniu FEC. Wyraźnie widać, że zastosowanie kodowania FEC znacznie poprawiło jakość przesyłanych sygnałów. W przypadku BER na poziomie 10^{-2} pojawiło się wiele zniekształceń i dźwięk był mocno zakłócony. Dobry balans osiągnięto na poziomie BER wynoszącym 10^{-5}. W tym systemie odpowiadało to współczynnikowi SNR równemu 4,5 dB. Różnica w jakości między BER na poziomie 10^{-5} i 10^{-6} okazała się nieznaczna. Wynosiła ona ok. 0,3 dB. Z kolei bitowa stopa błędów na poziomie 10^{-6} gwarantowała doskonałą jakość audio.

3. TRANSMISJA SYGNAŁÓW WIZJI

3.1. Opis symulacji toru wizji

Analogicznie jak w przypadku sygnałów fonii, opracowano model symulacyjny toru telekomunikacyjnego w środowisku Matlab/Simulink, przeznaczony do transmisji sygnałów wizji. Model ten umożliwiał również m.in. wyznaczenie bitowej stopy błędu (BER) dla transmitowanych plików wideo. Schemat blokowy dla części nadawczej i odbiorczej przedstawiono na rys. 3.



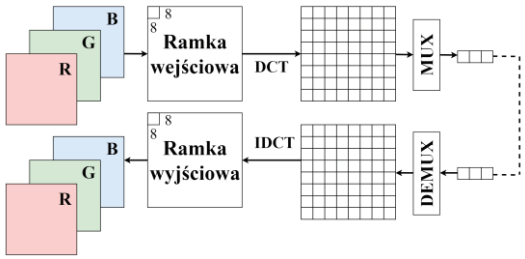
Rys. 3. Schemat blokowy symulowanego toru komunikacyjnego

Najważniejsze parametry zamodelowanego systemu przedstawiono w Tabeli 4.

Tab. 4. Parametry zamodelowanego toru komunikacyjnego dla plików wizyjnych

Parametr	*Wartość*
Rozdzielczość	480x270
Liczba klatek na sekundę	24
Model przestrzeni barw	RGB
Algorytm kompresji	[8x8] + DCT + kwantyzacja
Kodowanie protekcyjne	FEC 3/4 (BCH + LDPC 3/4)
Modulacja	256 QAM + OFDM
Liczba podnośnych OFDM	81
Odstęp pomiędzy podnośnymi	42 kHz
Szerokość kanału	3,4 MHz
Model kanału	AWGN

Materiał wideo wykorzystany w symulacji dostępny był w kolorze, w modelu przestrzeni barw RGB (Red, Green, Blue). Wariant ten najbardziej odpowiada właściwościom odbiorczym ludzkiego oka. Sygnał rozdzielny jest bowiem na trzy oddzielne kanały. Z tego względu zastosowano proces multipleksacji jak przedstawiono to na rys. 4.



Rys. 4. Proces multipleksacji dla transmisji plików wizji

Obraz na początku toru poddany był kompresji. Stosowany algorytm kompresji polegał na podziale obrazu wejściowego na mniejsze bloki o stałych rozmiarach (w przypadku zastosowanego algorytmu były to bloki 8x8).

Kolejnym krokiem była transformacja tych bloków. Wzięto pod uwagę dany w symulacji obraz *f* rozmiaru [MxN], który stanowił pojedynczą klatką obrazu wejściowego. W przestrzeni dwuwymiarowej pikseli o współrzędnych (x,y) oznaczono go jako $f(x,y)$ i aby przekształcić klatkę do obrazu w dziedzinę częstotliwości zastosowano algorytm DCT (Discrete Cosine Transform), który oblicza dwuwymiarową dyskretną transformatę obrazu. [4].

Taka operacja wykonywana była w części nadawczej. W przypadku części odbiorczej wykonywano operacją odwrotną, czyli IDCT (Inverse Discrete Cosine Transform) [4].

Kanały, które zostały poddane kompresji, były multipleksowane zgodnie ze standardem DVB-T2 (Digital Video Broadcasting – Terrestrial 2) [9], gdzie zastosowano kodowanie FEC o sprawności 3/4, jak przedstawiono to w Tabeli 5.

Tab. 5. Parametry kodowania blokowego dla sygnałów wizji

Sprawność kodowania	$n_{ldpc} = 64800(normalna\ ramka)$	
	k_{bch}	k_{ldpc}
3/4	48408	48600

Dane poddawano mechanizmowi przeplotu po stronie nadawczej, a następnie inwersji po stronie odbiorczej. W kolejnym kroku dane poddawane były przekształceniu na symbole za pomocą modulatora w paśmie podstawowym, zgodnie z modulacją 256-QAM (Quadrature Amplitude Modulation). W przypadku kanału transmisyjnego również zamodelowano wariant z białym szumem AWGN (Additive White Gaussian Noise).

3.2. Przetwarzanie sygnału wizji

Zgodnie z dokumentacją 3GPP [1], usługi komunikacji wizyjnej można podzielić na dwie grupy, o zmiennych wymaganiach dotyczących dopuszczalnego opóźnienia oraz stopy błędów, jak pokazano to w Tabeli 6.

Tab. 6. Wymagania jakościowe dotyczące transmisji sygnałów wizji

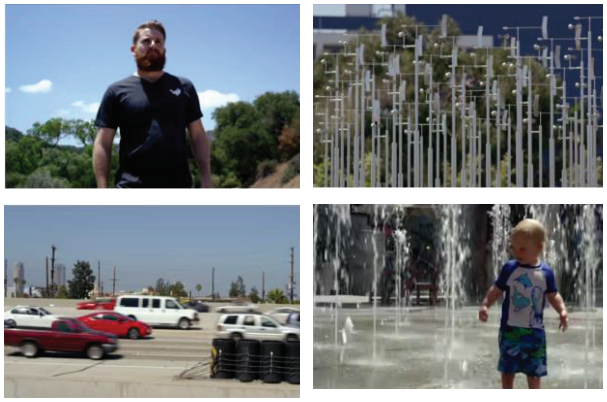
Typ usługi	*Dozwolone opóźnienie [ms]*	*Dozwolona stopa błędów*
Strumieniowanie (z buforowaniem)	300	10^{-6}
Strumieniowanie (na żywo)	150	10^{-3}

W tym konkretnym badaniu skupiamy się na BER w zakresie od 10^{-6} do 10^{-3}. Opóźnienia mogą zostać pominięte.

Pliki wideo pobrane z bazy [15] dostępne były w formacie MP4, kodowane w rozdzielczości 8-bitowej. W tym repozytorium dostępnych było szereg plików w różnej rozdzielczości obrazu, przy czym każdy plik składał się z pięciu różnych sekwencji o odmiennej dynamice i charakterze, w tym:

1) spacerujący mężczyzna;
2) obracające się wiatraki;
3) ruch uliczny (statyczne ujęcia z poruszającymi się samochodami na wielu pasach);
4) bawiące się dziecko przy fontannach;
5) podobne ujęcie jw. z tym, że nagrane w zwolnionym tempie (większa liczba klatek na sekundę);

jak pokazano to na rys. 5.

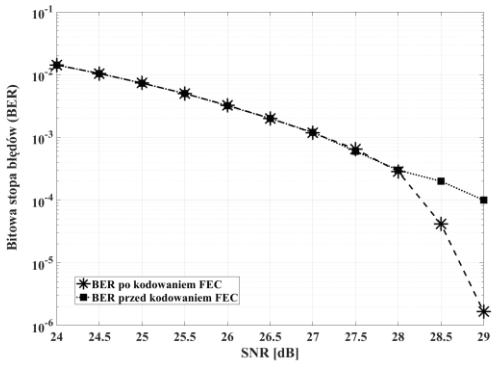


Rys. 5. Sekwencje w analizowanym pliku wideo

Na potrzeby badania wybrano plik o rozdzielczości 480x270 pikseli. Format ten jest popularny w przypadku wielu przenośnych odtwarzaczy multimedialnych, szczególnie małych urządzeń zaliczanych do grupy Internetu rzeczy IoT (Internet of Things).

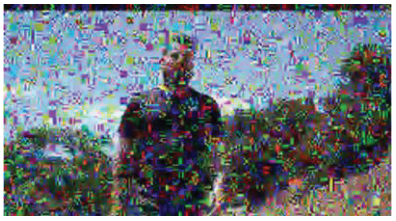
3.3. Wyniki

Na rys. 6 przedstawiono wyniki obliczeń bitowej stopy błędów (BER) w odniesieniu do SNR, w wariancie przed i po kodowaniu korekcyjnym.



Rys. 6. Wyniki zależności bitowej stopy błędów BER do składowej SNR przed i po zastosowaniu kodowania FEC

Zgodnie, z uzyskanymi wynikami kodowanie korekcyjne FEC ma zauważalny wpływ na końcowy wynik stopy błędów dla poziomu SNR powyżej 28 dB. Poziom BER zmienia się w tym przedziale. Dla SNR wynoszącego 29 dB poziom BER, który przed kodowaniem wynosił 10^{-4}, po kodowaniu FEC spadł do poziomu 10^{-6}.



a)

b)

c)

d)

Rys. 7. Przykładowa sekwencja wideo bez i z artefaktami i zniekształcenia w wyniku przejścia przez kanał z zakłóceniem dla: a) SNR=24 dB, b) SNR=27 dB, c) SNR=28,7 dB, d) SNR=30 dB.

Jak przedstawiono na rys. 7, w przypadku BER w przedziale od 10^{-3} do 10^{-2}, jakość filmów jest nie do zaakceptowania. Można zauważyć dużą liczbę artefaktów, które widoczne są w każdej ramce obrazu. W przypadku BER oscylującego w okolicy 10^{-5} widoczne są tylko pojedyncze artefakty i nieznaczne zniekształcenia kolorów.

4. WNIOSKI

Na podstawie przeprowadzonych badań można określić wpływ techniki OFDM i kodowania korekcyjnego dla treści audio i wideo. W przeciwieństwie do sygnałów wizji, sygnały fonii wymagają dużo niższej przepływności bitowej, dlatego można je przesyłać kanałem komunikacyjnym przy użyciu modulacji o niższej wartości. Jak pokazano, kodowanie korekcyjne FEC może znacząco podnieść jakość przesyłanych treści. Należy podkreślić, że każdy system i dedykowana usługa projektowane są z myślą o konkretnym zastosowaniu i docelowym użytkowniku końcowym. Wiąże się to z wieloma aspektami technicznymi, a przede wszystkim z niezawodnością. Z tego powodu bardzo interesujące byłoby zbadanie wpływu SNR i BER, związanych z obiektywną oceną jakości QoS, na subiektywne oceny użytkownika końcowego QoE (Quality of Experience). Taki eksperyment z pewnością byłby interesujący zarówno dla naukowców, jak i praktyków zajmujących się procesem tworzenia, obróbki i dostarczania treści multimedialnych.

LITERATURA

[1] 3GPP Technical Specification 123 203. 2013. *Technical specification group services and system aspects; Policy and charging control architecture*.

[2] Agarwal, Arun, S. Patra. 2012 „Performance analysis of OFDM based DAB systems using concatenated coding technique". *Advanced Materials Research*, 403 : 105–113.

[3] Aragon-Zavala, Alejandro, Pablo Angueira, Jon Montalban, Cesar Vargas-Rosales. 2021. „Radio propagation in terrestrial broadcasting television systems: a comprehensive survey". *IEEE Access* 9 : 34789–34817.

[4] Chen, Wen-Hsiung, C. H. Smith, S. Fralick. 1977. „A fast computational algorithm for the discrete cosine transform". *IEEE Transactions on Communications* 25 (9) : 1004–1009.

[5] Chen, Yanjiao, Kaishu Wu, Qian Zhang. 2014. „From QoS to QoE: A tutorial on video quality assessment". *IEEE Communications Surveys & Tutorials* 17 (2) : 1126–1165.

[6] Cioni, Stefano, Giovanni Emanuele Corazza, Massimo Neri, Alessandro Vanelli-Coralli. 2006. „On the use of OFDM radio interface for satellite digital multimedia broadcasting systems". *International Journal of Satellite Communications and Networking* 24 (2) : 153–167.

[7] Clark Jr., George C., J. Bibb Cain. 1981. *Error-correction coding for digital communications*. Springer.

[8] Debita, Grzegorz, Przemysław Falkowski-Gilski., Marcin Habrych, Grzegorz Wiśniewski, Bogdan Miedziński, Przemysław Jedlikowski, Agnieszka Waniewska, Jan Wandzio, Bartosz Polnik. 2020. „BPL-PLC voice communication system for the oil and mining industry". *Energies* 13 (18) : 4763.

[9] Eizmendi, Inaki, Manuel Velez, David Gómez-Barquero, Javier Morgade, Vincente Baena-Lecuyer, Mariem Slimani, Jan Zoellner. 2014. „DVB-T2: The second generation of terrestrial digital video broadcasting system". *IEEE Transactions on Broadcasting*". 60 (2) : 258–271.

[10] ETSI Standard EN 300 401. 2001. *Radio broadcasting systems; Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers*.

[11] Falkowski-Gilski, Przemysław. 2020. „On the consumption of multimedia content using mobile devices: A year to year user case study". *Archives of Acoustics* 45 (2) : 321–328.

[12] Harada, Hiroshi, Ramjee Prasad. 2002. *Simulation and software radio for mobile communications*. Artech House.

[13] ITU Recommendation P.501. 2017. *Test signals for telecommunication systems*.

[14] Kim, Geon, Yong-tae Lee, So-ra Park, Yong-hoon Lee, Kyungseok Kim. 2014. „Design and Implementation of a novel emergency wake-up alert system within a conventional T-DMB service network". *IEEE Transactions on Consumer Electronics* 60 (4) 574–579.

[15] Netflix Open Content. http://download.opencontent.netflix.com/?prefix=AV1/Chimera/Old/ (dostęp: 13.04.2022).

[16] Zhang, Hongsheng, Hongyun Wang, Guoyu Wang, Mingying Lu. 2017. „Design and implementation of the DAB/DMB transmitter identification information decoder". *International Journal of Circuits, Systems and Signal Processing* 11 : 59–64.

ANALIZA NIEZAWODNOŚCIOWO-EKSPLOATACYJNA INTELIGENTNYCH SYSTEMÓW TRANSPORTOWYCH Z UWZGLĘDNIENIEM ZAKŁÓCEŃ ELEKTROMAGNETYCZNYCH

RELIABILITY-EXPLOITATION ANALYSIS IN INTELLIGENT TRANSPORTATION SYSTEMS TAKING INTO ACCOUNT ELECTROMAGNETIC INTERFERENCE

Mirosław Siergiejczyk[1]; Jacek Paś[2]; Adam Rosiński[3]

[1] Politechnika Warszawska Wydział Transportu, Warszawa, miroslaw.siergiejczyk@pw.edu.pl
[2] Wojskowa Akademia Techniczna Wydział Elektroniki, Warszawa, jacek.pas@wat.edu.pl
[3] Politechnika Warszawska Wydział Transportu, Warszawa, adam.rosinski@pw.edu.pl

DOI: 10.15199/59.2022.4.27

Streszczenie**: W artykule przedstawiono zagadnienia związane z analizą niezawodnościowo-eksploatacyjną inteligentnych systemów transportowych. Są one eksploatowane w określonym środowisku transportowym. W artykule przedstawiono wyniki pomiarów pola magnetycznego emitowanego przez kable energetyczne dostarczające energię elektryczną na dworcu kolejowym do inteligentnych systemów transportowych. Zaprezentowana w artykule metodyka analizy niezawodnościowo-eksploatacyjnej inteligentnych systemów transportowych z uwzględnieniem zakłóceń elektromagnetycznych może być zastosowana podczas projektowania systemów elektronicznych użytkowanych w transporcie.**

Abstract**: The paper presents issues associated with the reliability-exploitation analysis of intelligent transportation systems (ITS) in rail transport. They operate in a defined transportation environment. The article presents the measurement results concerning a magnetic field emitted by electrical cables, which supply electricity to ITS stations at a train station. The reliability-operational analysis method which was shown in the paper, may be utilized when engineering electronic systems used in transport.**

Słowa kluczowe**: eksploatacja, inteligentne systemy transportowe, niezawodność, zakłócenia elektromagnetyczne.**

Keywords**: exploitation, intelligent transport systems, reliability, electromagnetic interference.**

1. WSTĘP

Inteligentne systemy transportowe ITS (ang. Intelligent Transportation Systems) to połączenie technologii telekomunikacyjnych i informacyjnych z infrastrukturą transportową i pojazdami w celu osiągnięcia korzyści, takich jak zwiększenia poziomu bezpieczeństwa, efektywności procesów transportowych oraz ochrony środowiska naturalnego. Zastosowanie ITS wpływa korzystnie na warunki podróżowania w zakresie multimodalnym – zarówno z zakresie prywatnych, jak i publicznych środków transportu drogowego, morskiego i lotniczego. Dzięki zastosowaniu ITS uzyskuje się bardziej ekonomiczną poprawę warunków komunikacyjnych niż gdyby rozbudowywać infrastrukturę komunikacyjną w dotychczasowej formie.

Inteligentne systemy transportowe funkcjonują najczęściej w zróżnicowanych warunkach eksploatacyjnych [6, 7]. Dzięki obserwacjom procesu ich użytkowania w środowisku transportowym (w szczególności kolejowym) potwierdzają zależność poprawnego działania od niezawodności części składowych oraz efektywnego zarządzania procesem ich eksploatacji. Zatem szczegółowa analiza zjawisk niezawodnościowo-eksploatacyjnych powinna uwzględniać nie tylko niezawodność ITS, ale także efektywność zarządzania eksploatacją. W tym celu autorzy artykułu dokonali odwzorowania zjawisk zachodzących w rzeczywistości (z uwzględnieniem zakłóceń elektromagnetycznych [2, 4, 5]) w modelu badawczym inteligentnego systemu transportowego.

Inteligentne systemy transportowe są obecnie elementem systemów transportowych. Ich prawidłowe funkcjonowanie ma istotny wpływ na bezpieczeństwo i sprawność procesu transportowego. Stan niezdatności urządzeń ITS oraz błędy w działaniu operatorów mogą prowadzić do stanów zagrożenia bezpieczeństwa. Teoria niezawodności zajmuje się m.in. analizą wpływu uszkodzeń na określone wskaźniki niezawodnościowe. Obszarem zainteresowań teorii bezpieczeństwa są skutki uszkodzeń i błędów, które prowadzą do zagrożenia bezpieczeństwa (szczególnie istotnego z punktu widzenia procesu przemieszczania osób i ładunków). Bardzo istotna jest zatem kwestia prawidłowego określenia, który ze stanów systemu można uznać za dopuszczalny lub niedopuszczalny z punktu widzenia bezpieczeństwa (stany częściowej zdatności).

W artykule zaprezentowano pomiary pola magnetycznego emitowanego przez kable energetyczne dostarczające energię elektryczną na dworcu kolejowym do inteligentnych systemów transportowych. Analiza otrzymanych wyników umożliwiła opracowanie modelu badawczego inteligentnego systemu transportowego, a następnie przeprowadzenie analizy niezawodnościowo-eksploatacyjnej z uwzględnieniem zakłóceń elektromagnetycznych.

2. ANALIZA PROCESU EKSPLOATACJI ITS Z UWZGLĘDNIENIEM ZAKŁÓCEŃ ELEKTROMAGNETYCZNYCH

Kluczowym zadaniem jaki spoczywa na systemie ITS jest przesłanianie informacji za pośrednictwem różnego typu środków łączności. Zastosowanie urządzeń telekomunikacyjnych i informatycznych powoduje, że systemy ITS są w rzeczywistości systemami teleinformatycznymi podatnymi na różnego rodzaju zagrożenia, w tym także te związane z zakłóceniami elektromagnetycznymi.

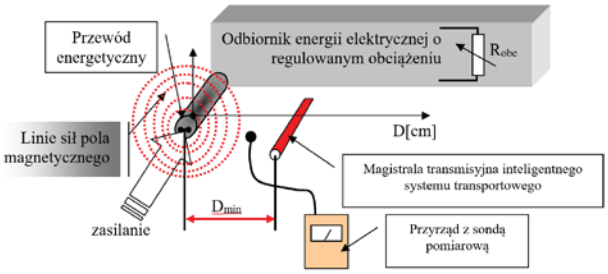
Wdrażanie systemów ITS wpływa korzystnie na:

- zwiększenie przepustowości sieci transportowej,
- zmniejszenie czas podróży i zużycie energii,
- poprawienie jakości środowiska naturalnego oraz poprawienie komfortu podróżowania i warunków ruchu maszynistów,
- redukowanie kosztów zarządzania taborem oraz związanych z utrzymaniem i renowacją nawierzchni.

Aby skutecznie realizować wymienione cele zarządzający systemem transportowym powinien mieć dostęp do aktualnych danych napływających automatycznie w odpowiednim przedziale czasu. Niektóre z danych muszą być dostępne w czasie rzeczywistym (np. alarmy, połączenia ratunkowe, obrazy z kamer). Inne dane mogą być przesyłane z niewielkimi opóźnieniami (zależnymi od priorytetu informacji). Część danych jest zazwyczaj gromadzona w urządzeniach i przekazywana dalej w określonych przedziałach czasu w postaci zagregowanej lub wstępnie przetworzonej. Dane pobierane z klasycznych czujników zawierają w sobie ograniczone zasoby informacyjne. W celu podejmowania racjonalnych decyzji zarządzający powinien mieć możliwości weryfikacji napływających danych i podglądu aktualnej sytuacji. Dlatego coraz częściej wprowadza się cyfrowe systemy monitoringu wizyjnego. Pozwalają one monitorować newralgiczne miejsca w czasie rzeczywistym oraz dokumentować zdarzenia i podjęte działania.

W inteligentnych systemach transportowych użytkowanych na rozległym obszarze kolejowym istotnym problemem jest okablowanie. Połączenia kablowe łączą czujniki, elementy wykonawcze systemu rozmieszczone na terenie dworca z centralą zarządzającą. W systemach ITS zainstalowanych na obszarze kolejowym kable zasilające, sygnałowe i magistrale transmisyjne przebiegają przez obszary na których wstępują stacjonarne źródła zakłóceń elektromagnetycznych [3, 9, 10]. Kable mogą krzyżować się lub przebiegać w pobliżu linii energetycznych zasilających systemy sterowania ruchem kolejowym, trakcję kolejową, systemy oświetlenia, itd. [1]. Przewody energetyczne, przez które płynie prąd zasilający wytwarzają wokół siebie niezamierzone pole elektryczne i magnetyczne proporcjonalne do napięcia oraz płynącego prądu. Pole to stanowi bardzo istotne źródło zakłóceń na dworcu kolejowym i w ogólnym przypadku może prowadzić do zakłócenia pracy inteligentnych systemach transportowych. W przypadku kabli energetycznych zasilających elektroniczne systemy bezpieczeństwa (ESB), systemy sterowania ruchem kolejowym, i inne urządzenia elektryczne użytkowane na rozległym obszarze kolejowym rozpatrujemy zakłócenia z zakresu małych częstotliwości – do 100 kHz. Ten zakres częstotliwości dodatkowo ze względów pomiarowych dzielony jest na dwa podzakresy: ELF (Extremely Low Frequencies) - od 5 Hz do 2 kHz oraz od 2 kHz do 100 kHz – VLF (Very Low Frequencies). W takich zakresach częstotliwości Autorzy artykułu wykonali pomiary zaburzeń elektromagnetycznych wytwarzanych w sposób niezamierzony na rozległym terenie kolejowym. Wykonano osobno pomiary natężenia **E** pola elektrycznego oraz indukcji **B** pola magnetycznego dla tych dwóch zakresów częstotliwości. Badając rozkłady **E** i **B** przyjęto charakterystyczne odległości od urządzeń elektronicznych systemów bezpieczeństwa które są instalowane w odległości co najmniej 30 cm od kabli zasilających energetycznych ze względu na istniejący poziom zakłóceń elektromagnetycznych. W wyniku przeprowadzonych pomiarów rozkładów E, B stwierdzono, że mogą występować pola elektryczne i magnetyczne wpływające niekorzystnie na ludzi oraz pracę urządzeń elektronicznych, wrażliwych na ich działanie (czujki, kamery, moduły sterujące ESB). Przykładowo, z danych literaturowych [2,11] i badań własnych wynika, że przy zewnętrznym polu magnetycznym o częstotliwości 50 Hz, którego wartość indukcji B jest większa niż 0,5 μT występują drgania obrazu na ekranie monitora komputerowego [12]. W kraju można się spotkać z zaleceniami dotyczącymi ochrony przed polami elektromagnetycznymi z zakresu małych częstotliwości, przyjmującymi jako korzystne następujące poziomy pola elektromagnetycznego - natężenia E pola elektrycznego na poziomie do 100 V/m oraz indukcji B nie większej niż 0,5 μT [12,13,14].

W celu określenia wskaźników niezawodnościowo-eksploatacyjnych inteligentnego systemu transportowego wykonano pomiary pola magnetycznego emitowanego przez kable energetyczne dostarczające energię elektryczną na dworcu kolejowym. W zależności od typu kabla oraz mocy pobieranej przez urządzenia zasilające powinny być spełnione warunki określające minimalną odległość dla niezakłóconej pracy systemu. Warunki i sposób wykonania pomiarów zostały przedstawione na rys. 1. Uzyskane wartości indukcji B pola magnetycznego dla zakresów częstotliwości ELF oraz VLF zostały przedstawione w tab. 1. Pomiary zostały wykonane przy zawartości współczynnika harmonicznych h sieci napięcia przemiennego 230 [V] 50 [Hz] równym h = 2,5 ÷ 3 [%].



Rys. 1. Schemat stanowiska pomiarowego

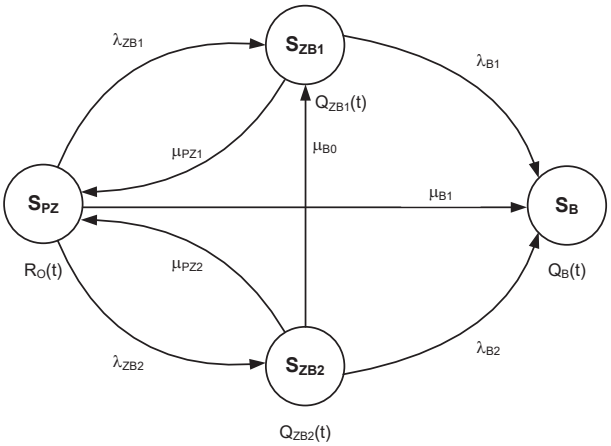
Analizując otrzymane wyniki pomiarów można stwierdzić, iż eksploatowany w środowisku transportowym (w którym występują zakłócenia elektromagne-

tyczne) inteligentny system transportowym może znajdować się w określonych stanach [8]. Zatem słuszna wydaje się idea, by dokonać analizy niezawodnościowo-eksploatacyjnej. Dzięki temu uzyskamy zależności pozwalające wyznaczyć wartości prawdopodobieństw przebywania systemu w wyróżnionych stanach eksploatacyjnych.

Przeprowadzając analizę funkcjonowania inteligentnego systemu transportowego, można zilustrować relacje zachodzące w tej strukturze, w aspekcie niezawodnościowo-eksploatacyjnym, tak jak przedstawia to rys. 2.

Tab. 1. Parametry indukcji B pola magnetycznego dla zakresu częstotliwości ELF i VLF w funkcji mocy pobieranej przez odbiornik

Parametr pola elektromagnetycznego określany podczas pomiarów				**Moc pobierana przez odbiornik**		
				P < 2kVA	5kVA > P > 2kVA	P > 5 kVA
Zakres częstotliwości	**ELF**	Indukacja B [μT] pola magnetycznego	*	0,23	0,24	0,29
			**	0,14	0,16	0,21
	VLF	Indukacja B [nT] pola magnetycznego	*	15,6	23,1	24,5
			**	14,1	21,4	19,4



Rys. 2. Relacje w inteligentnym systemie transportowym

Oznaczenia na rys. 2:

$R_O(t)$ – funkcja prawdopodobieństwa przebywania systemu w stanie pełnej zdatności S_{PZ},

$Q_{ZB1}(t)$ – funkcja prawdopodobieństwa przebywania systemu w stanie częściowej zdatności I S_{ZB1},

$Q_{ZB2}(t)$ – funkcja prawdopodobieństwa przebywania systemu w stanie częściowej zdatności II S_{ZB2},

$Q_B(t)$ – funkcja prawdopodobieństwa przebywania systemu w stanie niezdatności S_B,

λ_{ZB1} – intensywności przejścia ze stanu pełnej zdatności S_{PZ} do stanu częściowej zdatności I S_{ZB1},

λ_{ZB2} – intensywności przejścia ze stanu pełnej zdatności S_{PZ} do stanu częściowej zdatności II S_{ZB2},

μ_{PZ1} – intensywności przejścia ze stanu częściowej zdatności I S_{ZB1} do stanu pełnej zdatności S_{PZ},

μ_{PZ2} – intensywności przejścia ze stanu częściowej zdatności II S_{ZB2} do stanu pełnej zdatności S_{PZ},

μ_{B0} – intensywności przejścia ze stanu częściowej zdatności II S_{ZB2} do stanu częściowej zdatności I S_{ZB1},

μ_{B1} – intensywności przejścia ze stanu pełnej zdatności S_{PZ} do stanu niezdatności S_B,

λ_{B1} – intensywności przejścia ze stanu częściowej zdatności I S_{ZB1} do stanu niezdatności S_B,

λ_{B2} – intensywności przejścia ze stanu częściowej zdatności II S_{ZB2} do stanu niezdatności S_B.

Stan pełnej zdatności S_{PZ} jest stanem, w którym prawidłowo funkcjonuje inteligentny system transportowy. Stan częściowej zdatności S_{ZB1} jest stanem, w którym inteligentny system transportowy jest częściowo zdatny (zakłócenia pola magnetycznego przekraczają wartości tolerowane). Stan częściowej zdatności S_{ZB2} jest stanem, w którym inteligentny system transportowy jest częściowo zdatny (zakłócenia pola magnetycznego przekraczają wartości tolerowane, ale nie przekraczają wartości dopuszczalnych). Stan niezdatności bezpieczeństwa S_B jest stanem, w którym inteligentny system transportowy jest niezdatny (zakłócenia pola magnetycznego przekraczają wartości dopuszczalne).

Jeśli inteligentny system transportowy jest w stanie pełnej zdatności S_{PZ} i pojawia się zakłócenie pola magnetycznego (w zakresie tolerowanym przez system) to następuje przejście systemu do stanu częściowej zdatności S_{ZB1} z intensywnością λ_{ZB1}. Jeśli system jest w stanie częściowej zdatności S_{ZB1} to możliwe jest przejście do stanu pełnej zdatności S_{PZ} pod warunkiem podjęcia działań polegających na przywróceniu stanu zdatności.

Gdy istnieje stan częściowej zdatności S_{ZB1} i zakłócenia pola magnetycznego przekraczają wartości dopuszczalne, to następuje przejście do stanu niezdatności S_B z intensywnością λ_{B1}.

Jeśli inteligentny system transportowy jest w stanie pełnej zdatności S_{PZ} i pojawia się zakłócenie pola magnetycznego (w zakresie dopuszczalnym przez system) to następuje przejście systemu do stanu częściowej zdatności S_{ZB2} z intensywnością λ_{ZB2}. Jeśli system jest w stanie częściowej zdatności S_{ZB2} to możliwe jest przejście do stanu pełnej zdatności S_{PZ} pod warunkiem podjęcia działań polegających na przywróceniu stanu zdatności.

Gdy istnieje stan częściowej zdatności S_{ZB2} i zakłócenia pola magnetycznego przekraczają wartości dopuszczalne, to następuje przejście do stanu niezdatności S_B z intensywnością λ_{B2}.

Jeśli inteligentny system transportowy jest w stanie częściowej zdatności S_{ZB2} i nastąpi ograniczenie zakłóceń

pola magnetycznego (do poziomu tolerowanego), to następuje przejście systemu do stanu częściowej zdatności S_{ZB1} z intensywnością μ_{B0}.

Jeśli inteligentny system transportowy jest w stanie pełnej zdatności S_{PZ} i pojawia się zakłócenie pola magnetycznego (w zakresie przekraczającym wartości dopuszczalne przez system), to następuje przejście do stanu niezdatności S_B z intensywnością μ_{B1}.

System przedstawiony na rys. 2 może być opisany następującymi równaniami Kołmogorowa-Chapmana:

$$\begin{aligned}
&R_0'(t) = -\lambda_{ZB1} \cdot R_0(t) + \mu_{PZ1} \cdot Q_{ZB1}(t) - \lambda_{ZB2} \cdot R_0(t) + \\
&+\mu_{PZ2} \cdot Q_{ZB2}(t) - \mu_{B1} \cdot R_0(t) \\
&Q_{ZB1}'(t) = \lambda_{ZB1} \cdot R_0(t) - \mu_{PZ1} \cdot Q_{ZB1}(t) - \lambda_{B1} \cdot Q_{ZB1}(t) + \\
&+\mu_{B0} \cdot Q_{ZB2}(t) \\
&Q_{ZB2}'(t) = \lambda_{ZB2} \cdot R_0(t) - \mu_{PZ2} \cdot Q_{ZB2}(t) - \lambda_{B2} \cdot Q_{ZB2}(t) - \\
&-\mu_{B0} \cdot Q_{ZB2}(t) \\
&Q_B'(t) = \lambda_{B1} \cdot Q_{ZB1}(t) + \lambda_{B2} \cdot Q_{ZB2}(t) + \mu_{B1} \cdot R_0(t)
\end{aligned} \tag{1}$$

Przyjmując warunki początkowe:

$$\begin{aligned}
&R_0(0) = 1 \\
&Q_{ZB1}(0) = Q_{ZB2}(0) = Q_B(0) = 0
\end{aligned} \tag{2}$$

oraz stosując przekształcenie Laplace'a otrzymujemy:

$$R_0^*(s) = -\frac{b_1 \cdot b_2}{b_2 \cdot \lambda_{ZB1} \cdot \mu_{PZ1} - a \cdot b_1 \cdot b_2 + b_1 \cdot \lambda_{ZB2} \cdot \mu_{PZ2} + \mu_{B0} \cdot \mu_{PZ1} \cdot \lambda_{ZB2}}$$

$$Q_{ZB1}^*(s) = -\frac{b_2 \cdot \lambda_{ZB1} + \mu_{B0} \cdot \lambda_{ZB2}}{b_2 \cdot \lambda_{ZB1} \cdot \mu_{PZ1} - a \cdot b_1 \cdot b_2 + b_1 \cdot \lambda_{ZB2} \cdot \mu_{PZ2} + \mu_{B0} \cdot \mu_{PZ1} \cdot \lambda_{ZB2}}$$

$$Q_{ZB2}^*(s) = -\frac{b_1 \cdot \lambda_{ZB2}}{b_2 \cdot \lambda_{ZB1} \cdot \mu_{PZ1} - a \cdot b_1 \cdot b_2 + b_1 \cdot \lambda_{ZB2} \cdot \mu_{PZ2} + \mu_{B0} \cdot \mu_{PZ1} \cdot \lambda_{ZB2}}$$

$$Q_B^*(s) = -\frac{b_1 \cdot b_2 \cdot \mu_{B1} + b_2 \cdot \lambda_{B1} \cdot \lambda_{ZB1} + b_1 \cdot \lambda_{B2} \cdot \lambda_{ZB2} + \lambda_{B1} \cdot \mu_{B0} \cdot \lambda_{ZB2}}{b_2 \cdot c \cdot \lambda_{ZB1} \cdot \mu_{PZ1} - a \cdot b_1 \cdot b_2 \cdot c + b_1 \cdot c \cdot \lambda_{ZB2} \cdot \mu_{PZ2} + c \cdot \mu_{B0} \cdot \mu_{PZ1} \cdot \lambda_{ZB2}} \tag{3}$$

gdzie:

$$\begin{aligned}
&a = s + \lambda_{ZB1} + \lambda_{ZB2} + \mu_{B1} \\
&b_1 = s + \mu_{PZ1} + \lambda_{B1} \\
&b_2 = s + \mu_{PZ2} + \lambda_{B2} + \mu_{B0} \\
&c = s
\end{aligned} \tag{4}$$

Przeprowadzając dalszą analizę matematyczną otrzymuje się zależności pozwalające na wyznaczenie prawdopodobieństw przebywania całego systemu w stanach: pełnej zdatności S_{PZ}, częściowej zdatności S_{ZB1} i S_{ZB2} oraz niezdatności S_B.

3. WNIOSKI

Stosowanie układów elektronicznych w inteligentnych systemach transportowych skutkuje koniecznością funkcjonowania ich w różnorodnych systemach przy usytuowaniu bardzo blisko siebie. Skutkiem tego może być zwiększony poziom zakłóceń elektromagnetycznych, które wpływają na układ. Może to spowodować nieprawidłowe funkcjonowanie systemu ITS, a tym samym znajdowanie się jego w stanach częściowej zdatności lub niezdatności. Dlatego też projektując inteligentne systemy transportowe należy zaplanować ich funkcjonowanie w warunkach rzeczywistych, czyli w otoczeniu innych urządzeń elektronicznych i elektrycznych. Pomocna może w tym być przedstawiona analiza niezawodnościowo-eksploatacyjna inteligentnych systemach transportowych z uwzględnieniem zakłóceń elektromagnetycznych. Pozwala ona liczbowo określić wartości prawdopodobieństw przebywania systemu ITS w wyróżnionych stanach.

W dalszych badaniach tego zagadnienia autorzy planują opracowanie kolejnych modeli niezawodnościowo-eksploatacyjnych, w których to zostaną zróżnicowane stany częściowej zdatności. Umożliwi to dokładniejsze odwzorowanie funkcjonowania ITS użytkowanego w kolejowym środowisku elekromagnetycznym.

LITERATURA

[1] Billinton R., Allan R.N. 1996. „Reliability evaluation of power systems". New York: Plenum Press.

[2] Charoy A. 1999. „Interference in electronic equipment". Warsaw: WNT.

[3] Dabrowski T., Bednarek M., Fokow K., Wisnios M. 2012. „The method of threshold-comparative diagnosing insensitive on disturbances of diagnostic signals". *Przeglad Elektrotechniczny - Electrical Review* 88 (11A) : 93-97.

[4] Lheurette E. 2013. „Metamaterials and Wave Control". ISTE and Wiley.

[5] Ott H.W. 2009. „Electromagnetic compatibility engineering". Wiley.

[6] Paś J. 2015. „Operation of electronic transportation systems". Radom: Publishing House University of Technology and Humanities.

[7] Rosiński A. 2015. „Modelling the maintenance process of transport telematics systems". Warsaw: Publishing House Warsaw University of Technology.

[8] Siergiejczyk M,. Paś J., Rosiński A. 2015. „Issue of reliability–exploitation evaluation of electronic transport systems used in the railway environment with consideration of electromagnetic interference". *IET Intelligent Transport Systems* 10 (9) : 587–593.

[9] Siergiejczyk M., Rosiński A., Paś J. 2016. „Analysis of unintended electromagnetic fields generated by safety system control panels". Diagnostyka 17 (3) : 35-40.

[10] Vašata J., Doleček R. 2016. „Electromagnetic compatibility and lightning current impacts in the railway equipment buildings". 26th Conference Radioelektronika, Košice, Slovak Republic.
[11] Korniewicz H. 1998. „Pola elektromagnetyczne urządzeń biurowych. – Zawodowe i biurowe zagrożenia polami elektromagnetycznymi – Materiał szkoleniowy" – 19 lutego 1998, s 36-42.
[12] Paś J. 2014. „Linie napowietrzne wysokich napięć: środowisko elektromagnetyczne a ograniczenia w użytkowaniu terenów". *Biuletyn Wojskowej Akademii Technicznej*, Tom 63, Wydanie 3, pp. 51-63.
[13] Rozporządzenie Ministra Pracy i Polityki Społecznej z dn.29.11.2002 r. w sprawie dopuszczalnych stężeń i natężeń czynników szkodliwych dla zdrowia w środowisku pracy. Dz. U. Nr 217, poz. 1833.
[14] Paś, J.,Dyduch, J. 2009. „Impact of electromagnetic interference on transport security systems". *Measurements Automatics Robotics* 10.

OCENA BEZPIECZEŃSTWA ALGORYTMÓW ZAPEWNIAJĄCYCH POUFNOŚĆ ORAZ UWIERZYTELNIENIE DANYCH W SIECI 5G

SECURITY ASSESSMENT OF ALGORITHMS ENSURING DATA CONFIDENTIALITY AND AUTHENTICATION IN THE 5G NETWORK

Mariusz Borowski

Wojskowy Instytut Łączności – Państwowy Instytut Badawczy, 05-130 Zegrze Południowe, ul. Warszawska 22A
m.borowski@wil.waw.pl

DOI: 10.15199/59.2022.4.28

Streszczenie: **W artykule dokonano oceny bezpieczeństwa algorytmów zapewniających poufność oraz uwierzytelnienie danych wymienianych w sieci 5G na odcinku między urządzeniem mobilnym a stacją bazową. Algorytmy zapewniające poufność 128-NEA oraz algorytmy umożliwiające uwierzytelnienie danych 128-NIA są wystarczające do zapewnienia ochrony danych nieklasyfikowanych o różnych stopniach wrażliwości, przed adwersarzem klasycznym oraz wykorzystującym komputer kwantowy. Nie jest jednak możliwe ich użycie do ochrony krajowych informacji niejawnych w rozumieniu *Ustawy o ochronie informacji niejawnych*.**

Abstract: **The article assesses the security of algorithms that ensure confidentiality and authentication of data exchanged in the 5G network in the section between the mobile device and the base station. The 128-NEA confidentiality algorithms and 128-NIA data authentication algorithms are sufficient to protect unclassified data with varying degrees of sensitivity against classical and quantum computer adversaries. However, it is not possible to use them to protect national classified information according to *the Protection of Classified Information Act*.**

Słowa kluczowe: **algorytm Grovera, SNOW 3G, złożoność obliczeniowa, ZUC**

Keywords: **computational complexity, Grover algorithm, SNOW 3G, ZUC**

1. WSTĘP

Sieć 5G to najnowszy standard sieci komórkowej, piąta generacja technologii mobilnej, która zastąpi sieć 4G/LTE. Technologia 5G oferuje lepszą wydajność, niezawodność i zdecydowanie większą prędkość pobierania i wysyłania danych. Sieć 5G jest projektowana z myślą o podłączeniu do niej ogromnej liczby urządzeń, tzw. Internetu rzeczy. W związku z tym, że oferuje małe opóźnienia, umożliwi to automatyzację wielu dziedzin, takich jak przemysł, energetyka, transport czy medycyna. Zatem w sieci 5G będzie wymienianych wiele danych o różnym stopniu wrażliwości.

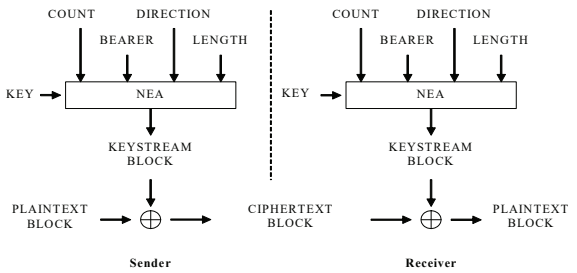
Twórcy standardu poświęcili wiele uwagi zapewnieniu efektywnych usług bezpieczeństwa w sieci 5G [1]. W niniejszym artykule dokonano oceny bezpieczeństwa algorytmów zapewniających poufność oraz uwierzytelnienie danych wymienianych w sieci 5G na odcinku między urządzeniem mobilnym a stacją bazową, na ataki adwersarza klasycznego, a także kryptoanalizę z wykorzystującą komputer kwantowy. Praca została wykonana w projekcie „Zaawansowane metody i techniki wykrywania i przeciwdziałania atakom na infrastrukturę dostępową i aplikacje sieci 5G", realizowanym w ramach programu CyberSecIdent - Cyberbezpieczeństwo i e-Tożsamość, współfinansowanym przez Narodowe Centrum Badań i Rozwoju.

2. ALGORYTMY WYKORZYSTANE DO ZAPEWNIENIA POUFNOŚCI W SIECI 5G

W celu zapewnienia poufności danych wymienianych w sieci 5G na odcinku między urządzeniem mobilnym a stacją bazową zdefiniowano następujące algorytmy kryptograficzne nazwane NEA (ang. New Encryption Algorithm) wykorzystujące klucz kryptograficzny o długości 128 bitów:

- 128-NEA1 [1]. Jako przekształcenie wewnętrzne zastosowano algorytm strumieniowy SNOW 3G [2];
- 128-NEA2 [1]. Jako przekształcenie wewnętrzne zastosowano algorytm AES-128 w trybie licznikowym;
- 128-NEA3 [1]. Jako przekształcenie wewnętrzne zastosowano algorytm strumieniowy ZUC [3].

Danymi wejściowymi do powyższych algorytmów szyfrujących NEA jest 128 bitowy klucz kryptograficzny KEY, 32 bitowy licznik ramek COUNT, 5 bitowy wskaźnik nadawcy BEARER, jednobitowa flaga kierunku transmisji DIRECTION, wymagana bitowa długość gamy pseudolosowej LENGTH. Flaga DIRECTION przyjmuje wartość 0 dla kierunku uplink zaś 1 dla kierunku downlink.



Rys. 1. Szyfrowanie danych z użyciem NEA

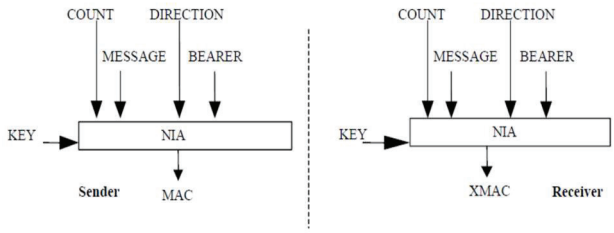
Rys. 1 przedstawia sposób wykorzystania algorytmu zapewniającego poufność NEA do szyfrowania danych przy wykorzystaniu gamy pseudolosowej, sumowanej binarnie z danymi odkrytymi. Tekst odkryty może następnie być odtworzony poprzez wygenerowanie takiej samej gamy pseudolosowej (wykorzystując te same dane wejściowe) oraz jej sumowanie binarne, bit po bicie, z tekstem zaszyfrowanym.

3. ALGORYTMY UŻYTE DO UWIERZYTELNIENIA DANYCH

W celu zapewnienia uwierzytelnienia danych wymienianych w sieci 5G na odcinku między urządzeniem mobilnym a stacją bazową zdefiniowano następujące algorytmy kryptograficzne nazwane NIA (ang. New Integrity Algorithm) wykorzystujące klucz kryptograficzny o długości 128 bitów:

- 128-NIA1 [1]. Jako przekształcenie wewnętrzne zastosowano algorytm strumieniowy SNOW 3G;
- 128-NIA2 [1]. Jako przekształcenie wewnętrzne zastosowano algorytm AES-128 w trybie CMAC;
- 128-NIA3 [1]. Jako przekształcenie wewnętrzne zastosowano algorytm strumieniowy ZUC.

Danymi wejściowymi do powyższych algorytmów zapewniających uwierzytelnienie danych NIA jest 128 bitowy klucz kryptograficzny KEY, 32 bitowy licznik COUNT, 5 bitowy wskaźnik nadawcy BEARER, sama wiadomość MESSAGE, o bitowej długości LENGTH; jednobitowa flaga kierunku transmisji DIRECTION. Rys. 2 przedstawia sposób wykorzystania algorytmu zapewniającego uwierzytelnienie danych NIA do wyznaczenia znacznika uwierzytelniającego MAC. Uwierzytelnienie jest poprawne, gdy odbiorca uzyska zgodność znacznika odebranego (MAC) oraz wyznaczonego (XMAC).



Rys. 2. Wyznaczanie znacznika MAC (XMAC) przez NIA

4. ZAGROŻENIA DLA ALGORYTMÓW ZAPEWNIAJĄCYCH POUFNOŚĆ I UWIERZYTELNIENIE ZE STRONY KRYPTOANALIZY KLASYCZNEJ

W niniejszym rozdziale przedstawiono ocenę bezpieczeństwa algorytmów zapewniających poufność 128-NEA (128-NEA1, 128-NEA2, 128-NEA3) oraz algorytmów umożliwiających uwierzytelnienie 128-NIA (128-NIA1, 128-NIA2, 128-NIA3) w sieci 5G. Pokazano zagrożenia ze strony kryptoanalizy klasycznej dla niniejszych przekształceń wewnętrznych: SNOW 3G [2], AES-128 oraz ZUC [3], gdyż podatności na ataki kryptoanalityczne przekształceń wewnętrznych, przekładają się negatywnie na bezpieczeństwo algorytmów zapewniającego poufność 128-NEA oraz algorytmów zapewniających uwierzytelnienie danych 128-NIA.

4.1. Ocena bezpieczeństwa algorytmów zapewniających poufność 128-NEA

Algorytm zapewniający poufność 128-NEA1 wykorzystuje szyfr SNOW 3G jako wewnętrzne przekształcenie kryptograficzne. Algorytm SNOW 3G jest synchronicznym szyfrem strumieniowym użytkującym 128 bitowy tajny klucz kryptograficzny oraz 128 bitowy wektor inicjalizujący. W Tab. 1. przedstawiono złożoności czasowo – pamięciowe oraz danych (wybranych tekstów jawnych, CP) najefektywniejszych ataków kryptoanalitycznych na szyfr strumieniowy SNOW 3G [9]. Wynika z niej, że dla tego szyfru istnieją ataki o złożoności czasowej dużo mniejszej (nawet $O(2^{53})$) niż złożoność ataku brutalnego wynosząca $O(2^{128})$. Konstruktorzy, w pewnym stopniu, zabezpieczyli się przed możliwością efektywnej kryptoanalizy poprzez ograniczenie długości szyfrowanych danych. Algorytm zapewniający poufność 128-NEA1 jest szyfrem strumieniowym szyfrującym /deszyfrującym bloki danych o długości do 20.000 bitów [2].

Tab. 1. Bezpieczeństwo algorytmu SNOW 3G względem klasycznego adwersarza

Algorytm	Typ ataku	Złożoność		
		Czas	Pamięć	Dane(CP)
SNOW 3G	Guess and Determine	2^{320}		$9 \cdot 2^{32}$
	Differential resynchronization attack	$2^{57.1}$	2^{25}	2^{33}
	Differential chosen IV attack	$2^{57.1}$	2^{25}	2^{33}
	Chosen IV resynchronization attack	2^{53}	-	2^{57}

Algorytm zapewniający poufność 128-NEA2 wykorzystuje szyfr AES-128 w trybie licznikowym CRT jako wewnętrzne przekształcenie kryptograficzne. Algorytm AES-128 w trybie CRT jest szyfrem strumieniowym użytkującym 128 bitowy tajny klucz kryptograficzny oraz 128 bitowy wektor inicjalizujący. W Tab. 2 zestawiono złożoności czasowo – pamięciowe oraz danych (wybranych tekstów jawnych, CP) najefektywniejszych ataków kryptoanalitycznych na algorytm AES-128 o różnej liczbie rund [6], [4]. Symetryczny algorytm blokowy AES-128 składa się z dziesięciu rund. Dla poprawnie skonstruowanego przekształcenia kryptograficznego (a takim jest AES-128), wraz ze wzrostem liczby wykonanych rund, rośnie jego odporność na ataku kryptoanalityczne realizowane przez klasycznego adwersarza.

Tab. 2. Bezpieczeństwo algorytmu AES-128 względem klasycznego adwersarza

Algorytm	Typ ataku	Liczba rund szyfru	Złożoność		
			Czas	Pamięć	Dane(CP)
AES-128	MitM	6	$2^{106.17}$	$2^{106.17}$	2^{8}
	MitM	7	$2^{126.47}$	$2^{126.47}$	2^{32}
	Impossible differentials	7	$2^{117.2}$	2^{106}	$2^{90.4}$
	SQUERE	7	2^{120}	-	2^{119}
	Bicliques	8	$2^{125.3}$	2^{8}	2^{88}
	Bicliques	10	$2^{126.2}$	2^{8}	2^{88}

W Tab. 2 pokazano, że dla algorytmu symetrycznego AES-128 o 7 rundach istnieje kilka ataków o złożoności czasowej nie niższej niż $O(2^{117})$. AES-128 o 8 rundach może być teoretycznie zaatakowany metodą Biclique [4] o złożoności czasowej $O(2^{125.3})$ oraz wymagającej 2^{88} wybranych tekstów jawnych. Teoretyczny atak na AES-128 (o dziesięciu rundach) ma złożoność niewiele mniejszą od złożoności ataku brutalnego, wynoszącej $O(2^{128})$.

Algorytm zapewniający poufność 128-NEA3 wykorzystuje szyfr ZUC jako wewnętrzne przekształcenie kryptograficzne. Algorytm ZUC jest synchronicznym szyfrem strumieniowym wykorzystującym 128 bitowy tajny klucz kryptograficzny oraz 128 bitowy wektor inicjalizujący. W Tab. 3 zestawiono złożoności czasowo – pamięciowe oraz danych (wybranych tekstów jawnych, CP) najefektywniejszych ataków kryptoanalitycznych na szyfr strumieniowy ZUC [9]. Widać z niej, że dla tego szyfru najefektywniejszym jest atak różnicowy z wybranym wektorem inicjalizującym IV o złożoności czasowej $O(2^{99.4})$ oraz wymagający $2^{13.3}$ wybranych tekstów jawnych. Inne ataki, np. Guess and Determine, mają złożoność czasową rzędu $O(2^{403})$ czyli większą niż złożoność ataku brutalnego. Konstruktorzy zabezpieczali się przed możliwością efektywnej kryptoanalizy poprzez ograniczenie długości szyfrowanych danych do 2^{32} bitów [3].

Tab. 3. Bezpieczeństwo algorytmu ZUC względem klasycznego adwersarza

Algorytm	Typ ataku	Złożoność		
		Czas	Pamięć	Dane(CP)
ZUC	Guess and Determine	2^{403}	-	$9 \cdot 2^{32}$
	Differential chosen IV attack	$2^{99.4}$	-	$2^{13.3}$

4.2. Ocena bezpieczeństwa algorytmów zapewniających uwierzytelnienie 128-NIA

Algorytmy zapewniające uwierzytelnienie 128-NIA wyznaczają 32 bitowy znacznik uwierzytelnienia MAC (ang. Message Authentication Code) przetwarzanej wiadomości przy wykorzystaniu $k = 128$ bitowego klucza kryptograficznego. Wyznaczenie kolizji dla funkcji skrótu bez klucza (np. SHA-2) zwracającej skrót o długości n-bitów ma złożoność $O(2^{n/2})$ ze względu na możliwość ataku urodzinowego. Atak na algorytm MAC wyznaczający znacznik uwierzytelnienia o długości $n/2$-bitów (o połowę krótszej niż w przypadku funkcji skrótu bez klucza) ma złożoność także $O(2^{n/2})$. Wyznaczenie kolizji dla algorytmu MAC (czyli także analizowanych 128-NIA) bez znajomości aktualnie wykorzystywanego klucza nie jest możliwe. Zatem złożoność ataku wykonywanego przez klasycznego adwersarza na uwierzytelnienie danych zabezpieczanych przy wykorzystaniu algorytmów 128-NIA wynosi $\min(2^k, 2^n)$ kroków czyli $O(2^{n=32})$. Bezpieczeństwo uwierzytelnienia danych o różnym poziomie wrażliwości, wymienianych w sieci 5G na odcinku między urządzeniem mobilnym a stacją bazową wynoszące $O(2^{32})$, jest na minimalnym dopuszczalnym poziomie.

Tab. 4. Bezpieczeństwo algorytmu ZUC względem klasycznego adwersarza

Algorytm	Dl. klucza k (bity)	Dl. MAC n (bity)	Bezpieczeństwo uwierzytelnienia
128-NIA1	128	32	$\min(2^k, 2^n) = O(2^{32})$
128-NIA2	128	32	$\min(2^k, 2^n) = O(2^{32})$
128-NIA3	128	32	$\min(2^k, 2^n) = O(2^{32})$

5. ZAGROŻENIA DLA ALGORYTMÓW ZAPEWNIAJĄCYCH POUFNOŚĆ I UWIERZYTELNIENIE ZE STRONY KRYPTOANALIZY PRZY UŻYCIU KOMPUTERÓW KWANTOWYCH

Algorytmy kryptografii symetrycznej wykorzystywane są do zapewnienia poufności danych, ich integralności a także uwierzytelnienia. Algorytmy te są efektywne obliczeniowo, a ich bezpieczeństwo opiera się na tajnym, współdzielonym kluczu kryptograficznym. Wraz z rozwojem kryptografii, obserwujemy postęp w metodach kryptoanalizy. W latach 90-ych XX wieku zostały opracowane algorytmy rozwiązywania problemów trudnych obliczeniowo w modelu ataku z użyciem komputerów klasycznych, które mają nawet złożoność wielomianową w modelu ataku z wykorzystaniem komputerów kwantowych [5]. Komputery kwantowe mają negatywny wpływ na bezpieczeństwo algorytmów kryptograficznych z kluczem symetrycznym, gdyż algorytm Grovera [7] oraz algorytm Simona [8] umożliwiają konstruowanie efektywnych, kwantowych ataków na ten typ przekształceń kryptograficznych. Do praktycznej realizacji algorytmów Shora, Grovera, Simona, brakuje (na razie) wydajnych komputerów kwantowych o przemysłowej skali zastosowania [5].

5.1. Redukcja bezpieczeństwa algorytmów zapewniających poufność 128-NEA

Algorytm Grovera jest implementacją poszukiwania elementu, który spełnia określone warunki, w zbiorze nieuporządkowanym o $N = 2^n$ elementach. Na przykład może to być wyszukiwanie nazwiska po numerze telefonu zapisanego w klasycznej książce telefonicznej. Jest to probabilistyczny problem obliczeniowy typu BQP (ang. Bounded-error Quantum Polynomial time). Klasyczne algorytmy wymagają $O(N)$ operacji, natomiast algorytm Grovera wymaga tylko $O(\sqrt{N})$ operacji, co daje kwadratowy przyrost wydajności. Weźmy pod uwagę symetryczny algorytm blokowy wykorzystujący klucz kryptograficzny o długości k bitów. Wtedy atak klasyczny ma złożoność $O(N = 2^k)$, natomiast atak z użyciem algorytmu Grovera wykonywany na komputerze kwantowym ma złożoność $O(N = 2^{k/2})$. W swoim działaniu algorytm Grovera wykorzystuje superpozycję stanów, ale głównym wyróżnikiem klasy algorytmów, do której należy, jest wzmacnianie amplitudy i wykorzystanie cech, które odróżniają amplitudę od prawdopodobieństwa [7].

Efektywność ataku z użyciem algorytmu Grovera nie zależy od aktualnie wykorzystywanego algorytmu symetrycznego zapewniającego poufność, a tylko od bitowej długości wykorzystywanego klucza kryptograficznego [7]. Wykorzystanie algorytmu Grovera do ataku na

algorytmy zapewniające poufność 128-NEA (128-NEA1, 128-NEA2, 128-NEA3) w sieci 5G, gwarantuje kwantowemu adwersarzowi kwadratowy przyrost efektywności prowadzonego ataku.

Tab. 5. Bezpieczeństwo algorytmów 128-NEA

Algorytm	Dl. klucza k (bity)	Efektywna długość klucza (bity)	
		Komputer klasyczny	Komputer kwantowy
128-NEA1	128	128	64
128-NEA2	128	128	64
128-NEA3	128	128	64

5.2. Bezpieczeństwa algorytmów 128-NIA1 oraz 128-NIA3 w ataku z użyciem komputera kwantowego

128-NIA1 oraz 128-NIA3 to algorytmy uwierzytelniające, generujące znacznik uwierzytelnienia o długości $n = 32$ bitów. Wykorzystują szyfry strumieniowe jako przekształcenia wewnętrzne, oraz $k = 128$ bitowy klucz symetryczny. Algorytmy 128-NIA1 oraz 128-NIA3 mogą zostać zaatakowane przez adwersarza kwantowego także z użyciem algorytmu Grovera. Zatem złożoność ataku wykonywanego przez kwantowego adwersarza na uwierzytelnienie danych zabezpieczanych przy wykorzystaniu algorytmów 128-NIA wynosi $\min(2^{k/2}, 2^n)$ kroków czyli w rozpatrywanym przypadku $O(2^{n=32})$.

Tab. 6. Bezpieczeństwo algorytmów 128-NIA

Algorytm	Dl. klucza k (bity)	Dl. MAC n (bity)	Bezpieczeństwo uwierzytelnienia	
			Komputer klasyczny	Komputer kwantowy
128-NIA1	128	32	$\min(2^k, 2^n) = O(2^{32})$	$\min(2^{k/2}, 2^n) = O(2^{32})$
128-NIA3	128	32	$\min(2^k, 2^n) = O(2^{32})$	$\min(2^{k/2}, 2^n) = O(2^{32})$

Porównując złożoność ataku metodą kryptoanalizy kwantowej do ataku metodą kryptoanalizy klasycznej można stwierdzić, że obydwa ataki na algorytmy uwierzytelniania 128-NIA1 oraz 128-NIA3 mają podobną złożoność wynoszącą $O(2^{32})$ kroków algorytmu.

5.3. Bezpieczeństwa algorytmu 128-NIA2

Tryby uwierzytelnienia danych mają na celu zagwarantowanie autentyczności wiadomości. Standardowy model bezpieczeństwa polega na tym, że trudno jest sfałszować wiadomość z ważnym znacznikiem, nawet mając dostęp do wyroczni (ang. *oracle*), która oblicza znacznik MAC (ang. *Message Authentication Code*) każdej wybranej wiadomości (oczywiście o sfałszowaną wiadomość wyrocznia nie może zostać zapytana). Algorytm Simona wykonywany na komputerze kwantowym pozwala na atak o złożoności wielomianowej na następujące tryby uwierzytelnienia wykorzystujące algorytmy symetryczne jako permutacje idealne: CBC-MAC, PMAC, **CMAC**, GMAC, GCM, OCB, powodując utratę autentyczności zabezpieczanych danych [8]. Złożoność obliczeniowa ataku to tylko $O(m)$, gdzie m jest długością w bitach przetwarzanego przez algorytm symetryczny bloku danych, i na ogół wynosi $m = 128$. Powodzenie ataku na tryb uwierzytelnienia nie zależy od aktualnie wykorzystywanego algorytmu symetrycznego.

Tab. 7. Bezpieczeństwo algorytmów 128-NIA2

Algorytm	Dl. klucza k (bity)	Dl. MAC n (bity)	Bezpieczeństwo uwierzytelnienia	
			Komputer klasyczny	Komputer kwantowy
128-NIA2	128	32	$\min(2^k, 2^n) = O(2^{32})$	$O(m=128=2^7)$

6. PODSUMOWANIE

Na podstawie danych zawartych w rozdziale 4, wykorzystując informacje zawarte w Tab. 1, Tab. 2, Tab. 3 oraz Tab. 4 można stwierdzić, że odporność analizowanych algorytmów zapewniających poufność 128-NEA oraz algorytmów umożliwiających uwierzytelnienie 128-NIA na ataki adwersarza klasycznego jest wystarczająca do zapewnienia ochrony danych nieklasyfikowanych o różnych stopniach wrażliwości. Nie jest jednak możliwe ich wykorzystanie do ochrony krajowych informacji niejawnych w rozumieniu *Ustawy z dnia 5 sierpnia 2010 r. o ochronie informacji niejawnych*. Szczegóły bezpieczeństwa algorytmów 128-NEA oraz 128-NIA względem adwersarza klasycznego przedstawiono w Tab. 8.

Tab. 8. Złożoność kryptoanalizy klasycznej algorytmów 128-NEA oraz 128-NIA

128-NEA – algorytmy zapewniające poufność		
128-NEA1	128-NEA2	128-NEA3
$O(2^{53})$	$O(2^{126.2})$	$O(2^{99.4})$
128-NIA – algorytmy zapewniające uwierzytelnienie		
128-NIA1	128-NEA2	128-NEA3
$O(2^{32})$	$O(2^{32})$	$O(2^{32})$

Na podstawie danych zawartych w rozdziale 5 wykorzystując informacje zawarte w Tab. 5, Tab. 6 oraz Tab. 7 można stwierdzić, że odporność analizowanych algorytmów zapewniających poufność 128-NEA oraz algorytmów umożliwiających uwierzytelnienie 128-NIA ataki realizowane przez adwersarza kwantowego nie jest zadowalająca, ale wystarczająca do zapewnienia ochrony danych nieklasyfikowanych o niskim stopniu wrażliwości. Szczegóły bezpieczeństwa algorytmów 128-NEA oraz 128-NIA względem adwersarza kwantowego przedstawiono w Tab. 9.

Tab. 9. Złożoność kryptoanalizy kwantowej algorytmów 128-NEA oraz 128-NIA

128-NEA – algorytmy zapewniające poufność		
128-NEA1	128-NEA2	128-NEA3
$O(2^{64})$	$O(2^{64})$	$O(2^{64})$
128-NIA – algorytmy zapewniające uwierzytelnienie		
128-NIA1	128-NIA2	128-NIA3
$O(2^{32})$	$O(2^7)$	$O(2^{32})$

LITERATURA

[1] 3GPP TS 33.401 v.15.10.0 (2022-01) *LTE Security architecture* (Release 15).

[2] 3GPP Confidentiality and Integrity Algorithms UEA2 & UIA2. *SNOW 3G Algorithm specification*, version 1.1. September 2006.

[3] 3GPP TS 35.222: "Confidentiality and Integrity Algorithms 128-EEA3 & 128-EIA3; Document 2: *ZUC specification*", version 1.6, June 2011.

[4] Bogdanov Andrey, Rechberger Christian.2011. *Biclique cryptanalysis of the full AES.*

[5] Borowski Mariusz, Wicik Robert. 2020. *Cryptographic protection of classified information in military radio communication faced with threats from quantum computers*, SPIE 11442.

[6] Derbez Patric, Fouque Pierre. 2015. *Exhausting Demirci-Selcuk Meet-in-the-Middle attack against reduced rounds AES*.
[7] Grover Lov Kumar. 1996. *A fast quantum mechanical algorithm for database search*, Bell Labs.
[8] Kaplan Marc, Naya-Plasencia Maria. 2016. *Breaking symmetric cryptosystems using quantum period finding*, Advances in Cryptology – CRYPTO 2016.
[9] Sulaiman Alyaa. 2014. *Comparative study on 4G/LTE cryptographic algorithms based on different factors*; International Journal of Computer and Telecommunication.

ANALIZA BEZPIECZEŃSTWA KOMUNIKACJI URZĄDZEŃ INTERNETU RZECZY WYKORZYSTUJĄCYCH PROTOKÓŁ ZIGBEE: NARZĘDZIA I SCENARIUSZE TESTOWANIA

SECURITY ANALYSIS OF INTERNET OF THINGS DEVICES USING ZIGBEE PROTOCOL: TOOLS AND TESTING SCENARIOS

Dominik Topyła[1]; Kacper Murat[1]; Krzysztof Zdulski[1]; Daniel Paczesny[1]; Jędrzej Bieniasz[1]

[1]Instytut Telekomunikacji, Politechnika Warszawska, Warszawa

DOI: 10.15199/59.2022.4.29

Streszczenie: **Artykuł prezentuje zagadnienie analizy bezpieczeństwa urządzeń Internetu Rzeczy w zakresie komunikacji bezprzewodowej protokołem ZigBee 802.15.4. Przeanalizowano referencyjną architekturę komunikacji ZigBee w zastosowaniu inteligentnego domu. Przygotowano platformę testową opartą na rozwiązaniu SDR – HackRF oraz wykonano przegląd narzędzi do wykonywania testów bezpieczeństwa protokołu ZigBee. Przygotowano scenariusze testowania bezpieczeństwa komunikacji z wykorzystaniem protokołu ZigBee w odniesieniu do znanego modelu bezpieczeństwa tego protokołu.**

Abstract: **This article presents the security analysis of IoT devices in scope of wireless communications over ZigBee 802.15.4 protocol stack. The reference ZigBee communication architecture in a smart home application was analyzed. A test platform was prepared based on the SDR - HackRF solution. Software solutions for performing security tests of the ZigBee protocol were reviewed. Security testing scenarios with the use of the ZigBee protocol were prepared with reference to the known security model of this protocol.**

Słowa kluczowe: **Internet Rzeczy, bezpieczeństwo sieci, sieci bezprzewodowe, ZigBee**

Keywords: **Internet of Things, network security, wireless networks, ZigBee**

1. WSTĘP

Urządzenia IoT (ang. Internet of Things) są coraz częściej powszechnie wykorzystywane, a ich liczba znacząco rośnie z roku na rok. Zwiększające się zapotrzebowanie na urządzenia IoT generuje potrzebę efektywnego zarządzania tymi urządzeniami, aby ułatwić ich konfigurację oraz sterowanie. Na rynku jest dostępnych wiele technologii i systemów zarówno komercyjnych, jak i ogólnie dostępnych do zarządzania urządzeniami IoT. Liczba aplikacji wykorzystujących IoT znacząco rośnie. Są one wykorzystywane w obszarach takich jak inteligentne budynki, inteligentne miasta czy monitorowanie stanu zdrowia. W takim przypadku IoT, a szerokim rozumieniu systemy cyber-fizyczne (ang. Cyber-Physical Systems, [4]) wprowadzają konwergencję świata fizycznego i cyfrowego. Koncepcja ta polega na tym, że systemy cyfrowe rozumiejąc świat fizyczny np. za pomocą czujników, mogą podejmować automatyczne decyzje, które następnie wywołują określone akcje sterujące i wpływające na świat fizyczny. Przykładem takiej automatyzacji może być system czujników w budynku, które po wykryciu ruchu włącza na pewien czas oświetlenie w danych sektorze budynku. Wspomniany system potrzebuje elementu zarządzającego poszczególnymi komponentami, umożliwiającemu dodawanie, usuwanie, konfigurację, nadzorowanie, aktualizację urządzeń. Wyróżnia się systemy, które są połączone z elementem zarządzającym przewodowo (np. platforma KNX) lub bezprzewodowo (np. platforma Home Assistant). W ramach niniejszego artykułu skupiono się na systemach wykorzystujących transmisję bezprzewodową jako tę, która szeroko jest upowszechniania w rozwiązaniach konsumenckich o wysokim poziomie popytu na nie.

Najpopularniejsze standardy wykorzystywane do lokalnej transmisji bezprzewodowej w urządzeniach IoT to Bluetooth, ZigBee oraz Wi-Fi 802.11 [14]. Każdy z wymienionych standardów ma swoje zastosowania w różnych scenariuszach IoT. Systemy wykorzystujące Bluetooth są stosunkowo tanie, łatwe i szybkie w implementacji, jednak wadą tej technologii jest stosunkowo mały zasięg. Urządzenia wykorzystujące protokół ZigBee cechują się niskim kosztem wytworzenia, niskim zużyciem energii, niewielkimi przepływnościami przesyłanych danych oraz niską stabilnością sieciową. Standard Wi-Fi 802.11 jest powszechnie stosowany w budynkach, więc możliwe jest wykorzystanie już działających urządzeń, a implementacja systemu oparta o ten standard jest łatwa. Urządzenia posiadające Wi-Fi 802.11 często mają też więcej funkcjonalności, jednak urządzenia te zużywają znacznie więcej energii niż urządzenia wykorzystujące protokół ZigBee czy Bluetooth. Należy też wspomnieć o technologiach łączności długodystansowej stosowanych do komunikacji między odległymi instalacjami IoT lub zdalnego dostępu do nich. Na te potrzeby odpowiadają m.in. sieci komórkowe piątej generacji – 5G, w ramach których scenariusz komunikacji IoT jest jednym z podstawowych [9]. Jako drugi standard należy wskazać LoRaWAN (ang. LoRa Wide Area Network). Przeznaczony jest dla urządzeń, których najważniejszymi cechami są niski pobór prądu i transmisja na daleką odległość (do 15 km).

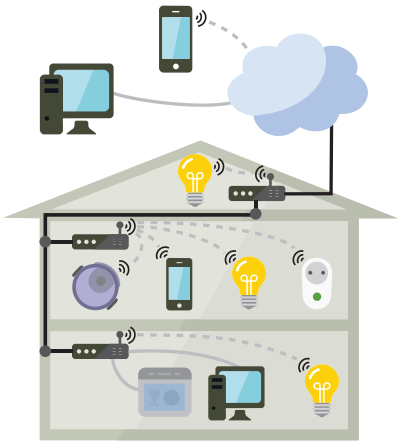
Ważnym aspektem projektowania systemu do zarządzania urządzeniami IoT jest cyberbezpieczeństwo. Często podczas projektowania urządzeń IoT i wykorzystywanych przez nie systemów, głównym celem jest umożliwienie łatwego wykorzystania funkcjonalności danego urządzenia czy całego systemu kosztem odpowiedniego

zabezpieczenia, które ma wtedy mniejszy priorytet. W niniejszym artykule zaprezentowano wstępną analizę aspektów bezpieczeństwa systemów IoT z punktu widzenia interfejsów bezprzewodowych, ze szczególnym uwzględnieniem łączności lokalnej ZigBee. Rozdział 2 opisuje referencyjne środowisko inteligentnego domu wraz ze wskazaniem potencjalnych scenariuszy cyberzagrożeń od strony interfejsu bezprzewodowego ZigBee. Rozdział 3 przedstawia narzędzia do badań interfejsów ZigBee oraz analizuje możliwości ich zastosowania do badań bezpieczeństwa komunikacji z wykorzystaniem tego protokołu. Rozdział 4 podsumowuje artykuł i przedstawia dalsze możliwe obszary rozwoju zagadnienia.

2. BEZPIECZEŃSTWO KOMUNIKACJI BEZPRZEWODOWEJ W ROZWIĄZANIACH IOT

2.1. Instalacje IoT dla inteligentnego budynku z wykorzystaniem interfejsów bezprzewodowych

W inteligentnych budynkach głównie wykorzystywaną topologią jest topologia gwiazdy. Głównym elementem tej topologii jest bramka, do której łączą się poszczególne urządzenia. Należy zaznaczyć, iż wszystkie urządzenia muszą znajdować się w zasięgu bramki. W przypadku bardziej rozbudowanej sieci urządzeń używana jest topologia rozproszonej gwiazdy, która zawiera bramki połączone są ze sobą (Rys. 1). Występuje wtedy bramka nadrzędna, która umożliwia zarządzanie i steruje pozostałymi bramkami. Użytkownik inteligentnego budynku może podłączyć się do bramki wykorzystując urządzenie mobilne (np. smartphone) i sterować siecią urządzeń będąc podłączonym lokalnie do bramki lub zdalnie w przypadku, gdy bramka jest podłączona z Internetem. W przypadku pierwszej opcji sieć inteligentnego budynku jest wydzielona od Internetu, dzięki temu ograniczone zostają ataki pochodzące z Internetu, jednak system może być nadal podatny na lokalne ataki przeprowadzane w bliskiej odległości inteligentnego budynku.



Rys. 1 Schemat ideowy komunikacji bezprzewodowej ZigBee w Inteligentnym Domu

Do zarządzania urządzeń IoT powszechnie wykorzystywane są platformy z mikrokontrolerami Arduino oraz Raspberry Pi. Główne ich zalety to niska cena, łatwa możliwość zaprogramowania, rozbudowana społeczność dzieląca się swoimi projektami oraz szeroka gama zastosowań. Arduino dobrze sprawdza się w małych, prostych projektach. Do bardziej rozbudowanych przypadków możliwe jest użycie Raspberry Pi, która ma znacznie większe możliwości i może wykonywać bardziej skomplikowane zadania.

W ramach niniejszego artykułu rozważana jest referencyjne środowisko stosujące rozwiązania IoT w scenariuszu inteligentnego domu. W ramach takiej architektury, urządzenia wchodzące w skład środowiska mogą opierać się na powszechnie dostępnych i rozwiązaniach IoT:

- Raspberry Pi 4B,
- bramka ZigBee ConBee II,
- router Linksys,
- regulator przyciemnienia Philips Hue,
- żarówki Philips Hue.

Urządzenia końcowe (regulator przyciemniania i żarówki) komunikują się wykorzystując protokół ZigBee z bramką, która pełni rolę koordynatora. Urządzenie końcowe może pełnić rolę przekaźnika, przekazując dane pomiędzy bramką i kolejnym urządzeniem końcowym zwiększając przy tym zasięg transmisji. Bramka ZigBee jest połączona poprzez port USB z platformą Raspberry Pi, na której zainstalowano oprogramowanie Home Assistant [7]. Platforma jest połączona kablem Ethernet do routera Linksys, który zapewnia dostęp do sieci poprzez wykorzystanie komunikacji bezprzewodowej Wi-Fi 802.11 lub kabla typu Ethernet. Użytkownik podłączony do routera może sterować systemem poprzez stronę internetową Home Assistant. Z jej poziomu możliwe jest dodawanie, usuwanie urządzeń i konfigurowanie urządzeń, aby reagowały w odpowiedni sposób na określone przez użytkownika akcje. Możliwe jest grupowanie urządzeń tak, aby jedną akcją wywołać reakcję grupy urządzeń.

2.2. Łączność ZigBee w systemach IoT oraz jej bezpieczeństwo

Protokół ZigBee jako realizacja standardu 802.15.4 [8] wykorzystywany jest do transmisji danych w sieciach bezprzewodowych, których główne cechy to: niewielka przepływność (do 250 kb/s dla pasma 2,4 GHz), niewielki pobór energii i niewielki zasięg (do 100 m). Protokół jest wykorzystywany w personalnych sieciach WPAN (ang. Wireless Private Area Network), sieciach sensorowych czy rozwiązaniach typu “smart home”. ZigBee pracuje w paśmie 2,4 GHz oraz w wybranych rejonach w pasmach 868 MHz i 915 MHz. W dalszej analizie skupiono się na paśmie 2,4 GHz. W tym paśmie dostępnych jest 16 kanałów o szerokości 5 MHz każdy. Używana metoda dostępu do medium w 802.15.4 to CSMA/CA, a przesyłane dane są modulowane przy pomocy O-QPSK i rozpraszane przy użyciu DSSS (ang. Direct Sequence Spread Spectrum).

ZigBee wspiera 3 typy topologii: gwiazdy, klastra oraz kraty. Protokół ZigBee w porównaniu do Bluetooth jest prostszy, ale również mniej bezpieczny. Bezpieczeństwo 802.15.4 opiera się na dwóch parach szyfrowanych kluczy 128 bitowych: sieciowym oraz połączeniowym. Klucz sieciowy jest taki sam dla całej sieci prywatnej

i wykorzystywany do komunikacji rozgłoszeniowej, natomiast klucz połączeniowy jest przypisywany do pojedynczego połączenia pomiędzy dwoma urządzeniami. W obrębie skonfigurowanej sieci ZigBee klucze są szyfrowane przed komunikacją pomiędzy urządzeniami. W przypadku gdy nowe, wcześniej nieskonfigurowane urządzenie próbuje podłączyć się do sieci ZigBee klucz sieciowy może zostać wysłany w sposób jawny, niezaszyfrowany [11]. Ta podatność może zostać wykorzystana do przeprowadzenia ataku na całą sieć prywatną.

ZigBee wspiera dwa modele bezpieczeństwa [1]: zcentralizowany i rozproszony. Pierwszy model może zawierać trzy typy urządzeń: router, koordynator oraz urządzenie końcowe. Router może przekazywać pakiety ZigBee pomiędzy urządzeniami końcowymi w obrębie danej sieci lub pomiędzy sieciami w przypadku, gdy jest połączony z kolejnymi routerami. Koordynator pracuje tylko w obrębie jednej sieci, przechowuje klucze bezpieczeństwa oraz umożliwia uwierzytelnianie urządzeń końcowych podczas dołączania ich do sieci. Urządzenia końcowe mogą być dodane do sieci ZigBee, lecz nie mogą same tworzyć sieci i przekazywać pakietów. Model rozproszony nie zawiera koordynatora, a klucze szyfrowania są przekazywane przez połączone ze sobą routery. To rozwiązanie jest mniej bezpieczne, ponieważ klucze szyfrujące nie są przechowywane przez jedno urządzenie.

3. PRZEGLĄD MOŻLIWOŚCI REALIZACJI BADAŃ BEZPIECZEŃSTWA KOMUNIKACJI PROTOKOŁEM ZIGBEE

3.1. Platforma fizyczna do testów ZigBee oparta na SDR

Przykładowe środowisko testowe składa się z 3 elementów: komputera, HackRF One i gniazdka sieciowe Tuya. Na komputerze zainstalowano system Kali Linux 2022.1, który jest przeznaczony do badania podatności i przeprowadzenia ataków. Do przechwytywania transmisji bezprzewodowej użyto urządzenia SDR (ang. Software Defined Radio) HackRF [6], które umożliwia przechwytywanie i transmisję sygnałów z zakresu częstotliwości od 1 MHz do 6 GHz. Ostatnim elementem topologii testowej jest gniazdko sieciowe TUYA IG019, które może być zarządzane przy wykorzystaniu protokołu ZigBee. Schemat ideowy środowiska testowego przedstawia Rys. 2.

Rys. 2 Schemat środowiska testowego do badań interfejsu ZigBee

Do pracy we wskazanym środowisku zastosowano odpowiednie oprogramowanie SDR, które umożliwi przechwytywanie sygnału z HackRF One i poddanie sygnału dalszym procesom. Wybrano oprogramowanie GNU Radio w wersji 3.9 ze względu na kompatybilność z wykorzystywanymi bibliotekami do badań ZigBee.

W przypadku pomyślnego zainstalowania oprogramowania głównej paczki GNU Radio, jednym z programów stowarzyszonych jest GNU Radio Companion. Jednakże, ta wersja oprogramowania nie instaluje domyślnie paczki *gr-osmosdr* [5], która jest niezbędna do pracy HackRF z protokołem ZigBee. Efektem będzie pojawienie się w GNU Radio Companion modułów *osmocom Source* i *osmocom Sink* umożliwiające odbieranie i transmisję sygnałów. Poprawną pracę HackRF z oprogramowaniem można potwierdzić uruchamiając przykładowy projekt odbiornika radia FM.

W dalszej części środowisko zostało uzupełnione o biblioteki do pracy z protokołem ZigBee. W tym celu wykorzystano ogólnodostępny projekt *gr-ieee802-15-4* z opracowanymi modułami ZigBee do GNU Radio [3]. Dodatkowo w celu możliwości przekazywania przechwyconych pakietów ZigBee do oprogramowania Wireshark wykorzystywana jest biblioteka *gr-foo* od tego samego Twórcy, co [3]. Dla poprawnej instalacji kluczowe jest użycie wersji bibliotek kompatybilnych z GNU Radio w odpowiedniej wersji.

3.2. Oprogramowanie do realizacji testów bezpieczeństwa komunikacji z wykorzystałem protokołu ZigBee

W przypadku łączności w sieciach ZigBee należy zwrócić uwagę na możliwości naruszenia podstawowych aspektów bezpieczeństwa komunikacji [1][12]. W zakresie integralności i poufności naruszenia te mogą dotyczyć podatności związanych z możliwością wykradzenia kluczy szyfrowania komunikacji. W zakresie kontroli dostępu to przede wszystkim możliwość wstrzykiwania pakietów ZigBee o dowolnej zawartości, gdy nie są one uwierzytelniane, tj. sprawdzane ich pochodzenie.

Przy pomocy wykradzionych kluczy szyfrujących nieuprawniona osoba jest w stanie odczytać dane przesyłane w pakietach ZigBee. Wstrzykiwanie pakietów ZigBee o dowolnej zawartości może prowadzić do m.in. przejęcia kontroli nad urządzeniami końcowymi, modyfikacji ich działania i ogólnego naruszenie bezpieczeństwa komunikacji ZigBee. Przygotowane środowisko w Rozdziale 3.1 może być wykorzystane do realizacji wspomnianych ataków na integralność, poufność i kontrolę dostępu za pomocą narzędzia Zigator [1][15]. Umożliwia ono realizację większości znanych i popularnych technik ataków na komunikację ZigBee:

- przechwycenie pakietów ZigBee w celu deszyfracji;
- możliwość wygenerowania arbitralnych pakietów, właściwie zaszyfrowanych i uwierzytelnionych;
- możliwość ustalenia kluczy z Trust Center oraz adresów MAC.

Razem oznacza to możliwość realizacji podsłuchu, wstrzykiwania, podszywania się i przechwytywania ko-

munikacji. W artykule [1] Autorzy narzędzia Zigator posumowali możliwości jego stosowania z wykorzystaniem innego rozwiązania SDR – USRP. W taki sposób można zastosować narzędzie HackRF wraz zainstalowanym środowiskiem GNU Radio oraz modułami odpowiedzialnymi za obsługę ZigBee, co jest istotnym wkładem prac prezentowanych w niniejszym artykule

Innymi narzędziami, które mogą być wykorzystanie w ramach środowiska z Rozdziału 3.1 są KillerBee [10] oraz jego rozwinięcie – SecBee [13]. KillerBee/SecBee stanowią zintegrowaną platformę narzędziową do testów i audytów bezpieczeństwa sieci ZigBee 802.15.4. SecBee uzupełnił możliwość wykorzystania KillerBee z rozwiązaniami SDR, np. HackRF. Wśród możliwości KillerBee/SecBee należy zwrócić uwagę na możliwość realizowania bardziej zaawansowanych scenariuszy ataków, integrujących kilka pojedynczych technik, np. w celu wywołania sztucznej kolizji w ramach sieci PAN (ang. Personal Area Networks). Ponadto oferuje podobne możliwości, co Zigator tj. podsłuch komunikacji, wstrzykiwanie spreparowanych pakietów czy wsparcie w pozyskaniu kluczy szyfrowania z komunikacji oraz urządzeń końcowych. Można uznać, że Zigator oraz platforma KillerBee/SecBee są rozwiązaniami komplementarnymi w testach bezpieczeństwa sieci i urządzeń komunikujących się protokołem ZigBee.

4. PODSUMOWANIE

Potrzeba realizacji i prowadzenia przekrojowych badań aspektów cyberbezpieczeństwa urządzeń Internetu Rzeczy staje się coraz bardziej kluczowa w zapewnianiu bezpieczeństwa cyberprzestrzeni, w której liczba takich urządzeń nieustająco rośnie. Ważnym aspektem jest, aby zespoły bezpieczeństwa miały ułatwiony dostęp do właściwych platform fizycznych i oprogramowania do prowadzenia takich testów, w szczególności takich, które wspierają realizację testów oddających realne scenariusze cyberzagrożeń, Jest to warunek konieczny, który pozwoli zaadaptować testy rozwiązań IoT już na etapie wytwarzania, zapewniając tym samym bezpieczny łańcuch wytwarzania (ang. Secure Development Life, SDLC), a co w szerszym kontekście porusza m.in. opracowanie ENISA [2]. W tym artykule przeanalizowano metody komunikacji bezprzewodowej urządzeń IoT, koncentrując się na protokole ZigBee 802.15.4. W Rozdziale 2 opisano referencyjny scenariusz komunikacji ZigBee w instalacjach IoT. W odniesieniu do tej architektury, zwrócono uwagę na podstawowe aspekty komunikacji ZigBee, które wpływają na sferę cyberbezpieczeństwa. W Rozdziale 3 przygotowano środowisko testowe oparte na rozwiązaniu SDR. Istotnym wkładem w tej części jest wykorzystanie rozwiązania HackRF, które jest popularnym i znanym rozwiązaniem do wykonywania testów penetracyjnych sieci bezprzewodowych. Następnie wskazano na możliwe techniki ataków w zakresie integralności, poufności i kontroli dostępu. Wybrano rozwiązania do realizacji zarówno pojedynczych technik, jak i całych scenariuszy ataków *end-to-end.* W kolejnych badaniach planowane jest przeprowadzenie testów różnych urządzeń IoT oraz rozszerzenie zestawu protokołów bezprzewodowych poddawanych testom. Zakłada się także wprowadzenie rozwiązań pomocniczych, które pozwolą na automatyzację części prac, co jest pożądane w kontekście wspomnianych procesów SDLC.

Badania prowadzono w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu CyberSecIdent IV, umowa nr CYBERSECIDENT/488240/IV/NCBR/2021

LITERATURA

[1] Dimitrios-Georgios Akestoridis, Madhumitha Harishankar, Michael Weber, and Patrick Tague. 2020. “Zigator: Analyzing the security of ZigBee-enabled smart homes.”, In Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '20). ACM USA, 77–88.

[2] European Union Agency for Cybersecurity (ENISA). „Guidelines for Securing the Internet of Things – Secure Supply Chain for IoT”. 2020. URL: https://www.enisa.europa.eu/publications/guidelines-for-securing-the-internet-of-things [Dostęp: 15.04.2022]

[3] gr-ieee-802-15-4. URL: https://github.com/bastibl/gr-ieee802-15-4 [Dostęp: 15.04.2022]

[4] Griffor, E., Greer, C., Wollman, D. and Burns, M. 2017. “Framework for Cyber-Physical Systems: Volume 1, Overview”, National Institute of Standards and Technology.

[5] gr-osmosdr. URL: https://github.com/osmocom/gr-osmosdr [Dostęp: 15.04.2022]

[6] HackRF One. URL: https://greatscottgadgets.com/hackrf/one/ [Dostęp: 15.04.2022]

[7] Home Assistant. „Home Assistant - Open source home automation”. URL: https://www.home-assistant.io/ [Dostęp 15.04.2022].

[8] IEEE. 2020. "IEEE Standard for Low-Rate Wireless Networks,", IEEE Std 802.15.4-2020 (Revision of IEEE Std 802.15.4-2015) , vol., no., pp.1-800.

[9] ITU-R. 2015. „M.2083-0, IMT Vision-Framework and overall objectives of the future development of IMT for 2020 and beyond M Series Mobile, radiodetermination, amateur and related satellite services”.

[10] KillerBee. URL: https://github.com/riverloopsec/killerbee [Dostęp: 15.04.2022]

[11] Kostadinov, D. 2015. „ Hacking ZigBee networks”. URL: https://resources.infosecinstitute.com/topic/hacking-zigbee-networks/ [Dostęp 15.04.2022].

[12] Meneghello, F., Calore, M., Zucchetto, D., Polese, M., & Zanella, A. 2019. “IoT: Internet of Threats? A Survey of Practical Security Vulnerabilities in Real IoT Devices”. IEEE Internet of Things Journal, 6, 8182-8201.

[13] SecBee. URL: https://github.com/Cognosec/SecBee [Dostęp: 15.04.2022]

[14] Stolojescu-Crisan, Cristina, Calin Crisan, and Bogdan-Petru Butunoi. 2021. "An IoT-Based Smart Home Automation System" Sensors 21, no. 11: 3784.

[15] Zigator. URL: https://github.com/akestoridis/zigator [Dostęp: 15.04.2022]

MODYFIKACJA OPROGRAMOWANIA URZĄDZEŃ INTERNETU RZECZY: ANALIZA MOŻLIWOŚCI I WPŁYWU NA BEZPIECZEŃSTWO

FIRMWARE MODIFICATIONS IN DEVICES OF INTERNET OF THINGS: EVALUATION OF SECURITY AND IMPACT

Krzysztof Zdulski[1]; Dominik Topyła[1]; Kacper Murat[1]; Daniel Paczesny[1]; Jędrzej Bieniasz[1]

[1]Instytut Telekomunikacji, Politechnika Warszawska, Warszawa

DOI: 10.15199/59.2022.4.30

Streszczenie: **W artykule przedstawiono zagadnienie modyfikacji oprogramowania urządzeń IoT i wynikających z tego potencjalnych zagrożeń dla bezpieczeństwa cyberprzestrzeni. Badania przeprowadzono z wykorzystaniem łatwo dostępnego i niskobudżetowego urządzenia IoT. Przeprowadzono badania wstępne oraz zrealizowano symulację ataku wykorzystującego komunikację Command&Control między urządzeniem IoT, a serwerem Atakującego. Przedyskutowano i podkreślono ważność problemu bezpieczeństwa oprogramowania i łańcucha dostaw dla urządzeń IoT.**

Abstract: **This article presents the issue of software modification of IoT devices and the potential impacts for cybersecurity. The research was carried out with the use of an easily available, low-cost IoT device. Initial research was carried out and an attack simulation using Command & Control communication between the IoT device and the attacker's server was executed. The importance of software security and the supply chain for IoT devices was discussed and emphasized.**

Słowa kluczowe: **bezpieczeństwo oprogramowania, cyberbezpieczeństwo, Internet Rzeczy, firmware**

Keywords: **software security, cybersecurity, Internet of Things, firmware**

1. WSTĘP

W ostatnich latach obserwuje się postępujący wzrost popytu na automatyzację wszelkich procesów gospodarczych i społecznych oraz integrację ich z sieciami internetowymi. Internet Rzeczy (ang. Internet of Things, IoT) jest prężnie rozwijającym się paradygmatem technologicznym będącym odpowiedzią na potrzeby rynku. Koncepcja ta opiera się na wykorzystaniu urządzeń o możliwości wzajemnej komunikacji do stworzenia sieci o pewnym stopniu inteligencji, podejmującej automatyczne decyzje w oparciu o zaprojektowane funkcje lub reagującej na działania użytkowników. Nieograniczony stopień swobody doboru urządzeń i sposobów wymiany informacji implikuje ogromny potencjał w wielu branżach takich jak ochrona zdrowia [2], przemysł samochodów autonomicznych [8] czy inteligentnych budynków [7].

Jednym z podstawowych cech wielu dostępnych urządzeń IoT jest możliwość ich modyfikacji przez Użytkowników. Można traktować ten aspekt zarówno jako dużą zaletę jak i potencjalne źródło problemów. Z jednej strony oczekiwane jest, aby Użytkownik otrzymywał określony stopień swobody w konfigurowaniu czy nawet modyfikowaniu urządzeń IoT pod swoje wymagania. Z drugiej strony, przy niskiej świadomości o ryzykach dla bezpieczeństwa cyberprzestrzeni otwiera to przestrzeń, którą mogą wykorzystywać Atakujący. Możliwość wpływania na oprogramowanie układowe urządzenia, tj. firmware stwarza możliwości do realizacji różnych scenariuszy cyberzagrożeń. Dzięki temu Atakujący mogą doprowadzić do podsłuchiwania ruchu w sieci lokalnej, podszywania się, czy utrudniania właściwej komunikacji. Modyfikacje firmware można traktować jako jeden z kroków umożliwiający realizację szeregu technik ataków określanych jako aktywne, co zostało usystematyzowane w [1].

W niniejszym artykule podjęto zagadnienie modyfikacji oprogramowania układowego tanich urządzeń IoT ze skutkami w sferze bezpieczeństwa cyberprzestrzeni. W Rozdziale 2 sprawdzono jakie są możliwości badań urządzeń IoT pod kątem ich bezpieczeństwa. Przeprowadzono referencyjną analizę jednego z dostępnych urządzeń IoT w zakresie zachowania się w sieci oraz możliwości ekstrakcji firmware. W Rozdziale 3 zaproponowano scenariusz zagrożenia powiązany z możliwością wykorzystania dostępnego otwarto-źródłowego oprogramowania do wybranego urządzenia IoT. Dokonano własnej modyfikacji wybranego oprogramowania i następnie zrealizowano zaproponowany scenariusz cyberzagrożenia. Na podstawie wyników przeprowadzono dyskusję nt. bezpieczeństwa oprogramowania, jego wytwarzania i sposobu dystrybucji dla urządzeń IoT umieszczając te zagadnienia w szerszym kontekście bezpieczeństwa łańcucha dostaw dla rozwiązań IoT. Rozdział 4 podsumowuje niniejszy artykuł i wskazuje na dalsze kierunki badań.

2. BADANIA WSTĘPNE WYBRANEGO URZĄDZENIA IOT

2.1. Analiza wstępna gniazdka Gosund SP111

Gniazdko Gosund SP111 to inteligentne gniazdko łączące się z chmurą chińskiej firmy Tuya. Poza możliwością włączenia i wyłączenia gniazdka pozwala ono na pomiar mocy podłączonego urządzenia wraz z historią zużycia energii. Firma Tuya pozwala producentom sprzętu w bardzo łatwy sposób zintegrować rozwiązania Internetu Rzeczy w ich produktach zapewniając zarówno

zasoby chmurowe do kontroli urządzeń, SDK do tworzenia aplikacji do kontroli urządzeń oraz oprogramowania układowego do tych urządzeń. Znakomita większość tanich urządzeń na rynku korzysta z rozwiązań Tuya, więc jest to bardzo ciekawy obiekt do analiz i badań cyberbezpieczeństwa.

Pierwszym krokiem w celu podłączenia urządzenia do chmury Tuya jest połączenie go z telefonem za pomocą aplikacji Tuya (lub innej stworzonej przy pomocy ich SDK). W ten sposób możemy skonfigurować urządzenie i podłączyć je do naszej sieci Wi-Fi 802.11, SSID oraz hasło podawane jest w aplikacji. Urządzenie wspiera jedynie sieci 2.4GHz. Z punktu widzenia analizy podstawowym narzędziem jest *Wireshark* do obserwacji komunikacji na wskazanych interfejsach sieciowych. Urządzenie po podłączeniu się do sieci z Wi-Fi 802.11 wykonuje procedurę rejestracji w chmurze. Sam proces rejestracji wymaga dokładniejszego zbadania w przyszłości, a część analizy oparta jest na przypuszczeniach. Urządzenie najprawdopodobniej pobiera listę serwerów MQTTS dla regionu, w którym się znajduje. Odpytuje przy tym kilka adresów URL, m.in.: https://a3.tuyacn.com/gw.json czy https://a3.tuyaeu.com/gw.json. Podjęto próbę zbadania wskazanych serwerów, jednakże nie było możliwe nawiązanie połączenia. Wynika to z tego, że serwer spodziewa się komunikacji zaszyfrowanej, w której część klucza szyfrowania została zaszyta wcześniej – Pre-Shared Key (PSK), tj. klucze symetryczne zostały wcześniej wymienione pomiędzy serwerem oraz urządzeniem końcowym. Na tym etapie dokonano pierwszych analiz oprogramowania wbudowanego umieszczonego w urządzeniu, które uzyskano w procedurze ekstrakcji za pomocą narzędzia *esptool* [3]. W pozyskanym pliku oprogramowania układowego zwrócono uwagę na ciąg znaków umieszczony pod adresem w pamięci `0xfb000`. Wskazany fragment pliku okazał się obiektem tekstowym w formacie JSON zawierającym 4 pary klucz-wartość (Tab. 1):

- *mac* - ostatnie 3 znaki adresu MAC mikro kontrolera ESP8266 (zapisane w formacie szesnastkowym)
- *prod_idx* - unikatowy identyfikator urządzenia
- *auz_key* - ciąg 192 bitów zakodowany w formacie base64 (pod adresem 0x70b14 znajduje się alfabet base64). Najprawdopodobniej wykorzystywany podczas procedury generowania klucza TLS-PSK.
- *prod_test* - zmienna boolowska, najprawdopodobniej wykorzystywana podczas testów po-produkcyjnych urządzenia

Ustalono zatem, że klucze te faktycznie początkowo zawarte są w oprogramowaniu układowym urządzenia [10], jednakże wymagane są dalsze testy, żeby ustalić ich format oraz zweryfikować możliwość odszyfrowania komunikacji.

Tab. 1. Obiekt JSON zawierający 4 parametry powiązane z szyfrowaniem komunikacji sieciowej gniazdka

```
{
  "mac": "e52",
  "prod_idx": "20538630",
  "auz_key":"b5XZaEhdmrMiZXirzpNuQDvRTFK335e",
  "prod_test": false
}
```

Dodatkowo zauważono, że w oprogramowaniu układowym zawarty jest alfabet *base64*, więc jest możliwe, że klucz zakodowany jest w tym formacie. Urządzenie po zarejestrowaniu się w chmurze Dostawcy oraz zaktualizowaniu listy serwerów MQTTS przechodzi w tryb operacyjny. Serwery MQTTS podobnie do serwerów HTTPS Dostawcy używają TLS-PSK, więc jednym z dalszych wyzwań jest przechwycenie klucza, aby poznać więcej szczegółów protokołu komunikacji. Port wykorzystywany przez MQTTS to 8886. Dodatkowo urządzenie regularnie wysyła pakiety typu broadcast protokołem UDP na port 6667 powiązane z TuyaAPI. Protokół ten może pozwolić na lokalną kontrolę urządzenia [5].

2.2. Wnioski dla dla dalszych badań bezpieczeństwa IoT

Kroki wykonane podczas analizy gniazdka Gosund SP111 można powtarzać dla wielu klas urządzeń. Analiza zachowań sieciowych przy pomocy narzędzia Wireshark oraz proste analizy oprogramowania układowego są elementarnymi krokami w tworzonych procedurach analizy urządzeń Internetu Rzeczy. Analiza oprogramowania układowego może zostać przeprowadzona w przyszłości w sposób zautomatyzowany, z wykorzystaniem prostych skryptów napisanych w języku powłoki, lub bardziej zaawansowanych w języku Python. Analiza zachowań sieciowych też może zostać zautomatyzowana, jednak nadal potrzebna jest interakcja z urządzeniem w celu zebrania tych danych.

Kolejnymi ścieżkami w analizie gniazdka mogą być próby przełamania zabezpieczeń, w tym odszyfrowania komunikacji z serwerami HTTP i MQTT. Do tego celu konieczna jest analiza wsteczna oprogramowania układowego zgranego z gniazdka i zrozumienia funkcji wyprowadzania klucza PSK służącego do szyfrowania. Jednakże w ramach dalszej części artykułu zajęto się możliwościami modyfikacji samego oprogramowania układowego. Jednym z rozpatrywanych scenariuszy cyberzagrożenia jest możliwość wgrania zmodyfikowanego otwartoźródłowego oprogramowania układowego, np. takiego jak Tasmota [9], które mogłoby wykradać SSID oraz hasło do sieci bezprzewodowej Wi-Fi, w której zostało podłączone urządzenie ze zmodyfikowanym oprogramowaniem. W ten sposób możliwe byłoby uzyskiwane dostępów do wielu sieci Wi-Fi 802.11, gdzie proste gniazdko IoT stawałoby się *tylną furtką* do tej sieci.

3. BADANIA MOŻLIWOŚCI MODYFIKACJI OPROGRAMOWANIA URZĄDZEŃ IOT

3.1. Ekstrakcja i ponowne wgranie oprogramowania układowego

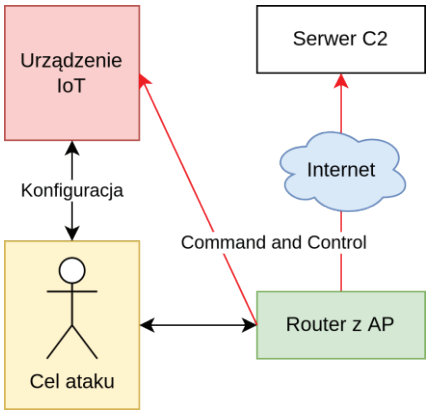
W eksperymentach badawczych Autorów artykułu dokonano testów mających na celu ustalenie możliwości zgrania oraz wgrania oprogramowania układowego wskazanego gniazdka Gosund SP111, w ogólności będącym urządzeniem Internetu Rzeczy opartym na mikrokontrolerze ESP8266. Wykorzystuje ono standardowy program rozruchowy znajdujący się w pamięci typu Flash urządzenia. Eksperyment ten zakończył się sukcesem, możliwe okazało się zgranie oraz ponowne wgranie oprogramowania układowego wykorzystując standardowe narzędzie *esptool* [3].

3.2. Przygotowanie eksperymentu

Eksperyment polegał na przetestowaniu możliwości modyfikacji oprogramowania układowego w sposób, który pozwoliłby na przesłanie identyfikatora oraz hasła sieci bezprzewodowej do serwera Command&Control (C2) Atakującego, znajdujący się poza siecią lokalną, w Internecie. Scenariusz ataku realizowany jest bez wiedzy Ofiary („Cel ataku") i przy jej działaniach polegających na konfiguracji zmodyfikowanego urządzenia IoT zgodnie z instrukcją. Na symulacji serwera Atakującego została uruchomiona usługa oczekująca na połączenia na arbitralnie wybranym porcie TCP i logująca odbierane dane. Całościowy scenariusz ataku przedstawiony został na Rys. 1. Na etapie badań procedura zakłada brak urządzeń blokujących lub w inny sposób utrudniających komunikację urządzeń w sieci lokalnej z resztą Internetu, takich jak firewall lub IDS/IPS (Intrusion Detection/Intrusion Prevention System). W takim otwartym podejściu możliwe będzie także sformułowanie właściwych wniosków oraz rekomendacji dla budowy bezpiecznej architektury sieci IoT.

3.3. Przygotowanie środowiska testowego

Pierwszym krokiem podczas przygotowania środowiska testowego jest rozkręcenie inteligentnego gniazdka korzystając ze śruby umiejscowionej w otworze na bolec uziemienia. Następnie podłączamy programator. Na komputerze służącym do programowania zainstalowano narzędzie *platformio* [6] ułatwiające kompilację, wgrywanie oprogramowania oraz debugowanie.



Rys. 1 Scenariusz komunikacji C2 z wykorzystaniem gniazdka GoSound

3.4. Modyfikacja oprogramowania układowego

Modyfikacja skompilowanego oprogramowania układowego jest problematyczna, więc podstawą opracowania zmodyfikowanej wersji firmware było zastosowanie otwarto-źródłowego oprogramowanie Tasmota [9]. Jest to uniwersalne oprogramowanie na platformę ESP8266, stworzone w oparciu o popularny framework Arduino, pozwalające na obsługę wielu urządzeń IoT. Jego popularność wynika z niskiego progu wejścia do zastosowania, gdyż posługuje się mechanizmem deklaratywnych szablonów do opisu funkcji, które dane złącza pełnią w zadanym urządzeniu.

Na potrzeby eksperymentu przygotowano modyfikację oprogramowania układowego w taki sposób, aby wysyłało ono identyfikator sieci oraz hasło na serwer atakującego w momencie startu wewnętrznego serwera HTTP. Wybór warunków wykonania tej akcji nie zostało wybrane przypadkowo. Wynika ono z analiz wstępnych urządzenia, gdy eksperymentalnie potwierdzono, że w tym momencie urządzenie posiada już dostęp do Internetu, co jest niezbędne do nawiązania symulowanej komunikacji C2. Kod modyfikacji zawiera Tab. 2.

Tab. 2 Propozycja modyfikacji firmware na potrzeby realizacji scenariusza komunikacji C2

```
/* tasmota/xdrv_01_webserver.ino */
void StartWebserver(int type, IPAddress ipweb)
{
  ...
  Wi-FiClient client;
  if(client.connect("<IP OR HOSTNAME>", PORT))
  {
    client.print(SettingsText(SET_STASSID1 +
        Settings->sta_active));
    client.print(":");
    client.println(SettingsText(SET_STAPWD1 +
        Settings->sta_active));
    client.flush();
    client.stop();
    Serial.println("[C2] Success");
  }
  else {
    Serial.println("[C2] Error!");
}
...
```

Następnie potwierdzono prawidłową transmisję danych po stronie serwera C2 obserwując logi.

3.5. Wnioski z eksperymentu dla cyberbezpieczeństwa

Opisany w Rozdziale 3.4 eksperyment pokazał znaczącą lukę w architekturze konsumenckich rozwiązań IoT, takich jak badane gniazdko Gosund SP111, ale również innych opartych o platformę ESP. Przede wszystkim, w prosty sposób możliwe jest odkrycie interfejsu reprogramującego to urządzanie, a tym samym wgranie zmodyfikowanego oprogramowanie układowego nie stanowi problemu. Ponadto, istnieją duże społeczności entuzjastów tych systemów, którzy udostępniają wiele rad i instrukcji do przeprowadzania takich modyfikacji. Można przy tym zwrócić uwagę, że Użytkownicy rzadko weryfikują pobierane oprogramowanie do takich urządzeń, co wpływa na wzrost ryzyka dotarcia z tak modyfikowanym oprogramowaniem do odbiorów. Innym sposobem dystrybucji może być odsprzedaż takich gniazdek z drugiej ręki, gdy dokonano już zmian w oprogramowaniu. Tym samym można łatwo uzyskania nieuprawniony dostęp do sieci bezprzewodowych w różnych sieciach. Warto zwrócić uwagę, że w ramach eksperymentu oprócz identyfikatora oraz hasło sieci bezprzewodowej, wykradziono dodatkowo adres IP, z którego nawiązane zostało połączenie. Należy też rozpatrywać, że możliwe jest opracowanie bardziej zawansowanych modyfikacji pozwalających zamienić gniazdko w urządzenie stanowiące punkt przesiadkowy lub ogólnie serwer komunikacji zwrotnej dla Atakującego. Jest to szczególnie przydatne do omijania firewalli w sieciach, które przepuszczają ruch wejściowy tylko wtedy, gdy połączenie zostało nawiązane

z wewnątrz. Prowadzi to do możliwości realizacji bardziej zaawansowanych scenariuszy, w których przeprowadzane mogą być cyberataki na inne urządzenia dostępne w sieci lokalnej. W ostatnich latach w ten sposób wykorzystywano między innymi kamery IoT [11].

Uwidacznia się tym samym ogólnie określany problem cyberbezpieczeństwa łańcucha dostaw dla rozwiązań IoT [4] w powiązaniu z szerokim problem świadomości Użytkowników cyberprzestrzeni. Otwartość rozwiązań z jednej strony ma wiele zalet jak wysokie zainteresowanie społeczności, dostępność wsparcia czy odblokowywanie nowych funkcji. Z drugiej strony należy zwiększać świadomość w zakresie możliwości wykorzystania tak prostych, tanich i łatwo dostępnych systemów IoT jako elementu realizacji większych scenariuszy cyberzagrożeń.. W zakresie tego tematu powstało w ostatnim czasie kilka ważnych opracowań, w tym propozycji standardów prowadzących do podniesienia cyberbezpieczeństwa rozwiązań końcowych, w tym IoT jak przytoczone opracowanie ENISA [4].

4. PODSUMOWANIE

Badania bezpieczeństwa urządzeń IoT można prowadzić w łatwy i dostępny sposób. Istotne jest, aby testy bezpieczeństwa stawały się obowiązkowym etapem wytarzania rozwiązań IoT. Dzięki temu można rozwiązywać problemy bezpieczeństwa zanim dojdzie do realizacji faktycznych scenariuszy zagrożeń w cyberprzestrzeni. W niniejszym artykule zademonstrowano kolejne kroki badań wybranego urządzenia IoT w kierunku zbadania możliwości modyfikacji oprogramowania układowego, tj. firmware. Wskazano, że firmware pobrany wprost z urządzenia stanowi wyzwanie inżynierii wstecznej, natomiast możliwe jest jego łatwe zastąpienie przez dostępne rozwiązania otwarto-źródłowe. W Rozdziale 3 pokazano możliwość przygotowania firmware w oparciu na otwarto-źródłowym projekcie Tasmota do realizacji scenariusza z utworzeniem kanału Command&Control (C2) dla potencjalnego złośliwego oprogramowania dla urządzeń IoT. Osiągnięte wyniki przedyskutowano w szerszym kontekście bezpieczeństwa wytwarzania i dostarczania urządzeń IoT w różnych zastosowaniach. W kolejnych latach musi rosnąć świadomość twórców rozwiązań IoT w tym zakresie. Takie opracowania jak niniejszy artykuł czy inne dostępne przewodniki [4] przybliżają i upowszechniają wiedzę o tych problemach, a także sposoby ich rozwiązywania. W kolejnych badaniach planowane jest uwzględnienie innych urządzeń IoT stosowanych w zastosowaniach inteligentnego domu. Zostaną też podjęte kroki związane z inżynierią wsteczną oryginalnego firmware, w szczególności pod kątem zbadania poprawności stosowania mechanizmów bezpieczeństwa w protokołach komunikacyjnych.

Badania prowadzono w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu CyberSecIdent IV, umowa nr CYBERSECIDENT/488240/IV/NCBR/2021

LITERATURA

[1] Butun, I., Osterberg, P., & Song, H. 2020. "Security of the Internet of Things: Vulnerabilities, Attacks, and Countermeasures". In IEEE Communications Surveys and Tutorials (Vol. 22, Issue 1, pp. 616–644).

[2] De Michele, R., Furini, M. 2019. "IoT Healthcare: Benefits, Issues and Challenges". ACM International Conference Proceeding Series, s. 160–164.

[3] esptool. URL: https://github.com/espressif/esptool [Dostęp 08.03.2022].

[4] European Union Agency for Cybersecurity (ENISA). „Guidelines for Securing the Internet of Things – Secure Supply Chain for IoT". 2020. URL: https://www.enisa.europa.eu/publications/guidelines-for-securing-the-internet-of-things [Dostęp: 08.03.2022]

[5] Localtuya. "A Home Assistant custom Integration for local handling of Tuya-based devices." URL: https://github.com/rospogrigio/localtuya [Dostęp 08.03.2022].

[6] platformio. „PlatformIO - Professional collaborative platform for embedded development". URL: https://platformio.org/ [Dostęp 08.03.2022].

[7] Risteska Stojkoska, B.L.,Trivodaliev, K. V. 2017. "A review of Internet of Things for smart home: Challenges and solutions", Journal of Cleaner Production, vol. 140, pp. 1454–1464.

[8] Syafrudin, M., Alfian, G., Fitriyani, N.L., Rhee, J. 2018. "Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing", Sensors, vol. 18, no 9.

[9] Tasmota. "Alternative firmware for ESP8266 and ESP32 based devices with easy configuration." URL: https://github.com/arendst/tasmota [Dostęp 08.03.2022].

[10] Tuya. 2022. "Tuya Convert. Collaboration document for PSK Identity" URL: https://github.com/ct-OpenSource/tuya-convert/wiki/Collaboration-document-for-PSK-Identity-02 [Dostęp 08.03.2022].

[11] Fagan, M. , Megas, K. , Marron, J. , Brady, K. , Cuthill, B. , Herold, R. , Lemire, D. and Hoehn, N. 2021. "IoT Device Cybersecurity Guidance for the Federal Government: Establishing IoT Device Cybersecurity Requirements", National Institute of Standards and Technology.

[12] ZDNet „ Critical IoT security camera vulnerability allows attackers to remotely watch live video - and gain access to networks". 2021. URL: https://www.zdnet.com/article/critical-iot-security-camera-vulnerability-allows-attackers-to-remotely-watch-live-video-and-gain-access-to-networks/ [Dostęp: 08.03.2022]

KSZTAŁCENIE W ZAKRESIE CYBERBEZPIECZEŃSTWA NA POLITECHNICE WARSZAWSKIEJ

EDUCATION IN THE FIELD OF CYBERSECURITY AT WARSAW UNIVERSITY OF TECHNOLOGY

Andrzej Kraśniewski[1], Krzysztof Szczypiorski[2]

[1] Politechnika Warszawska, Warszawa, e-mail: andrzej.krasniewski@pw.edu.pl
[2] Politechnika Warszawska, Warszawa, e-mail: krzysztof.szczypiorski@pw.edu.pl

DOI: 10.15199/59.2022.4.31

Streszczenie: **Przedstawiono spektrum działań związanych z kształtowaniem oferty edukacyjnej Politechniki Warszawskiej w zakresie cyberbezpieczeństwa. Szczególnie dużo uwagi poświęcono innowacyjnemu podejściu do opracowania programów studiów pierwszego i drugiego stopnia oraz pozytywnym doświadczeniom, potwierdzonym opiniami szerokiego grona interesariuszy, związanym z koncepcją studiów pierwszego stopnia i jej praktyczną realizacją.**

Abstract: **The development of the education offer in cybersecurity at the Warsaw University of Technology is presented. Particular attention is given to the innovative approach underlying the development of the first- and second-cycle degree programmes. A positive perception of the conceptual framework and the actual implementation of the first-cycle programme, confirmed by opinions of a wide group of stakeholders, is also reported.**

Słowa kluczowe: **cyberbezpieczeństwo, kształcenie, Politechnika Warszawska, studia.**

Keywords: **cybersecurity, degree programmes, education, Warsaw University of Technology.**

1. WSTĘP

Historia kształcenia w zakresie cyberbezpieczeństwa na Politechnice Warszawskiej sięga końca ubiegłego wieku, kiedy to na Wydziale Elektroniki i Technik Informacyjnych kilka przedmiotów tematycznie związanych głównie z kryptografią włączono do oferty przeznaczonej dla studentów studiów pierwszego i drugiego stopnia na kierunkach *informatyka* oraz *telekomunikacja*. W tym okresie realizowano też dość liczne prace dyplomowe – inżynierskie i magisterskie, pod kierunkiem osób prowadzących badania w wybranych obszarach cyberbezpieczeństwa. Liczba studentów zainteresowanych realizacją takich pracami oraz przedmiotami z zakresu cyberbezpieczeństwa szybko rosła, wskazując na potrzebę znacznego wzbogacenia oferty edukacyjnej w tym obszarze.

Istotnym impulsem do rozszerzenia takiej oferty było utworzenie w 2015 roku w Instytucie Telekomunikacji na Wydziale Elektroniki i Technik Informacyjnych PW Zakładu Cyberbezpieczeństwa.

W roku akademickim 2017/2018, w ramach studiów drugiego stopnia na kierunku *telekomunikacja*, uruchomiono specjalność *teleinformatyka i cyberbezpieczeństwo*, przeznaczoną dla absolwentów studiów pierwszego stopnia na kierunku *telekomunikacja* i kierunkach pokrewnych. Wśród kadry akademickiej narastało jednak przekonanie, że nadszedł czas na bardziej zasadnicze przedsięwzięcia związane z kształceniem w zakresie cyberbezpieczeństwa. Ich prezentacja jest przedmiotem referatu.

Omówiono realizowane w okresie ostatnich kilku lat działania związane z kształtowaniem oferty edukacyjnej Politechniki Warszawskiej w zakresie cyberbezpieczeństwa, koncentrując uwagę na innowacyjnym podejściu do opracowania programów studiów pierwszego i drugiego stopnia na kierunku *cyberbezpieczeństwo* prowadzonych na Wydziale Elektroniki i Technik Informacyjnych

2. STUDIA PIERWSZEGO STOPNIA

Program studiów pierwszego stopnia na kierunku *cyberbezpieczeństwo* opracowano w ramach realizowanego przez PW, finansowanego przez NCBR programu NERW (Nauka – Edukacja – Rozwój – Współpraca). Prace rozpoczęto z początkiem 2018 roku, a w październiku 2019 roku pierwsza grupa 60 studentów rozpoczęła kształcenie na tym nowym kierunku studiów prowadzonym na Wydziale Elektroniki i Technik Informacyjnych. W referacie omówiono jedynie podstawowe cechy tych studiów; więcej informacji można znaleźć na poświęconej im stronie internetowej [1].

Zakres wiedzy i umiejętności absolwenta oraz program studiów prowadzący do ich uzyskania odpowiada najnowszym globalnym rozwiązaniom opublikowanego w 2017 roku raportu *National Initiative for Cybersecurity Education* (USA) [4], wymagań zdefiniowanych przez międzynarodowe organizacje i stowarzyszenia inżynierskie oraz najbardziej rozpoznawalną w świecie agencję akredytującą programy studiów technicznych – ABET (obszerną analizę sposobu definiowania kompetencji absolwentów studiów w zakresie cyberbezpieczeństwa można znaleźć w [2]).

Program powstał we współpracy ze wszystkimi istotnymi interesariuszami Jako podstawową zasadę w pracach programowych przyjęto, że opracowany program ma odpowiadać przede wszystkim potrzebom i interesom odbiorców (studentów i pracodawców), a w mniejszym stopniu – interesom kadry nauczającej.

Przy tworzeniu programu uwzględniono uwagi przekazane przez przedstawicieli kilkunastu działających w kraju instytucji zainteresowanych zatrudnieniem absolwentów studiów na tym kierunku – instytucji administracji centralnej, instytucji finansowych, firm informatycznych i operatorów sieci telekomunikacyjnych (Asseco, Centralne Laboratorium Kryminalistyczne Policji, Cert Polska, Cryptomage, IBM, KPMG, Narodowe Centrum Kryptologii, Orange, PKO BP, Play, T–Mobile, VT Cyber, WB Electronics). Stałymi konsultantami zespołu twórców programu byli: Dyrektor Departamentu Cyberbezpieczeństwa w ówczesnym Ministerstwie Cyfryzacji, Dyrektor ds. infrastruktury krytycznej i cyberbezpieczeństwa w Thales Polska, prezes Fundacji Bezpieczna Cyberprzestrzeń oraz zastępca dyrektora Naukowo-Akademickiej Sieci Komputerowej.

W pracach programowych – w całym 10-miesięcznym okresie ich trwania – uczestniczyli studenci. Ich udział w tych pracach nie ograniczał się do opiniowania propozycji przedstawianych przez kadrę akademicką. Opierając się na doświadczeniach wynikających z pracy zawodowej oraz uczestnictwa w programach międzynarodowej wymiany studentów (ERASMUS+, ATHENS), studenci przedstawiali konkretne propozycje dotyczące pożądanych treści programowych, wykorzystywanych narzędzi (w tym języków programowania i narzędzi programistycznych) oraz form realizacji zajęć. Propozycje te wpłynęły w znacznym stopniu na ostateczny kształt programu, co stanowiło przykład skutecznego wdrożenia koncepcji współdziałania kadry akademickiej i studentów w procesie kształtowania oferty edukacyjnej, stanowiącej jedną z najlepszych metod doskonalenia jakości kształcenia [3].

Program studiów charakteryzuje się kilkoma oryginalnymi, niezbyt powszechnie występującymi w praktyce polskich uczelni rozwiązaniami. W szczególności, zestaw zajęć na pierwszym semestrze został zaprojektowany z myślą o ułatwieniu „aklimatyzacji" nowo przyjętym studentom, ich wzajemnej integracji, wzmocnieniu zainteresowania studiami technicznymi, motywacji do studiowania na Wydziale, a zarazem z intencją wyposażenia ich w istotne kompetencje uniwersalne, przydatne w dalszym kształceniu. Obejmuje on m.in. następujące przedmioty/moduły zajęć:

- *Wprowadzenie do cyberbezpieczeństwa* – przedmiot, na którym studenci współdziałając z prowadzącymi, realizują swoje pierwsze, proste projekty związane z zagadnieniami cyberbezpieczeństwa,
- *Pozatechniczne aspekty pracy inżyniera* – zajęcia, mające charakter warsztatowo-treningowy, w ramach których studenci rozwiązują problemy, analizując studia przypadku, biorąc pod uwagę społeczne, ekonomiczne, prawne, etyczne i inne pozatechniczne uwarunkowania działalności zawodowej inżyniera, pogłębiają wiedzę poprzez samodzielne wyszukanie i analizę treści odpowiednich materiałów źródłowych i przedstawiają wyniki swoich prac w formie opracowań i prezentacji, przygotowanych zgodnie z omówionymi na zajęciach zasadami,
- *Szybkie prototypowanie inżynierskie* – zajęcia, w trakcie których studenci, początkowo indywidualnie, a następnie w zespołach, konstruują mini-roboty, rywalizując o osiągnięcie jak najlepszych parametrów swoich projektów.

Tak więc już na pierwszym semestrze studenci projektują i konstruują fizyczne obiekty, odnosząc w ten sposób pierwsze „realne" sukcesy w karierze inżyniera.

Znamienną cechą programu jest ograniczony wymiar wykładów. Dominują zajęcia prowadzone w formach aktywizujących studentów, ukierunkowane na zdobywanie wiedzy i umiejętności poprzez rozwiązywanie problemów i projektowanie. Dobrze ilustruje to przykład przedmiotów matematycznych. Dominują ćwiczenia oraz zajęcia laboratoryjne i projektowe, a rozwiązywane zadania i realizowane projekty odnoszą się do realnych problemów związanych z zagadnieniami cyberbezpieczeństwa.

20 maja 2019 r. nowo opracowany program studiów na kierunku *cyberbezpieczeństwo* został zaprezentowany na zorganizowanej przez Ministerstwo Nauki i Szkolnictwa Wyższego konferencji NKN Forum – Kształcenie, która zgromadziła ok. 350 uczestników. "Zamówiona" przez organizatorów prezentacja programu (jedyna prezentacja ilustrowana multimedialnie podczas obrad plenarnych) koncentrowała się na prostudenckich aspektach programu (nagranie prezentacji można znaleźć na stronie internetowej studiów [1]).

Program został objęty patronatem ówczesnego Ministerstwa Cyfryzacji. Na stronie internetowej Rządu RP znalazły się m.in. następujące informacje[1]:

„Ministerstwo Cyfryzacji objęło patronat nad nowym kierunkiem studiów Politechniki Warszawskiej. Naszym zdaniem ich program może być uznawany przez inne uczelnie za modelowy. Mamy nadzieję, że zachęci je to do uruchamiania bliźniaczych kierunków."

„To jest pewne: studenci cyberbezpieczeństwa – dzięki temu, czego nauczą się w trakcie studiów – staną się pożądanymi przez pracodawców kandydatami do pracy w m.in. administracji publicznej, instytucjach sektora finansowego, instytucjach zajmujących się różnymi aspektami bezpieczeństwa państwa, czy w firmach komercyjnych."

Program wzbudził i nadal wzbudza ogromne zainteresowanie kandydatów na studia. W pierwszej rekrutacji, przeprowadzonej w 2019 roku, chęć studiowania *cyberbezpieczeństwa* zadeklarowało ponad 1800 osób, co przy limicie rekrutacyjnym 60 oznaczało, że o jedno miejsce ubiegało się ponad 30 kandydatów (drugi najpopularniejszy kierunek na PW). W rekrutacji przeprowadzonej w 2021 roku o jedno miejsce ubiegało się 22 kandydatów, co oznaczało, że *cyberbezpieczeństwo* stało się najpopularniejszym kierunkiem studiów na PW.

Innowacyjny i prostudencki charakter studiów na kierunku *cyberbezpieczeństwo* potwierdzają m.in. wyniki ankiety przeprowadzonej w styczniu 2020 roku wśród

[1] https://www.gov.pl/web/cyfryzacja/bezpieczenstwo-przede-wszystkim-zostan-studentem-pionierskiego-kierunku- [dostęp: 2.05.2022]

studentów pierwszego semestru tych studiów. Spośród 52 studentów, którzy wypełnili ankietę dotyczącą m.in. oceny programu studiów (stopa zwrotu ponad 90%), 37 studentów (71%) oceniło, że studia na kierunku *cyberbezpieczeństwo* wyróżniają się pozytywnie wśród innych studiów prowadzonych na Wydziale[2] (pozostałe odpowiedzi: 1 – wyróżniają się negatywnie, 2 – nie wyróżniają się w istotny sposób; 11 – nie mam dostatecznej wiedzy, aby dokonać oceny, 1 – brak odpowiedzi). Ponadto, 87% studentów stwierdziło, że pierwszy semestr utwierdził ich w słuszności wyboru kierunku studiów, udzielając odpowiedzi „tak" lub „raczej tak" na pytanie dotyczące tej kwestii.

Nowy program studiów wpisuje się w strategię rozwoju Uczelni, związaną m.in. z uzyskaniem statusu uczelni badawczej. *Cyberbezpieczeństwo i analiza danych* stanowi jeden z 7 wyodrębnionych przez PW Priorytetowych Obszarów Badawczych[3]. Pozyskane w ostatnim okresie przez pracowników Wydziału duże projekty badawcze związane z cyberbezpieczeństwem stanowią gwarancję odpowiedniej tematyki i jakości projektów prac dyplomowych realizowanych przez studentów.

W lutym 2021 roku zespół twórców programu otrzymał prestiżową nagrodę Ministra Edukacji i Nauki za znaczące osiągnięcia w zakresie działalności dydaktycznej. W tym kontekście warto odnotować, że:

- wniosek o nagrodę został wsparty m.in. opiniami Przewodniczącego Parlamentu Studentów RP, przewodniczącego Komisji Kształcenia Rady Głównej Nauki i Szkolnictwa Wyższego oraz Dyrektora Departamentu Cyberbezpieczeństwa w ówczesnym Ministerstwie Cyfryzacji,
- wśród nagrodzonych znalazło się dwoje studentów, co stanowi ewenement, biorąc pod uwagę, że nagrody ministra za osiągnięcia dydaktyczne przeznaczone są z zasady dla nauczycieli akademickich.

3. STUDIA DRUGIEGO STOPNIA

W lutym 2023 roku spodziewani są pierwsi absolwenci opisanych wyżej studiów pierwszego stopnia. Większość z nich pragnie kontynuować kształcenie na studiach drugiego stopnia, pogłębiając wiedzę i umiejętności związane z cyberbezpieczeństwem. Dla nich, a także dla absolwentów innych kierunków, prowadzonych na PW i nie tylko na PW, stworzono program studiów drugiego stopnia na kierunku *cyberbezpieczeństwo*, który po zakończeniu formalnej procedury akceptacji przez organy Uczelni (jest ona obecnie w końcowym stadium) zostanie uruchomiony w semestrze letnim roku akademickiego 2022/2023.

Znamienna dla nowego programu jest jego elastyczność, przejawiająca się w możliwości realizacji przez studenta zindywidualizowanego programu, łączącego w różnym stopniu dwie uzupełniające się wizje ścieżki kształcenia: naukową i przedsiębiorczą. Ścieżka naukowa jest ściśle ukierunkowana na perspektywę kontynuowania kształcenia w szkole doktorskiej z silnym naciskiem na aktywne zaangażowanie studentów w badania naukowe. Ścieżka przedsiębiorcza jest zorientowana na praktyczne zastosowania cyberbezpieczeństwa i skoordynowana z realizacją projektów badawczych mających zastosowanie w otoczeniu społeczno-gospodarczym.

Taka koncepcja kształcenia wychodzi naprzeciw strategicznym interesom państwa, oczekiwaniom społecznym oraz trudnym do zaspokojenia potrzebom rynku pracy. Jest ona także zgodna z misją i wizją rozwoju Uczelni, zawartą w nowej „Strategii rozwoju Politechniki Warszawskiej do roku 2030", przyjętej przez Senat PW w 2021 roku. Podobnie jak w przypadku studiów pierwszego stopnia jest ona także zgodna z wizją rozwoju PW realizowaną w ramach programu „Inicjatywa Doskonałości – Uczelnia Badawcza" (ID-UB). Ponadto, można oczekiwać, że uruchomienie studiów drugiego stopnia na kierunku *cyberbezpieczeństwo* przyniesie Uczelni konkretne korzyści:

- będzie miało bezpośredni wpływ na cyberbezpieczeństwo PW poprzez angażowanie studentów w praktyczne zadania związane z zapewnieniem bezpieczeństwa teleinformatycznego Uczelni,
- będzie kuźnią kadr naukowych PW w zakresie cyberbezpieczeństwa.

Jako jedno z planowanych w ramach realizacji projektu ID-UB działań, służących podnoszeniu jakości kształcenia, wskazano reorientację celów i metod kształcenia, mającą na celu stymulowanie innowacyjnych i przedsiębiorczych postaw studentów oraz przygotowanie ich do zespołowego rozwiązywania problemów interdyscyplinarnych i prowadzenia badań.

W przypadku studiów drugiego stopnia na kierunku *cyberbezpieczeństwo* elementami tej reorientacji jest m.in. koncepcja realizacji programu zakładająca odejście od kształcenia opartego na biernym uczestnictwie w zajęciach (narzucającym pozyskiwanie wiedzy teoretycznej i pasywne jej odtwarzanie na sprawdzianach) na rzecz stosowania metod opartych na rozwiązywaniu problemów i realizacji projektów powiązanych z badaniami, zwłaszcza realizowanymi wspólnie z instytucjami z otoczenia społeczno-gospodarczego oraz innych formach prowadzenia zajęć aktywizujących studentów.

Takie podejście przejawia się m.in. wprowadzeniem do programu dwóch dużych modułów PBL (Project Based Learning) o wymiarze 8 i 12 ECTS (dotyczących analizy danych w kontekście cyberbezpieczeństwa oraz zapewnieniu bezpiecznego funkcjonowania systemów internetu rzeczy), przy czym w przypadku pierwszego z tych modułów jego dopełnieniem (do 12 ECTS) są dwa przedmioty służące wzbogaceniu wiedzy studenta o elementy niezbędne do realizacji projektu o zaawansowanym charakterze.

[2] Ocenę tę należy rozpatrywać w kontekście faktu, że studia na 3 z 5 pozostałych kierunków znajdujących się w ofercie Wydziału w rankingu kierunków technicznych miesięcznika Perspektywy (2019) zajęły 1. miejsce.

[3] Jest to jedyny przypadek wśród polskich uczelni, kiedy cyberbezpieczeństwo występuje w nazwie POB.

Inne, warte odnotowania cechy programu to:

- specjalny przedmiot poświęcony metodologicznym podstawom prowadzenia badań naukowych,
- interdyscyplinarny charakter – wiele elementów programu dotyczących pozatechnicznych aspektów zapewniania bezpieczeństwa systemów teleinformatycznych,
- nowa koncepcja seminarium dyplomowego (z wymaganiem napisania artykułu naukowego),
- ograniczenie liczby przedmiotów w poszczególnych semestrach i w całym cyklu studiów,
- mała liczba egzaminów (łącznie 3 w przedmiotach obowiązkowych programu),
- rozwijanie umiejętności wyszukiwania informacji i samodzielnej krytycznej analizy materiałów źródłowych w niemal wszystkich przedmiotach,
- praca zespołowa w wielu przedmiotach,
- sposób podejścia do osiągania poziomu B2+ znajomości języka (3 przedmioty obowiązkowe prowadzone w języku angielskim).

Realizacja opisanej koncepcji kształcenia jest możliwa m.in. dzięki licznym prowadzonym na Wydziale projektom badawczym związanym z cyberbezpieczeństwem, stanowiącym podstawę do formułowania tematów projektów realizowanych w ramach poszczególnych przedmiotów, a także tematów prac dyplomowych.

4. INNE INICJATYWY

Mimo że główne działania edukacyjne PW w obszarze cyberbezpieczeństwa koncentrują się na Wydziale Elektroniki i Technik Informacyjnych, także studenci innych jednostek mają możliwość rozwijania kompetencji w tym obszarze:

- Wydział Elektryczny prowadzi specjalność *cyberbezpieczeństwo* na studiach drugiego stopnia na kierunku *informatyka stosowana*,
- Wydział Matematyki i Nauk Informacyjnych prowadzi specjalność *matematyka w cyberbezpieczeństwie* na studiach drugiego stopnia na kierunku *matematyka*,
- Wydział Zarządzania prowadzi studia pierwszego i drugiego stopnia na kierunku *zarządzanie bezpieczeństwem infrastruktury krytycznej*,
- Wydział Administracji i Nauk Społecznych włącza do swojej oferty przedmiotów humanistycznych i społecznych, przeznaczonych dla wszystkich studentów PW, coraz więcej zajęć dotyczących pozatechnicznych aspektów cyberbezpieczeństwa.

Oprócz rozwijania kompetencji w zakresie cyberbezpieczeństwa poprzez studia, studenci szczególnie zainteresowani tą dziedziną mogą realizować swoją pasję, zdobywać dodatkowe doświadczenie i rozwijać innowacyjne pomysły w ramach formalnie utworzonych, jako organizacje studenckie, kół naukowych. Obecnie dwie takie grupy, czyli Studenckie Koło Naukowe „Bezpieczeństwo informacji" oraz Studenckie Koło Naukowe „Cyberbezpieczeństwo", kierowane przez pracowników Wydziału Elektroniki i Technik Informacyjnych, zajmują się różnymi aspektami cyberbezpieczeństwa. Ta forma zajęć pozaprogramowych jest dostępna dla wszystkich studentów PW, niezależnie od kierunku studiów.

W odpowiedzi na rosnące zainteresowanie bezpieczeństwem systemów teleinformatycznych w ramach oferty edukacyjnej PW przeznaczonej dla osób uczących się „przez całe życie" prowadzone są od dawna studia podyplomowe obejmujące zagadnienia z tego zakresu. Obecnie dostępne są następujące programy tego typu, bezpośrednio odnoszące się do cyberbezpieczeństwa:

- *Ochrona informacji w sieciach i systemach teleinformatycznych: projektowanie i audyt zabezpieczeń*, na Wydziale Elektroniki i Technik Informacyjnych,
- *Bezpieczeństwo systemów informatycznych* na Wydziale Elektrycznym,
- *MBA Cybersecurity Management* w Szkole Biznesu PW (program zostanie uruchomiony w semestrze zimowym roku akademickiego 2022/2023).

5. PODSUMOWANIE

Przedstawione działania związane z kształtowaniem oferty edukacyjnej PW w zakresie cyberbezpieczeństwa, zgodne z misją i wizją rozwoju Uczelni, realizowaną m.in. w ramach programu „Inicjatywa Doskonałości – Uczelnia Badawcza", wychodzą naprzeciw strategicznym interesom państwa oraz oczekiwaniom społecznym, w tym coraz trudniejszym do zaspokojenia potrzebom rynku pracy. Są jednocześnie polem eksperymentów związanych z wprowadzeniem do procesu kształcenia istotnych innowacji, wysoko ocenianych przez kluczową grupę interesariuszy – studentów.

LITERATURA

[1] *Cyberbezpieczeństwo: Studia I stopnia na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej*. https://cyber.elka.pw.edu.pl [dostęp: 6.06.2022].

[2] Kraśniewski Andrzej. 2021. *Cybersecurity Research, Education and Management: University Perspective*. Oficyna Wydawnicza PW. ISBN 978-83-8156-317-8.

[3] Lubicz-Nawrocka Tanya. 2017. "Co-Creation of the Curriculum: Challenging the Status Quo to Embed Partnership", *Journal of Educational Innovation, Partnership and Change* 2(2). doi: 10.21100/jeipc.v3i2.529

[4] Newhouse William, Stephanie Keith, Benjamin Scribner, Gregory Witte. 2017. *National Initiative for Cybersecurity Education (NICE): Cybersecurity Workforce Framework*. NIST Special Publication 800-181, National Institute of Standards and Technology, USA. doi: 10.6028/NIST.SP.800-181.

WSPARCIE PROCESU ZARZĄDZANIA PODATNOŚCIAMI PRZY UŻYCIU KONWERSJI OCENY BAZOWEJ CVSS 2.0 DO 3.X

SUPPORT FOR THE VULNERABILITY MANAGEMENT PROCESS USING CONVERSION CVSS BASE SCORE 2.0 TO 3.X

Maciej Nowak[1]; Michał Walkowski[2]; Sławomir Sujecki[3]

[1] Politechnika Wrocławska, Wrocław, maciej.nowak@pwr.edu.pl
[2] Politechnika Wrocławska, Wrocław, michal.walkowski@pwr.edu.pl
[3] Politechnika Wrocławska, Wrocław, slawomir.sujecki@pwr.edu.pl

DOI: 10.15199/59.2022.4.32

Streszczenie: **Przedstawiamy koncepcję oraz analizę użycia algorytmów uczenia maszynowego do konwersji oceny bazowej standardu Common Vulnerability Scoring System (CVSS) z wersji 2.0 do 3.x. CVSS jest międzynarodowym branżowym standardem opisującym luki w oprogramowaniu oraz podającym mierzalne wskaźniki zagrożenia. Nie wszystkie publicznie znane podatności posiadają ocenę krytyczności w CVSS 3.x, co wymusza na organizacjach jednoczesne używanie dwóch wersji standardu. To stanowi problem w jednoznacznej priorytetyzacji podatności.**

Abstract: **We present the concept and analysis of the use of machine learning algorithms to convert the base score of the Common Vulnerability Scoring System (CVSS) from version 2.0 to 3.x. CVSS is an international industry standard that describes software vulnerabilities and provides measurable risk indicators. Not all publicly known vulnerabilities have a criticality rating in CVSS 3.x, which forces organizations to use two versions of the standard simultaneously. This causes a problem in unambiguous prioritization of vulnerabilities.**

Słowa kluczowe: **bezpieczeństwo systemów IT, standard CVSS, uczenie maszynowe.**

Keywords: **CVSS standard, machine learning, security of IT systems..**

1. WSTĘP

Proces zarządzania podatnościami stanowi ważny element w działalności każdej renomowanej firmy związanej z branżą teleinformatyczną. Tematyka priorytetyzacji zagrożeń od dawna poruszana jest również w literaturze naukowej [4, 13, 17]. Jednakowoż, każda organizacja podchodzi do problemu inaczej, co stanowi jej wewnętrzną politykę bezpieczeństwa [10]. Komercyjne systemy zarządzania podatnościami są bardzo drogie i nie informują użytkowników o szczegółach procedury priorytetyzacji podatności. Dodatkowo, w celu mierzenia stanu bezpieczeństwa infrastruktury teleinformatycznej, organizacje wykorzystują metryki zaproponowane w 2010 roku przez Center for Internet Security (CIS) [2]. Aby spełnić wymagania CIS, najczęściej wybierany jest standard oparty na ocenie bazowej CVSS 2.0, natomiast model oceny CVSS 3.x. jest pomijany. Common Vulnerability Scoring System (CVSS) jest standardem opisującym charakterystykę podatności oprogramowania [1] i oceniającym powagę zagrożenia w skali od 0 do 10. Standard CVSS 2.0 wprowadzono w 2007 roku [6], natomiast kolejne wersje - CVSS 3.0 [7] w 2015 – problematyczny ze względu na możliwość różnej interpretacji [12] oraz CVSS 3.1 [8] w 2019. Ze względu na dużą liczbę luk, nie wszystkie podatności zostały przekonwertowane z CVSS 2.0 do 3.x, pomimo że CVSS 3.x lepiej ocenia istotę podatności i skuteczniej szacuje zagrożenia [10]. Niezależnie od wersji, na ocenę podatności wpływ mają trzy składowe - bazowa (BS), czasowa (TS) oraz środowiskowa (ES), przy czym dwie ostatnie są opcjonalne i stosowane w przypadku bardzo wnikliwej analizy. W naszych rozważaniach zajmujemy się oceną bazową (BS) dla dwóch niekompatybilnych wersji 2.0 oraz 3.x.

Najważniejsze różnice pomiędzy standardami zestawiliśmy w tabeli 1. Pomimo zmian w nomenklaturze obu standardów, parametry zapisane w poszczególnych wierszach tabeli są funkcjonalnie zgodne. Wektor bazowy CVSS 2.0 opisany jest przez 6 składowych, które mogą przyjmować jedną z trzech wartości, natomiast w CVSS 3.x potrzebna jest znajomość aż 8 parametrów – dodatkowo S i UI. Kolejną różnicą jest liczba wartości metrycznych jakie mogą przyjmować parametry AV oraz AC wektora bazowego CVSS 3.x – odpowiednio 4 i 2.

Tab. 1. Różnice w budowie wektorów bazowych CVSS 2.0 oraz CVSS 3.x

Standard CVSS 2.0	*Standard CVSS 3.x*
Wektor dostępu (AV)	Wektor ataku (AV)
Złożoność dostępu (AC)	Złożoność ataku (AC)
Wymagane poświadczenie (Au)	Wymagane uprawnienia (PR)
	Interakcja z użytkownikiem (UI)
Wpływ na poufność (C)	Wpływ na poufność (C)
Wpływ na integralność (I)	Wpływ na integralność (I)
Wpływ na dostępność (A)	Wpływ na dostępność (A)
	Zakres (S)

Występuje również rozbieżność pomiędzy kategoryzacją podatności (ang. Qualitative Severity Rating Scale, QSRS), którą przedstawiamy w tabeli 2.

Tab. 2. Różnice w kategoriach podatności dla CVSS 2.0 oraz CVSS 3.x

Wartość BS	*Standard CVSS 2.0*	*Standard CVSS 3.x*
0,0	Niska	Informacyjna
0,1-3,9		Niska
4,0-6,9	Średnia	Średnia
7,0-8,9	Wysoka	Wysoka
9,0-10,0		Krytyczna

Pomimo niejednoznacznej oceny podatności, oba standardy CVSS 2.0 oraz CVSS 3.x są jednocześnie używane przez operatorów sieci teleinformatycznych [10].

2. Koncepcja badawcza

W tej sekcji opisujemy koncepcję realizacji konwersji wartości bazowych standardu CVSS 2.0 do 3.x. Strategia obejmuje 4 kroki – akwizycję danych, procedurę doboru zbioru uczącego, zestawienie algorytmów uczących oraz procedurę ich selekcji - testowanie. Ostateczną wartość BS CVSS 3.x uzyskuje się poprzez powiązanie wyznaczonych w konwersji 8 składowych wektora CVSS 3.x wzorami opisanymi w specyfikacjach [8, 12].

2.1. Akwizycja danych

Pierwszy etap to pozyskanie danych o znanych słabościach i podatnościach przy użyciu otwartego oprogramowania Vulnerability Management Center (VMC) [9, 11]. Baza dostarcza informacje o wszystkich znanych 73179 podatnościach posiadających dwie oceny BS CVSS 2.0 i 3.x. (to stanowi około 49% wszystkich zarejestrowanych podatności, pozostałe nie zawierają oceny CVSS 3.x). Następnie przy użyciu skryptu zaimplementowanego w języku programowania Python następuje tworzenie wektorów wierszowych BS 2.0 i 3.x zawierających parametry z tabeli 1 oraz przypisaną wartość BS. Wartości parametrów, z wyłączeniem BS, zostają przekonwertowane na indeksy klas, służące w kolejnych krokach do statystycznej klasyfikacji. Wstępne badania pokazały błędne mapowanie oryginalnych wektorów bazowych CVSS 2.0 na 3.x, dlatego w dalszym etapie oryginalny wektor CVSS 2.0 rozszerzono o dodatkowe 50 oraz 100 elementów. Dodatkowe elementy pozyskiwane są z opisów podatności i stanowią ilość wystąpień słów kluczowych. Przetwarzanie tekstu zrealizowano przy użyciu biblioteki Natural Language Toolkit (NLTK). Skrypty oraz dane wejściowe zamieściliśmy w ogólnodostępnym repozytorium [15].

2.2. Procedura tworzenia zbioru uczącego

Pozyskane w poprzednim etapie dane są niezbalansowane, dlatego wykorzystano autorską metodę podpróbkowania opisaną w [16]. Z rozszerzonych wektorów CVSS 2.0 tworzony jest wspólny zbiór uczący, przy założeniu, że dla każdej klasy parametru CVSS 3.x pozyskujemy jedynie 80 reprezentacyjnych wektorów opisujących tę klasę (najbardziej skorelowane z medianą). W zależności od kombinacji długości wektorów użytych do uczenia wszystkich klasyfikatorów – 57 lub/i 107, w praktyce liczba elementów wspólnego zbioru uczącego zawiera się w przedziale od 1736 do 1968.

2.3. Algorytmy uczenia maszynowego

Do badań wybrano 5 algorytmów uczenia maszynowego – k-najbliższych sąsiadów (kNN) z metryką euklidesową (E) i podobieństwem cosinusowym (C), naiwny klasyfikator Bayesa (NB), probabilistyczną sieć neuronową (PNN), miękkie niezależne modelowanie analogii klas (SIMCA) oraz maszynę wektorów podpierających z funkcją jądrową (KSVM) w dwóch konfiguracjach - z liniową funkcją jądra (LIN) oraz z gaussowską radialną funkcją bazową (GRBF) [14]. KSVM GRBF w pętli uczenia wykorzystuje 2-krotną walidację krzyżową, współczynnik błędnych klasyfikacji (ang. Misclassification Ratio, MCR), algorytm Nelder-Mead z zestawem 50 punktów początkowych i kNN z metryką cosinusową (C) do końcowej klasyfikacji nierozpoznanych próbek. Dodatkowo w przypadku KSVM TRGBF zastosowano podejście One-vs-One (OvO) dla parametrów o co najmniej 3 klasach.

2.4. Procedura selekcji algorytmów - testowanie

Parametry wszystkich algorytmów uczenia maszynowego optymalizowano dla klasyfikacji poszczególnych składowych wektora bazowego CVSS 3.x. Uwzględniono różne długości wektorów uczących oraz liczbę komponentów dla wstępnego przetwarzania realizowanego za pomocą analizy głównych składowych (PCA). Procedura zgrubnej selekcji obejmowała wyznaczenie dokładności klasyfikacji algorytmów przy walidacji zbiorem 100 losowo wybranych wektorów (niezależne od zbioru uczącego obserwacje). Zadanie powtarzano 10-cio krotnie, a do dalszej części badań kwalifikowano maksymalnie 3 algorytmy. W następnym kroku oceniano jakość modeli, przeprowadzając test obejmujący 5-cio krotną walidację krzyżową zbioru uczącego opisanego w podrozdziale 2.2. Na podstawie uśrednionych wartości MCR oraz czułości, precyzji i średniej harmonicznej (F1-score), dopasowano najlepsze modele klasyfikacji. Potwierdzeniem skuteczności oraz powtarzalności konwersji wektorów CVSS 2.0 na klasy parametrów wektora CVSS 3.x było trzyetapowe testowanie na zbiorach o różnej liczności - 1000, 10000 oraz 71000 elementów, z powtórzeniem zadania, odpowiednio 50, 50 oraz 10 razy.

3. WYNIKI

Wykonując procedurę opisaną w podrozdziale 2.4, w tabeli 3 zestawiono finalne przypisanie algorytmów uczenia maszynowego do klasyfikacji poszczególnych składowych wektora bazowego CVSS 3.x. Dodatkowo umieszczono informacje przy jakich długościach wektorów uczących oraz liczbie komponentów PC uzyskano maksymalizację dokładności i średniej harmonicznej oraz wysoką powtarzalność. Zbiór uczący liczył 1968 wektorów.

Wyniki średniej dokładności klasyfikacji, z uwzględnieniem liczności zbioru testującego (1000, 10000, 71000 wektorów) oraz krotności wykonywanych prób testowych, przedstawiono w tabeli 4. Pogrubioną czcionką wyróżniono najwyższe wartości średnich. Można zauważyć, że średnie wartości w obrębie tej samej składowej wektora CVSS 3.x różnią się od siebie o mniej niż 0,75%, natomiast dla mniej licznych zbiorów testujących mamy większe odchylenie standardowe (poza

Tab. 3. Przypisanie algorytmów uczenia maszynowego do klasyfikacji składowych wektora bazowego CVSS 3.x z podaniem liczby komponentów PC oraz długości wektorów uczących

Składowe wektora bazowego CVSS 3.x	*Algorytm*	*Liczba komponentów PC*	*Długość wektorów Uczących*
AV	KSVM GRBF	35	107
AC	PNN	25	107
PR	KSVM GRBF	24	57
S	NB	38	107
UI	PNN	41	107
C	KSVM GRBF	43	57
I	KSVM GRBF	32	107
A	PNN	48	107

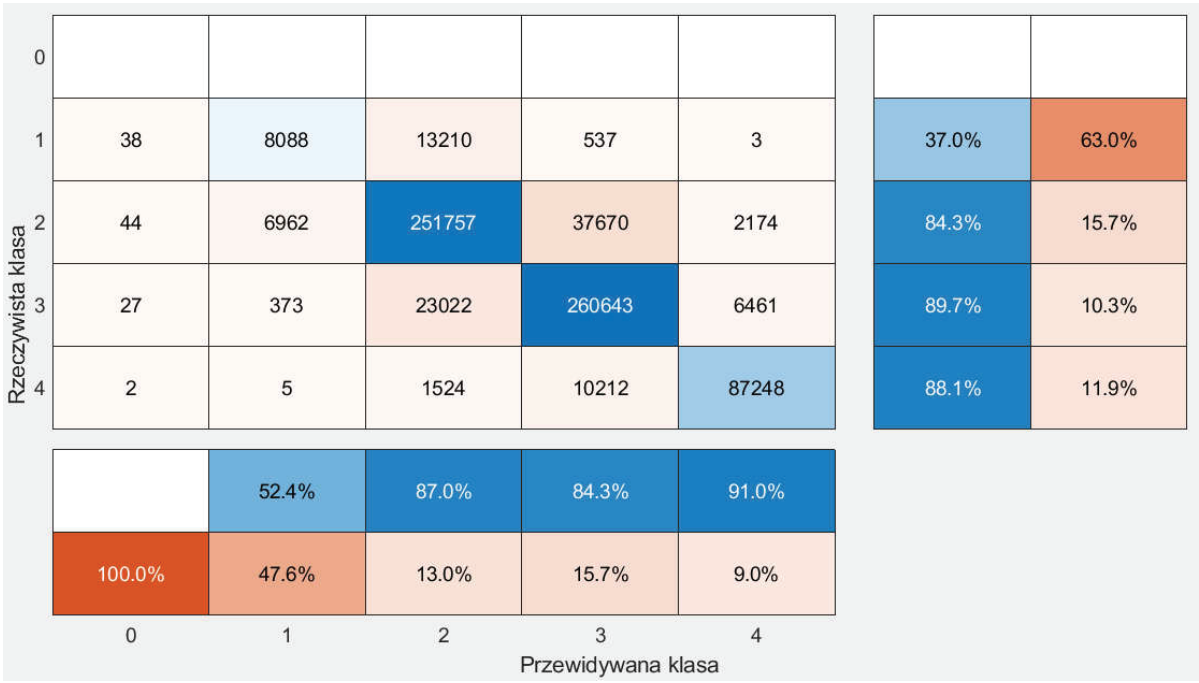
Tab. 4. Średnia dokładność klasyfikacji składowych wektora bazowego CVSS 3.x wyznaczona na podstawie testowania wektorami liczności 1000, 10000 oraz 71000 oraz krotności testowania odpowiednio 50, 50, 10 razy

Składowe wektora bazowego CVSS 3.x	*Średnia dokładność wyznaczona dla zbioru testującego liczności*		
	1000 wektorów	*10000 wektorów*	*71000 Wektorów*
AV	**90,11**±1,17	**90,11**±0,42	89,73±0,66
AC	**94,40**±0,75	94,22±0,19	94,21±0,06
PR	90,53±1,74	90,39±1,60	**90,89**±1,49
S	94,19±0,75	94,18±0,24	**94,84**±0,05
UI	93,90±0,81	**94,00**±0,23	93,95±0,07
C	90,93±0,87	90,88±0,99	**91,62**±0,29
I	92,55±0,88	**92,65**±0,25	92,53±0,11
A	94,74±0,75	94,87±0,22	**94,90**±0,05

parametrami AV i C z licznością zbioru testującego 1000). Dodatkowo, niezależnie od parametru oraz liczności zbioru testującego, mediana dokładności klasyfikacji zawsze jest większa od 90%. Uzyskane wyniki są bardzo powtarzalne, z wyjątkiem parametrów AV i PR (największy rozrzut wartości), gdzie w nielicznych przypadkach mogą pojawić się wartości odstające.

Wartość oceny bazowej CVSS 3.x uzyskiwana jest przez powiązanie parametrów wektora CVSS 3.x wzorami matematycznym opisanymi w standardzie. To obniża końcową dokładność konwersji do przedziału wartości (62,4 – 62,8)%. W przybliżeniu inżynierskim, wartość BS CVSS 3.x jest iloczynem dokładności klasyfikacji poszczególnych parametrów plus 5%. Kluczowym parametrem jest zakres (S), który determinuje użycie jednego z dwóch sposobów obliczania wartości bazowej. Należy zwrócić uwagę, że liczba kombinacji indeksów klas występująca w wektorze bazowym CVSS 3.x wynosi 2592, natomiast wartość tego wektora zawiera się w przedziale 0-10 z rozdzielczością 0,1 (101 wartości). Dodatkowo liczba kombinacji klas odpowiadająca poszczególnym wartościom oceny bazowej jest różna, a więc nie mamy do czynienia z rozkładem jednostajnym dyskretnym. Przykładowo, w przedziale oceny bazowej 0,1-1,5 oraz dla BS = 9,5 oraz BS = 9,7 nie istnieje żadna kombinacja, dla BS = 9,8 mamy jedną kombinację, natomiast dla BS =5,7 istnieje aż 79 różnych wektorów bazowych. BS = 0, może osiągnąć 96 różnych wektorów, w których spełniony jest warunek C = I = A = 0.

Na rysunku 1 przedstawiono macierz błędnych klasyfikacji kategorii podatności z oceny bazowej CVSS 2.0 do CVSS 3.x. uzyskane dla 71000 podatności



Rys. 1. Macierz błędów klasyfikacji kategorii podatności oceny bazowej CVSS 2.0 do 3.x

Konwersję wykonano 10 razy, dlatego macierz przedstawia skumulowane wartości (710000 obserwacji). Klasa 0 oznacza podatność o klasyfikacji informacyjnej; 1 - niskiej; 2 - średniej; 3 - wysokiej; oraz 4 - krytycznej. Na podstawie wartości macierzy błędów, widać że zastosowane rozwiązanie najlepiej radzi sobie z klasyfikacją oceny bazowej CVSS 3.x dla kategorii krytycznej (precyzja 91%, czułość 88,1%), a najgorzej z klasyfikacją dla kategorii informacyjnej (precyzja 0) oraz niskiej (precyzja 52,4%, czułość 37%). Kategoria informacyjna stanowi jedynie 0,016% obserwacji w zbiorze testującym. Żaden z wektorów należący do tej kategorii nie został sklasyfikowany poprawnie, jak również żaden wektor z klas 1-4 nie został błędnie przypisany do kategorii informacyjnej. Dla kategorii niskiej (2,17% obserwacji w zbiorze testującym), 45,13% obserwacji błędnie przypisano do kategorii średniej, 2,42% do wysokiej a tylko 0,03% do krytycznej. Oznacza to, że metoda zawyża kategorię podatności dla klas 0 i 1, co z punktu widzenia bezpieczeństwa systemów IT jest lepszym rozwiązaniem niż jej zaniżanie. W przypadku klas 2 i 3, widać że najwięcej błędnych klasyfikacji występuje w obrębie sąsiadujących kategorii. Dla klasy 2, 7,95% obserwacji błędnie przypisano do klasy 3, a w 4,56% przypadków zaniżono kategorię. W wypadku klasy 3, 12,19% obserwacji zaniżono o jedną kategorię, natomiast 3,3% przypisano kategorię krytyczną. Średnia dokładność klasyfikacji kategorii podatności dla przeprowadzonych testów wynosi 85,6%.

4. PODSUMOWANIE

Przedstawione wyniki badań wskazują na wysoką skuteczność stosowania algorytmów uczenia maszynowego do klasyfikacji parametrów wektora oceny bazowej CVSS 3.x. Mediana dokładności klasyfikacji dla najgorszego przypadku – parametru AV – wynosi 90%. Końcowa wartość BS CVSS 3.x wyliczana jest przy użyciu wzorów matematycznych wiążących 8 składowych wektora, co stanowi główny powód obniżenia dokładności konwersji do poziomu ~(62-63)%. Opracowana przez nas metoda pozwala na dobrą klasyfikację kategorii podatności, szczególnie dla kategorii krytycznej (91% precyzji). Dla obserwacji należących do dwóch najniższych kategorii (informacyjna oraz niska) metoda woli przypisać wyższą kategorię niż pozwolić sobie na pominięcie zagrożenia. Średnia dokładność klasyfikacji kategorii podatności wynosi 85,6%. Oznacza to, że z powodzeniem można zastosować opisane rozwiązanie do wsparcia procesu zarządzania podatnościami.

LITERATURA

[1] Booth, Harold, Rike Doug,. Witte Gregory A. 2013. „The national vulnerability database (nvd): Ove-rview". NVD.

[2] CIS, The CIS Security Metrics. 2010. Dostęp 25.04.2022.

[3] DSecure.me. 2021. VMC: Vulnerability Management Center. Dostęp 01.05.2022. https://github.com/DSecureMe/vmc

[4] Eschelbeck, Gerhard. 2005. „The Laws of Vulnerabili-ties: Which security vulnerabilities really matter?" In-formation Security Technical Report, 10(4) : 213-219.

[5] Fall Doudou, Kadobayashi Youki. 2019. „The common vulnerability scoring system vs. rock star vulnerabilities: Why the discrepancy?". ICISSP, 405-411.

[6] FIRST. 2017. A Complete Guide to the Common Vulnerability Scoring System Version 2.0. Dostęp 01.05.2022. https://www.first.org/cvss/v2/guide

[7] FIRST. 2017. Common Vulnerability Scoring System v3.0: Specification Document. Dostęp 01.05.2022. https://www.first.org/cvss/v3.0/specification-document

[8] FIRST. 2019. Common Vulnerability Scoring System v3.1: Speci-fication Document. Dostęp 01.05.2022. https://www.first.org/cvss/v3.1/specification-document

[9] F-Secure. 2021. Vulnerability Management Tool. Dostęp 28.04.2022.

[10] Gartner. 2019. A Guidance Framework for Developing and Implementing Vulnerability Management. Dostęp 01.05.2022. https://www.gartner.com/en/documents/3970669

[11] IBM IBM X-Force Threat Intelligence. Dostęp 28.05.2022.

[12] Klinedinst Dan J. 2015. „*CVSS and the Internet of Things*". Dostęp 01.05.2022. https://insights.sei.cmu.edu/blog/cvss-and-the-internet-of-things/

[13] Lai, Yeu-Pong, Hsia Po-Lun. 2007. „Using the vulnerability information of computer systems to improve the network security". Computer Communications, 30(9) : 2032–2047.

[14] Nowak Maciej Roman, Zdunek Rafał, Edward Pliński, Świątek Piotr, Strzelecka Magdalena, Malinka Wiesław, Plińska Stanisława. 2019. „Recognition of pharmacological bi-heterocyclic compounds by using terahertz time domain spectroscopy and chemometrics". Sensor. 19(15) : 3349.

[15] Nowak Maciej, Walkowski Michał, Sujecki Sławomir. 2021. „CVSS 2.0 extended vector database". Dostęp 01.05.2022. https://github.com/mwalkowski/cvss-2-extended-vector-database

[16] Nowak Maciej, Walkowski Michał, Sujecki Sławomir. 2021. „Machine learning algorithms for conver-sion of cvss base score from 2.0 to 3.x". Springer International Conference on Computational Scien-ce, 255-269.

[17] Rieke Roland. 2006. „Modelling and analysing network security policies in a given vulnerability setting". Springer International Workshop on Critical Information Infrastructures Security, 67-78.

ANALIZA PORÓWNAWCZA ŚRODOWISK EMULACYJNYCH KATHARÁ I NETKIT

COMPARATIVE ANALYSIS OF KATHARÁ AND NETKIT EMULATION ENVIRONMENTS

Krzysztof Kurczab[1]; Katarzyna Kosek-Szott[2]

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica, Kraków, krzysztofkur@student.agh.edu.pl

[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica, Kraków, katarzyna.kosek-szott@agh.edu.pl

DOI: 10.15199/59.2022.4.33

Streszczenie**: Artykuł zawiera analizę porównawczą dwóch środowisk, Netkit oraz Kathará, pozwalających na emulację sieci teleinformatycznych złożonych z routerów i stacji końcowych wyposażonych w system operacyjny Linux. W artykule przedstawiono główne zalety jakimi cechuje się każde z tych środowisk oraz różnice jakie ich dzielą. Analiza wyników z przeprowadzonych badań koncentruje się na szybkości działania obu środowisk, sposobie zarządzania demonami routingu oraz poprawności wpisów w tablicach routingu.**

Abstract**: The article presents a comparative analysis of two environments, Netkit and Kathará, which enable emulation of ICT networks composed of Linux routers and client stations. The article discusses the main advantages and disadvantages of each environment. The analysis of the gathered results concentrates on the operation speed, management of routing daemons, and the correctness of routing table entries.**

Słowa kluczowe**: Sieci teleinformatyczne, Kathará, Netkit, emulator, trasowanie**

Keywords**: ICT networks, Kathará, Netkit, emulator, routing**

1. WSTĘP

W dobie „wszystko jako usługa" (ang. *everything-as-a-service*) wirtualizacja urządzeń sieciowych odgrywa bardzo istotną rolę. Rozwiązanie to daje administratorom i inżynierom sieci m.in. możliwość testowania nowych funkcji przed wdrożeniem w sieci produkcyjnej, ulepszenia sposobów nawiązywania połączeń między urządzeniami, eksperymentowania z różnymi protokołami routingu, itd. Dodatkowo, wirtualizacja sieci przyczynia się do zmniejszenia kosztów utrzymania oraz zużycia energii.

W celu ułatwienia wirtualizacji urządzeń oraz wygodnego tworzenia i testowania złożonych topologii sieciowych powstały środowiska emulacyjne Netkit (oparte na wirtualizacji sprzętowej) oraz Kathará (oparte na kontenerach). Oba środowiska są szeroko stosowane m.in. w prowadzeniu zajęć dydaktycznych[1]. Umożliwiają one studentom wygodne i bezpłatne rozwijanie umiejętności technicznych. Oba środowiska mogą zostać także wykorzystane do prowadzenia badań związanych z działaniem różnych protokołów routingu. Dużą zaletą obu środowisk jest możliwość konfigurowania urządzeń wyposażonych w system operacyjny Linux oraz korzystanie z oprogramowania Quagga[2], które implementuje różne protokoły routingu (m.in., RIP, OSPF, BGP) i posiada wiersz poleceń podobny do wiersza poleceń urządzeń firmy CISCO.

Niniejszy artykuł przedstawia analizę porównawczą wyżej wymienionych środowisk emulacyjnych. W rozdziale 2 porównano wirtualizacę i konteneryzację. W rozdziale 3 i 4 przedstawiono odpowiednio środowisko Netkit oraz Kathará. Rozdział 5 przedstawia topologię użytą w badaniach. Rozdział 6 prezentuje wyniki analizy porównawczej emulatorów. Rozdział 7 zawiera najważniejsze wnioski.

2. WIRTUALIZACJA vs KONTENERYZACJA

Tradycyjny sposób wirtualizacji sprzętowej z nadzorcą (ang. *hypervisor, virtual machine monitor*) sprawia, że tworzone maszyny wirtualne są dużych rozmiarów, proces uruchamiania zajmuje dużo czasu, a wydajność pogarsza się podczas jednoczesnego uruchamiania wielu maszyn wirtualnych.

Wraz z postępującą ewolucją przetwarzania w chmurze, wprowadzono nową lekką (kontenerową) metodę wirtualizacji. Zapewnia ona lepszą wydajność, rozmiary obrazów maszyn wirtualnych są zdecydowanie mniejsze, a jednocześnie jest w stanie obsłużyć gęstsze sieci w porównaniu z wirtualizacją z nadzorcą. Konteneryzacja polega na wirtualizacji systemu operacyjnego (ang. *OS-level virtualization*) i tworzeniu oddzielnych instancji w przestrzeni użytkownika (ang. *user space*), tzw. kontenerów. Pozwala na znaczne przyspieszenie wykonywanych procesów.

Przykładem oprogramowania wspierającego konteneryzację jest Docker. Obrazy Dockera pozwalają na wykonywanie kodu wewnątrz przenośnego kontenera, ze wszystkimi niezbędnymi zależnościami (bibliotekami, plikami konfiguracyjnymi itd.) [3]. Pozwala to na umieszczenie aplikacji w izolowanym środowisku (kontenerze). Dodatkowo, kontenery Docker mogą być

[1] https://www.kathara.org/stories.html

[2] https://www.quagga.net/

zarządzane przez platformę Kubernetes[3], która pozwala na uruchamianie i zarządzanie rozproszonych klastrów urządzeń, złożonych nawet z tysięcy węzłów [4]. Kubernetes pozwala na rozmieszczenie kontenerów na maszynach w sposób, który zapewnia optymalizację dostarczonych zasobów.

3. ŚRODOWISKO NETKIT

Netkit[4] to środowisko emulacyjne dedykowane dla systemu operacyjnego Linux, którego celem jest dostarczenie możliwości emulacji złożonych scenariuszy sieciowych i zachowania urządzeń sieciowych (routerów i urządzeń końcowych) oraz konfiguracji tych urządzeń (z użyciem narzędzi dostępnych w systemie operacyjnym Linux) z wykorzystaniem wirtualizacji sprzętowej. Netkit to oprogramowanie typu open source, które pozwala na dogłębne poznanie i testowanie wykorzystywanych w produkcji protokołów i systemów (np. BGP, DNS). Każda konfiguracja może zostać zapisana jako zwykły plik ASCII, który można łatwo przenosić i współdzielić.

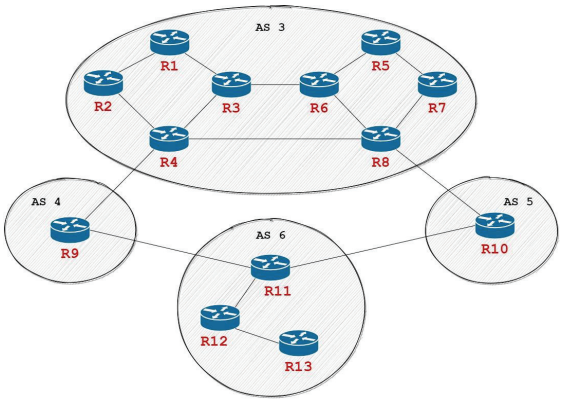
4. ŚRODOWISKO KATHARÁ

Oprogramowanie Kathará (dedykowane dla systemów Windows, Linux oraz Mac) jest następcą oprogramowania Netkit i zostało zaimplementowane z użyciem języka Python. Wspiera ono platformy Docker oraz Kubernetes [1]. Podobnie jak Netkit, jest to narzędzie do emulacji sieci teleinformatycznych. Jest to jednak rozwiązanie dużo lżejsze i zdecydowanie szybsze w działaniu [3]. Emulator Kathará, podobnie jak Netkit, umożliwia konfigurację tradycyjnych urządzeń sieciowych za pomocą plików konfiguracyjnych. Każde z urządzeń jest jednak emulowane za pomocą kontenera, który jest połączony z innymi urządzeniami za pośrednictwem wirtualnych sieci LAN warstwy drugiej modelu ISO/OSI. Dodatkowo, Kathará wspiera sieci sterowane programowo (ang., *software defined networks*).

5. TOPOLOGIA SIECIOWA

Rys. 1 przedstawia uproszczony diagram topologii sieciowej, która została użyta do przeprowadzenia analizy porównawczej środowisk Netkit 2.8 i Kathará 3.2.0. W badaniach zostały użyte złożone scenariusze sieciowe, uwzględniające trzy różne konfiguracje protokołu BGP: *Full Mesh*, *Route Reflectors* i *Confederations*. Pliki konfiguracyjne wszystkich scenariuszy są dostępne na platformie GitHub[5].

Wewnątrz systemów autonomicznych AS3 i AS6 został skonfigurowany jednoobszarowy protokół OSPF. Dodatkowo, pomiędzy poszczególnymi systemami autonomicznymi został skonfigurowany protokół eBGP. Konfiguracja *Full Mesh* zawierała dodatkowo podstawową konfigurację protokołu iBGP, w której wszystkie routery BGP systemów autonomicznych AS3 i AS6 tworzyły pełną siatkę połączeń iBGP. W przypadku konfiguracji *Route Reflectors*, jako reflektory tras zostały wybrane routery R2 i R7, znajdujące się w najbardziej złożonym systemie autonomicznym AS3. Natomiast w scenariuszu *Confederations*, system autonomiczny AS3 został podzielony na dwa podsystemy autonomiczne - routery R1-R4 tworzyły pierwszy podsystem autonomiczny, natomiast routery R5-R8 tworzyły drugi podsystem autonomiczny. W ten sposób wewnątrz systemu AS3 powstała konfederacja dwóch mniejszych podsystemów autonomicznych. Dodatkowo, każdy z routerów znajdujących się w AS3 posiadał interfejs loopback o adresie IP skojarzonym z numerem występującym w jego nazwie, np. R1 miał adres 1.1.1.1.



Rys. 1. Uproszczony diagram topologii sieciowej wykorzystanej do przeprowadzenia badań

6. WYNIKI BADAŃ

Niniejszy rozdział przedstawia porównanie środowisk Netkit i Kathará pod względem: szybkości działania, szybkości restartu demonów routingu oraz poprawności zawartości tablic routingu. Słupki błędów przedstawione na wykresach reprezentują odchylenia standardowe z sześciu pomiarów wykonanych dla każdego z badanych scenariuszy sieciowych.

6.1. Czas uruchomienia środowiska

Szybkie uruchomienie środowiska emulacyjnego (tj. uruchomienie wszystkich maszyn wirtualnych) jest szczególnie ważne, gdy scenariusze są złożone, zawierają zróżnicowane konfiguracje lub kiedy analizie poddawane są różne przypadki tego samego scenariusza, co wymaga wielokrotnego restartowania środowiska. Rys. 2 pokazuje, że czas uruchamiania środowiska Kathará jest około 3 razy krótszy niż w przypadku środowiska Netkit.

6.2. Czas nawiązania połączeń

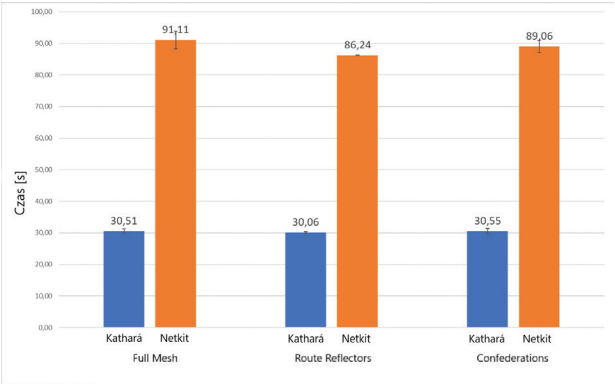
Rys. 3 przedstawia szybkość nawiązywania połączeń między urządzeniami, pozwalających na pomyślne przesłanie zapytania `ping` pomiędzy

[3] https://kubernetes.io
[4] https://www.netkit.org/
[5] Pliki konfiguracyjne użyte w trakcie badań: https://github.com/kkurczab/BGP-Route-Reflectors-and-Confederations-Kathara

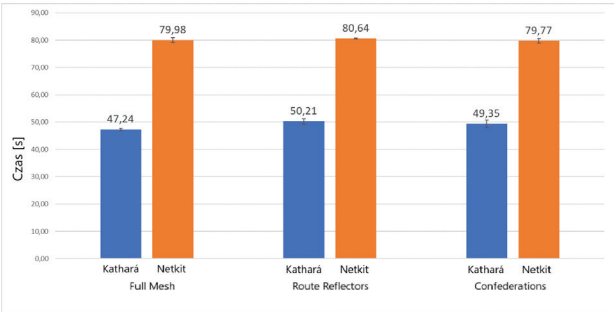
odległymi routerami: R13 i R2. W badanym przykładzie, ten czas jest około 1,6 razy krótszy dla środowiska Kathará niż dla środowiska Netkit.

6.3. Czas zamknięcia środowiska

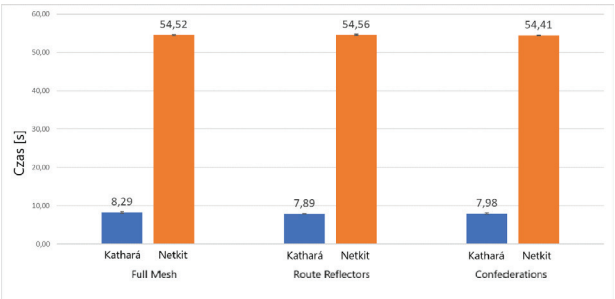
Czas potrzebny do wyłączenia wszystkich terminali wirtualnych w przypadku środowiska Kathará jest około 6,8 razy krótszy niż w przypadku środowiska Netkit, co prezentuje Rys. 4.



Rys. 2. Czas uruchamiania środowisk emulacyjnych Netkit i Kathará



Rys. 3. Czas potrzebny do nawiązania połączeń między odległymi routerami w środowisku Netkit i Kathará



Rys. 4. Czas potrzebny do zamknięcia środowisk Netkit i Kathará dla 3 różnych scenariuszy sieciowych

6.4. Czas restartu demonów Quagga

Następnie przetestowano szybkość ponownego uruchamiania demonów routingu oprogramowania Quagga na jednej z utworzonych maszyn wirtualnych. Czas restartu demonów zostały zmierzone za pomocą polecenia:

```
# time /etc/init.d/quagga restart.
```

Polecenie to pozwoliło na wyświetlenie podsumowania czasu rzeczywistego, czasu procesora użytkownika i czasu pracy systemu. Dodatkowo, w przypadku obu środowisk zmieniano priorytet uruchamiania demonów (im wyższy priorytet, tym ważniejszy jest proces dla jądra systemu) korzystając z polecenia:

```
#/etc/init.d/quagga [priorytet]
```

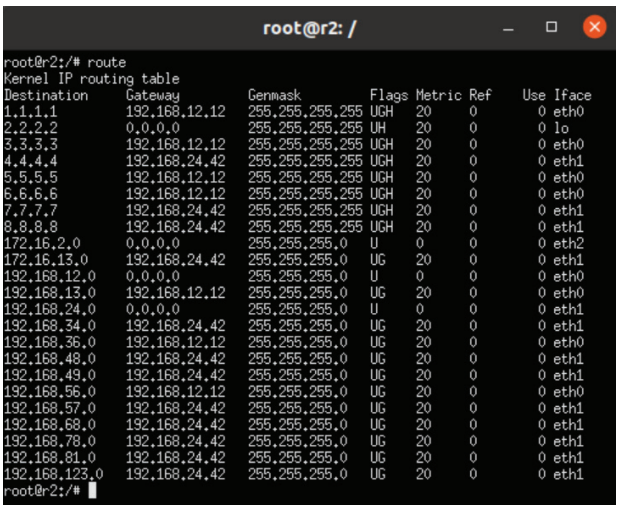
Tab. 1 przedstawia czas (w sekundach) potrzebny do ponownego uruchomienia demonów routingu w środowiskach Kathará i Netkit w zależności od wybranego priorytetu. W przypadku domyślnych, tj. niskich wartości priorytetu, działanie środowiska Netkit jest znacząco szybsze. Jednak po zmianie priorytetu na co najmniej 3, Kathará okazuje się być około 4 razy szybsza. W tym przypadku przewaga środowiska Kathará jest jednak ledwo zauważalna.

Tab. 1. Pomiar czasu restartu demonów routingu [s]

Priorytet	*Netkit*	*Kathará*
0	0,11	121,53
1	0,439	361,071
5	0,343	0,063
8	0,23	0,061
10	0,24	0,061

6.5. Tablice routingu w systemie Linux

Porównując tablicę routingu przykładowego urządzenia uruchomionego w środowiskach Kathará i Netkit, można zaobserwować istotną różnicę. W przypadku środowiska Kathará, wartość metryki jest niepoprawna. Jest ona taka sama dla wszystkich miejsc docelowych (równa 20), w przeciwieństwie do wartości wyświetlanych w środowisku Netkit (Rys. 6), które są zbieżne z informacjami przechowywanymi przez demona routingu (Rys. 8). Jest to więc funkcjonalność, która wymaga poprawy. Co ciekawe, zaobserwowane różnice w wyświetlanych wartościach metryk, nie mają wpływu na sposób działania urządzeń.



```
root@r2: /
root@r2:/# route
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
1.1.1.1         192.168.12.12   255.255.255.255 UGH   20     0        0 eth0
2.2.2.2         0.0.0.0         255.255.255.255 UH    20     0        0 lo
3.3.3.3         192.168.12.12   255.255.255.255 UGH   20     0        0 eth0
4.4.4.4         192.168.24.42   255.255.255.255 UGH   20     0        0 eth1
5.5.5.5         192.168.12.12   255.255.255.255 UGH   20     0        0 eth0
6.6.6.6         192.168.12.12   255.255.255.255 UGH   20     0        0 eth0
7.7.7.7         192.168.24.42   255.255.255.255 UGH   20     0        0 eth1
8.8.8.8         192.168.24.42   255.255.255.255 UGH   20     0        0 eth1
172.16.2.0      0.0.0.0         255.255.255.0   U     0      0        0 eth2
172.16.13.0     192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.12.0    0.0.0.0         255.255.255.0   U     0      0        0 eth0
192.168.13.0    192.168.12.12   255.255.255.0   UG    20     0        0 eth0
192.168.24.0    0.0.0.0         255.255.255.0   U     0      0        0 eth1
192.168.34.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.36.0    192.168.12.12   255.255.255.0   UG    20     0        0 eth0
192.168.48.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.49.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.56.0    192.168.12.12   255.255.255.0   UG    20     0        0 eth0
192.168.57.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.68.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.78.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.81.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.123.0   192.168.24.42   255.255.255.0   UG    20     0        0 eth1
root@r2:/#
```

Rys. 5. Przykładowa tablica routingu środowiska Kathará

6.6. Tablice demona routingu

W obu środowiskach, wartości metryk są wyświetlane poprawnie w tablicach routingu demona routingu Quagga, co zostało przedstawione na Rys. 7 i 8. Wyjaśnia to dlaczego, pomimo różnic występujących w tablicach routingu systemu Linux, zachowanie protokołów routingu w obu przypadkach jest takie samo.

```
r2
r2:~# route
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
7.7.7.7         192.168.24.42   255.255.255.255 UGH   40     0        0 eth1
5.5.5.5         192.168.12.12   255.255.255.255 UGH   50     0        0 eth0
6.6.6.6         192.168.12.12   255.255.255.255 UGH   40     0        0 eth0
1.1.1.1         192.168.12.12   255.255.255.255 UGH   20     0        0 eth0
3.3.3.3         192.168.12.12   255.255.255.255 UGH   30     0        0 eth0
8.8.8.8         192.168.24.42   255.255.255.255 UGH   30     0        0 eth1
4.4.4.4         192.168.24.42   255.255.255.255 UGH   20     0        0 eth1
192.168.101.0   192.168.24.42   255.255.255.0   UG    0      0        0 eth1
192.168.68.0    192.168.24.42   255.255.255.0   UG    30     0        0 eth1
192.168.36.0    192.168.12.12   255.255.255.0   UG    30     0        0 eth0
192.168.81.0    192.168.24.42   255.255.255.0   UG    30     0        0 eth1
192.168.34.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
172.16.2.0      *               255.255.255.0   U     0      0        0 eth2
192.168.48.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.49.0    192.168.24.42   255.255.255.0   UG    20     0        0 eth1
192.168.13.0    192.168.12.12   255.255.255.0   UG    20     0        0 eth0
192.168.78.0    192.168.24.42   255.255.255.0   UG    30     0        0 eth1
172.16.13.0     192.168.24.42   255.255.255.0   UG    0      0        0 eth1
192.168.12.0    *               255.255.255.0   U     0      0        0 eth0
192.168.91.0    192.168.24.42   255.255.255.0   UG    0      0        0 eth1
192.168.56.0    192.168.12.12   255.255.255.0   UG    40     0        0 eth0
192.168.57.0    192.168.24.42   255.255.255.0   UG    40     0        0 eth1
192.168.123.0   192.168.24.42   255.255.255.0   UG    0      0        0 eth1
192.168.24.0    *               255.255.255.0   U     0      0        0 eth1
r2:~#
```

Rys. 6. Przykładowa tablica routingu środowiska Netkit

```
root@r2: /
r2# sh ip route
Codes: K - kernel route, C - connected, S - static, R - RIP,
       O - OSPF, I - IS-IS, B - BGP, P - PIM, A - Babel,
       > - selected route, * - FIB route

O>* 1.1.1.1/32 [110/20] via 192.168.12.12, eth0, 00:00:48
C>* 2.2.2.0/24 is directly connected, lo
O>* 2.2.2.2/32 [110/10] is directly connected, lo, 00:01:30
O>* 3.3.3.3/32 [110/30] via 192.168.12.12, eth0, 00:00:43
  *                     via 192.168.24.42, eth1, 00:00:43
O>* 4.4.4.4/32 [110/20] via 192.168.24.42, eth1, 00:00:43
O>* 5.5.5.5/32 [110/50] via 192.168.12.12, eth0, 00:00:39
  *                     via 192.168.24.42, eth1, 00:00:39
O>* 6.6.6.6/32 [110/40] via 192.168.12.12, eth0, 00:00:39
  *                     via 192.168.24.42, eth1, 00:00:39
O>* 7.7.7.7/32 [110/40] via 192.168.24.42, eth1, 00:00:25
O>* 8.8.8.8/32 [110/30] via 192.168.24.42, eth1, 00:00:25
B   172.16.2.0/24 [200/0] via 4.4.4.4, 00:00:41
O   172.16.2.0/24 [110/10] is directly connected, eth2, 00:01:30
C>* 172.16.2.0/24 is directly connected, eth2
B>  172.16.13.0/24 [200/0] via 4.4.4.4 (recursive), 00:00:41
  *                       via 192.168.24.42, eth1, 00:00:41
O   192.168.12.0/24 [110/10] is directly connected, eth0, 00:01:30
C>* 192.168.12.0/24 is directly connected, eth0
O>* 192.168.13.0/24 [110/20] via 192.168.12.12, eth0, 00:00:48
B   192.168.24.0/24 [200/0] via 4.4.4.4, 00:00:41
O   192.168.24.0/24 [110/10] is directly connected, eth1, 00:00:47
C>* 192.168.24.0/24 is directly connected, eth1
O>* 192.168.34.0/24 [110/20] via 192.168.24.42, eth1, 00:00:43
O>* 192.168.36.0/24 [110/30] via 192.168.12.12, eth0, 00:00:43
  *                       via 192.168.24.42, eth1, 00:00:43
O>* 192.168.48.0/24 [110/20] via 192.168.24.42, eth1, 00:00:43
O>* 192.168.49.0/24 [110/20] via 192.168.24.42, eth1, 00:00:43
O>* 192.168.56.0/24 [110/40] via 192.168.12.12, eth0, 00:00:39
  *                       via 192.168.24.42, eth1, 00:00:39
O>* 192.168.57.0/24 [110/40] via 192.168.24.42, eth1, 00:00:25
O>* 192.168.68.0/24 [110/30] via 192.168.24.42, eth1, 00:00:25
O>* 192.168.78.0/24 [110/30] via 192.168.24.42, eth1, 00:00:25
O>* 192.168.81.0/24 [110/30] via 192.168.24.42, eth1, 00:00:25
B>  192.168.123.0/24 [200/0] via 4.4.4.4 (recursive), 00:00:41
  *                         via 192.168.24.42, eth1, 00:00:41
```

Rys. 7. Przykładowa tablica routingu demona Quagga w środowisku Kathará

7. WNIOSKI

Zaprezentowane w artykule porównanie środowisk emulacyjnych Kathará i Netkit wskazuje na przewagę pierwszego z nich. Dzięki zastosowanej konteneryzacji, środowisko Kathará charakteryzuje się szybkim działaniem oraz zmniejszonym wykorzystaniem zasobów sprzętowych w porównaniu ze środowiskiem Netkit. Przeprowadzone testy środowiska Kathará pozwoliły jednocześnie zaobserwować, że zawartość tablic routingu systemu operacyjnego Linux zwirtualizowanych urządzeń nie jest poprawna. Dzieje się tak, pomimo faktu, że zawartość tablic demonów routingu jest w pełni poprawna. Jest to wada, która zdecydowanie wymaga poprawy.

```
r2
r2# sh ip route
Codes: K - kernel route, C - connected, S - static, R - RIP, O - OSPF,
       I - ISIS, B - BGP, > - selected route, * - FIB route

O>* 1.1.1.1/32 [110/20] via 192.168.12.12, eth0, 00:15:32
C>* 2.2.2.0/24 is directly connected, lo
O>* 3.3.3.3/32 [110/30] via 192.168.12.12, eth0, 00:14:35
  *                     via 192.168.24.42, eth1, 00:14:35
O>* 4.4.4.4/32 [110/20] via 192.168.24.42, eth1, 00:14:35
O>* 5.5.5.5/32 [110/50] via 192.168.12.12, eth0, 00:14:17
  *                     via 192.168.24.42, eth1, 00:14:17
O>* 6.6.6.6/32 [110/40] via 192.168.12.12, eth0, 00:14:18
  *                     via 192.168.24.42, eth1, 00:14:18
O>* 7.7.7.7/32 [110/40] via 192.168.24.42, eth1, 00:13:56
O>* 8.8.8.8/32 [110/30] via 192.168.24.42, eth1, 00:13:56
B   172.16.2.0/24 [200/0] via 4.4.4.4, 00:13:32
O   172.16.2.0/24 [110/10] is directly connected, eth2, 00:15:36
C>* 172.16.2.0/24 is directly connected, eth2
B>* 172.16.13.0/24 [200/0] via 4.4.4.4 (recursive via 192.168.24.42, eth1), 00:13:32
O   192.168.12.0/24 [110/10] is directly connected, eth0, 00:15:36
C>* 192.168.12.0/24 is directly connected, eth0
O>* 192.168.13.0/24 [110/20] via 192.168.12.12, eth0, 00:15:32
B   192.168.24.0/24 [200/0] via 4.4.4.4, 00:13:32
O   192.168.24.0/24 [110/10] is directly connected, eth1, 00:15:36
C>* 192.168.24.0/24 is directly connected, eth1
O>* 192.168.34.0/24 [110/20] via 192.168.24.42, eth1, 00:14:35
O>* 192.168.36.0/24 [110/30] via 192.168.12.12, eth0, 00:14:35
  *                     via 192.168.24.42, eth1, 00:14:35
O>* 192.168.48.0/24 [110/20] via 192.168.24.42, eth1, 00:14:35
O>* 192.168.49.0/24 [110/20] via 192.168.24.42, eth1, 00:14:35
O>* 192.168.56.0/24 [110/40] via 192.168.12.12, eth0, 00:14:18
  *                     via 192.168.24.42, eth1, 00:14:18
O>* 192.168.57.0/24 [110/40] via 192.168.24.42, eth1, 00:13:56
O>* 192.168.68.0/24 [110/30] via 192.168.24.42, eth1, 00:13:56
O>* 192.168.78.0/24 [110/30] via 192.168.24.42, eth1, 00:13:56
O>* 192.168.81.0/24 [110/30] via 192.168.24.42, eth1, 00:13:56
B>* 192.168.123.0/24 [200/0] via 4.4.4.4 (recursive via 192.168.24.42, eth1), 00:13:
32
```

Rys. 8. Przykładowa tablica routingu demona Quagga w środowisku Netkit

PODZIĘKOWANIA

Praca została sfinansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji naukowej (nr 16.16.230.434) dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH w Krakowie.

LITERATURA

[1] Bonofiglio, Gaetano et al. 2018. “Kathará: A container-based framework for implementing network function virtualization and software defined networks.”. IEEE/IFIP Network Operations and Management Symposium: 1-3.

[2] Morabito, Roberto et al. 2015. “Hypervisors vs. lightweight virtualization: A performance comparison.”. In 2015 IEEE International Conference on Cloud Engineering: 386–393.

[3] Potdar, Amit M et al. 2020. “Performance evaluation of docker container and virtual machine.” Procedia Computer Science 171: 1419–1428.

[4] Scazzariello, Mariano, et al. "Megalos: A Scalable Architecture for the Virtualization of Large Network Scenarios." *Future Internet* 13.9 (2021): 227.

TECHNOLOGIA DYSTRYBUCJI KLUCZY KWANTOWYCH W SIECIACH 5G OPARTYCH NA ROZWIĄZANIACH OD WIELU DOSTAWCÓW

QUANTUM KEY DISTRIBUTION TECHNOLOGY IN MULTI-PROVIDER 5G NETWORKS

Jordi Mongay Batalla; Sławomir Sujecki; Tomasz Wichary

[1] Politechnika Warszawska, Warszawa, jordi.mongay.batalla@pw.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa i Politechnika Wrocławska, Wrocław, slawomir.sujecki@wat.edu.pl
[3] Politechnika Warszawska, Warszawa, tomasz.wichary.dokt@pw.edu.pl

DOI: 10.15199/59.2022.4.34

Streszczenie: **W niniejszym artykule rozważono zastosowanie technologii *Dystrybucji Kluczy Kwantowych* (ang. Quantum Key Distribution) do rozwiązywania problemów bezpiecznego przesyłania danych w heterogenicznych mobilnych sieciach z wieloma dostawcami. W sieciach mobilnych, powszechnie przyjętym rozwiązaniem do szyfrowania transmisji danych jest zastosowania kluczy asymetrycznych. W niedalekiej przyszłości takie rozwiązanie może stracić na znaczeniu z powodu szybkiego rozwoju technologii komputerów kwantowych. Podjęcie prac badawczo-rozwojowych nad wykorzystaniem nowych rozwiązań opartych na technologii kwantowej w sieciach mobilnych staje się koniecznością. W szczególności dotyczy to sieci 5G, 5G-Advanced, które będą użytkowane jeszcze co najmniej przez najbliższą dekadę. W artykule zostało przedstawione rozwiązanie przesyłania kluczy za pomocą kanałów kwantowych do bezpiecznej transmisji danych pomiędzy elementami sieci. Otrzymane wyniki potwierdzają zasadność nowego podejścia do bezwarunkowego bezpieczeństwa przesyłania danych przy wykorzystaniu obecnej infrastruktury sieci komórkowych.**

Słowa kluczowe: **Bezpieczeństwo sieci, dystrybucja kluczy kwantowych, 5G, sieć wielu dostawców.**

Keywords: **Network security, Quantum Key Distribution, 5G, multi-vendor network.**

1. WSTĘP

Sieci mobilne budowane są przy wykorzystaniu różnych technologii. Każda nowa generacja sieci przynosi nowe bardziej skomplikowane rozwiązania. Różnorodność potrzeb i możliwych ich rozwiązań powoduje konieczność budowy infrastruktury w oparciu o wielu dostawców. Klasycznym podejściem do budowy sieci jest wykorzystanie homogenicznych rozwiązań od jednego lub dwóch dostawców sprzętu. Firmy te dostarczają całość infrastruktury, własne rozwiązania sprzętowe i oprogramowanie (tzw. „purpose-built"). Innym podejściem jest rozdzielenie dostawców z podziałem na sprzęt i oprogramowanie (ang. commercial off-the-shelf, COTS). Rozwiązanie to powoduje zwiększenie liczby dostawców, zmniejszenie kosztów zakupu sprzętu, ale również konieczność koordynowania łańcuchów dostaw oraz całego procesu utrzymania sieci przez operatorów.

Realizowanie infrastruktury 5G na bazie sprzętu pochodzącego od wielu dostawców może okazać się korzystne, ponieważ zwiększa konkurencyjność rozwiązań oraz powoduje obniżenie kosztów samego rozwiązania. Niestety mogą wzrosnąć inne koszty np. koszty utrzymania sieci (mamy wiele różnych rozwiązań od różnych dostawców) lub utrzymania dodatkowego personelu technicznego. Przyjęty model budowy infrastruktury dla sieci 5G zakłada możliwość wyboru rozwiązań od mniejszych dostawców specjalizujących się w swoich dziedzinach, np. w elementach infrastruktury krytycznej. Operatorzy mogą także uelastycznić swoją ofertę, zwiększyć konkurencyjność i pozyskiwać nowych klientów.

Z punktu widzenia bezpieczeństwa sieci, zatomizowana infrastruktura, która powstaje w przypadku implementacji technologii 5G, przy wykorzystaniu sprzętu pochodzącego od wielu dostawców, wprowadza nowe zagrożenia. Do takich zagrożeń zalicza się np. rozwiązania chmurowe, a w szczególności wykorzystywanie zewnętrznych zasobów publicznych, brak kontroli nad wszystkimi warstwami sieci oraz zwiększona podatność, m.in. interfejsów obsługujących zewnętrzną komunikację. Infrastruktura narażona na zewnętrzne ataki wymaga nowego podejścia do zabezpieczenia i dostępu do zasobów.

W tej publikacji analizujemy wyzwania związane z bezpieczeństwem sieci 5G realizowanych w oparciu o sprzęt pochodzący od wielu dostawców (Sekcja 2) i pokazujemy, że niektóre modyfikacje obecnej architektury sieci przy wykorzystaniu technologii dystrybucji kluczy kwantowych (QKD) mogą być odpowiedzią na problemy bezpieczeństwa w takich sieciach. Przedstawiamy rozwiązanie QKD dla sieci 5G, które można wdrożyć w oparciu o istniejącą infrastrukturę transportową. W Sekcji 3 analizujemy ograniczenia takiego systemu QKD i stwierdzamy (w Sekcji 4), że może on znaleźć zastosowanie w obecnie rozwijanych sieciach 5G.

2. BEZPIECZEŃSTWO W SIECIACH OPARTYCH NA ROZWIĄZANIACH OD WIELU DOSTAWCÓW

Wiele zaleceń dotyczących budowy bezpiecznych sieci 5G zostało opracowane przez organizacje standaryzacyjne, takie jak 3GPP [1]. Zalecenia dotyczą możliwych zabezpieczeń funkcji sieciowych do bezpiecznej realizacji powierzonych im zadań, natomiast wiele elementów takich jak zabezpieczenie samej infrastruktury fizycznej nie jest rozważanych w standardach 3GPP.

Inna organizacja standaryzacyjna ETSI Industry Specification Group (ISG) definiuje zalecenia dotyczące wykorzystania infrastruktury klucza publicznego (PKI), dystrybucji kluczy do zabezpieczenia komunikacji w środowisku wirtualnym oraz interfejsów przeznaczonych do zarządzania i orkiestracji [2].

Dodatkowe wyzwanie w zabezpieczeniu infrastruktury i procesów zarządzania sieciami zbudowanych w oparciu o wielu dostawców, stanowi oddzielenie warstwy aplikacji od infrastruktury sprzętowej, realizacja funkcji sieciowych przez rozwiązania od różnych dostawców np. planista zadań (ang. scheduler) i inne metody implementacji oraz przez przyjęcie innych kryteriów do realizacji tego samego zadania [3].

Inicjatywa GSMA (ang. Global System for Mobile Communications), skupiająca operatorów telekomunikacyjnych, proponuje kompleksowy przegląd zagrożeń oraz technik ograniczania potencjalnych skutków ataków związanych z bezpieczeństwem całej infrastruktury [4]. Projekt „Cloud iNfrastructure Telco Taskforce" (CNTT), prowadzony wspólnie przez GSMA oraz The Linux Foundation zajmuje się standaryzacją infrastruktury chmurowej dla rozwiązań operatorskich. Niestety ta inicjatywa ogranicza się do zdefiniowania wzorcowej architektury, ujednolicenia interfejsów i ich właściwości bazując na modelach abstrakcyjnych. Zabezpieczenie infrastruktury, w tym rozwiązań opartych na chmurze, wdrożenie, pozostaje w gestii operatora sieci 5G [5].

Sieć 5G realizowana w oparciu o sprzęt pochodzący od wielu dostawców jest siecią rozproszoną, która komunikuje się z innymi fragmentami sieci i jest często zarządzaną przez otwarte interfejsy. Taki model sieci powoduje większą podatność na cyberataki oraz konieczność dodatkowych zabezpieczeń [6]. Dlatego, aby zwiększyć bezpieczeństwo sieci mobilnych 5G, proponujemy wprowadzenie dystrybucji kluczy zaufanych za pośrednictwem kanałów kryptograficznych QKD. Do realizacji takiego rozwiązania wymagane jest wydzielenie infrastruktury do kwantowej dystrybucji kluczy QKD, niezależnej od klasycznego kanału transportowego (rys. 1).

Przy wdrażaniu sieci 5G, operator sieci komórkowej (ang. Mobile Network Operator, MNO) definiuje akceptowalny poziom bezpieczeństwa oraz mechanizmy kontroli dostępu, filtrowania ruchu lub całkowitej izolacji usługi lub fragmentu sieci. W oparciu o świadczone usługi definiuje zabezpieczenia, np. na granicach chronionych obszarów (ang. perimeter protection). W przypadku rozwiązań chmurowych wymagane są dodatkowe mechanizmy zabezpieczeń.

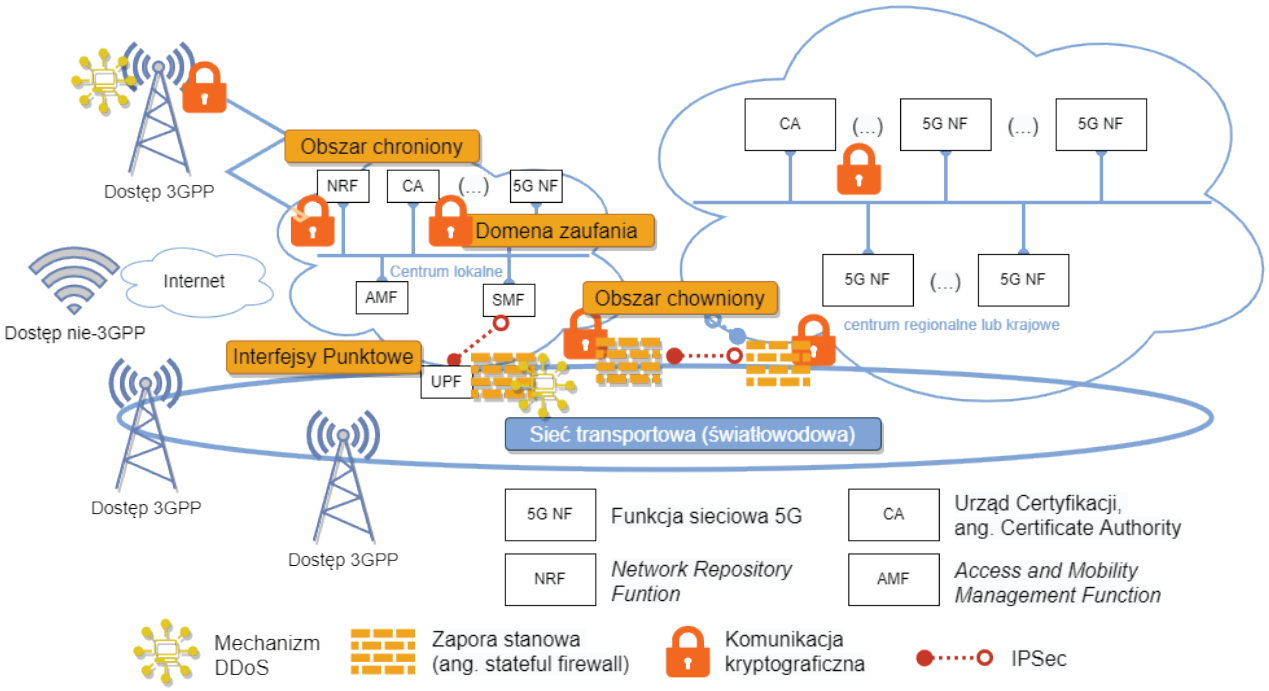
Jedną z metod zabezpieczenia sieci 5G jest pogrupowanie elementów sieciowych np. pod względem przetwarzania danych wrażliwych lub innych kryteriów takich jak np. możliwość wystawienia interfejsów na zewnątrz sieci. Przykładem takich interfejsów jest ruch roamingowy, dostęp do Internetu lub dostęp radiowy oparty na standardach 3GPP (np. sieci radiowe 4G, 5G) oraz pozostałych sposobach dostępu (np. Wi-Fi, LAN itp.) . Wyznaczenie takich obszarów powoduje ograniczenie skutków potencjalnego ataku. Skutki cyberataku będą dotyczyły tylko do tego obszaru. Możemy wyróżnić dwa typy komunikacji, które występują sieciach operatorów:

- Komunikacja wewnętrzna to komunikacja pomiędzy elementami, funkcjami sieciowymi, zasobami chmury. Do zabezpieczenia i monitoringu ruchu oraz przechowywania danych wykorzystuje się mechanizmy dostępne w danym środowisku, np. obszary (ang. namespace) w środowisku Kubernetesa, które definiuje obszary zaufane i polityki sieciowe do kontroli dostępu do takiego obszaru.
- Komunikacja zewnętrzna to komunikacja pomiędzy różnymi obszarami oraz ze środowiskiem zewnętrznym. Do zabezpieczenia stosuje dodatkowe, często dedykowane rozwiązania.

Sieć mobilna 5G zbudowana w oparciu o sprzęt wielu dostawców jest często bardzo skomplikowana i musi stawiać czoło różnym wyzwaniom pod względem bezpieczeństwa infrastruktury i świadczonych usług. Usługi wymagają przetwarzania danych w czasie rzeczywistym, wymagają bezpieczeństwa zintegrowanego w każdym elemencie sieci oraz infrastruktury chmurowej. Bardzo ważne jest, aby zapewnić dla usług integralność, poufność danych oraz kontrolować autentyczność i przepływ ruchu dla obu typów komunikacji wewnętrznej i zewnętrznej. Takim rozwiązaniem może zastosowanie uwierzytelniania za pomocą certyfikatów oraz protokołu IPsec.

W przypadku infrastruktury purpose-built, wiarygodność certyfikatów opiera się o urzędy certyfikacji (ang. Certificate Authorities, CA), oraz preinstalowanych rozwiązań od dostawcy sprzętu [7]. Dostawcy sprzętu przekazują swoje rozwiązania zarządzania certyfikatami do operatora. Przy wielu dostawcach, nie istnieje jeden CA dla wszystkich elementów i dlatego wymagane są różne urzędy certyfikacji. Urzędy certyfikacji wymagają wymiary informacji między sobą, w szczególności weryfikowania certyfikatów. Przykładową architekturę sieci MNO zlokalizowanej w różnych miejscach przedstawiliśmy na rysunku 1. W tym rozwiązaniu urząd rejestracyjny (ang. Registration Authority, RA) wymagany jest do weryfikacji rejestracja lub odnowienie certyfikatów i przesyłania ich dalej do CA.

Sieci transportowe operatorów przeważnie zbudowane są w technologii światłowodowej. Proponujemy wykorzystanie istniejących sieci do stworzenia równoległych kanałów komunikacyjnych dla QKD. Rozwiązanie to może zapewnić bezpieczną dystrybucję certyfikatów dla różnych CA. Zastosowanie łączy kwantowych pozwoli na uniknięcie ataków kryptograficznych polegających na podsłuchu i modyfikacji przesłanych certyfikatów (ang. man-in-the-middle, MITM) wewnątrz sieci (komunikacja wewnętrzna). W rozdziale III przedstawiamy szczegóły dotyczące tworzenia QKD w oparciu o istniejącą infrastrukturę światłowodową, która może pełnić rolę bezpiecznego, uwierzytelnionego kanału komunikacji.



Rys. 1. Bloki bezpieczeństwa i przykład dystrybuowanej (wiele lokalizacji) infrastruktury opartej na wiele dostawców

W sieciach 5G, stosuje się architekturę opartą na poziomie usług dla warstwy sygnalizacji (ang. Service-Based Architecture, SBA). SBA opiera się na łączeniu funkcji sieciowych 5G (ang. Network Function, NF) przez wspólną magistralę. Magistrala traktowana jest jako zaufana domena komunikacyjna, gdzie interfejsy każdego NF podłączonego do magistrali tworzą obszar chroniony.

W ramach tej koncepcji wyróżniamy następujące sposoby komunikacji:

- bezpośrednia lub pośrednia komunikacja między NF;
- proces identyfikacji NF przez jedną z usług NRF (ang. Network Repository Function) oraz rejestracji profili usług, wyszukiwania usług świadczonych przez inne funkcje sieciowe;
- rolę maskowania i ochronny obszaru chronionego pełni funkcja SCP (ang. Service Communication Proxy, SCP).

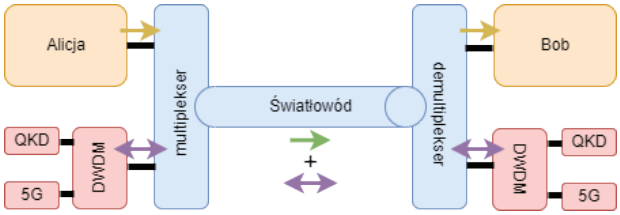
Pojedynczą magistralę SBA traktujemy jako domenę zaufania (ang. trust domain) [8]. Zaufanie pomiędzy 5G NF możemy zbudować przy doborze odpowiednich mechanizmów bezpieczeństwa, np. gdy (i) zapewniona jest poufność, integralność oraz ochrona komunikacji, oraz (ii) poprzez uwierzytelnianie i autoryzację samych NF. Wymagania bezpieczeństwa mogą zastać spełnione przy zastosowaniu: protokołu TLS, który zapewnia poufność, prywatność sieci oraz integralność transmisji oraz za pomocą protokołu OAuth2.0, do identyfikacji i autoryzacji przy użyciu tokenów. Jedną z usług funkcji sieciowej NRF jest funkcjonalność OAuth2.0 serwera. W połączeniu z wykorzystaniem protokołu mTLS oraz infrastrukturą PKI jest w stanie zapewnić ochronę integralności i poufności dla warstwy sygnalizacji. Protokół TLS jest podatny na ataki typu DDoS (ang. Distributed Denial of Service), który może wysycić zasoby przeznaczone do szyfrowania danych, dlatego konieczne jest zastosowanie dodatkowego elementu, który ograniczy wpływ ataku i nie dopuści do przeciążenia zasobów (ang. overload resources).

Problemem przy zastosowaniu protokołu TLS może być konieczność weryfikacji szyfrowanej sesji i diagnostyka zaszyfrowanych kanałów. Rozwiązaniem wspomnianych wcześniej problemów jest zastosowanie funkcji sieciowej SCP, która pełni rolę proxy dla połączeń TLS. SCP może odszyfrować kanał TLS, bezpiecznie przetworzyć dane i wysłać do odbiorcy w nowym zaszyfrowanym połączeniu. Dodatkowo, gdy rozważymy uwierzytelnianie oparte na certyfikatach (np. za pomocą procedury Client Credentials Assertion) możemy wyeliminować potrzebę posiadania użytkowników i haseł. Problem z użyciem uwierzytelniania opartego na certyfikatach polega na tym, że dwie części muszą wymieniać klucze publiczne w celu odszyfrowania danych, które zostały podpisane elektronicznie. W przypadku, gdy NF znajdują się w różnych obszarach chronionych, użycie QKD zapewni im bezpieczną wymianę kluczy, unikając ataków typu MITM. W następnej sekcji rozważamy zastosowanie QKD przy wykorzystaniu jednego włókna światłowodowego do wymiany kluczy pomiędzy dwoma urządzeniami końcowymi sieci 5G.

3. DYSTRUBUCJA KLUCZY KWANTOWYCH

Dystrybucja kluczy kwantowych (ang. Quantum Key Distribution – QKD) jest obecnie przedmiotem intensywnych prac badawczych w wielu ośrodkach na całym świecie. Jednym z istotnych powodów wzrostu zainteresowania tą tematyką jest szybki rozwój technologii komputerów kwantowych oraz odkrycie efektywnego algo-

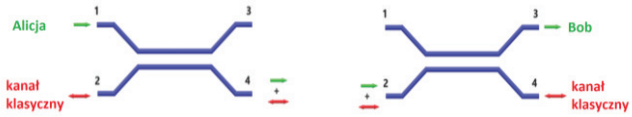
rytmu faktoryzacji przez Shore'a [9]. W tej publikacji rozważamy zastosowanie QKD w sieciach 5G opartych na wielu dostawcach sprzętu. W szczególności przedmiotem rozważań jest łącze światłowodowe, które umożliwia wymianę kluczy kwantowych oraz komunikację 5G przy wykorzystaniu tylko jednego włókna światłowodwego. Schemat takiego łącza jest przedstawiony na Rys. 2. Alicja i Bob wykorzystują pasmo O do przesyłu sygnału kwantowego, zaś pasmo C jest wykorzystywane do przesyłu sygnałów klasycznych, zarówno komunikacji 5G jak i QKD. Sygnały z obu pasm telekomunikacyjnych są łączone po stronie Alicji multiplekserem i rozdzielane po stronie Boba przy zastosowaniu demultipleksera.



Rys. 2. Schemat infrastruktury fotonicznej przeznaczonej do wymiany kluczy kwantowych między dwoma urządzeniami końcowymi sieci 5G.

Schematy multipleksera i demultipleksera są przedstawione na Rys. 3. Do realizacji tych urządzeń można zastosować odpowiednio zaprojektowane sprzęgacze kierunkowe. Urządzenia tego typu były dotychczas powszechnie wykorzystywane na przykład do łączenia sygnału optycznego i pompy we wzmacniaczach EDFA.

a) b)



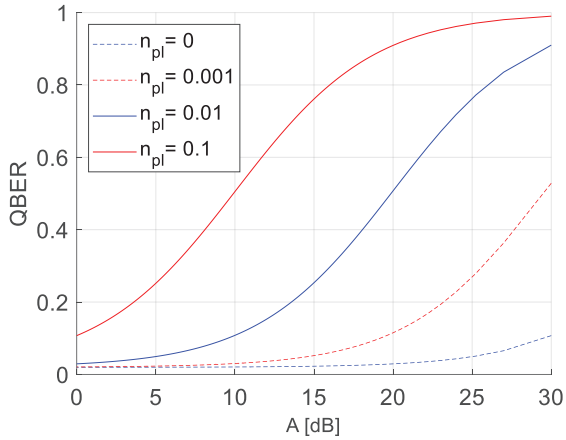
Rys. 3. Schemat multipleksera (a) i demultipleksera (b) infrastruktury fotonicznej QKD przedstawionej na Rys. 2.

W celu obliczenia stopy błędu transmisji kwantowej (ang. Quantum BER – QBER) można sie posłużyć następującą zależnością [10]:

$$QBER = \frac{P_{dc}+f_r A\eta P_{ap}+f_r \eta n_{pl}}{f_r A\eta+P_{dc}+f_r A\eta P_{ap}+f_r \eta n_{pl}} \qquad (1)$$

przy czym P_{dc} oznacza prawdopodobieństwo detekcji spowodowanej przez prąd ciemny fotodiody, f_r częstotliwość przesyłania bitów kwantowych, A -tłumienie w kanale kwantowym, η - sprawność fotodiody, P_{ap} – prawdopodobieństwo emisji dodatkowgo impulsu po detekcji fotonu, n_{pl} liczba fotonów, które przeniknęły z kanału klasycznego do kanału kwantowego w czasie jednego okresu transmisji kwantowej równej $1/f_r$.

Rys. 4 przedstawia zależność QBER od tłumienności łącza kwantowego dla kilku wybranych wartości n_{pl}, która została wyznaczona na podstawie (1) dla typowych parametrów łącza przy zastosowaniu fotodiody lawinowej jako fotodetektora w kanale kwantowym. Wartości QBER poniżej 0.276 pozwalają na wymianę kluczy kwantowych [11]. Przy n_{pl} = 0.001 można zatem tolerować straty około 25 dB. To odpowiada odległości ponad 60 km przy tłumienności światłowodu 0.4 dB/km w paśmie O.



Rys. 4. Zależność QBER od tłumienności łącza kwantowego dla wybranych wartości n_{pl}. multipleksera

4. PODSUMOWANIE

Przedstawiono wyzwania bezpieczeństwa w sieciach mobilnych budowanych na podstawie sprzętu dostarczanego przez wiele dostawców. Zaproponowano wprowadzenie kanału kwantowego QKD do rozwiązania niektórych z ww. wyzwań przy założeniu, że jest on zrealizowany przy wykorzystaniu dostępnej infrastruktury światłowodowej. Wyniki analizy QBER w łączu światłowodowym potwierdzają, że technologia QKD może zostać wykorzystana do wymiany kluczy kryptograficznych w sieciach 5G.

PODZIĘKOWANIE

Składamy podziękowanie partnerom projektu MPSEC5G „Struktura bezpieczeństwa sieci 5G opartej na wielu dostawcach: specyfikacja, wdrożenie i rozwój procesu ewaluacji" dofinansowanego przez Narodowe Centrum Badań i Rozwoju, za wymianę pomysłów i owocne dyskusje.

LITERATURA

[1] Security architecture and procedures for 5G System. 3GPP Technical Specification 33.501, Oct. 2018
[2] Network Functions Virtualisation (NFV); Trust; Report on Certificate Management, ETSI Group Report NFV-SEC 005, July 2021
[3] A. Garcia-Saavedra and X. Costa-Pérez, "O-RAN: Disrupting the Virtualized RAN Ecosystem," IEEE Communications Standards Magazine, vol. 5, no. 4, pp. 96-103, Dec. 2021
[4] Network Function Virtualisation (NFV) Threats Analysis, GSMA Association official Document FS.33, March 2020
[5] J. Mongay Batalla, E. Andrukiewicz, G. Peinado Gomez, P. Sapiecha, C. X. Mavromoustakis, G. Mastorakis, J. Zurek, M. Imran, Security Risk Assessment for 5G networks - national perspective.

IEEE Wireless communications, vol. 27, no. 4, pp. 16-22, Aug. 2020
[6] T. Wichary, J. Mongay Batalla, C. X. Mavromoustakis, J. Żurek and G. Mastorakis, Network Slicing Security Controls and Assurance for Verticals, Electronics 2022, 11(2), 222
[7] L. Harn and J. Ren, "Generalized Digital Certificate for User Authentication and Key Establishment for Secure Communications," IEEE Transactions on Wireless Communications, vol. 10, no. 7, pp. 2372-2379, July 2011
[8] H. C. Rudolph, A. Kunz, L. L. Iacono and H. V. Nguyen, "Security Challenges of the 3GPP 5G Service Based Architecture," IEEE Communications Standards Magazine, vol. 3, no. 1, pp. 60-65, March 2019
[9] R. Romaniuk, “Kwantowa Alicja i Bob,” Elektronika - konstrukcje, technologie, zastosowania, vol. 63, no. 1, pp. 18-26, 2021.
[10] G.Ribordy, J.D. Gautier, N. Gisin, O. Guinnard, H.Zbinden, ”Fast and user friendly quantum key distribution”, arXiv:quant-ph/9905056.
[11] H. F. Chau, “Practical Scheme To Share A Secret Key Through An Up To 27.6% Bit Error Rate Quantum Channel”, arXiv:quant-ph/0205060v3.

ANALIZA ZALEŻNOŚĆI POMIĘDZY PARAMETRAMI SIECI TPM A CZASEM SYNCHRONIZACJI

ANALYSIS OF THE RELATIONSHIP BETWEEN TPM NETWORK PARAMETERS AND ITS SYNCHRONIZATION TIME

Miłosz Stypiński[1]

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

DOI: 10.15199/59.2022.4.35

Streszczenie: **Kryptografia z wykorzystaniem sieci neuronowych to nowy trend w cyberbezpieczeństwie. Sieci typu Tree Parity Machine oferują alternatywę dla powszechnie stosowanych protokołów uzgodnienia klucza kryptograficznego. Celem tego artykułu jest analiza zależności pomiędzy parametrami sieci TPM, a czasem synchronizacji oraz jej bezpieczeństwem.**
Słowa kluczowe: **bezpieczeństwo, kryptografia, sieci neuronowe, uzgodnienie klucza kryptograficznego**

Abstract: **Neural cryptography is a new trend in cybersecurity. Tree Parity Machine (TPM) networks offer an alternative to commonly used cryptographic key agreement protocols. The purpose of this article is to analyze the relationship between TPM network parameters and the synchronization time and its security.**
Keywords: **cryptography, neural networks, secure key agreement, security**

1. WSTĘP

Zmierzając ku coraz powszechniejszej cyfryzacji, bezpieczeństwo staje się coraz bardziej kluczową składową systemów i aplikacji. Niezależnie od dziedziny zainteresowania, cyberbezpieczeństwo nie może zostać odsunięte na dalszy plan. W innym wypadku dochodzi do bardzo medialnego wykorzystania podatności, tak jak to miało miejsce w 2018 roku w przypadku firmy Facebook, gdzie zostałby wykradzione dane 30 milionów użytkowników [1].

Algorytmy bezpieczeństwa stanowią nierozłączny element dzisiejszych rozwiązań informatycznych, gdzie wśród nich znajdują się algorytmy szyfrowania asymetrycznego. Najpopularniejsze algorytmy szyfrowania asymetrycznego stosowanie obecnie to algorytmy RSA i pochodne; oraz rodzina algorytmów opartych na krzywych eliptycznych. Niestety wymienione grupy algorytmów są bezpieczne warunkowo, to znaczy, że opierają się na problemie obliczeniowym, który na dzień dzisiejszy nie posiada efektywnego algorytmu będącego w stanie odwrócić odpowiednich operacji matematycznych w racjonalnym czasie. Problemy na których oparte są wspomniane algorytmy to odpowiednio faktoryzacja liczb1 pierwszych oraz „dodawanie" punktów na krzywej eliptycznej w pierścieniu modulo. Niestety, oba te problemy mogą być efektywnie rozwiązywanie w przypadku implementacji algorytmu Shor'a na komputerze kwantowym [6]. Aby zapobiec załamaniu bezpieczeństwa wszystkich istniejących systemów informatycznych, powstają algorytmy odporne na działanie komputera kwantowego. Jednym z nich jest uzgodnienie klucza z wykorzystaniem sieci typu Tree Parity Machine (TPM).

2. Sieć TPM

W celu przedstawienia algorytmu synchronizacji dwóch sieci TPM, co pozwala na uzgodnienie klucza kryptograficznego, najpierw zostanie przedstawiona jej struktura, następnie protokół uzgodnienia klucza, a na koniec rozdziału przedstawione zostaną reguły uczące iteratywnie aktualizujące wagi sieci TPM, tworzące klucz kryptograficzny.

2.1. Struktura sieci TPM

Sieć TPM to struktura przypominająca dwuwarstwowy perceptron. W celu dokładnego opisania sieci TPM wykorzystuje się trzy parametry: *K, L, N,* które są kolejno odpowiedzialne za liczbę wejść do każdego z neuronów pierwszej warstwy; za granicę przedziału ograniczającego wartości wag oraz za liczbę neuronów w pierwszej warstwie sieci [2]. Sieć jest uczona binarnym wektorem wejściowym, którego składowe przyjmują wartości -1 lub 1. Na wyjściu sieci znajduje się tylko jeden neuron.

Znaczącą różnicą pomiędzy modelem perceptronu, a strukturą TPM jest fakt, że wagi w sieci TPM mogą przyjmować jedynie wartości całkowitoliczbowe. Ponadto, funkcje aktywacji zostały zmodyfikowane odpowiednio, tak aby zwracały wartości binarne -1 lub 1. Funkcja aktywacji w neuronach Θ pierwszej warstwy jest zmodyfikowaną funkcją signum i ma następującą postać:

$$\sigma = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \tag{1}$$

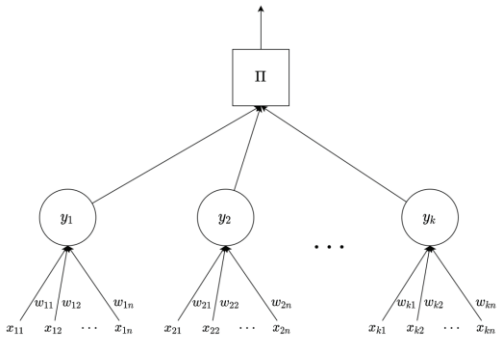
Z wykorzystaniem funkcji (1) możemy przedstawić wzór na wartość wyjściową neuronu pierwszej warstwy (2).

$$y_k = \sigma\left(\sum_{n=1}^{N} x_{kn} \cdot w_{kn}\right) \tag{2}$$

Neuron drugiej warstwy nie posiada powiązanych z nim wag, a funkcja aktywacji jest tożsama z iloczynem wartości w wektorze wejściowym. Z uwagi na fakt, że wektor wejściowy składa się z liczb -1 lub 1, w rezultacie wyjście całej sieci też jest równe -1 lub 1, co jest kluczowe w uzgodnieniu klucza kryptograficznego.

Dodatkowo w [7] autorzy zaproponowali nowy parametr M, odpowiedzialny za dywersyfikację wartości, które może przyjmować wektor wejściowy. Parametr M tworzy przedział wartości $< -M, M >$, który ogranicza wartości poszczególnych elementów wektora wejściowego. Jednak, te wartości muszą pozostać całkowitoliczbowe, aby umożliwić uzgodnienie klucza kryptograficznego. Wprowadzenie parametru M nie ma wpływu na protokół uzgodnienia klucza, a skutkuje przyspieszeniem synchronizacji, a co za tym idzie zwiększa bezpieczeństwo całego systemu. Dalej w badaniach rozważane są tylko sieci TPM o parametrach K, L, M i N. Analiza parametrów sieci TPM bez parametru M jest dostępna między innymi w [4].

Schemat sieci TPM został przedstawiony na Rysunku 1.



Rys. 1 Struktura sieci TPM

2.2. Uzgodnienie klucza kryptograficznego w sieci TPM

Celem synchronizacji dwóch sieci TPM jest uzgodnienie klucza kryptograficznego w bezpieczny sposób, co jest tożsame z tym, że żaden postronny podmiot nie będzie w stanie odtworzyć uzgodnionego sekretu na podstawie wymienionych informacji.

Protokół wykorzystujący sieci TPM do uzgodnienia sekretu został przedstawiony poniżej:

1. strony uzgadniające klucz jawnie określają parametry K, L, M, N sieci TPM;
2. jedna ze stron wybiera losowy wektor X składający się z wartości $x_i \in < -N, N >$;
3. każda ze stron oblicza odpowiedź sieci dla wcześniej ustalonego wektora X i ujawnia odpowiedź publicznie;
4. w przypadku, gdy odpowiedzi obu sieci się zgadzają, każda ze stron aktualizuje wagi swojej sieci stosując odpowiednią regułę uczącą;
5. punkty od 2 do 4 powtarzane są do momentu pełnej synchronizacji obu sieci.

Aby powyższy protokół mógł działać wymagane są reguły uczące, których zastosowanie skutkuje pełną synchronizacją sieci w skończonym czasie. W [5] opisane zostały trzy różne reguły uczące, które zostały zaprezentowane poniżej:

- reguła Hebbian

$$w_{kn}(t+1) = w_{kn}(t) + O(t)x_{kn}(t)\Theta\big(y_k(t), O(t)\big) \quad (3)$$

- reguła anti-Hebbian

$$w_{kn}(t+1) = w_{kn}(t) - O(t)x_{kn}(t)\Theta\big(y_k(t), O(t)\big) \quad (4)$$

- reguła losowego spaceru

$$w_{kn}(t+1) = w_{kn}(t) + x_{kn}(t)\Theta\big(y_k(t), O(t)\big) \quad (5)$$

W powyższych wzorach t oznacza iterację protokołu uzgodnienia klucza, k oznacza k-ty neuron sieci TPM, n oznacz n-te wejście neuronu, a y_n jest wyjściem n-tego neuronu pierwszej warstwy sieci.

W [3] autorzy udowodnili, że stosowanie powyższego protokołu skutkuje synchronizacją sieci TPM w skończonym czasie oraz, co jest równie ważne, w czasie krótszym niż osoby postronne próbujące wykorzystać wymienione dane do odtworzenia uzgodnionego klucza kryptograficznego.

Długość uzgodnionego klucza kryptograficznego zależy od parametrów sieci TPM i wyraża się następującym wzorem:

$$Długość\ klucza = K \times N \times log_2(L) \quad (6)$$

Trzeba jednak pamiętać, że wybór parametrów K, L, M, N również wpływa na rozkład wartości wag podczas synchronizacji, więc efektywna długość klucza kryptograficznego będzie odgórnie ograniczona przez (6) [7].

3. WPŁYW PARAMETRÓW NA CZAS SYNCHRONIZACJI

W tej sekcji przedstawione są metodologia badań, opis środowiska symulacyjnego oraz otrzymane wyniki.

3.1. Metodologia badań

W celu sprawdzenia wpływu zmian wartości parametru M na bezpieczeństwo uzgodnionego klucza wykorzystane zostało dedykowane środowisko symulacyjne. W każdym uruchomieniu synchronizowane były 3 różne sieci TPM. Dwie sieci należące do stron chcących uzgodnić kluczy kryptograficzny oraz dodatkowa jedna sieć TPM próbująca odtworzyć klucz kryptograficzny na podstawie wymienianych wektorów wejściowych. Ostatnia z wymienionych sieci w dalszej części artykułu nazywana jest złośliwą siecią TPM.

Przeprowadzone symulacje wykorzystują sieci TPM z parametrami K i N równymi odpowiednio 3 i 50. Parametry M i L były zależne od symulowanego scenariusza i przybierały odpowiednio wartości z następujących zbiorów: $\{1,2,3,4,5,6,7\}$ oraz $\{8,9,10,11\}$. Dla każdego zestawu parametrów przeprowadzono 100 symulacji, podczas których badana była potrzebna liczba iteracji do pełnej synchronizacji oraz podobieństwo złośliwej sieci TPM według następującego wzoru:

$$S_{score} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N}\Theta\big(w_{kn}, w_{kn}^{A}\big)}{K \times N}, \quad (6)$$

gdzie Θ to funkcja zwracająca 1 w przypadku, gdy jej argumenty są równe, a w przeciwnym przypadku zwracająca 0.

Tab.1 Średnia liczba iteracji oraz średni poziom synchronizacji złośliwej sieci TPM

		Synchronizowane sieci TPM				Złośliwa sieć TPM			
L	M	Średnia	Mediana	Min	Max	Średnia	Mediana	Min	Max
8	1	$1699,6 \pm 784,4$	1631	1010	2614	$0,133 \pm 0,076$	0,127	0,067	0,207
	2	$690,5 \pm 310,8$	695	388	1092	$0,128 \pm 0,087$	0,127	0,067	0,260
	3	$415,4 \pm 236,8$	406	619	619	$0,147 \pm 0,089$	0,140	0,080	0,234
	4	$240,1 \pm 144,3$	228	141	429	$0,150 \pm 0,0882$	0,140	0,073	0,226
	5	$205,1 \pm 191,9$	178	84	580	$0,184 \pm 0,177$	0,160	0,080	0,494
	6	$131,2 \pm 97,6$	119	75	251	$0,168 \pm 0,140$	0,153	0,067	0,326
	7	$118,6 \pm 83,6$	110	71	229	$0,210 \pm 0,216$	0,180	0,053	0,467
9	1	$2337,0 \pm 1298,7$	2095	1456	4290	$0,119 \pm 0,082$	0,107	0,053	0,200
	2	$910,0 \pm 506,4$	869	491	1462	$0,129 \pm 0,097$	0,120	0,067	0,253
	3	$495,6 \pm 275,6$	518	272	518	$0,130 \pm 0,122$	0,120	0,393	0,393
	4	$328,6 \pm 166,0$	310	164	470	$0,150 \pm 0,108$	0,133	0,080	0,287
	5	$266,2 \pm 225,4$	258	110	596	$0,141 \pm 0,101$	0,140	0,073	0,340
	6	$158,7 \pm 106,4$	145	69	299	$0,152 \pm 0,108$	0,147	0,087	0,360
	7	$124,6 \pm 72,2$	114	68	215	$0,209 \pm 0,334$	0,187	0,067	0,980
10	1	$2552,4 \pm 1288,6$	2399	1812	4681	$0,116 \pm 0,045$	0,113	0,073	0,160
	2	$1246,2 \pm 866,8$	1114	700	2175	$0,114 \pm 0,067$	0,113	0,053	0,173
	3	$611,1 \pm 280,6$	584	314	912	$0,113 \pm 0,065$	0,113	0,033	0,200
	4	$363,3 \pm 221,1$	350	201	694	$0,141 \pm 0,115$	0,133	0,040	0,314
	5	$275,28 \pm 174,6$	259	158	470	$0,134 \pm 0,102$	0,120	0,080	0,340
	6	$264,6 \pm 164,0$	253	110	550	$0,150 \pm 0,131$	0,133	0,087	0,326
	7	$168,3 \pm 95,9$	156	101	262	$0,152 \pm 0,108$	0,147	0,053	0,280
11	1	$3019,8 \pm 1453,6$	3002	1946	4665	$0,121 \pm 0,054$	0,120	0,067	0,180
	2	$1333,3 \pm 745,9$	1290	571	2493	$0,112 \pm 0,073$	0,100	0,060	0,233
	3	$774,0 \pm 444,7$	706	377	1248	$0,103 \pm 0,062$	0,100	0,027	0,166
	4	$528,2 \pm 219,6$	528	311	789	$0,116 \pm 0,090$	0,113	0,060	0,240
	5	$350,56 \pm 199,0$	351	159	547	$0,127 \pm 0,117$	0,113	0,053	0,374
	6	$264,6 \pm 164,0$	253	115	427	$0,150 \pm 0,131$	0,133	0,087	0,326
	7	$218,4 \pm 162,9$	188	107	427	$0,154 \pm 0,140$	0,127	0,073	0,427

3.1. Analiza wyników

W wyniku symulacji można zauważyć, że wraz ze wzrostem parametru *L* mediana potrzebnych iteracji do pełnej synchronizacji rośnie. Należy jednak pamiętać, że zwiększenie parametru *L* skutkuje również wydłużeniem otrzymanego klucza kryptograficznego zgodnie ze wzorem (6). Ponadto, można zauważyć, że zwiększenie wartości parametru M znacząco redukuje medianę potrzebnych iteracji protokołu. Jednak, negatywnym skutkiem wzrostu wartości parametru M jest również szybsza synchronizacja złośliwej sieci TPM.

Można też zauważyć, że dla wartości parametru M równego 7, maksymalna synchronizacja złośliwej sieci TPM zbliżała się do wartości 0,5. W przypadku, gdy strona atakująca zwielokrotniłaby atak poprzez synchronizowanie wielu sieci na raz, istnieje prawdopodobieństwo, że klucz został by uzyskany przez stronę atakującą, co skutkowałoby kompromitacją całego systemu bezpieczeństwa.

Wyniki wszystkich symulacji zostały przedstawione w Tabeli 1.

4. PODSUMOWANIE

W artykule przedstawione zostało działanie protokołu uzgodnienia klucza kryptograficznego wykorzystującego sieci TPM. Rozważone zostały tylko sieci TPM z dodatkowym parametrem M. Następnie dla sieci TPM o różnych zestawach parametrów przeprowadzona została analiza zmian wartości parametrów na średnią długość uzgodnienia klucza wyrażoną w iteracja oraz średnią synchronizację złośliwej sieci TPM dla każdego ze scenariuszy. Całość została zakończona analizą otrzymanych wyników.

Sieci TPM mogą stać się konkurencyjnym rozwiązaniem w momencie udanej implementacji algorytmu Shor'a na komputerze kwantowym, ze względu na odporność na działanie tego algorytmu. Przedstawiona analiza wykazała, że kluczowym aspektem tworzenia systemu kryptograficznego wykorzystującego sieci TPM jest dobór jej parametrów, tak aby zapewnić wymaganą długość klucza projektowanego systemu kryptograficznego przy jednoczesnej maksymalizacji bezpieczeństwa wymiany klucza kryptograficznego z wykorzystaniem sieci TPM.

LITERATURA

[1] Blog aplikacji Facebook, https://about.fb.com/news/2018/10/update-on-security-issue/, data dostępu: 2022-05-09

[2] Kanter I., Kinzel W., i Kanter E., "Secure exchange of information by synchronization of neural networks," EPL (Europhysics Letters), vol. 57, 02 2002.

[3] Kinzel W., "Theory of interacting neural networks", ch. 9, pp. 199–217. John Wiley & Sons, Ltd, 2002

[4] Rosen-Zvi M., Kanter I., i Kinzel W., "Cryptography based on neural networks analytical results," Journal of Physics A: Mathematical and General, vol. 35, pp. L707–L713, 10 2002.

[5] Ruttor A., Neural Synchronization and Cryptography. Rozprawa doktorska, Uniwersytet w Wurzburgu, 2007.

[6] Shor P. W., "Algorithms for quantum computation: discrete logarithms and factoring," Proceedings 35th Annual Symposium on Foundations of Computer Science, 1994, pp. 124-134, doi: 10.1109/SFCS.1994.365700.

[7] M. Stypiński and M. Niemiec, "Synchronization of Tree Parity Machines Using Nonbinary Input Vectors," in IEEE Transactions on Neural Networks and Learning Systems, 2022, doi: 10.1109/TNNLS.2022.3180197.

PORÓWNANIE ARCHITEKTUR APLIKACJI SIECIOWYCH: MONOLITYCZNA I MIKROSERWISÓW

COMPARISON OF NETWORK APPLICATION ARCHITECTURES: MONOLITHIC AND MICROSERVICES

Sebastian Kulig[1]; Andrzej Matiolański[1]

[1] AGH Akademia Górniczo-Hutnicza, Wydział Informatyki, Elektroniki i Telekomunikacji, Instytut Telekomunikacji Kraków, Polska

e-mail: andrzej.matiolanski@agh.edu.pl

DOI: 10.15199/59.2022.4.36

Streszczenie: **Architektura monolityczna oraz mikroserwisów to dwa popularne sposoby budowania aplikacji sieciowych. Artykuł zawiera ich porównania na podstawie przeprowadzonych badan wydajnościowych. Podsumowanie zawiera rekomendacje do stosowania określonego rodzaju architektury w konkretnym przypadku.**

Abstract: **Monolithic architecture and microservices are two popular ways to build web applications. The article contains their comparisons based on the conducted performance research. The summary includes recommendations for the use of a specific type of architecture in a specific case.**

Słowa kluczowe: **architektury sieciowe, architektura mikroserwisów, architektura monolityczna, skalowalność, wydajność**

Keywords: **network architectures, microservice architecture, monolithic architecture, scalability, performance**

1. WPROWADZENIE

Architektura oprogramowania określa sposoby organizacji komponentów systemu, ich budowę oraz rozwój tak, aby spełniały założone cele. Jej wybór jest jednym z pierwszych etapów w procesie tworzenia oprogramowania, który odbywa się po zebraniu wymagań od klienta. Wybór właściwej architektury jest jednym z kluczowych zagadnień w procesie tworzenia oprogramowania. Znacząco wpływa na funkcjonowanie aplikacji, szybkość wdrażania nowych funkcjonalności i bezpieczeństwo. Migracja pomiędzy różnymi stylami architektur jest procesem długotrwałym i kosztownym [1, 3].

Jak stwierdzono w standardzie ISO/IEC/IEEE 42030 [5], definiującym sposoby opisu i organizowania systemów, jednymi z najważniejszych aspektów, jakie należy brać pod uwagę przy ocenie architektury, są efektywne wykorzystanie dostępnych zasobów oraz wydajność aplikacji z punktu widzenia użytkownika końcowego. W artykule została porównana architektura monolityczna oraz mikroserwisów. Pierwsze z powyższych rozwiązań od dawna jest standardowym wyborem w wielu projektach i nawet takie firmy jak Spotify, LinkedIn, Twitter, Amazon czy Netflix w początkowej fazie swojej działalności opierały tworzone platformy właśnie o architekturę monolityczną. Zapewnia ona satysfakcjonującą wydajność w przypadku małych i średnich aplikacji, jednak wraz ze wzrostem liczby funkcjonalności i skomplikowania tworzonego oprogramowania, uwydatniają się jej wady. Rosnąca ilość kodu w obrębie jednego projektu, dokładanie kolejnych technologii, coraz większa liczba użytkowników aplikacji, a co za tym idzie, większe zapotrzebowanie na zasoby sprawia, że konieczne stało się znalezienie nowego rozwiązania. Naprzeciw tym wymaganiom wychodzi architektura rozproszona. Jednym z prekursorów zastępowania architektury monolitycznej grupą wyspecjalizowanych i lżejszych serwisów była firma Netflix. Obecnie już wszystkie z wyżej wymienionych przedsiębiorstw dokonały migracji swoich systemów do architektury mikroserwisów. Architektura ta, mimo iż rozwiązuje wiele problemów, które występowały w klasycznych systemach, to wciąż jej wpływ na wydajność aplikacji w porównaniu do wersji monolitycznej jest przedmiotem wielu badań [2 ,4, 6, 7].

Celem artykułu jest porównanie, jak wybór jednego z wyżej wymienionych podejść wpływa na efektywność systemu w różnych scenariuszach. W tym celu przygotowane zostały dwie aplikacje. Jedna w architekturze monolitycznej, a druga z wykorzystaniem mikroserwisów. Obie oferują identyczne funkcjonalności i zostały wykonane, w miarę możliwości, z wykorzystaniem tych samych technologii. Następnie przeprowadzono szereg testów obciążeniowych w celu zbadania podstawowych parametrów związanych z funkcjonowaniem oprogramowania w zależności od natężenia ruchu i dostępnych zasobów.

2. CHARAKTERYSTYKA ARCHITEKTUR

Architektura monolityczna polega na umieszczeniu rdzenia funkcjonalności w jednym projekcie, przy czym możliwa jest interakcja z innymi usługami czy bazami danych. Na przestrzeni lat pojawiło się wiele wersji tej architektury, od pojedynczego bloku kodu, do systemów określanych jako N-warstwowe. Zwykle jednak większość aplikacji komercyjnych budowanych jest według trzywarstwowego schematu: interfejsu użytkownika, logiki biznesowej oraz warstwy dostępu do danych.

Architektura mikroserwisów polega na dekompozycji pojedynczej aplikacji, na szereg mniejszych wyspecjalizowanych serwisów, będących niezależnymi jednostkami. Poszczególne moduły powinny być luźno ze sobą powiązane, co umożliwia równoległą pracę nad różnymi usługami. Komunikacja pomiędzy poszczególnymi komponentami może odbywać się za pomocą protokołów synchronicznych, takich jak HTTP/REST lub asynchronicznych np. AMQP.

3. ŚRODOWISKO TESTOWE

W celu przeprowadzenia testów obie wersje aplikacji zostały umieszczone w kontenerach z systemem operacyjnym Linux na serwerze posiadającym 16 GB pamięci RAM, procesor Intel Core i5-7400 oraz dysk SSD z prędkością zapisu 940 MB/s. W przypadku zastosowania innej konfiguracji wartości liczbowe dla przeprowadzonych eksperymentów mogą się różnić, jednak zachowanie architektur będzie zbliżone i pozwoli na wyciągnięcie podobnych wniosków. Serwer połączony został z drugim komputerem znajdującym się w tej samej sieci lokalnej, na którym uruchomiono program JMeter. Monitoring zasobów przeprowadzono za pomocą narzędzi open-source: Grafana, Prometheus, cAdvisor.

Te dwie aplikacje (o architekturze monolitycznej i mikroserwisów) wykonywały szereg typowych dla aplikacji internetowych operacji: Create account (1), Authenticate (2), Get all products (3), Get one product details (4), Make an order (5).

4. EKSPERYMENTY

W ramach pracy rozważone zostały trzy scenariusze eksperymentalne dla obu rodzajów architektur:

4.1. Praca pod umiarkowanym obciążeniem

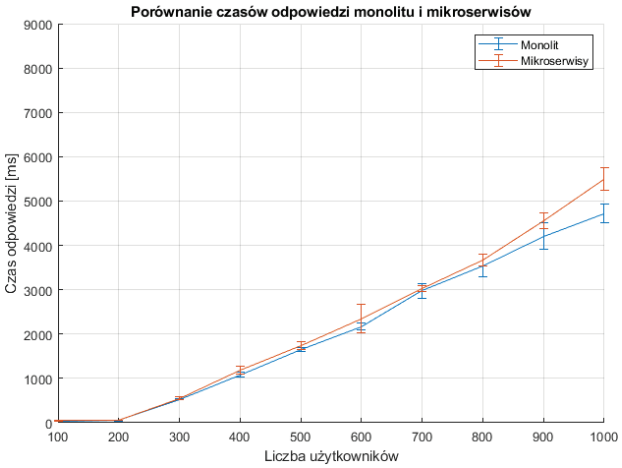
Scenariusz testowy przewidywał wykorzystanie obciążenia aplikacji bez zbliżania się do limitów ich możliwości. W tym celu zasymulowano działanie 100 użytkowników. Co sekundę dodawanych było 10 wirtualnych użytkowników. Sumarycznie podczas testu wykonano 5000 zapytań w każdej z aplikacji.
Dane zawarte w tabeli 1. i 2. przedstawiają statystykę dla zapytań w całym eksperymencie. Wartości średnie są zbliżone, dla architektury mikroserwisów czasy maksymalne są zdecydowanie wyższe. Na rysunku 2. widzimy podobne czasy odpowiedzi przez większość eksperymentu.

Tab. 1. Czasy wykonania operacji dla architektury monolitycznej dla 1000 prób w milisekundach

Operacja	*Śr.*	*Min.*	*Max.*	*Odch.*
(1)	82	68	223	19,31
(2)	83	73	233	12,71
(3)	15	9	78	8,09
(4)	11	7	62	4,90
(5)	24	13	129	11,70
Razem (1-5)	43	7	234	3,03

Tab. 2. Czasy wykonania operacji dla architektury mikroserwisów dla 1000 prób w milisekundach

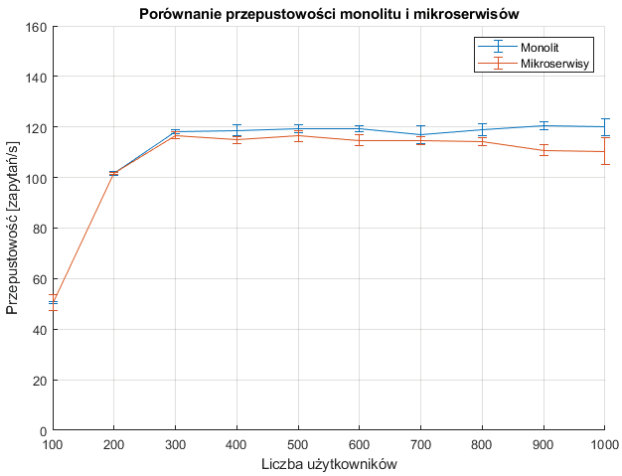
Operacja	*Śr.*	*Min.*	*Max.*	*Odch.*
(1)	88	67	402	30,48
(2)	88	72	442	32,97
(3)	18	11	103	10,95
(4)	14	8	65	6,61
(5)	20	12	288	18,05
Razem (1-5)	46	8	442	11,01



Rys. 1. Analiza czasów odpowiedzi

4.2. Praca przy dużej liczbie użytkowników

W celu zbadania zachowania obu architektur pod wpływem rosnącego obciążenia przeprowadzono serię pomiarów, zwiększając stopniowo liczbę użytkowników. Testy rozpoczęto od 100 użytkowników i kontynuowano dodawanie kolejnych (10/sekundę), aż do 1000 użytkowników. Jest to sytuacja, w której pomimo zwiększania ruchu przestaje rosnąć przepustowość. Analizie podlegała wydajność całej aplikacji, jak i również parametry związane z poszczególnymi rodzajami zapytań.



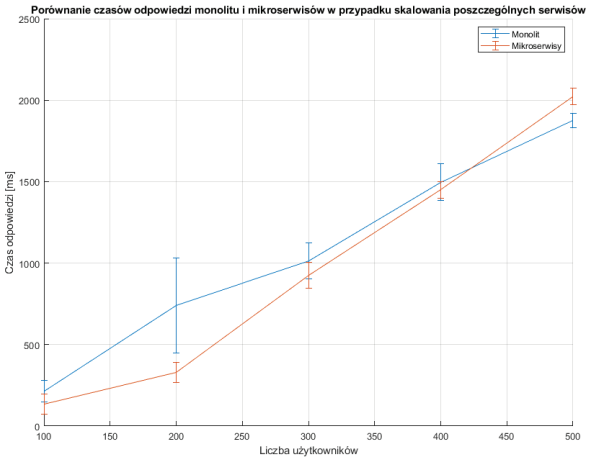
Rys. 2. Analiza czasów odpowiedzi

Na rysunku 2. widzimy, że do pewnego momentu przyrostu użytkowników przepustowość rośnie. W punkcie wysycenia widzimy widoczny spadek przepustowości w przypadku wykorzystania mikroserwisów.

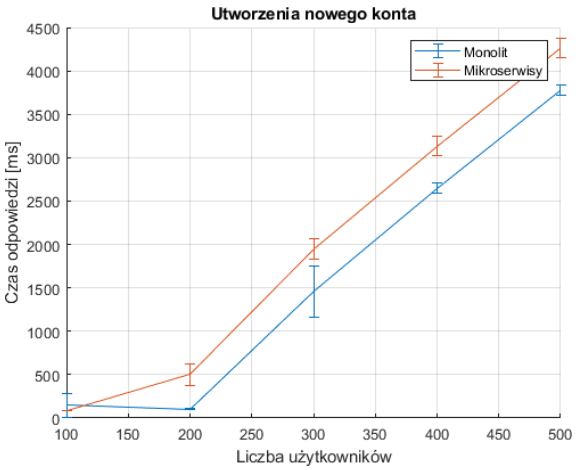
4.3. Porównanie skalowalności obciążonej aplikacji dla obu rodzajów architektur

W trzecim z testowanych scenariuszy ocenie podlegały potencjalne zyski wynikające z możliwości punktowego skalowania architektury mikroserwisów. W większości systemów obciążenie rozkłada się nierównomiernie na poszczególnych komponentach. W przypadku wersji monolitycznych jedyną możliwością poprawy wydajności, bez ingerencji w kod jest zwiększenie ilości zasobów dla całej aplikacji lub dodanie kolejnych jej instancji i zastosowanie równoważenia obciążenia. Architektura mikroserwisów dzięki swojej granularności, pozwalają na skalowanie pojedynczych serwisów, które stanowią wąskie gardła systemu. W ten sposób możemy w lepszy sposób rozdysponowywać dostępne zasoby.

W ramach tego scenariusza ograniczono ilość pamięci RAM do 4 GB i zezwolono na dostęp do 25% całości mocy obliczeniowej procesora. Zasoby te zostały w całości przydzielone do aplikacji monolitycznej i taka sama ich ilość została rozdysponowana pomiędzy poszczególne serwisy drugiej wersji aplikacji zgodnie z ich zużyciem.

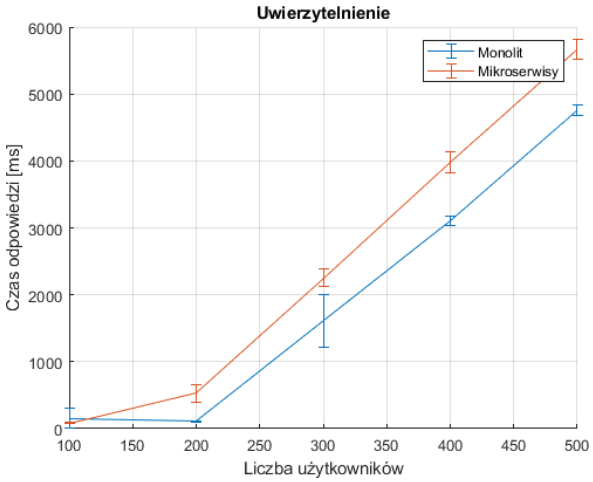


Rys. 3. Analiza przepustowości pod kątem ilości użytkowników

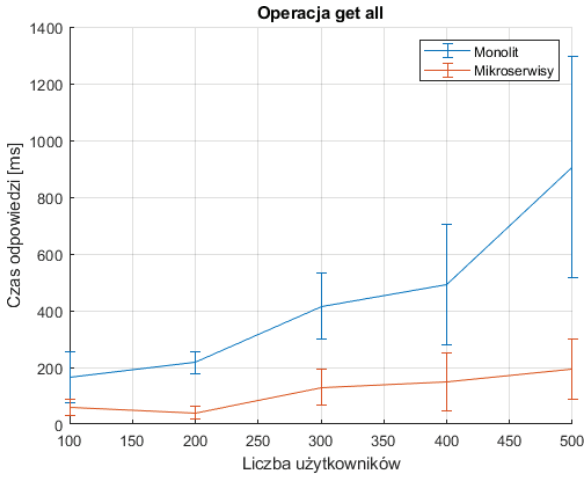


Rys. 4. Analiza przepustowości operacji utworzenia nowego konta użytkownika

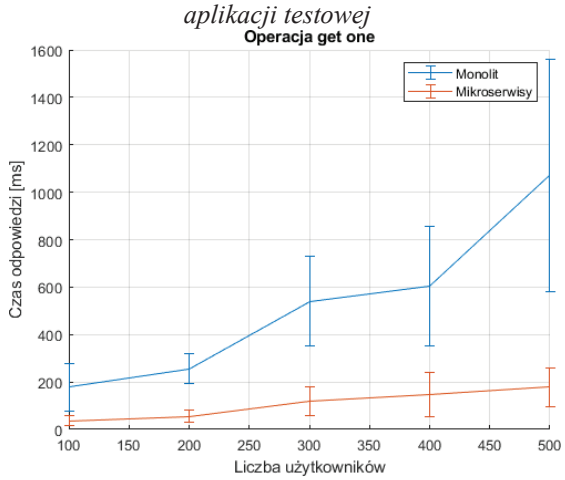
W trzecim scenariuszu analizowana przepustowość obu wariantów aplikacji pokazuje różnice w działaniu architektur. Na rysunku 3 przedstawiono zmianę czasu odpowiedzi wraz ze wzrostem użytkowników dla ograniczonych zasobów. Na rysunkach 4-8 przedstawiono szczegółowe wartości z podziałem na rodzaje operacji. Można zauważyć istotne różnice w ich działaniu dla obu typów architektur.



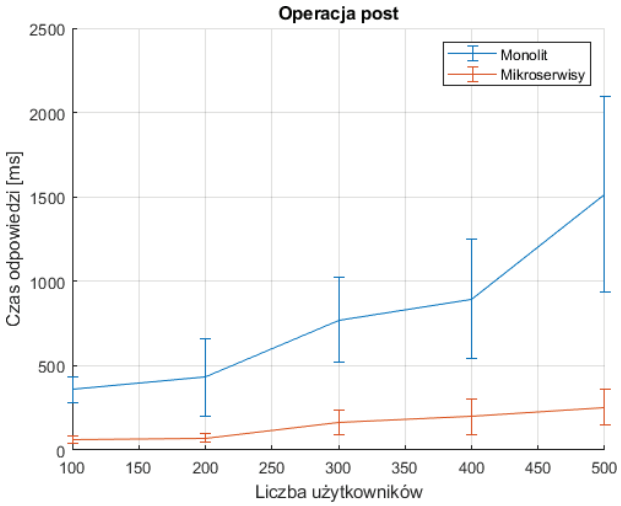
Rys. 5. Analiza przepustowości operacji uwierzytelnienia użytkownika w aplikacji



Rys. 6. Analiza przepustowości operacji typu get_all w aplikacji testowej



Rys. 7. Analiza przepustowości operacji typu get_one w aplikacji testowej



Rys. 8. Analiza przepustowości operacji typu post w aplikacji testowej

5. WNIOSKI

Na podstawie przestawionych badań oraz literatury omawianego przedmiotu sporządzona została poniższa lista wad, oraz zalet poszczególnych rozwiązań.

Architektura monolityczna jest rozwiązaniem prostym, przez co w nieskomplikowanych projektach, pozwala na szybkie wprowadzenie produktu na rynek. Sam proces wdrożenia również jest tutaj uproszczony, gdyż może się ograniczyć do uruchomienia (nawet) pojedynczego procesu na serwerze. Oferuje większe bezpieczeństwo komunikacji wewnątrz aplikacji, opierającej się na wywołaniach funkcji w porównaniu do zapytań poprzez sieć używanych w mikroserwisach. Architektura monolityczna to także większa wydajność przy małych i średnich projektach. Przeprowadzone eksperymenty wykazały, że architektura monolityczna osiąga niższe opóźnienia oraz większe przepustowości, przy jednoczesnym mniejszym zużyciu zasobów. Jednak wraz z rozwojem budowanego oprogramowania, ilość kodu, która jest tworzona przez zespół projektu, staje się trudna do opanowania, nawet dla najbardziej doświadczonych programistów. W takiej sytuacji nietrudno o błędy, a w tym przypadku, awaria pojedynczego elementu, oznacza zawieszenie funkcjonowania całej aplikacji. Architektura monolityczna posiada słabsze możliwości skalowania aplikacji. W warstwie sprzętowej, nawet jeśli tylko niektóre elementy systemu potrzebują większej ilości zasobów, to nie jesteśmy w stanie przydzielić ich tylko dla tych części aplikacji, które wymagają tego najbardziej.

Architektura mikroserwisów umożliwia punktowe skalowanie poszczególnych serwisów w zależności od tego, który aktualnie jest bardziej obciążony. Pozwala to na lepsze rozdysponowanie dostępnych zasobów. Dodanie funkcjonalności do systemu może zostać dokonane w stosunkowo szybkim czasie, gdyż najczęściej będzie oznaczało dodanie nowego serwisu bez dużej ingerencji w już istniejące. Błąd, który pojawi się w jednym z serwisów, nie musi oznaczać wyłączenia z użytkowania całej aplikacji, a jedynie pewnych funkcjonalności. Modularność i niezależność serwisów, pozwala na kontrolowanie długu technologicznego, przez większą elastyczność w doborze technologii i języków do konkretnego problemu. Systemy tworzone w tej architekturze przystosowane są do pracy pod dużym obciążeniem i są tworzone z myślą o jednoczesnej obsłudze wielu użytkowników. W przypadku mikroserwisów zapewnienie odpowiedniego poziomu bezpieczeństwa jest utrudnione ze względu na komunikację poprzez sieć. Przygotowanie architektury mikroserwisów jest procesem dużo bardziej skomplikowanym niż w przypadku monolitu. Każda usługa wymaga odpowiedniej konfiguracji, wdrożenia i skalowania. W związku z tym konieczne jest zaangażowania dużej liczby specjalistów z tej dziedziny. Znacznemu utrudnieniu ulega też proces monitorowania i zarządzania pracą takiej aplikacji i znajdowania przyczyn awarii. Ze względu na rozproszoną budowę systemu utrudnione jest zapewnienie transakcyjności. Konieczne staje się zagwarantowanie spójności danych w sposób programowy w przypadku usług, które aktualizują wiele baz danych. Proces testowania jest bardziej złożony niż w wersji monolitycznej, gdyż wymaga uruchomienia wszystkich usług powiązanych.

Jak widać, wybór architektury, w jakiej rozwijana będzie dana aplikacja, nie jest zadaniem trywialnym i należy wziąć pod uwagę mnóstwo aspektów. Przedstawione wyniki eksperymentalne są zgodne z doniesieniami literaturowymi, natomiast pokazują zachowanie się aplikacji w środowisku ograniczonych zasobów.

LITERATURA

[1] Al-Debagy O., Martinek P. "A Comparative Review of Microservices and Monolithic Architectures". *2018 IEEE 18th International Symposium on Com- putational Intelligence and Informatics (CINTI). 2018, s. 000149–000154. DOI: 10. 1109/CINTI.2018.8928192*.

[2] Buzato F. H. L., Goldman A., Batista A. "Efficient Resources Utilization by Different Microservices Deployment Models". *2018 IEEE 17th International Symposium on Network Computing and Applications (NCA). 2018, s. 1–4. DOI: 10.1109/NCA.2018.8548346*

[3] Fritzsch J. 2019. „Microservices Migration in Industry: Intentions, Strategies, and Challenges". *2019 IEEE International Conference of Softwre Maintenance and Evolution (ICSME), s. 481-490*

[4] Gos K., Zabierowski W. "The comparison of microservice and mono- lithic architecture". *2020 IEEE XVIth International Conference on the Per- spective Technologies and Methods in MEMS Design, s. 150–153.*

[5] „*International Standard – Software, systems and enterprise – Architecture Evaluation Framework*", ISO.IEC/IEEE 42030:2019(E)

[6] Saransig A., Freddy T. L. "Performance Analysis of Monolithic and Micro Service Architectures" *International Conference on Software Process Improvement 2019, s. 270–279. DOI: 10.1007/978-3-030-01171-0_25*

[7] Ueda T., Nakaike T., Ohara M. "Workload characterization for microservices". *2016 IEEE International Symposium on Workload Characteri- zation (IISWC). 2016, s. 1–10. DOI: 10.1109/IISWC.2016.7581269*.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

ŚRODOWISKO TESTOWE FUNKCJI CORE BASELINE DLA FRAMEWORKU IOTRUST

TEST ENVIRONMENT FOR CORE BASELINE FUNCTIONS FOR THE IOTRUST FRAMEWORK

Marcin Golański, Paweł Tomaszewicz, Mariusz Rawski

WEiTI IT, PW, Warszawa
{m.golanski, p.tomaszewicz, m.rawski}@tele.pw.edu.pl

DOI: 10.15199/59.2022.4.37

Streszczenie**: W artykule omówiono środowisko testowe dla sieci Interne-tu Rzeczy wykorzystującej szkielet IoTrust. Celem badań było testowanie funkcji związanych z bezpieczeństwem w oparciu o zalecenia NIST.**
Abstract**: This paper presents testing a proving environment for an Internet of Things network using the IoTrust framework. The purpose of the research was focused on testing securi-ty-related functions based on NIST recommendations.**

Słowa kluczowe**: Docker, IoT, Node-RED, testbed.**

Keywords**: Docker, IoT, Node-RED, testbed.**

1. WSTĘP

W referacie omówiono koncepcję środowiska programistyczno-sprzętowego służącego do testowania rozwiązań proponowanych przez szkielet IoTrust. Zadaniem, które należało przeprowadzić, była weryfikacja badanych i projektowanych elementów. Opracowane środowisko wykonawcze w całości nazwano środowi-skiem testowym testbed. Praca była współfinansowana przez Narodowe Centrum Badań i Rozwoju w ramach projektu CYBERSECIDENT/456446/III/NCBR/2020.

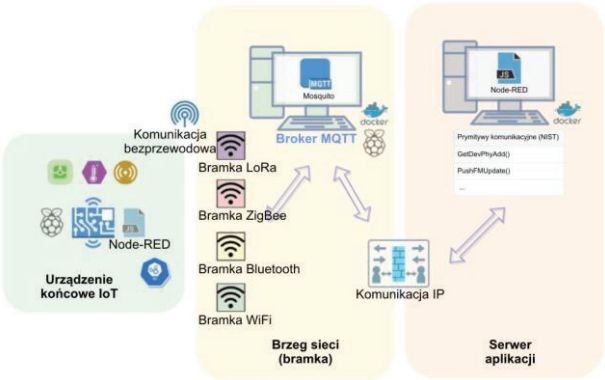
2. OPIS ŚRODOWISKA TESTOWEGO TESTBED

W ramach projektu przygotowano architekturę testową *testbed*, która pozwala zweryfikować funkcjonowanie wszystkich elementów Sieci IoT. Założenia zgodne z wymaganiami National Institute of Standards and Technology (NIST) powinny zostać zrealizowane w formie mikro-kodu aplikacji, który pozwoli na sprawdzenie poprawności funkcjonowania wskazanych modułów.

2.1. Architektura rozwiązania

Na rys. 1 przedstawiono podstawową strukturę *testbed*. Całość można podzielić na trzy warstwy: (a) serwer aplikacji, (b) brzeg sieci, który stanowi bramka, (c) urządzenie końcowe IoT. Jest zgodna z typowymi rozwiązaniami i zawiera wszystkie kluczowe elementy charakteryzujące Internet Rzeczy. Podział ten ułatwia dekompozycję zadania, tak aby można było szybko i łatwo przypisać komponenty sprzętowe do poszczególnych zadań. Ponadto wspomaga wybór rozwiązań w warstwie aplikacyjnej.

Warto zauważyć, że zrezygnowano z dużych, gotowych do użycia platform testowych IoT, gdyż w tym przypadku dążono do uzyskania jak największej elastyczności. Innym słowy każdy element przedstawiony na diagramie może zostać niskim kosztem wymieniony na inny, o ile zachowamy zgodność interfejsów i potraktujemy każdy z modułów jak czarną skrzynkę.



Rys. 1. Architektura testowa testbed

2.2. Serwer aplikacji

Pierwsza z warstw, tj. serwer aplikacji, stanowi centrum zarządzania testami. Prace rozpoczęto od zdefiniowania tzw. prymitywów komunikacyjnych np. *GetDevPhyAdd()*. Zostały one wyznaczone na podstawie wymagań NIST [1]. Tworzą zbiór funkcji, które testują bezpieczeństwo rozwiązania. Innymi słowy, pozwalają ocenić stopień zgodności z założeniami dotyczącymi aspektów zabezpieczenia platformy przed niepożądanymi działaniami.

Aby zrealizować prymitywy posłużono się platformą Node-RED [2]. Jest to narzędzie programistycznym do programowania wizualnego, opracowane pierwotnie przez IBM w celu łączenia urządzeń, sprzętowych interfejsów API i usług online Internetu Rzeczy. Platforma wykonawcza jest zbudowana z wykorzystaniem Node.js. Jednostki utworzone w Node-RED są przechowywane przy użyciu składni JSON.

Aplikacjami w Node-RED nazywa się przepływy składające się z węzłów połączonych ścieżkami. Na pulpicie roboczym węzły wejściowe mogą subskrybować usługi zewnętrzne, rozpocząć nasłuchiwanie danych na porcie TCP np. rozpocząć przetwarzanie żądań HTTP. Mogą także komunikować się z systemem operacyjnym.

Powyższy, dosyć skrótowo opisany mechanizm, ujawnia podstawowe zalety Node-RED. Stosowanie architektury przepływu z zapewnieniem możliwości generowania zdarzeń na każdym jego poziomie jest największą przewagą nad konkurencyjnymi rozwiązaniami. Należy zauważyć, że przepływy potrafią połączyć środowisko zewnętrzne z niskopoziomowymi funkcjami systemu operacyjnego. Jest to szczególnie istotne, dostarcza bowiem bardzo dobrych mechanizmów kontroli i separacji przy jednoczesnej możliwości odwoływania się bezpośrednio do urządzeń.

Platforma Node-RED została osadzona w środowisku wirtualizacyjnym Docker [3]. Stanowi ona zestaw produktów typu platforma jako usługa PaaS [4], które wykorzystują wirtualizację na poziomie systemu operacyjnego, aby dostarczać oprogramowanie w pakietach zwanych kontenerami. Są one odizolowane i łączą własne oprogramowanie, biblioteki i pliki konfiguracyjne. Ponieważ wszystkie kontenery korzystają z usług jednego jądra systemu operacyjnego, zużywają mniej zasobów niż tradycyjne maszyny wirtualne.

Podsumowując, serwer aplikacji został przygotowany tak, aby ułatwić szybką implementację wymaganej funkcjonalności Node-RED. Może ponadto zostać swobodnie zaimplementowany w dowolnym środowisku sprzętowym dzięki konteneryzacji.

Z punktu widzenia rozwoju aplikacji serwerowej przyjęto model dwukierunkowy. Dzięki opisanym wcześniej mechanizmom można zarządzać poziomo zmianami całego systemu oraz pionowo rozwojem aplikacji w kilku równoległych wersjach.

Jako system operacyjny dla maszyn wirtualnych zaproponowano lekki system obsługujący system wątków POSIX. Dobrymi kandydatami są niewielkie dystrybucje Linux tj. takie, które pozwalają na uruchomienie podstawowych funkcji typowego systemu operacyjnego w wersji jak najbardziej zbliżonej do rozwiązania typu *bare metal* [5]. Autorzy brali pod uwagę oprogramowanie Debian Linux lub Alpine Linux.

Wymienione dystrybucje spełniają dobrze wymagania dotyczące szybkości działania w środowisku o ograniczonych zasobach. Podstawowe wymagania w przypadku Debian Linux to do 512 MB pamięci RAM, 2 GB przestrzeni dyskowej. Z kolej dla Alpine Linux odpowiednio co najmniej 100 MB pamięci RAM i do 700 MB przestrzeni dyskowej.

Parametry wyraźnie pokazują, że maszyny wirtualne, które powinny zostać przygotowane dla prawidłowego funkcjonowania aplikacji serwerowej, będą optymalizowane pod kątem zasobów. Ostateczny wybór powinien zatem uwzględniać zakres dostępnego wsparcia technicznego, gotowe do użycia biblioteki systemowe i pozostałe źródła wspomagające szybkość procesu prototypowania.

Biorąc pod uwagę wymagania dla środowiska testowego, bardziej wskazany jest Debian Linux. W szczególności, jest to dystrybucja, na której bazuje system operacyjny mikrokomputera Raspberry Pi opisanego w dalszej części referatu.

2.3. Brzeg sieci

Druga warstwa tj. brzeg sieci stanowiła wyzwanie sprzętowe. Należy zwrócić uwagę, że komunikacja pomiędzy warstwami pierwszą i drugą odbywa się przy wykorzystaniu typowych protokołów internetowych. Przekraczając brzeg sieci, musimy liczyć się ze zmianą standardu. Na rys. 1 wyróżniono kolejno bramki: LoRa, ZigBee, Bluetooth, WiFi.

Praktycznie tylko ostatnia z wymienionych będzie korzystać z komunikacji TCP/IP. Pozostałe działają na zasadzie przekazywania enkapsulowanych pakietów. Wiąże się z tym wiele problemów. Najważniejszym z nich jest obniżenie dostępnego pasma komunikacji. Na przykład dla LoRa będzie to spadek transmisji GB na poziom pojedynczych bitów.

Ponieważ jako podstawowy układ elektroniczny węzła brzegowego zdecydowano się na wykorzystanie mikrokomputera ze złączem GPIO, dlatego najprostszą metodą wbudowania pożądanej metody komunikacji jest skorzystanie z szeroko dostępnych nakładek. Jednocześnie warto zauważyć, że kolejno montowane układy powinny posiadać biblioteki programistyczne, które można wykorzystać w systemie operacyjnym dla Raspberry Pi. Zgodnie z najlepszą wiedzą autorów, pożądane jest oprogramowanie tworzone w językach interpretowalnych tj. Python czy JavaScript. Dla obydwu przykładów są dostępne maszyny wykonawcze działające w środowisku Linux.

Zaczynając od pierwszej wymienionej technologii radiowej LoRa, dostępne są nakładki o różnej funkcjonalności. Autorzy sugerują wykorzystanie układu Dragino SX127X GPS HAT. W aspekcie oprogramowania opisywany układ można programować zarówno w języku Python jak i korzystając z modułów Node.

Kolejnym urządzeniem, które jest pożądane ze względu na wzbogacenia możliwości środowiska testowego, jest nakładka ZigBee. W tym przypadku paleta możliwości sprowadza się do urządzeń komunikujących się za pomocą portu USB. Po raz kolejny o wyborze układu dla bazowego węzła brzegowego decydowało oprogramowanie systemowe, umożliwiające szybkie zespolenie elektroniki wykorzystując dobrze zdefiniowane i opisane interfejsy komunikacyjne. W tym przypadku Raspberry Pi pełni rolę stacji bazowej, która łączy się z węzłami końcowymi za pomocą protokołu komunikacyjnego ZigBee. Do komunikacji bezprzewodowej i protokołu sieciowego multihop, można wykorzystać moduł XBee serii S2 firmy Digi International.

Podobnie jak w przypadku nakładki LoRa tutaj również można użyć zarówno bibliotek języka Python, jak i JavaScript. W szczególności pierwsza z nich została rozbudowana o obsługę Xbee dzięki indeksowi pakietów Pythona o nazwie PyPI. Należy wspomnieć również modułach Node-RED takich jak *node-red-contrib-xbee*, co pozwala stwierdzić, że moduł Zig-Bee wybrany do implementacji w architekturze środowiska testowego nie powinien stanowić przeszkody w jego szybkiej budowie.

Pozostałe dwa standardy komunikacyjne, czyli Bluetooth i WiFi są wbudowane sprzętowo i obsługiwane na-

tywnie przez mikrokomputer. Z punktu widzenia dewelopera systemu interfejs WiFi jest obsługiwany jak każdy typowy port do komunikacji TCP/IP. Nie wymagane są zatem specjalizowane biblioteki z oprogramowaniem systemowym. Z kolei Bluetooth wymaga jedynie zainstalowania biblioteki *libbluetooth-dev*, co daje możliwość korzystania ze skryptów systemowych jako interfejsu dla komunikujących się urządzeń.

Typowy węzeł IoT ma ograniczone zasoby energetyczne i dlatego pożądane jest, aby wysyłanie danych do serwera aplikacji odbywało się przy jak najmniejszym zużyciu baterii. Łączność z urządzeniem jest nieciągła, przeważnie charakteryzuje się gwałtownym natężeniem wymiany informacji. Dostępne pasmo jest ograniczone, biorąc pod uwagę możliwość wykorzystania dostępnej przepustowości. Pożądane jest, aby dane wysyłane kanałem informacyjnym były natychmiast aktualizowane. Jest to wymóg szczególnie trudny do spełnienia, jeśli uwzględnimy, że z jednej strony ważna jest energooszczędność komunikacji a z drugiej nie można pozwolić na błędy pomiarowe. Z punktu widzenia serwera aplikacji wymagane jest, aby wiadomości były wysyłane w możliwie najkrótszym czasie. Postulat ten wymusza wprost zastosowanie dodatkowego pośrednika komunikacji, który weźmie na siebie obowiązek szybkiego reagowania i uaktualniania informacji. Warto również zwrócić uwagę na reaktywny charakter takiego rozwiązania.

WiFi jak i Bluetooth są technologiami szerokopasmowymi. Wiąże się z tym zagadnienie buforowania transmisji wynikające z konieczności sprzęgania interfejsów o skrajnie różnych możliwościach transmisyjnych. Odpowiedzią na tak postawiony problem jest zastosowanie specjalnego pośrednika tj. brokera MQTT. Wspomniany element jest odpowiedzialny za obsługę wszelkich kolejek pakietów, dba o zabezpieczenie przed utratą informacji, innymi słowy pozwala łączyć rozmaite technologie komunikacyjne.

MQTT zapewnia trzy tryby transferu w zależności od wymaganej niezawodności: (a) QoS0 - przekazywanie danych bez gwarancji niezawodności, (b) QoS1 - przekazywanie danych z gwarancją niezawodności oraz (c) QoS2 (pełna obsługa aplikacji). QoS1 jest najbardziej podobny do HTTP z punktu widzenia niezawodności.

Aby zaimplementować broker wiadomości, wykorzystano pakiet Eclipse Mosquitto. Autorzy projektu opracowali niewielką serwerową implementację protokołu MQTT, co oznacza, że: (a) dołączone zostały tylko niezbędne funkcje, (b) funkcje zostały oprogramowane tak wydajnie, jak to tylko było możliwe, (c) elementy zewnętrzne zostały maksymalnie uproszczone.

Biorąc pod uwagę powyższe zalety w ostatecznej implementacji środowiska testowego, nie brano pod uwagę innych rozwiązań, stawiając na najlepiej udokumentowane oraz przetestowane przez społeczność programistów.

Podobnie jak w przypadku pierwszej warstwy postanowiono wykorzystać konteneryzacje rozwiązania. Jednak w tym przypadku całość została osadzona na mikrokomputerze klasy Rapsberry Pi.

2.4. Urządzenie końcowe

Trzecia warstwa to urządzenie końcowe IoT. Jako wzorcowy węzeł ponownie przygotowano płytkę Raspberry Pi. Została ona wyposażona w zestaw czujników w formie nakładki wykorzystującej złącze GPIO. W praktyce możliwa jest obsługa dowolnych sensorów, o ile posiadają one kompatybilny interfejs. Na platformie testowej mają one funkcje wyłącznie demonstracyjne. Nie zakłada się optymalizacji sprzętowej, koniecznej w przypadku instalacji urządzeń w docelowym środowisku działania.

Uwzględniając założenie o sprawnym prototypowaniu jako najbardziej odpowiedni i zarazem gotowy do użycia zestaw czujników wytypowano płytkę Sense Hat. Jest to zaawansowana bezprzewodowa płytka sensorowa, współpracująca z mikrokomputerem Raspberry Pi oraz zintegrowana z rozmaitymi czujnikami elektronicznymi takimi jak: czujnik temperatury, czujnik wilgotności, barometr, żyroskop, akcelerometr, magnetometr.

Lista możliwych pomiarów jest w tym przypadku na tyle duża, że węzeł końcowy zbudowany przy pomocy wymienionych komponentów powinien zagwarantować bogatą funkcjonalność, biorąc pod uwagę środowisko testowe. Dodatkowo dzięki wyświetlaczowi możliwe jest symulowanie działania aktuatora, co z kolei pozwoli sprawdzić bezpieczeństwo kanału komunikacji w dół od aplikacji Internetu Rzeczy do urządzenia końcowego. Autorzy zwracają uwagę, że pomimo iż około 80% rozproszonych aplikacji IoT bazuje na jednokierunkowej telemetrii, to pozostałe 20% ma jakościowo olbrzymie znaczenie, gdyż zazwyczaj pozwala zdalnie sterować elementami wykonawczymi.

Jeśli chodzi o sposób realizacji oprogramowania pokładowego węzła końcowego tak jak w przypadku pozostałych warstw zdecydowano o wykorzystaniu konteneryzacji Docker z realizacją API dostępowego za pomocą Node-RED.

Zbudowanie wysokowydajnej platformy opartej na małych komputerach jednopłytkowych jest wyzwaniem. W szczególności maszyny te mają bardzo ograniczoną wydajność strumieni wejście/wyjście. W wielu badaniach autorzy koncentrują się na problemie pobierania i wdrażania kontenerów Docker. Zakłada się, że węzły końcowe mają ograniczoną pojemność pamięci masowej, w związku z tym nie można oczekiwać, że będą przechowywały w pamięci podręcznej obrazy kontenerów dla wielu aplikacji.

Wdrażanie obrazów kontenerów może być nieakceptowalnie powolne, mowa o opóźnieniach rzędu wielu minut, w zależności od rozmiaru obrazu kontenera i stanu sieci. Cały czas należy pamiętać, że jeżeli z powodu wybranej technologii węzła końcowego będziemy korzystać

np. z czujników LoRa, to kanał komunikacyjny będzie miał pojemność wielokrotnie mniejszą od powszechnie stosowanych połączeń dla transmisji danych w sieciach IP. Opóźnienia, o których mowa są nie do przyjęcia w scenariuszach takich jak aplikacje Internetu Rzeczy wspomagane przetwarzaniem na krawędzi sieci, gdzie węzły końcowe mogą być mobilne, a nowe kontenery muszą być dynamicznie tworzone, gdy czujnik przemieszcza się do nowej lokalizacji. Dlatego też maksymalne skrócenie czasu wdrażania ma zasadnicze znaczenie dla zapewnienia satysfakcjonujących opinii o systemie z punktu widzenia użytkownika końcowego.

W literaturze można spotkać rozmaite techniki optymalizacyjne, których celem jest zwiększenie poziomu równoległości procesu wdrażania. Zazwyczaj pozwalają one zredukować czas wdrażania o 10-50% w zależności od zawartości i struktury obrazu kontenera oraz dostępnego pasma sieciowego. W połączeniu ze sobą sprawiają, że osadzanie kontenerów może być nawet do czterech razy szybsze niż natywna implementacja Dockera, pozostając przy tym całkowicie kompatybilnymi z niezmodyfikowanymi obrazami.

Pracując nad architekturą środowiska testowego, autorzy referatu rozważali możliwość skorzystania ze zoptymalizowanej kompilacji kodu Dockera, jednak ostatecznie nie zdecydowano się na to. Wynika to głównie z paradygmatu zachowania jak największej kompatybilności z powszechnie dostępnymi na rynku rozwiązaniami.

W pełni funkcjonalny węzeł końcowy, pracujący bez zakłóceń w środowisku testowym wymaga odpowiednio przygotowanego interfejsu programisty. W tym przypadku również zdecydowano się na eksperyment. W miejsce niskopoziomowych języków programowania, które nie pozwalają na szybkie prototypowanie na poziomie logiki, zdecydowano się na ponowne wykorzystanie platformy Node-RED.

Obrazy Node-RED przeznaczone dla kontenerów Docker są oparte na oficjalnych obrazach systemu operacyjnego Alpine. Jest to konieczne, aby utrzymać je tak małe, jak to tylko możliwe. Użycie Alpine Linux zmniejsza rozmiar budowanego obrazu, ale usuwa standardowe zależności, które są wymagane do natywnej kompilacji modułów. Aby temu zapobiec, trzeba rozszerzyć obraz Node-RED o brakujące pakiety bezpośrednio w działających kontenerach lub zbudować nowe obrazy. Również w tym przypadku proponowane jest zastosowanie standardowych obrazów i zainstalowanie modułów w zależności od potrzeb.

Po uruchomieniu Node-RED za pomocą Docker należy upewnić się, że dodane węzły lub przepływy nie zostaną utracone, jeśli kontener zostanie zniszczony. Dane aplikacji mogą być przechowywane przez zamontowanie katalogu danych na woluminie poza kontenerem. Oznacza to, że węzeł powinien zostać wyposażony w odpowiednią pamięć nieulotną. Może to być np. odpowiednia karta pamięci SD.

Ostatnią kwestią jest połączenie warstwy programistycznej, która zostaje zamknięta w przestrzeni kontenera z urządzeniami będącymi na wyposażeniu węzła. Jak wspomniano, w celu dodania płytki sensorowej wykorzystane zostanie złącze GPIO. Autorzy platformy Node-RED zapewnili odpowiednie biblioteki o nazwie *node-red-node-pi-gpiod*.

Aspekt wirtualizacji za pomocą kontenerów ma przewagę z punktu widzenia wymagań narzucanych przez NIST. Można założyć, że przy dobrze zaplanowanej architekturze możliwe będzie wydzielenie obszarów szczególnie istotnych, przechowujących dane wrażliwie, tak aby znalazły się w obrębie jednego kontenera. Wówczas nastąpi optymalizacja zasobów wynikająca wyłącznie z dobrze zaprojektowanego rozwiązania. Na przykład wymiana metody szyfrowania nie musi już oznaczać, że jesteśmy zmuszeni przesyłać duże ilości danych. Wystarczy wymienić wspomniany kontener lub wręcz zbudować na jego bazie nowy przesyłając informację o różnicach w kodzie. Jest to szczególnie ważne, jeżeli mówimy o węzłach końcowych, gdyż to właśnie one są często narażone na bezpośrednie działanie adwersarzy.

3. PODSUMOWANIE

Uwagi i wnioski, zamieszczone kolejno w ostatnich akapitach rozdziałów, zgodnie z najlepszą wiedzą autorów, świadczą o elastyczności zastosowanych rozwiązań. Pozwalają one skalować te same technologie, począwszy od bogatych w zasoby serwerów w warstwie pierwszej, a skończywszy na mikro-platformach w warstwie trzeciej. Unifikacja taka ma duże znaczenie, w szczególności przy platformach testowych, gdyż pozwala uniknąć wszelkich problemów wynikających z niespójności sprzętowo-aplikacyjnej modułów komunikujących się w ramach jednego środowiska.

LITERATURA

[1] Iorga, Michaela, Larry Feldman, Robert Barton, Michael Martin, Nedim Goren, Charif Mahmoudi. 2017. *The nist definition of fog computing*. No. NIST Special Publication (SP) 800-191 (Draft). National Institute of Standards and Technology.

[2] Lekić, Milica, Gordana Gardašević. 2018. „IoT sensor integration to Node-RED platform". In *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, IEEE: 1-5.

[3] Turnbull, James. 2014. *The Docker Book: Containerization is the new virtualization*.

[4] Dua, Rajdeep, A. Reddy Raja, Dharmesh Kakadia. 2014. „Virtualization vs containerization to support paas". In *2014 IEEE International Conference on Cloud Engineering*, IEEE: 610-614.

[5] Bernstein, David. 2014. „Containers and cloud: From lxc to docker to Kubernetes". *IEEE Cloud Computing 1*, no. 3: 81-84.

ZASTOSOWANIE KONCEPCJI SDN W WARSTWIE TRANSPORTOWEJ SIECI IMS/NGN

APPLICATION OF THE SDN CONCEPT IN IMS/NGN TRANSPORT STRATUM

Kamil Bachorski[1]; Sylwester Kaczmarek[2]; Maciej Sac[3]

[1] ADVA Optical Networking, Gdynia, kbachorski@adva.com
[2] Politechnika Gdańska, Wydział ETI, Gdańsk, kasyl@eti.pg.edu.pl
[3] Politechnika Gdańska, Wydział ETI, Gdańsk, maciej.sac@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.38

Streszczenie: **Koncepcja sieci Next Generation Network, wykorzystująca w warstwie usług architekturę IP Multimedia Subsystem, zakłada centralizację funkcji sterowania zasobami sieciowymi w warstwie transportowej. Analogiczne założenie jest podstawą koncepcji Software Defined Networks. Na podstawie tej analogii, zaproponowano integrację obu koncepcji, przedstawiono architekturę zintegrowanej sieci, scenariusze sterowania zasobami oraz przygotowano środowisko zwirtualizowanych, otwartych projektów realizujących funkcjonalność obu koncepcji. Następnie przeprowadzono ich integrację poprzez dostosowanie oprogramowania oraz wykonano testowanie zintegrowanej sieci.**

Abstract: **The concept of the Next Generation Network, using the IP Multimedia Subsystem architecture in service stratum, assumes centralization of network resource control functions in transport stratum. A similar assumption is the basis for the Software Defined Networks concept. Based on this analogy, integration of both concepts was proposed, architecture of the integrated network was presented along with resource control scenarios. Moreover, an environment of virtualized, open source projects implementing the functionality of both concepts was prepared. Their integration was performed by adapting the software and followed by testing the integrated network.**

Słowa kluczowe: **IMS, NGN, ONOS, SDN.**

Keywords: **IMS, NGN, ONOS, SDN.**

1. WSTĘP

Dynamiczny rozwój usług telekomunikacyjnych determinuje konieczność udoskonalenia pracy systemów telekomunikacyjnych przeznaczonych do ich świadczenia oraz optymalizację wykorzystania zasobów sieciowych. W te założenia wpisuje się koncepcja Software Defined Networks (SDN) [1], która umożliwia zwiększenie poziomu automatyzacji sieci poprzez możliwość sterowania zasobami w sposób programowy. Sprzyja temu założenie rozdzielenia płaszczyzny sterowania od płaszczyzny danych w ramach urządzeń sieciowych. Zgodnie z tą koncepcją, w ramach płaszczyzny sterowania funkcjonują sterowniki SDN zarządzające zasobami sieciowymi, z kolei płaszczyzna danych zawiera programowalne przełączniki. Architektura SDN zapewnia otwartość rozwiązań sterowników i przełączników oraz protokołów stosowanych do komunikacji pomiędzy tymi urządzeniami. Popularne sterowniki SDN to między innymi Open Networking Operating System (ONOS) [2] czy Floodlight [3]. Natomiast jednym z najbardziej rozpowszechnionych protokołów służących do wymiany informacji pomiędzy płaszczyzną sterowania i danych, w celu zarządzania i sterowania przepływami, jest protokół OpenFlow [4].

Centralizacja funkcji sterowania zasobami sieciowymi jest również założeniem koncepcji sieci Next Generation Network (NGN) [5] wykorzystującej w warstwie usług architekturę IP Multimedia Subsystem (IMS) [6], w dalszej części artykułu nazywanej po prostu siecią IMS/NGN. Podstawowym modelem odniesienia dla niniejszej pracy jest architektura sieci IMS/NGN opracowana przez International Telecommunication Union Telecommunication Standardization Sector (ITU-T). Sieć IMS/NGN składa się z dwóch zasadniczych warstw. Pierwszą z nich jest warstwa usługowa (service stratum), w ramach której operują funkcje architektury IMS, komunikujące się między sobą przy wykorzystaniu protokołów SIP i Diameter. Główną funkcją IMS jest Call Session Control Function (CSCF). Steruje ona procesem sygnalizacji i scenariuszami począwszy od użytkownika końcowego mającego dostęp do usług zarówno w sieci IMS/NGN jak i innych sieci. Umożliwia także przekazywanie informacji związanej z usługami do drugiej z warstw, warstwy transportowej (transport stratum), a konkretnie do elementu Resource and Admission Control Function (RACF) przy wykorzystaniu protokołu Diameter. RACF jest centralnym punktem sterowania zasobami sieciowymi, zdolnym do interpretacji wiadomości w ramach protokołu Diameter oraz na ich podstawie zarządzania i sterowania zasobami warstwy transportowej.

Przegląd rozwiązań dotyczących koncepcji SDN oraz IMS/NGN, analiza ich architektur oraz funkcjonalności, umożliwiła zaproponowanie ich zintegrowania oraz określenie warunków ich docelowej współpracy. Propozycja zakłada przede wszystkim wykorzystanie sterownika SDN w warstwie transportowej sieci IMS/NGN jako elementu sterującego zasobami sieciowymi w postaci programowalnych przełączników. Założono również pozostawienie funkcjonalności elementu RACF zdolnego do interpretacji wiadomości Diameter otrzymywanych z Proxy-CSCF (P-CSCF), a następnie na ich podstawie wywoływanie API wybranego sterownika SDN.

Przeglądając prace w tym obszarze działalności badawczo-rozwojowej, można odnaleźć przykłady koncepcji integracji oraz prac prowadzonych w celu zapewnienia współpracy koncepcji IMS/NGN oraz SDN. Ich

wspólnymi cechami są: skupienie się na architekturze IMS (bez uwzględniania standardów ITU-T dotyczących sieci NGN) oraz brak dostępu do kodu źródłowego oprogramowania. Przykładem jest projekt OSIMS zrealizowany w Uniwersytecie w Patras [7]. Autorzy przygotowali środowisko testowe, w ramach którego zintegrowali funkcjonalność sieci IMS oraz SDN i zaimplementowali komunikację pomiędzy nimi. Do realizacji funkcjonalności rdzenia sieci IMS wykorzystali projekt Open IMS Core. Natomiast sieć SDN zrealizowali z wykorzystaniem sterownika Floodlight sterującego programowalnymi przełącznikami Open vSwitch. Projekt analogiczny do OSIMS, wykorzystujący te same otwarte oprogramowanie, został utworzony przez ośrodki NIPiP oraz UH [8].

Mając na uwadze brak dostępnych do wykorzystania podczas badań rozwiązań, autorzy tego artykułu podjęli decyzję o realizacji własnego projektu integracji koncepcji IMS/NGN oraz SDN, z uwzględnieniem nie tylko standardów dotyczących IMS (3GPP, ETSI, itd.), ale także sieci NGN (głównie ITU-T). Projekt realizowany na Politechnice Gdańskiej (PG), opisywany w tym artykule, zakłada wykorzystanie częściowo innych niż opisano w [7,8] otwartych rozwiązań oprogramowania. Jako sterownik SDN wykorzystano Open Networking Operating System (ONOS) [2] rozwijany przez konsorcjum Open Networking Foundation (ONF). Natomiast funkcjonalność sieci IMS jest realizowana przez aplikację CDiameterPeer [9] umożliwiającą między innymi generację wiadomości protokołu Diameter. Aplikacja CDiameterPeer jest częścią projektu Open IMS Core, co umożliwia dalszą integrację z oprogramowaniem tego projektu. W projekcie realizowanym na PG podobnie jak w przypadku projektów z pozostałych ośrodków, jako klaster programowalnych przełączników funkcjonuje Open vSwitch [10].

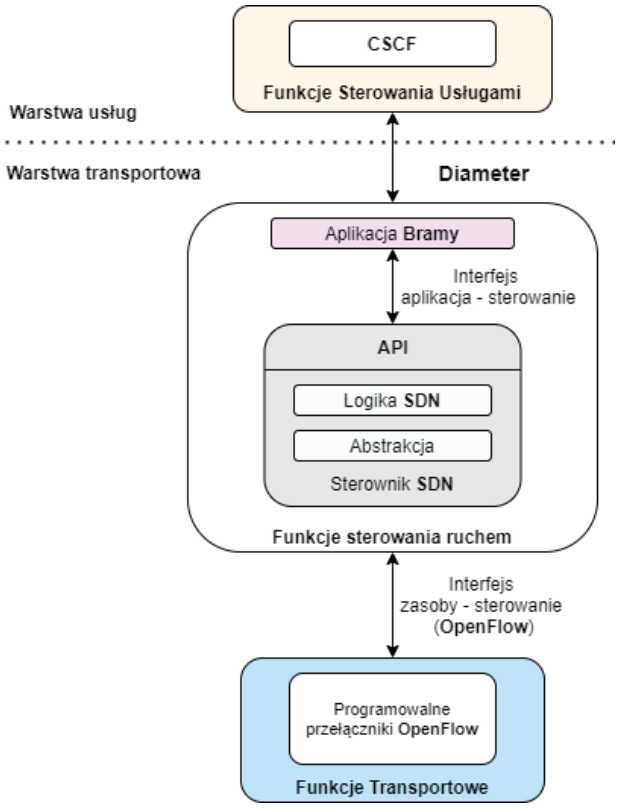
Niniejszy artykuł zawiera opis wyników prac nad omawianym zagadnieniem realizowanych na PG. Zasadnicza część artykułu składa się z dwóch kolejnych rozdziałów. W ramach drugiego rozdziału opisano propozycję architektury sieci IMS/NGN wykorzystującej w warstwie transportowej koncepcję SDN, założenia współpracy obu koncepcji oraz skrócony opis scenariuszy realizacji usług dla zintegrowanej sieci. Zawiera on również konieczne etapy niezbędne do realizacji zaproponowanej koncepcji. Trzeci rozdział dotyczy szczegółów związanych z praktyczną implementacją bazującą na otwartym oprogramowaniu i wynikających stąd trudności oraz konsekwencji. W rozdziale trzecim zamieszczono także wyniki przeprowadzonych testów. Artykuł kończy rozdział czwarty w postaci krótkiego podsumowania oraz planowanego zakresu dalszych prac.

2. KONCEPCJA ZASTOSOWANIA SDN W WARSTWIE TRANSPORTOWEJ SIECI IMS/NGN

Przeprowadzona analiza koncepcji SDN oraz IMS/NGN, a także rozwiązań wykorzystywanych w ramach innych projektów umożliwiła rozpoczęcie prac nad autorską koncepcją integracji IMS/NGN i SDN. Efektem prac jest propozycja architektury zintegrowanej sieci oraz realizowanych w niej scenariuszy sterowania zasobami, które opisano w niniejszym rozdziale. Jest to podstawa do prac nad praktyczną realizacją koncepcji opisaną w rozdziale trzecim.

2.1. Architektura sieci IMS/NGN wykorzystująca koncepcję SDN

Podstawowym i pierwszym zadaniem projektu było określenie miejsca dla elementów funkcjonalnych sieci SDN w sieci IMS/NGN oraz ustalenie zasad współpracy pomiędzy elementami sieci IMS/NGN oraz sieci SDN. Efektem rozważań i przeprowadzonych analiz było przygotowanie propozycji architektury sieci IMS/NGN z uwzględnieniem w warstwie transportowej składników sieci SDN (rys. 1).



Rys. 1. Propozycja architektury IMS/NGN z wykorzystaniem koncepcji SDN w warstwie transportowej

Z punktu widzenia projektu najistotniejszym elementem warstwy usług sieci IMS/NGN jest zbiór serwerów CSCF, a szczególnie serwer P-CSCF. Przekazuje on wiadomości dotyczące parametrów żądanych usług do warstwy transportowej przy wykorzystaniu protokołu Diameter. Bezpośrednim odbiorcą tych wiadomości jest element RACF widoczny z poziomu warstwy usług jako Aplikacja Bramy (AB). Funkcjonalność AB obejmuje interpretację wiadomości protokołu Diameter, a następnie wywoływanie API sterownika SDN, który na podstawie dostarczonych danych odpowiednio wysterowuje zasoby sieciowe. Aplikacja Bramy oraz sterownik SDN wraz ze swoimi wbudowanymi składnikami odpowiedzialnymi między innymi za logikę sterowania SDN oraz warstwę abstrakcji, pod którą ukryte są zasoby sieciowe, zostały ujęte w proponowanej architekturze jako funkcje sterowania ruchem i ulokowane w warstwie transportowej. Zbiór funkcji sterowania ruchem komunikuje się przy wykorzystaniu interfejsu zasoby–sterowanie (protokół OpenFlow) z funkcjami transportowymi zawierającymi programowalne przełączniki.

Spośród wskazanych styków, należy nadmienić, że styk pomiędzy funkcją CSCF oraz Aplikacją Bramy jest zestandaryzowany i można wykorzystać do jego realizacji (po odpowiednim dostosowaniu) dostępne aplikacje oferujące komunikację z wykorzystaniem protokołu Diameter. Taka sama sytuacja ma miejsce z interfejsem zasoby-sterowanie i wymianą informacji przy użyciu protokołu OpenFlow. Jedynym stykiem, który nie jest zestandaryzowany, jest interfejs aplikacja-sterowanie pomiędzy Aplikacją Bramy oraz API sterownika SDN. Jest to główny element prac, poza Aplikacją Bramy, który należy zaimplementować w celu uzyskania pełnej funkcjonalności przedstawionej architektury.

Dla zaproponowanej architektury zostały opracowane scenariusze realizacji usług. Zawierają one dwa przypadki uwzględniające różny stan dostępności zasobów w sieci i zostały opisane w punkcie 2.2.

2.2. Scenariusze realizacji usług w zintegrowanej sieci

Dla zobrazowania komunikacji na wskazanych w zintegrowanej architekturze interfejsach, przygotowane zostały dwa scenariusze realizacji usług. Bazując na nich można zaimplementować funkcjonalność Aplikacji Bramy integrującej koncepcje IMS/NGN oraz SDN.

Przygotowane scenariusze realizacji usług umożliwiają przede wszystkim lepszy pogląd na etapy implementacji oprogramowania i testowania sekwencji wiadomości wykorzystywanych w projekcie protokołów (Diameter i OpenFlow). Wraz z przygotowaną architekturą środowiska stanowią podstawę do realizacji koncepcji integracji IMS/NGN z SDN. Pierwszy z przygotowanych scenariuszy zakłada przypadek braku zasobów dla obsługi połączenia. Natomiast drugi z nich zakłada dostępność zasobów i zakończone powodzeniem zestawienie połączenia pomiędzy dwoma urządzeniami końcowymi sieci IMS/NGN. Scenariusz ten został przedstawiony na rys. 2.

W obu scenariuszach uwzględniono elementy sieci IMS takie jak P-CSCF i Serving-CSCF (S-CSCF) oraz użytkowników końcowych sieci IMS/NGN (User Equipment) oznaczonych jako UE 1 i UE 2. Ponadto uwzględniono w nim elementy sieci SDN – sterownik oraz programowalne przełączniki. Scenariusz obejmuje kompleksowy proces realizacji połączenia od generacji jego żądania do rezerwacji zasobów, a następnie ich zwolnienia i w efekcie rozłączenia. Zawiera wiadomości sygnalizacyjne protokołów SIP oraz Diameter. Uwzględnia również proces sterowania zasobami sieciowymi przy wykorzystaniu OpenFlow z wykorzystaniem operacji na tablicach przepływu. Scenariusz nie uwzględnia wywoływania API sterownika SDN przez Aplikację Bramy, ponieważ jest to styk wewnątrz elementu (rys. 1).

W punkcie 2.3 opisano kolejny etap realizacji koncepcji dotyczący wyboru otwartego oprogramowania umożliwiającego implementację założonej funkcjonalności obu integrowanych sieci. Otwarte oprogramowanie wymagało adaptacji i z uwagi na jej złożoność zostało podzielone na etapy.

2.3. Wybór oprogramowania i etapy realizacji koncepcji

Rozpoczęcie prac w ramach opisywanego projektu wymagało wybrania otwartego oprogramowania, umożliwiającego realizację funkcjonalności sieci IMS/NGN oraz SDN podlegającej integracji. Pierwszym wybranym projektem open source był CDiameterPeer. Jest to aplikacja działająca w topologii peer to peer. Udostępnia możliwość wysyłania oraz interpretacji wiadomości protokołu Diameter. Zatem wpisuje się w założenia projektu, jako element P-CSCF generujący wiadomości Diameter z warstwy usług oraz jako składnik Aplikacja Bramy. Sieć SDN została zbudowana z dwóch otwartych projektów. Pierwszym z nich jest wspomniany sterownik SDN – ONOS. Zarządza on zasobami sieciowymi realizowanymi przez programowalne przełączniki Open vSwitch emulowane w środowisku Mininet. Komunikacja na styku SDN odbywa się przy wykorzystaniu protokołu OpenFlow.

Ze względu na duży nakład prac związanych z realizacją projektu oraz ich złożoność, prace te zostały podzielone na następujące cztery etapy (które zostaną opisane szerzej w dalszej części artykułu):

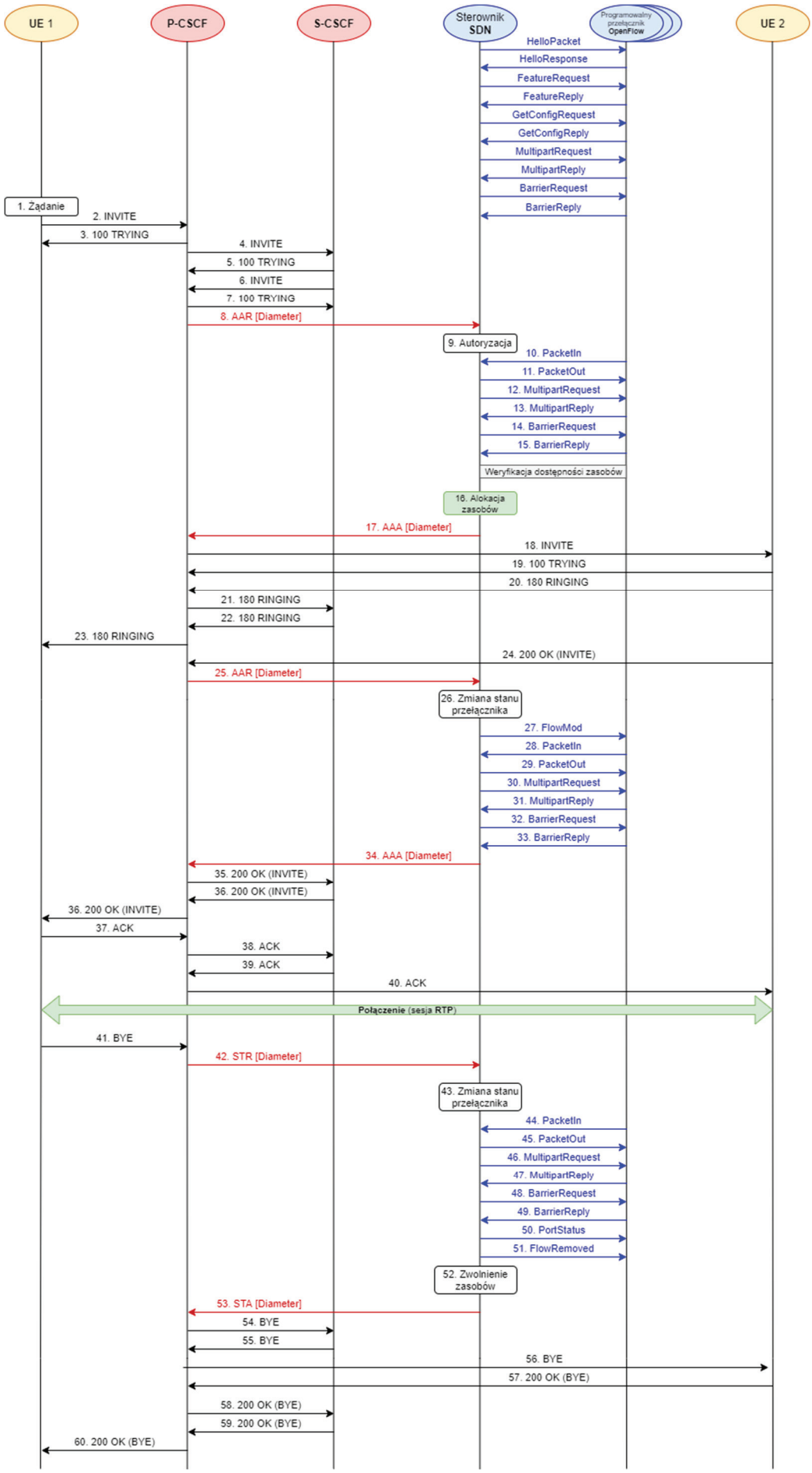
1. Przygotowanie środowiska IMS/NGN i SDN.
2. Konfiguracja podstawowej komunikacji Diameter oraz OpenFlow.
3. Implementacja styku z protokołem Diameter.
4. Implementacja styku z protokołem HTTP REST/JSON.

3. REALIZACJA KONCEPCJI ORAZ TESTY OPROGRAMOWANIA

Na podstawie zaproponowanej koncepcji (rys. 1) oraz scenariuszy rozpoczęto prace nad praktyczną implementacją projektu według architektury przedstawionej na rys. 3. Kolejne etapy prac przedstawiono w tym rozdziale.

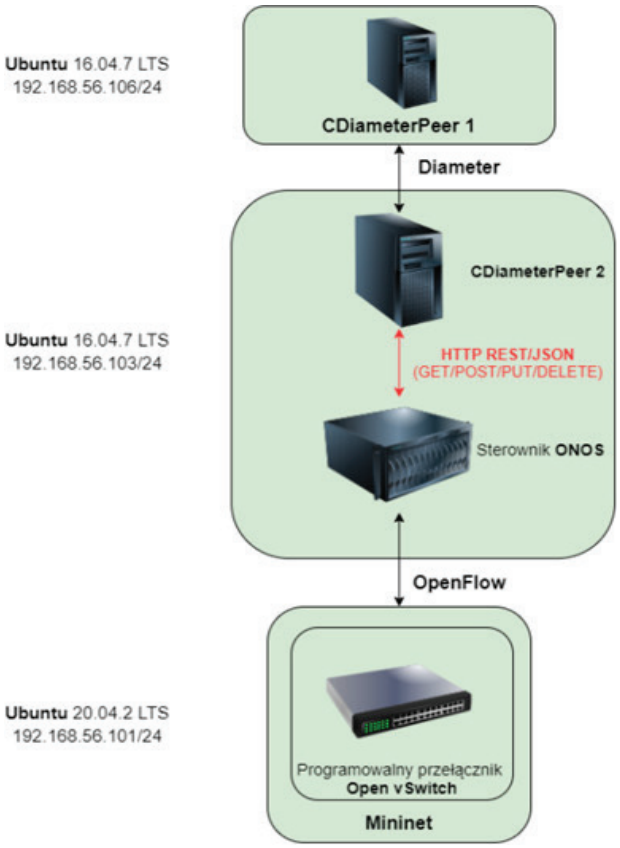
3.1. Przygotowanie środowiska IMS/NGN i SDN

Struktura przygotowywanego środowiska została zobrazowana na rys. 3. Etap przygotowania rozpoczęto od utworzenia trzech maszyn wirtualnych z poziomu aplikacji hipernadzorcy - Oracle VM VirtualBox i zainstalowania na nich dystrybucji systemu Linux. Systemom zapewniono dostęp do Internetu oraz skonfigurowano wewnętrzną sieć IP do ich wzajemnej komunikacji. Na pierwszej z maszyn, oznaczonej jako CDiameterPeer 1 została zainstalowana dystrybucja Ubuntu w wersji 16.04.7 LTS. System zawiera zainstalowane oprogramowanie z projektu CDiameterPeer, które służy do generowania wiadomości Diameter i realizuje wybraną funkcjonalność serwera P-CSCF. Druga z maszyn wirtualnych zawiera identyczny system operacyjny co pierwsza, na którym zainstalowano dwa otwarte oprogramowania. Podobnie jak na pierwszej maszynie jest to CDiameterPeer, oznaczony cyfrą 2, działający jako składnik Aplikacji Bramy oraz sterownik ONOS. Trzecia z maszyn działa bazując na systemie operacyjnym Ubuntu 20.04.2 LTS. Zainstalowano w niej środowisko Mininet [11] emulujące programowalne przełączniki Open vSwitch.



Rys. 2. Scenariusz realizacji usługi w zintegrowanej sieci dla przypadku dostępności zasobów

W ten sposób utworzone środowisko (rys. 3) wymaga komunikacji przy wykorzystaniu protokołów wyższych warstw na trzech stykach. Na pierwszym ze styków wykorzystywany jest protokół Diameter służący do komunikacji pomiędzy CDiameterPeer 1 oraz CDiameterPeer 2. Na drugim ze styków używany jest protokół HTTP REST/JSON wykorzystywany do wywoływania API sterownika ONOS przez CDiameterPeer 2. Na trzecim ze styków, styku sieci SDN, wykorzystywany jest protokół OpenFlow służący do zarządzania przełącznikami Open vSwitch przez sterownik ONOS. W następnym podrozdziale opisano proces uruchomienia podstawowej komunikacji na dwóch stykach - Diameter oraz OpenFlow.



Rys. 3. Architektura przygotowanego środowiska

3.2. Konfiguracja podstawowej komunikacji Diameter oraz OpenFlow

Przygotowanie styku z protokołem Diameter w pierwszej kolejności polegało na zainstalowaniu i uruchomieniu aplikacji CDiameterPeer 1 i 2 zgodnie z rys. 3. Obie aplikacje z założenia działają w topologii peer to peer. Są napisane w języku C++ i do ich uruchomienia niezbędny był kompilator GCC zainstalowany wcześniej w zasobach systemu. Aplikacja po uruchomieniu udostępnia listing z debuggera umożliwiający identyfikację poprawności działania poszczególnych jej procesów.

Konfiguracja poszczególnych peerów Diameter odbyła się poprzez modyfikację plików konfiguracyjnych przechowywanych w formacie XML. W każdym z plików trzeba było przede wszystkim wskazać Full Quality Domain Name (FQDN) w formacie adresu IP oraz identyfikator dostawcy o wartości 11502 powiązanego z identyfikatorem aplikacji 16777235. Ostatnie dwie wartości, zgodnie ze standardem ITU-T [12], służą do oznaczenia styku Diameter Rs umiejscowionego pomiędzy P-CSCF oraz RACF. W pliku konfiguracyjnym wskazano również detale dotyczące sąsiedniego peera Diameter w postaci jego adresu IP oraz portu warstwy 4 ISO/OSI.

Modyfikacja plików XML po dwóch stronach (CDiameterPeer 1 i 2) umożliwiła nawiązanie podstawowej komunikacji Diameter pomiędzy peerami. Poprawność tego zadania stwierdzono bazując na listingu z debuggera (rys. 4) po dwóch stronach, zawierającego informację o nawiązaniu połączenia i otwarciu portu 3868 (domyślny port Diameter). Potwierdzenie uzyskano również w aplikacji sniffera sieciowego - Wireshark. Po uruchomieniu monitorowania po dwóch stronach, przechwycono wymianę pakietów Capabilities-Exchange służących do negocjacji parametrów komunikacji Diameter (w tym obsługiwanych aplikacji protokołu Diameter).

```
0(9670) --- Peer List: ---
0(9670)  S[Closed] 192.168.56.103:3868 D[ ]
0(9670) ------------------
0(9671) INFO:accept_connection(): new tcp connection accepted!
0(9671) INFO:accept_connection(): Receiver process forked [9693]
0(9693) INFO:receiver_process(): [4] Receiver process starting up...
0(9693) DBG:sm_process(): Peer 192.168.56.103  State Closed    Event
0(9693) --- Peer List: ---
0(9693)  S[R_Open] 192.168.56.103:3868 D[ ]
0(9693)          [16777235,11502]
0(9693)          [16777235,0]
0(9693)          [16777235,10415]
```

Rys. 4. Nawiązanie połączenia pomiędzy peerami Diameter

Następnie przystąpiono do przygotowania styku z protokołem OpenFlow (rys. 3). Początkowo zainstalowano sterownik ONOS w wersji 2.6.0.b przy wykorzystaniu repozytorium GitHub. Do instalacji wymagane było wyposażenie systemu Ubuntu w moduły języka Python, system kontroli wersji Git oraz narzędzia do obsługi formatu kompresji ZIP i biblioteki OSGI, do czego wykorzystano narzędzie apt-get. Warto nadmienić, że po zainstalowaniu wszystkich niezbędnych modułów, napotkano na problemy podczas kompilacji i budowy obrazu sterownika ONOS związane ze współpracą z daną wersją systemu Linux. Przetestowano najnowsze wersje Ubuntu, tzn. 20 oraz 18, jednak nie udało się na nich tego wykonać. Ostatecznie obraz został zbudowany na wersji Ubuntu 16. Opisane problemy kompatybilności nie zostały uwzględnione w dokumentacji sterownika ONOS.

Po zbudowaniu obrazu sterownika ONOS, przystąpiono do jego konfiguracji. Ustawiono między innymi adres IP oraz uruchomiono na nim aplikacje z poziomu udostępnianego interfejsu CLI. Są one niezbędne do dalszej pracy i umożliwiają m. in. wsparcie protokołu OpenFlow czy włączenie interfejsu REST API. Należało również uruchomić przynajmniej aplikację reaktywnego przekazywania czyli mechanizm instalowania na żądanie przepływów w tablicach programowalnych przełączników.

Drugim przygotowanym elementem sieci SDN były programowalne przełączniki Open vSwitch w wersji 2.3.0

emulowane w środowisku Mininet w wersji 2.13.3. Do instalacji aplikacji Mininet wykorzystano ponownie repozytorium GitHub. Open vSwitch jest wbudowany w Mininet domyślnie. Należało go jedynie uruchomić i sprawdzić status odpowiednich procesów z poziomu linii komend systemu Linux.

Proces integracji Open vSwitch i Mininet ze sterownikiem ONOS (rys. 5) jest zautomatyzowany. W wierszu poleceń systemu Linux wydano komendę mn z argumentami definiującymi między innymi adres IP sterownika z domyślnym portem 6653, wersję protokołu OpenFlow 1.3 (przy najwyższej obsługiwanej wersji 1.5 były problemy ze współpracą Mininet ze sterownikiem ONOS, o czym nie poinformowano w dokumentacji) oraz docelową topologię sieci przełączników i ich liczbę (dwuwymiarowa topologia toroidalna o łącznej liczbie dziewięciu przełączników). Po wykonaniu tej komendy otrzymano listing z informacją o powstaniu żądanej topologii oraz pomyślnym połączeniu ze sterownikiem na drugiej maszynie. Następnie w wierszu poleceń Mininet przetestowano komunikację pomiędzy zwirtualizowanymi przełącznikami przez wykonanie testu protokołu ICMP (polecenie pingall; rys. 5). W tym samym czasie w interfejsie przeglądarkowym sterownika ONOS zaobserwowano utworzenie żądanej topologii. Za pomocą programu Wireshark przechwycono również pakiety OpenFlow służące do komunikacji pomiędzy sterownikiem i przełącznikami (rys. 6). Proces komunikacji i kolejność wymiany specyficznych wiadomości były zgodne z oczekiwaniami.

```
mininet@mininet-vm:~$ sudo mn --controller remote,ip=192.168.56.103
protocols=OpenFlow13 --topo torus,3,3
*** Creating network
*** Adding controller
Connecting to remote controller at 192.168.56.103:6653
*** Adding hosts:
h1x1 h1x2 h1x3 h2x1 h2x2 h2x3 h3x1 h3x2 h3x3
*** Adding switches:
s1x1 s1x2 s1x3 s2x1 s2x2 s2x3 s3x1 s3x2 s3x3
*** Adding links:
(h1x1, s1x1) (h1x2, s1x2) (h1x3, s1x3) (h2x1, s2x1) (h2x2, s2x2) (h2
s3x1) (h3x2, s3x2) (h3x3, s3x3) (s1x1, s1x2) (s1x1, s2x1) (s1x2, s1x
s1x3, s1x1) (s1x3, s2x3) (s2x1, s2x2) (s2x1, s3x1) (s2x2, s2x3) (s2x
2x1) (s2x3, s3x3) (s3x1, s1x1) (s3x1, s3x2) (s3x2, s1x2) (s3x2, s3x3
3x3, s3x1)
*** Configuring hosts
h1x1 h1x2 h1x3 h2x1 h2x2 h2x3 h3x1 h3x2 h3x3
*** Starting controller
c0
*** Starting 9 switches
s1x1 s1x2 s1x3 s2x1 s2x2 s2x3 s3x1 s3x2 s3x3 ...
*** Starting CLI:
mininet> pingall
*** Ping: testing ping reachability
h1x1 -> h1x2 h1x3 h2x1 h2x2 h2x3 h3x1 h3x2 h3x3
h1x2 -> h1x1 h1x3 h2x1 h2x2 h2x3 h3x1 h3x2 h3x3
                                  ...
h3x3 -> h1x1 h1x2 h1x3 h2x1 h2x2 h2x3 h3x1 h3x2
*** Results: 0% dropped (72/72 received)
```

Rys. 5. Wynik testowania środowiska Mininet

Osiągnięte rezultaty potwierdziły poprawność podstawowej komunikacji na stykach z protokołami Diameter oraz OpenFlow. W następnym podrozdziale skupiono się na oprogramowaniu działania styku z protokołem Diameter (rys. 3) zgodnie z zaleceniem ITU-T [12] dla styku Rs umiejscowionego pomiędzy P-CSCF i RACF.

3.3. Implementacja styku z protokołem Diameter

Po uruchomieniu i przetestowaniu podstawowej funkcjonalności styku Diameter, należało zaimplementować jego działanie na potrzeby sterowania zasobami transportowymi zgodnie z zaleceniem ITU-T [12]. Jak zilustrowano na rys. 2, na styku tym wymieniane są następujące wiadomości: żądania rezerwacji zasobów transportowych na potrzeby usługi (wstępna rezerwacja i jej potwierdzenie – wiadomości AAR) oraz odpowiedzi na te żądania (wiadomości AAA), żądania zwolnienia zasobów transportowych na potrzeby usługi (wiadomości STR) oraz odpowiedzi na te żądania (wiadomości STA).

	Time	Source	Destination	Protocol	Len	Info
11	0.17…	192.168.56.101	192.168.56.103	OpenFlow	82	Type: OFPT_HELLO
13	0.18…	192.168.56.103	192.168.56.101	OpenFlow	90	Type: OFPT_FEATURES_REQUEST
15	0.19…	192.168.56.101	192.168.56.103	OpenFlow	98	Type: OFPT_FEATURES_REPLY
16	0.19…	192.168.56.103	192.168.56.101	OpenFlow	82	Type: OFPT_MULTIPART_REQUEST, (
18	0.19…	192.168.56.101	192.168.56.103	OpenFlow	274	Type: OFPT_MULTIPART_REPLY, OFP
19	0.20…	192.168.56.103	192.168.56.101	OpenFlow	82	Type: OFPT_GET_CONFIG_REQUEST
21	0.20…	192.168.56.101	192.168.56.103	OpenFlow	74	Type: OFPT_BARRIER_REPLY
22	0.20…	192.168.56.101	192.168.56.103	OpenFlow	78	Type: OFPT_GET_CONFIG_REPLY
24	0.20…	192.168.56.103	192.168.56.101	OpenFlow	82	Type: OFPT_MULTIPART_REQUEST, (
25	0.20…	192.168.56.101	192.168.56.103	OpenFlow	98	Type: OFPT_MULTIPART_REPLY, OFP
26	0.21…	192.168.56.103	192.168.56.101	OpenFlow	82	Type: OFPT_MULTIPART_REQUEST, (
27	0.21…	192.168.56.101	192.168.56.103	OpenFlow	11…	Type: OFPT_MULTIPART_REPLY, OFP
29	0.36…	192.168.56.103	192.168.56.101	OpenFlow	90	Type: OFPT_ROLE_REQUEST
30	0.36…	192.168.56.103	192.168.56.101	OpenFlow	90	Type: OFPT_ROLE_REQUEST
31	0.36…	192.168.56.101	192.168.56.103	OpenFlow	90	Type: OFPT_ROLE_REPLY
33	0.36…	192.168.56.101	192.168.56.103	OpenFlow	90	Type: OFPT_ROLE_REPLY
35	0.39…	192.168.56.103	192.168.56.101	OpenFlow	12…	Type: OFPT_BARRIER_REQUEST
36	0.39…	192.168.56.101	192.168.56.103	OpenFlow	274	Type: OFPT_MULTIPART_REPLY, OFP
37	0.39…	192.168.56.101	192.168.56.103	OpenFlow	74	Type: OFPT_BARRIER_REPLY
38	0.39…	192.168.56.101	192.168.56.103	OpenFlow	274	Type: OFPT_MULTIPART_REPLY, OFP
39	0.39…	192.168.56.101	192.168.56.103	OpenFlow	74	Type: OFPT_BARRIER_REPLY

Rys. 6. Zrzut pakietów na styku OpenFlow

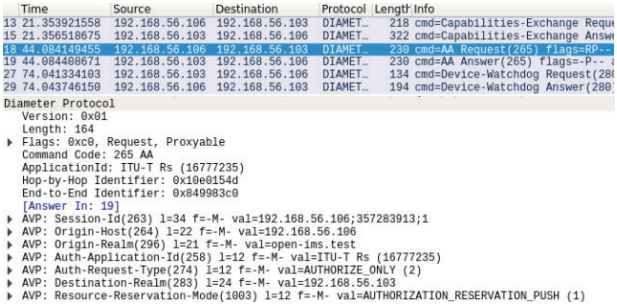
Generowanie poszczególnych rodzajów żądań Diameter (wiadomości AAR, STR) odbywa się poprzez maszynę wirtualną CDiameterPeer 1 z rys. 3, która reaguje na sygnały uniksowe wysyłane do jej procesu głównego aplikacji CDiameterPeer. W tym celu w pliku main.c tej aplikacji utworzono funkcję request_init(), która w zależności od otrzymanego sygnał uniksowego (np. SIGUSR2) generuje wiadomość AAR albo STR z opisanymi przez standard [12] polami Attribute–Value Pair (AVP). Przykładowo dla wiadomości AAR z wstępną rezerwacją zasobów poza domyślnie wysyłanymi przez CDiameterPeer polami AVP należało dodać następujące pola: „Auth-Application-Id(258)", „Auth-Request-Type(274)", „Resource-Reservation-Mode(1003)", „Destination-Realm(283)". Przetwarzanie wiadomości AAR oraz STR przez drugą maszynę wirtualną z rys. 3 (zawierającą oprogramowanie CDiameterPeer 2) wymaga edycji pliku server.c jej aplikacji CDiameterPeer. Ma ona na celu wywoływanie REST API sterownika ONOS w reakcji na otrzymywane wiadomości Diameter (szczegóły w następnym podrozdziale) oraz przesyłanie odpowiedzi na te wiadomości (zawierające rezultat rezerwacji – AAA, bądź zwolnienia – STA – zasobów transportowych).

Testowanie poprawności implementacji styku Diameter polegało na analizie listingów oprogramowania CDiameterPeer 1 i 2, a także zrzutów przesyłanych wiadomości Diameter za pomocą sniffera Wireshark dla wszystkich rodzajów generowanych żądań. Zrzut wiadomości Diameter towarzyszący generowaniu wiadomości AAR zilustrowano na rys. 7. Poza wiadomościami charakterystycznymi dla protokołu Diameter (Capabilities-Exchange Request/Answer, Device-Watchdog Request Answer) widać na nim także wiadomość AAR z zawartymi w niej zestandaryzowanymi polami AVP oraz wiadomość AAA, która jest odpowiedzią na AAR. Wszystkie przeprowadzone testy potwierdziły poprawność implementacji styku Diameter w opisywanym projekcie.

3.4. Implementacja styku z protokołem HTTP REST/JSON

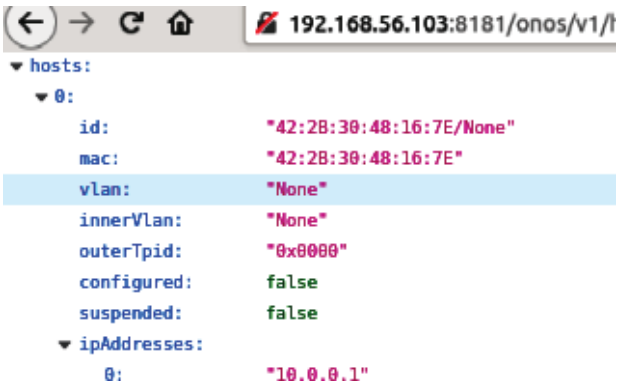
Jak już wspomniano, w reakcji na otrzymane wiadomości Diameter AAR oraz STR oprogramowanie CDiameterPeer 2 (druga maszyna wirtualna na rys. 3) generuje

wiadomości HTTP REST/JSON do sterownika ONOS, w celu rezerwacji lub zwolnienia zasobów transportowych. Przy tym logika styku z protokołem HTTP REST/JSON powinna zostać zaimplementowana w pliku server.c aplikacji CDiameterPeer 2.



Rys. 7. Zrzut wiadomości Diameter towarzyszący generowaniu wiadomości AAR

W ramach przygotowania do tej implementacji dokonano rozpoznania REST API sterownika ONOS za pomocą przeglądarki internetowej (rys. 8). Aktualnie trwają prace nad zastosowaniem bibliotek C++ libcurl [13] (protokół HTTP) oraz JsonCpp [14] (format wymiany danych JSON), z którymi autorzy artykułu mieli doświadczenie w innych projektach [15], do oprogramowania styku HTTP REST/JSON w ramach opisywanego projektu.



Rys. 8. Testowanie REST API sterownika ONOS

4. PODSUMOWANIE

Przedstawiona w artykule koncepcja odpowiada kierunkom prac prowadzonych w innych ośrodkach i pokazuje możliwość integracji aktualnych technologii telekomunikacyjnych jakimi są IMS/NGN i SDN. Wpisuje się to w trendy dzisiejszych prac badawczych nad optymalizacją zarządzania i sterowania systemów telekomunikacyjnych poprzez zwiększenie poziomu ich automatyzacji. Bardzo ważna jest tu także możliwość zastosowania rozwiązań z otwartych projektów, co obniża koszty realizacji i skraca czas od koncepcji do wdrożenia.

W ramach projektu wykorzystano znaczną liczbę otwartych projektów realizujących konkretną funkcjonalność sieci IMS/NGN (jak CDiameterPeer) oraz sieci SDN (jak sterownik ONOS z programowalnymi przełącznikami Open vSwitch emulowanymi w Mininet). Poza tym wykorzystano narzędzia wpisujące się w najnowsze trendy wykorzystania wirtualizacji, jak aplikacja hipernadzorcy Oracle VM VirtualBox. Ta różnorodność otwartego oprogramowania i narzędzi oraz braki w dokumentacji wymagają dużej biegłości w rozpoznawaniu problemów konfiguracji i modyfikacji oprogramowania.

Z czterech etapów prac związanych z realizacją opisywanego projektu w pełni wykonano trzy pierwsze: przygotowanie środowiska IMS/NGN i SDN; konfiguracja podstawowej komunikacji Diameter oraz OpenFlow; implementacja styku z protokołem Diameter. Czwarty i zarazem ostatni etap (implementacja styku z protokołem HTTP REST/JSON) jest w trakcie realizacji z wykorzystaniem bibliotek do obsługi wiadomości protokołu HTTP oraz formatu wymiany danych JSON, które autorzy tego artykułu znają i stosowali w innych projektach. Osiągnięte rezultaty pozwalają już zatem na postawienie tezy, że koncepcja SDN może być z powodzeniem zintegrowana i zastosowana do współpracy z siecią IMS/NGN.

LITERATURA

[1] ITU-T Rec. Y.3300. 06/2014. *Framework of software-defined networking*.
[2] *Open Network Operating System (ONOS) SDN Controller*, https://opennetworking.org/onos/.
[3] *Floodlight Controller*, https://floodlight.atlassian.net/wiki/spaces/floodlightcontroller/overview.
[4] ONF TS-025. 04/2015. *OpenFlow Switch Specification Version 1.5.1*.
[5] ITU-T Rec. Y.2012. 04/2010. *Functional requirements and architecture of next generation networks*.
[6] 3GPP TS 23.228. 12/2021. *IP Multimedia Subsystem (IMS); Stage 2 (Release 17)*. v17.3.0.
[7] *Open Source IMS (OSIMS)*, http://nam.ece.upatras.gr/ppe/?q=node/2.
[8] Khairi Sara, Brahim Raouyane, Mostafa Bellafkih. 2017. "Towards Enhanced QoS Management SDN-based for Next Generation Networks with QoE Evaluation: IMS Use Case". *Journal of Mobile Multimedia* 13 (3-4) : 183-196.
[9] *Open IMS Core CSCFs: The CDiameterPeer Module (cdp)*, http://OpenIMSCore.sourceforge.net/docs/ser_ims/CDP.html.
[10] *Open vSwitch*, https://www.openvswitch.org/.
[11] *Mininet*, http://mininet.org/.
[12] ITU-T Rec. Q.3301. 08/2013. *Resource control protocol No. 1, version 3 – Protocol at the Rs interface between service control entities and the policy decision physical entity*.
[13] *libcurl*, https://curl.se/libcurl/.
[14] *JsonCpp*, https://github.com/open-source-parsers/jsoncpp.
[15] Blok Marek, Sylwester Kaczmarek, Marcin Narloch, Maciej Sac. 2017. "The central server of the Border Guard's distributed multimedia system for monitoring and visualisation of ongoing and archival events". *Journal of Marine Engineering & Technology* 16 (4) : 257-272.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

HEURYSTYCZNY ALGORYTM ULOKOWANIA ŁAŃCUCHÓW FUNKCJI OBSŁUGI W PROGRAMOWEJ SIECI 5G

HEURISTIC ALGORITHM FOR PLACEMENT OF SERVICE FUNCTION CHAINS IN SOFTWARIZED 5G NETWORK

Jerzy Martyna
Wydział Matematyki i Informatyki UJ
ul. Prof. S. Łojasiewicza 6, 30-348 Kraków
e-mail: jerzy.martyna@uj.edu.pl

DOI: 10.15199/59.2022.4.39

Streszczenie: **Wirtualizacja sieci 5G dostarcza usługi sieciowej określanej jako łańcuchy funkcji obsługi. Ich ulokowanie w rzeczywistej sieci winno spełniać wymagania użytkowników i ograniczenia systemowe. W artykule przedstawiono metodę ulokowania SFC, która minimalizuje koszty przetwarzania w systemie 5G. Przybliżone rozwiązanie tego problemu może być uzyskane m.in. przy użyciu podanego algorytmu heurystycznego. Badania symulacyjne potwierdziły efektywność proponowanego rozwiązania.**

Abstract: **Virtualization of the fifth-generation network delivers the service function chains (SFCs). Their placement in the real network should meet user requirements and system restrictions. The article presents the SFC deployment method that minimizes the processing costs in the 5G system. An approximate solution to this problem can be obtained, among others, using the given heuristic algorithm. Simulation tests confirmed the effectiveness the proposed solution.**

Słowa kluczowe: **Łańcuch funkcji obsługi, sieć komórkowa 5G, niezawodność k-tego terminala, alokacja zasobów radiowych, sieć zdefiniowana programowo.**

Keywords: **Service Function Chain. 5G cellular networ, k-terminal reliability, radio resource allocation, software defined network.**

1. WPROWADZENIE

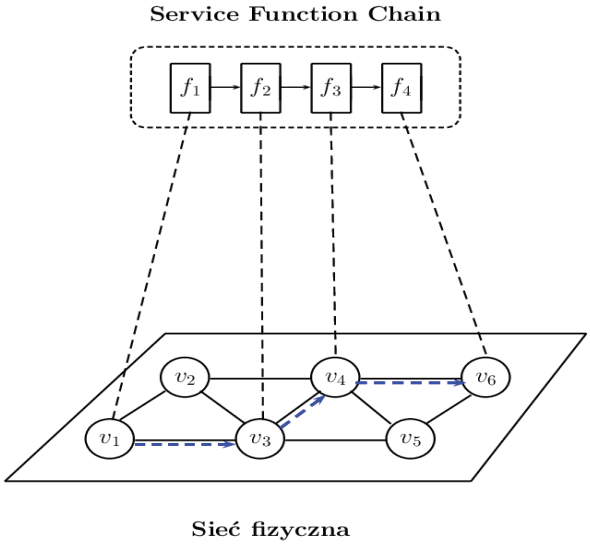
Programowo zdefiniowane sieci (Software Defined Networks, SDNs) wraz z łańcuchami funkcji obsługi (ang. service function chains, SFCs) są istotnymi technologiami, które pozwalają na przezwyciężenie szeregu ograniczeń sieci 5G [1]. Między innymi, dzięki nim można zapewnić sprawdzanie aktualnego wykorzystania zasobów lub umożliwić ponowne użytkowania pól nagłówka w protokołach płaszczyzny danych. Co więcej, dzięki wirtualizacji sieci 5G można dużo łatwiej je projektować i i zarządzać ich funkcjami.

Koncepcja SFC została m.in. wprowadzona w pracach [2], [3]. Zakłada ona m.in., że łańcuch SFC składa się z zestawu wirtualnych funkcji sieciowych (VNF) połączonych poprzez łącza logiczne. Wiele SFCs dla różnych klientów może współdzielić te same zasoby obliczeniowe i sieciowe w celu poprawy wykorzystanie zasobów. Pomimo to, w miarę wzrostu wymagań dla sieci 5G, m.in. przy tworzenia chmur radiowych czy umożliwiania mobilności, rozmieszczenie SFC staje się dużym wyzwaniem dla projektantów tych sieci. Stąd też efektywne odwzowanie SFC w sieciach mobilnych 5G może zminimalizować koszty użytkowania zasobów i łączy do nich.

Znalezienie optymalnego rozmieszczenia łańcuchów SFC w fizycznych zasobach sieciowych jest problemem NP-trudnym. Wielu autorów podejmowało się znalezienia rozwiązywania tego problemu. Między innymi Cohen i in. [4] zaproponowali algorytmy aproksymacyjne, które mogą obniżyć koszty eksploatacji systemu. Autorzy pracy [5] badali problem wdrażanie VNF i wyboru ścieżki w celu lepszego wykorzystania łączy i serwerów, co również poprawia efektywność wykorzystania systemu. Z kolei, aby zoptymalizować rozmieszczenie VNF dla zmniejszenia kosztów operacyjnych sieci i poprawy jej wykorzystania, autorzy pracy [6] zaproponował heurystykę opartą na programowaniu dynamicznym. Pomimo to, zaproponowana też heurystyka głównie optymalizuje wykorzystanie serwerów i ignoruje koszt implementacji ścieżek.

W pracy [7] przedstawiono algorytm rozmieszczania SFC, który zapewnia usługi o wymaganych parametrach QoS. W innym artykule, jego autorzy Park in. [8] zaproponowali rozwiązanie dostarczające usługi gwarancji bezpieczeństwa. W innym artykule autorzy [9] zaprezentowali algorytm rozmieszczania SFC gwarantujący wymagane opóźnienia czasowe. Pomimo to, żadna z powyżej podanych prac nie bierze pod uwagę kosztów wdrożenia VNF. Uwzględniono to m.in. w pracy [10], w której przedstawiono minimalizację długości ścieżki całego SFC poprzez łączenie VNF i mapowanie tymczasowych łączy oraz minimalizację kosztów zasobów obliczeniowych poprzez maksymalne wykorzystanie istniejących maszyn wirtualnych dla tego samego VNF.

Głównym celem artykułu jest przedstawienie nowej metody rozmieszczania łańcuchów funkcji obsługi w programowanych sieciach 5G, która pozwala na uwzględnienie kosztów przetwarzania oraz minimalizację kosztów użycia zasobów. Dzięki niej jest możliwe znalezienie rozwiązania, które jest bliskie optymalnemu.



Rys. 1. Żądanie usługi z użyciem przykładowego łańcucha funkcji obsługi.

Badania symulacyjne potwierdziły efektywność tej metody.

W kolejnym rozdziale przedstawiono model systemu. W rozdziale 3 opisano ulokowanie SFC w sieci fizycznej, które uwzględnia niezawodność i koszty. Rozdział 4 przedstawia heurystyczny algorytm takiego umieszczania SFC w sieci fizycznej. Wyniki badania efektywności proponowanego algorytmu podano w rozdziale 5. Podsumowanie zamieszczono w ostatnim rozdziale.

2. MODEL SYSTEMU

W przyjętym modelu systemu założono, że sieć fizyczna (nazywana też siecią podłoża (ang. *substrate network*) zdefiniowana jest jako nieskierowany graf fizyczny $G = (V, E)$, gdzie V jest zbiorem N krawędzi, będącymi węzłami fizycznymi, zapisywanymi jako $V = \{v_1, ..., v_N\}$, a E jest zbiorem połączeń fizycznych L między węzłami oznaczonymi $E = \{e_1, ..., e_L\}$. Każdy węzeł fizyczny $v, v \in V$ należy scharakteryzować przez moc obliczeniową, rozmiar pamięci itp. Niech zbiór łańcuchów usług będzie określony przez S. Pojedynczy łańcuch usług składa się z listy VNF i jest reprezentowany przez graf skierowany $SF = (F, M)$, gdzie F jest odpowiednio zbiorem funkcji wirtualnych, a M jest zbiorem łączy wirtualnych. Funkcje wirtualne składają się z żądań mocy obliczeniowej, rozmiaru pamięci itp. Z kolei każde łącze wirtualne $m, m \in M$ składa się z żądań przepustowości, szybkości transmisji danych itp. Rys. 1 przedstawia przykład skonstruowanej SFC, który składa się z czterech VNF połączonych łączami wirtualnymi.

Rozmieszczanie SFC w sieci fizycznej przebuega następująco. Łańcuch usług s musi zostać przydzielony do jednego lub więcej VNF f, pod warunkiem, że spełnione są następujące wymagania. Każdy VNF f może być umieszczony tylko w jednym węźle fizycznym v, który spełnia przyjęte ograniczenia. Analogicznie, każde łącze m może być odwzorowane tylko na jedną fizyczną ścieżkę, która także spełnia przyjęte ograniczenia. Ponadto, ze względu na przyjęte tutaj podejście, należy wziąć pod uwagę takie żądania, jak pewne wartości niezawodności i kosztów.

3. NIEZAWODNOŚĆ SIECI I KOSZT PRZETWARZANIA

Przyjmuje się, niezawodność sieci jest prawdopodobieństwem, że system pomyślnie wykonuje zamierzone funkcje przez określony czas [11]. W przeciwnym razie, jeśli sieć działa pomyślnie w czasie t_0, niezawodność sieci daje prawdopodobieństwo, że w przedziale od 0 do t_0 nie było żadnych awarii. Zakłada się, że prawdopodobieństwo udanej komunikacji między węzłem źródłowym a docelowym następuje, gdy istnieje ścieżka pomiędzy nimi i złożona jest z k węzłów. Prawdopodobieństwo to nazywane jest też niezawodnością k-tego terminala.

Niezawodność k-tego terminala dla k węzłów $\{n_1, \ldots, n_k\} \subset V(G, p)$ dla grafu (G, p), gdzie p jest prawdopodobieństwem awarii łącza, można podać jako [12]:

$$R_C^{n_1,\ldots,n_k}(G,p) = \sum_{i=w_{n_1},\ldots,n_k}^{\omega} T_i^{n_1,\ldots,n_k}(G) p^{\omega-i}(1-p)^i \tag{1}$$

gdzie $\omega = \mid E(G) \mid$ to rozmiar grafu, $T_i^{n_1,\ldots,n_k}(G)$ oznacza liczbę podgrafów łączących węzły $n_1, \ldots, n_k$ z i-tymi krawędziami. oraz $w_{n_1}, \ldots, w_{n_k}$ to rozmiar minimalnego zbioru łączącego węzły $n_1, \ldots, n_k$.

Niezawodność k-tego terminala może być również wyrażona w następujący sposób:

$$R_C^{n_1,\ldots,n_k}(G,p) = 1 - \sum_{i=\beta(G)}^{\omega} C_i^{n_1,\ldots,n_k}(G) p^i (1-p)^{\omega-1} \tag{2}$$

gdzie $C_i^{n_1,\ldots,n_k}(G)$ oznacza liczbę krawędzi o liczności i oraz $\beta(G)$ podaje spójność.

Całkowity koszt przetwarzania w systemie 5G jest złożony z kosztu użycia zasobów $\theta(r)$ w węźle, kosztu fizycznych węzłów $\theta(w)$ oraz kosztu komunikacji $\theta(c)$, co można przedstawić następująco:

$$\Theta(r, w, c) = \theta(r) + \theta(w) + \theta(c) \tag{3}$$

Ponieważ fizyczne węzły mogą być różnorodne, to każdy VNF f w odniesieniu do poszczególnych węzłów może mieć inną wartość kosztu. Stąd też koszt użycia zasobów $\theta(r)$ w węźle wynosi:

$$\theta(r) = \sum_{s \in S} \sum_{v \in V} \sum_{w \in F} (cost_{sf}^{v_{CPU}} + cost_{sf}^{v_{mem}} + cost_{sf}^{v_{sto}}) X_{sf}^{v} \tag{4}$$

gdzie $cost_{sf}^{v_{CPU}}$ jest kosztem CPU w fizycznym węźle v używanym przez VNF f, $cost_{sf}^{v_{mem}}$ jest kosztem pamięci

fizycznego węzła v używanego przez VNF f, $cost_{sf}^{v_{sto}}$ jest kosztem pamięci zewnętrznej fizycznego węzła v używanego przez VNF f. Zmienna binarna X_{sf}^{v} jest równa 1, jeśli VNF f z łańcucha s jest umieszczony w fizycznym węźle v. W przeciwnym razie $X_{sf}^{v} = 0$.

Koszt fizycznych węzłów wiąże się z kosztami prądu zużywanymi przez ten węzeł. Jego wartość wynosi

$$\theta(w) = \sum_{v \in V} (p_v^{max} + (p_v^{max} - p_v^{min}) \frac{1}{cap_{CPU}^{v}} \sum_{f \in F} d_{CPU,f}^{v} X_f^{v}) \cdot Z^v \quad (5)$$

gdzie p_v^{min}, p_v^{max} są odpowiednio minimalną i maksymalną wartością zużycia energii węzła $v \in V$, cap_{CPU}^{v} jest średnią szybkością przetwarzania CPU węźle v, $d_{CPU,f}^{v}$ jest wymaganym czasem obsługi przez CPU w węźle v należącym do VNF f. Zmienna binarna $Z^v = 1$, jeśli węzeł fizyczny v jest aktywny, w przeciwnym razie $Z^v = 0$.

Koszt komunikacji $\theta(c)$ jest sumą kosztów zajętych pasm dla ścieżek fizycznych przez łącza wirtualne, na których rozmieszczane są łącza wirtualne. Można go przedstawić jako

$$\theta(c) = \sum_{s \in S} \sum_{v \in V} \sum_{m \in M} (cost_{sm}^{e_{band}} \cdot Q_{sm}^{e}) \quad (6)$$

gdzie $cost_{sm}^{e_{band}}$ jest kosztem zajętego pasma dła ścieżki fizycznej używanych przez łącze wirtualne m. Zmienna binarna $Q_{sm} = 1$, jeśli łańcuch s jest rozmieszczony w fizycznym łączu $e \in E$. W przeciwnym razie zmienna ta jest równa 0.

4. HEURYSTYCZNY ALGORYTM ROZMIESZCZENIA ŁAŃCUCHÓW FUNKCJI OBSŁUGI

Algorytm zapewnia niezależnie od fizycznej topologii sieci 5G, że wszystkie elementy tej sieci będą udostępnione, o ile wystąpi żądanie ich udostępnienia. Ponadto bierze on pod uwagę redundancję poszczególnych elementów tej sieci, w tym łączy i węzłów. Uwzględnia on też żądania usługi transmisji danych z ograniczeniami czasowymi danych i wymaganiami dotyczącymi szybkości przepływu. Pseudokod dla powyższego algorytmu przedstawiono na rys. 2. Jego działanie polega na znajdowaniu dla danej topologii fizycznej 5G najkrótszych ścieżek między podanym korzeniem i liściami drzew. Następnie, ze względu na ograniczenia czasowe, niektóre z nich spełniają wymagania niezawodności i kosztów przetwarzania. W tym celu wykorzystywane są wartości niezawodności k-tego terminala i kosztów przetwarzania. W przypadku, gdy takiej ścieżki brakuje, sprawdzana jest możliwość redundancji poszczególnych elementów. Dzięki redundancji tych elementów możliwe jest znalezienie ścieżek spełniających żądane parametry. Znalezione w ten sposób ścieżki są przypisywane do danych żądań SFC.

Algorithm 1 Algorytm rozmieszczania SFC

```
procedure ALOKACJA ZASOBÓW RADIOWYCH
Require: R_con, D_con, Θ_con;
    [bufor] ← ∅;
    for ∀f ∈ VNF do
        sort ścieżki(f);
        put bufor[posortowane_ścieżki];
    end for
    for ∀s ∈ S do
        for ∀f ∈ VNF do
            oblicz R_f oraz Θ(r, w, c) dla
            k-tego terminala
            while (R_f > R_con ∧ Θ(r, w, c) < Θ_con) do
                oblicz end_to_end_delay_f
                if end_to_end_delay_f < D_con then
                    odwzoruj_s_do_
                    bufor[posortowane_ścieżki];
                else
                    znajdź_ścieżkę_redundantną_s'
                end if
                odwzoruj_s_oraz_s'_do_
                bufor[posortowane_ścieżki]
            end while
        end for
    end for
end procedure
```

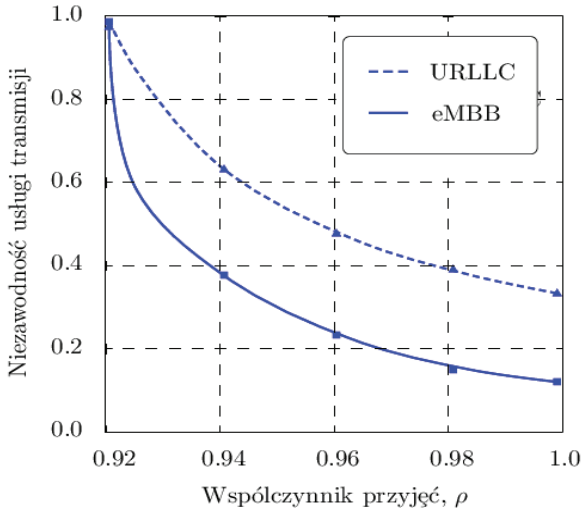
Rys. 2. Pseudokod algorytmu dla rozmieszczania ścieżek SFC.

5. WYNIKI BADAń SYMULACYJNYCH

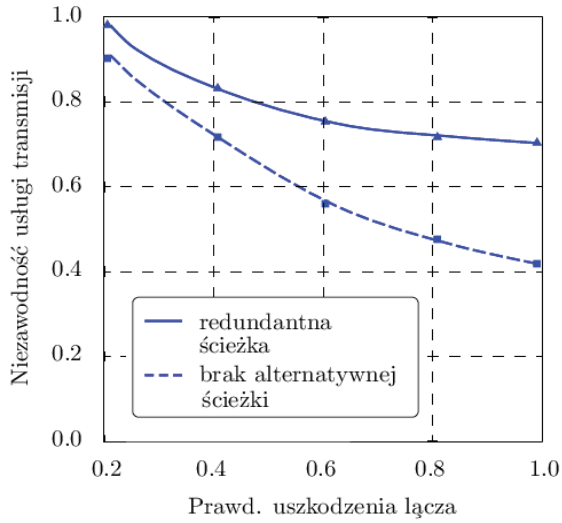
Sieć fizyczna jest tutaj reprezentowana przez sieć 5G o konfiguracji dwupoziomowej z trzema węzłami. Każdy węzeł sieci ma jedno centrum danych o pojemności równej 2000 jednostek. W modelu założono dwie klasy przepływuy: URLLC i eMBB. Przepływ URLLC wymaga wyjątkowo niezawodnego i małego opóźnienia w transmisji, czyli dla jednokierunkowej transmisji radiowej z dopuszczalnym opóźnieniem 1 ms. Przepływ eMBB zapewnia duże możliwości w zakresie łączności dla Internetu Rzeczy. Każde z centrów danych może zapewnić od sześciu do ośmiu funkcji sieci wirtualnej. Przyjęto, że współczynnik błędów pakietów (PER) wynosi 10^{-3} [13].

Dla każdego żądania łańcucha usług w sieci zaakceptowano tylko 4 VNF. Każdy z nich wymaga dwóch zasobów (pojemność mocy, wielkość pamięci) a każde łącze wirtualne ma wymaganą przepustowość równą 200, 300, 400 Gb/s z jednakowym prawdopodobieństwem. Wartość opóźnienia przetwarzania VNF jest tutaj równe 60 − 120 μs.

Za pomocą programu symulacyjnego dokonano oceny zaproponowanego algorytmu pod kątem oceny działania. W pierwszym scenariuszu zbadano wiarygodność wymagań dla dostarczonego algorytmu rozmieszczenia żądań z łańcuchów usług. Następnie zbadano dostępność dla obu strumieni danych z przyjętym algorytmem ulokowania i jego brakiem.



Rys. 3. Niezawodność usługi transmisji w zależności od współczynnika przyjęć ρ.



Rys. 4. Niezawodność usługi transmisji w zależności od prawd. uszkodzenia łącza.

Na rys. 3 przedstawiono niezawodności usługi transmisji w zależności od współczynnika przyjęć ρ. Użyty tu współczynnik przyjęć został określony jako iloraz liczby zaakceptowanych żądań transmisji o wymaganej niezawodności do liczby wszystkich żądań transmisji. Wykres wskazuje, że zastosowany algorytm może znacząco poprawić niezawodność transmisji.

Rys. 4 przedstawia niezawodność usługi transmisji w zależności od prawdopodobieństwa uszkodzenia łącza przy użyciu nadmiarowej ścieżki lub jej braku. Jest widoczne, że zastosowanie nadmiarowej ścieżki powodujw wzrost niezawodności przepływu danych dla obydwu klas przepływu.

6. WNIOSKI

W pracy przedstawiono metodę rozmieszczanie łańcuchów funkcji usług w zwirtualizowanych sieciach 5G pozwala na znajdowanie ścieżek spełniających wymaganą niezawodność i koszty. Przedstawiony algorytm może być zastosowany online w czasie rzeczywistym. Ewentualne uwzględnienie redundantnych ścieżek pozwala na dodatkowy wzrost efektywności zwirtualizowanej sieci. Uzyskane wyniki wskazują na skuteczność w osiąganiu kluczowych celów projektowych przy pomocy algorytmu rozmieszczania ścieżek w fizycznej sieci 5G.

Literatura

[1] F. Z. Yousaf, M. Bredel, S. Schaller, F. Schneider, NFV and SDN - Key Technology Enablers for 5G Networks, IEEE J. on Sel. Areas in Comm., vol. 35, no. 11, pp. 2468 - 2478, 2017.

[2] P. Quinn, T. Nadeau, Problem Statement for Service Function Chaining, RFC 7498, Tech. Rep. 7498, Apr. 2015. [Online]. Available: https://rfc-editor.org/rfc/rfc7498.txt

[3] A. M. Medhat i in., Service Function Chaining in Next Generation Networks: State of the Art and Research Challenges, IEEE Comm. Mag., vol. 55, no. 2, pp. 216 - 223, 2017.

[4] R. Cohen, L. Lewin-Eytan, J. S. Naor, D. Raz, Near optimal placement of virtual network functions, in Proc. IEEE INFOCOM 2015, pp. 1346 - 1354.

[5] T.-W. Kuo i in., Deploying chains of virtual network functions: On the relation between link and server usage, in Proc. IEEE INFOCOM 2016, pp. 1 - 9.

[6] M. F. Bari, S. Chowdhury, R. Ahmed, and R. Boutaba. (Mar. 2015). On orchestrating virtual network functions in NFV. [Online]. https://arxiv.org/abs/1503.06377

[7] T. Kim i in., A QoS assured network service chaining algorithm in network function virtualization architecture, in Proc. IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput. (CCGrid), May 2015, pp. 1221–1224.

[8] T. Park i in., QoSE: Quality of Security a network security framework with distributed NFV, in Proc. IEEE ICC, May 2016, pp. 1–6.

[9] Y. Li, L. T. X. Phan, B. T. Loo, Network functions virtualization with soft real-time guarantees, in Proc. IEEE INFOCOM, Apr. 2016, pp. 1 – 9.

[10] D. Zhao, J. Ren, R. Lin, S. Xu, V. Chang, On orchestrating service function chains in 5G mobile network. IEEE Access, vol. 7, pp. 39402-39416, 2019.

[11] K. Sahner, K. S. Trivedi, A. Puliafito, Performance and Reliability Analysis of Computer Systems: An Example-Based Approach Using the SHARPE Software Package, Boston: Kluwer Academic Publishers, 1996.

[12] G. Egeland, P. E. Engelstad. The reliability of wireless backhaul mesh networks. Proc. of the 2008 IEEE Int. Symp. on Wireless Comm. Syst., pp. 178 - 183, 2008.

[13] P. Popovski, K. F. Trillingsgaard, O. Simeone, 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View, IEEE Access, Vol. 6 pp. 55765-55779, 2018.

MECHANIZM ZESTAWIANIA ŚCIEŻEK OPTYCZNYCH TYPU BYPASS W SIECIACH IP/EON

MECHANISM OF SETTING LIGHTPATHS CALLED BYPASSES IN IP/EON NETWORKS

Edyta Biernacka[1]; Jerzy Domżał[2]

[1] AGH w Krakowie, edyta.biernacka@agh.edu.pl
[2] AGH w Krakowie, jerzy.domzal@agh.edu.pl

DOI: 10.15199/59.2022.4.40

Streszczenie: **Artykuł przedstawia analizę wydajności autorskiego mechanizmu zestawienia dodatkowych ścieżek (zwanych bypass) w sieciach wielowarstwowych złożonych z warstwy wirtualnej (IP) oraz z warstwy elastycznej sieci optycznej (EON). Zaproponowano współczynnik wykorzystania zasobów podczas ich dynamicznej alokacji dla mechanizmu bypass. Przeprowadzono eksperymenty symulacyjne dla dwóch topologii sieci oraz uwzględniono dwie klasy ruchu. Wyniki potwierdzają wysoką wydajność zaproponowanego mechanizmu z uwzględnieniem priorytetyzacji ruchu.**

Abstract: **The article presents an analysis of the efficiency of the proprietary mechanism of setting up additional paths (called bypass) in multilayer networks composed of a virtual layer (IP) and a layer of elastic optical network (EON). The resource utilization ratio was proposed during its dynamic allocation for the bypass mechanism. Simulation experiments were conducted for two network topologies and two classes of traffic were taken into account. The results confirm the high efficiency of the proposed mechanism, taking into account traffic prioritization.**

Słowa kluczowe: **ruting, dynamiczna alokacja zasobów, elastyczna sieć optyczna, sieć wielowarstwowa.**

Keywords: **routing, dynamic resource allocation, elastic optical network, multi-layer network.**

1. WSTĘP

W artykule przedstawiono problem dynamicznej alokacji zasobów w sieci wielowarstwowej złożonej z warstwy IP (wirtualnej) i warstwy optycznej. Problem ten rozważa się odrębnie dla ruchu o wysokim i niskim priorytecie.

Podstawowym celem artykułu jest przedstawienie analizy wydajności mechanizmu zestawienia dodatkowych ścieżek optycznych w sieciach wielowarstwowych z uwzględnieniem klas ruchu (priorytetów) oraz zajętości zasobów warstwy optycznej. Należy podkreślić, że w poprzednich pracach [1], [2] i [3] poddano analizie mechanizm zestawienia dodatkowych ścieżek optycznych, którego celem było zmniejszenie prawdopodobieństwa blokowania przepustowości. W sytuacji wystąpienia przeciążenia w warstwie wirtualnej zestawiana jest dodatkowa ścieżka bez uwzględniania priorytetu żądania czy też zajętości zasobów warstwy optycznej. W tym artykule przedstawiono mechanizm zestawienia dodatkowych ścieżek optycznych z uwzględnieniem klasy ruchu i zajętości łączy podczas alokacji. Eksperymenty symulacyjne przeprowadzono z użyciem środowiska OMNeT++ [8], a także przeprowadzono i dokonano oceny BBP (Bandwidth Blocking Probability) dla różnych klas ruchu w sieci IP/EON (Internet Protocol/Elastic Optical Network). Badano wydajność mechanizmów zestawienia dodatkowych ścieżek dla żądań o wysokim i niskim priorytecie, a także uwzględniono współczynnik wykorzystania zasobów dla ruchu o niskim priorytecie. W rozdziale 2 opisano mechanizm zestawienia dodatkowych ścieżek oraz alokacji zasobów dla różnych klas ruchu, natomiast wyniki badań symulacyjnych oraz analizę wyników zawarto w rozdziale 3. Podsumowanie przedstawiono w rozdziale 4.

2. MECHANIZM ZESTAWIENIA DODATKOWCH ŚCIEŻEK OPTYCZNYCH

W tej części referatu przedstawiono koncepcję mechanizmu zestawienia dodatkowych ścieżek optycznych oraz zaproponowane ulepszenie dla tego mechanizmu. Ze względu na ograniczoną wielkość referatu został pominięty dokładny opis warstwy optycznej (szczegółowe informacje znajdują się w [1]) oraz architektura sieci wielowarstwowej, która została przedstawiona w pracach [2] i [3].

2.1. Koncepcja

Przedstawiony mechanizm stanowi rozwiązanie problemu dynamicznej obsługi żądań ruchu o różnych priorytetach w sieciach wielowarstwowych złożonych z warstw IP i optycznej. Sieci te są szeroko stosowane przez operatorów. Dodatkowo współpraca między warstwami sieci jest zapewniona przez płaszczyznę sterowania zgodną z koncepcją sieci sterowanych programowo (SDN - Software-Defined Network), a centralny sterownik monitoruje użycie zasobów w sieci. W warstwie optycznej zastosowana jest elastyczna sieć optyczna (EON), która zapewnia efektywne wykorzystanie zasobów poprzez drobnoziarnisty charakter widma optycznego i adaptacyjne szybkości transmisji (zastosowanie zaawansowanych formatów modulacji) [7]. Aby zestawić ścieżkę optyczną w sieci EON należy rozwiązać problem RSA (Routing and Spectrum Allocation) [4]. Oznacza to, że ścieżka optyczna musi spełniać warunki ciągłości, nie nakładania się i sąsiedztwa widma wzdłuż ścieżki.

Mechanizm zestawienia dodatkowych ścieżek tzw. bypass zakłada dwa rodzaje zasobów optycznych dla sieci wielowarstwowej. Pierwszy rodzaj zasobów jest widoczny dla warstwy IP, a ścieżki optyczne alokujące te zasoby traktuje się jako wirtualne łącza w warstwie IP. Natomiast drugi rodzaj zasobów (zwanych w dalszej części referatu jako EON) jest ukryty dla warstwy IP i zarezerwowany do obsługi ruchu wyłącznie w warstwie optycznej. Ruch jest kierowany główną ścieżką przez warstwę wirtualną, jeśli zasoby w tej warstwie zapewniają zadaną przepustowość. W przeciwnym razie proponowany algorytm próbuje zestawić dodatkową ścieżkę optyczną (nazywaną bypass) alokującą ukryte zasoby. Aby ścieżka bypass mogła być zestawiona musi być spełniony warunek RSA, a liczba n sąsiednich szczelin zajmowanego widma przez bypass, zgodnie z [4], wynosi:

$$n = \left\lceil \frac{bw_{demand}}{M_p \cdot C_{FS}} \right\rceil + GB \tag{1}$$

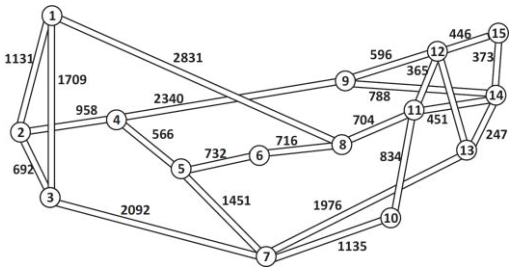
gdzie: bw_{demand} [Gbit/s] – żądana przepływność; M_p [bit/s/Hz] – efektywność widmowa modulacji wybranej dla ścieżki *bypass*; C_{FS} [Gbit/s] – pojemność szczeliny dla modulacji BPSK; GB – liczba szczelin przeznaczona do oddzielenia sąsiednich transmisji optycznych.

2.2. Alokacja zasobów EON dla różnych klas ruchu

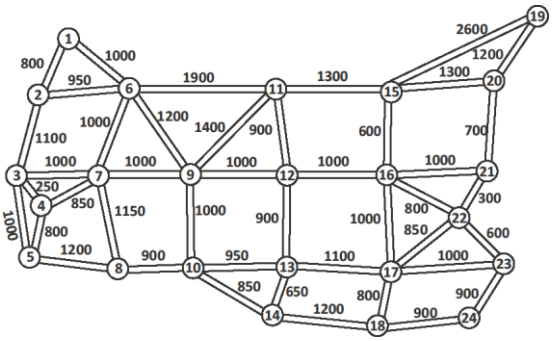
Proponowaną nowością w tym referacie jest umożliwienie zestawianie dodatkowych ścieżek w zależności od priorytetu zgłoszenia. W chwili wystąpienia przeciążenia w warstwie wirtualnej proponowany mechanizm na podstawie priorytetu ruchu zestawia dodatkowe ścieżki uwzględniając zajętość widma optycznego. Dla ruchu o wysokim priorytecie będzie zestawiał najkrótsze ścieżki w celu minimalizacji opóźnień. Natomiast dla ruchu o niskim priorytecie proponowany mechanizm zestawia ścieżki optyczne wzdłuż łączy, których zasoby są zajęte poniżej zakładanego współczynnika wykorzystania. Biorąc pod uwagę ruch generowany w sieci złożony z 20% ruchu o wysokim priorytecie i 80% ruchu o niskim priorytecie, ustalono współczynnik wykorzystania równy 80%. Ruch o wysokim priorytecie ma dostępne całe spektrum zarezerwowane na dynamiczne ścieżki. Dla tego ruchu wybierane są najkrótsze ścieżki, dla których możliwe jest zestawienie ścieżki bypass. Dla żądań o niskim priorytecie mechanizm alokuje zasoby łączy, których spektrum zajęte jest poniżej 80%.

3. ANALIZA SYMULACYJNA

W celu oszacowania wydajności proponowanego mechanizmu przeprowadzono liczne eksperymenty symulacyjne. Pomiarów dokonano w sieci IP/EON dla 3 przypadków: bez mechanizmu bypass („tylko warstwa IP"), z mechanizmem podstawowym bypass oraz z proponowanym ulepszeniem mechanizmu bypass (wybór spektrum do zestawienia ścieżki optycznej na podstawie klasy ruchu i zajętości spektrum). W przeprowadzonych eksperymentach symulacyjnych użyto topologii sieci NSF (15 węzłów, 46 skierowanych łączy) oraz UBN (24 węzły, 86 łączy skierowanych). Daną topologię przyjęto jako topologię warstwy IP oraz topologię warstwy optycznej. Rysunki 1 i 2 przedstawiają kolejno topologie NSF i UBN. Liczby opisujące łącza określają ich długość w kilometrach. Tabela 1 zawiera parametry charakterystyczne dla topologii sieciowych. Warto zauważyć, że sieć UBN ma wyższy średni stopień węzła w porównaniu z siecią NSF.



Rys. 1. Topologia sieci NSF



Rys. 2. Topologia sieci UBN

Dostępne pasmo optyczne włókna światłowodowego (pasmo C, 4 THz) zostało podzielone na 320 szczelin, każda o szerokości równej 12,5 GHz [5]. Do oddzielenia sąsiednich transmisji przeznaczono pasmo ochronne *GB* równe jednej szczelinie. Format modulacji wybrany dla sygnału optycznego wyznaczono na podstawie odległości na jaką przesyłany był sygnał. Dla sygnału optycznego rozważano cztery formaty modulacji: BPSK (o efektywności widmowej wyrażonej w bit/s/Hz M_p=1), QPSK (M_p=2), 8QAM (M_p=3), 16QAM (M_p=4). Podobnie jak w [1], [2] i [3] dla modulacji BPSK przyjęto maksymalny zasięg optyczny równy 9600 km, dla QPSK 4800 km, dla 8QAM 2400 km, a dla 16-QAM 1200 km.

Tab. 1. Parametry topologii sieci

Parametr	*NSF*	*UBN*
Liczba węzłów	15	24
Liczba łączy	46	86
Średnia długość łącza	1022 km	997 km
Średni stopień węzła	3.07	3.58

Dla mechanizmów bypass 160 szczelin zostało zarezerwowanych dla zestawienia dynamicznych ścieżek i 160 szczelin widocznych dla warstwy IP do zestawienia wirtualnej topologii. Dla przypadku „tylko warstwa IP" 320 szczelin było dostępnych do zestawienia wirtualnej topologii. Korzystając z zasobów widocznych dla IP zestawiono ścieżki optyczne będące łączami wirtualnymi. Przepustowość poszczególnych łączy C_{IP} [Gbit/s] obliczono na podstawie (2):

$$C_{IP} = (N_{IP} - GB) \cdot M_p \cdot C_{FS} \tag{2}$$

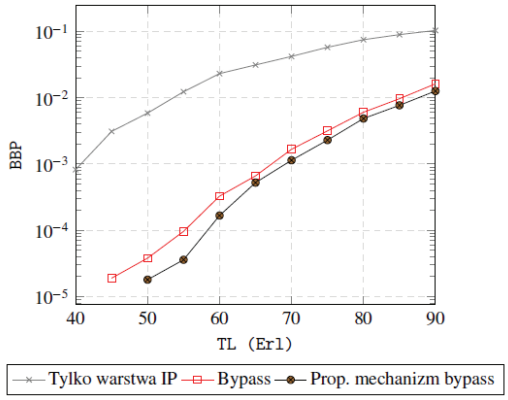
gdzie: N_{IP} – liczba alokowanych szczelin przez optyczną ścieżkę, która tworzy wirtualne łącze;

M_p [bit/s/Hz] – efektywność widmowa modulacji użytej dla optycznej ścieżki; C_{FS} [Gbit/s] – pojemność szczeliny dla modulacji BPSK; *GB* – pasmo ochronne. Należy zauważyć, że ścieżki zestawione dla warstwy wirtualnej mają długość łącza fizycznego. W przyjętym modelu transmisji podczas zestawienia wirtualnej topologii w NSF dla 34 ścieżek użyto modulacji 16QAM, dla 10 ścieżek 8QAM oraz dla 2 ścieżek modulacji QPSK. W UBN odpowiednio były to 76, 8 i 2 ścieżki. W warstwie IP zastosowano protokół rutingu OSPF. W przypadku dynamicznie zestawianej ścieżki typu bypass dla rutingu i alokacji zasobów ukrytych wykorzystano metodę SPF (Shortest Path First) i FFSA (First Fit Slot Allocation) z uwzględnieniem 10 najkrótszych ścieżek. Możliwe było zestawienie dodatkowej ścieżki bypass o długości jednego skoku IP.

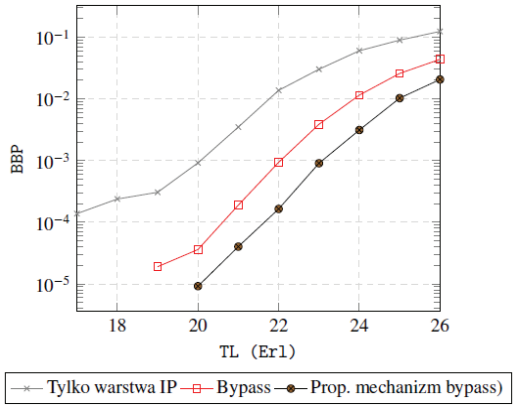
Ruch tła (stały) generowany był przez każdą parę węzłów w sieci i wynosił odpowiednio 200 Gbit/s dla NSF, 80 Gbit/s dla UBN. Ruch tła przesyłany był w warstwie wirtualnej przy użyciu protokołu OSPF. Dla wspomnianego ruchu tła, najbardziej obciążone łącze w warstwie wirtualnej w sieci gdzie wykorzystano mechanizmy bypass (160 szczelin widocznych dla IP i 160 szczelin dla bypass) było zajęte w 67% całkowitej pojemności łącza wirtualnego dla obu topologii. Ruch zmienny był generowany dynamicznie pomiędzy wybranymi węzłami *V*, dla których średnia długość najkrótszych ścieżek do dowolnego innego węzła była najmniejsza [5]. Ruch zmienny generowały węzły o numerach $V=\{12,13,14,15\}$ w NSF, w UBN o numerach $V=\{9,10,12,13,16\}$. Żądania nadchodziły do sieci zgodnie z rozkładem Poissona, ze średnią intensywnością λ zgłoszeń na jednostkę czasu i były równomiernie rozłożone wśród skierowanych par węzłów *V*. Żądana przepływność była zmienną losową o rozkładzie jednostajnym w przedziale od 50 Gbit/s do 1 Tbit/s z krokiem 50 Gbit/s [9]. Dla czasu trwania transmisji zastosowano rozkład wykładniczy o średniej $1/\mu$. Wyniki otrzymano z pomiaru 10^5 zgłoszeń w warunkach, w których sieć osiągnęła stabilny stan pracy (po czasie równym 5000 zgłoszeń). Symulacje powtórzono dla każdego punktu pomiarowego TL (obciążenia sieci, Traffic Load), obliczonego na podstawie λ/μ. Symulacje wykonano korzystając z dyskretnego dziennika zdarzeń środowiska OMNeT++ [8]. Ocenie poddano prawdopodobieństwo blokady przepływności (BBP) dla wszystkich żądań. Obliczono je jako iloraz sumy przepływności odrzuconych żądań do całkowitej żądanej przepływności dla danego typu ruchu.

Rysunki 3 i 4 przedstawiają wyniki symulacji mierzonego parametru BBP w skali logarytmicznej dla sieci NSF (rys. 3), dla sieci UBN (rys. 4). Zgodnie z rys. 3, implementacje mechanizmów bypass (bypass i proponowany mechanizm bypass) pozwoliły zmniejszyć BBP w sieci w porównaniu z referencyjnym modelem „tylko warstwa IP". Dla przypadku „tylko warstwa IP" pierwsza blokada w sieci wystąpiła dla wartości TL równej 40 Erl. Dla przypadku bypass było to 45 Erl, a dla proponowanego mechanizmu pierwsza blokada wystąpiła najpóźniej dla wartości TL równej 50 Erl. Wraz ze wzrostem TL, BBP dla „tylko warstwa IP" pozostaje nadal największe. Podobne zachowanie zaobserwowano dla UBN (rys. 4) – odpowiednio wartości 17 Erl, 18 Erl i 20 Erl. Dla przypadku „tylko warstwa IP" BBP jest największe zarówno w NSF jak i UBN. Spowodowane jest to brakiem możliwości przekierowania ruchu na inne ścieżki w chwili wystąpienia przeciążenia. Mechanizmy bypass pozwalają na zestawienie dodatkowych ścieżek w chwili wystąpienia przeciążenia i obsłużenia większej liczby żądań, co wpływa na redukcję BBP w sieci.

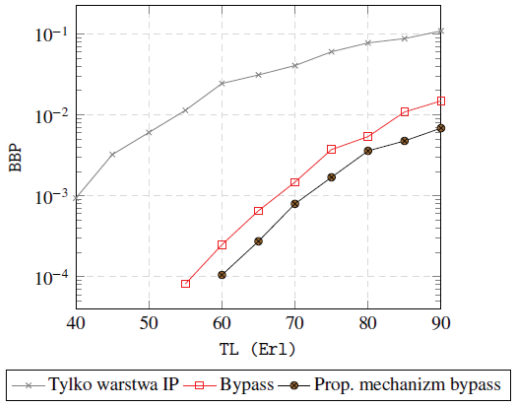
Rysunki 5 i 6 przedstawiają wyniki symulacji mierzonego parametru BBP dla ruchu o wysokim priorytecie dla sieci NSF (rys. 5) oraz dla sieci UBN (rys. 6). Rysunki 7 i 8 przedstawiają wyniki symulacji BBP dla ruchu o niskim priorytecie. Na rysunkach BBP jest przedstawione w skali logarytmicznej. Analizując uzyskane wyniki można stwierdzić, że pierwsze blokady żądań o wysokim



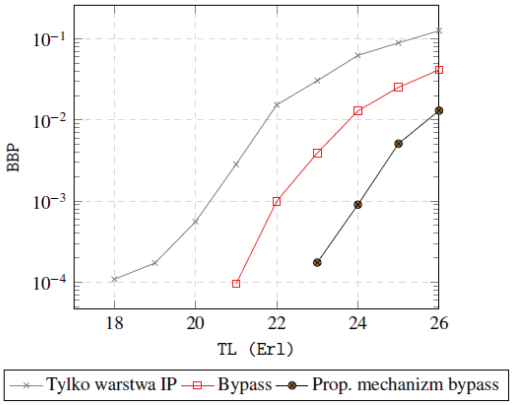
Rys. 3. Wyniki symulacji BBP w funkcji TL dla sieci NSF



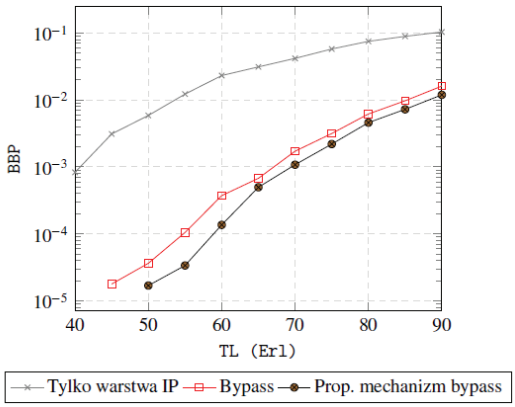
Rys. 4. Wyniki symulacji BBP w funkcji TL dla sieci UBN



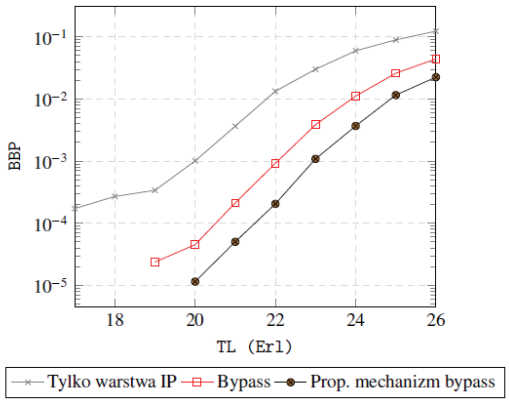
Rys. 5. Wyniki symulacji BBP dla ruchu o wysokim priorytecie w funkcji TL dla sieci NSF



Rys. 6. Wyniki symulacji BBP dla ruchu o wysokim priorytecie w funkcji TL dla sieci UBN



Rys. 7. Wyniki symulacji BBP dla ruchu o niskim priorytecie w funkcji TL dla sieci NSF



Rys. 8. Wyniki symulacji BBP dla ruchu o niskim priorytecie w funkcji TL dla sieci UBN

priorytecie jak i o niskim priorytecie pojawiają się najpóźniej dla proponowanego mechanizmu bypass, najwcześniej dla przypadku „tylko warstwa IP". Porównując wyniki otrzymane dla mechanizmów bypass, można stwierdzić, że proponowany bypass równoważy obciążenie w warstwie EON poprzez alokacje żądań ruchu o niskim priorytecie wzdłuż łączy o zajętości widma poniżej 80%. Dzięki temu więcej zgłoszeń ruchu o wysokim priorytecie zostaje przyjętych (rys. 5 i 6), a BBP dla ruchu o niskim priorytecie pozostaje co najmniej na tym samym poziomie co dla przypadku „zwykłego" bypass (rys. 7 i 8). Największy zysk w postaci zmniejszonego BBP dla ruchu priorytetowego można zaobserwować w sieci UBN (rys. 6). Sieć UBN ma wyższy stopień węzła w porównaniu do NSF (Tabela 1). Dlatego też w UBN można zaobserwować większy zysk dla proponowanego mechanizmu bypass wynikający z równoważenia obciążenia.

4. WNIOSKI

W pracy przedstawiono analizę wydajności mechanizmów zestawienia dodatkowych ścieżek optycznych w sieci IP/EON. Symulacje obejmowały dwie topologie: NSF i UBN. Poddano ocenie wydajność mechanizmów bypass w sieci dla ruchu o wysokim i niskim priorytecie. Analiza wykazała zysk w postaci zmniejszonego BBP w sieciach z mechanizmem bypass. Co więcej pokazano, że wprowadzone ulepszenie uwzględniające klasy ruchu i zajętość zasobów zwiększa wydajność mechanizmu zestawienia dodatkowych ścieżek. Wprowadzona modyfikacja prowadzi do równomiernej alokacji zasobów w sieci. Dzięki temu zmniejszone zostaje BBP oraz opóźnienia pojawienia się pierwszych blokad w sieci. Zysk ten widoczny jest szczególnie w sieci UBN, której stopień węzła jest wyższy w porównaniu do NSF. W przyszłości będą opracowane kolejne scenariusze symulacyjne i optymalizacja topologii warstwy wirtualnej.

LITERATURA

[1] Biernacka Edyta, Domżał Jerzy, Jurkiewicz Piotr, Wójcik Robert. 2017. „Model wielowarstwowej współpracy sieci IP z elastyczną siecią optyczną". *Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne,* nr 8-9 s. 984–987.

[2] Biernacka Edyta, Domżał Jerzy, Boryło Piotr. 2018. „Mechanizm bypass dla dynamicznego ruchu przesyłanego pomiędzy centrami danych w sieciach IP/EON". *Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne*, nr 8–9, s. 764–767.

[3] Biernacka Edyta, Domżał Jerzy, Wójcik Robert.019. „Analiza użyteczności mechanizmu bypass dla ruchu priorytetowego w sieciach SDN IP/EON. *Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne*, nr 7, s. 631–634.

[4] Christodoulopoulos K., Tomkos I., Varvarigos E. 2011. „Elastic bandwidth allocation in flexible OFDM-based optical networks". *Journal of Lightwave Technology* 29 (9) (2011) 1354-1366.

[5] Goścień Róża, Walkowiak Krzysztof. 2015. "Comparison of different data center location policies in survivable elastic optical networks". *7th International Workshop on Reliable Networks Design and Modeling (RNDM)*, 2015, pp. 48-55.

[6] ITU-T, Spectral grids for WDM applications: WDM frequency grid (G.694.1), ITU-T Recommendation G.694.1.

[7] Jinno Masahiko. 2017. „Elastic optical networking: Roles and benefits in beyond 100-gb/s era". *Journal of Lightwave Technology* 35 (5) 1116-1124.

[8] Omnet++ http://www.omnetpp.org.

[9] Klinkowski Mirosław, Walkowiak Krzysztof. 2019. "An Efficient Optimization Framework for Solving RSSA Problems in Spectrally and Spatially Flexible Optical Networks". *IEEE/ACM Transactions on Networking 1–13.*

PRZEGLĄD OTWARTO-ŹRÓDŁOWYCH NARZĘDZI DO ZARZĄDZANIA BEZPRZEWODOWYMI SIECIAMI LOKALNYMI[1]

REVIEW OF OPEN-SOURCE TOOLS
FOR THE MANAGEMENT OF LOCAL WIRELESS NETWORKS

Piotr Augustyniak[1]; Jakub Skóra[1]; Piotr Zwierzykowski[1];

[1]Politechnika Poznańska, Poznań, email: piotr.augustyniak@o2.pl
[2]Politechnika Poznańska, Poznań, email: jakub.skora@student.put.poznan.pl
[3]Politechnika Poznańska, Poznań, email: piotr zwierzykowski@put.poznan.pl

DOI: 10.15199/59.2022.4.41

Streszczenie**: W artykule przedstawiono przegląd otwarto-źródłowych narzędzi administracyjnych dla bezprzewodowych sieci lokalnych. W ramach pracy zaproponowano zbiór narzędzi administratora sieci WLAN.**
Abstract**: This article provides an overview of open source administrative tools for wireless LANs. As part of the work, an optimal set of WLAN administrator tools was proposed.**
Słowa kluczowe**: automatyzacja, zarządzanie bezprzewodowymi sieciami WLAN, kontrola wersji i oprogramowania, tworzenie kopii zapasowych i odzyskiwanie po awarii**
Keywords**: automation, WLAN network management, version and software control, backup and disaster recovery.**

1. WSTĘP

Dynamiczny i nieprzerwany rozwój sieci komputerowych nieustannie stawia nowe wyzwania przed administratorami sieci. Lawinowo rosnąca liczba urządzeń mobilnych, które mogą się ze sobą komunikować, rodzi wiele nowych wyzwań. Istnieje wiele topologii fizycznych i logicznych, a kombinacji i możliwych konfiguracji sprzętu oraz oprogramowania jest jeszcze więcej, dlatego wybór przez administratora odpowiednich narzędzi jest trudnym zadaniem. Na rynku oprogramowania do zarządzania siecią można wyróżnić dwie grupy programów. Programy komercyjne przeznaczone dla konkretnych urządzeń, lub rozwiązań, danego producenta oraz rozwiązania opensourcowe. Otwarto-źródłowe narzędzia są w większości przypadków darmowe i bardziej uniwersalne, ale przez to mniej rozbudowane.
Programy należące do każdej z tych grup, można podzielić według wielu różnych kryteriów. Ważna jest zasada i szybkość działania, liczba i zakres oferowanych funkcji i możliwości. Duże znaczenie ma również skalowalność, niezawodność, bezpieczeństwo oraz poziom trudności obsługi. Nie można zapominać o tym jakie systemy są wspieranie, jakie jest wsparcie twórców i/lub społeczności oraz jakie są wymagania sprzętowe. Dla każdego rozwiązania przedstawiono również słabe i mocne strony.

Artykuł powstał w oparciu o doświadczenia zebrane podczas przygotowania pracy dyplomowej przez jednego z autorów artykułu[2]. Głównym celem artykułu jest przedstawienie propozycji uniwersalnego zestawu narzędzi dla administratora bezprzewodowej sieci lokalnej. Przyjęto, że wszystkie proponowane rozwiązania będą oparte na oprogramowaniu o otwarto-źródłowej licencji.

Artykuł podzielono na 4 rozdziały. Rozdział 2 przedstawia zestawienie porównywanych rozwiązań pod kątem podstawowej funkcjonalności administracyjnej. W rozdziale 3 przedstawiono krótkie podsumowanie tendencji rozwoju narzędzi do zarządzania sieciami WLAN. Artykuł kończy krótkie podsumowanie.

Tab. 1. Porównanie programów realizujących funkcję zarządzania

Program	OpenWRT Remote Manager	OpenWISP	Cardinal
GUI/CLI	CLI	GUI	GUI
Wspierane urządzenia	Instancje OpenWRT	Instancje OpenWRT + możliwość pisania backendów konfiguracyjnego dla dowolnego typu sieciowego systemu operacyjnego	Punkty dostępowe Cisco w trybie autonomicznym z włączonym SSH
Główne funkcje	Zarządzanie zdalnymi instalacjami OpenWRT, które działają jako routery	Zarządzanie siecią zbudowaną na bazie OpenWRT	Kontroler punktów dostępowych
Mocne strony	Łatwa możliwość zarządzania portami, użytkownikami i interfejsami sieciowymi, tworzenie kopii zapasowych konfiguracji, całość napisana w Pythonie	Integracja z RADIUS, tworzenie tuneli VPN, możliwość konfiguracji sieci mesh oraz bezprzewodowych, wsparcie dla WPA2, WPA3 802.1x i innych, wielodrogowe powiadomienia, modułowa budowa	Przyjemny w obsłudze interfejs użytkownika, prosta instalacja, do działania wymaga tylko MySQL, Python3 i Nginx
Słabe strony	Brak wsparcia, rozwoju i GUI	Początkowo przytłaczająca ilość modułów, złożona instalacja	Wsparcie tylko do punktów dostępowych Cisco w trybie autonomicznym
Dokumentacja i wsparcie społeczności	Średnie	Duże	Małe

2. RÓŻNORODNOŚĆ FUNKCJONALNOŚCI

Badanie przeprowadzono w następujących obszarach: zarządzanie, automatyzacja, tworzenie kopii zapasowych i odzyskiwanie po awarii, kontrola wersji i oprogramowania, mapowanie zasobów, serwer wydruku, monitorowanie, autoryzacja, wydajność łącza, zarządzanie logami, zarządzanie zgłoszeniami, analiza pakietów, skanowanie sieci oraz oprogramowanie dla routerów.

Do **zarządzania** wybierano OpenWISP [1] (Tab.1). Jest to najbardziej kompletne spośród analizowanych rozwiązań [2][3]. Posiada znaczne wsparcie i społeczność, co jest dużym jego dużą zaletą, szczególne dla mało doświadczonych administratorów. Jest to system modułowy, dzięki czemu można go swobodnie rozbudowywać dodając kolejne, niekoniecznie związane z zarządzaniem,

[1] *Badania zostały sfinansowane ze środków Ministerstwa Edukacji i Nauki (Projekt 0313/SBAD/1307)*
[2] *W artykule wykorzystano m.in. doświadczania Jakub Skóra: „Otwarto źródłowa skrzynka narzędziowa administratora sieci WIFI - ujęcie administracyjne", Politechnika Poznańska, praca dyplomowa 2022*

moduły. Posiada przejrzysty i intuicyjny interfejs użytkownika.

Tab. 2. Porównanie programów realizujących funkcję automatyzacji

Program	Ansible	Terraform	Chef	Puppet	SaltStack
Wspierane platformy – serwer	Uniksowe	Linux	Linux	Linux	Linux
Wspierane platformy – agent	Nie wymaga	Nie wymaga	Uniksowe, Windows	Uniksowe, Windows	Uniksowe, Windows
Najlepiej stosować do	Zarządzanie konfiguracją	Zarządzanie zasobami infrastruktury	Zarządzanie konfiguracją	Zarządzanie konfiguracją	Zarządzanie konfiguracją
Infrastruktura	Zmienna	Niezmienna	Zmienna	Zmienna	Zmienna
Język	Proceduralny	Deklaratywny	Proceduralny	Deklaratywny	Deklaratywny
Mocne strony	Łatwy do nauczenia, brak agenta, Playbooks oparte na YAML, portal Ansible Galaxy – portal z treściami do Ansible	Duża elastyczność, możliwość generowania wykresów zależności	Dobra integracja z chmurą, duża elastyczność i dojrzałość	Solidny do zarządzania stanem serwerów, łatwe i solidne testowanie modułów	Wysoka wydajność dla dużych wdrożeń, obsługa na bazie agentów i bez
Słabe strony	Niewystarczający interfejs użytkownika	Brak funkcji cofania zmian	W obecnie sytuacji dobrze dostosowany tylko do AWS, zalecana znajomość Ruby	Brak kompleksowych funkcji raportowania, trudny dla początkujących zalecana znajomość Ruby	Interfejs sieciowy oferuję ograniczone możliwości i funkcje
Dokumentacja i wsparcie społeczności	Duże	Średnie	Średnie	Duże	Średnie

Do **automatyzacji** wybrano Ansible'a [4]. Główną zaletą tego rozwiązania jest to, że do działania nie potrzebuje agentów na systemach końcowych, łączy się z nimi za pomocą SSH lub PowerShell. Dodatkowo używa bardzo prostego formatu zapisu (YAML). Korzyścią płynąca z zastosowania Ansible jest również idempotentność stanu docelowego (Tab.2)

Tworzenie kopii zapasowych wybrano **Bacula** [5] (Tab. 3). Dzięki dobrej skalowalności **Bacula** może być stosowana w dużych i małych środowiskach. Jest to rozwiązanie dojrzałe rozwiązaniem (cechuje się dobrą skalowalnością, obsługuje szyfrowanie TLS, pozwala na kompresje danych a także weryfikację ich integralności), które może konkurować z rozwiązaniami komercyjnymi. Dużą jego zaletą jest dostępność zaawansowanych funkcji zarządzania pamięciom masową, które pomagają w odzyskiwaniu uszkodzonych i/lub utraconych plików.

Tab. 3. Porównanie programów realizujących kopii zapasowej i odzyskiwania po awarii

Program	Amanda	Duplicati	Bacula	BackupPC	luckyBackup	Box Backup	BURP
GUI/CLI	CLI	GUI, CLI	GUI (Baculum), CLI	GUI	GUI, CLI	GUI (Boxs), CLI	CLI
Wspierane platformy – klient	Windows, Linux, macOS	Windows, Linux, macOS	Windows, Linux, macOS	Windows, Linux, macOS	Linux	Windows, Linux, macOS	Windows, Linux, macOS
Wspierane platformy – serwer	Linux	Linux	Linux	Linux	Linux	Linux	Linux
Wymagane oprogramowanie po stronie klienta	Tak	Tak	Tak	Nie	Tak	Tak	Tak
Harmonogram kopii zapasowych	Tak	Tak	Tak	Tak	Tak	Tak	Tak
Mocne strony	Duże możliwości konfiguracji	Można pominąć szyfrowanie	Własna implementacja tworzonych plików	Można wznowić w przypadku przerwania, nie wymaga oprogramowania po stronie klienta	Bardzo łatwa obsługa	Zapewnia niezawodność bez skomplikowanej konfiguracji lub drogiego sprzętu	Można wznowić w przypadku przerwania
Słabe strony	Czasem niedopracowany interfejs	Słabe wsparcie	Duża złożoność konfiguracji, brak kopii przyrostowych	Tylko GUI	Nie jest już rozwijany	Brak bardziej zaawansowanych funkcji dostępnych u innych	Brak bardziej zaawansowanych funkcji dostępnych u innych
Dokumentacja i wsparcie społeczności	Średnie	Średnie	Duże	Średnie	Małe	Małe	Średnie

Przegląd programów umożliwiających **kontrolę wersji i oprogramowania** przedstawiono w Tab. 4. W tej kategorii wybrano program Oxidized [6]. Dużą zaletą tego rozwiązania jest łatwość dodawania nowych modeli urządzeń oraz dobra integracja z narzędziami do monitorowania sieci.

Do **mapowania zasobów** wybrano Sambę [7] (Tab. 6). Rozwiązanie to jest wieloplatformowe, łatwe w konfiguracji, korzysta z bardzo znanego, szeroko stosowanego protokołu SMB. TrueNAS CORE natomiast będzie wybrany jako system operacyjny dla NAS [8]. Posiada modułową budowę, dzięki czemu może być rozbudowywany o dodatkową funkcjonalność. Wspiera wszystkie popularne systemy plików oraz OpenZFS [9].

Tab. 4. Porównanie programów realizujących funkcję kontroli wersji i oprogramowania

Program	Rancid	Oxidized	rConfig v3	Netshot
GUI/CLI	CLI	CLI, częściowo GUI	GUI	GUI
Wspierane platformy	Uniksowe	Linux, Docker	Centos 7+	Linux, Docker
Liczba wspieranych sieciowych systemów operacyjnych	Ponad 30	Ponad 130	Ponad 15	Ponad 25
Mocne strony	Bardzo dojrzały i dobrze działający program, dobra skalowalność	Wsparcie dla bardzo dużej ilości rozwiązań, wspiera pull i push dla zmian konfiguracji, dobra skalowalność, częste aktualizacje	Przejrzysty interfejs webowy, wbudowane raportowanie, monitorowanie pod kątem zgodności	Intuicyjny interfejs webowy, szybko rozwiązywane problemy, aktywnie rozwijany, możliwość integracji – REST API
Słabe strony	Działa tylko na modelu pull, brak GUI	Brak GUI	Nowsza wersja jest płatna, brak wsparcia i aktualizacji dla tej wersji	Niepłynne i problematyczne konfigurowanie niektórych urządzeń
Dokumentacja i wsparcie społeczności	Duże	Średnie	Małe	Średnie

Tab. 5. Porównanie programów realizujących funkcję mapowania zasobów – programy i NAS

Program	Samba	OpenSSH + SSHFS	OpenSSH + Filezilla	OpenSSH + WinSCP	NFS
Główne zadanie	Mapowanie zasobów	Mapowanie zasobów	Mapowanie zasobów	Mapowanie zasobów	Mapowanie zasobów
GUI/CLI	CLI	CLI	GUI, CLI	GUI, CLI	CLI
Wspierane platformy	Windows, Linux, macOS	Uniksowe, (istnieje port na Windows)	Windows, Linux, macOS	Windows	Windows, Linux, macOS
Wspierane protokoły	SMB	SFTP	SFTP, FTPS, FTP	SFTP, SCP, FTP	NFS
Mocne strony	Łatwa konfiguracja dla Windowsa jak i Linuxa	Wystarczy dostęp przez SSH do działania	Nie wszystkie zaawansowane funkcje w wersji płatnej	Zintegrowany edytor tekstu, przyjazny interfejs	Bardzo łatwa konfiguracja
Słabe strony	Brak wsparcia producenta, użycie netbios – mało wydajny protokół	Niewydajne działanie dla dużego opóźnienia	Wyższy próg wejścia niż u innych	Niska wydajność	Nie jest zbyt bezpieczny bez zapory
Dokumentacja i wsparcie społeczności	Średnie	Średnie	Duże	Średnie	Małe

TrueNAS CORE	OpenMediaVault	XigmaNAS
NAS	NAS	NAS
GUI, CLI	GUI, CLI	GUI, CLI
VM	VM	VM
SMB, AFP, NFS, iSCSI, SSH, rsync, FTP/TFTP	SMB, FTP, NFS, SSH, rsync, iSCSI, AFP, TFTP	SMB, AFP, NFS, FTP/TFTP, rsync, SCP, iSCSI, Unison
Szyfrowanie na wolumenach ZFS, oferuje wiele przydatnych wtyczek	Obsługuję agregację łączy, IPv6 i Wake on LAN	Szyfrowanie dysków z obsługą kryptografiiczmej akceleracji sprzętowej
Wymaga dużej ilości pamięci RAM, zwłaszcza przy używaniu ZFS	Brak funkcji Enterprise-NAS	Interfejs nie jest przyjazny dla użytkownika
Duże	Średnie	Średnie

Obsługę **serwera wydruku** powierzono CUPS [10]. Program ten co prawda działa tylko na systemach unixowych, ale za pomocą Samby może również obsługiwać drukarki na Windows. Pozwala komputerowi stać się serwerem wydruku i zarządzać podstawowymi zadaniami drukowania. Posiada graficzny interfejs webowy, dzięki któremu konfiguracja jest intuicyjna i łatwa. Warto dodać, że program ten oferuje najwięcej funkcji i posiada obszerną dokumentację.

Tab. 6. Porównanie programów realizujących funkcję serwera wydruku

Program	CUPS	LPRng	PyKota
Wspierane platformy	Uniksowe	Uniksowe	Uniksowe
Główne funkcje	Zapewnia funkcje serwera wydruku	Zapewnia funkcje serwera wydruku	Zarządzanie drukowaniem
Działanie z drukarkami z Windows	Przy pomocy Samby	Przy pomocy Samby	Przy pomocy Samby
Mocne strony	Posiada GUI do obsługi, integracja wielu metoda uwierzytelniania, kompatybilność z innymi systemami drukowania, modułowa architektura, integracja z LDAP, obsługa szyfrowania SSL3 i TLS	Posiada interfejs graficzny (LPRng Tool), automatyczne utrzymywanie pracy, bardzo szczegółowa diagnostyka, wiele drukarek obsługujących jedną kolejkę, kompatybilność z innymi buforami wydruku i drukarkami sieciowymi, które używają LPR.	Działa z CUPS i LPRng, zarządzanie przydziałami, kolejkami, oraz użytkownikami, rodlicznie drukowanie
Słabe strony	Płatna wersja na urządzenia firmy Apple, konfiguracja z linii poleceń	GUI nie działa idealnie, czasem nie chce drukować z teoretycznie wspieranych drukarek	Brak wsparcia, narzędzie nie jest już rozwijane
Dokumentacja i wsparcie społeczności	Duże	Średnie	Małe

Do **monitowania** sieci wybrano Zabbix [11] (Tab.7). Zbiera, analizuję i wizualizuje dane. Wykrywa anomalie i powiadamia administratora za pomocą alarmów. Instalacja i konfiguracja programu nie jest trudna dzięki obszernej dokumentacji oraz wsparciu społeczności użytkowników programu.

Tab. 7. Porównanie programów realizujących funkcję monitorowania

Program	Zabbix	Cacti	Prometheus	Nagios Core	Icinga	LibreNMS	Checkmk Raw Edition	NetXMS	OpenNMS
Wspierane platformy – agenci	Windows, Linux, macOS, Docker i inne (SNMP)	Windows, Linux i inne (SNMP)	Windows, Linux i inne (SNMP)	Windows, Linux i inne (SNMP)	Windows, Linux i inne (SNMP)	Windows, Linux i inne (SNMP)	Windows, Linux, macOS i inne (SNMP)	Windows, Linux i inne (SNMP)	SNMP
Platforma serwera	Linux, Docker	Linux	Linux, Docker	Linux	Windows i Linux	Linux, VM	Linux, Docker, VM	Windows i Linux	Linux, Docker
Mocne strony	Bardzo wydajny na niskich poziomach	Bardzo wiele możliwości dostosowywania	Idealny dla sieci kontenerowych, bardzo szybka baza danych	Bardzo duża biblioteka wtyczek, prosty interfejs	Ulepszony technicznie Nagios, działa większość wtyczek i rozszerzeń Nagiosa	Wysoce konfigurowalny moduł ostrzegania, dobre skalowanie	Konfiguracja i rozwiązywanie problemów bardziej intuicyjne niż u innych	Łatwa instalacja, przyjemny i elastyczny interfejs użytkownika	Elastyczne i wydajne narzędzie, dobrze integruje się z systemami alertów
Słabe strony	Gorsza wydajność powyżej 1000 węzłów, ograniczone wsparcie	Słabo się skaluje, brakuje obecnych u innych popularnych funkcji	Kłopotliwa i czasochłonna pierwsza konfiguracja, głównie używany dla metryk	Jakościowo gorszy niż płatna wersja	Wyższa, stroma krzywa uczenia się niż w innych	Stosunkowo duże zużycie pamięci	Braki i nieaktualna dokumentacja	Brak wersji komercyjnej, czasem mało intuicyjny	Stroma krzywa uczenia się, nieintuicyjny interfejs użytkownika
Dokumentacja i wsparcie społeczności	Duże	Średnie	Średnie	Duże (bezpłatna pomoc przy wdrażaniu)	Duże	Średnie	Średnie	Duże	Duże

Wszystkie dane są wizualizowane na czytelnych dashboardach. Zabbix obsługuję wiele protokołów łączności oraz daje możliwość uruchamiania skryptów. Można go zintegrować z innymi rozwiązaniami np. do wizualizacji danych.

Za **autoryzację** odpowiedzialny będzie FreeRadius [12]. Jest to najczęściej wdrażany i najpopularniejszy serwer RADIUS. Posiada najwięcej funkcji oraz największą dokumentację spośród wszystkich analizowanych rozwiązań. FreeRadius jest szybki, bogaty w funkcje, dobrze się skaluję i ma budowę modułową (Tab.8).

Tab. 8. Porównanie programów realizujących funkcję autoryzacji

Program	FreeRadius	freeDiameter	tac_plus	Kerberos
Wspierane platformy – serwer	Uniksowe	Uniksowe	Linux	Uniksowe, Windows
Główne zadanie	Uwierzytelnianie, autoryzacja i kontrola dostępu	Uwierzytelnianie, autoryzacja i kontrola dostępu	Uwierzytelnianie, autoryzacja i kontrola dostępu	Uwierzytelnianie
Protokół	RADIUS	Diameter	TACACS+	Kerberos
Mocne strony	Możliwość integracji i innymi narzędziami ułatwiającymi zarządzanie, lepsza kontrola dostępu niż w TACACS+, Vendor Specific Attributes dla prawie stu dostawców, obsługa przełączania awaryjnego, równoważenia obciążenia	Wsparcie dla SCTP i TCP, architektura peer-to-peer zapewnia duża elastyczność, wbudowane mechanizmy przełączania awaryjnego, obsługa tłumaczenia RADIUS/Diameter	Obsługa przez większość producentów, dobra skalowalność, używa protokołu TCP, cały pakiet jest szyfrowany	Wsparcie dla wielu systemów operacyjnych, wzajemnie uwierzytelnianie, silna kryptografia, udostępniane są tajne klucze – jest bardziej wydajne niż udostępnianie kluczy publicznych.
Słabe strony	Mniejsza niezawodność niż TACACS+ - UDP, szyfrowanie tylko hasła	Brak aktywnego wsparcia i aktualizacji, brak szyfrowania – bezpieczna transmisja za pomocą TLS lub IPSec	Tylko Linux, wsparcie dla kontroli dostępu niż w RADIUS, małe wsparcie	Ściśle wymagania czasowe, długa konfiguracja, tylko uwierzytelnianie
Dokumentacja i wsparcie społeczności	Duże	Małe	Małe	Średnie

Wydajność łącza będzie badana za pomocą Iperf3 (Tab.9) . Jest to bardzo często używane narzędzie, które może mierzyć przepustowość TCP, UDP i SCTP. Program ten ma stosunkowo małe wymagania sprzętowe i jest łatwy w obsłudze. Działa na wielu platformach.

Tab. 9. Porównanie programów realizujących funkcję badania wydajności łącza

Program	Iperf3	Netperf	Packet Sender	Mausezahn	nping
GUI/CLI	GUI(JPerf), CLI	CLI	GUI	CLI	CLI
Wspierane platformy	Linux, Windows (nieoficjalnie)	Windows, Linux, macOS	Windows, Linux, macOS	Linux	Windows, Linux, macOS
Główne zadanie	Pomiar wydajności i przepustowości sieci	Pomiar wydajności i przepustowości sieci	Wysyłanie i odbieranie pakietów w celach testowych	Tworzenie pakietów w celach testowych	Wysyłanie pakietów w celach i analiza odpowiedzi testowych
Mocne strony	Bardzo prosta obsługa, wiele możliwości dostosowywania parametrów, wielowątkowość, obsługa połączeń multicastowych oraz IPv6	Zapewnia szereg predefiniowanych testów, wsparcie dla IPv6, podaje dane dot. wykorzystania procesora	Automatyzacja i skryptowanie środowiska sieciowego, wbudowane czytniki ASCII i HEX, skalowanie w górę i w dół	Możliwość tworzenia prawie każdego możliwego i niemożliwego pakietu ,rozwijany w ramach netsniff-ng	Niestandardowe generowanie pakietów ARP, TCP, UDP i ICMP, możliwość śledzenia trasy, elastyczny i łatwy do dostosowywania
Słabe strony	Brak kompatybilności wstecznej z Iperf2	Małe wsparcie, brak aktualizacji dodających nowe funkcje	Mniej niezawodne niż rozwiązania komercyjne, nie dostosowywany do potrzeb biznesowych	Dostępny tylko na Linux oraz brak w planach wersji na Windows	Eksperymentalne wsparcie dla IPv6
Dokumentacja i wsparcie społeczności	Duże	Małe	Średnie	Małe	Średnie

Logi będą zarządzane przez Graylog [13]. Jest wiodące i niezawodne narzędzie do scentralizowanego przechwytywania, przechowywania i analizy logów. Pozwala ono na łatwe zarządzanie nieustrukturyzowanymi i ustrukturowanymi logami. Dużą zaletą programu jest wykorzystanie bazy danych Elasticsearch, która bardzo szybko przeszukuję rekordy (Tab. 10).

Do **zarządzania zgłoszeniami** wybrano OSTicket [14]. Program ten jest jednym z najłatwiejszych w użytkowaniu i konfiguracji. Posiada wiele funkcji pomagających w komunikacji ze zgłaszającymi oraz pomagających w zarządzaniu zgłoszeniami. Wiele z nich jest zautomatyzowanych m.in wysyłanie potwierdzeń po otwarciu zgłoszenia. Pozwala na dostosowywanie wszystkiego czego potrzebujemy od szablonów wiadomości poczty elektronicznej, aż po uprawnienia agentów obsługujących zgłoszenia. Dobrze się skaluje.

Tab. 10. Porównanie programów realizujących funkcję zarządzania logami

Program	Octopussy	Graylog	AlienVault OSSIM	SIEMonster Community Edition	Prelude OSS
Wspierane platformy	Linux	Linux, Docker	VM i fizyczne	VM, Docker	Uniksowe
Główne zadanie	Zarządzanie zdarzeniami	Zarządzanie zdarzeniami	SIEM	SIEM	SIEM
Mocne strony	Obsługa wtyczek, wielojęzyczność, niestandardowe powiadomienia, rozbudowane wbudowane narzędzie do tworzenia raportu o incydentach	Duża elastyczność, i szybkość działania, możliwość integracji z innymi rozwiązaniami, łatwe skalowanie, potrafi przetwarzać duże ilości danych	Korelacja zdarzeń SIEM, wykrywanie zasobów, ocena podatności, monitorowanie zachowania, wykrywanie włamań	Analiza zagrożeń i raportowanie w czasie rzeczywistym	Obsługuję wiele formatów logów, korelacja, filtrowanie, ostrzeganie, możliwości analizy i wizualizacji
Słabe strony	Mała pomoc, nie jest już rozwijany, ma błędy związane z bezpieczeństwem	Dashboardy i wykresy nie są już tak elastyczne, nie mogą korzystać z całego ekosystemu ELK	Mało funkcji zarządzania dziennikami, problemy z wydajnością na dużą skalę, możliwość wdrożenia tylko dla jednego serwera	Limit do 100 urządzeń końcowych, w wersji community brak wielu funkcji, wybrakowana dokumentacja online	Słabsza wydajność, skalowalność i mniej funkcji niż inne SIEMy, przeznaczony głównie do testów i/lub celów badawczych
Dokumentacja i wsparcie społeczności	Małe	Duże	Duże	Małe	Średnie

Tab. 11. Porównanie programów realizujących funkcję zarządzania zgłoszeniami

Program	OTRS Community Edition	OTOBO	osTicket	Zammand	UVdesk
Wspierane platformy	Uniksowe	Linux, Docker	Wszystkie	Linux, Docker	Windows, Linux, macOS
Mocne strony	Automatyzacja procesów i przepływu pracy, przyjazny dla użytkowników mobilnych, konfigurowalne portale klientów i agentów, obsługuję 38 różnych języków	Dashboardy z wieloma ważnymi informacjami, kompaktowe i przejrzyste przeglądanie ticketów, dużo statystyk i możliwości raportowania, łatwe dostosowywanie do własnych potrzeb, możliwość przejścia z OTRS	Duże możliwości filtrowania i sortowania, małe w formacie HTML, niestandardowe pola w zgłoszeniach, konfigurowalny schemat automatycznych odpowiedzi	Częste aktualizacje, zewnętrzne uwierzytelnianie, historia zmian w zgłoszeniach, duża wydajność i elastyczność, pozwala na generowanie raportów	Łatwa instalacja i konfiguracja, duża elastyczność, szybkość i możliwość konfiguracji, wiele kanałów kontaktu
Słabe strony	W styczniu 2021 rozwój i wsparcie kończą się, trochę stary interfejs użytkownika, wiele przydatnych funkcji w płatnej wersji	Nowe oprogramowanie - stosunkowo mało dokumentacji, wiele przydatnych funkcji w płatnej wersji	Każda zmiana personalizacji wymagana ponownego wypełnienia formularza, brak raportowania	Skomplikowana konfiguracja, brak wersji płatnej, stosunkowo nowe oprogramowanie	Słaba identyfikacja spamu
Dokumentacja i wsparcie społeczności	Średnie	Średnie	Duże	Duże (wsparcie w ponad 30 językach)	Średnie

Najlepszym rozwiązaniem do **analizy pakietów** będzie Wireshark [15] (Tab. 12). Posiada bardzo rozbudowany i czytelny interfejs graficzny oraz odpowiednik uruchamiany z poziomu linii poleceń (TShark). Przechwycony ruch można sortować i odfiltrowywać wedle szukanych parametrów.

Nmap [16]. Program posiada interfejs graficzny o nazwie Zenmap. Jest bardzo elastyczny i ma składnie łatwą do nauczenia. Jego działanie polega na tworzeniu pakietów i wysyłaniu ich do celu oraz na analizie odpowiedzi na nie. Dzięki temu może m.in. wykrywać hosty w sieci, wyliczać otwarte porty oraz wykrywać systemy operacyjny celów.

Tab. 12. Porównanie programów realizujących funkcję analizy pakietów

Program	Wireshark	tcpdump	EtterCap	netsniff-ng	Ngrep	EtherApe	TCPTrace	Xplico
GUI/CLI	GUI, CLI (TShark)	CLI	GUI, CLI	CLI	CLI	GUI	CLI	GUI, CLI
Wspierane platformy	Uniksowe, Windows	Uniksowe, Windows (WinDump)	Uniksowe, Windows	Linux	Uniksowe, Windows	Uniksowe	Linux	Linux, VM
Główne zadanie	Przechwytywanie i analiza pakietów	Przechwytywanie i analiza pakietów	Przechwytywanie i analiza pakietów	Analizator ruchu + zestaw narzędzi sieciowych	Analiza pakietów sieciowych	Graficzna reprezentacja ruchu + monitorowanie	Analiza plików zrzutów TCP i UDP	Analiza śledcza sieci
Mocne strony	Zbieranie i analiza pakietów w jednym, przyjazny interfejs	Lekka i szybka aplikacja, wiele możliwość dostosowania parametrów, nie wymaga GUI	Testy penetracyjne, oferuje metody izolowania punktów końcowych	Wszechstronne narzędzie składające się z wielu programów, szybkie działanie analizatora	Składnia kompatybilna z grep, dane prezentowane w przyjazny sposób	Graficzne podejście do pokazania natężenia ruchu, odczytywanie ruchu na żywo i z pliku	Generowanie kilku typów danych wejściowych, tworzenie wykresów	Limit na dane to limit dysku twardego, możliwości dostosowywania
Słabe strony	Stroma krzywa uczenia się, domyślnie zbiera wszystko, co może być przytłaczające	Skomplikowany język zapytań filtrowania	Lekko przestarzały interfejs	Działa tylko na Linux, brak GUI	Powoduje większą liczbę fałszywych alarmów, wolniejsze działanie od innych snifferów	Nie jest łatwy w użytkowaniu	Brak wsparcia i aktualizacji, minimalne przetwarzanie UDP od wersji 5	Działa tylko na Linux, ograniczenie w postaci czasu dostępu do dysku twardego
Dokumentacja i wsparcie społeczności	Duże	Duże	Średnie	Średnie	Średnie	Średnie	Małe	Średnie

Skanowanie sieci (Tab. 13) będzie wykonywane przez

W kategorii **oprogramowaniem routerów** wybrano OpenWRT [17] (Tab. 14). Jest to najstarszy alternatywy system operacyjny dla routerów SOHO. Bazuje na Linuxie. Posiada wiele funkcji różnorodnych funkcji tj. obsługa tuneli VPN, serwer danych, obsługa zewnętrznego modemu 3G, obsługa QoS, czy zbieranie statystyk dotyczących pracy urządzenia. Jego dużą zaleta jest wielkość społeczność użytkowników i jest regularna aktualizacja.

Tab. 13. Porównanie programów realizujących funkcję skanowania sieci

Program	Nmap	Zmap	p0f	Masscan
GUI/CLI	GUI(Zenmap), CLI	GUI, CLI	CLI	CLI
Wspierane platformy	Uniksowe, Windows	Linux, macOS	Uniksowe, Windows	Windows, Linux, macOS
Tryb skanowania	TCP SYN, TCP, UDP	Bezstanowy	Pasywny	Bezstanowy
Typowe zastosowania	Wykrywanie hostów, skanowanie otwartych portów	Skanowanie i znajdowanie otwartego portu – pojedynczego portu	Odcisk palca stosu TCP/IP	Skanowanie i znajdowanie otwartego portu – wielu portów
Szczegółowość skanowania	Duża	Niska	Średnia	Niska
Mocne strony	Bardzo wszechstronne narzędzie, może być używany do audytu systemów sieciowych, omijanie zapory i/lub IDS,	Bardzo szybkie skanowanie – pełna przestrzeń adresowa w czasie poniżej 45 minut (gigabajtowe łącze) lub 5 minut (łącze 10 Gigabajt i PF_RING)	Nie generuje ruchu, przez co jest trudniejszy lub nie możliwy do wykrycia, obsługa pełnych wyrażeń filtrujących w stylu tcpdump, duża szybkość	Może przesyłać do 10 miliomów pakietów na sekundę, kompatybilny z Nmap
Słabe strony	Skanowanie portów zostawia po sobie ślady – jest „głośne"	Sondowanie tylko jednego portu, może skanować tylko przestrzeni IPv4	Zdalny/skanowany host musi zainicjować połączenie TCP – skanowanie pasywne	Obsługuje tylko adresy IP i proste zakresy (nie obsługuje DNS)
Dokumentacja i wsparcie społeczności	Duże	Średnie	Średnie	Średnie

Tab. 14. Porównanie programów realizujących funkcję oprogramowania dla routerów

Program	DD-WRT	Tomato	OpenWRT
Mocne strony	Obsługuje dużą liczbę routerów, ma wiele opcji i funkcji do zaawansowanego dostosowywania, obsługuje OpenVPN i QoS	Szybkie działanie, nowoczesny interfejs, obsługa OpenVPN i monitorowania na żywo, zajmują mało miejsca	Duża liczba opcji kontroli i dostosowywania, obsługa OpenVPN i QoS
Słabe strony	Brak nowych wersji, może być przytłaczający dla nowych użytkowników	Ograniczona funkcjonalność, obsługa mniejszej ilości routerów niż konkurencja	Mniejsza liczba obsługiwanych routerów niż DD-WRT, mniej przyjazny dla użytkownika niż inne
Dokumentacja i wsparcie społeczności	Duża	Średnia	Duża

3. TRENDY I KIERUNKI ZMIAN

Liczba projektów otwarto-źródłowych będzie w przyszłości rosła, taka postać modelu biznesowego okazała się skuteczna dla wielu dostawców. Dążenie do otwartego oprogramowania będzie kontynuowane, zapewne nadal znaczna ilość firm będzie rozpowszechniać podstawową wersję swojego oprogramowania w taki sposób, dając możliwość zakupu płatnych dodatków i wsparcia. W nowych rozwiązaniach coraz większe znaczenie ma łatwość korzystania, a co za tym idzie przejrzystość i funkcjonalność interfejsu graficznego. Od wielu lat można zaobserwować wśród wielu administratorów tendencję do preferowania interfejsu graficznego zamiast tekstowego (CLI). Wynika to przede wszystkich z łatwości pracy, dlatego jako jeden z głównym czynników przemawiających za konkretnym rozwiązaniem jest intuicyjność obsługi interfejsu graficznego. Odbywa to się czasem kosztem liczby funkcji, czy wydajności. Wybierając rozwiązania otwarte warto pamiętać, że często społeczności tworzące otwarte oprogramowanie są źródłem wielu innowacji, dla przykładu: platforma konteneryzacji - Docker; platforma do zarządzania, automatyzacji i skalowania aplikacji kontenerowych - Kubernetes; środowisko GNOME 3; chmura i rozwiązania chmurowe; system operacyjny Chrome OS; czy wszechobecność w rozwiązaniach IoT. Obecnie wiele znanych firm o zasięgu globalnym inwestuję w oprogramowanie i rozwiązania otwarto-źródłowe. Programy obecnie używane stają się coraz starsze co często przekłada się na ich lepsze działanie, większą społeczność i mnogość funkcji. Również coraz większego znaczenia nabiera chmura, ważniejsza staję się integracja z nią.

4. PODSUMOWNIE

Celem artykułu było przedstawienie porównania i dokonanie wyboru aplikacji otwarto źródłowych, które mogą stanowić zbiór aplikacji wykorzystywanych przez administratora bezprzewodowych sieci lokalnych. W tym celu autorzy korzystając z własnych doświadczeń zdefiniowali obszary działań administratorów, a także określili rodzaje i zakresy funkcji poszczególnych narzędzi. Następnie wskazali rozwiązania spełniające większość przedstawionych warunków.

Wybierając aplikacje otwarto źródłowe warto pamiętać, że wiele spośród nich rozwija się dynamicznie. Czasem upływ kilku miesięcy może spowodować, że jedne propozycje stają się przestarzałe, czy też gorsze. Inne natomiast zyskują na popularności i zdobywają coraz to większe zainteresowanie w branży, co znajduje swoje odbicie we wzroście renomy i możliwości rozwoju. Niektóre nowe narzędzia opierają się na starszych, dobrze znanych rozwiązaniach, ale w nowszej, odświeżonej, zazwyczaj rozbudowanej i dużo przystępniejszej formie, która może idealnie odpowiadać obecnym trendom rozwoju technologicznego. Z tego powodu, przy wyborze danych programów kierowano się m.in. łatwością obsługi i konfiguracji, wspieranymi platformami oraz możliwością dostosowywania rozwiązania do bieżących potrzeb, a więc elastycznością i adaptacyjnością.

Opisując poszczególne narzędzia autorzy starali się prezentować naturalne i obiektywne podejście do demonstrowanych programów/narzędzi, tak by nie sugerować znanych i już wykorzystywanych rozwiązań. Warto również podkreślić, że dodatkowym celem prowadzonych prac była chęć przetestowania rozwiązań innych od tych powszechnie stosownych w sieciach produkcyjnym. Autorzy mają nadzieję, że takie nieco odmienne od powszechnie stosowanego podejścia, pozwoli na zwrócenie uwagi na rozwiązania mniej popularne o obiecujących cechach.

LITERATURA

[1] https://medium.com/@williamchoudhury/openwisp-blog-post-ac6987d6b734 [dostęp 02.05.2022]

[2] https://github.com/jumpscale7/openwrt-remote-manager [dostęp 02.05.2022]

[3] https://dataiku-research.github.io/cardinal/ [dostęp 02.05.2022]

[4] https://www.ansible.com/ [dostęp 02.05.2022]

[5] https://www.bacula.org/ [dostęp 02.05.2022]

[6] https://github.com/ytti/oxidized [dostęp 02.05.2022]

[7] https://www.samba.org/ [dostęp 02.05.2022]

[8] https://www.truenas.com/truenas-core/ [dostęp 02.05.2022]

[9] https://www.openzfs.org [dostęp 02.05.2022]

[10] https://www.cups.org/ [dostęp 02.05.2022]

[11] https://www.zabbix.com/ [dostęp 02.05.2022]

[12] https://freeradius.org/ [dostęp 02.05.2022]

[13] https://www.graylog.org/ [dostęp 02.05.2022]

[14] https://osticket.com/ [dostęp 02.05.2022]

[15] https://www.wireshark.org/ [dostęp 02.05.2022]

[16] https://nmap.org/ [dostęp 02.05.2022]

[17] https://openwrt.org/ [dostęp 02.05.2022]

PROBLEMY BADAWCZE I MODELOWANIE SIECI SENSOROWYCH W ŚRODOWISKACH MIEJSKICH*

RESEARCH PROBLEMS AND MODELING OF SENSOR NETWORKS IN URBAN ENVIRONMENTS

Bartosz Musznicki[1]; Maciej Piechowiak[2]; Piotr Zwierzykowski[3]

[1] Politechnika Poznańska, Poznań, rsrch@musznicki.com
[2] Uniwersytet Kazimierza Wielkiego, Bydgoszcz, maciej.piechowiak@ukw.edu.pl
[3] Politechnika Poznańska, Poznań, piotr.zwierzykowski@put.poznan.pl

DOI: 10.15199/59.2022.4.42

Streszczenie: **Badania nad topologią bezprzewodowych sieci sensorowych (WSN) koncentrują się głównie na modelowaniu i analizie metod optymalizacji struktury połączeń. Tego typu podejście jest jednak rzadko wykorzystywane przez projektantów sieci, a powstające systemy czujników w środowiskach miejskich często odbiegają od wzorców proponowanych w literaturze i dostępnych modelach. Zakłada się, że otwarty dostęp do rzeczywistych danych pochodzących z różnych typów urządzeń pozwoli zrozumieć naturę tych danych, a także wyodrębnić i rozwiązać problemy występujące w heterogenicznych sieciach. Analizowanie topologii takich sieci i diagnozowanie problemów powinno zatem opierać się na danych uzyskanych z rzeczywistych sieci. Dzięki takiemu podejściu możliwe jest zrozumienie zmian w typologii sieci oraz identyfikacja problemów w heterogenicznych sieciach bezprzewodowych. W artykule przedstawiono wybrane najważniejsze i zarazem aktualne tematy badawcze związane z miejskimi sieciami czujników. Zaprezentowano również koncepcję modelowania sieci w oparciu o publicznie dostępne dane o położeniu geograficznym węzłów w czasie rzeczywistym.**

Słowa kluczowe: **bezprzewodowe sieci sensorowe, sieci miejskie, routing oportunistyczny.**

Keywords: **wireless sensor networks, urban networks, opportunistic routing.**

1. WSTĘP

Bezprzewodowe sieci sensorowe (ang. *Wireless Sensor Networks*, *WSN*) składają się z wielu węzłów rozmieszczonych na pewnym obszarze w celu wykonania wspólnego zadania [6]. Podstawowymi elementami takich sieci są urządzenia wyposażone w czujniki, które monitorują zmienność zjawisk i wielkości fizycznych, takich jak: temperatura, wilgotność, obecność obiektu, dźwięk, ciśnienie, ruch, stopień zanieczyszczenia powietrza pyłem zawieszonym itp. Typowe zastosowania czujników w sieciach miejskich obejmują między innymi: koordynację pojazdów specjalistycznych (karetek pogotowia, pojazdów ratunkowych), logistykę transportu publicznego (w celu optymalizacji ruchu i wykorzystania zasobów instytucjonalnych), zarządzanie ruchem, monitorowanie parametrów środowiskowych określających jakość powietrza i wody, monitorowanie miejskich pojazdów wypożyczanych (samochodów elektrycznych, skuterów, rowerów).

Badania nad sterowaniem i kontrolą topologii bezprzewodowych sieci czujników koncentrują się na modelowaniu i analizowaniu metod, które mogą być potencjalnie wykorzystane do optymalizacji struktury połączeń (np. w celu ograniczenia energii elektrycznej). Jednak w praktyce pomysły i koncepcje przedstawione przez środowiska badawcze są rzadko wykorzystywane przez projektantów sieci, a systemy czujników, które zostały już wdrożone (i są w trakcie opracowywania w środowiskach miejskich), rzadko korzystają z dostępnych modeli badawczych opracowanych i przedstawionych w literaturze. Co więcej, powszechny dostęp do różnorodnych technologii bezprzewodowych, umożliwiających empiryczny rozwój i popularyzację nowych rozwiązań, również sprzyja umacnianiu się tego szczególnego trendu [5].

Analizowanie topologii takich sieci i diagnozowanie problemów powinno więc opierać się na danych uzyskanych z sieci rzeczywistych. Dane uzyskane z różnych typów urządzeń pozwolą zrozumieć zmiany zachodzące w typologii sieci oraz zidentyfikować problemy występujące w takich heterogenicznych sieciach. W tym artykule autorzy przedstawiają zagadnienia badawcze dotyczące wydajności transmisji sieci bezprzewodowych w środowisku miejskim, na podstawie otwartych baz danych dostarczanych przez czujniki zainstalowane w miejskiej infrastrukturze sieciowej.

Artykuł jest podzielony na 3 części. W pierwszej części przedstawiono problemy badawcze związane z miejskimi sieciami sensorowymi. Druga część przedstawia badania mające na celu wykazanie możliwości wykorzystania dostępnych zbiorów danych w badaniach efektywności przesyłu danych w sieciach miejskich. Artykuł kończy się podsumowaniem i opisem dalszych prac.

2. PROBLEMY BADAWCZE W MIEJSKICH SIECIACH SENSOROWYCH

Złożoność zagadnień związanych z transmisją danych w sieciach sensorowych stworzyła szereg problemów badawczych. Najważniejsze z nich dotyczą w

* Badania zostały sfinansowane ze środków Ministerstwa Edukacji i Nauki (projekt nr 0313/SBAD/1307).

szczególności: modelowania sieci z wykorzystaniem grafów (w sposób podobny do modelowania przewodowych i bezprzewodowych sieci telekomunikacyjnych oraz teleinformatycznych), a także wizualizacji tych sieci (w tym mobilności węzłów), społecznych trendów w mobilności (współdzielenie danych uzyskiwanych z pojazdów transportu publicznego, zwykle poprzez otwarte, dobrze udokumentowane interfejsy programowania aplikacji – API), problemów routingu (ze szczególnym uwzględnieniem nowych technik, w tym routingu oportunistycznego), agregacji danych, odciążania i detekcji mobilnej.

2.1. Routing oportunistyczny

Trasowanie jest szczególnie trudnym zadaniem w sieciach połączonych czujników. Protokoły routingu dla takich sieci różnią się od protokołów przeznaczonych dla sieci tradycyjnych. W przypadku sieci WSN należy wziąć pod uwagę wymóg oszczędzania energii, zwłaszcza przy autonomicznym zasilaniu, oraz odpowiedniego zarządzania zasobami węzła. Wybór trasy polega na wybraniu najlepszej drogi pomiędzy dwoma węzłami. Transmisja danych odbywa się poprzez wybieranie kolejnych węzłów. Tradycyjne metody routingu w sieciach bezprzewodowych typu multi-hop proaktywnie wybierają węzły po stronie nadawcy przed transmisją.

Koncepcja sieci oportunistycznych wywodzi się z badań nad sieciami tolerującymi opóźnienia (ang. *Delay Tolerant Networks*, DTN), które doprowadziły do opracowania architektury DTN. Składa się ona z niezależnych partycji sieciowych, między którymi istnieją jedynie sporadyczne możliwości komunikacji, czasem zaplanowane w czasie, a czasem – zupełnie przypadkowe. Niezależne, oddalone od siebie sieci tworzą tzw. regiony sieci DTN, a system bram jest odpowiedzialny za zapewnienie połączeń między nimi. Model ten pasuje do charakterystyki sieci miejskich – oddzielne wyspy (np. pojazdy) zawierające czujniki przesyłają zebrane dane o położeniu i inne dane telemetryczne w zaplanowanym czasie lub gdy poruszają się w pobliżu bramy (np. przystanku autobusowego) [7].

2.2. Agregacja danych

Inteligentne miasta i sieci miejskie wykorzystują dane uzyskane z różnych czujników, aby oferować lepsze usługi. Dane te są zwykle przesyłane do węzłów końcowych (np. bramy DTN). Jednak pokrycie całego obszaru miasta statycznymi czujnikami i zapewnienie routingu między nimi może być kosztowne. W literaturze rozważa się wykorzystanie systemu transportu publicznego jako mobilnej platformy dla węzłów czujników oraz przystanków autobusowych pełniących rolę ujść (ang. *sink*) [3]. Platforma taka może poprawić zasięg sieci czujników i wykorzystać komunikację oportunistyczną. Rozwiązanie to może być wykorzystywane w aplikacjach o ograniczonych opóźnieniach, dlatego zaproponowano algorytm wyboru węzłów-ujść na przystankach autobusowych w inteligentnym mieście w sposób umożliwiający minimalizowanie maksymalnego opóźnienia podczas dostarczania wiadomości. Eksperymenty pokazują, że wykorzystując tylko 16% przystanków autobusowych jako odbiorników, można uzyskać blisko 10-procentowy wzrost maksymalnego opóźnienia sieci nie tracąc znacząco na zasięgu przestrzennym.

W artykule [9] opisano koncepcję *Internetu rowerów* (*IoB - Internet of Bikes*), sieci czujników opartej na systemie rowerów miejskich. Zaproponowano wydajny protokół routingu IoB-DTN oparty na agregacji danych, który stosuje paradygmat DTN do aplikacji Internetu Rzeczy. Dane odczytywane przez rowery są przesyłane w tzw. trybie *store-and-forward* i agregowane przez zestaw równoważnych odbiorników. Protokół IoB-DTN może być postrzegany jako uproszczona wersja różnych protokołów DTN typu *n-copy*, zoptymalizowana pod kątem urządzeń IoT (m.in. usunięto niektóre funkcje routingu, które są bezużyteczne w sieciach DTN).

2.3. Modelowanie i wizualizacja sieci

Model mobilności powinien naśladować zmiany zachodzące w rzeczywistej sieci. W literaturze przedmiotu zdefiniowano szereg modeli mobilności dla sieci ad-hoc wraz z analitycznym podejściem do mobilności pojedynczego węzła i grupy. W pracy [4] stwierdzono, że w celu uwzględnienia dynamiki sieci w modelu mobilności, w procesie optymalizacji można zastosować np. metodę zstępowania gradientów (zamiast popularnego prawa ruchu Newtona znanego z dynamiki klasycznej). Syntetyczne sieci z modelami mobilności są wykorzystywane w literaturze do testowania nowych algorytmów routingu i protokołów dla sieci sensorowych [8]. Od niedawna dostępne są rzeczywiste źródła danych o lokalizacji węzłów sensorowych w środowisku miejskim, które wspierają prowadzenie badań nad strukturami takich sieci. Eliminuje to konieczność sięgania po modele syntetyczne i pozwala na wykorzystanie istniejących modeli w rzeczywistych topologiach, a także tworzenie nowych modeli.

Mobilność współdzielona to nowa, odrębna kategoria mobilności miejskiej, która obejmuje różnego rodzaju pojazdy (np. rowery, skutery, hulajnogi, samochody) przeznaczone do indywidualnego użytku. Najczęściej są one wynajmowane za pośrednictwem aplikacji mobilnych. Dedykowane portale udostępniające dane w określonym formacie są ważnym źródłem informacji o węzłach w środowisku miejskim. Po zakończeniu podróży pojazd staje się dostępny dla kolejnych użytkowników [12]. W Polsce zagregowane informacje o pojazdach transportu publicznego są udostępniane w czasie rzeczywistym przez gminy i inne podmioty [14,16]. Jest to jeden z niewielu otwartych i szeroko udokumentowanych zbiorów danych generowanych przez floty pojazdów transportu publicznego. Oprócz aktualnej lokalizacji geograficznej (współrzędne) czy numeru bocznego pojazdu, zawiera także szczegółowe dane, jak: prędkość pojazdu, typ pojazdu (np. niskopodłogowy), wyposażenie (rampa, klimatyzacja, system zapowiedzi głosowych, automat biletowy, ładowarka USB) oraz czy przewidziane jest miejsce do przewozu rowerów. Są to informacje, które mogą być przydatne także dla twórców aplikacji, a ostatecznie dla mieszkańców. Narzędzia takie umożliwiają również prowadzenie badań nad strukturą sieci i wydajnością transmisji danych w sieciach miejskich z wykorzystaniem rzeczywistych danych, a nie syntetycznie tworzonych zbiorów danych.

Dane z takich urządzeń są udostępniane w różnych formatach i strukturach w oparciu o popularne interfejsy API (np. *REpresentational State Transfer*, REST) i mogą

być wykorzystywane do tworzenia specyficznych map sieci, uwzględniających także ich dynamiczną strukturę. W tym procesie często wykorzystuje się graficzną reprezentację tych sieci. Do tego celu często wykorzystuje się format JSON (ang. *JavaScript Object Notation*), który oferuje niezależna od języka (tekstową) składnię służąca do definiowania formatów wymiany danych [2]. Wykorzystywany jest także otwarty format *GeoJSON* zaprojektowany do reprezentowania prostych obiektów geograficznych wraz z ich atrybutami nieprzestrzennymi [1]. Stosuje się również mechanizm *Protocol Buffers,* który jest neutralnym językowo i platformowo, rozszerzalnym mechanizmem Google do serializacji danych strukturalnych i stanowi jednocześnie sposób na zminimalizowanie rozmiaru wysyłanej wiadomości (pliku), przy jednoczesnym zachowaniu jej pełnej zawartości [15]. W tym rozwiązaniu oryginalna wiadomość jest kodowana i dekodowana za pomocą ustalonego (dla danego protokołu) klucza, co pozwala ograniczyć jej zawartość tylko do wartości konkretnych parametrów bez podawania pełnej struktury pliku i nazw parametrów, jak ma to miejsce w formacie XML (*Extensible Markup Language*), czy JSON. Powoduje to to znaczne zmniejszenie rozmiaru przesyłanego pliku. Z kolei *GTFS Realtime* to specyfikacja, która pozwala agencjom transportu publicznego dostarczać programistom aplikacji aktualizacje dotyczące ich floty w czasie rzeczywistym [11]. Jest to rozszerzenie GTFS (*General Transit Feed Specification*) – otwartego formatu danych dla rozkładów jazdy transportu publicznego i związanych z nimi informacji geograficznych.

3. PRZYKŁAD MODELOWANIA SIECI

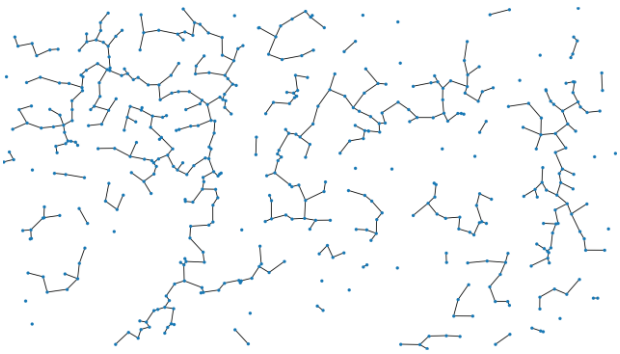
Na podstawie obserwacji omówionych w poprzednich rozdziałach przeprowadzono studium wykonalności, aby odpowiedzieć na pytanie, czy publicznie dostępne zbiory danych mogą być efektywnie wykorzystane w badaniach efektywności transmisji danych w miejskich sieciach czujników. W tym celu wybrano narzędzia programowe wraz z uznanymi bibliotekami do przetwarzania danych zgodnie ze specyfikacją żądań dostawców zbiorów danych.

W pierwszym etapie prowadzonych prac badawczych testowano interfejsy API udostępniane przez dostawców danych o położeniu geograficznym węzłów. W badaniach wykorzystano otwarte źródła danych udostępnione przez *Zarząd Transportu Miejskiego w Poznaniu* (informacje o lokalizacji pojazdów [17]), infrastrukturę *Smart City Poznań* (lokalizacja przystanków komunikacji miejskiej, parkometrów i biletomatów) oraz infrastrukturę projektu *Airly* [10]. Opracowano niezbędne oprogramowanie do zautomatyzowania procesu pozyskiwania, czyszczenia i wstępnej wizualizacji danych.

W ramach pojedynczego, statycznego *zdjęcia* sieci, wykonanego 27 listopada 2019 r. o godz. 15:00, zidentyfikowano 499 węzłów (465 stacjonarnych i 34 mobilne). W trakcie badania przyjęto następujące założenia:

- połączenie komunikacyjne między dwoma węzłami w sieci istnieje, gdy odległość między nimi jest nie większa niż 100 metrów.

Rys. 1. Lokalizacja założonych węzłów sensorycznych i łączy radiowych w centrum Poznania w dniu 27.11.2019



Rys. 2. Minimalny las rozpinający założonych węzłów sensorycznych i łączy radiowych w centrum Poznania w dniu 27.11.2019

- obszar miasta został ograniczony według następujących współrzędnych: maksymalna szerokość geograficzna: 52.415, minimalna szerokość geograficzna: 52.400, minimalna długość geograficzna: 16.898, maksymalna długość geograficzna: 16.942,
- zasięg radiowy każdego węzła ma charakter dookólny,

W ten sposób otrzymano graf (Rys. 1) nałożony na mapę miasta o następujących parametrach [13]:

- liczba elementów (połączonych podgrafów): 57,
- liczba krawędzi: 1005,
- całkowity koszt: 67232 metrów,
- średni stopień węzła: 4,03.

Stopień węzła wynika z liczby innych węzłów znajdujących się w zasięgu radiowym tego węzła. Sieć jest gęstsza w obszarach o większym zagęszczeniu elementów infrastruktury (biletomaty, parkomaty, pojazdy itp.). Struktura ta jest w dużej mierze związana z przebiegiem tras komunikacyjnych i rozmieszczeniem infrastruktury pomocniczej. Przykładowa modelowana sieć ma zmienną liczbę spójnych elementów o zmiennej w czasie strukturze, ale jest przedstawiona w danym momencie czasowym. Wizualizacja przedstawia więc chwilowy model sieci przy przyjętych założeniach.

W kolejnym kroku wyznaczono minimalny las rozpinający (tj. zbiór minimalnych drzew rozpinających wyznaczonych dla każdego połączonego podgrafu)

powyższego grafu za pomocą algorytmu Kruskala, gdy wagą krawędzi jest odległość geograficzna między węzłami wyrażona w metrach (Rys. 2). Takie podejście może być celowe do wyznaczenia optymalnego rozmieszczenia bramek (ujść) w obrębie modelowanych wysp DTN. W rezultacie otrzymano następujące parametry sieci:

- liczba elementów (połączonych podgrafów): 57,
- liczba krawędzi: 442,
- całkowity koszt: 23834 metrów,
- średni stopień węzła: 1,77.

Obszary zastosowań tego podejścia obejmują: rozgłaszanie (broadcast); transmisję grupową (multicast) informacji sterujących lub aktualizacji oprogramowania oraz budowę optymalnego drzewa łączącego wszystkie spójne drzewa z centralnym węzłem klastra (implementacja topologii hierarchicznej łączącej lokalne poddrzewa). Stacjonarne węzły pełnią rolę agregatorów transmisji, lokalnych buforów oraz nadajników do wyższych warstw hierarchii sieciowej poprzez wykorzystanie dedykowanych połączeń (w innych technologiach lub o większej mocy). Ze względu na to, że zasoby (obliczeniowe, pamięciowe, zasilające i połączeniowe) mogą znacznie przekraczać zasoby prostych węzłów, to będą one również pełniły rolę tzw. lokalnego magazynu i przekaźnika w modelu *store-and-forward* z wykorzystaniem protokołów routingu operujących na kompletnych porcjach danych (tzw. *bundle*).

Dodatkowo, wykorzystując mobilny węzeł przekaźnikowy w modelu *store, carry, and forward*, można łączyć komponenty w bardziej spójną strukturę sieci, rozciągniętą w czasie. Połączenia te mogą być tworzone oportunistycznie, w zależności od ruchów niezależnego agenta mobilnego. Ruchy te mogą wynikać z zaplanowanej trasy mobilnego przekaźnika (agenta).

4. WNIOSKI I DALSZE PRACE

Wykorzystanie połączonych węzłów sensorycznych w środowiskach miejskich oraz powszechna dostępność dostarczanych przez nie danych stwarza ogromne możliwości i otwiera nowe obszary badań. Jak udowodniono w niniejszym artykule, dzięki publicznie dostępnym interfejsom (API), możliwe stało się zbadanie koncepcji routingu oportunistycznego i DTN z wykorzystaniem danych o rzeczywistym geograficznym położeniu węzłów, a także wyznaczenie nowych trendów w badaniach, np. efektywnego projektowania algorytmów rozmieszczania bramek (ujść) lub integracji DTN z routingiem w sieciach MANET.

Wstępne wyniki przedstawione w niniejszym artykule sugerują, że ze względu na złożoną dynamikę sieci w sieciach miejskich nie można zbudować stałej rozproszonej sieci bezprzewodowej transmisji danych. Dlatego też prace badawcze powinny dotyczyć opracowania wydajnych algorytmów mających na celu zaprojektowanie statycznego rozmieszczenia lokalnych bramek dostępowych lub zaplanowanie optymalnych tras dla urządzeń mobilnych w celu zbierania informacji z wysp DTN lub rozpowszechniania wiadomości w sytuacjach kryzysowych i klęsk żywiołowych. W kolejnych pracach autorzy zamierzają wykorzystać dane uzyskane z miejskich sieci sensorowych do badań nad oportunistycznym routingiem oraz nad efektywnym projektowaniem algorytmów rozmieszczania bram DTN.

LITERATURA

[1] Butler H., Daly M., Doyle A., Gillies S., Schaub T., Hagen S. 2016. "The GeoJSON Format". *RFC 7946*.

[2] Crockford D. 2006. "The application/json media type for JavaScript Object Notation (JSON)". *IETF, RFC 4627*, 7.

[3] Cruz P., Couto R.S., Costa L.H.M. 2019. "An algorithm for sink positioning in bus-assisted smart city sensing," *Future Generation Computer Systems*. Vol. 93, pp. 761–769.

[4] Ma K., Zhang Y., Trappe W. 2008. "Managing the mobility of a mobile sensor network using network dynamics". *IEEE Transactions on Parallel and Distributed Systems*, Vol. 19, no. 1, pp. 106–120.

[5] Musznicki B. 2019. "Empirical approach in topology control of sensor networks for urban environment". *Journal of Telecommunications and Information Technology*. No. 1, pp. 47–57.

[6] Musznicki B., Zwierzykowski P. 2012. "Survey of Simulators for Wireless Sensor Networks". *International Journal of Grid and Distributed Computing*. Vol. 5, no. 3, pp. 23–50.

[7] Pelusi L., Passarella A., Conti M. 2006. "Opportunistic networking: Data forwarding in disconnected mobile ad hoc networks". I. Vol. 44, pp. 134 – 141.

[8] Santos B.P., Goussevskaia O., Vieira L.F., Vieira M.A., Loureiro A.A. 2018. "Mobile matrix: Routing under mobility in IoT, IoMT, and Social IoT". *Ad Hoc Networks*. Vol. 78, pp. 84–98.

[9] Zguira Y., Rivano H., Meddeb A. 2018. "Internet of bikes: A DTN protocol with data aggregation for urban data collection". *Sensors*, Vol. 18, no. 9.

[10] "Airly API". Dostęp: styczeń 2022. https://developer.airly.org/pl/docs, 2022.

[11] "GTFS Realtime Overview". Dostęp: styczeń 2022. *https://developers.google.com/transit/gtfs-realtime*.

[12] "On the threshold of a breakthrough. Shared mobility in Poland". Dostęp: grudzień 2022. *https://smartride.pl/wp-content/uploads/2020/02/RaportSharedMobility2019PLmaly.pdf*.

[13] "OpenStreetMap". Dostęp: luty 2022. *https://wiki.openstreetmap.org/wiki/AboutOpenStreetMap*.

[14] "Otwarty Gdańsk". Dostęp: luty 2022. *https://ckan.multimediagdansk.pl/*.

[15] "Protocol Buffers". Dostęp: styczeń 2022. *https://developers.google.com/protocol-buffers/*.

[16] "Smart City Poznan". Dostęp: luty 2022. *https://www.poznan.pl/mim/smartcity/api-dane-przestrzenne, p,25877,38305.html*.

[17] "ZTM Poznan, for developers". Dostęp: marzec 2022. *https://www.ztm.poznan.pl/pl/dla-deweloperow/gtfsRtFiles*.

ANALIZA BEZPIECZEŃSTWA PROTOKOŁÓW KOMUNIKACYJNYCH URZĄDZEŃ INTELIGENTNEGO DOMU OPARTYCH NA MIKROKONTROLERZE ESP

SECURITY EVALUATION OF COMMUNICATION PROTOCOLS FOR SMART HOME APPLIANCES BASED ON ESP MICROCHIPS

Kacper Murat[1]; Krzysztof Zdulski[1]; Dominik Topyła[1]; Michał Marzęcki[1]; Jędrzej Bieniasz[1]

[1]Instytut Telekomunikacji, Politechnika Warszawska, Warszawa

DOI: 10.15199/59.2022.4.43

Streszczenie: **W artykule przedstawiono zagadnienie badań bezpieczeństwa protokołów komunikacyjnych urządzeń Internetu Rzeczy. Platformą badań były niskobudżetowe urządzenia IoT przeznaczone do realizacji instalacji inteligentnego domu oparte na popularnych mikrokontrolerach ESP. Zbadano komunikację przewodową (UART) oraz bezprzewodową (Wi-Fi 802.11). Sformułowano potencjalne scenariusze zagrożeń dla bezpieczeństwa cyberprzestrzeni w obszarze urządzeń IoT, a także rekomendacje dotyczące rozwiązań tych problemów.**

Abstract: **This article presents research on the security of communication protocols for Internet of Things devices. The research platform was low-cost IoT devices designed to implement smart home installations based on popular ESP microcontrollers. Wired (UART) and wireless (Wi-Fi 802.11) communication tests were carried out. Potential scenarios of threats to cyberspace security in the area of IoT devices were formulated, as well as recommendations to mitigate these problems.**

Słowa kluczowe: **analiza protokołów, cyberbezpieczeństwo, Internet Rzeczy, ESP**

Keywords: **protocol analysis, cybersecurity, Internet of Things, ESP**

1. WSTĘP

Koncepcja Internetu Rzeczy (ang. *Internet of Things*, IoT) w ostatniej dekadzie zanotowała znaczący wzrost zainteresowania wśród klientów indywidualnych [5]. Wraz z rosnącym popytem na niskobudżetowe rozwiązania powstało wiele programowalnych platform i komponentów oferujących niezbędne funkcje w zastosowaniu jako węzły sieci IoT [9], [10]. Zapotrzebowanie ilościowe na rozwiązania IoT w powiązaniu z oczekiwaniem niskiej ceny urządzeń skutkuje tym, że następuje unifikacja produkcji platform sprzętowych IoT, a produkcja ich zostaje skupiona przez kilku większych dostawców. Platformy końcowe różnią się między sobą jedynie wgranym oprogramowaniem końcowym. Takie podejście wiąże się z istnieniem dużego ryzyka z kategorii cyberbezpieczeństwa łańcucha dostaw (ang. *supply chain*). W przypadku odkrycia podatności dla danego rozwiązania IoT od pozornie niszowego dostawcy może się okazać, że dotyczy ona dużo większej liczby urządzeń. Tym samym, wpływ (ang. *impact*) takiego odkrycia będzie wyznaczany jako co najmniej *średni* [11] w analizie ryzyka. Ponadto, takie podejście do produkcji urządzeń IoT może oznaczać łatwe powielanie podatności w rozwiązaniach końcowych.

Jedną z najpopularniejszych, a zarazem tanich platform wykorzystywanych w urządzeniach typu *Smart Home* jest mikrokontroler ESP w wariancie 8266EX. Jest to prosty układ typu SoC (ang. *System on a Chip*) wyposażony w zintegrowany moduł Wi-Fi. Jego następcą jest platforma ESP32 dodatkowo oferująca moduł Bluetooth. Jest ona jednak droższa, co determinuje mniejszy udział jako komponent na rynku tanich urządzeń IoT. Producent mikrokontrolera, firma Espressif, umożliwił programowanie jej w prosty sposób wykorzystując interfejs UART, który jak się okazuje, nie został w żaden sposób zabezpieczony. W ten sposób, ekstrakcja kodu źródłowego urządzenia jest bardzo łatwa, a jedyne zabezpieczenie urządzenia istnieją na poziomie oprogramowania wbudowanego tj. firmware. Takie podejście nie gwarantuje w żaden sposób bezpieczeństwa w zakresie integralności styku warstwy fizycznej i logicznej urządzenia. Jednym z rozpatrywanych scenariuszy zagrożeń może być np. łatwe umieszczenie szkodliwego kodu wykonywalnego powodującego, że atakujący uzyskuje dostęp do hasła sieci Wi-Fi, do której podłączają się te urządzenia.

W niniejszym artykule zaprezentowano wstępne badania bezpieczeństwa niskobudżetowych urządzeń IoT opartych o mikrokontrolery ESP na przykładzie gniazdka IoT Gosund SP111. Badaniami objęto komunikację przewodową interfejsem UART, komunikację bezprzewodową Wi-Fi 802.11 oraz analizę oprogramowania układowego pozyskanego z urządzenia. Przedyskutowano możliwe scenariusze cyberzagrożeń. Rozdział 2 przedstawia procedurę badawczą i rezultaty badań nad jednym z urządzeń IoT opartym na mikrokontrolerze ESP. W ramach tej procedury zbadano kanały przewodowe i bezprzewodowe komunikacji urządzenia oraz przeanalizowano pobrany z urządzenia firmware w postaci pliku binarnego. Rozdział 3 podsumowuje znalezione problemy bezpieczeństwa i rozważa możliwe scenariusze cyberzagrożeń dla urządzeń inteligentnego domu. Rozdział 4 podsumowuje artykuł i przedstawia dalsze możliwe obszary rozwoju zagadnienia.

2. BADANIA BEZPIECZEŃSTWA URZĄDZEŃ INTELIGENTNEGO DOMU

2.1. Procedura badawcza

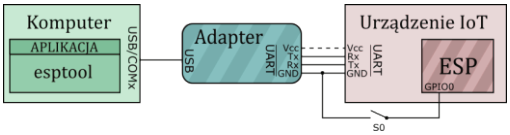
Procedura badawcza została podzielona na kilka następujących po sobie etapów:

1) Przygotowanie środowiska.
2) Eksport firmware'u z urządzenia i analiza pliku wyjściowego.
3) Import firmware'u do urządzenia
4) Integracja urządzenia z siecią inteligentnego domu wraz z automatyczną aktualizacją oprogramowania wbudowanego do najnowszej wersji udostępnianej przez producenta. W ramach tego kroku odbywa się śledzenie komunikacji na interfejsach UART i Ethernet.
5) Analiza zmian w firmware po aktualizacji oprogramowania.
6) Dekompilacja kodu.

Szczegóły realizacji etapu 1) przedstawia Rozdział 2.2, etapy 2) i 3) opisuje Rozdział 2.3. Etap 4) opisano w Rozdziale 2.4. Etap 5) nie wskazał żadnych istotnych zmian. Etap 6) nie został uwzględniony w tej części badań prezentowanych w niniejszym artykule.

2.2. Środowisko i przygotowanie do badań

Stanowisko testowe zostało przedstawione na Rys. 1.



Rys. 1 Środowisko testowe do badań urządzeń IoT

Istotą prowadzonych eksperymentów było podłączenie interfejsu UART urządzenia IoT z portem uniwersalnej magistrali szeregowej (USB) komputera. W tym celu wykorzystany został adapter USB-UART PL2303. Umożliwia on również zasilanie urządzeń napięciem 3,3V lub 5V zależnie od ustawienia fizycznego przełącznika (zworki). Obiektem badań było inteligentne gniazdko elektryczne Gosund SP111. Jednostką obliczeniową tego urządzenia jest mikrokontroler ESP8266EX [6]. Takie gniazdko może być zintegrowane z siecią inteligentnego domu za pomocą łącza Wi-Fi. Inicjalizacja (sparowanie) możliwa jest z wykorzystaniem aplikacji na urządzenia mobilne *Tuya Smart*. Na komputerze zainstalowana została aplikacja *esptool*. Jest to oprogramowanie typu open-source dedykowane komunikacji z mikrokontrolerami ESP. Za jego pomocą możliwy jest m.in. import i eksport firmware'u, a także wiele innych funkcji [3].

2.3. Ekstrakcja i import firmware w badaniach wybranego urządzenia

Oprogramowanie *esptool* ułatwia pracę badawczą m.in. poprzez automatycznie wykrywanie podłączonych urządzeń. Jeżeli ten krok zostanie zrealizowany poprawnie, to ekstrakcja firmware odbywa się za pomocą prostej komendy. W przypadku problemów istnieje możliwość wprowadzenia ręcznej konfiguracji, aby wykryć urządzenie do badań. Jest to niezbędne przy podłączaniu wielu urządzeń równolegle do badań, gdy korzysta się z różnych portów komputera. Gniazdko będące przedmiotem badania zostało podłączone do portu COM16 i skonfigurowano przepływność symbolową 115.2 kbps, aby następnie dokonać ekstrakcji firmware'u do pliku binarnego. W tym celu należało podać adresy komórek w rejestrze w którym znajduje się kod. Adres początkowy jest stały i wynosi 0x000000, jednak adres końcowy zależy od wielkości firmware'u. *Esptool* pozwolił ustalić ten rozmiar dla testowanego urządzenia o wielkości 1MB, co oznacza, że adres ostatniego miejsca w pamięci wynosi 0x100000. Po wykonaniu ekstrakcji otrzymano plik binarny firmware'u w skompilowanej wersji. Zwrócono uwagę, że ciągi znaków z tablicy ASCII nie są w żaden sposób konwertowane, więc możliwe jest ich łatwy przegląd. Wyróżniający się obszar pamięci o adresach od `000fb000` do `000fb17f` zaprezentowano na Rys. 2.

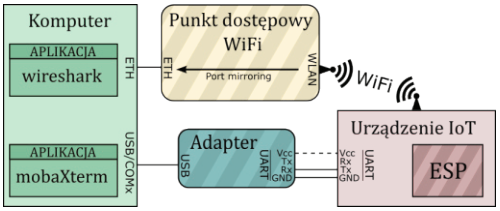
```
000faff0: ffff ffff ffff ffff ffff ffff ffff ffff  ................
000fb000: 7b22 6d61 6322 3a22 3765 3022 2c22 7072  {"mac":"7e0","pr
000fb010: 6f64 5f69 6478 223a 2232 3035 3338 3633  od_idx":"2053863
000fb020: 3022 2c22 6175 7a5f 6b65 7922 3a22 4c68  0","auz_key":"Lh
000fb030: 3936 4877 424d 6a76 5759 7071 4b51 6577  96HwBMjvWYpqKQew
000fb040: 5558 4330 4872 6831 6631 4c56 4a54 222c  UXC0Hrh1f1LVJT",
000fb050: 2270 726f 645f 7465 7374 223a 6661 6c73  "prod_test":fals
000fb060: 657d 0000 0000 0000 0000 0000 0000 0000  e}..............
000fb070: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb080: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb090: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb0a0: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb0b0: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb0c0: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb0d0: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb0e0: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb0f0: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb100: 7b22 6170 5f73 7369 6422 3a22 536d 6172  {"ap_ssid":"Smar
000fb110: 744c 6966 6522 2c22 6170 5f70 7764 223a  tLife","ap_pwd":
000fb120: 6e75 6c6c 7d00 0000 0000 0000 0000 0000  null}..........
000fb130: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb140: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb150: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb160: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb170: 0000 0000 0000 0000 0000 0000 0000 0000  ................
000fb180: ffff ffff ffff ffff ffff ffff ffff ffff  ................
```

Rys. 2 Zrzut z obszaru pamięci pod adresami od 000fb000 do 000fb17f

Można zauważyć, że występują w nim struktury tekstowe w formacie JSON. Pierwsza z takich struktur zawiera informacje o 12 ostatnich bitach adresu MAC urządzenia (3 znaki w zapisie heksadecymalnym), obiekt identyfikatora wraz z kluczem autoryzacyjnym oraz zmienną typu *bool* o nazwie *prod_test*. Druga struktura zawiera domyślne SSID i hasło do nieznanego punktu dostępowego Wi-Fi. Wstępnym wnioskiem na tym etapie jest to, że odkryte struktury mogą służyć do awaryjnego dostępu do urządzenia. Poprawność adresu MAC potwierdziły również komunikaty z interfejsu UART. Import firmware'u odbywa się w sposób analogiczny jak ekstrakcja.

2.4. Integracja badanego urządzenia w scenariuszu inteligentnego domu

Wybrane gniazdko IoT do badań współpracuje z chmurą Tuya, a samo parowanie opiera się na wpisaniu SSID i hasła do sieci Wi-Fi w dedykowanej aplikacji i przesłaniu ich do urządzenia IoT. Następnie urządzenie uwierzytelnia się pozyskanymi danymi. Ze względu na to, że ta komunikacja odbywa się bezprzewodowo za pomocą Wi-Fi, do dalszych badań niezbędnym było rozbudowanie stanowiska testowego przedstawionego na Rys. 1. Rozszerzenie do scenariusza w konfiguracji inteligentnego domu prezentuje Rys. 3.



Rys. 3 Scenariusz inteligentnego domu z badanym urządzeniem IoT

Zainstalowano bezprzewodowy punkt dostępowy Mikrotik z funkcją monitorowania ruchu na interfejsach, tak aby kopiować ruch sieciowy do analizy w czasie rzeczywistym. Komunikację przewodową interfejsu szeregowego obserwowano przy wykorzystaniu stanowiska komputerowego z adapterem USB-UART. Zapis komunikacji przewodowej prezentuje Rys. 4.

```
    [notice]mqtt_client.c:590 gw wifi stat is:1
    …
1 → [notice]gw_intf.c:420 Authorization success
    mode : sta(70:03:9f:5d:07:e0)
    add if0
    [notice]device.c:729 ID:(null)
    [err]app_dltj.c:193 app_dltj_init
    scandone
    state: 0 -> 2 (b0)
    state: 2 -> 3 (0)
    [notice]mqtt_client.c:590 gw wifi stat is:5
    state: 3 -> 5 (10)
    add 0
    aid 2
    pm open phy_2,type:2 0 0
    cnt
2 → connected with LaVA_Monitored_2.4, channel 1
    dhcp client start...
    ip:192.168.88.248,mask:255.255.255.0,gw:192.168.88.1
3 → [notice]mqtt_client.c:590 gw wifi stat is:5
    [notice]mqtt_client.c:590 gw wifi stat is:6
    [notice]mqtt_client.c:633 DNS START 03-09 17:55:54
    [notice]mqtt_client.c:660 DNS END 03-09 17:55:54
    [notice]mqtt_client.c:666 MQTT CONN START 03-09 17:55:54
    [notice]mqtt_client.c:689 MQTT CONN END 03-09 17:55:55
    [notice]mqtt_client.c:788 mqtt connect success
    [err]smart_wf_frame.c:2243 unsupport frame type:64
4 → [notice]smart_wf_frame.c:3823 firmware self detect upgrade start...
5 → [notice]smart_wf_frame.c:3833 get fw ug info op_ret:2
```

Rys. 4 Zapis z interfejsu UART dla badanego gniazdka

Jak można zauważyć na Rys. 4, firmware wysyła na port UART wszystkie informacje serwisowe zgodnie z określonym katalogiem oznaczeń. Te komunikaty pozwalają wyciągnąć różne wnioski na temat struktury kodu i operacji przeprowadzanych przez urządzenie, m.in.: widoczność wielu nazw plików kodu źródłowego i miejsca w których wykonywane są akcje. W linii (1) z Rys. 4 widoczne jest powodzenie autoryzacji urządzenia mobilnego w sieci gniazdka, a w linii (2) z Rys. 4 powodzenie autoryzacji gniazdka do punktu dostępowego Wi-Fi. Między tymi liniami widoczne są zmiany stanów urządzenia, które po otrzymaniu SSID i hasła rozłącza sesję ze smartphone'm przechodzi do roli serwera Wi-Fi na klienta. Następnie szuka punktu dostępowego i łączy się z nim wykorzystując wcześniej pozyskane dane. Równolegle na interfejs UART został również przesłany adres MAC gniazdka. Kolejne obserwowane akcje to włączenie usługi klienta DHCP i pozyskanie z routera adresu IP. Począwszy od linii (3) na Rys. 4 widoczny jest proces nawiązywania połączenia z chmurą poprzez protokół MQTT. Następnie linia (4) informuje o rozpoczęciu sprawdzania dostępności aktualizacji firmware'u, a linia (5) o jej rozpoczęciu.

Warto odnotować, że wszystkie logi wyświetlały się w krótkim okresie czasu (kilku sekund). Po podłączeniu się urządzenia do Internetu jest ono w stanie pobrać kod ze strony producenta, nie dając nawet możliwości potwierdzenia tego procesu. Biorąc pod uwagę, że gniazdko ma pełen dostęp do sieci prywatnej, należy zwrócić uwagę na zagrożenie związane z nieuczciwymi praktykami dostawców rozwiązań chmurowych. W [12] zwrócono uwagę na możliwość wykorzystania takiego kanału aktualizacji do zaimportowania złośliwego oprogramowania do wszystkich podłączanych urządzeń tworząc w ten sposób złośliwy *botnet.* Z punktu widzenia monitorowania i analizy danych dla cyberbezpieczeństwa taki firmware o rozmiarze 1MB referencyjnie do skali ruchu przetwarzanego przez urządzenia sieciowe jest znikomo mały i przesyłany w ruchu zaszyfrowanym. Oznacza to, nie można polegać na rozwiązaniu wykrywania zagrożeń w sieci, więc tym bardziej takie badania jak prezentowane w niniejszym artykule oraz wnioski dla cyberbezpieczeństwa są szczególnie istotnie. Na uwagę zasługuje również fakt, że interfejs UART będący kanałem do debugowania urządzenia nie jest zablokowany w użyciu produkcyjnym. Mimo, że w wiadomościach nie występują hasła, to po skorelowaniu z komunikacją bezprzewodową mogą stanowić dodatkowe, nieszyfrowane źródło referencyjne do ich wyciągania. Inną obserwacją jest zauważenie, że interfejs przekazuje informacje o błędach podczas interakcji użytkownika z urządzeniem. Równolegle przeanalizowano zrzut ruchu sieciowego z komunikacji Wi-Fi 802.11 z wykorzystaniem programu Wireshark, co przedstawia Rys. 5.

No.	Source	Destination	Protocol	Lengt	Info
10036	192.168.88.248	35.156.42.116	TCP	105	14381 → 8886 [SYN] Seq=0 Win=4380 Len=0 MSS=1460
10051	35.156.42.116	192.168.88.248	TCP	105	8886 → 14381 [SYN, ACK] Seq=0 Ack=1 Win=2048 Len=0 MSS=1460
10052	192.168.88.248	35.156.42.116	TCP	101	14381 → 8886 [ACK] Seq=1 Ack=1 Win=4380 Len=0
10057	192.168.88.248	35.156.42.116	TLSv1.2	168	Client Hello
10062	35.156.42.116	192.168.88.248	TCP	101	8886 → 14381 [ACK] Seq=1 Ack=68 Win=2048 Len=0
10063	35.156.42.116	192.168.88.248	TLSv1.2	249	Server Hello, Server Key Exchange, Server Hello Done
10071	192.168.88.248	35.156.42.116	TLSv1.2	162	Client Key Exchange
10092	35.156.42.116	192.168.88.248	TCP	101	8886 → 14381 [ACK] Seq=149 Ack=129 Win=2048 Len=0
10095	192.168.88.248	35.156.42.116	TLSv1.2	107	Change Cipher Spec
10102	35.156.42.116	192.168.88.248	TCP	101	8886 → 14381 [ACK] Seq=149 Ack=135 Win=2048 Len=0
10103	192.168.88.248	35.156.42.116	TLSv1.2	186	Encrypted Handshake Message
10112	35.156.42.116	192.168.88.248	TCP	101	8886 → 14381 [ACK] Seq=149 Ack=220 Win=2048 Len=0
10113	35.156.42.116	192.168.88.248	TLSv1.2	192	Change Cipher Spec, Encrypted Handshake Message
10121	192.168.88.248	35.156.42.116	TLSv1.2	234	Application Data
10131	35.156.42.116	192.168.88.248	TLSv1.2	170	Application Data
10134	192.168.88.248	35.156.42.116	TCP	101	14381 → 8886 [ACK] Seq=353 Ack=309 Win=4072 Len=0
10143	192.168.88.248	35.156.42.116	TLSv1.2	202	Application Data
10151	35.156.42.116	192.168.88.248	TLSv1.2	170	Application Data
10260	192.168.88.248	35.156.42.116	TCP	101	14381 → 8886 [ACK] Seq=454 Ack=378 Win=4003 Len=0

Rys. 5 Komunikacja TLS 1.2 gniazdka z chmurą

Na podstawie tej analizy można potwierdzić, że komunikacja z chmurą producenta jest zabezpieczona za pomocą sesji protokołu TLS (ang. *Transport Layer Security*). Stosowany jest protokół w wersji 1.2 [2], dla której znane są podatności takie jak np. *Poodle* [7] i *Raccon* [8].

3. ANALIZA CYBERBEZPIECZEŃSTWA URZĄDZEŃ INTELIGENTNEGO DOMU

Odkryte problemy bezpieczeństwa, które wpływają na rozszerzanie powierzchni do realizacji cyberataków (ang. *cyber attack surface*) można sklasyfikować jako:

- ***problemy bezpieczeństwa w ramach interfejsów komunikacyjnych***. Interfejs UART badanego urządzenia umożliwia uzyskanie informacji o akcjach wykonywanych na urządzeniu. Komunikacja bezprzewodowa jest szyfrowana, ale odbywa się to jednak przy pomocy protokołu TLS w wersji 1.2
- ***problemy bezpieczeństwa z oprogramowaniem wbudowanym***. Sam kod wykonywalny nie jest łatwo dostępny, więc jakiekolwiek monitorowanie jego bezpieczeństwa jest bardzo utrudnione

lub niemożliwe. W połączeniu z możliwością wgrania swojego oprogramowania ryzyko istnienia dodatkowego, złośliwego kodu sprawia, że jedynym skutecznym zabezpieczeniem jest wydzielona sieć dla takich urządzeń bez dostępu do Internetu, ale taki dostęp jest jednak wymagany w procesie inicjalizacji w sieci inteligentnego domu.

Znalezione i wskazane luki mogą również stanowić możliwość realizacji różnych scenariuszy cyberzagrożeń, a w ramach niniejszego artykułu rozważono teoretycznie dwa scenariusze referencyjne.

Pierwszy analizowany scenariusz dotyczy wykonania podmiany oprogramowania w gniazdkach tak, aby udostępniały hasło i SSID sieci Wi-Fi na niezabezpieczony interfejs UART. Dalej, te dane mogłyby być przekazane do modułu komunikacji LoRa, a to już oznacza możliwość przekazywania takich danych na zewnątrz. Wysyłkowy sposób dystrybucji może zapewnić znajomość przybliżonego adresu umiejscowienia punktu dostępowego Wi-Fi. Moduł LoRa gwarantuje daleki zasięg odbioru wiadomości, więc wystarczy kilka stacji bazowych LoRa w obsługiwanych miastach. Atak umożliwia namierzenie lokalizacji i dostęp do prywatnej sieci ofiary. Dalszym krokiem może być podsłuchiwanie ruchu lokalnego, podszycie, zagłuszanie, czy wysycanie łącz. Te zagrożenia wskazywane są w literaturze, jako jedne z ważniejszych w przypadku rozwiązań IoT [1].

Drugi scenariusz zagrożenia opiera się na wykorzystaniu możliwości ekosystemów oprogramowania open source i jednoczesnym założeniu, że wielu Użytkowników wykorzystuje takie oprogramowanie bez żadnej weryfikacji. W ramach takiego zagrożenia można w łatwy sposób przygotować zmodyfikowany firmware o dowolnie złośliwych funkcjach. Na ten problem wskazywano w Rozdziale 3.4. Uogólniając, widoczny jest wyraźny problem ze świadomością cyberbezpieczeństwa producentów rozwiązań IoT i ich użytkowników skutkujący brakiem stosowania właściwych mechanizmów bezpieczeństwa. Te zagadnienia wpisują się w klasyfikację zagrożeń opisywaną m.in. w [4], łącząc różne wymiary bezpieczeństwa łańcucha dostaw.

4. PODSUMOWANIE

Badania bezpieczeństwa urządzeń inteligentnego domu stają się co raz bardziej istotne, aby wykrywać i rozwiązywać problemy bezpieczeństwa zanim dojdzie zrealizowania się cyberzagrożeń. Niniejszy artykuł przedstawił jak można metodycznie prowadzić takie badania. W ramach zaprezentowanych wyników dokonano eksportu firmware z urządzenia IoT opartego o mikrokontroler ESP do pliku binarnego i ponownego jego importu. Została udowodniona możliwość podsłuchania informacji serwisowych przekazywanych poprzez dostępny na płytce interfejs UART. Przeanalizowano także komunikację urządzenia po interfejsie Wi-Fi. W Rozdziale 3 podsumowano odkryte podstawowe problemy bezpieczeństwa z badanym urządzeniem IoT. Rekomendowane jest, aby takie urządzenia trafiające do sprzedaży nie posiadały możliwości wykorzystania interfejsu UART lub co najmniej zapewniono brak możliwości nadpisania oprogramowania wbudowanego. W kolejnych badaniach planowane jest rozszerzenie bazy urządzeń inteligentnego domu, które będą poddawane procesom badawczym. Zakłada się także wprowadzenie narzędzi pomocniczych, które pozwolą na automatyzację części prac. Ważnym krokiem uzupełniającym badania w przyszłości będzie zrealizowanie Etapu 6) z procesu przedstawionego w Rozdziale 3.1, tj. możliwość dekompilacji firmware'u do kodu źródłowego.

Badania prowadzono w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu CyberSecIdent IV, umowa nr CYBERSECIDENT/488240/IV/NCBR/2021

LITERATURA

[1] Butun, I., Osterberg, P., & Song, H. 2020. "Security of the Internet of Things: Vulnerabilities, Attacks, and Countermeasures". In IEEE Communications Surveys and Tutorials (Vol. 22, Issue 1, pp. 616–644).

[2] Dierks, T., Rescorla, E. 2008. "The Transport Layer Security (TLS) Protocol Version 1.2". RFC Editor. URL: http://www.rfc-editor.org/rfc/rfc5246.txt [Dostęp: 07.04.2022]

[3] esptool. URL: https://github.com/espressif/esptool [Dostęp 07.04.2022].

[4] European Union Agency for Cybersecurity (ENISA). „Guidelines for Securing the Internet of Things – Secure Supply Chain for IoT". 2020. URL: https://www.enisa.europa.eu/publications/guidelines-for-securing-the-internet-of-things [Dostęp: 07.04.2022]

[5] Hanes, D., Salgueiro, G., Grossetete, P., Barton, R., Henry, J. 2017. "IoT Fundamentals: Networking Technologies, Protocols and Use Cases for the Internet of Things". no. 3491. Cisco Press.

[6] Microcontrollerslab, "ESP8266 Pinout Reference." URL: https://microcontrollerslab.com/esp8266-pinout-reference-gpio-pins/ [Dostęp 07.04.2022].

[7] NVD-CVE-2014-3566. URL: https://nvd.nist.gov/vuln/detail/cve-2014-3566 [Dostęp: 07.04.2022].

[8] NVD-CVE-2020-1968. URL: https://nvd.nist.gov/vuln/detail/CVE-2020-1968 [Dostęp: 07.06.2022].

[9] Ojo, M. O., Giordano, S., Procissi, G., Seitanidis, I. N. 2018. "A Review of Low-End, Middle-End, High-End Iot Devices," IEEE Access, vol. 6.

[10] Raza, A., Ikram, A. A., Amin, A. and Ikram, A. J. 2016. "A review of low cost and power efficient development boards for IoT applications," in 2016 Future Technologies Conference (FTC), pp. 786–790.

[11] Ross, R. 2012. "Guide for Conducting Risk Assessments", National Institute of Standards and Technology.

[12] Zhang, X., Upton, O., Beebe, N. L. Choo, K. K. R. 2020. "IoT Botnet Forensics: A Comprehensive Digital Forensic Case Study on Mirai Botnet Servers," Forensic Science International: Digital Investigation, vol. 32, 2020.

SYSTEM WERYFIKACJI TOŻSAMOŚCI NA BAZIE KALDI

KALDI BASED SPEAKER VERIFICATION SYSTEM

Adam Dustor[1]

[1] Politechnika Śląska, Gliwice, adam.dustor@polsl.pl

DOI: 10.15199/59.2022.4.44

Streszczenie: **W artykule omówiono system weryfikacji tożsamości zbudowany w oparciu o specjalistyczny i otwarty pakiet oprogramowania Kaldi oraz przetestowany na dużej i darmowej bazie VoxCeleb przy wykorzystaniu wektorów tożsamości do modelowania głosu. Omówiono także problematykę zasobów mowy wykorzystywanych do testowania głosowych systemów biometrycznych. Przedstawiono wpływ wymiarowości wektorów tożsamości na błędy weryfikacji.**

Abstract: **This paper discusses a speaker verification system built with the help of free software Kaldi and tested with the VoxCeleb dataset. The problems with the speaker recognition datasets are also discussed. The results for speaker verification system based on i-vectors are shown as well as the influence of dimensionality of i-vectors on verification errors.**

Słowa kluczowe: **bezpieczeństwo, biometria, weryfikacja mówcy.**

Keywords: **biometrics, security, speaker verification.**

1. WSTĘP

Rozpoznawanie mówcy w warunkach szumu i bez ograniczeń jest niezwykle trudnym zagadnieniem. Zastosowania rozpoznawania mówcy są bardzo liczne i różnorodne, począwszy od uwierzytelniania w systemach o wysokim poziomie bezpieczeństwa, w kryminalistyce, aż po wyszukiwanie osób w dużych zbiorach danych. Wszystkie te zadania wymagają wysokiej skuteczności rozpoznawania mówcy w warunkach rzeczywistych. Jest to niezwykle trudne zadanie z powodu wielu czynników zarówno zewnętrznych jak zakłócenia typu rozmowy i muzyka w tle, śmiech, pogłos, efekty kanałowe oraz wewnętrznych, czyli czynników związanych z mówcą (wiek, akcent, intonacja, sposób mówienia).

Istotnym problemem na drodze rozwoju systemów rozpoznawania mówcy jest brak ogólnodostępnych zasobów mowy o dużej liczbie mówców i danych uczących. Istnieje zaledwie kilka darmowych baz możliwych do wykorzystania w biometrii głosu. Najważniejsze z nich to SITW, Forensic Comparison i VoxCeleb. Istnieje co prawda wiele baz dostępnych za opłatą jak TIMIT, YOHO, NIST SRE stworzonych przez LDC. Z uwagi jednak na ich stosunkowo wysoki koszt, zbiory te nie są swobodnie dostępne dla społeczności badawczej. Otwarty dostęp do wielkich zbiorów uczących pozwolił na znaczne postępy w dziedzinie rozpoznawania mowy dzięki wykorzystaniu głębokich sieci neuronowych. Dostępność do takich zbiorów uczących dla naukowców zajmujących się rozpoznawaniem głosu z pewnością przyczyniłaby się do znacznego postępu w tej dziedzinie. Na szczególną uwagę zasługuje baza VoxCeleb, cechująca się bardzo dużą liczbą mówców i wypowiedzi a przy tym darmową.

W artykule przedstawiono wyniki badań nad zagadnieniem weryfikacji mówcy przy wykorzystaniu wektorów tożsamości (i-vector) oraz bazy VoxCeleb. Całość uczenia oraz testowania zrealizowano w oparciu o specjalistyczny pakiet Kaldi.

2. OPROGRAMOWANIE

Badania nad systemami rozpoznawania głosu można prowadzić na wiele sposobów. Najtrudniejsze i najbardziej czasochłonne podejście to samodzielna implementacja wszystkich funkcjonalności i algorytmów rozpoznawania. Takie podejście daje badaczowi pełny wgląd i wiedzę o tym co dzieje się we wnętrzu systemu biometrycznego. Niestety jest to obarczone dużą podatnością na błędy implementacyjne złożonych algorytmów w wybranym języku programowania. W rezultacie system rozpoznawania składa się ze zbioru wysoce specyficznych i skomplikowanych podprogramów czy też funkcji i klas, które są trudne do zrozumienia przez innych badaczy zajmujących się tą tematyką. Stąd częściej spotykanym podejściem jest implementacja całego systemu w oparciu o gotowe i otwarte oprogramowanie rozwijane przez szeroką społeczność badaczy z danej dziedziny. W dziedzinie przetwarzania sygnału mowy takie narzędzia są łatwo dostępne. Część z tych narzędzi jest dedykowana do badań i ich główną cechą jest elastyczność i łatwość użytkowania i modyfikacji, inne z kolei skupiają się na efektywności. Najpopularniejsze pakiety oprogramowania wykorzystywane w badaniach nad przetwarzaniem mowy to:

- Kaldi [12] – otwartoźródłowy pakiet dedykowany przede wszystkim do rozpoznawania mowy, szeroko wspierany przez liczną grupę programistów i stale rozwijany. Napisany w języku C++ zawiera bogatą bibliotekę algorytmów i wykorzystuje otwarte biblioteki BLAS i LAPACK przez co jest bardzo szybki i posiada bardzo dobrą dokumentację z gotowymi „przepisami" pozwalającymi na zbudowanie kompletnego systemu rozpoznawania mowy czy głosu.

- MSR Identity Toolbox [15] – stanowi zbiór skryptów i funkcji w języku MATLAB. Tworzenie modeli głosu możliwe jest tylko dla metod GMM-UBM i wekto-

rów tożsamości i-vector. Wykorzystanie płatnego narzędzia jakim jest Matlab stanowi poważną przeszkodę w komercyjnym wykorzystaniu tego pakietu.

- ALIZE [7] – otwartoźródłowa platforma programistyczna do badań nad systemami biometrycznymi. Dostarcza narzędzi pozwalających na budowę komercyjnych aplikacji głosowych. Pakiet napisano w C++ co przekłada się na jego dużą szybkość, ale jednocześnie jakiekolwiek modyfikacje pakietu wymagają dogłębnej znajomości oprogramowania. Poważną wadą ALIZE jest brak narzędzi do ekstrakcji parametrów z sygnału mowy jak i wizualizacji wyników działania systemu rozpoznawania. Ekstrakcja cech jest zwykle realizowana za pomocą SPro.

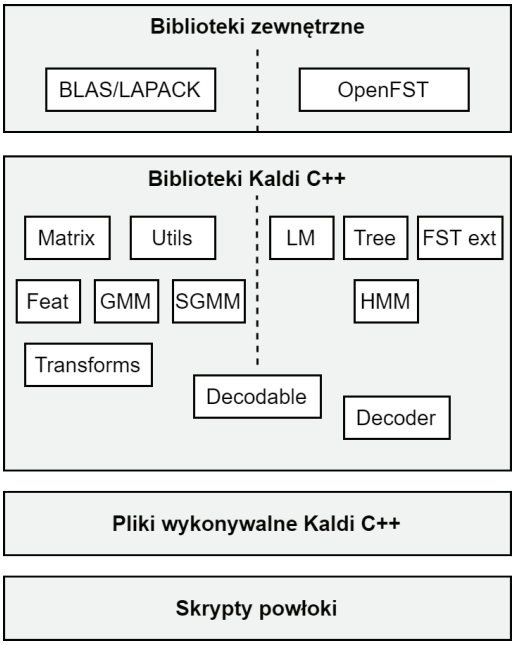
- SPro [4] – otwartoźródłowy pakiet w języku C do przetwarzania sygnału mowy z naciskiem na ekstrakcję cech. Stanowi zbiór funkcji do wczytywania i zapisu plików audio, przetwarzania sygnału (FFT, analiza LPC), przetwarzania cech (liftracja, normalizacja cepstrum, parametry delta). W pakiecie brak wysokopoziomowych funkcji pozwalających na bezpośrednią konwersję sygnału mowy w ciąg parametrów. Przy korzystaniu z tego pakietu należy uważać na rodzaj wykorzystywanej wersji SPro, gdyż wersja 64 bitowa ma poważny błąd objawiający się nieprawidłowymi wektorami cech. Wersja poprawna jest do pobrania ze strony www Uniwersytetu Avignon [5].

- Spear-BOB [6] – pakiet napisany zarówno w C++ (niskopoziomowe funkcje) i Python (warstwa wyższa) przez co jest szybki a jednocześnie pozwala na łatwe modyfikacje i testowanie różnych scenariuszy rozpoznawania głosu.

- SIDEKIT [8] – pakiet napisany całkowicie w języku Python. W pakiecie położono nacisk na kompatybilność z innymi narzędziami. SIDEKIT umożliwia odczyt i zapis plików w formacie SPro i HTK (Hidden Markow Model Toolkit), pozwala na wczytywanie modeli GMM w formacie ALIZE i HTK. Możliwy jest również zapis cech w formacie HDF5. Pakiet napisano z zamysłem minimalnej zależności od innych pakietów Pythona a w celu zwiększenia czytelności przy wykorzystaniu minimalnej możliwej liczby klas. Z uwagi na specyfikę języka Python pakiet ten jest wyraźnie wolniejszy w działaniu od Kaldi a do bezproblemowego prowadzenia obliczeń na wielu rdzeniach procesora zaleca się system Linux. W systemie Windows występują poważne problemy związane ze zrównolegleniem obliczeń (moduł multiprocessing należy zastąpić multithreading i dokonać odpowiednich zmian w kodzie pakietu).

2.1. Pakiet Kaldi

Kaldi [12] to otwartoźródłowy zestaw narzędzi do rozpoznawania mowy napisany w języku C++ i udostępniony na licencji Apache License v2.0. Charakterystyczną cechą tego pakietu oprogramowania jest nowoczesny i elastyczny kod, który jest łatwy do zrozumienia, modyfikacji i rozbudowy. Kaldi dysponuje bardzo dobrą dokumentacją, która pozwala na dokładne zaznajomienie się z dostępnymi funkcjonalnościami pakietu i dogłębne zrozumienie jego działania. Oprogramowanie powstało głównie do badań nad rozpoznawaniem mowy, jednakże z czasem stało się popularne również wśród osób zajmujących się biometrią głosu. Oprogramowanie działa w systemie Linux a jego instalacja wymaga pobrania źródeł z repozytorium i kompilacji zgodnie z instrukcjami. Złożoność pakietu Kaldi jest na tyle duża, że konieczne jest wcześniejsze zaznajomienie się z podstawami działania pakietu. Na rysunku 1 przedstawiono uproszczoną strukturę pakietu.



Rys. 1. Komponenty Kaldi [12]

Do działania pakiet wymaga dwóch bibliotek zewnętrznych jak OpenFST (konieczna, jeśli chcemy skorzystać przykładowo z modułu do modelowania HMM) oraz BLAS/LAPACK wymaganej przy operacjach macierzowych i trenowaniu modeli GMM. Moduły można podzielić na dwie odrębne połowy, z których każda zależy tylko od jednej z bibliotek zewnętrznych. Pojedynczy moduł Decodable stanowi pomost między tymi dwiema połówkami. Dostęp do funkcji bibliotek zapewniają narzędzia napisane w języku C++ i pracujące w trybie wiersza poleceń i wywoływane poprzez skrypty powłoki systemu Linux. Każde narzędzie ma bardzo specyficzną funkcjonalność, którą można uzyskać za pomocą niewielkiego zestawu argumentów wywoływanych z poziomu wiersza poleceń.

3. BAZA VOXCELEB

Baza VoxCeleb1 [10] zawiera 153516 wypowiedzi dla 1251 osób, pobranych z filmów umieszczonych w serwisie YouTube. Zbiór danych jest zrównoważony pod względem płci (55% mówców płci męskiej). Mówcy reprezentują szeroki wachlarz różnych grup etnicznych, akcentów, zawodów i grup wiekowych. W bazie zawarto również informacje o narodowości i płci każdego mówcy. Baza oprócz nagrań audio zawiera również filmy wideo z których te wypowiedzi pobrano więc może również służyć do badań nad rozpoznawaniem twarzy. Filmy zawarte w zbiorze danych zostały nagrane w wielu wymagających środowiskach akustycznych jak przykładowo: stadion, ciche wywiady w studio, przemówienia wygłaszane przed

dużą publicznością. Zawiera wycinki z nagrań profesjonalnych jak i zarejestrowanych za pomocą urządzeń przenośnych. Co najważniejsze, we wszystkich tych nagraniach występują zakłócenia jak rozmowy w tle, śmiech, nakładające się wypowiedzi, akustyka pomieszczenia. Mamy do czynienia również z różnej jakości sprzętem nagrywającym. Średnia liczba wypowiedzi na mówcę to 116 a długość wypowiedzi od 4 do 145 sekund. Baza została podzielona na część treningową (1211 mówców, 148642 wypowiedzi) oraz testową (40 mówców, 4874 nagrania).

Pewnego rodzaju sukces bazy VoxCeleb1 do którego przyczyniły się jej darmowość i obszerność były przyczynkiem do powstania jeszcze większego zasobu jakim jest baza VoxCeleb2 [1], która to zawiera 1128246 wypowiedzi 6112 mówców. Baza ta również jest podzielona na zbiór wykorzystywany do uczenia (5994 mówców i 1092009 nagrań) oraz na część testową (118 mówców i 36237 wypowiedzi). Średnia liczba wypowiedzi na mówcę to 185 a średnia długość wypowiedzi to 7,8 sekundy.

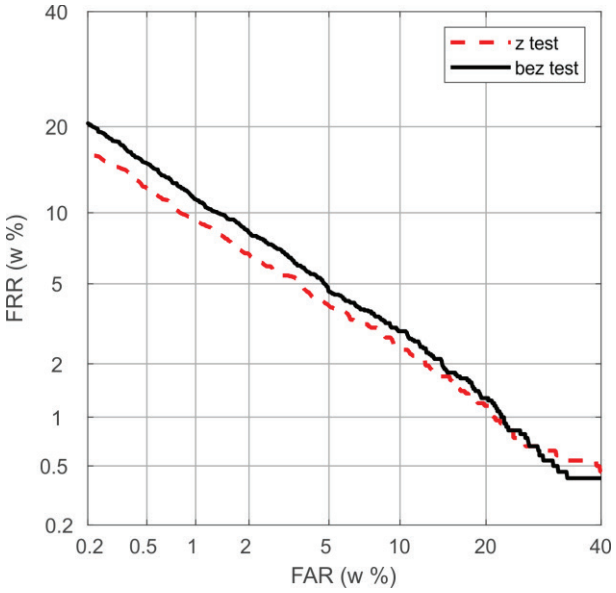
Łączna długość nagrań w obu bazach to odpowiednio 352 i 2442 godzin. Tak obszerny materiał pozwala na uczenie nie tylko klasyfikatorów opartych o metody klasyczne jak GMM, ale i bardziej wymagające w kwestii ilości danych uczących jak głębokie sieci neuronowe.

4. WERYFIKACJA GŁOSU W KALDI

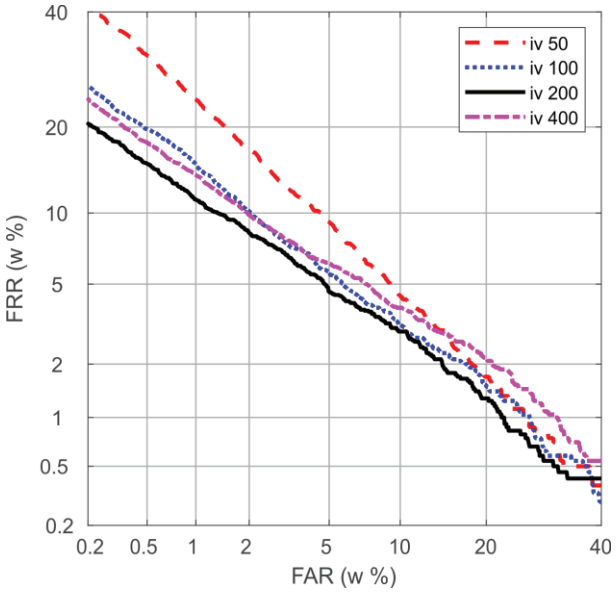
W ramach badań przetestowano weryfikację mówcy w oparciu o parametry MFCC [13], model GMM-UBM [14] oraz wektory tożsamości (i-vectors) [2] z analizą PLDA. Do badań wykorzystano całą bazę VoxCeleb1 oraz część testową bazy VoxCeleb2. Na dane uczące składały się część testowa i treningowa VoxCeleb2 i połowa testowej VoxCeleb1. Do testów wykorzystano pozostałą połowę części testowej bazy VoxCeleb1. W rezultacie łączny rozmiar danych wykorzystanych do treningu i testów zostało ograniczony do około 40GB. Badania zrealizowano na komputerze z procesorem Intel i9-9900k wyposażonym w 128GB pamięci RAM oraz dysk 4TB. Całe środowisko obliczeniowe działało pod kontrolą systemy Linux (Ubuntu 20.04 LTS). Ponieważ pakiet Kaldi nie dysponuje narzędziem do generacji krzywych DET [9] w oparciu, o które najczęściej obrazuje się działanie systemu weryfikacji, wykorzystano pakiet funkcji dostępnych na stronie NIST napisanych w języku MATLAB. We wszystkich testach wykorzystano model GMM-UBM 512. W niektórych testach (co odpowiednio zaznaczono) wymiarowość wektorów tożsamości była redukowana w oparciu o analizę LDA.

W ramach badań wstępnych sprawdzono czy dołączenie do zbioru uczącego części testowej spowoduje zmniejszenie wartości błędów systemu i w jakim stopniu. W prawidłowo przeprowadzonych testach nigdy nie wolno dopuścić, aby te same wypowiedzi wystąpiły w zbiorach uczących i testowych. Na rysunku 2 pokazano, w jakim stopniu takie nieprawidłowe podejście sztucznie poprawia wyniki. Wartość błędu EER zmalała z wartości 4,9% na 4,25%. Kolejne badania były już przeprowadzone całkowicie zgodnie ze sztuką (zbiory testowe i uczące były rozdzielne).

W kolejnym badaniu sprawdzono wpływ wymiarowości wektorów tożsamości na wartość błędu EER co przedstawiono na rysunku 3 dla odpowiednio wymiarowości 50, 100, 200 i 400. Minimalna wartość błędu EER wyniosła 4,89% dla wymiarowości 200.

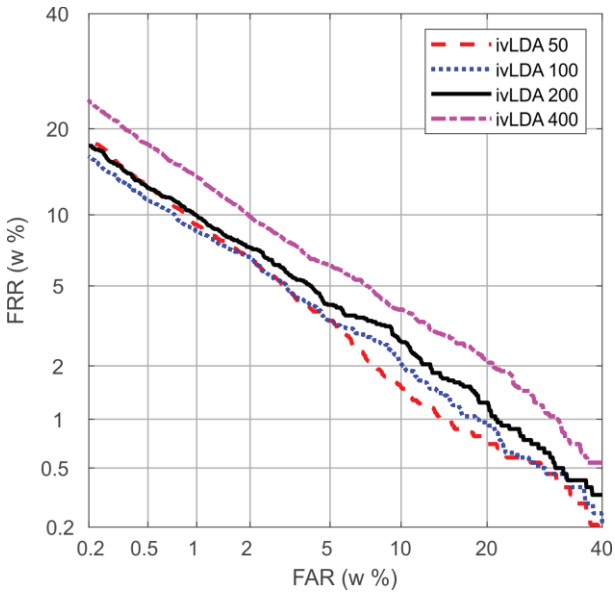


Rys.2. Wpływ dołączenia do zbioru uczącego części testowej (ivector 200, ivectorLDA 200)



Rys. 3. Wpływ wymiarowości wektorów tożsamości

W ostatnim badaniu sprawdzono czy zredukowanie w oparciu o analizę LDA wymiarowości wektorów tożsamości ma wpływ na poziom błędów weryfikacji. Badanie przeprowadzono dla wektora tożsamości o 400 współczynnikach. W oparciu o analizę LDA uzyskano wektory o odpowiednio 50, 100, 200 i 400 współczynnikach. Wyniki zaprezentowano na rysunku 4. Najniższą wartość błędu $EER \approx 4\%$ uzyskano dla parametrów ivectorLDA 50. Był to jednocześnie najlepszy rezultat osiągnięty w trakcie badań.



Rys. 4. Wpływ redukcji wymiarowości wektorów tożsamości na poziom błędów

5. PODSUMOWANIE

W artykule przedstawiono system weryfikacji zrealizowany w oparciu o dedykowany do tego celu pakiet Kaldi. Badania zrealizowano na bardzo obszernej i ogólnodostępnej bazie VoxCeleb 1 i 2 zawierającej mowę nagraną w realnych warunkach. W zwięzły sposób omówiono problematykę zasobów mowy wykorzystywanych w biometrii głosu oraz dedykowane do tego typu badań oprogramowanie. Warto zauważyć, że wyniki osiągnięte dla bazy VoxCeleb są istotnie gorsze niż dla znacznie mniejszej bazy TIMIT w której dla klasycznej metody modelowania opartego o mieszaniny Gaussa jest możliwe osiągnięcie poziomu EER rzędu 1% [3].

Osiągnięte wyniki wskazują na niezwykle istotny fakt jakim jest konieczność prowadzenia badań na maksymalnie dużych bazach testowych. Ponieważ jednak większość takich baz jest dostępna za opłatą z LDC to tym większego znaczenia nabiera fakt, że tak duża baza jaką jest VoxCeleb jest dostępna bez jakichkolwiek opłat. W dalszym badaniach planowane jest wykorzystanie jej do testów systemu opartego na głębokich sieciach neuronowych [11].

LITERATURA

[1] Chung J. S., Nagrani A., Zisserman A. 2018. „VoxCeleb2: Deep Speaker Recognition". *INTERSPEECH*.

[2] Dehak N., Kenny P. J., Dehak R., Dumouchel P., Ouellet P. 2011. „Frontend factor analysis for speaker verification". *IEEE Transactions on Audio, Speech, and Language Processing* 19(4).

[3] Dustor Adam. 2020. „Speaker verification with TIMIT corpus – some remarks on classical methods". *Proceedings of the 24th International Conference on Signal Processing Algorithms, Architectures, Arrangements and Applications (SPA)*.

[4] Gravier G. „Spro: a free speech signal processing toolkit". http://www.irisa.fr/metiss/guig/spro/.

[5] Gravier G. „SPro – without 64bit bug" https://alize.univ-avignon.fr/spro-5.0-157.tar.gz

[6] Khoury E., Shafey L., Marcel S. 2014. „Spear: An open-source toolbox for speaker recognition based on bob". *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing ICASSP*.

[7] Larcher A., Bonastre J.-F., Fauve B., Lee K. A., Lévy C., Li H., Mason J., Parfait, J.-Y. 2013. „ALIZE 3.0 - Open source toolkit for state-of-the-art speaker recognition". *Proceedings of the Annual Conference of the International Speech Communication Association INTERSPEECH*: 2768-2772.

[8] Larcher A., Lee K. A., Meignier S. 2016. „An extensible speaker identification SIDEKIT in Python". *International Conference on Audio Speech and Signal Processing ICASSP*.

[9] Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M. 1997. „The DET curve in assesment of detection task performance". *Proceedings of the EUROSPEECH*: 1895–1898.

[10] Nagrani A., Chung J. S., Zisserman A. 2017. „VoxCeleb: a large-scale speaker identification dataset". *INTERSPEECH*.

[11] Ohi Q., Mridha M. F., Hamid M. A., Monowa M. 2021. „Deep Speaker Recognition: Process, Progress, and Challenges". *IEEE Access* 9: 89619-89643.

[12] Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K. 2011. „The kaldi speech recognition toolkit". *IEEE Workshop on Automatic Speech Recognition and Understanding*.

[13] Rabiner L. R., Juang B. H. 1993. „Fundamentals of speech recognition". Prentice Hall.

[14] Reynolds D. A., Quatieri T. F. 2000. „Speaker Verification Using Adapted Gaussian Mixture Models". *Digital Signal Processing* 10: 19–41.

[15] Sadjadi S.O., Slaney M., Heck L. 2013. „MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research". *IEEE SLTC Newsletter* 1(4).

APLIKACYJNA ZAPORA SIECIOWA Z ELEMENTAMI OBRONY MIMICZNEJ
WEB APPLICATION FIREWALL BASED ON CYBER MIMIC DEFENSE APPROACH

Mariusz Sepczuk[1]

[1] Politechnika Warszawska, Warszawska, mariusz.sepczuk@pw.edu.p

DOI: 10.15199/59.2022.4.45

Streszczenie: **Artykuł przedstawia wyniki prac realizowanych w ramach konkursu na granty badawcze CyberiADa-2. Celem projektu realizowanego w ramach konkursu jest stworzenie aplikacyjnej zapory sieciowej z elementami obrony mimicznej przeznaczonej do ochrony aplikacji www. Artykuł przedstawia kolejne etapy realizacji zapory sieciowej oraz uzyskane, wstępne wyniki.**
Abstract: **The paper presents the results of project works performed under the CyberiADa-2 research grant. The project aims to create a web application firewall with elements of a mimic defense approach designed to protect the web applications. The article presents the successive stages of the firewall implementation and the obtained preliminary results.**

Słowa kluczowe: **obrona mimiczna, aplikacyjna zapora sieciowa, dynamiczna rekonstrukcja zasobów, system decyzyjny.**

Keywords: **mimic defense, web application firewall, dynamic structure reconstruction, decision system.**

1. WSTĘP

Współczesny rozwój Internetu i ogólna dostępność aplikacji www spowodowały, że stały się one łatwym celem dla atakujących. W związku z powyższym wprowadzono rozwiązania chroniące takie aplikacje, analogicznie jak to ma miejsce w przypadku ochrony na poziomie warstwy sieci. Typowa zapora sieciowa przeznaczona do ochrony aplikacji www (z ang. Web Application Firewall-WAF) [7] może wykrywać zagrożenia na trzy sposoby: na podstawie sygnatur, na podstawie logiki aplikacji i na podstawie podejścia hybrydowego. Każde z tym podejść ma swoje zalety, ale i też ograniczenia, które mają wpływ na działanie. W szczególności podejście sygnaturowe jest szybkie z punktu widzenia wdrożenia i wykrywania anomalii ruchu www, ale jest obarczone sporą liczbą próbek ruchu zaklasyfikowanych jako ruch dozwolony, podczas gdy jest to atak. Jedną z metod stosowanych do zwiększenia bezpieczeństwa w sieciach jest koncepcja obrony mimicznej. Wykorzystanie tej techniki w tym kontekście ma co najmniej dwie zalety: pozwala na wzrost wykrywalności ataków (w przypadku działania zapory sieciowej w trybie ochrony sygnaturowej) oraz zasilenie zapory o nowe reguły blokujące próby ataków wcześniej wykrytych przez elementy obrony mimicznej.

Celem projektu realizowanego w ramach konkursu na granty badawcze jest weryfikacja zastosowania koncepcji obrony mimicznej do zwiększenia poziomu bezpieczeństwa stron internetowych w sieci z działającą aplikacyjną zaporą sieciową. Niniejsza praca przedstawia kolejne etapy prac zrealizowanych w ramach projektu. Rozdział 2 artykułu prezentuje szczegóły koncepcji obrony mimicznej. Rozdział 3 zawiera opis środowiska testowego wykorzystywanego podczas badań. Rozdział 4 przedstawia uzyskane wyniki i ich analizę. W Rozdziale 5 znajduje się podsumowanie i opis przyszłych prac badawczych.

2. KONCEPCJA OBRONY MIMICZNEJ

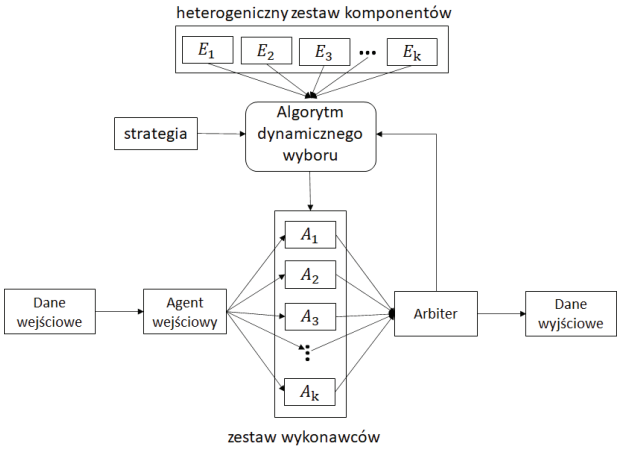
2.1. Geneza koncepcji obrony mimicznej

Wiele koncepcji w dziedzinie cyberbezpieczeństwa jest inspirowanych przez szeroko rozumianą przyrodą. Tak jest w przypadku nazewnictwa złośliwego oprogramowania (np. wirusy i robaki), jak i w przypadku kryptografii DNA (wykorzystanie molekuł sekwencji DNA w ochronie danych), czy systemów IDS/IPS (naturalny układ odpornościowy jest prototypowym przykładem naturalnego układu wspierającego obronę przed intruzami). Nie inaczej jest z obroną mimiczną.

Koncepcja obrony mimicznej wywodzi się od cech gatunku ośmiornicy zwanej z ang. mimic octopus [3]. Ośmiornica ta potrafi naśladować kilkanaście gatunków zwierząt (zazwyczaj zdolności zwierząt w tym kierunku ograniczają się do jednego) m.in.: płaszczkę, meduzę, rybę solę, węża morskiego. W zależności od kontekstu sytuacji, w którym znajduje się zwierzę (zagrożenia lub ataku), potrafi w szybki sposób dostosować się do otoczenia, zmieniając naśladownictwo jednego gatunku w drugie. Cechy te, tj. szybkie i różnorodne przystosowanie się do otoczenia w danej sytuacji zainspirowały badaczy do stworzenia koncepcji obrony mimicznej.

2.2. Wykorzystanie obrony mimicznej w cyberbezpieczeństwie

Idea obrony mimicznej wykorzystuje nowatorską teorię i metody obrony w celu wykrywania zagrożeń takich jak nieznane luki bezpieczeństwa w różnych obszarach cyberprzestrzeni. W szczególności mogą to być podatności w aplikacjach www. Głównym założeniem tej koncepcji jest wprowadzenie heterogenicznej i redundantnej architektury stosującej techniki kamuflażu w systemie docelowym w celu zwiększenia dynamiki, losowości i różnorodności tego systemu. Rysunek 1 przedstawia ogólny schemat architektury DHR (z ang. dynamic, heterogeneity and redundancy architecture) [5][9], która realizuje podejście obrony mimicznej.



Rys. 1. Architektura DHR

Na początku, dane wejściowe od użytkownika trafiają do Agenta wejściowego, który je powiela i rozsyła do wszystkich wykonawców (zestaw wykonawców), aby zostały odpowiednio przetworzone. Każdy z wykonawców składa się z innego zestawu komponentów w celu zapewnienia różnorodności i łatwiejszej detekcji podatnego komponentu. Następnie wykonawcy przesyłają odpowiedzi na dane wejściowe do modułu Arbitra, która na ich podstawie decyduje, czy nastąpił atak i wysyła odpowiednią wiadomość do użytkownika. Co więcej, w przypadku stwierdzenia ataku, stosowna informacja jest wysyłana do modułu wybierającego heterogeniczne komponenty na podstawie ustalonej strategii wyboru. Z wybranych komponentów tworzony jest nowy wykonawca i następnie podmieniany z tym, który według Arbitra dał odpowiedź wskazującą na atak.

Kiedy zestaw wykonawców jest dostatecznie duży, Agent wejściowy decyduje, do których mają trafić dane wejściowe. Oznacza to wprowadzenie pewnej losowości w uzyskiwaniu przez Arbitra odpowiedzi pochodzących od wykonawców w danym momencie. W szczególności, w chwili t_1 odpowiedzi może być n (liczba wylosowanych wykonawców, do których trafią dane wejściowe), a liczba wszystkich możliwych wariantów zestawów odpowiedzi wynosi C_k^n (k to liczba wszystkich wykonawców), podczas gdy w chwili $t_2 \neq t_1$ odpowiedzi może być m (nowa liczba wylosowanych wykonawców, do których trafią dane wejściowe) i mogą pochodzić od zupełnie innych wykonawców. Z punku widzenia bezpieczeństwa jest to utrudnienie dla atakującego, ponieważ nie jest w stanie łatwo przewidzieć, na podstawie jakiego zestawu odpowiedzi pochodzących od wykonawców i składających się na nich komponentów arbiter podejmował decyzję.

3. ZASTOSOWANIE KONCEPCJI OBRONY MIMICZNEJ

Realizacja celu postawionego we wniosku projektowym składała się z następujących etapów:

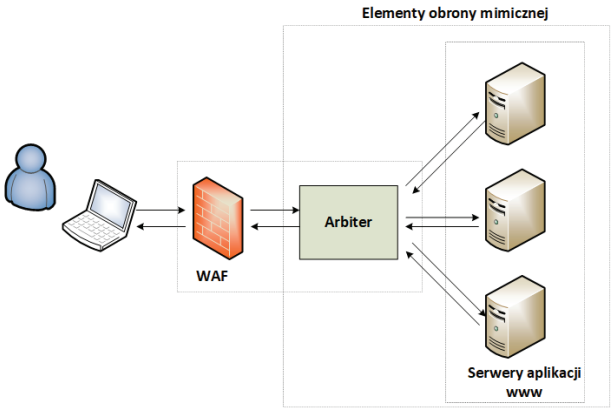
1) Propozycja architektury sieciowej z aplikacjami www chornionymi przez aplikacyjną zaporę sieciową z elementami obrochy mimicznej;
2) Stworzenie środowiska odzwierciedlajacego zaproponowaną architekturę;
3) Przeprowadzenie pomiarów, zebranie wyników i wyciągnięcie wniosków.

W tym rozdziale zostana omówione dwa pierwsze punkty, a w następnym punkt trzeci.

3.1. Idea aplikacyjnej zapory sieciowej wspieranej przez element ochrony mimicznej

Aby zwiększyć bezpieczeństwo zapewniane przez podejście sygnaturowe wykrywania ataków, aplikacyjna zapora sieciowa może zostać wzmocniona przez elementy obrony mimicznej, które będą wykrywać ataki wcześniej niewykryte (bez sygnatur). Na Rysunku nr 2 przedstawiono schemat architektury proponowanego rozwiązania. Badana sieć składa się czterech obszarów:

1) Użytkownika, który wysyła zapytania na stronę www;
2) Aplikacyjnej zapory sieciowej, która sprawdza, czy zapytanie wysłane przez użytkownika nie zawiera ataku;
3) Arbitra, elementu obrony mimicznej, który decyduje, czy nastąpiła próba ataku, którego nie wykryła aplikacyjna zapora sieciowa;
4) Serwerów aplikacyjnych www, na których uruchomiany są przeglądane przez użytkownika strony internetowe.



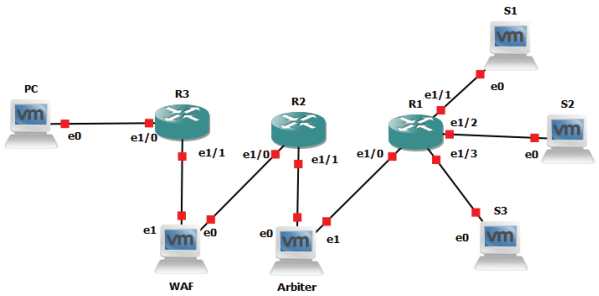
Rys. 2. Architektura sieciowa aplikacji www chronionej przez aplikacyjną zaporę sieciową z elementami obrony mimicznej

Idea działania architektury jest następująca. Zapytanie HTTP, które zostało przeanalizowane przez aplikacyjną zaporę sieciową, trafia do modułu Arbitra. Moduł ten kopiuje takie zapytanie i wysyła je do wszystkich (w tym przypadku trzech) serwerów aplikacyjnych. Każdy z serwerów ma uruchomioną tę samą aplikację, jednak w różnej konfiguracji (np. na danym serwerze znajduje się zainstalowane inne oprogramowanie serwera aplikacyjnego, w ramach którego działa aplikacja, inne wersje bibliotek programistycznych, czy też inne silniki baz danych). Każdy serwer po otrzymaniu zapytania przesyła odpowiedź ponownie do Arbitra. Arbiter na podstawie otrzymanych odpowiedzi decyduje, czy mogło dość do ataku, czy też nie. Jeśli atak nie został wykryty, to do użytkownika przesyłana jest wybrana odpowiedź na jego

wcześniejsze żądanie. W przeciwnym wypadku wysyłany jest ogólny komunikat o błędzie, a Arbiter tworzy regułę pozwalającą wykryć próbę ataku, która jest przesyłana i instalowana na WAF. Dzięki temu, gdy ktoś ponownie spróbuje wysłać zapytanie HTTP zawierający charakterystyczny, złośliwy ciąg znaków, zostanie ono wykryte już przez WAF i zablokowane.

3.2. Stworzone środowisko testowe

Omawiana we wcześniejszym podrozdziale koncepcja zostatała zaimplementowana w postaci wirtualnego środowiska w narzędziu GNS3 [1] (por. Rysunek 3), które jest graficznym emulatorem sieci, pozwalającym tworzyć i testować sieci składające się z wirtualnego sprzętu. Jako WAF zostało wybrane oprogramowanie ModSec w trybie serwera pośredniczącego (z ang. reverse-proxy) protokołu HTTP. Z kolei PC to maszyna wirtualna z systemem operacyjnym Kali Linux, która świetnia sprawdza się do przeprowadzania testów bezpieczeństwa, w szczególności testów stron www. Moduł Arbitra został napisany w języku java. Wybór języka był podyktowany dość dobrą prędkością przetwarzania danych (wolniejszą niż języki C i C++, ale wciąż dobrą do zastosowań wykonywania zadań w czasie rzeczywistym) [2][7]. Serwery S1, S2 i S3 to trzy instancje tej samej aplikacji (trzech wykonawców), składające się z różnych komponentów. W szczególności każdy serwer miał uruchomioną aplikację www w innej wersji, ale z punktu widzenia działania, tj. przesyłanego kodu aplikacji do wyświetlenia przez przeglądarkę, była to ta sama aplikacja.



Rys. 3. Schemat sieciowy testowanego środowiska

4. SCENARIUSZE TESTOWE i UZYSKANE WYNIKI

Po stworzeniu odpowiedniego środowiska do testów opracowane zostały scenariusze testowe, które miały potwierdzić lub zaprzeczyć zasadności wykorzystania elementow obrony mimicznej przy wykrywaniu ataków pzez aplilacyjną zaporę sieciową.

4.1. Scenariusze testowe

Zdefiniowanie scenariuszów testowych uwarunkowane było przez dwa czynniki: wybór odpowiedniego zestawu popularnych ataków na aplikacje www oraz istniejących rozwiązań do zarządzania treścią, które będą dostępne w kilku kolejnych wersjach. W kontekście pierwszego kryterium zdecydowano się na następujące grupy ataków i przykładowe realizacje [4]8]:

1) Wstrzyknięcia:
   a) Cross site scriptng (XSS) – podatność związana ze wstrzyknięciem złośliwego kodu javascript w przekazywanych przez użytkownika danych;
   b) SQL Injection – podatność związana ze wstrzyknięciem specjalnych instrukcji SQL celowo zmieniających całościowy sens nowego zapytania.
2) Nieautoryzowany dostęp do plików:
   a) Path traversal – podatność umożliwiająca atakującemu odczytanie dowolnych plików na serwerze, na którym uruchomiona jest aplikacja www
3) Wyciek danych – pozyskanie informacji, które nie powinny być dostępne dla każdego użytkownika,
4) Gromadzenia informacji – pozyskanie cennych danych pozwalających atakującemu na stworzenie kolejnych scenariuszów ataków.

Dużo trudniejsze okazało się wypełnienie kryterium drugiego. Zazwyczaj, jeśli aplikacja www posiada jakąś podatność, to jest ona naprawiana w kolejnej jej wersji. Nowa wersja jest upubliczniana, a stara jest usuwana, aby jak najmniejsze grono ludzi mogło z niej skorzystać. Często zdarza się, że wykryta podatność nawet po kilku naprawach ciągle występuje. W takim przypadku podatny komponent lub funkcja jest całkowicie usuwana z aplikacji. Do dość długich poszukiwaniach udało się znaleźć dwie aplikacje www: GetSimple CMS oraz Piwigo, które są dostępne publicznie w kolejnych wersjach, a każda z nich poprawia błędy bezpieczeństwa występujące w poprzednim wariancie. Aplikacje te zostały zainstalowane na serwerach w środowisku testowym w trzech różnych wersjach wraz z dodatkowymi, różniącymi się komponentami (np. baza danych, wersja PHP, itp.).

Przeprowadzone eksperymentu dotyczyły dwóch przypadków:

1) Atak był przeprowadzany w sieci chronionej przez aplikacyjną zaporę sieciową z aktualną bazą sygnatur ataków, ale bez elementów obrony mimicznej,
2) Atak był przeprowadzony w sieci chronionej przez aplikacyjną zaporę sieciową z aktualną bazą sygnatur ataków oraz dodatkowym wsparciem elementami ochrony mimicznej.

4.2. Uzyskane wyniki

Dla dwóch wymienionych w poprzednim podrozdziale przypadków zweryfikowano, czy atak został wykryty, a następnie zablokowany. Wyniki eksperymentu znajdują się w Tabeli 1 (sieć bez elementów ochrony mimicznej) i Tabeli 2 (sieć z elementami obrony mimicznej).

Tab. 1. Wykrywanie ataku tylko za pomocą aplikacyjnej zapory sieciowej

Typ ataku	*Czy zablokowano atak?*
Cross site scripting	×
SQL Injection	×
Path traversal	×
Wyciek danych	×
Gromadzenie informacji	×

Tab. 2. Wykrywanie ataku za pomocą aplikacyjnej zapory sieciowej z elementami obrony mimicznej

Typ ataku	*Czy zablokowano atak?*
Cross site scripting	✓
SQL Injection	✓
Path traversal	×
Wyciek danych	✓
Gromadzenie informacji	✓

Wyniki z Tabeli 1 wskazują, że próba ataku w żadnym przypadku nie została wykryta. Nie oznacza to, że aplikacyjna zapora sieciowa nie działała lub w ogóle nie wykrywała żadnych ataków. Z punktu widzenia atakującego proste próby ataku zazwyczaj są wykrywane, więc należy utrudnić wykrycie takiego ruchu poprzez wykorzystanie technik zaciemnienia wprowadzanego ładunku. W tym konkretnym przypadku udało się ominąć wykrycie i wprowadzone ładunki traktowane były jako dozwolony ruch. Z kolei wyniki w Tabeli 2 wskazują, że 4 próby ataków zostały wykryte i zablokowane. Jest to dość wyraźna poprawa wykrywania prób ataków w stosunku do wariantu tylko z ochroną zapewnianą przez aplikacyjną zaporę sieciową. Oczywiście dalsze badania ze zwiększoną liczbą scenariuszów testowych mogą dokładniej określić, w jakich przypadkach ochrona mimiczna wyraźnie zwiększa bezpieczeństwa aplikacji www.

Oprócz samego faktu wykrycia ataku i jego blokady, Arbiter tworzy odpowiednią regułę, zawierającą sygnaturę i przesyła ją do lokalnej bazy reguł zapory WAF. Dzięki temu przy kolejnej próbie ataku z tą samą sygnaturą zostanie ona zablokowana już na etapie analizy przez WAF i nie będzie nawet badana przez elementy obrony mimicznej.

Przy tej okazji warto wyjaśnić, w jaki sposób atak zostaje wykryty. Zgodnie z koncepcją obrony mimicznej decyzję czy próbka ruchu posiada wzorzec ataku, jest podejmowana przez arbitra. Arbiter na podstawie uzyskanych odpowiedzi od trzech serwerów aplikacyjnych, porównuje je i określa ich zgodność. Jeśli wszystkie trzy odpowiedzi są takie same prawdopodobnie nie ma ataku i użytkownik może otrzymać odpowiedź na wysłane wcześniej żądanie HTTP. W przypadku, kiedy odpowiedzi się różnią, prawdopodobnie wystąpił atak i taki ruch należy zablokować. Na podstawie takiego szybkiego i prostego porównania można określić, czy doszło do ataku i czy ruch powinien być zablokowany.

Co więcej w Tabeli 2 widać, że jeden z ataków nie został wykryty. Stało się tak, ponieważ kolejne wersje aplikacji www nie wprowadziły poprawek dla podatności Path traversal. W związku z tym Arbiter otrzymał trzy takie same odpowiedzi i nie wykrył próby ataku. Przypadek ten pokazuje sytuację, kiedy proste porównywanie odpowiedzi przez Arbitra może powodować wystąpienie decyzji fałszywie pozytywnych. W takim przypadku wymagane jest dodatkowe sprawdzenie potwierdzające pierwszą decyzję.

5. PODSUMOWANIE

W pracy tej przedstawiono kolejne wyniki prac wykonanych w ramach realizowanego konkursu na granty badawcze dotyczącego weryfikacji zastosowania koncepcji obrony mimicznej do zwiększenia bezpieczeństwa stron internetowych w sieci z działającą aplikacyjną zaporą sieciową.

Na początku omówiono ogólną ideę obrony mimicznej i jej praktyczną realizację w pracach projektowych. Następnie przedstawiono opis przeprowadzonych eksperymentów oraz zebrano otrzymane rezultaty. Uzyskane wyniki dowodzą, że zastosowanie elementów obrony mimicznej w architekturze sieciowej zwiększa bezpieczeństwo w porównaniu do tego, które jest świadczone jedynie przez aplikacyjną zaporę sieciową wykorzystującą podejście sygnaturowe.

Ponadto w pracy pokazano przypadek, dla którego podejście obrony mimicznej wymaga dodatkowego sprawdzenia analizowanego ruchu, co może stanowić pole do dalszych badań w tym obszarze.

LITERATURA

[1] Gil, Pablo, et al. 2014. "Computer networks virtualization with GNS3: Evaluating a solution to optimize resources and achieve a distance learning." *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. IEEE, strony: 1-4.

[2] Gunavaran Brihadiswaran. A Performance Comparison Between C, Java, and Python. https://medium.com/swlh/a-performance-comparison-between-c-java-and-python-df3890545f6d (dostęp: 01.05.2022)

[3] Hanlon, Roger T., Anya C. Watson, and Alexandra Barbosa. 2010. "A “mimic octopus” in the Atlantic: flatfish mimicry and camouflage by Macrotritopus defilippi." *The Biological Bulletin* 218.1, strony: 15-24.

[4] Kaur, Daljit, and Parminder Kaur. 2016. "Empirical analysis of web attacks." *Procedia Computer Science* 78, strony: 298-306.

[5] Li, Guangsong, et al. "A framework for mimic defense system in cyberspace. 2021." *Journal of Signal Processing Systems* 93.2 strony: 169-185.

[6] Nikishkov, Gennadiy P., Yu G. Nikishkov, and Vladimir V. Savchenko. 2003. "Comparison of C and Java performance in finite element computations." *Computers & structures* 81.24-25, strony: 2401-2408.

[7] Prandl, Stefan, Mihai Lazarescu, and Duc-Son Pham. 2015. "A study of web application firewall solutions." *International Conference on Information Systems Security*. Springer, Cham, strony: 501-510.

[8] OWASP Top 10 Vulnerabilities, https://www.veracode.com/security/owasp-top-10 (dostęp: 01.05.2022)

[9] Xu, Junjie. 2021. "Research on cyberspace mimic defense based on dynamic heterogeneous redundancy mechanism." *Journal of Computer and Communications* 9.7, strony: 1-7.

DETEKCJA STEGANOGRAFII OBRAZOWEJ Z UŻYCIEM PROSTYCH SIECI NEURONOWYCH

DETECTION OF IMAGE STEGANOGRAPHY USING SIMPLE NEURAL NETWORKS

Mikołaj Płachta[1]; Artur Janicki[2]

[1] Politechnika Warszawska, Warszawa, mikolaj.plachta.dokt@pw.edu.pl
[2] Politechnika Warszawska, Warszawa, artur.janicki@pw.edu.pl

DOI: 10.15199/59.2022.4.46

Streszczenie: **W tym referacie poruszamy problem wykrywania danych ukrytych steganograficznie w plikach JPEG. Zostały przedstawione najpopularniejsze algorytmy służące do ukrywania informacji w obrazach, a następnie zostały omówione algorytmy służące do ekstrakcji cech z obrazów cyfrowych. Pokazane zostały metody opisane w literaturze bazujące na ekstrakcji cech DCTR, GFR oraz PHARM oraz sieciach neuronowych w architekturze Convolution-Batch-Normalization-Dense. Zostało zaproponowane nowe rozwiązanie z wykorzystaniem prostej sieci neuronowej, przeprowadzone zostały badania dokładności oraz innych metryk dla najlepszej konfiguracji sieci.**

Abstract: **In this paper we deal with the problem of detection steganographically hidden data in JPEG files. The most popular algorithms for hiding information in images are presented and next we discussed algorithms for extracting features from digital images. We presented the methods described in the literature, which are based on the extraction of DCTR, GFR and PHARM features and using neural networks in the Convolution-BatchNormalization-Dense scheme. A new solution with a simple neural network was proposed and tested.**

Słowa kluczowe: **steganografia, głębokie uczenie maszynowe, przetwarzanie obrazów, baza BOSS**

Keywords: **steganography, deep machine learning, image processing, BOSS database**

1. WSTĘP

Steganografia to metoda ukrywania informacji niejawnych w materiałach jawnych. Innymi słowy, można ukryć tajną wiadomość w danych, które publicznie są wysyłane lub dostarczane, ukrywając samo istnienie tajnej komunikacji. Metody steganograficzne stanowią poważne zagrożenie dla użytkowników, ponieważ mogą być wykorzystywane do rozpowszechniania złośliwego oprogramowania lub mogą być wykorzystywane przez takie złośliwe oprogramowanie (tzw. stegomalware) na przykład do komunikacji C&C lub do wycieków danych wrażliwych. Znaczna część metod steganograficznych wykorzystuje jako nośnik dane multimedialne, w tym obrazy. Metody te są często określane odpowiednio jako steganografia mediów cyfrowych i steganografia obrazowa. Jednym z przykładów może być ukrywanie informacji w strukturze plików GIF [1], co jest dość innowacyjne ze względu na binarną złożoność struktury GIF.

Eksperymenty przedstawione w tym referacie są częścią zakończonego już projektu Secure Intelligent Methods for Advanced RecoGnition of malware and stegomalware (SIMARGL), finansowanego w ramach programu EU Horyzont 2020, grant nr 833042. Celem badań było znalezienie najskuteczniejszych automatycznych metod wykrywania cyfrowej steganografii w obrazach JPEG. Kompresja JPEG jest powszechnie stosowana do przechowywania i przesyłania obrazów, dlatego można ją łatwo wykorzystać do złośliwych celów. W tym badaniu przeanalizowano różne kombinacje modeli opartych na sieciach neuronowych zdolnych do wykrywania steganograficznie ukrytych treści, które mogą być potencjalnie wykorzystane przez złośliwe oprogramowanie. Tego typu metody mogą być wbudowane w oprogramowanie antymalware lub dowolne inne systemy zajmujące się skanowaniem plików w celach bezpieczeństwa (np. systemy przesyłania wiadomości).

2. POWIĄZANE PRACE

W tym referacie skupiamy na steganografii obrazowej pod kątem obrazów w formacie JPEG. Popularność tego formatu plików zaowocowała wieloma metodami ukrywania danych, a także różnymi metodami ich wykrywania, które przedstawiono w skrócie poniżej.

2.1. Steganografia obrazowa oparta na algorytmie JPEG

Wiele algorytmów steganograficznych działa w domenie przestrzennej, niektóre wprowadzają zmiany także na poziomie współczynników dyskretnej transformaty kosinusowej (DCT) przechowywanych w plikach JPEG. Często mają one na celu zminimalizowanie prawdopodobieństwa wykrycia dzięki wykorzystaniu adaptacyjności treści: osadzają one dane głównie w mniej przewidywalnych regionach, gdzie zmiany są trudniejsze do zidentyfikowania. Takie modyfikacje są najtrudniejsze do wykrycia; dlatego wybrano je do badania. Na podstawie innych badań wybrano metody: nsF5 [2], JPEG Universal Wavelet Relative Distortion (J-Uniward) [3] i Uniform Embedding Revisited Distortion (UERD) [4].

Ideą nsF5 jest wprowadzanie danych poprzez modyfikację najmniej znaczących bitów współczynników AC transformaty DCT danego pliku JPEG. Następnie ukrywanie wiadomości odbywa się poprzez tzw. kodowanie syndromów, czyli specjalnej informacji w postaci macierzy, która jest kluczem szyfrującym oraz odszyfrowującym. J-Uniward polega na modelowaniu zniekształceń steganograficznych spowodowanych wprowadzaniem

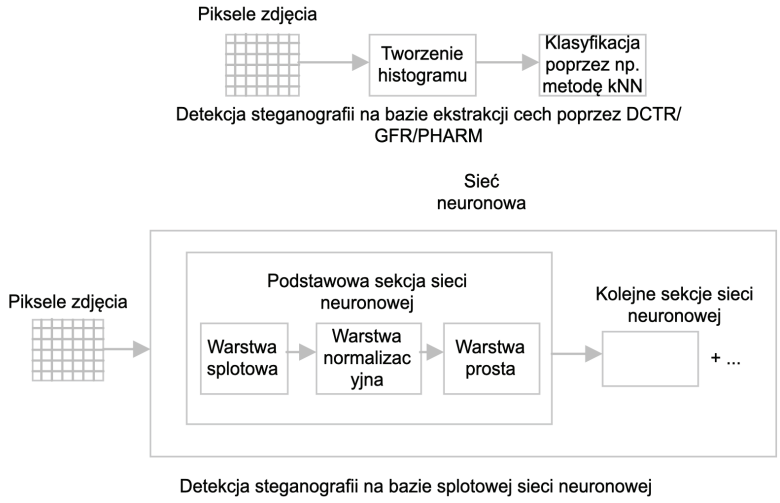
dodatkowych informacji. Modyfikowanie bazowych informacji wprowadza zmiany, które koder powinien starać się zminimalizować lub ukryć w obszarach, gdzie są one trudniejsze do wykrycia. Ostatnim badanym algorytmem jest UERD, którego celem jest zminimalizowanie prawdopodobieństwa wykrycia obecności informacji zakodowanej steganograficzne poprzez zminimalizowanie wpływu wprowadzonych danych na parametry statystyczne danych w obrazie. Osiąga to się poprzez analizę parametrów współczynników DCT danych modów, a także całych bloków DCT i ich sąsiadów.

2.3 Metody ekstrakcji cech steganograficznych

W literaturze przebadano kilka przestrzeni cech do steganalizy obrazu. Jedna z nich opiera się na metodzie Discrete Cosine Transform Residuals (DCTR) [5], której głównym celem jest analiza danych wynikających z uzyskania wartości DCT dla danego obrazu. Najpierw w tej metodzie tworzony jest losowy filtr 8x8 pikseli, który zostanie zastosowany później do przefiltrowania całego obrazu. Następnie tworzony jest histogram po zastosowaniu funkcji splotu z wcześniej wspomnianym filtrem, iterując po każdym fragmencie analizowanego obrazu. Inną odmianą tego podejścia jest metoda oparta na parametrach Gabor Filter Residuals (GFR) [6]. Działa w bardzo podobny sposób do DCTR, ale zamiast losowego filtra 8x8 stosuje się filtry Gabora. Kolejnym podejściem do parametryzacji jest użycie parametrów PHase Aware pRojection Model (PHARM) [7]. Stosuje się tu różne filtry liniowe i nieliniowe, podczas gdy histogram konstruowany jest z rzutu wartości dla każdej reszty fragmentu obrazu.

Ostatnio do wykrywania steganograficznie ukrytych danych w obrazach wykorzystywane są sieci neuronowe. Jako dane wejściowe wykorzystywano zwykle wyodrębnione parametry obrazu oparte na zdekompresowanych wartościach DCT, które wstępnie są filtrowane i wprowadzane do pierwszej warstwy splotowej sieci.

W tym celu najczęściej wykorzystywane są zastrzeżone warianty sieci splotowych, takie jak XuNet [8] czy ResNet [9]. Wspólną cechą tych sieci jest łączenie struktur Convolution-BatchNormalization-Dense (C-BN-D), tj. funkcji splotowej, warstwy normalizacyjnej i podstawowej warstwy neuronów z odpowiednią funkcją aktywacji. Wykorzystywane są funkcje takie jak sigmoidalna, TLU (Threshold Linear Unit) i gaussowska, ale najczęstszymi są Rectified Linear Unit (ReLU) lub TanH. Modele predykcyjne oparte na ekstrakcji cech oraz schemacie C-BN-D zostały przedstawione na Rys. 1.

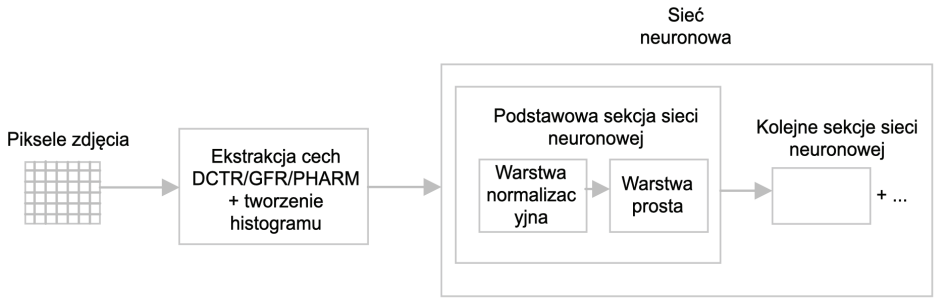


Rys. 1 Przykłady modeli predykcyjnych

3. PRZEPROWADZONE BADANIA

3.1. Zbiór danych

Wykorzystano kolekcję obrazów „Break Our Steganograhic System" (BOSS) [10], która zawiera 10 000 czarno-białych zdjęć (bez ukrytych danych). Zdjęcia zostały przekonwertowane na format JPEG ze współczynnikiem jakości 75. Następnie wygenerowano trzy inne zestawy obrazów, ukrywając tam losowe dane o gęstości (bpnzac, czyli liczby bitów na każdy niezerowy współczynnik AC) 0,4 lub 0,1, używając trzech różnych algorytmów steganograficznych: J-Uniward, nsF5 i



Rys. 2 Zaproponowany schemat modelu do badań

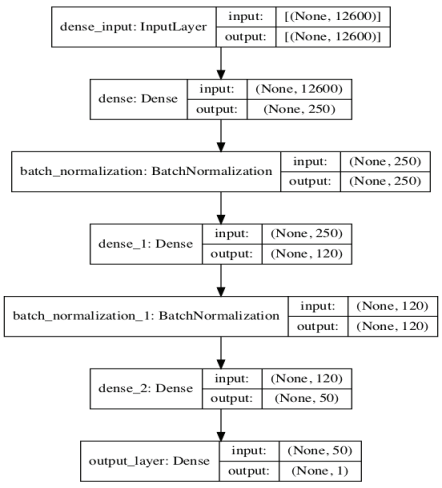
2.4 Metody detekcji steganografii

UERD. Każdy zbiór danych został następnie podzielony równolegle na podzbiory uczące i testowe, w stosunku 90:10.

3.2 Zaproponowana metoda detekcji

Zaproponowana architektura sieci została oparta głównie na strukturze Dense-BatchNormalization, ale nie korzystała z części splotowej, jak opisano w dostępnej literaturze. Część splotowa została zastąpiona ekstrakcją cech za pomocą parametryzacji DCTR, GFR lub PHARM. Schemat tej koncepcji został przedstawiony na Rys. 2. Przetestowaliśmy również różne funkcje aktywacji dla warstwy Dense, ale najlepsze wyniki uzyskano dla funkcji ReLU. Wykorzystaliśmy dwa optymalizatory: Adaptive Moment Estimation (Adam) i Stochastic Gradient Descent (SGD) które dawały różne wyniki w zależności od rodzaju parametrów wejściowych. Ostatnim parametrem, który znacząco wpłynął na efektywność uczenia się modelu, był współczynnik uczenia. Odkryto, że obniżenie go dało bardzo obiecujące wyniki bez zmiany architektury sieci i optymalizatora.

Środowisko sieci neuronowej zostało oparte na bibliotece Keras i TensorFlow ze względu na prostotę definicji modelu. Jedna z używanych konfiguracji sieci jest pokazana na Rys. 3.



Rys. 3. Przykładowy schemat użytej sieci neuronowej.

3.3 Scenariusze testowe

Do eksperymentów podsumowujących badania wybrano dwie architektury sieciowe: pierwszy z trzema warstwami z funkcją aktywacji ReLU, z 250 neuronami w pierwszej, 120 w drugiej i 50 w trzeciej, wykorzystane w czterech modelach referencyjnych, natomiast drugi z dwiema warstwami również z funkcją ReLU, mające 500 neuronów w pierwszej warstwie i 250 w drugiej, użyte w ostatnim (piątym) modelu referencyjnym. Oprócz prostych warstw sieci Dense, pomiędzy nimi zostały zastosowane dodatkowe warstwy normalizacyjne (BatchNormalization). Dwa trójwarstwowe modele wykorzystywały optymalizator Adam, a pozostałe optymalizator SGD, podczas gdy dwuwarstwowy model wykorzystywał tylko optymalizator Adam. W przypadku współczynnika uczenia dla optymalizatora Adama zastosowano wartości $1e^{-4}$ lub $1e^{-5}$, natomiast dla SGD $1e^{-3}$ lub $1e^{-4}$. W sumie przetestowano pięć różnych konfiguracji sieci neuronowych pod kątem wykrywania użycia metod steganograficznych.

3.4 Ocena efektywności modeli

Do oceny stworzonych modeli wykorzystano powszechnie stosowane metryki, takie jak dokładność, precyzja, czułość, F1-score oraz wartość AUC. Przedstawiono również same krzywe ROC, ponieważ wizualnie przedstawiają one skuteczność modelu detekcji. Ponieważ zestaw testowy jest idealnie zrównoważony, wynik dokładności nie jest obciążony i dobrze odzwierciedla zdolność detekcji danego klasyfikatora.

4 Wyniki

Tab. 1 przedstawia uzyskane wyniki dokładności podczas przeprowadzonej weryfikacji na zbiorach testowych dla wszystkich konfiguracji modeli. Najniższą dokładność uzyskano dla zbioru opartego na J-Uniward. Lepsze wyniki pod względem dokładności uzyskano dla zestawów opartych na UERD, a najlepsze osiągnięto dla nsF5. Najgorzej wypadły konfiguracje z optymalizatorem SGD, natomiast konfiguracje z optymalizatorem Adam wypadły korzystniej. Wyniki dla konfiguracji opartej na trzech warstwach i dwóch warstwach dla optymalizatora Adam są dość podobne. Modele sieci neuronowych wykorzystujących parametry PHARM uzyskiwały stosunkowo najgorsze wyniki.

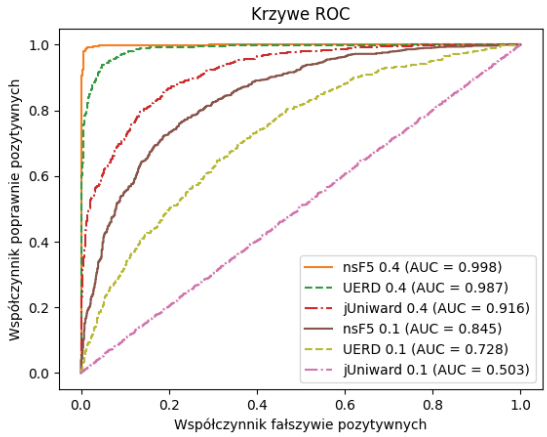
Tab. 1. Wyniki dokładności dla pięciu konfiguracji

Architektura sieci	Optymalizator	Parametry	J-Uniward		nsF5		UERD		Średnia
			0.1	0.4	0.1	0.4	0.1	0.4	
250xBNx120xBNx50 (3 warstwy)	Adam $1e^{-4}$	DCTR	–	83.1	**76.3**	98.8	**66.5**	**94.5**	**78.3**
		GFR	–	86.5	68.3	95.5	63.4	92.9	76.1
		PHARM	–	74.7	62.3	95.9	51.4	88.5	70.5
	Adam $1e^{-5}$	DCTR	–	83.0	74.2	**99.7**	64.7	93.1	77.5
		GFR	–	**88.4**	68.0	98.2	62.6	92.5	76.6
		PHARM	–	76.1	66.1	93.4	55.5	89.4	71.8
	SGD $1e^{-3}$	DCTR	–	77.0	73.8	99.6	62.8	91.4	75.8
		GFR	–	78.6	68.8	97.5	58.5	91.9	74.2
		PHARM	–	58.4	51.4	59.8	50.6	61.5	55.3
	SGD $1e^{-4}$	DCTR	–	73.3	52.1	99.1	51.6	91.9	69.7
		GFR	–	82.2	58.6	97.6	52.1	91.9	72.1
		PHARM	–	60.9	–	69.7	50.6	68.9	58.4
500xBNx250 (2 warstwy)	Adam $1e^{-5}$	DCTR	–	80.8	73.5	99.6	61.9	93.5	76.6
		GFR	**53.6**	86.4	67.6	97.4	64.2	91.9	76.9
		PHARM	–	75.0	54.1	94.2	54.0	87.9	69.2

Z drugiej strony, wyniki uzyskane dla parametrów DCTR i GFR dla wszystkich kombinacji były znacznie lepsze i raczej zbliżone, co oznacza, że najprawdopodobniej można je stosować zamiennie w narzędziach steganalitycznych dla JPEG. Cechy GFR zwykle dawały nieco lepsze wyniki dla wyższego wskaźnika gęstości (0,4 bpnzac), podczas gdy dla niższego wskaźnika gęstości (0,1 bpnzac) to właśnie przestrzeń cech DCTR okazywała się nieco lepsza dla większości testowanych modeli. Trójwarstwowa sieć z optymalizatorem Adam przy współczynniku uczenia $1e^{-4}$ osiągnęła najlepsze wyniki i to ona została wybrana do dalszych badań. W Tab. 2 przedstawiono wyniki wszystkich metryk dla najlepszej konfiguracji.

Tab. 2. Wyniki wszystkich metryk dla najlepszej konfiguracji

Metryka	J-Uniward		nsF5		UERD		Średnia
	0.1	0.4	0.1	0.4	0.1	0.4	
Dokładność	50.2	83.1	76.3	98.8	66.5	94.5	78.2
Precyzja	50.2	80.2	74.5	98.5	63.6	94.6	76.9
Czułość	46.1	87.7	80.7	99.0	78.7	94.3	81.1
F1-score	48.1	83.8	77.3	98.8	70.1	94.4	78.8
AUC	50.3	91.6	84.5	99.8	72.8	98.7	83.0



Rys. 4. Krzywe ROC dla najlepszej konfiguracji

5 Podsumowanie

Podczas eksperymentów z różnymi rodzajami warstw sieci neuronowych zauważono, że dodanie warstwy normalizacyjnej znacznie poprawiło efektywność modelu. Dla przykładu, przy metodzie nsF5 poprawiło to wyniki od około 15 do 20 pkt. procentowych, podczas gdy dla algorytmu J-Uniward 0,4 umożliwiło zbudowanie przyzwoitego modelu klasyfikującego. Bez tej warstwy sieć miała tendencję do klasyfikowania wszystkich obrazów do jednej klasy. Wskazuje to, że warstwy normalizacyjne w modelu C-BN-D są niezbędne, w przeciwieństwie do warstw splotowych, których brak można skompensować, np. wybierając inną przestrzeń cech wejściowych.

Wyniki uzyskane w naszym badaniu dla konfiguracji sieci trójwarstwowej i dwuwarstwowej były dość podobne dla optymalizatora Adam. Może to wskazywać, że powiększanie architektury sieci neuronowej nie ma sensu, gdyż może mieć jedynie negatywny wpływ na wydajność obliczeniową modelu neuronowego.

W tym referacie przeanalizowano skuteczność wykrywania ukrytych treści w obrazach JPEG przy użyciu prostych sieci neuronowych. Pokazano, że wydajność w dużym stopniu zależała od zastosowanej metody steganograficznej i gęstości ukrytych danych. Podczas gdy wykrywanie obecności treści ukrytych za pomocą algorytmu nsF5 przy gęstości 0,4 bpnzac jest prawie bezbłędne, wykrycie danych ukrytych za pomocą J-Uniward przy 0,1 bpnzac jest prawie niemożliwe, niezależnie od zastosowanej metody analizy. Jednym z celów naszego badania było znalezienie najlepszej przestrzeni cech do steganalizy obrazu. Parametry DCTR i GFR dały najlepsze wyniki, natomiast przestrzeń zbudowana na parametrach PHARM dała najgorsze wyniki. Dlatego dla celów wykrywania ukrytych danych zaleca się zastosowanie parametrów DCTR lub GFR. Szersze podsumowanie badań razem z porównaniem z płytkimi metodami uczenia maszynowego zostało przedstawione w artykule [11].

LITERATURA

[1] Puchalski, D.; Caviglione, L.; Kozik, R.; Marzecki, A.; Krawczyk, S.; Choraś, M. *Stegomalware Detection through Structural Analysis of Media Files*. Proc. 15th International Conference on Availability, Reliability and Security; Association for Computing Machinery: New York, NY, USA, 2020; ARES ’20.

[2] Fridrich, J.; Pevný, T.; Kodovský, J. *Statistically undetectable JPEG steganography*: Dead ends, challenges, and opportunities. the 9th ACM Multimedia & Security Workshop. Association for Computing Machinery, 2007, p. 3–14.

[3] Holub, V.; Fridrich, J.; Denemark, T. *Universal distortion function for steganography in an arbitrary domain*. EURASIP Journal on Multimedia and Information Security 2014

[4] Guo, L.; Ni, J.; Su, W.; Tang, C.; Shi, Y.Q. *Using Statistical Image Model for JPEG Steganography*: Uniform Embedding Revisited. IEEE transactions on information forensics and security 2015

[5] Holub, V.; Fridrich, J. *Low-complexity features for JPEG steganalysis using undecimated DCT*. IEEE Transactions on Information Forensics and Security 2015

[6] Song, X.; Liu, F.; Yang, C.; Luo, X.; Zhang, Y. *Steganalysis of adaptive JPEG steganography using 2D Gabor filters. Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security; Association for Computing Machinery*: New York, NY, USA, 2015; IH&MMSec ’15,

[7] Holub, V.; Fridrich, J. *Phase-aware projection model for steganalysis of JPEG images.* Media Watermarking, Security, and Forensics 2015; Alattar, A.M.; Memon, N.D.; Heitzenrater, C.D., Eds. International Society for Optics and Photonics, SPIE, 2015, pp. 259 – 269

[8] Xu, G.; Wu, H.Z.; Shi, Y.Q. *Structural Design of Convolutional Neural Networks for Steganalysis*. IEEE Signal Processing Letters 2016, 23, 708–712

[9] He, K.; Zhang, X.; Ren, S.; Sun, J. *Deep Residual Learning for Image Recognition.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] *Break Our Steganographic System Base webpage* (BossBase). http://agents.fel.cvut.cz/boss/. Accessed: 2022-01-18

[11] Płachta, M.; Krzemień, M.; Szczypiorski, K.; Janicki, A. Detection of Image Steganography Using Deep Learning and Ensemble Classifiers. Electronics 2022, 11, 1565. https://doi.org/10.3390/electronics11101565

Implementacja w FPGA modułu funkcji fizycznie nieklonowalnej bazującej na oscylatorach pierścieniowych

FPGA implementation of physically unclonable function module based on ring oscillators

Radosław Giermakowski[1], Mateusz Biernacki[2], Mariusz Rawski[3]

[1] Politechnika Warszawska, Warszawa, r.giermakowski@tele.pw.edu.pl
[2] Politechnika Warszawska, Warszawa, m.biernacki@tele.pw.edu.pl
[3] Politechnika Warszawska, Warszawa, m.rawski@tele.pw.edu.pl

DOI: 10.15199/59.2022.4.47

Streszczenie: **W artykule omówiono realizację modułu funkcji fizycznie nieklonowalnej PUF w układach programowalnych FPGA. Ponadto przeprowadzono szczegółową analizę wpływu liczby implementowanych inwerterów na właściwości statystyczne generowanej odpowiedzi oraz wykorzystanie zasobów FPGA. Zaproponowana implementacja ma stanowić rozwiązanie typy *lightweight* dla celów uwierzytelniania i generacji kluczy dla rozwiązań IoT.**

Abstract: **This paper discusses the implementation of a physically unclonable function PUF module in FPGA programmable circuits. In addition, a detailed analysis of the effect of the number of implemented inverters on the statistical properties of the generated response and FPGA resource utilization was performed. The proposed implementation is intended to be a lightweight solution for authentication and key generation purposes for IoT solutions.**

Słowa kluczowe: **FPGA, funkcja fizycznie nieklonowalna, IoT, oscylator pierścieniowy**

Keywords: **FPGA, IoT, physical unclonable function, ring oscillator**

1. WSTĘP

Powszechne wykorzystanie urządzeń takich jak telefony komórkowe czy karty kredytowe sprawiło, iż stały się one celem ataków, co rodzi poważne problemy związane z bezpieczeństwem i ochroną prywatności. Konwencjonalna kryptografia oferuje różne środki przeciwdziałania tym problemom, jednak większość z nich opiera się na koncepcji klucza tajnego. Zakłada się bowiem, że urządzenie jest w stanie przechowywać tajne informacje, które przy pomocy szyfrowania pozostają ukryte dla atakującego. Tajne klucze kryptograficzne przechowywane są najczęściej w pamięci nieulotnej urządzania, co wiąże się z istnieniem wielu sposobów naruszenia bezpieczeństwa. Coraz groźniejsze stają się popularne ataki typu *side-channel*, które przy stosunkowo krótkim czasie i niewielkich kosztach są w stanie doprowadzić do kradzieży klucza tajnego. Stąd też istnieje kilka reguł, które muszą być brane pod uwagę przy opracowywaniu rozwiązań mających zapewnić bezpieczeństwo wrażliwym danym:

1. bezpieczny i unikalny klucz szyfrujący powinien być generowany dla każdej instancji algorytmu szyfrowania,
2. klucz musi być przekazywany, przechowywany i pobierany w bezpieczny sposób bez możliwości jego ujawnienia,
3. algorytm bądź instancja odpowiedzialna za wygenerowanie klucza musi zagwarantować brak ujawnienia informacji o kluczu oraz wynikach pośrednich jego generowania.

1.1. Funkcja fizycznie nieklonowalna

Jednym z najbardziej obiecujących rozwiązań spełniającym wszystkie z powyższych reguł jest funkcja fizycznie nieklonowalna **PUF** (*Physical Unclonable Function*) [1]. Jest to technika w zabezpieczeniach sprzętowych, która wykorzystuje nieodłączne różnice w urządzeniach do wytworzenia nieklonowalnej, unikalnej odpowiedzi na dane wejściowe - wyzwanie (*challenge*). Jest funkcją fizyczną, ponieważ relacja par wyzwanie-odpowiedź jest określona przez złożone właściwości materiału fizycznego takiego jak zmienność procesu produkcyjnego urządzeń CMOS. Nieklonowalność natomiast wynika z faktu, że właściwości tych nie można w sposób kontrolowany odtworzyć, co sprawia, że każde urządzenie jest w rzeczywistości niepowtarzalne.

Kluczową cechą PUF jest determinizm i powtarzalność generowanej odpowiedzi w obrębie jednego urządzenia oraz ich unikatowość pomiędzy różnymi jednostkami. Pozwala to zrekonstruować losowy ciąg danych za każdym razem, gdy zachodzi taka konieczność, a więc nie ma potrzeby przetrzymywania go w pamięci urządzenia [2]. Dzięki temu wartość ta jest niewidzialna dla atakującego, co może zapewnić najwyższy poziom bezpieczeństwa i być użyte jako klucz kryptograficzny. Ponadto odpowiedź wygenerowana przez PUF może być unikatowym identyfikatorem danego urządzania. Istnieje zatem znaczne podobieństwo do biometrii człowieka, na przykład odcisku palca. Tak jak każda osoba może być zidentyfikowana na jego podstawie, tak możliwe jest zidentyfikowanie każdego urządzenia elektronicznego na podstawie jego unikalnych właściwości fizycznych. Spotykanym zastosowaniem jest także użycie PUF jako ziarna do generatorów liczb losowych.

W artykule omówiono implementację PUF wykorzystującą oscylatory pierścieniowe RO PUF (*Ring Oscillator PUF*). Zaproponowano implementację umożliwiającą redukcję zasobów układu, umożliwiając tym samym

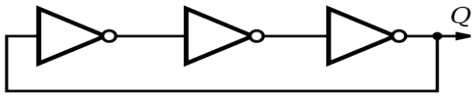
jego wykorzystanie w urządzeniach IoT do realizacji cyberbezpieczeństwa w ramach frameworku IoTrust [4]. Ponadto zaproponowano sposób realizacji RO PUF w układach FPGA firmy Xilinx. Przeprowadzono także analizę dotyczącą wpływu liczby inwerterów w oscylatorach pierścieniowych na właściwości statystyczne generowanej odpowiedzi.

2. PUF bazujący na oscylatorze pierścieniowym

RO PUF zaliczany jest do grupy funkcji fizycznie nieklonowalnych bazujących na pomiarze opóźnień. Oznacza to bowiem, że generowane losowe bity odpowiedzi są efektem porównania par strukturalnie identycznych elementów obwodu składających się z układu logicznego i połączeń międzysystemowych. W przypadku RO PUF porównywane są częstotliwości par oscylatorów pierścieniowych, gdzie źródłem opóźnień stają się stosowane inwertery.

2.1. Oscylator pierścieniowy

Oscylator pierścieniowy jest obwodem składającą się z szeregu połączonych ze sobą inwerterów, gdzie sygnał wyjściowy podawany jest z powrotem na wejście w pętli sprzężenia zwrotnego (Rys.1). Wzrost liczby użytych inwerterów powoduje zmniejszenie częstotliwości oscylacji. Losowe zmiany w opóźnieniach bramek logicznych i połączeń między nimi odzwierciedlane są w mierzonych częstotliwościach, przy pomocy asynchronicznego licznika oraz zegara referencyjnego. W większości przypadków nie jest konieczna znajomość konkretnej częstotliwości każdego z oscylatorów PUF.



Rys. 1. Oscylator pierścieniowy

2.2. Ograniczenie zajętości układu

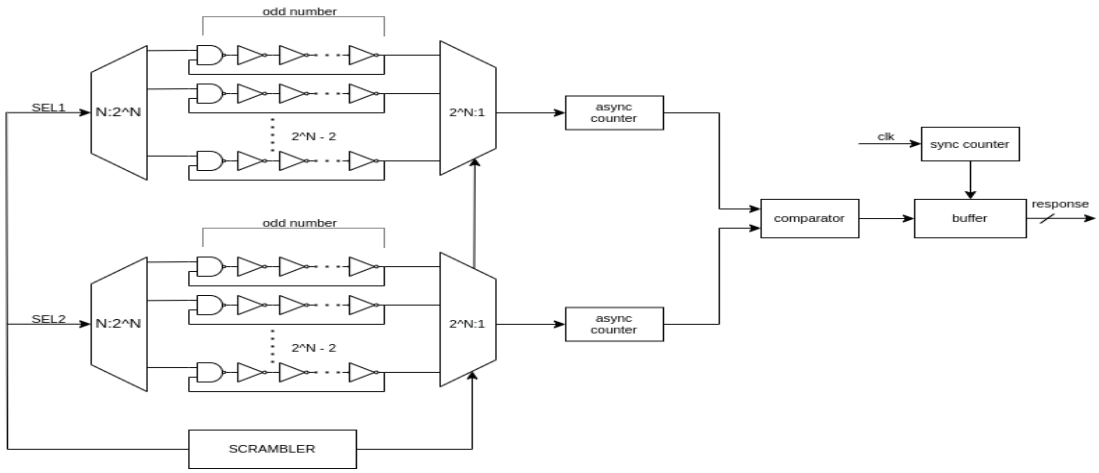
Głównym problemem RO PUF jest wysokie zużycie zasobów umożliwiających wygenerowania zaledwie jednego bitu odpowiedzi. Jest to bowiem konsekwencja wykorzystania architektury równoległej, która zakłada powielenie pojedynczych instancji PUF w zależności od oczekiwanej rozdzielczości bitowej odpowiedzi. Główną zaletą takiego rozwiązania jest jednak uzyskanie wysokiego poziomu losowości i unikalności generowanej odpowiedzi.

Mając na uwadze główne przeznaczenie implementowanego PUF należało znaleźć kompromis pomiędzy właściwościami statystycznymi a zajętością powierzchni. Dlatego też zaproponowano szeregową architekturę PUF przedstawioną na Rys.2. Główną zaletą takiego rozwiązania jest ograniczenie ilości stosowanych oscylatorów pierścieniowych, a co za tym idzie ograniczenie zużycia zasobów układu. Konstrukcja ta była możliwa dzięki wykorzystaniu modułu *scramblera*. Jego działanie polega na wygenerowaniu serii wyzwań dla PUF na podstawie początkowej wartości. Zatem każdy kolejny bit odpowiedzi jest generowany na podstawie innego wyzwania. Najprostsza realizacja scramblera może polegać na zastosowaniu zwykłego licznika, który będzie inkrementował lub dekrementował przyjętą wartość wyzwania. Prowadzi to jednak do tworzenia skorelowanych wartości wyjściowych ze względu na liniowość licznika.

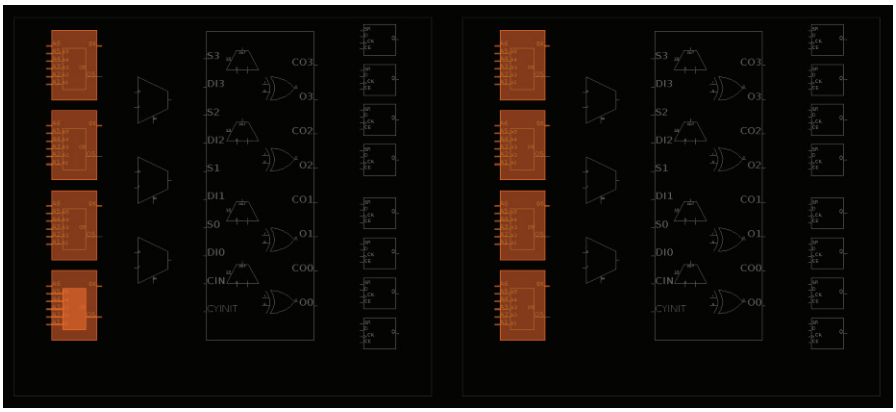
W celu uniknięcia wysokiego skorelowania bitów odpowiedzi, a co za tym idzie zwiększenia poziomu unikalności odpowiedzi dla różnych wyzwań, zaproponowano strukturę scramblera opartą na rejestrze przesuwnym z liniowym sprzężeniem zwrotnym LFSR, wykorzystywanego w pseudolosowych generatorach. Prowadzić to jednak będzie do skorelowania niektórych bitów odpowiedzi, szczególnie w przypadku źle wybranych wartości wyzwań. Aby rozwiązać ten problem należy wprowadzić pewien stopień nieliniowości układu, co może być zrealizowane poprzez obliczenie prostej operacji XOR, wartości z LFSR z początkową wartością wyzwania. Kolejne wyzwania z LFSR obliczane są natomiast na podstawie wielomianu generującego.

2.3. Propozycja architektury RO PUF

Zaproponowano architekturę (Rys. 2), która zakłada wykorzystanie dwóch demultiplekserów, które sterują bramkami NAND oscylatorów pierścieniowych, co pozwala na ich włączenie/wyłączenie. Zabieg ten jest stoso-



Rys. 2. Propozycja architektury RO PUF



Rys. 3. Implementacja RO w fizycznych strukturach FPGA

wany w celu ograniczenia zużycia energii układu. Następnie implementowanych jest szereg oscylatorów pierścieniowych, których konkretna liczba jest w pełni konfigurowalna. Na podstawie wartości ze scramblera, multipleksery wybierają odpowiednią parę oscylatorów, których to narastające zbocza porównywane są w asynchronicznych licznikach. Implementowany jest także dodatkowy synchroniczny zegar referencyjny odpowiedzialny za kontrolowanie czasu porównania. Na wyjściu zaś znajduje się bufor, który w formie rejestru przesuwnego zatrzaskuje kolejne bity odpowiedzi.

3. Implementacja oscylatora pierścieniowego

Najważniejszym elementem PUF, a zarazem sprawiającym największe problemy jest oscylator pierścieniowy. Jego implantacja w języku opisu sprzętu HDL, polegająca na szeregowym zestawieniu bramek NOT wiąże się z wieloma problemami. Przede wszystkim wykorzystanie narzędzia CAD do syntezy i implementacji takiego projektu prowadzi do zoptymalizowania struktury oscylatora i usunięcia zbędnych, z punktu widzenia wartości logicznych uzyskanych na wyjściu, inwerterów.

Użycie makr narzędzia CAD nie rozwiązuje problemu, ponieważ proces ich implementacji w fizycznych strukturach FPGA prowadzi do występowania znaczących różnic w strukturze każdego z pojedynczych oscylatorów pierścieniowych. Efektem takiej implementacji jest ograniczenie powtarzalności generowanej odpowiedzi.

3.1. Propozycja implementacji

W układach programowalnych firmy Xilinx głównym zasobem służącym do implementacji układów kombinacyjnych czy sekwencyjnych są konfigurowalne bloki logiczne **CLB** (*configurable logic block*). W przypadku architektury ZYNQ-7000 każdy blok CLB sklada się z dwóch bloków SLICE. Pojedynczy zaś blok SLICE zbudowany jest między innymi z czterech 6-wejściowych tablic **LUT**. To własnie ich wykorzystanie okazuje się kluczowe w opisywanym rozwiązaniu. Użycie predefiniowanych instancji LUT6 jako bramek NOT czy NAND sprowadza się jedynie do opisania tablicy przejść-wyjść. Zastsowanie takiego rozwiązania oraz dodatkowych makr narzędzia CAD umożliwiających kontrolę nad procesem syntezy sprawia, że każda z bramek zostanie zaimplementowana identycznie, a żadna z nich nie ulegnie zoptymalizowaniu.

Zaproponowane rozwiązanie jednak nie jest kompletne ze względu na fakt, iż narzędzie CAD Vivado podczas fazy implementacji losowo umieszcza bramki po całej matrycy FPGA, nie tworząc identycznych struktur oscylatorów. Idealnym rozwiązaniem jest doprowadzenie do sytuacji, w której każdy pojedynczy RO może zostać umieszczony w pojedynczym bloku SLICE. W docelowym rozwiązaniu ściśle zdefiniowano lokalizację każdej tablicy LUT, przypisując jej odpowiednie miejsce w danym bloku SLICE, jak i określając lokalizację SLICE na matrycy FPGA. W tym celu zdefiniowano plik *.xdc, który przy wykorzystaniu dedykowanych makr umożliwiających kontrolę nad procesem odwzorowania technologicznego gwarantował identyczną strukturę każdego oscylatora. Przykład odwzorowania w fizycznych zasobach układu FPGA oscylatora pierścieniowego składającego się z czterech bramek NOT, bramki NAND i dodatkowej tablicy LUT6 realizującej sprzężenie zwrotne przedstawiono na Rys. 3.

4. Analiza uzyskanych wyników

4.1. Podstawowe parametry

Ocenę miary jakości zaimplementowanego RO PUF można dokonać za pomocą czterech podstawowych parametrów [5]. Pierwszym z nich jest równomierność (*uniformity*), która określa jak równomierny jest rozkład „0” i „1” w bitach odpowiedzi PUF. W idealnych warunkach, proporcja ta powinna wynosić 50%. Kolejny parametr to unikalność (*uniqueness*) wewnątrz urządzenia (*intra-chip*), który określa stopień unikatowości generowanych odpowiedzi dla różnych wartości wyzwań. Unikalność pomiędzy urządzniami (*inter-chip*) reprezentuje natomiast zdolność PUF do odróżnienia konkretnego urządzenia od innych urządzeń tego samego typu. Obliczana jest na podstawie odległość Hamminga pomiędzy odpowiedziami PUF różnych urządzeń i tej samej wartości wyzwania. Ostatnim parametrem jest niezawodność (*reliability*) określająca jak powtarzalna jest generowana odpowiedź PUF. Aby obiektywnie ją ocenić obliczana jest odległość Hamminga dla wielu

Tab. 1. Analiza wpływu liczby zaimplementowanych inwerterów na generowaną odpowiedź

Liczba inwerterów	Równomierność [%]	Unikalność (inter-chip) [%]	Unikalność (intra-chip) [%]	Niezawodność [%]
3	48,239	37,018	49,612	96,685
5	50,661	30,612	49,889	97,773
7	51,509	32,612	49,302	97,826

odpowiedzi generowanych na podstawie tej samej wartości wyzwania.

4.2. Omówienie wyników

Na podstawie opisanych powyżej parametrów dokonano analizy zależności jakości generowanej odpowiedzi RO PUF od liczby użytych inwerterów w oscylatorach pierścieniowych. Ponadto zbadano zajętość zasobów układu FPGA dla każdego z testowanych wariantów. Architektura implementowanego układu składała się z 64 oscylatorów pierścieniowych o liczbie inwerterów kolejno: 3, 5 oraz 7. Układ generował 64-bitową odpowiedź.

Dane przedstawione w Tab. 1 ukazują, że jedynym parametrem, na który kluczowy wpływ ma liczba zaimplementowanych inwerterów inwerterów to unikalność wygenerowanej odpowiedzi pomiędzy różnymi urządzeniami. Wartości te jednak są dalekie od 50%, co jest związane z faktem, iż zmienność procesu technologicznego w nowych urządzeniach jest coraz niższa. Rzutuje to bowiem na zmniejszenie różnic w częstotliwościach danych oscylatorów na różnych urządzeniach. Nie zmienia to jednak faktu, iż wartość unikalności dla 3 inwerterów pozwala z powodzeniem na stosowanie RO PUF jako uniklanego identyfikatora danego urządzenia. Pozostałe parametry ukazują niemal idealne zachowanie testowanego modułu. Dla każdego przypadku generowana odpowiedź charakteryzuje się prawie równomiernym rozkładem zer i jedynek oraz unikalnością mierzoną dla pojedynczego urządzenia bliską 50%. Niezawodność mimo, iż cechuje się wartością około 98% wyklucza użycie PUF do generowania klucza kryptograficznego. Jest to typowy problem PUF bazującego na pomiarze opóźnień. Wynika on z faktu fluktuacji częstotliwości oscylatorów pierścieniowych w zależności od czynników środowiskowych, takich jak na przykład temperatura. Problem ten jest możliwy do rozwiązania dzięki zastosowaniu korekcji błędów oraz algorytmu danych pomocniczych [3].

Tabela 2. przedstawia natomiast zużycie zasobów układu FPGA Artix-7. Uzyskane wartości jednoznacznie potwierdzają, że zaimplementowano moduł PUF o znikomym zużyciu elementów logicznych układu, co jest zgodne z jego przeznaczeniem w urządzeniach o ograniczonych zasobach. Zastosowanie szeregowej struktury układu pozwoliło znacząco zredukować liczbę użytych inwerterów, co ograniczyło wykorzystanie bloków SLICE.

Podsumowując, na podstawie uzyskanych danych można wyciągnąć wniosek, iż najkorzystniejszym rozwiązaniem jest implementacja oscylatorów pierścieniowych składających się z trzech inwerterów. Cechują się one najwyższą częstotliwością oscylacji, co zapewnia wysoki poziom unikalności generowanej odpowiedzi, tym samym ogranicza zużycie zasobów układu przy jednoczesnym niewielkim obniżeniu niezawodności. Jednak zakładając, iż każdy z przypadków wymaga poddania odpowiedzi PUF korekcji błędów, niewielki spadek wartości niezawodności nie niesie za sobą żadnych konsekwencji.

Tab. 2. Wykorzystanie zasobów układu FPGA: Artix-7

Liczba inwerterów	SLICE LUTS (17600)	SLICE REGISTERS (35200)	SLICE (4400)
3	474	148	174
5	603	148	232
7	730	148	235

FINANSOWANIE

Praca współfinansowana przez Narodowe Centrum Badań i Rozwoju w ramach projektu Nr CYBERSECIDENT/456446/III/NCBR/2020.

LITERATURA

[1] Anagnostopoulos, N., A., Katzenbeisser, S., Chandy, J. and Tehranipoor, F. 2018 .“An overview of DRAM-based security primitives. Cryptography”, 2(2), p.7.

[2] Herder, C., Yu M., Koushanfar F. and Devadas, S., 2014. "Physical Unclonable Functions and Applications: A Tutorial," in Proceedings of the IEEE, vol. 102, no. 8, pp. 1126-1141.

[3] Delvaux, J., Gu, D., Schellekens D. and Verbauwhede, I. 2015. "Helper Data Algorithms for PUF-Based Key Generation: Overview and Analysis," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 6, pp. 889-902.

[4] Chmiel, M., Korona, M., Kozioł, F., Szczypiorski, K. and Rawski, M. 2021. „Discussion on IoT Security Recommendations against the State-of-the-Art Solutions. Electronics”, 10(15).

[5] Maiti, A., Gunreddy, V. and Schaumont, P., 2013. A Systematic Method to Evaluate and Compare the Performance of Physical Unclonable Functions. In: Athanas, P., Pnevmatikatos, D., Sklavos, N. (eds) Embedded Systems Design with FPGAs. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-1362-2_11.

IMPLEMENTACJA ARCHITEKTURY SYSTEMU SYMEC

DEVELOPMENT OF SYMEC ARCHITECTURE

Andrzej Bęben[1]; Wojciech Burakowski[1]; Maciej Sosnowski[1,2]; Dariusz Więcek[2]; Paweł Olender[2], Bartosz Belter[3]; Bartosz Krakowiak[3]; Michał Hoeft[4]; Mariusz Miszewski[5]; Adrian Chomicki[5]

[1] Politechnika Warszawska, Warszawa, {abeben, wojtek, m.sosnowski}@tele.pw.edu.pl
[2] Instytut Łączności - PIB Warszawa, {d.wiecek, p.olender}@il-pib.pl
[3] Poznańskie Centrum Superkomputerowo Sieciowe, Poznań {bartosz.belter, bartosz.krakowiak}@man.poznan.pl
[4] Politechnika Gdańska, Gdańsk, michal.hoeft@pg.edu.pl
[5] DGT Sp. z o.o, Gdańsk, {mariusz.miszewski, adrian.chomicki}@dgt.pl

DOI: 10.15199/59.2022.4.48

Streszczenie**: W artykule przedstawiono docelową architekturę systemu obliczeń na brzegu sieci, która została opracowana, zaimplementowana i wdrożona w ramach projektu SyMEC. W szczególności przedstawiono główne elementy opracowanego systemu, podstawowe realizowane procesy dotyczące zarządzania cyklem życia oferowanych aplikacji i usług MEC, a także doświadczenia wynikające z implementacji prototypu systemu SyMEC i jego wdrożenia w krajowej sieci badawczej PL-LAB 2020. W podsumowaniu przedstawiono kierunki dalszego rozwoju system SyMEC.**

Abstract**: The article presents the final design of the edge computing system developed by the SyMEC project. We describe the main elements of the developed system, the fundamental processes related to the lifecycle management of the MEC applications and services, and the experiences coming from the system deployment in the PL-LAB 2020 research network. The summary presents further research directions.**

Słowa kluczowe**: technika obliczeń na brzegu sieci, MEC, OpenNESS, SyMEC, prototypowanie i testowanie**

Keywords**: edge computing, MEC, OpenNESS, SyMEC Project, prototyping and testing**

1. WSTĘP

Systemy obliczeń na brzegu sieci, nazywane systemami MEC (*Multi-access Edge Computing*) [1] lub „edge/fog compuiting", są powszechnie uważane za jeden z podstawowych elementów funkcjonalnych sieci 5G/6G oraz przyszłej zintegrowanej infrastruktury sieciowo-obliczeniowej. Systemy MEC oferują zasoby obliczeniowe blisko użytkowników końcowych, np. przy stacjach bazowych lub w innych punktach agregacji ruchu, co pozwalana na przetwarzanie danych u ich źródła. W porównaniu do systemów chmurowych, systemy MEC pozwalają na istotne zmniejszenie opóźnienia związanego z przekazem i przetwarzaniem danych oraz zmniejszenie obciążenia sieci. Z tego względu infrastruktura MEC jest szczególnie atrakcyjnym środowiskiem dla wdrożenia aplikacji rozszerzonej/mieszanej rzeczywistości, systemów sterowania autonomicznymi pojazdami czy też zastosowań Przemysłu 4.0.

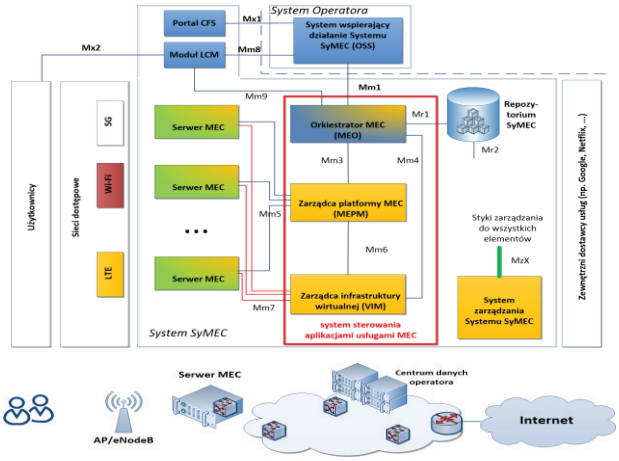
Wynikiem prac projektu jest specyfikacja, implementacja, integracja, testy i wdrożenie instalacji pilotowej prototypu systemu SyMEC realizującego obliczenia na brzegu sieci. Jego centralnym elementem jest system sterowania i zarządzania aplikacjami MEC [3]. Odpowiada on za zarządzanie cyklem życia aplikacji i usług oferowanych użytkownikom i realizuje procesy: 1) rejestracji, udostępnienia oraz usunięcia obrazu aplikacji z systemu. Procesy te są inicjowane przez usługodawców, którzy chcą świadczyć swoje usługi wykorzystując infrastrukturę SyMEC; 2) uruchomienia i zatrzymania instancji aplikacji MEC, wywoływane bezpośrednio przez użytkowników. W ramach projektu zbudowano serwer SyMEC, który został zaprojektowany, aby sprzętowo wspierać zarówno funkcje infrastruktury MEC, np. wirtualizację, przekaz danych, czy też zapewnienie bezpieczeństwa, jak również wspierać aplikacje MEC w realizacji m.in. analizy i kodowania obrazów wideo, obliczeń równoległych, algorytmów sztucznej inteligencji [2]. Ponadto elementami systemu SyMEC są również: repozytorium aplikacji MEC umożliwiające przechowywanie obrazów aplikacji MEC, a także platforma MEC wspierająca aplikacje MEC i moduły systemu zarządzania, zapewniające integrację systemu SyMEC z systemami OSS/BSS (*Operations/ Business Support System*) operatora [3]. Ważnymi elementami systemu SyMEC są również moduły współpracy z sieciami dostępowymi WLAN, 4G czy 5G. Moduły te gromadzą informacje kontekstowe o stanie sieci dostępowej, wykorzystywanych zasobach, częstotliwościach, obciążeniu sieci, dostępności użytkowników, oferowanych usługach itp. Informacje te są następnie udostępniane aplikacjom MEC za pomocą funkcji API udostępnionych przez platformę MEC.

Przedstawiony prototyp systemu SyMEC został zainstalowany, przetestowany i udostępniony w środowisku ogólnopolskiej, rozległej sieci badawczej PL-LAB 2020.

Organizacja artykułu jest następująca. W rozdziale 2 przedstawiono architekturę systemu SyMEC oraz omówiono podstawowe realizowane procesy. Rozdział 3 przedstawia prototyp systemu SyMEC, który został uruchomiony w środowisku Kubernetes (k8s). Instalację prototypu w sieci PL-LAB 2020 przedstawiono w rozdziale 4. W rozdziale 5 podsumowano artykuł i przedstawiono kierunki dalszych prac.

2. ARCHITEKTURA SYSTEMU SYMEC

Opracowaną architekturę systemu SyMEC przedstawiono na Rys. 1. Architektura ta jest wzorowana na architekturze systemu obliczeń na brzegu sieci zaproponowanej przez grupę standaryzacyjną ETSI MEC [1].



Rys. 1. Architektura systemu SyMEC.

Architektura SyMEC obejmuje sześć podstawowych elementów:

1. System sterowania i zarządzania aplikacjami i usługami SyMEC, który składa się z: orkiestratora MEC odpowiedzialnego za obsługę sygnalizacji wykorzystywanej w procesie obsługi żądań usługodawców i użytkowników, wybór odpowiedniej lokalizacji instancji MEC, a także za zarządzanie zasobami infrastruktury MEC, wykorzystując w tym celu jej abstrakcyjny model; zarządcy platformy MEC - MEPM (*MEC Platform Manager*) odpowiedzialnego za koordynację uruchomienia/zatrzymania instancji aplikacji MEC w rozproszonej infrastrukturze MEC, a także translację formatu żądań do postaci wymaganej przez zastosowaną platformę wirtualizacją; oraz zarządcy wirtualnej infrastruktury VIM (*Virtual Infrastructure Manager*), który wykonuje żądania Zarządcy MEPM i bezpośrednio uruchamia instancje aplikacji MEC na uprzednio wybranym serwerze.

2. Serwer MEC, który wraz z zaimplementowaną na serwerze platformą wirtualizacją udostępnia zasoby obliczeniowe, tj. moc obliczeniową procesorów CPU i akceleratorów GPU, pamięć operacyjną i masową, oraz umożliwia uruchomienie i udostępnienie instancji aplikacji MEC dla użytkowników, usług platformy MEC oraz modułów systemu sterowania. Ponadto serwer MEC zawiera przełącznik sieciowy, który umożliwia utworzenie niezawodnej i bezpiecznej sieci serwerów MEC, zapewniającej komunikację pomiędzy 1) elementami systemu SyMEC, 2) uruchomionymi instancjami aplikacji MEC a ich użytkownikami dołączonymi do sieci dostępowych, a także 3) elementami otoczenia, np. wybranymi adresami w sieci operatora lub sieci Internet.

3. Repozytorium aplikacji MEC odpowiada za zarządzanie obrazami aplikacji MEC. Repozytorium przechowuje te obrazy i udostępnia je użytkownikom systemu SyMEC do zainstalowania na serwerze. Stan repozytorium jest uaktualniany w procesach rejestracji, udostępnienia oraz usunięcia obrazu aplikacji. Ważną funkcją repozytorium jest również zarządzanie i przechowywanie deskryptorów aplikacji MEC zawierających pełen opis wymagań aplikacji. W szczególności deskryptory te zawierają unikalny identyfikator aplikacji, metadane opisujące aplikację, opis zasobów wymaganych do uruchomienia instancji aplikacji (w tym wymaganą architekturę procesora), wymagania dotyczące przekazu danych i dołączenia użytkownika, m.in. rodzaj wykorzystywanego protokołu, numery wykorzystywanych portów, preferowaną nazwę symboliczną DNS (*Domain Name System*). Deskryptory aplikacji są wykorzystywane przez system m.in. do wybrania odpowiedniego serwera i skonfigurowania uruchomionej aplikacji.

4. Platforma MEC jest zbiorem usług oferowanych w ramach klastra obliczeniowego SyMEC, które wspierają działanie uruchomionych instancji aplikacji MEC, oferując funkcje automatycznej rejestracji i wykrywania instancji aplikacji, uwierzytelniania i zarządzania certyfikatami bezpieczeństwa, zarządzanie dostępem do aplikacji MEC przez użytkowników, oraz rejestrowanie i rozgłaszanie adresów domenowych dla udostępnionych aplikacji MEC. Ponadto platforma MEC udostępnia bezpieczne kanały wymiany wiadomości w trybie „publikuj/subskrybuj" bezpośrednio pomiędzy instancjami aplikacji MEC, oraz pomiędzy aplikacjami a systemem SyMEC, w ramach których, zgodnie z przydzielonymi uprawnieniami, są udostępniane informacje o stanie systemu i sieci dostępowej czy informacje kontekstowe związane z lokalizacją użytkowników.

5. Moduły współpracy z sieciami dostępowymi WLAN, 4G, 5G zbierają informacje o stanie sieci dostępowej i realizowanych połączeń oraz przekazują te informacje platformie MEC. Platforma udostępnia te informacje w zestandaryzowanej formie w postaci funkcji API wywoływanych przez aplikacje MEC. Należy zwrócić uwagę, że jest koniecznym opracowanie specjalizowanych modułów współpracy dla każdej z technik sieci dostępowej ze względu na istotne różnice w zbiorach charakteryzujących je parametrów. Moduły te powinny być również dostosowane do możliwości urządzeń zastosowanych w sieci dostępowej. W tym kontekście istotną funkcją modułów współpracy jest również zapewnienie możliwości przekazu danych i dostosowanie formatu przesyłanych wiadomości pomiędzy terminalem użytkownika dołączonym do punktu świadczenia usług a aplikacją MEC uruchomioną na serwerze SyMEC.

6. System zarządzania SyMEC obejmuje zestaw modułów, które umożliwiają pełną integrację systemu SyMEC, jako podsystemu oferującego usługi i aplikacje MEC, w ekosystemie OSS/BSS operatora. Z tego względu moduły zarządzania opracowano, zapewniając ich pełną zgodność ze standardami TeleManagement Forum (TMF) [4]. Opracowano następujące moduły: i) moduł zarządzania zasobami, który umożliwia gromadzenie oraz przechowywanie informacji o zasobach obliczeniowych serwerów SyMEC, lokalizacji punktów dostępowych oraz topologii połączeń wraz z charakterystyką przekazu danych pomiędzy serwerami i punktami dostępowymi, ii) moduł monitorowania, który zbiera i udostępnia informacje o stanie serwerów, ich obciążeniu i wykorzystaniu oraz statusie uruchomionych aplikacji MEC.

Na podstawie doświadczeń związanych z zaprojektowaniem systemu SyMEC należy zwrócić uwagę, iż architektura ETSI MEC jest jedynie pewnym abstrakcyjnym modelem systemu obliczeń na brzegu sieci. Model ten wymaga istotnych modyfikacji i uzupełnień, co wynika z faktu, iż system MEC jest jednym z podsystemów złożonego ekosystemu infrastruktury operatora i musi zostać w pełni zintegrowany z pozostałymi podsystemami tej infrastruktury takimi jak podsystem OSS/BSS, czy też podsystem zarządzania. Ponadto, zastosowanie abstrakcyjnego modelu zasobów jest korzystne dla działania orkiestratora czy interfejsu użytkownika, ponieważ umożliwia ukrycie fizycznej realizacji obiektów, co upraszcza metody sterowania zastosowane w heterogenicznym systemie złożonym z różnych klastrów obliczeniowych. Jednakże, specyfikacja i implementacja zarządcy platformy MEC, zarządcy VIM, repozytorium aplikacji czy modułów współpracy z sieciami dostępowymi w znacznym stopniu zależy od zastosowanej platformy wirtualizacji oraz wykorzystywanych technik sieciowych. Podstawowym elementem, który odwzorowuje świat abstrakcyjnych obiektów używanych w modelu na ich fizyczną reprezentację, jest zarządca platformy MEC wraz z modułami systemu zarządzania. Należy również podkreślić, że zastosowana w prototypie platforma wirtualizacji kontenerowej bazująca na narzędziach Docker [6] i Kubernetes (k8s) [7] wymusiła również rozszerzenie deskryptora aplikacji MEC względem opisu zdefiniowanego przez ETSI oraz wpłynęła na repozytorium obrazów aplikacji.

W celu zapewnienia współpracy z pozostałymi systemami operatora w architekturze SyMEC zachowano pełną zgodność styków zewnętrznych ze standardami ETSI MEC oraz TMF. W szczególności zachowano zgodność styku Mm1 pomiędzy orkiestratorem a systemem OSS/BSS, który umożliwia inicjowanie uruchomienia/zatrzymania instancji aplikacji MEC lub rejestrowanie/usunięcie nowej usługi/aplikacji MEC. Również zachowano zgodność opisu zasobów systemu SyMEC ze standardami TMF, co umożliwia łatwą integrację systemu SyMEC z systemami zarządzania operatora. Ponadto zachowano zgodność interfejsu programowego API aplikacji MEC ze standardami ETSI [5], aby umożliwić łatwe wdrożenie aplikacji MEC opracowanych przez niezależnych twórców. Pozostałe styki zdefiniowano wykorzystując otwarte interfejsy REST API lub, jak w przypadku styku Mm6, wykorzystano API zdefiniowane w narzędziu Kubernetes.

Podsumowując, architektura SyMEC, w porównaniu do architektury ETSI MEC, została rozbudowana o następujące elementy: i) moduły współpracy z systemami operatora, m.in. z systemem OSS/BSS, a także z systemem zarządzania siecią operatora; ii) repozytorium obrazów aplikacji MEC zarejestrowanych w systemie, iii) moduły sieci sterowanej programowo, która zapewnia bezpieczną i niezawodną komunikację w rozproszonym klastrze obliczeniowym utworzonym przez serwery MEC zlokalizowane bezpośrednio na brzegu sieci (*far edge*), ale również w lokalnych centrach obliczeniowych (*near edge*) jak i regionalnych/centralnych centrach obliczeniowych operatora, a także iv) moduły zapewnienia komunikacji pomiędzy terminalem użytkownika dołączonym do danego punktu dostępowego sieci dostępowej a dedykowaną instancją aplikacji MEC uruchomioną na wybranym serwerze rozproszonego klastra obliczeniowego.

3. PROTOTYP SYSTEMU SyMEC

W ramach projektu opracowano prototyp systemu SyMEC, który obejmuje: serwer SyMEC, implementację wszystkich elementów architektury SyMEC, integrację systemu z sieciami dostępowymi WLAN, LAN/Ethernet, 4G i 5G, a także implementację i integrację przykładowych aplikacji MEC, nazwanych aplikacjami SyMEC. W prototypie wykorzystano oprogramowanie platformy MEC opracowane w ramach projektu OpenNESS [8] i udostępnione na licencji otwartego oprogramowania. Wykorzystanie tego oprogramowania pozwoliło na zachowanie pełnej zgodności styku API aplikacji MEC oraz usług platformy ze standardami ETSI MEC [5], co jest szczególnie ważne z punktu widzenia sukcesu przyszłych wdrożeń systemu SyMEC i możliwości oferowania w nim aplikacji MEC opracowanych przez niezależnych twórców. Ponadto opracowany prototyp został zintegrowany z produkcyjnym systemem OSS/BSS wykorzystywanym w sieci GEANT [8]. W ten sposób uzyskano w pełni funkcjonalny system MEC.

Wszystkie wytworzone moduły oprogramowania zostały przygotowane w postaci kontenerów Docker, a komunikacja pomiędzy modułami jest realizowana przez interfejsy REST API. Umożliwia to uruchomienie systemu SyMEC w dowolnym środowisku obliczeniowym zarządzanym przez Kubernetes, a także umożliwia skalowanie wydajności elementów systemu. Oprogramowanie wytworzono zgodnie z podejściem Ciągłej Integracji i Ciągłego Wdrażania CI/CD (*Continuous Integration/Continuous Development*). W tym celu przygotowano środowisko automatycznej instalacji, uruchomienia i testowania bazujące na narzędziu GitLab [9] oraz lokalnej instancji klastra obliczeniowego k8s. Moduły oprogramowania po integracji i pozytywnych testach przeprowadzonych w zautomatyzowany sposób w lokalnym klastrze k8s były następnie automatycznie instalowane w środowisku ogólnopolskiej sieci badawczej PL-LAB 2020. Testy przeprowadzane w tej sieci pozwoliły na weryfikację poprawności działania prototypu, ocenę jego wydajności oraz współpracy z heterogenicznymi sieciami dostępowymi w warunkach zbliżonych do warunków pracy sieci operatora.

Szczególnym wyzwaniem w opracowaniu prototypu systemu SyMEC było:

1. Zbudowanie serwera SyMEC i jego integracja z platformami wirtualizacji Docker/k8s i KVM oraz akceleratorem GPU. Serwer SyMEC opracowano bazując na procesorze ARM64 serii CAVIUM/Marvell CN8370. Procesor z tej rodziny charakteryzuje się wysoką wydajnością i efektywnością energetyczną. Ponadto, w przeciwieństwie do typowych procesorów stosowanych w serwerach obliczeniowych, zapewnia on sprzętową obsługę interfejsów sieciowych (aż do 120 Gbit/s), modułów szyfrowania, a także wspiera wirtualizację i narzędzia DPDK, umożliwiając równoległą obsługę potoków przetwarzania wykorzystując wiele rdzeni procesora. Serwer ten wyposażono w akcelerator graficzny GPU NVIDIA A5000, który zwiększa efektywność działania aplikacji MEC dzięki sprzętowej realizacji analizy i kodowania obrazów wideo, obliczeń równoległych i algorytmów uczenia maszynowego wspieranych 8 tysiącami rdzeni CUDA i 256

rdzeniami Tensor [2]. Ważnymi osiągnięciami przy opracowaniu serwera SyMEC było zaprojektowanie i wykonanie wielowarstwowej płyty głównej serwera, opracowanie sterowników urządzeń, m.in. interfejsów sieciowych, karty GPU, a także integracja platform witalizacyjnych Docker/k8s i KVM oraz przełącznika programowego Open vSwitch. Niestety, pomimo deklaracji dostawców oprogramowania np. NVIDIA, Docker, k8s, czy też Centos o pełnym wsparciu architektury ARM64, uruchomienie sterowników oraz narzędzi wirtualizacyjnych wymagało modyfikacji/korekt kodu źródłowego oraz specyficznej konfiguracji jądra systemu operacyjnego. Na serwerze zainstalowano system operacyjny Centos 7 z jądrem w wersji 4.14.76.

2. Integracja systemu SyMEC z platformą OpenNESS. System SyMEC w pełni zintegrowano z platformą MEC opracowaną w ramach projektu OpenNESS (*Open Network Edge Services Software*) firmy Intel [8]. Pomimo, iż platforma OpenNESS, podobnie do systemu SyMEC, została zaimplementowana w postaci kontenerów Docker, integracja tych systemów wymagała znaczącego nakładu pracy. Główną barierą była konieczność migracji modułów OpenNESSa z platformy Intel x86, dla której zostały opracowane, na platformę ARM64 właściwą dla serwera SyMEC. Dodatkowo istotnym problemem są częste zmiany architektury platformy OpenNESS wprowadzane praktycznie z każdym kolejnym jej wydaniem i brak zachowania wstecznej kompatybilności. Obecnie z systemem SyMEC zintegrowano platformę OpenNESS w wersji 21.09. Aktualizacja platformy OpenNESS do nowej wersji, nazywanej Intel® Smart Edge Open, wymaga przedefiniowania opracowanych w projekcie modułów współpracy z sieciami dostępowymi, które ściśle współpracują z platformą MEC.

3. Implementacja systemu sterowania i zarządzania aplikacjami MEC wraz z repozytorium MEC. System sterowania zaprojektowano wzorując się na architekturze ETSI MEC, rozszerzając ją o repozytorium aplikacji MEC. Moduł ten nie został explicite zdefiniowany w standardach ETSI, gdyż jego funkcje mogą pełnić zewnętrzne repozytoria. Natomiast, z punktu widzenia wymagań operatora, jest istotnym utrzymanie własnego, zaufanego repozytorium, w którym będą przechowywane obrazy aplikacji MEC, zweryfikowane w trakcie procesu rejestracji aplikacji MEC pod względem zgodności z infrastrukturą, bezpieczeństwa itp.. Ponadto istotnym wyzwaniem w projektowaniu modułu zarządcy platformy MEC jest konieczność wspierania wielu heterogenicznych zarządców wirtualnej infrastruktury VIM. W przypadku dużych wdrożeń należy oczekiwać, iż system SyMEC będzie zarządzał wieloma regionalnymi lub lokalnymi klastrami obliczeniowym, z których każdy może być zrealizowany w oparciu o inną platformę wirtualizacji, np. Kubernetes vs. OpenStack, bądź też w oparciu o inną wersję danej platformy. Dodatkowo, niezbędnym jest zapewnienie niezawodnej komunikacji pomiędzy instancjami aplikacji uruchomionymi w różnych klastrach. Ze względu na znaczące różnice w zastosowanej technice wirtualizacji nie jest możliwe utrzymanie jednolitego interfejsu pomiędzy zarządcą platformy MEC a zarządcą wirtualnej infrastruktury VIM.

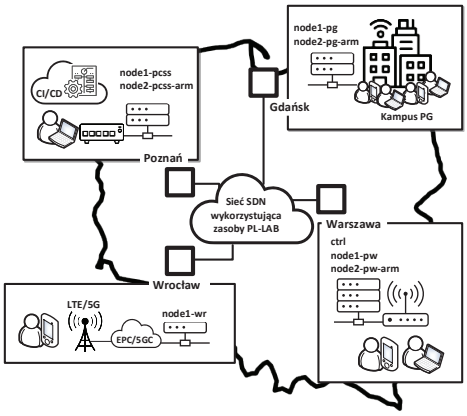
4. Integracja systemu SyMEC z systemem OSS/BSS operatora. System SyMEC zintegrowano z przykładowym systemem OSS/BSS, aby przetestować możliwość świadczenia usług użytkownikom oraz usługodawcom. Należy podkreślić, że użytkownik postrzega usługi i aplikacje MEC jako nowy zbiór usług udostępnionych dla niego w portalu zarządzania usługami CFS (*Customer Facing Service*). Zatem to system OSS/BSS jest stroną inicjującą akcje w systemie SyMEC w imieniu użytkownika lub dostawcy usług. Z tego względu istotne jest zachowanie pełnej zgodności styku mm1, zdefiniowanego pomiędzy systemem OSS/BSS a orkiestratorem, ze standardami. Ponadto ważnym jest również przedstawienie i udostępnienie zasobów oferowanych przez system SyMEC w sposób zrozumiały dla pozostałych podsystemów złożonego ekosystemu infrastruktury operatora.

5. Implementacja i integracja aplikacji MEC. W ramach prototypu przygotowano zbiór przykładowych aplikacji MEC, które umożliwiają przeprowadzenie testów oraz pokazów działania systemu SyMEC. Przeanalizowano ponad 60 aplikacji dostępnych na licencji otwartego oprogramowania i wybrano zbiór aplikacji o zróżnicowanej funkcjonalności, które udostępniono w systemie. Aplikacje te obejmują: gry interaktywne, adaptacyjne strumieniowanie obrazów wideo, analizę i transkodowanie obrazów wideo w czasie zbliżonym do rzeczywistego, zapewnienie bezpieczeństwa (indywidualny firewall), czy też analizy stanu sieci dostępowych. Udostępnione aplikacje wykorzystują akcelerator GPU dla zwiększenia wydajności analizy i kodowania obrazów wideo oraz algorytmów sztucznej inteligencji. Należy zwrócić uwagę, że aplikacje te nie były projektowane dla systemów typu MEC, dlatego jedynie w ograniczonym stopniu wykorzystują funkcje oferowane przez platformę MEC.

4. WDROŻENIE SYSTEMU W KRAJOWEJ SIECI BADAWCZEJ PL-LAB 2020

Opracowany prototyp systemu SyMEC został wdrożony w ogólnopolskiej sieci badawczej PL-LAB 2020, co umożliwiło zintegrowanie węzłów MEC z różnymi sieciami dostępowymi i przeprowadzenie testów w warunkach zbliżonych do warunków panujących w sieciach operacyjnych. Prototyp zainstalowano w czterech lokalizacjach, tj.: w Gdańsku (Politechnika Gdańska), Poznaniu (Poznańskie Centrum Superkomputerowo-Sieciowe), Wrocławiu (Instytut Łączności - Państwowy - Instytut Badawczy) oraz w Warszawie (Politechnika Warszawska), tak jak to przedstawiono na Rys. 2.

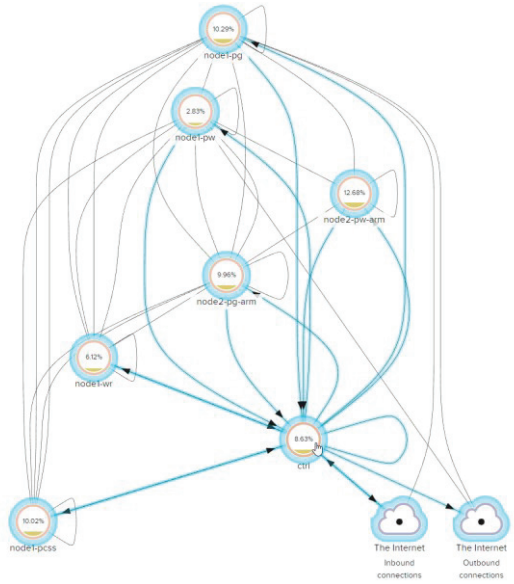
W każdej lokalizacji zostały zainstalowane dwa serwery MEC. Jeden z nich to opracowany w projekcie serwer MEC wyposażony w procesor ARM64, a drugi serwer MEC zbudowano wykorzystując typowy serwer z procesorem x86. Heterogeniczne środowisko klastra obliczeniowego pozwoliło na instalację, zbadanie działania oraz możliwości współpracy aplikacji MEC zaprojektowanych dla systemów ARM64 oraz x86. W szczególności, węzeł SyMEC w Gdańsku zintegrowano z rozległą siecią WLAN zainstalowaną w ramach kampusu Politechniki Gdańskiej. Węzeł SyMEC we Wrocławiu został zintegrowany z eksperymentalnymi sieciami LTE (4G) oraz 5G zbudowanymi z wykorzystaniem oprogramowania



Rys. 2. Wdrożenie systemu SyMEC w ogólnopolskiej sieci badawczej PL-LAB 2020.

o otwartym kodzie [11]. W tym przypadku, ze względu na ograniczenia dotyczące licencjonowanego pasma częstotliwości, stacje bazowe oraz terminale końcowe zostały zainstalowane w komorze bezechowej. Ponadto w węzłach zlokalizowanych w Warszawie i Poznaniu serwery SyMEC zintegrowano z sieciami lokalnymi LAN/WLAN, udostępniając aplikacje MEC i zasoby systemu SyMEC użytkownikom tych sieci do badań dotyczących techniki MEC.

Przykładowy obraz stanu klastra obliczeniowego SyMEC, uwzględniający utworzone połączenia pomiędzy serwerami i stan ich obciążenia, przedstawiono na Rys. 3.



Rys. 3. Przykładowy widok stanu systemu SyMEC.

Opracowana instalacja obejmuje pełną funkcjonalność opracowanego systemu SyMEC, co umożliwia prowadzenie badań dotyczących techniki MEC. Należy podkreślić, iż opracowany prototyp i instalacja pilotażowa obejmują pełną funkcjonalność opracowanego systemu SyMEC, integrują serwery MEC opracowane w ramach projektu (bazujące na procesorach arm64), jak również serwery bazujące na procesorach x86, umożliwiają wykorzystanie przez aplikacje MEC akceleracji sprzętowej i zrównoleglenia obliczeń w oparciu o procesory GPU.

5. PODSUMOWANIE

W artykule przedstawiono docelową architekturę opracowanego systemu obliczeń na brzegu sieci, nazywanego systemem SyMEC oraz omówiono rolę poszczególnych elementów tego systemu. Ponadto przedstawiono zaimplementowany prototyp, a także doświadczenia wynikające z jego implementacji i wdrożenia w krajowej sieci badawczej PL-LAB 2020. Wdrożona instalacja pozwala na prowadzenie badań dotyczących rozwoju techniki MEC, opracowania nowych metod i algorytmów sterowania, orkiestracji, zarządzania, a także nowych aplikacji MEC i ich zastosowań. Dalsze prace badawcze będą związane z opracowaniem zasad współpracy w wielodomenowych systemach MEC, nowych metod zarządzania zasobami oraz nowych zastosowań w obszarze usług wirtualnej rozszerzonej rzeczywistości, Przemysłu 4.0, czy zastosowań multimedialnych.

Badania, których wyniki są przedstawione w tym artykule, były prowadzone w ramach projektu POIR 04.01.02.00-00-0107/17, pt. „System MEC dla wspierania zaawansowanych aplikacji w środowisku sieci przewodowych i bezprzewodowych 3G/4G/5G", współfinansowanego w ramach Programu Operacyjnego Innowacyjny Rozwój na lata 2014-2020.

LITERATURA

[1] ETSI standard, Multi-access Edge Computing (MEC); Framework and Reference Architecture, ETSI GS MEC 003 V3.1.1, marzec 2022.
[2] M. Sosnowski, A. Bęben, W. Jóźwiak, M. Miszewski, „Serwer i platforma MEC", KSTiT 2022, 7-9 września 2022.
[3] B. Krakowiak, B. Belter, A. Bęben, W.Burakowski, M. Sosnowski, W. Jóźwiak, P. Olender, „System sterowania i zarządzania SyMEC", KSTiT 2022, 7-9 września 2022.
[4] TeleManagement Forum: https://www.tmforum.org (dostęp: 30.05.2022)
[5] ETSI standard, Multi-access Edge Computing (MEC); General principles, patterns and common aspects of MEC Service APIs, ETSI GS MEC 009 V3.1.1, czerwiec 2021.
[6] Docker: https://www.docker.com/ (dostęp: 30.05.2022)
[7] Kubernetess "Production-Grade Container Orchestration", https://kubernetes.io, (dostęp: 7.06.2022)
[8] OpenNESS, Open Network Edge Services Software, Intel, https://www.openness.org/ (dostęp 6.06.2022)
[9] Deliverable D6.2 - Automation and Orchestration of Services in the GÉANT Community, Projekt GEANT, 30.09.2019,
[10] GitLab, The One DevOps Platform, https://about.gitlab.com/ (dostęp: 8.06.2022)
[11] Open Air Interfeces, https://openairinterface.org/oai-5g-ran-project/ (dostęp: 6.06.2022)

SERWER I PLATFORMA MEC
MEC SERVER AND MEC PLATFORM

Maciej Sosnowski[1]; Andrzej Bęben[2]; Witold Jóźwiak[2], Mariusz Miszewski[3]

[1] Instytut Łączności - PIB, Politechnika Warszawska, Warszawa, m.sosnowski@tele.pw.edu.pl
[2] Politechnika Warszawska, Warszawa, {abeben,w.jozwiak}@tele.pw.edu.pl
[3] DGT Sp. z o.o, Gdańsk, mariusz.miszewski@dgt.pl

DOI: 10.15199/59.2022.4.49

Streszczenie**: W artykule omówiono serwer MEC oraz Platformę MEC, które zostały opracowane w projekcie SyMEC. Na serwerze są uruchamiane aplikacje MEC, z których korzystają użytkownicy systemu. Aplikacje te są wspomagane i konfigurowane przez funkcje Platformy MEC. Przedstawione zostały najważniejsze charakterystyki serwera MEC oraz funkcje realizowane przez Moduły Platformy MEC.**

Abstract**: The article describes the MEC server and the MEC Platform that were developed by the SyMEC project. The end-users applications run on the MEC server. The applications may utilize functions provided by the MEC Platform. The most important MEC server characteristics and the functions provided by the MEC Platform Modules were shortly summarized.**

Słowa kluczowe**: OpenNESS, Platforma MEC, Serwer MEC, Projekt SyMEC**

Keywords**: MEC Platform, MEC Server, OpenNESS, SyMEC Project**

1. WSTĘP

Celem projektu SyMEC było zaprojektowanie i implementacja systemu obliczeń na brzegu sieci MEC (ang. *Multi-access Edge Computing*), przeznaczonego do wdrożenia w infrastrukturze operatora telekomunikacyjnego. System ten pozwala na uruchomienie aplikacji blisko użytkownika i wykorzystanie zalet wynikających m.in. z możliwości uzyskania małego opóźnienia w sieci czy też większej dostępnej przepływności bitowej. Często aplikacje te mają znaczenie lokalne, np. aplikacje, które umożliwiają obejrzenie powtórki akcji z meczu na stadionie czy wybór widoku z danej kamery na koncercie. Jednocześnie zmniejsza się obciążenie sieci operatora, gdyż duże ilości danych (np. strumieni wideo) mogą być przetwarzane na brzegu sieci, bez potrzeby przesyłania ich do centrum danych.

Technika przetwarzania na brzegu sieci wymaga umiejscowienia w sieci dostępowej, lub w jej bezpośrednim sąsiedztwie, serwerów, na których będą uruchamiane aplikacje dla użytkowników. Należy oczekiwać, że na brzegu sieci, dla jednej sieci dostępowej, przeznaczone będą pojedyncze sztuki serwerów MEC. Typowym miejscem, w którym instaluje się serwer MEC jest szafa teletechniczna przy stacji bazowej. Z tego powodu serwer MEC powinien charakteryzować się niskim poborem prądu oraz wydajnym systemem chłodzenia. Większość serwerów obliczeniowych jest zbudowana na bazie procesorów o architekturze x86 (Intel), a ich docelowym miejscem instalacji jest klimatyzowana serwerownia. Tym samym było konieczne opracowanie serwera, który byłby w stanie pracować wydajnie poza centrum danych.

Jednym z elementów architektury Systemu MEC jest tzw. Platforma MEC. Jej celem jest dostarczanie usług, z których mogą korzystać aplikacje użytkowników. Usługi te zwiększają funkcjonalność aplikacji. Przykładowo, Platforma MEC może agregować informacje o użytkownikach sieci (ich lokalizację czy dostępną przepływność bitową) i udostępniać je aplikacjom, które na podstawie tych danych mogą modyfikować swoje działanie. W skład Platformy MEC wchodzą 1) Zarządca Platformy MEC (ang. *MEC Platform Manager*, MEPM) oraz 2) Moduły Platformy MEC. Na podstawie przeprowadzonej analizy stanu sztuki w dziedzinie techniki MEC przyjęto, że do implementacji Platformy MEC w Systemie SyMEC wykorzystane zostaną Moduły Platformy MEC wypracowane w ramach projektu *OpenNESS* [1], które są publicznie dostępne w repozytorium *github* [2] jako otwarte oprogramowanie.

Organizacja artykułu jest następująca. W rozdziale drugim przedstawiono opis opracowanego serwera MEC, w rozdziale trzecim – Platformę MEC oraz moduły wchodzące w jej skład. Rozdział czwarty podsumowuje artykuł.

2. SERWER MEC

W projekcie SyMEC zbudowano i zintegrowano serwer MEC. Serwer ten jest wyposażony w interfejsy i moduły komunikacyjne wymagane w Systemie MEC. Serwer MEC ma kompaktowe gabaryty, możliwość zastosowania różnych interfejsów i modułów komunikacyjnych oraz wykorzystuje najnowsze osiągnięcia techniki ARM. Technika ta cechuje się wysokim stosunkiem mocy obliczeniowej procesora do zużywanej energii [3], tym samym nie jest konieczne umiejscowienie go w klimatyzowanym pomieszczeniu. Serwer taki posiada również sprzętowe wsparcie dla technik wirtualizacji. Jednym z głównych zastosowań serwera MEC są aplikacje związane z przetwarzaniem obrazu, a zatem serwer MEC został wyposażony w kartę graficzną, którą można udostępnić tym aplikacjom.

Najbardziej istotne charakterystyki serwera MEC są następujące:

Procesor: CAVIUM/Marvell CN8370-1800BG1676-SCP-G, moc obliczeniowa 24 Cores@1,8GHz, konfiguracja z CPU + GPU, architektura ARM64 v8.1, RAM 128GB.

Interfejsy sieciowe: 2x 40GbETH (QSFP+), 4x 10GbETH (SFP+), 120Gbps całkowitej przepustowości.

Interfejsy zarządzania: RJ-11, RS232 (UART0) – lokalne zarządzanie, RJ-45, Fast ETH – zdalne zarządzanie.

Interfejsy pozostałe: 2x USB 3.0 Host Mode , 2x SATA 3.0 do podłączenia HDD, PCIe 3.0 x8 do podłączenia karty graficznej (GPU) lub kart rozszerzeń, TWSI0 (I2C) do komunikacji i sterowania, pamięci DDR4.

Podzespoły zainstalowane: 2xHDD 8TB Toshiba MG06ACA800E, GPU NVIDIA RTX A5000 (8192 CUDA Cores), RAM 2x32GB DDR4 MTA36ASF4G72PZ-2G6E.

Na serwerze został zainstalowany system operacyjny Centos, a jądro zostało dostosowane do wymagań narzędzi wirtualizacyjnych – KVM, docker, Kubernetes, oraz Open vSwitch. Został zainstalowany także zestaw narzędzi dla wirtualizacji karty graficznej (*Nvidia Container Toolkit* [4]), który umożliwia przekazanie dostępu do karty do aplikacji uruchomionych w postaci kontenerów zarządzanych przez Kubernetes, a sama karta widoczna jest jako zasób klastra.

Zdjęcia serwera MEC są zamieszczone na Rys 1.

Rys. 1. Serwer MEC

3. PLATFORMA MEC

Rolą Platformy MEC jest dostarczenie usług, pozwalających na zwiększenie funkcjonalności aplikacji. Platforma programowa MEC to specjalizowane oprogramowanie uruchomione wewnątrz serwera MEC, które jest odpowiedzialne, m.in. za:

1) Utrzymanie środowiska, w którym aplikacje MEC mogą rejestrować, wyszukiwać, rozgłaszać i świadczyć usługi na rzecz innych aplikacji MEC, włączając w to usługi MEC dostępne na innych serwerach MEC. Przykładem takich usług są usługi udostępnienia informacji o aktualnym stanie sieci, lokalizacji użytkownika, synchronizacji zegarów, dystrybucji wiadomości w trybie „publikuj/subskrybuj", czy też informacji kontekstowych dotyczących obiektów lub usług zlokalizowanych w otoczeniu serwera MEC. Ponadto aplikacje MEC obejmują również komponenty świadczące usługi na rzecz innych aplikacji, a nie wyłącznie usługi świadczone na rzecz użytkowników;

2) Wyznaczenie zasad przekazu ruchu pomiędzy wewnętrznymi i zewnętrznymi obiektami wirtualnymi, aplikacjami MEC, usługami MEC i otoczeniem.

3) Utrzymanie rejestru usług i ich adresów w ramach lokalnego serwera nazw DNS (ang. *Domain Name System*) oraz konfigurację utworzonych obiektów wirtualnych. Za konfigurację modułów Platformy MEC odpowiada Zarządca Platformy MEC [5].

Poniżej omówione zostały poszczególne moduły Platformy MEC.

3.1. Edge Application Agent (EAA)

EAA [6] to usługa, która umożliwia wymianę komunikatów między aplikacjami zgodnie z modelem *publish-subscribe*. Informacje są produkowane przez aplikacje, które nazywane są Producentami (ang. *Producers*), a wykorzystywane przez aplikacje nazywane Konsumentami (ang. *Consumers*). Konsumenci to aplikacje obsługujące żądania użytkowników, które w procesie obsługi tych żądań korzystają z informacji przesyłanych przez Producentów. Producenci nie są powiązani z żadną konkretną aplikacją użytkownika – informacje, które dostarczają do Platformy MEC, mogą być wykorzystywane przez wielu Konsumentów. Jednocześnie dany Konsument może korzystać z informacji dostarczanych przez wielu Producentów.

W Projekcie zostały zaimplementowane dwa moduły współpracy z sieciami 1) WLAN oraz 2) 4G/5G [7], które pełnią rolę Producentów informacji o stanie tych sieci oraz o stanie użytkowników do nich dołączonych. Informacje te mogą być wykorzystane np. przez aplikacje strumieniujące wideo, aby adaptować jakość przekazu do dostępnej przepływności czy też np. do wyświetlenia odpowiedniej reklamy kontekstowej dopasowanej do lokalizacji użytkownika.

3.2. Certsigner / Certrequester

Moduły *Certsigner / Certrequester* [8] realizują usługę generowania certyfikatów TLS dla aplikacji, które takiego certyfikatu wymagają. Usługa ta pozwala usunąć proces generowania certyfikatu z aplikacji. Usługa zapewnia, że aplikacja posiada już taki certyfikat w momencie uruchomienia. Ponieważ aplikacja nie jest już odpowiedzialna za wygenerowanie swojego certyfikatu, to nie musi zawierać w swoim obrazie narzędzi do tego potrzebnych (np. *openssl, easyrsa*) ani związanej z tym logiki. Skutkuje to zmniejszeniem rozmiaru obrazu aplikacji oraz uproszczeniem jej działania. Dodatkowo jeśli przyjmiemy, że aplikacja odpowiadałaby za podpisanie swojego certyfikatu kluczem głównym, wspólnym dla całego Systemu, to

należałoby do aplikacji przekazać ten klucz, co byłoby niepożądane ze względów bezpieczeństwa.

Biorąc powyższe pod uwagę, przyjęto rozwiązanie, że w początkowej fazie uruchamiania *poda* z aplikacją uruchamiany jest *init-container* [9] o nazwie *Certrequester*, który generuje parę kluczy (publiczny / prywatny) i żądanie CSR (ang. *Certificate Signing Request*), które następnie zostaje przekazane do centralnego *Certsigner'a*. Po zatwierdzeniu żądania przez administratora klastra, moduł *Certsigner* generuje certyfikat i zwraca go. Certyfikat jest zapisywany w folderze współdzielonym przez *Certrequester* oraz docelową aplikację. Następnie *init-container Certrequester* kończy działanie i jest uruchamiany kontener z docelową aplikacją. Tym samym właściwa aplikacja zostanie uruchomiona dopiero w momencie, gdy certyfikat dla niej jest już gotowy.

Aplikacja sama mogłaby wygenerować i wysłać żądanie CSR do *Certsigner'a*, jednak zastosowanie opisanego wyżej mechanizmu wynosi fazę generowania kluczy, żądania CSR oraz komunikacji z *Certsigner'em* do innego kontenera, który odpowiada jedynie za to zadanie. Faza generowania certyfikatu jest *de facto* powtarzalna dla wszystkich aplikacji, które takiego certyfikatu wymagają, a zastosowanie takiego generycznego mechanizmu jest możliwe po prostej modyfkacji w definicji opisu danej aplikacji oraz przygotowaniu pliku z informacjami, które mają zostać zawarte w CSR (np. nazwa domenowa aplikacji).

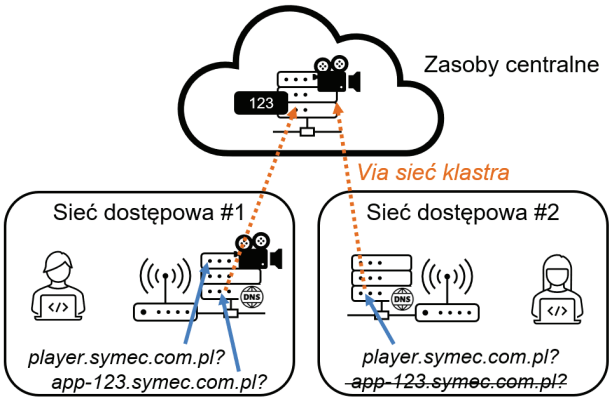
3.3. EdgeDNS

Usługa *EdgeDNS* [10] odpowiada za odwzorowanie nazw domenowych aplikacji MEC na adresy IP, pod którymi są one dostępne dla użytkowników danej sieci dostępowej. Usługa EdgeDNS jest uruchomiona jako kilka niezależnych instancji DNS, działających na wybranych serwerach MEC, najczęściej na jednym serwerze MEC per sieć dostępowa. Użytkownicy danej sieci będą w stanie odwzorować nazwę domenową danej instancji aplikacji, np. odtwarzacza wideo *player.symec.com.pl*, na adres IP serwera MEC, najczęściej na adres z puli prywatnej używanej w tej sieci. Gdy użytkownik dołączy się do innej sieci dostępowej, aplikacja może być nadal dostępna pod tą samą nazwą domenową, ale za pośrednictwem innego serwera MEC, pod innym adresem IP. Aplikacja może być uruchomiona bezpośrednio na serwerze, na który kieruje EdgeDNS, lub na serwerze w innej lokalizacji (np. w zasobach centralnych) – wówczas serwer ten pełni jedynie rolę bramy do sieci klastra.

Aby skorzystać z mechanizmu EdgeDNS, w punkcie dostępowym do danej sieci należy jako jeden z adresów serwerów DNS podać adres IP EdgeDNS'a, który ma być w tej sieci używany. W chwili uruchomienia aplikacji o danej nazwie domenowej, Zarządca Platformy MEC dodaje odpowiednie wpisy do poszczególnych EdgeDNS'ów. W szczególności konkretna nazwa domenowa, rozwiązywana w różnych sieciach dostępowych, może kierować do innych instancji aplikacji lub nie być powiązana z żadnym adresem IP (tj. aplikacja z tej sieci jest nieosiągalna).

Przykładowo na Rys. 2. przedstawiono dwie sieci dostępowe, do których dołączono serwery MEC oraz uruchomiono instancje EdgeDNS'ów. W sieci dostępowej #1 adres *player.symec.com.pl* kieruje na lokalny serwer, na którym dostępna jest ta aplikacja, natomiast w sieci #2 na lokalny serwer, pełniący funkcję bramy do sieci klastra, a aplikacja uruchomiona jest w zasobach centralnych. Z kolei dostęp do aplikacji „123", która została uruchomiona w zasobach centralnych, możliwy jest przez nazwę *app-123.symec.com.pl*, ale tylko z sieci #1.

Zapewnienie komunikacji po nazwach domenowych jest szczególnie istotne w przypadku aplikacji webowych, gdyż umożliwia wykorzystanie *Ingressu*, tj. usługi nasłuchującej, np. na jednym adresie IP i dwóch portach HTTP/HTTPS (standardowo TCP:80/443), która przekierowuje żądania do odpowiedniej instancji serwera http (np. *apache2, nginx*), właśnie na podstawie nazwy domenowej zawartej w żądaniu. Integracja usługi EdgeDNS z siecią dostępową zapewnia, że (dynamicznie tworzone) aplikacje będą osiągalne przez tych użytkowników, dla których są one przeznaczone.



Rys. 2. Odwzorowanie nazw domenowych na adresy IP w dwóch sieciach dostępowych

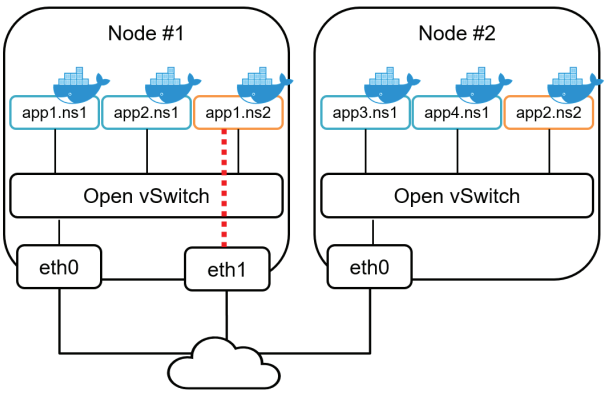
3.4. Interface Service

Usługa *Interface Service* [11] pozwala tworzyć nowe sieci, do których dołączane są aplikacje MEC oraz fizyczne interfejsy sieciowe serwera. Standardowo, komunikacja między aplikacjami rozmieszczonymi na różnych serwerach jest realizowana przez tzw. *sieć klastra*. Sieć ta jest utworzona w oparciu o SDN i Open vSwitch (OvS), którego instancje są uruchomione na wszystkich serwerach tworzących klaster. Sieć klastra jest siecią wewnętrzną, do której nie można dołączyć się za pośrednictwem żadnego interfejsu fizycznego. Rozciąga się ona między poszczególnymi serwerami dzięki zastosowaniu tuneli GRE, które są konfigurowane przez centralny sterownik SDN.

Dostęp do aplikacji z sieci zewnętrznej (czyli spoza klastra, np. dla użytkowników) jest możliwy np. poprzez zastosowanie mechanizmu *NodePort*, tj. udostępnieniu aplikacji na konkretnym porcie na wybranych adresach IP wybranych serwerów. Ruch odbierany na takim porcie, poprzez reguły *iptables*, jest przekierowywany na odpowiedni adres i port w sieci klastra. W niektórych zastosowaniach jest to jednak rozwiązanie niepraktyczne np. w przypadku , kiedy aplikacja korzysta z dużej liczby portów. Wówczas w ramach sieci SDN można utworzyć nową sieć i dołączyć do niej wybrane aplikacje oraz fizyczne interfejsy danego serwera. Takie rozwiązanie

pozwala na zrealizowanie dostępu do tych aplikacji bezpośrednio przez interfejs fizyczny bez konieczności udostępniania ich na zewnątrz innymi mechanizmami (np. *NodePort*).

Przykładowo na Rys. 3. przedstawiono sieć klastra złożonego z dwóch serwerów o dwóch przestrzeniach nazw (*ns1* i *ns2*), gdzie widoczność między aplikacjami zapewniona jest tylko w ramach danej przestrzeni nazw. Ruch sieciowy między serwerami jest przesyłany interfejsami *eth0*, natomiast interfejs *eth1* serwera #1 został bezpośrednio dołączony do sieci współdzielonej z *app1.ns2*.



Rys. 3. Sieć SDN - widoczność w ramach przestrzeni nazw ns1 / ns2, interfejs eth1 dołączony do sieci klastra, dla aplikacji app1.ns2

3.5. Network Feature Discovery (NFD)

Usługa *NFD* [12] zbiera szeroki zakres informacji o poszczególnych serwerach w Systemie SyMEC oraz o funkcjach przez nie wspieranych. Lista rozpoznawanych funkcji została przedstawiona w [13] i obejmuje np. nazwę serwera (*hostname*), rodzaj i wersję uruchomionego systemu operacyjnego, wersję używanego jądra, typ dysku czy też informacje o funkcjach wspieranych przez hardware, w szczególności procesor czy karty sieciowe.

Informacje te są zapisywane w bazie danych Kubernetesa w postaci etykiet (ang. *labels*) opisujących dany serwer. Etykiety posiadają swoją nazwę i wartość, np. etykieta określająca architekturę serwera może mieć postać *kubernetes.io/arch=arm64*. W definicji aplikacji można określić, jakimi etykietami musi być oznaczony serwer, aby Kubernetes rozważał go jako miejsce uruchomiania aplikacji. Przykładowo, niektóre aplikacje mogą być uruchamiane wyłącznie na serwerach z architekturą ARM64, a inne tylko na X86.

W szczególności, można jednoznacznie wskazać konkretny serwer poprzez etykietę określającą nazwę serwera i w ten sposób jest to zrealizowane w Systemie SyMEC. Wybór serwera, na którym uruchomiona będzie aplikacja, dokonywany jest przez orkiestrator, gdyż w tym wyborze -zazwyczaj- kluczowa jest odległość do użytkownika, dla którego uruchamiana jest aplikacja. Potencjalnie jednak, w przypadku dostępności kilku serwerów MEC w jednej sieci dostępowej, orkiestrator może korzystać z informacji zebranych przez NFD.

3.6. Telemetria

W ramach Platformy MEC działa zestaw narzędzi umożliwiających monitorowanie stanu urządzeń fizycznych oraz poszczególnych aplikacji, tak jak na Rys. 4. Na każdym serwerze działają instancje aplikacji *collectd* oraz *cAdvisor*, które monitorują wykorzystanie zasobów, odpowiednio, na serwerze fizycznym oraz przez poszczególne aplikacje (kontenery). Zbierane informacje zapisywane są do bazy danych szeregów czasowych *Prometheus* oraz wizualizowane w *Grafanie*.

Rys. 4. Zestaw narzędzi do telemetrii

4. PODSUMOWANIE

W artykule opisano serwer MEC oraz Platformę MEC, które tworzą środowisko uruchomieniowe dla aplikacji użytkowników w Systemie SyMEC. Zaprezentowano najważniejsze charakterystyki serwera MEC, wykorzystany hardware oraz oprogramowanie, a także funkcjonalność podstawowych Modułów Platformy MEC opracowanej i udostępnionej przez projekt OpenNESS. Moduły te wspomagają działanie aplikacji użytkowników.

Badania, których wyniki są przedstawione w tym artykule, były prowadzone w ramach projektu POIR 04.01.02.00-00-0107/17,pt. „System MEC dla wspierania zaawansowanych aplikacji w środowisku sieci przewodowych i bezprzewodowych 3G/4G/5G”, współfinansowanego w ramach Programu Operacyjnego Innowacyjny Rozwój na lata 2014-2020.

LITERATURA

[1] Strona domowa projektu OpenNESS - Open Network Edge Services Software: https://www.openness.org/ (dostępna 06.06.2022)

[2] Repozytorium kodu projektu OpenNESS (aktualnie przemianowane do Smart Edge Open): https://github.com/smart-edge-open (dostępne 06.06.2022)

[3] Zhonghong Ou i inni. 2012. „Energy- and Cost-Efficiency Analysis of ARM-Based Clusters”. *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID '12)*. IEEE Computer Society, USA, 115–123.

[4] Dokumentacja oprogramowania NVIDIA Container Toolkit: https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/overview.html (dostępna 06.06.2022)

[5] Krakowiak Bartosz i inni. 2022. „System sterowania i zarządzania SyMEC”. *KSTIT 2022*.

[6] Dokumentacja modułu Edge Application Agent: https://github.com/smart-edge-

open/specs/blob/master/doc/applications-onboard/smartedge-open-eaa.md (dostępna 06.06.2022)
[7] Hoeft Michał i inni. 2020. „Moduł platformy SyMEC odpowiedzialny za współpracę serwera MEC z sieciami komórkowymi i bezprzewodowymi sieciami dostępowymi WLAN". *Przegląd Telekomunikacyjny+Wiadomości Telekomunikacyjne*, 7-8: 149-154.
[8] Dokumentacja modułów Certsigner / Certrequester: https://github.com/smart-edge-open/specs/blob/master/doc/applications-onboard/smartedge-open-certsigner.md (dostępna 06.06.2022)
[9] Dokumentacja init-containers: https://kubernetes.io/docs/concepts/workloads/pods/init-containers/ (dostępna 06.06.2022)
[10] Dokumentacja modułu EdgeDNS: https://github.com/smart-edge-open/specs/blob/master/doc/applications-onboard/smartedge-open-edgedns.md (dostępna 06.06.2022)
[11] Dokumentacja modułu Interface Service: https://github.com/smart-edge-open/specs/blob/master/doc/applications-onboard/smartedge-open-interface-service.md (dostępna 06.06.2022)
[12] Dokumentacja modułu NFD: https://smart-edge-open.github.io/docs/components/resource-management/node-feature-discovery/ (dostępna 06.06.2022)
[13] Spis etykiet dla NFD: https://kubernetes-sigs.github.io/node-feature-discovery/v0.11/get-started/features.html (dostępny 06.06.2022)

SYSTEM STEROWANIA I ZARZĄDZANIA SYMEC

SYMEC CONTROL AND MANAGEMENT SYSTEM

Bartosz Krakowiak[1]; Bartosz Belter[1]; Andrzej Bęben[2]; Wojciech Burakowski[2]; Witold Jóźwiak[2]; Maciej Sosnowski[2,3]; Paweł Olender[3]

[1] Poznańskie Centrum Superkomputerowo-Sieciowe, Poznań, bkrakowiak@man.poznan.pl
[2] Politechnika Warszawska, Warszawa, {abeben, m.sosnowski, w.jozwiak}@tele.pw.edu.pl
[3] Instytut Łączności – PIB, Warszawa, p.olender@il-pib.pl

DOI: 10.15199/59.2022.4.50

Streszczenie: **Artykuł przedstawia główne elementy systemu sterowania i zarządzania systemu SyMEC. Elementy te realizują proces zarządzania cyklem życia aplikacji MEC, który obejmuje między innymi dodanie/usunięcie aplikacji do/z systemu, przydzielenie zasobów i uruchomienie/zatrzymanie instancji aplikacji w węzłach MEC. Ponadto przedstawiono opracowany system zarządzania i monitorowania wykorzystania zasobów infrastruktury MEC. Przedstawione elementy zostały zaimplementowane i zintegrowane w pełni funkcjonalnym prototypie systemu SyMEC.**

Abstract: **The paper presents main elements of the SyMEC control and management system. It is responsible for MEC application lifecycle management which include, among others, creation of application package, resource allocation and instantiation. The management part of the system provides monitoring and management features for edge infrastructure resources. Presented solution has proven by development of the fully featured prototype of the SyMEC system.**

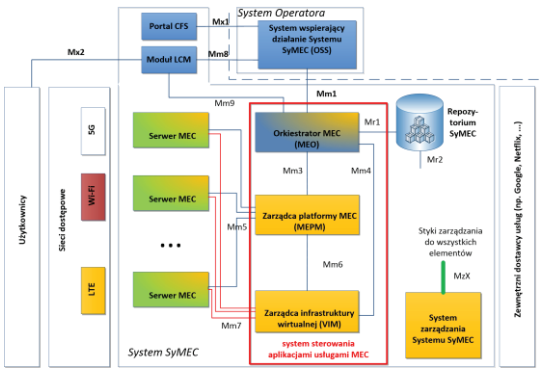
Słowa kluczowe: **system SyMEC, orkiestrator MEC, system zarządzania systemem SyMEC**

Keywords: **SyMEC system, MEC orchestration, SyMEC control and management system**

1. WSTĘP

Technika przetwarzania danych na brzegu sieci MEC (*Multi-Access Edge Computing)* oferuje nowy sposób świadczenia usług polegający na udostępnieniu serwerów MEC na brzegu sieci, możliwie najbliżej użytkowników. Rozwiązanie to rozszerza infrastrukturę operatora telekomunikacyjnego o węzły obliczeniowe umieszczone w części dostępowej sieci (na brzegu sieci), co umożliwia analizę, przechowywanie i udostępnianie danych w pobliżu ich źródła. Rozwiązanie MEC pozwala na oferowanie aplikacji wymagających znacznej mocy obliczeniowej i małych opóźnień przekazu danych np. aplikacji rozszerzonej/mieszanej rzeczywistości. Ponadto, technika MEC przyczynia się do zmniejszenia obciążenia sieci telekomunikacyjnych dzięki lokalnemu przetwarzaniu danych.

Zarządzanie cyklem życia aplikacji MEC jest jednym z głównych zadań opracowanego podsystemu systemu sterowania i zarządzania systemu SyMEC. Na podstawie przeprowadzonej analizy rozwiązań i standardów MEC, zaprojektowano system sterowania i zarządzania SyMEC, który składa się z: orkiestratora MEC, zarządcy platformy MEC, zarządcy infrastruktury wirtualnej oraz systemu zarządzania. System ten udostępnia aplikacje MEC na żądanie użytkowników. W procesie tym orkiestrator MEC realizuje funkcje związane z zarządzaniem cyklem życia aplikacji i usług MEC obejmujące m.in. rejestrowanie aplikacji, umieszczenie obrazu aplikacji w repozytorium MEC, tworzenie i usuwanie instancji aplikacji MEC na życzenie użytkowników. Ponadto, orkiestrator MEC odpowiada za przydzielenie aplikacji MEC wymaganych zasobów oraz efektywne wykorzystanie zasobów obliczeniowych serwerów MEC oraz zasobów sieci. Zarządca platformy MEC dostosowuje żądania orkiestratora do postaci wymaganej przez zarządców wirtualnej infrastruktury odpowiedzialnych za dany obszar systemu. Ponadto, koordynuje on działanie usług platformy MEC oferowanych w rozproszonej infrastrukturze serwerów MEC. Zarządca wirtualnej infrastruktury jest odpowiedzialny za uruchomienie/zatrzymanie instancji aplikacji MEC w wybranym przez orkiestrator serwerze MEC, zarządzanie wirtualnymi obiektami/usługami w serwerach MEC oraz zasobami wirtualnej infrastruktury SyMEC. System zarządzania SyMEC wspiera procesy zarządzania systemem SyMEC współpracując z systemami OSS/BSS (*Operations/ Business Support System*) operatora. Obejmuje to głównie proces zarządzania urządzeniami fizycznymi systemu SyMEC, a częściowo także proces zarządzania usługami MEC (w zakresie monitoringu). Architekturę systemu SyMEC przedstawiono na Rys. 1.



Rys. 1. Architektura systemu SyMEC.

W kolejnych rozdziałach szczegółowo opisano budowę i funkcjonalność poszczególnych modułów systemu sterowania i zarządzania SyMEC.

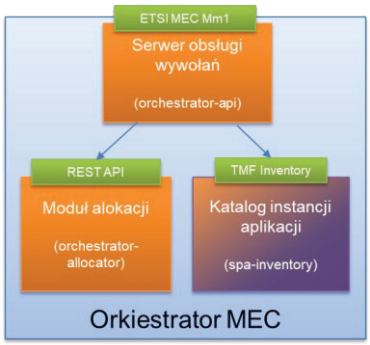
2. ORKIESTRATOR MEC

Orkiestrator MEC jest centralnym elementem architektury systemu SyMEC, który koordynuje komunikację z zewnętrznymi elementami systemu SyMEC, tj. systemem OSS/BSS operatora oraz odpowiada za poprawną realizację funkcji związanych z zarządzaniem cyklem życia aplikacji i usług MEC. Obejmuje to m.in. rejestrowanie aplikacji, umieszczenie obrazu aplikacji w repozytorium MEC, utworzenie instancji aplikacji MEC, jak również zapewnienie optymalnego wykorzystania zasobów serwerów MEC oraz zasobów sieci. Rolą orkiestratora jest koordynacja i delegowanie wykonywania operacji przez pozostałe komponenty systemu. Orkiestrator MEC deleguje zadania związane z obsługą pakietów aplikacji do repozytorium MEC oraz zadania dotyczące instancji aplikacji do Zarządcy platformy MEC.

2.1. Architektura orkiestratora MEC

W projekcie SyMEC funkcjonalność orkiestratora MEC została zaimplementowana za pomocą trzech modułów oprogramowania, przedstawionych na Rys. 2:

- Serwera obsługi wywołań - odpowiedzialnego za przyjmowanie, delegację i nadzór nad realizacją wszystkich żądań przychodzących do systemu SyMEC; moduł ten udostępnia na zewnątrz interfejs Mm1, który jest zgodny ze standardem ETSI MEC [1];
- Modułu alokacji - podejmującego decyzję o lokalizacji uruchamianej instancji aplikacji MEC i rezerwacji odpowiednich zasobów na serwerach MEC; moduł ten udostępnia własny interfejs REST API;
- Katalogu instancji aplikacji - przechowującego informacje o wszystkich instancjach aplikacji MEC uruchomionych w systemie SyMEC; moduł ten udostępnia standaryzowany interfejs TMF638 Service Inventory Management API [4] na zewnątrz systemu SyMEC;



Rys. 2. Struktura orkiestratora MEC.

2.2. Repozytorium aplikacji MEC

Repozytorium aplikacji jest obowiązkowym elementem biorącym udział w procesach cyklu życia aplikacji MEC, choć de facto nie jest częścią orkiestratora MEC. Repozytorium służy do przechowywania aplikacji MEC i udostępniania pakietów aplikacji MEC na potrzeby ich instalacji na serwerach MEC. Pakietem aplikacji nazywamy obiekt składający się z deskryptora aplikacji, obrazu aplikacji oraz identyfikatora pakietu. Deskryptor aplikacji służy do opisu wymagań aplikacji, takich jak m.in. ilość wymaganych zasobów procesora, rozmiar pamięci czy też wykorzystywane porty sieciowe. Obraz aplikacji jest obrazem kontenera zgodnym z platformą Docker [6]. Repozytorium implementuje tę część funkcjonalności, która jest związaną z cyklem życia pakietów, czyli odpowiada za dodanie i usunięcie pakietu oraz udostępnianie informacji o pakiecie i obrazach aplikacji pozostałym komponentom systemu SyMEC. Architektura systemu SyMEC rozdziela styki zarządzania cyklem życia pakietu i aplikacji pomiędzy osobne elementy funkcjonalne: orkiestrator oraz repozytorium aplikacji. Spowodowane jest to potrzebą odseparowania logiki funkcjonowania systemu od sposobów przechowania danych aplikacji.

2.3. Zarządzanie cyklem życia aplikacji MEC

Orkiestrator MEC zarządza cyklem życia aplikacji MEC oraz deleguje przeprowadzenie procesów cyklu życia pakietów aplikacji MEC do repozytorium aplikacji. Instancja aplikacji jest obiektem, który łączy jednostkę aplikacji uruchomioną w platformie MEC z rezerwacją zasobów oraz pakietem aplikacji.

Podstawowymi procesami cyklu życia instancji aplikacji MEC są utworzenie, uruchomienie, zatrzymanie i usunięcie. Proces uruchomienia aplikacji rozpoczyna się od utworzenia instancji aplikacji. Utworzenie instancji aplikacji MEC wymaga poprzedniego umieszczenia w systemie pakietu aplikacji MEC, który dostarcza informacje o ustawieniach i wymaganiach aplikacji oraz obraz kontenera aplikacji. Następnie należy wywołać żądanie uruchomienia instancji aplikacji. Orkiestrator po otrzymaniu żądania weryfikuje poprawność żądania oraz powiązanego z instancją pakietu aplikacji, po czym za pomocą modułu alokacji wybiera najlepszy serwer MEC oraz rezerwuje zasoby w abstrakcyjnym modelu systemu. Po wykonaniu rezerwacji zasobów orkiestrator wysyła żądanie uruchomienia aplikacji do zarządcy platformy MEC. Proces zatrzymania i usunięcia jest analogiczny i składa się z zatrzymania aplikacji przez zarządcę platformy, zwolnienia zasobów, a na końcu usunięcia instancji aplikacji.

2.4. Alokacja zasobów przez orkiestrator MEC

Moduł alokacji jako część orkiestratora uczestniczy w procesach rezerwacji i zwolnienia zasobów wymaganych dla uruchomienia/wyłączenia instancji aplikacji. Moduł ten wykorzystuje bazę danych, w której jest przechowywany abstrakcyjny model zasobów serwerów MEC wraz z informacją o topologii systemu. Ponadto, baza ta przechowuje informację o aktualnie uruchomionych instancjach aplikacji oraz zajmowanych zasobach. Na podstawie tych danych są podejmowane decyzje o alokacji kolejnych zasobów. Komunikacja z modułem alokacji odbywa się poprzez interfejs o architekturze REST API.

Moduł alokacji oferuje trzy algorytmy wyboru lokalizacji instancji aplikacji, tj. prosty, dokładny oraz heurystyczny. Algorytm prosty polega na wybraniu

serwera MEC spośród listy węzłów spełniających wymagania, tj. dysponującymi zasobami wymaganymi dla uruchomienia żądanej instancji aplikacji MEC oraz umożliwiającymi spełnienie wymagań dotyczących przekazu danych pomiędzy punktem dostępowym do którego jest dołączony użytkownik a rozważanym węzłem MEC (zwracanych w ramach kwerendy lokalizacji węzłów i punktów dostępowych). Algorytm dokładny wykorzystuje metody optymalizacji bazujące na programowaniu liniowym MILP (*Mixed Integer Linear Programming*) do wybrania najlepszego węzła względem przyjętej funkcji celu uwzględniającej m.in. wykorzystanie zasobów, odległość, czy też koszt. Algorytm ten dopuszcza również ograniczoną realokację uprzednio uruchomionych instancji aplikacji, jeśli działanie to pozwoli na uruchomienie żądanej instancji aplikacji MEC. Algorytm zaimplementowano wykorzystując narzędzie MiniZinc [2] zintegrowane z kodem modułu alokacji. Algorytm heurystyczny przyjmuje te same kryteria co algorytm dokładny i pozwala znaleźć rozwiązanie w znacząco krótszym czasie w porównaniu do algorytmu dokładnego, jednakże nie daje gwarancji, że rozwiązanie jest optymalne.

2.5. Interfejs REST API orkiestratora MEC

Orkiestrator MEC oferuje interfejsy zarządzania instancjami aplikacji oraz pakietami aplikacji w architekturze REST API. Zasoby oraz modele danych zostały oparte o standardy ETSI MEC. Do funkcji interfejsu należą operacje tworzenia, pobierania i usuwania pakietów aplikacji, uruchamianie i wyłączanie instancji aplikacji oraz aktualizacja stanu aplikacji lub pakietu.

2.6. Integracja orkiestratora SyMEC z systemem OSS/BSS

W celu umożliwienia łatwej integracji orkiestratora SyMEC z systemami OSS/BSS operatorów zastosowano standard TMF639 - Resource Inventory Management API [5] do przechowywania obiektów wykorzystywanych zarówno przez orkiestrator jak i przez repozytorium SyMEC. Ponadto, zintegrowano orkiestrator z przykładowym systemem OSS/BSS wykorzystującym standardy TMF638 Service Inventory Management API, TMF639 Resource Inventory Management API i TMF640 Service Configuration and Activation API [13]. Przygotowany system OSS/BSS składa się z modułu orkiestracji usług, modułu aktywacji usług oraz portalu użytkownika. Opracowane oprogramowanie pozwala na pełne zademonstrowanie integracji systemu SyMEC w środowisku operatora telekomunikacyjnego.

3. ZARZĄDCA PLATFORMY MEC

Zarządca platformy MEC, MEPM (*MEC Platform Manager*), odpowiada za uruchomienie/zatrzymanie instancji aplikacji MEC w serwerach MEC i zapewnienie komunikacji w wirtualnej infrastrukturze systemu. Po otrzymaniu żądania od orkiestratora, zarządca MEPM weryfikuje jego poprawność, a następnie przekształca to żądanie do postaci zrozumiałej dla zarządcy infrastruktury wirtualnej VIM (*Virtual Infrastructure Manager*). Należy zwrócić uwagę, iż w systemie SyMEC mogą być zastosowane różne platformy wirtualizacji, a rolą zarządcy MEPM jest dostosowanie żądań orkiestratora do odpowiedniej składni i funkcjonalności oferowanej przez zarządcę VIM. W opracowanym prototypie systemu SyMEC rolę zarządcy VIM pełni Kubernetes (k8s) [7], do którego zarządca MEPM przesyła żądania z wykorzystaniem interfejsu REST API. Ponadto, zarządca platformy MEC udostępnia aplikacjom MEC usługi platformy MEC, a także konfiguruje dostęp do tworzonych aplikacji MEC dla użytkowników końcowych.

Główne funkcje realizowane przez zarządcę MEPM w procesie obsługi żądania utworzenia nowej instancji aplikacji MEC są następujące:

- Utworzenie definicji kontenera, w którym uruchomiona zostanie aplikacja MEC;
- Utworzenie trwałej przestrzeni dyskowej dla kontenera (persistent volume), o ile aplikacja tego wymaga;
- Utworzenie w klastrze k8s wirtualnego obiektu, typu Deployment, w którym zostanie uruchomiony kontener z tworzoną aplikacją MEC;
- Utworzenie w klastrze k8s usługi udostępnienia portów uruchomionej instancji aplikacji MEC dla użytkowników;
- W przypadku aplikacji wykorzystujących protokół HTTP/HTTPS, utworzenie w Kubernetesie wpisu filtru wejściowego umożliwiającego dostęp użytkownika do uruchomionej instancji aplikacji na typowych portach;
- Utworzenie wpisu nazwy symbolicznej dla uruchomionej instancji aplikacji w lokalnym serwerze nazw platformy MEC, która zostaje udostępniona w lokalnej sieci użytkownika.

Po wykonaniu opisanych powyżej operacji zarządca MEPM wysyła do orkiestratora potwierdzenie poprawnego wykonania żądania, a także przekazuje identyfikator tworzonej instancji, numery portów oraz nadaną nazwę symboliczną w serwerze nazw. Następnie inicjowany jest proces monitorowania stanu aplikacji MEC. Gdy wszystkie kontenery wymagane przez aplikację MEC zostaną poprawnie uruchomione, zarządca MEPM asynchronicznie powiadamia orkiestrator o zakończeniu procesu utworzenia nowej instancji. Proces zatrzymania instancji przebiega podobnie do procesu tworzenia nowej instancji, przy czym następuje zwolnienie zasobów, usunięcie nazw symbolicznych i zatrzymanie usług udostępnienia zatrzymanej instancji.

4. SYSTEM ZARZĄDZANIA SYMEC

System zarządzania SyMEC to zestaw modułów, które mają za zadanie wspierać procesy zarządzania systemem SyMEC przez operatora. Moduły te współpracują z elementami systemu OSS/BSS, niejako pośrednicząc w realizacji typowych funkcji zarządzania. Przede wszystkim dotyczy to funkcji zarządzania urządzeniami fizycznymi systemu SyMEC, a częściowo także funkcji zarządzania usługami MEC (w zakresie monitoringu). Przyjęto założenie, że moduły systemu

zarządzania SyMEC będą integrowane z OSS/BSS, a wręcz mogą stać się elementami OSS/BSS.

W ramach prototypu systemu SyMEC zaimplementowano podzbiór funkcjonalności obejmujący moduły zarządzania zasobami oraz monitoringu.

4.1. Moduł zarządzania zasobami

Moduł zarządzania zasobami umożliwia gromadzenie oraz przechowywanie informacji o zasobach, które są używane w systemie SyMEC. W tym celu moduł wykorzystuje bazę danych i udostępnia interfejs do obsługi żądań informacji o zasobach oraz modyfikacji zasobów (dodanie, usunięcie, aktualizacja). Ze względu na konieczność integracji modułów systemu zarządzania SyMEC z systemami OSS/BSS operatora, moduł zarządzania zasobami implementuje interfejs REST API, który stanowi adaptację interfejsu zdefiniowanego przez stowarzyszenie TeleManagement Forum (TMForum) [3] w Resource Inventory API REST Specification (TMF639) [5]. TMForum to stowarzyszenie, które opracowuje standardy i zalecenia dla systemów teleinformatycznych dostawców usług i operatorów telekomunikacyjnych. Standard TMF639 definiuje generyczny model zasobów, który umożliwia elastyczną reprezentację różnego typu zasobów.

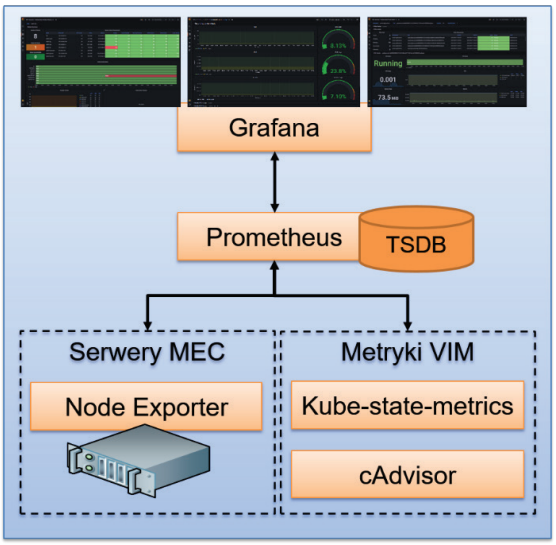
Moduł zarządzania zasobami zarządza informacjami o serwerach MEC i punktach dostępowych. Obiekty punktów dostępowych zawierają także informacje o istniejących połączeniach z węzłami MEC. Dane te stanowią źródło informacji dla modułu alokacji orkiestratora MEC, który na ich podstawie wyznacza ilość dostępnych zasobów w systemie SyMEC, np. przy podejmowaniu decyzji o uruchomieniu aplikacji MEC. Z danych tych korzystają także moduły systemu OSS/BSS.

4.2. Moduł monitoringu

Moduł monitoringu zapewnia ogląd działania systemu SyMEC w różnych aspektach. Zakres monitoringu obejmuje stan serwerów, ich obciążenia i wykorzystania zasobów oraz status uruchomionych obiektów wirtualnych (m.in. aplikacje uruchamiane w SyMEC). Monitoring wykorzystania zasobów prowadzony jest zarówno pod kątem faktycznego ich użycia, jak i stopnia rezerwacji zasobów.

Moduł monitoringu jest zbudowany pod kątem uruchamiania w kontenerach Docker [6] na platformie Kubernetes [7]. Ma to na celu ułatwienie monitorowania elementów SyMEC, który także jest uruchamiany na platformie Kubernetes. Z tego względu moduł monitoringu jest zaimplementowany w postaci plików (skryptów) w formacie YAML, w których zdefiniowane są obiekty Kubernetes odpowiadające poszczególnym składowym modułu monitoringu.

W strukturze modułu monitoringu, przedstawionej na Rys. 3, można wyróżnić część backend odpowiedzialną za zbieranie i przechowywanie danych oraz część frontend odpowiedzialną za prezentację danych użytkownikowi.



Rys. 3. Struktura modułu monitoringu

Centralny element części backend stanowi serwis (system) Prometheus [8], który pobiera metryki ze wskazanych celów i gromadzi je w bazie szeregów czasowych (*Time Series Database*). Baza ta może być następnie odpytana za pomocą dostarczanego przez serwis Prometheus języka zapytań PromQL (*Prometheus Query Language*). Do części backend należą także komponenty będące źródłem metryk. Jednym z nich jest serwis Node exporter [9], który jest uruchamiany na platformie Kubernetes jako DeamonSet (czyli jedno urządzenie wirtualne pod na każdym węźle klastra). Komponent ten przekształca i udostępnia metryki sprzętowe z poszczególnych węzłów do formatu wymaganego przez serwis Prometheus. Drugim źródłem metryk jest serwis kube-state-metrics [10], który na podstawie danych z API Kubernetes generuje metryki o stanie obiektów obecnych na platformie Kubernetes. Kolejnym źródłem metryk jest serwis cAdvisor [11], który jest zintegrowany z platformą Kubernetes (dokładniej z kubelet będącym agentem działającym na każdym węźle). Komponent ten udostępnia metryki działających kontenerów.

Cześć frontend modułu monitoringu jest realizowana za pomocą aplikacji internetowej Grafana [12], która umożliwia pobieranie metryk zgromadzonych w bazie serwisu Prometheus i tworzenie na ich podstawie pulpitów menedżerskich (*dashboard*) z interaktywnymi wizualizacjami. W ramach projektu opracowano dedykowane pulpity menedżerskie, które wizualizują m.in.: status węzłów i ich obciążenie, faktyczne i alokowane wykorzystanie zasobów obliczeniowych (CPU, RAM, przestrzeń dyskowa) oraz stan i charakterystykę utworzonych obiektów pod (w tym aplikacji uruchamianych w SyMEC).

5. PODSUMOWANIE

W artykule zaprezentowano elementy systemu sterowania i zarządzania SyMEC. Jednocześnie opisano architekturę i implementację tego systemu z uzasadnieniem wyboru konkretnych rozwiązań. Zaprezentowane rozwiązanie jest częścią systemu

SyMEC i tworzy w pełni funkcjonalny system MEC, który jest przystosowany do wdrożenia w środowisku operatora telekomunikacyjnego. Opracowane oprogramowanie zostało wdrożone i przetestowane w ogólnopolskiej sieci badawczej PL-LAB 2020. Instalacja ta jest podstawą do dalszych prac nad technologią MEC.

Badania, których wyniki są przedstawione w tym artykule, były prowadzone w ramach projektu POIR 04.01.02.00-00-0107/17, pt. „System MEC dla wspierania zaawansowanych aplikacji w środowisku sieci przewodowych i bezprzewodowych 3G/4G/5G", współfinansowanego w ramach Programu Operacyjnego Innowacyjny Rozwój na lata 2014-2020.

LITERATURA

[1] ETSI standard, „Multi-access Edge Computing (MEC); MEC Management; Part 2: Application lifecycle, rules and requirements management", ETSI GS MEC 0 010-2, v2.1.1, November 2019.
[2] MiniZinc: https://www.minizinc.org/ (dostępna: 30.05.2022)
[3] TeleManagement Forum: https://www.tmforum.org (dostępna: 30.05.2022)
[4] TM Forum Specification, Service Inventory API REST Specification (TMF638): https://projects.tmforum.org/wiki/display/API/TMF638+Service+Inventory+Management+API+REST+Specification+R16.5.1 (dostępna: 30.05.2022)
[5] TM Forum Specification, Resource Inventory API REST Specification (TMF639): https://projects.tmforum.org/wiki/display/API/TMF639+Resource+Inventory+Management+API+REST+Specification+R17.0.1 (dostępna: 30.05.2022)
[6] Docker: https://www.docker.com/ (dostępna: 30.05.2022)
[7] Kubernetes: https://kubernetes.io/ (dostępna: 30.05.2022)
[8] Prometheus: https://prometheus.io/docs/introduction/overview/ (dostępna: 30.05.2022)
[9] Node exporter: https://github.com/prometheus/node_exporter (dostępna: 30.05.2022)
[10] Kube-state-metrics: https://github.com/kubernetes/kube-state-metrics (dostępna: 30.05.2022)
[11] cAdvisor: https://github.com/google/cadvisor (dostępna: 30.05.2022)
[12] Grafana: The open observability platform: https://grafana.com/docs/ (dostępna: 30.05.2022)
[13] TM Forum Specification, Activation and Configuration API REST Specification (TMF640): https://projects.tmforum.org/wiki/display/API/TMF640+Activation+and+Configuration+API+REST+Specification+R15.5.1 (dostępna: 30.05.2022)

Wdrożenie systemu MEC w sieciach WLAN: Moduł współpracy serwera MEC z siecią dostępową WLAN

MEC system implementation in WLAN networks:
MEC server module providing integration with WLAN access networks

Michał Hoeft, Krzysztof Gierłowski, Józef Woźniak, Wojciech Gumiński, Krzysztof Nowicki, Jacek Rak

Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska, Gdańsk
michal.hoeft@pg.edu.pl, krzgierl@pg.edu.pl, jowoz@eti.pg.edu.pl, wogumin@pg.edu.pl, know@eti.pg.edu.pl, jrak@pg.edu.pl

DOI: 10.15199/59.2022.4.51

Streszczenie: **W artykule zaprezentowano zasady funkcjonowania modułu współpracy sieci dostępowych WLAN z serwerem MEC oraz pozostałymi komponentami ekosystemu, budowanego w ramach projektu SyMEC. Zaproponowana architektura pozwala na ujednolicenie dostępnej funkcjonalności oraz komunikację zewnętrznych aplikacji z podmodułami odpowiedzialnymi za zbieranie i przekazywanie informacji o sieci dostępowej. Jako przykładową aplikację zademonstrowano wizualizację szeregu parametrów sieci dostępowej WLAN.**

Abstract: **The paper presents specification and implementation details of the MEC server module providing integration with WLAN access networks, as well as its cooperation with other components of the ecosystem built as elements of SyMEC project. The proposed architecture allows to unify the available functionality and provides communication of edge applications with submodules responsible for collecting and disseminating information about the access network. As an example, the application demonstrating visualization of a number of important WLAN parameters was prepared.**

Słowa kluczowe: **serwery MEC, sieci dostępowe WLAN, moduł współpracy z MEC**

Keywords: **MEC server, WLAN access network, MEC integration module**

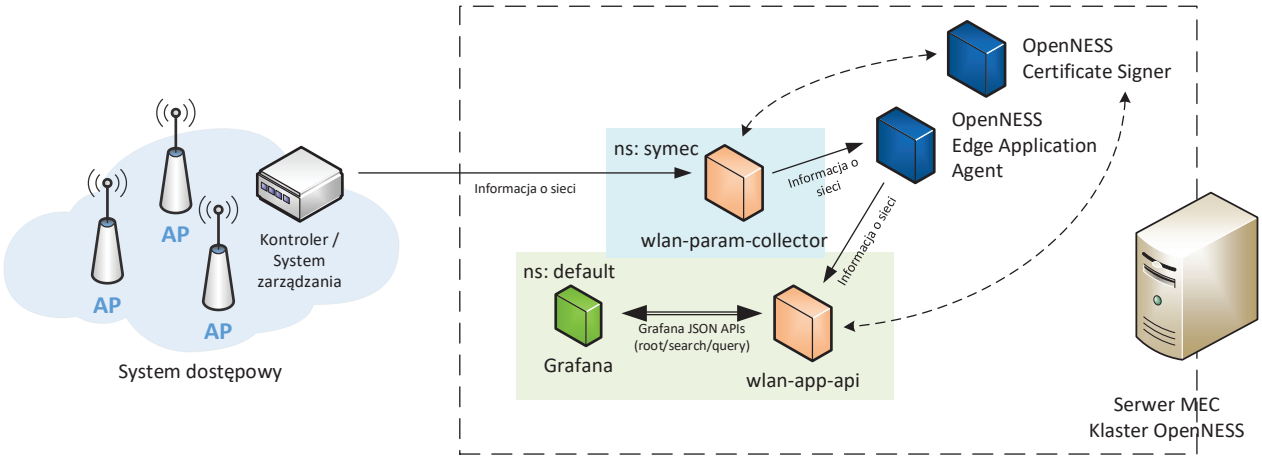
1. WSTĘP

Jedną z najważniejszych technik zapewniających zmniejszenie opóźnienia i ograniczenie ruchu w sieci szkieletowej staje się w ostatnim okresie Multi-access Edge Computing (MEC). Koncepcja ta związana jest z przenoszeniem części zasobów obliczeniowych i/lub aplikacji usługowych z chmury obliczeniowej, czyli z odległych centrów danych, na brzeg sieci, blisko punktów dostępu radiowego. Technika MEC pozwala tym samym na poprawę efektywności przetwarzania i świadczenia szeregu usług w środowisku chmurowym. Przeniesienie dotyczy przy tym głównie aplikacji wrażliwych na opóźnienia, bądź też aplikacji zasobochłonnych, jak np. aplikacje wymagające strumieniowania z odległych serwerów do wielu użytkowników jednej sieci dostępowej. Implementacja serwerów MEC wprowadza zatem wręcz rewolucję w zarządzaniu siecią [1].

Jako korzyści wynikające z zastosowania serwerów typu MEC w bezprzewodowych sieciach dostępowych wskazuje się możliwość odciążenia terminali mobilnych poprzez delegowanie złożonych zadań obliczeniowych i wyniesienia ich do ulokowanych stosunkowo niedaleko elementów chmury obliczeniowej [2,3]. Część z tych zdań może być specyficzna dla środowiska wdrożenia i dotyczyć np. optymalizacji przełączania w niejednorodnej, heterogenicznej sieci dostępowej [4], czy optymalizacji sposobu dostarczania treści w sieciach typu VANET [5].

Aby usystematyzować, zapewnić kompatybilność oraz możliwości wykorzystania nowej architektury poprzez implementowane aplikacje uruchamiane na brzegu sieci niezbędne jest opracowanie standaryzowanych interfejsów oraz powiązań pomiędzy blokami funkcjonalnymi całej architektury. W zakresie tym intensywne prace prowadzi ETSI, w ramach grupy roboczej ETSI MEC ISG która opracowała szereg dokumentów definiujących nowe usługi MEC dostępne dla aplikacji uruchamianych na brzegu sieci [6]. Jedną z takich usług jest dostarczanie informacji o aktualnym stanie sieci dostępowej WLAN [7]. Dokumenty standaryzujące precyzują sposób dostępu do tych informacji, ale sama kwestia implementacji komponentu, który będzie w stanie takie dane zbierać, gromadzić i przetwarzać pozostaje nadal otwarta. Pojawia się więc konieczność opracowanie modułów współpracy sieci dostępowych, z uwzględnieniem ich specyfiki, z ulokowanymi na brzegu sieci serwerami MEC. Taki właśnie komponent dedykowany do pracy z sieciami dostępowymi WLAN został zaprojektowany i zaimplementowany jako jeden z elementów architektury systemu SyMEC [8,9,10].

Istotną korzyścią zastosowania takiego modułu jest możliwość dostarczania szerokiego zestawu informacji opisujących aktualny stan sieci dostępowej, znacząco wykraczający poza zestaw danych dostępnych po stronie klienta. Wcześniejsze prace dotyczące realizacji usług transmisji głosowych oraz transmisji wideo [11,12,13,14,15], wskazywały na istotne korzyści wynikające z wykorzystania dodatkowych informacji o sieci dostępowej. Wprowadzenie ujednoliconych zasad ich dostarczania oraz sposobu współpracy komponentów



Rys. 1. Architektura modułu współpracy z sieciami WLAN

zbierających i przetwarzających te dane pozwoli na wykorzystanie wyników tych prac w szerszym kontekście np. w sieciach 5G.

W artykule, w kolejnych jego rozdziałach, przedstawiono architekturę i opis cech funkcjonalnych modułu współpracy serwera MEC z sieciami dostępowymi WLAN oraz przykładowe zasady monitoringu i wizualizacji mierzonych parametrów (aplikacja Grafana [16]) specyficznych dla sieci WLAN, a także związanych z wymaganiami poprawnej realizacji oferowanych przez serwer MEC aplikacji.

2. ARCHITEKTURA MODUŁU WSPÓŁPRACY Z SIECIAMI WLAN

W ramach przeprowadzonych prac zaprojektowano, zaimplementowano oraz przetestowano moduł odpowiedzialny za współpracę z sieciami dostępowymi klasy Wireless Local Area Network (WLAN). Moduł ten przeznaczony jest do uruchamiania w środowisku systemu SyMEC, wykorzystującego platformę OpenNESS [17]. Moduł oferuje możliwość odczytu wartości szeregu parametrów opisujących aktualne warunki pracy systemu dostępowego WLAN, a następnie udostępnienia ich aplikacjom uruchamianym w systemie SyMEC.

Architektura modułu została przedstawiona na Rys. 1 i obejmuje dwa rodzaje podmodułów:

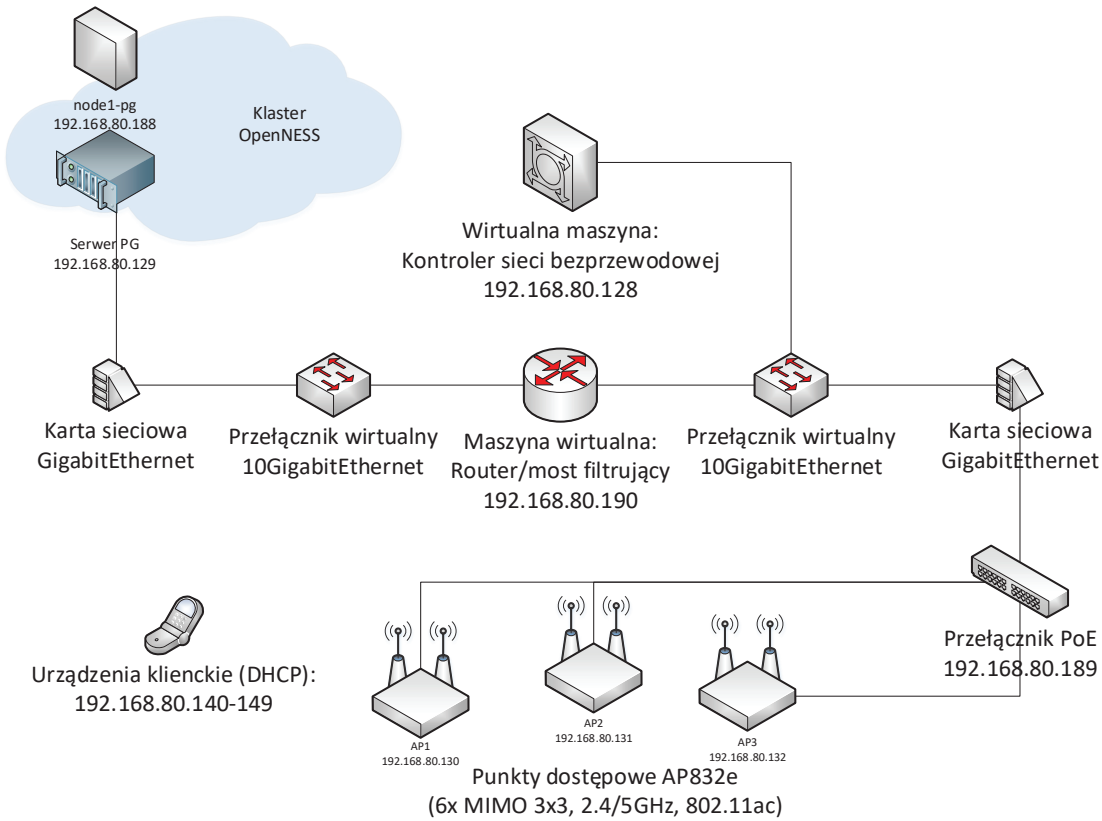
- podmoduł zbierający informacje (*wlan-param-collector*) – odpowiedzialny za komunikację z systemem zarządzania sieci dostępowej i odczytywanie parametrów jej pracy,
- podmoduł udostępniający dane aplikacjom (*wlan-app-api*) – możliwe jest jednoczesne wykorzystanie wielu rodzajów podmodułów tego typu, w celu udostępnienia danych o stanie sieci WLAN w szeregu formatów i z użyciem różnorakich interfejsów logicznych.

Powyższe podmoduły komunikują się przy wykorzystania komponentu Edge Application Agent (EAA), stanowiącego część platformy OpenNESS i działającego na zasadzie *publish-subscribe* [18]. Powyższe podejście pozwala na osiągnięcie skalowalności poziomej podmodułów udostępniania danych, gdyż ich liczba nie wpływa na obciążenie podmodułu zbierającego dane – publikuje on informacje jednokrotnie, a moduł EAA zajmuje się ich dostarczeniem do wszystkich zainteresowanych odbiorców.

Wykorzystanie modułu EAA wiąże się z koniecznością rejestracji w systemie OpenNESS wszystkich wykorzystujących go elementów (zarówno podmodułu zbierającego dane, jak i podmodułów udostępniania) oraz uzyskania przez nie cyfrowych certyfikatów w formacie x.509. Certyfikaty te są następnie wykorzystywane do uwierzytelniania podmodułów wobec modułu EAA oraz do kryptograficznej ochrony wymienianego z nim ruchu sieciowego.

Zaimplementowane moduły w pełni obsługują powyższe mechanizmy, wprowadzone w platformie OpenNESS od wersji 2020.12 w postaci Edge Application API i Edge Application Authentication API.

Zaimplementowany podmoduł zbierający informacje *wlan-param-collector* został zaprojektowany w sposób umożliwiający łatwe jego dostosowanie do współpracy z różnymi rodzajami sieci dostępowych, o ile dostęp do ich interfejsów zarządzania możliwy jest przy wykorzystaniu narzędzi linii poleceń (command-line), zwracających informacje w postaci tekstu. Rozwiązanie takie oferuje daleko posuniętą elastyczność, gdyż mechanizmy zarządzania nowoczesnymi systemami WLAN w przeważającej większości oferują narzędzia klienckie tego typu, niezależnie od logicznego interfejsu udostępnianego przez system zarządzania w przypadku konkretnej implementacji.



Rys . 2. Schemat systemu testowego

Prace implementacyjne dotyczące modułu współpracy z sieciami dostępowymi WLAN prowadzone były z wykorzystaniem środowiska weryfikacyjnego, umożliwiającego prowadzenie bieżącej kontroli poprawności funkcjonowania implementowanych mechanizmów. Środowisko to obejmowało część witalizowaną w postaci w pełni funkcjonalnego, wielowęzłowego wdrożenia platformy OpenNESS kolejno w wersjach 2020.9, 2020.12 i 2021.3, a także fizyczną – system dostępowy WLAN, wykorzystujący urządzenia klasy korporacyjnej, firmy Fortinet. Ogólna architektura powyższego systemu dostępowego została przedstawiona na Rys. 2

Środowisko wykorzystuje serwer wirtualizacyjny VMWare ESXi, służący do uruchomienia kontrolera sieci bezprzewodowej Fortinet MC3200-V oraz routera/mostu filtrującego. Za ich pośrednictwem podłączono fizyczną sieć dostępową złożoną z 3 punktów dostępowych AP832e, wyposażonych łącznie w 6 interfejsów radiowych 2.4/5 GHz, obsługujących komunikację z klientami bezprzewodowymi zgodnie ze standardem IEEE 802.11ac MIMO 3x3.

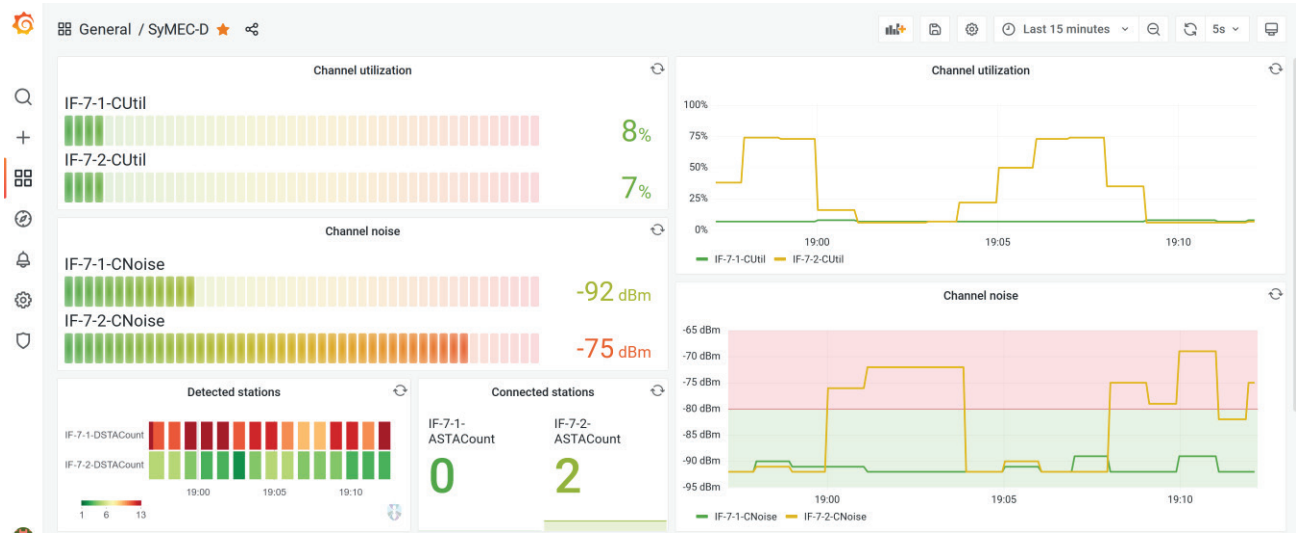
3. DEMONSTRATOR WSPÓŁPACY Z APLIKACJĄ ZEWNĘTRZNĄ

Prezentowana implementacja podmodułu zbierającego informacje udostępnia dane dotyczące trzech grup wskaźników funkcjonowania systemu dostępowego:

- wskaźników dotyczących infrastruktury sieci dostępowej,
- wskaźników dotyczących urządzeń klienckich,
- wskaźników dotyczących otoczenia sieci dostępowej.

Zaimplementowany podmoduł udostępniania danych aplikacjom (*wlan-app-api*), w przygotowanym demonstratorze, umożliwia dostęp do odczytanych informacji popularnej aplikacji Grafana, przeznaczonej do analizy i wizualizacji danych. Podmoduł rejestruje się jako odbiorca danych publikowanych przez podmoduł zbierający informacje, w wyniku czego okresowo otrzymuje aktualizacje wartości parametrów pracy sieci dostępowej. Informacje te przechowywane są przez konfigurowalny okres, umożliwiając podmodułowi analizę stanu sieci WLAN i jego zmian w powyższym okresie. Wyniki analizy udostępniane są aplikacji Grafana przy wykorzystaniu interfejsu HTTP-REST i formatu JSON, zgodnie ze specyfikacją Grafana JSON Root/Search/Query API. Przykładowy widok aplikacji demonstrującej wykorzystanie modułu został zaprezentowany na Rys. 3.

Wgląd do aktualnych danych opisujących sieć dostępową może być korzystny z puntu widzenia usług świadczonych w sieci, dla których informacje dostarczane przez standardowe narzędzia i sterowniki systemów operacyjnych mogą być niewystarczające aby oszacować aktualny stan sieci i potencjalną jakość usługi realizowaną przy jej pomocy. Jedną z niewielu informacji, dotyczącą sieci

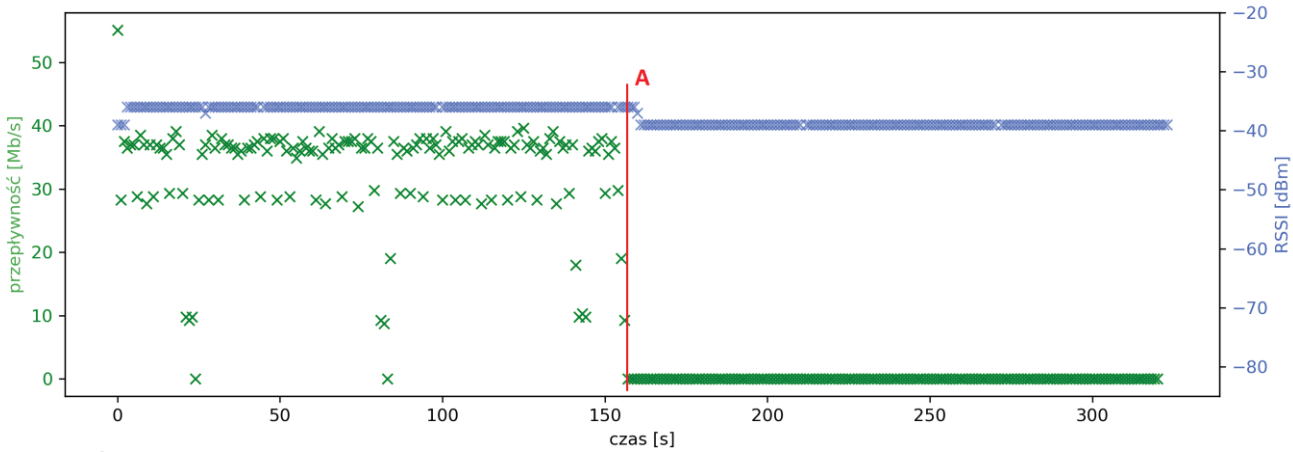


Rys. 3. Przykładowy wizualizacja danych z modułu współpracy z sieciami WLAN

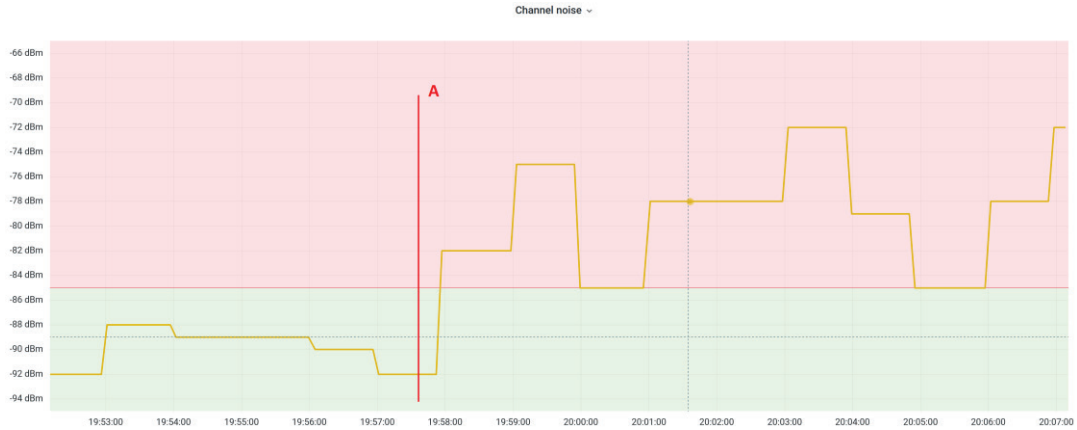
WLAN, dostępną po stronie klienta jest wskaźnik RSSI. Wykres wartości tego parametru raportowanych przez interfejs bezprzewodowy pracujący pod kontrolą systemu operacyjnego Linux oraz przepływność testowego połączenia TCP zaprezentowano na Rys 4.

W trakcie opisywanego testu włączono równolegle, na tym samym kanele, konkurencyjną transmisję wykorzystującą technikę TDMA (moment włączenia zaznaczono czerwoną linią oznaczoną jako A), co spowodowało, że mechanizmy unikania kolizji sieci Wi-Fi nie były w stanie dostosować się do koegzystencji z transmisją innego typu. Widoczna jest niewielka zmienność raportowanych wartości RSSI w zakresie 3 dB, która dla tego poziomu wartości RSSI (-36 dBm) nie wyjaśnia spadku przepływności analizowanego połączenia. Zastosowanie dodatkowego modułu opisanego w tym artykule, dzięki możliwości analizy większej liczby parametrów, pozwala spojrzeć na całą sytuację z szerszej perspektywy (przykład takich informacji został zaprezentowany na Rys.5).

Wizualizacja danych dostarczanych z kontrolera sieci dostępowej Wi-Fi pozwala zidentyfikować niezgodne ze standardowymi sieciami Wi-Fi transmisje, które w tym przypadku będą wykrywane jako zwiększenie średniej wartości wskaźnika mocy szumów występujących w kanale (*Channel noise*). Interpretacja dodatkowych danych pozwala na identyfikację problemu i wyjaśnienie przyczyny spadku przepływności obserwowanych dla testowego połączenia TCP.



Rys. 4. Średnia wartość przepływności testowego połączenia TCP (kolor zielony, oś lewa) oraz wartości parametru RSSI (kolor niebieski, oś prawa).



Rys. 5. Wizualizacja parametru channel noise pokazująca powód spadku przepływności połączenia testowego

4. PODSUMOWANIE

W artykule zaprezentowano sposób integracji systemu MEC z bezprzewodową siecią dostępową WLAN. Jednocześnie opisano architekturę tego modułu i sposoby integracji z innymi komponentami systemu SyMEC. Wskazane zostały korzyści wynikające z takiego wdrożenia pozwalające na przekazywanie szczegółowych informacji o stanie sieci dostępowej elementom aplikacyjnym realizującym usługi, a ulokowanym na brzegu sieci. Wdrożenie zostało zrealizowane w ogólnopolskiej sieci badawczej PL-LAB i może być wykorzystane w dalszych pracach realizowanych w tym środowisku.

Badania, których wyniki są przedstawione w tym artykule, były prowadzone w ramach projektu POIR 04.01.02.00-00-0107/17,pt. „System MEC dla wspierania zaawansowanych aplikacji w środowisku sieci przewodowych i bezprzewodowych 3G/4G/5G”, współfinansowanego w ramach Programu Operacyjnego Innowacyjny Rozwój na lata 2014-2020.

LITERATURA

[1] *Tendencje w rozwoju polskiej i światowej telekomunikacji i teleinformatyki. (Red. J. Woźniak), Wydawnictwo WAT, 2020.*
[2] Y. Li, G. Xu, J. Ge, X. Fu and P. Liu, "Communication and Computation Cooperation in Wireless Network for Mobile Edge Computing," in *IEEE Access*, vol. 7, s.106260-106274, 2019
[3] J. Saeid i in: Mobility-Aware Offloading Decision for Multi-Access Edge Computing in 5G Networks, *Sensors* 22, nr 7: 2692, 2022
[4] T. M. Ho i. in: Joint Server Selection, Cooperative Offloading and Handover in Multi-Access Edge Computing Wireless Network: A Deep Reinforcement Learning Approach, in *IEEE Transactions on Mobile Computing*, vol. 21, nr 7, s. 2421-2435, 2022
[5] J. Al-Badarneh i in.: Cooperative mobile edge computing system for VANET-based software-defined content delivery, Computers Electrical Engineering, t. 71. s. 388–397,, 2018
[6] ETSI GS MEC 003: Multi-access Edge Computing; Framework and Reference Architecture, 2022
[7] ETSI GS MEC 028: Multi-access Edge Computing (MEC); WLAN Information API, 2020
[8] A. Bęben, W. Burakowski, B. Belter, D. Parniewicz: Architektura systemu MEC dla wspierania zaawansowanych aplikacji w środowisku sieci przewodowych i bezprzewodowych 3G/4G/5G. Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, 2020, Nr 7-8, s.144-148
[9] Strona internetowa projektu SyMEC symec.com.pl (dostępna 06.06.2022)
[10] M. Hoeft, i in.: Moduł platformy SyMEC odpowiedzialny za współpracę serwera MEC z sieciami komórkowymi i bezprzewodowymi sieciami dostępowymi WLAN. Przegląd Telekomunikacyjny+Wiadomości Telekomunikacyjne, 2020, Nr 7-8, s.149-154.
[11] A. Kostuch, K. Gierłowski, J. Wozniak: Performance analysis of multicast video streaming in IEEE 802.11 b/g/n testbed environment. IFIP WMNC, 2009, Springer, Berlin, s. 92-105
[12] K. Gierłowski, A. Kostuch, J. Woźniak, K. Nowicki: Testbed analysis of video and VoIP transmission performance in IEEE 802.11 b/g/n networks. Telecommunication Systems, 2011, Vol.48/3, s. 247-260
[13] M. Hoeft, i in.: Measurements of QoS/QoE parameters for media streaming in a PMIPv6 testbed with 802.11 b/g/n WLANs. Metrology and Measurement Systems, 2012, Vol.18/2, s. 283-294
[14] W. U. Rahman, C. S. Hong and E. Huh, "Edge Computing Assisted Joint Quality Adaptation for Mobile Video Streaming," in *IEEE Access*, vol. 7, pp. 129082-129094, 2019
[15] M. Maghsoud,, J. Noll. 2021. "Estimating PQoS of Video Streaming on Wi-Fi Networks Using Machine Learning" *Sensors* 21, no. 2: 62
[16] Grafana: The open observability platform: https://grafana.com/ (dostępna 06.06.2022)
[17] OpenNESS - Open Network Edge Services Software: https://www.openness.org/ (dostępna 06.06.2022)
[18] Dokumentacja modułu Edge Application Agent: https://github.com/smart-edge-open/specs/blob/master/doc/applications-onboard/smartedge-open-eaa.md (dostępna 06.06.2022)

Uruchomienie systemu MEC w sieciach 3GPP: Moduł współpracy serwera MEC z sieciami 4G i 5G

MEC system implementation in 3GPP networks:
MEC server module providing integration with 4G and 5G networks

Dariusz Więcek, Tomasz Anuszczyk; Igor Michalski; Krzysztof Rzeźniczak

Instytut Łączności – Państwowy Instytut Badawczy, Wrocław
d.wiecek@il-pib.pl, t.anuszczyk@il-pib.pl, i.michalski@il-pib.pl, k.rzezniczak@il-pib.pl

DOI: 10.15199/59.2022.4.52

Streszczenie: **W niniejszym artykule przedstawiono sposób integracji, uruchomienia i testowania modułu MEC w sieciach 3GPP: 4G i 5G. Wykorzystano do tego celu środowisko OpenNESS oraz w części radiowej RAN SDR srsRAN i 5G OpenAirInterface. Skonfigurowano środowisko sieciowe 4G/5G, oprogramowano karty SIM użyte w smartfonach i modemach oraz wykonano praktyczne podłączenie i testy serwera MEC w środowisku 4G i 5G.**

Abstract: **This article describes how to integrate, run and test the MEC system module in the 3GPP: 4G and 5G networks. The OpenNESS environment was used for this purpose together with the RAN SDR srsRAN and 5G OpenAirInterface radio access. The 4G/5G network environments were set up, SIM cards were programmed for the smartphones and modems and the MEC server was connected and tested in a 4G and 5G environment.**

Słowa kluczowe: **serwery MEC, sieci 4G, 5G, moduł współpracy z MEC**

Keywords: **MEC server, 4G and 5G networks, MEC integration module**

1. WSTĘP

W celu zmniejszenia opóźnień, ograniczenia ruchu w sieci szkieletowej czy efektywnego zarządzania ruchem na brzegu sieci – w pobliżu stacji bazowej – do której dołączony jest dany użytkownik, rozwijana jest koncepcja Multi-access Edge Computing (MEC). Oznacza ona dynamiczne wykorzystywanie przez serwery aplikacji zarządzania ruchem w zasobach będących blisko użytkownika, na brzegu sieci, zamiast w tradycyjnych zasobach chmury obliczeniowej i odległych centrach danych. Techniki MEC pozwalają na wiele nowych sposobów przetwarzania z wykorzystaniem danych dostępnych na brzegu sieci – co pozwala na szybką i efektywną pracę w celu realizacji usług, a także na pojawienie się zupełnie nowych usług korzystających bezpośrednio z danych dostępnych na brzegu sieci [1].

Aby zapewnić odpowiednią współpracę serwera MEC z daną siecią 3GPP konieczne jest zdefiniowanie odpowiedniej architektury oraz interfejsów dostępu. W ramach standaryzacji ETSI w grupie roboczej ETSI MEC ISG opracowano wiele różnych zaleceń i norm dotyczących współpracy MEC z sieciami 3GPP. Podstawowe sposoby włączenia serwera MEC do sieci 4G i 5G wskazano w [2]. Do serwera MEC przekazywane są różne dane dostępne na brzegu sieci celem efektywniejszego zarządzania ruchem i realizacją konkretnych usług specyficznych dla danych warunków istniejących w lokalnym otoczeniu stacji bazowej. Jedną z takich usług jest dostarczanie informacji o aktualnym stanie sieci radiowej RAN (Radio Access Network) 3GPP za pomocą Radio Network Information (RNI) [3]. Standaryzacja ETSI pozwala na zdefiniowanie podstawowych informacji o interfejsie wymiany danych i sposobach ich przekazywania, jednak szczegółowa implementacja interfejsów i usług MEC nie jest standaryzowana i pozostawiona do decyzji wykonawcy serwera usług MEC [4, 5, 6]. Ważną korzyścią stosowania informacji RNI dostępnej na brzegu sieci jest dostęp do aktualnego stanu środowiska radiowego sieci 3GPP – zmieniającego się dynamicznie na bieżąco i możliwości wykorzystania tych danych do realizacji odpowiednich usług użytkowników w sposób odpowiadający na aktualne potrzeby użytkowników tj. np. do realizacji adaptacyjnej jakości usługi transmisji video [7].

W artykule przedstawiono sposoby implementacji interfejsów serwera MEC w sieciach 3GPP, przedstawiono architekturę i opis cech funkcjonalnych opracowanego modułu współpracy serwera MEC z sieciami 4G i 5G oraz pokazano sposoby pomiaru i wizualizacji mierzonych parametrów sieci przy użyciu aplikacji Grafana [8].

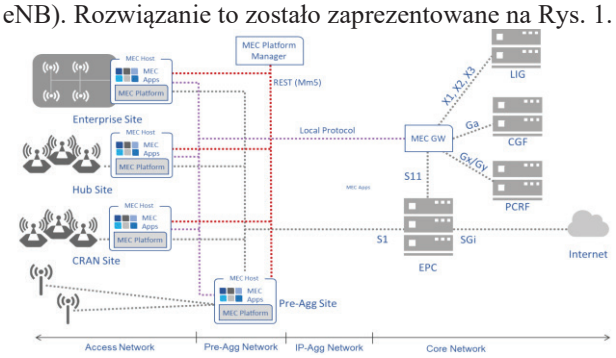
2. ARCHITEKTURA MODUŁU WSPÓŁPRACY Z SIECIAMI 3GPP

W ramach prac projektowych zaimplementowano oraz przetestowano moduł odpowiedzialny za współpracę z sieciami 3GPP. Moduł ten przeznaczony jest do uruchamiania w środowisku MEC, wykorzystując platformę OpenNESS [9]. Moduł oferuje możliwość odczytu wartości parametrów systemu 3GPP zgodnie z wymaganiami RNI, a następnie udostępnienia ich aplikacjom uruchamianym w systemie MEC.

Zarówno sposób działania serwera MEC, jak i sposób jego integracji z rdzeniem sieci 3GPP podlega ścisłej standaryzacji. Integracja serwera MEC może być jednak

realizowana na kilka sposobów, które zostały opracowane przez ETSI [2]. Opisane tam scenariusze integracji umożliwiają dostosowanie jego wdrożenia do potrzeb aplikacji (np. poprzez umiejscowienie go możliwie blisko użytkownika lub trasy przesyłania zagregowanego ruchu z wielu lokalizacji) lub też do wymogów narzucanych przez strukturę systemu (np. obecność systemów prywatnych 5G zintegrowanych z systemem ogólnokrajowym).

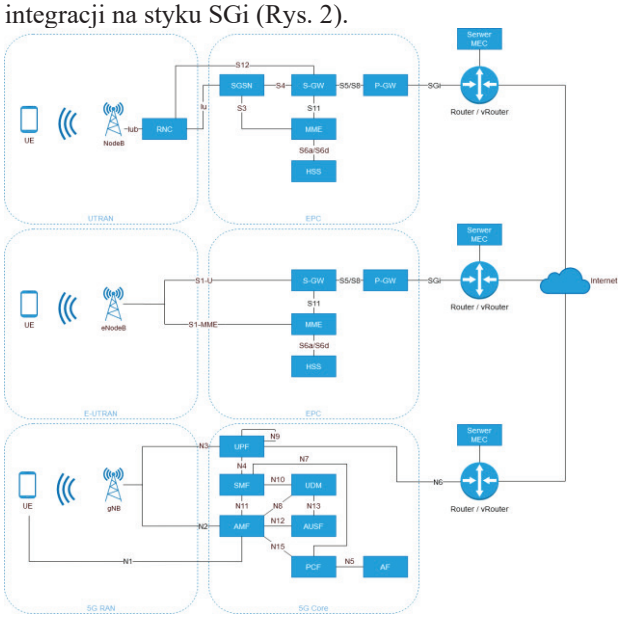
Jeden z prezentowanych scenariuszy nazwany „Bump in the wire" zakłada ścisłą integrację serwera MEC z elementami sieci dostępowej (np. platformami eNB). Rozwiązanie to zostało zaprezentowane na Rys. 1.



Rys. 1. Scenariusz „Bump in the wire" określający jeden ze sposobów wdrożenia serwera MEC [2]

Jest to scenariusz, który niesie korzyści wynikające z ulokowania serwera MEC bardzo blisko użytkownika końcowego, ale znacząco komplikuje sposób jego wdrożenia – np. poprzez konieczność dekapsulacji danych użytkownika na poziomie tuneli GTP-U czy wprowadzenia dodatkowej bramy MEC pozwalającej na inspekcję ruchu wychodzącego z sieci.

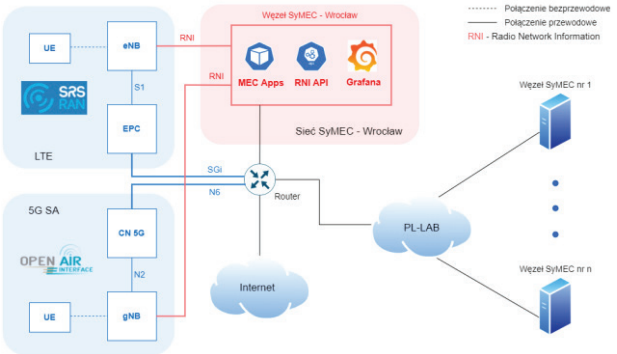
Odmiennym scenariuszem wdrożenia serwera MEC w systemach 3GPP opisanych w [2] jest scenariusz integracji na styku SGi (Rys. 2).



Rys. 2. Lokalizacja serwera MEC we współpracy z sieciami 3GPPza pomocą interfejsu SGi/N6

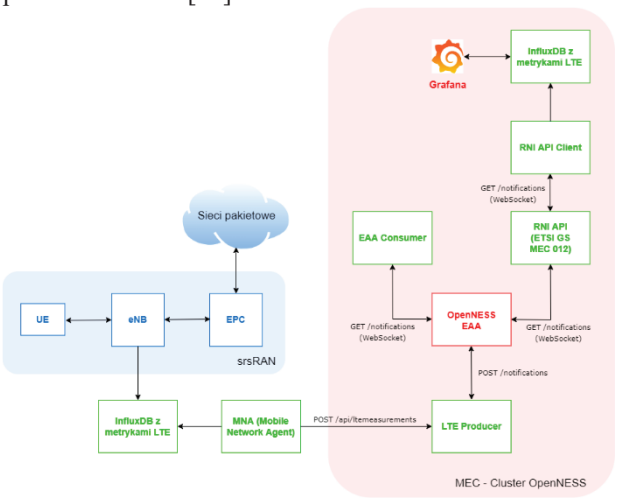
Scenariusz ten wymaga mniej zmian w sieci operatora, ponieważ standardowe elementy architektury 3GPP i interfejsy są wykorzystywane do zarządzania sesjami, naliczania opłat, itp. Ten typ wdrożenia wykorzystywany będzie szczególnie w scenariuszach uwzględniających zamknięty system oferujący dedykowane usługi np. Push to Talk czy komunikacji typu Maszyna-Maszyna (M2M). Jest on również rozwiązaniem, które może zostać zastosowane w przypadku platformy MEC będącej odrębnym produktem, co było przedmiotem prac w projekcie SyMEC [5]. Architektura rozwiązania została przedstawiona na Rys. 3. Zastosowano osobne interfejsy do sieci 4G (SGi) i 5G (N6) oraz dwa interfejsy przekazywania danych RNI (z sieci 4G i 5G). Jako rozwiązanie sieci 4G zastosowano system srsRAN [10] a w przypadku sieci 5G SA OpenAirInterface [11].

Informacja przekazywana z brzegu sieci 4G i 5G za pomocą RNI API/NR API przetwarzana jest i przekazywana do serwera MEC dostępnego w danym węźle sieci SyMEC.



Rys. 3. Architektura modułów współpracy z sieciami 3GPP

Moduły komunikują się przy wykorzystaniu komponentu Edge Application Agent (EAA), stanowiącego część platformy OpenNESS i działającego w oparciu o Kubernetes. Interfejs API EAA pełni rolę magistrali komunikacji między aplikacjami w modelu publish-subcribe [12].



Rys. 4. Architektura dostępu do sieci 4G/LTE

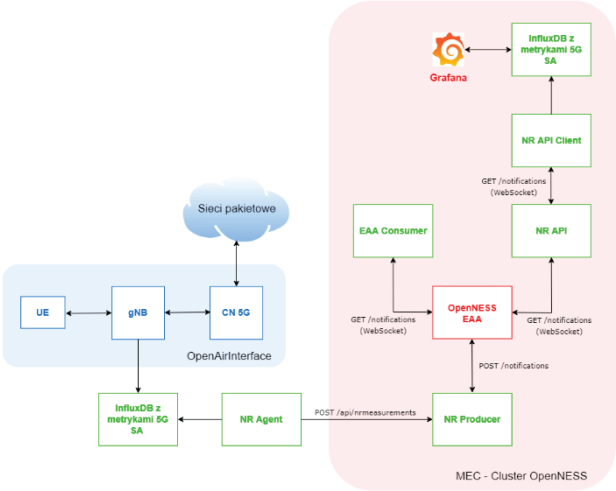
W przypadku systemu 4G (Rys. 4) opracowano podmoduły służące do przekazywania danych o stanie sieci, których podstawowe cechy podano w Tabeli 1.

Tab. 1. Komponenty systemu modułu dostępu do 4G/LTE

Komponent	Realizowane funkcje
InfluxDB z metrykami LTE	Baza danych przechowująca bieżące metryki sieci LTE

MNA (Mobile Network Agent)	Aplikacja odczytująca w interwałach czasowych metryki z bazy danych i przesyłająca je do LTEProducer
LTE Producer	Aplikacja udostępniająca punkt końcowy do przesyłania informacji o sieci LTE. Wysyła odebrane informacje o sieci LTE do OpenNESS EAA.
OpenNESS EAA	Edge Application Agent jest usługą, która pełni rolę magistrali komunikacji między aplikacjami w modelu publish-subcribe
EAA Consumer	Aplikacja subskrybuje EAA w celu uzyskania informacji o sieciach dostępowych.
RNI API	Udostępnia informacje o sieci LTE poprzez API aplikacją MEC
RNI API Client	Aplikacja odczytująca w interwałach czasowych aktualne wartości parametrów sieci LTE. Odczytane dane zapisywane są do bazy danych.
Grafana	Odpowiada za wizualizację informacji o stanie sieci LTE

Poniżej na Rys. 5. przedstawiono architekturę modułu dostępu do sieci 5G. Przepływ informacji odbywa



Rys. 5. Architektura dostępu do sieci 5G NR

się analogicznie jak dla sieci LTE poprzez wykorzystanie modułu Edge Application Agent (EAA) z OpenNESS.

Tab. 2. Komponenty systemu modułu dostępu do 5G NR

Komponent	Realizowane funkcje
InfluxDB z metrykami 5G (przy OAI)	Baza danych InfluxDB przechowująca bieżące informacje o sieci 5G
NR Agent	Aplikacja odczytująca w interwałach czasowych metryki z bazy danych i przesyłająca je do NR Producer
NR Producer	Aplikacja udostępniająca punkt końcowy do przesyłania informacji o sieci 5G. Wysyła odebrane informacje o sieci 5G do OpenNESS EAA
OpenNESS EAA	Edge Application Agent jest usługą, która pełni rolę magistrali komunikacji między aplikacjami w modelu publish-subcribe
EAA Consumer	Aplikacja subskrybuje EAA w celu uzyskania informacji o sieciach dostępowych.
NR API	Udostępnia informacje o sieci 5G poprzez API aplikacją MEC
NR API Client	Aplikacja odczytująca w interwałach czasowych aktualne wartości parametrów sieci 5G. Odczytane dane zapisywane są do bazy danych.
InfluxDB z metrykami 5G	Baza danych InfluxDB przechowująca informacje o sieci 5G na potrzeby wizualizacji metryk w aplikacji Grafana
Grafana	Odpowiada za wizualizację informacji o stanie sieci 5G

3. DEMONSTRACJA WSPÓŁPACY Z SIECIAMI 4G i 5G

W ramach integracji i testowania sieci 4G, jako implementację sieci wybrano oprogramowanie srsRAN oraz sprzęt USRP (Universal Software Radio Peripheral) B210 firmy Ettus Research [13] wraz z antenami zewnętrznymi. Jako UE wykorzystano PC z modemem Huawei i smartphone Xiaomi Redmi Note 7. Rozwiązanie srsRAN to darmowy pakiet oprogramowania radiowego typu open-source, umożliwiający realizację sieci telefonii komórkowej 4G LTE i 5G NR NSA w wersji SDR (Software Defined Radio) z wykorzystaniem urządzeń USRP. Kompleksowa sieć LTE w srsRAN składa się z trzech głównych komponentów:

- srsEPC, czyli implementacji kompletnej sieci szkieletowej EPC (Evolved Packet Core). Aplikacja srsEPC działa jako pojedynczy plik binarny i zapewnia usługi bazy danych użytkowników HSS (Home Subscriber Server), jednostki zarządzania mobilnością MME (Mobility Management Entity), głównej bramy użytkowników S-GW (Serving Gateway) oraz bramy pakietów P-GW (Packet Data Network Gateway),

- srsENB, czyli stacji bazowej eNodeB, która jako aplikacja łączy się z dowolną siecią szkieletową EPC i tworzy lokalną komórkę LTE. Do przesyłania i odbierania sygnałów drogą radiową wymagany jest sprzęt SDR, taki jak Ettus Research USRP,

- terminala użytkownika UE z odpowiednio zaprogramowaną kartą SIM.

W przypadku sieci 5G NR jako rozwiązanie sieci przyjęto oprogramowanie radiowe OpenAirInterface, również typu open-source. Pozwala ono na realizację sieci w trybie Stand Alone (SA) przez wdrożenie technologii 3GPP na sprzęcie SDR, takim jak urządzenie USRP. W projekcie wykorzystano urządzenie Ettus USRP B210, które zostało podłączone do komputera z systemem Linux za pomocą USB 3.0. Kompletna sieć 5G SA w OAI składa się z trzech komponentów:

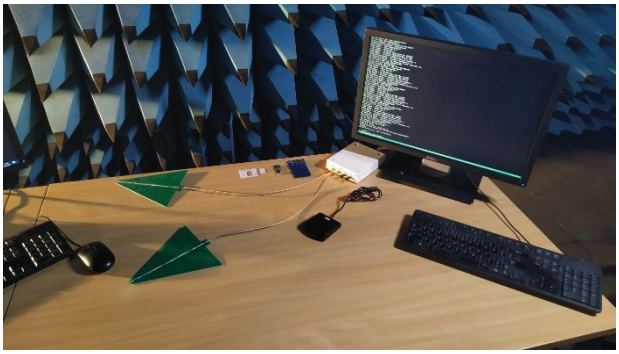
- oprogramowania OAI CN 5G, czyli realizacji sieci szkieletowej 5G,

- oprogramowania OAI gNB, czyli realizacji RAN sieci 5G,

- terminala użytkownika z zaprogramowaną kartą SIM.

Jako UE w sieci 5G SA wykorzystano moduł 5G RM500Q-GL firmy Quectel z osprzętem. Dla powyższego terminala użytkownika użyto dedykowaną kartę SIM dla OAI firmy Open-Cells i kompatybilny programator uicc.

Kompletne rozwiązania sieciowe zostały zaimplementowane i zintegrowane w środowisku SyMEC. Dokonano koniecznej integracji, przetwarzania, certyfikacji i wysyłania danych z OpenNESS EAA 2020.12. Usługi w celu nawiązania połączenia z EAA muszą przeprowadzić wzajemne uwierzytelnianie TLS. Aplikacje komunikujące się z EAA wykorzystują certyfikat wygenerowany za pomocą Certificate Signer. CSR po przesłaniu musi zostać zatwierdzony przez administratora Kubernetes. Zaimplementowane moduły w pełni obsługują powyższe mechanizmy, wprowadzone w platformie OpenNESS od wersji 2020.12 w postaci Edge Application API i Edge Application Authentication API. Na Rys. 5 i 6 przedstawiono fizycznie zintegrowane stanowisko sieci 4G i 5G, wykonane, uruchomione i przetestowane zgodnie z przedstawioną wyżej architekturą. Ze względu na wykorzystywanie do emisji częstotliwości w pasmach licencjonowanych (2,5 GHz w przypadku 4G oraz 3,6 GHz w przypadku 5G NR) uruchomienie i emisja prowadzona była w komorze bezodbiciowej.



Rys. 5 Stanowisko demonstracyjne dla sieci 4G

Rys. 6. Stanowisko demonstracyjne dla sieci 5G NR

Informacje o stanie parametrów sieciowych na brzegu sieci LTE i 5G SA, które zbierano w trakcie trwania transmisji na stanowisku demonstracyjnym zostały zwizualizowane w narzędziu do monitorowania Grafana. Integracja sieci dostępowych 3GPP z systemem SyMEC została zweryfikowana i zaprezentowana poprzez uruchomienie aplikacji MEC na terminalu użytkownika.

4. PODSUMOWANIE

W artykule zaprezentowano sposób integracji systemu MEC z sieciami 3GPP: 4G i 5G. Opisano architekturę tego modułu oraz sposób integracji z innymi komponentami w ramach systemu SyMEC. Dzięki dostępowi do bieżących informacji o stanie sieci radiowej 3GPP mierzonymi na brzegu tej sieci możliwe jest dynamiczne zarządzanie ruchem i wykorzystywanie ich do realizacji usług istotnych z punktu widzenia użytkownika znajdującego się blisko danej stacji bazowej. Uruchomienie modułów zostało zrealizowane w ramach projektu SyMEC w sieci PL-LAB i może być wykorzystywane do dalszych badań i rozwoju aplikacji i usług MEC realizowanych na brzegu sieci.

5. PODZIĘKOWANIE

Badania te były prowadzone w ramach projektu POIR 04.01.02.00-00-0107/17,pt. „System MEC dla wspierania zaawansowanych aplikacji w środowisku sieci przewodowych i bezprzewodowych 3G/4G/5G”, współfinansowanego w ramach Programu Operacyjnego Innowacyjny Rozwój na lata 2014-2020.

LITERATURA

[1] ETSI GS MEC 003: Multi-access Edge Computing; Framework and Reference Architecture, 2022
[2] ETSI, MEC Deployments in 4G and Evolution Towards 5G. 2018. White Paper No. 24.
[3] ETSI GS MEC 012 V2.1.1 (2019-12).
[4] A. Bęben, W. Burakowski, B. Belter, D. Parniewicz: Architektura systemu MEC dla wspierania zaawansowanych aplikacji w środowisku sieci przewodowych i bezprzewodowych 3G/4G/5G. Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, 2020, Nr 7-8, s.144-148
[5] Strona internetowa projektu SyMEC symec.com.pl
[6] M. Hoeft, D. Więcek i in.: Moduł platformy SyMEC odpowiedzialny za współpracę serwera MEC z sieciami komórkowymi i bezprzewodowymi sieciami dostępowymi WLAN. Przegląd Telekomunikacyjny+Wiadomości Telekomunikacyjne, 2020, Nr 7-8, s.149-154.
[7] W. U. Rahman, C. S. Hong and E. Huh, "Edge Computing Assisted Joint Quality Adaptation for Mobile Video Streaming," in *IEEE Access*, vol. 7, pp. 129082-129094, 2019
[8] Grafana: The open observability platform: https://grafana.com/
[9] OpenNESS - Open Network Edge Services Software: https://www.openness.org/
[10] https://www.srslte.com/
[11] https://openairinterface.org/
[12] Edge Application Agent: https://github.com/smart-edge-open/specs/blob/master/doc/applications-onboard/smartedge-open-eaa.md
[13] https://www.ettus.com/all-products/ub210-kit/

Aplikacje i usługi systemu SyMEC

Applications and services of SyMEC system

Tomasz Łysiak, Jakub Kabat

Instytut Chemii Bioorganicznej Polskiej Akademii Nauk, Poznańskie Centrum Superkomputerowo-Sieciowe
jkabat@man.poznan.pl, tlysiak@man.poznan.pl

DOI: 10.15199/59.2022.4.53

Streszczenie: **W artykule scharakteryzowano aplikacje, które były testowane w systemie SyMEC. Opisano, w jaki sposób dostosowano te aplikacje, aby można je było uruchomić w systemie. Porównano działanie tych aplikacji ze scenariuszem, gdyby one były udostępniane w środowisku chmurowym. Ponadto, w artykule opisano opracowany w projekcie portal obsługi użytkowników.**

Abstract: **The article describes tested applications as well as how to adjust them to launching in the system. These applications were compared with applications that run in the cloud. Moreover, the paper describes portal, which was developed for user support.**

Słowa kluczowe: **Aplikacje, usługi SyMEC, portal obsługi użytkowników**

Keywords: **applications, SyMEC services, users portal,**

1. WSTĘP

Technika MEC (*Multi-access Edge Computing*) [5] stanowi jeden z kluczowych elementów w sieciach nowych generacji, w tym w sieciach 5G. Zastosowanie tej techniki pozwala na istotne zmniejszenie opóźnień w przekazie danych, co zapewnia dzięki zastosowaniu brzegowych serwerów obliczeniowych zlokalizowanych blisko użytkowników końcowych. Dzięki takiemu rozwiązaniu część obliczeń realizowanych do tej pory w chmurze, może być udostępniana na serwerach brzegowych. Operacje obliczeniowe realizowane na terminalach użytkowników, również mogą być wykonywane na serwerach MEC. Dodatkową zaletą tej techniki jest zmniejszenie ruchu w sieci szkieletowej.

Z punktu widzenia operatora sieci, zastosowanie techniki MEC pozwala na sterowanie ruchem w sieci, a ponad to może udostępniać użytkownikom lokalne informacje o parametrach sieci oraz samemu korzystać z informacji udostępnionych przez użytkowników. Informacje takie jak lokalizacja użytkownika lub aktualne obciążenie sieci, mogą zostać wykorzystane przez aplikacje do dostosowania swojego działania do aktualnie panujących warunków ruchowych.

Połączenie techniki MEC z sieciami nowych generacji, pozwala na tworzenie nowych aplikacji i usług, korzystających z bardzo niskich opóźnień i dodatkowych informacji o stanie sieci. Aplikacje wirtualnej/rozszerzonej rzeczywistości, strumieniowania wideo lub gry wymagające bardzo niskich opóźnień w pełni korzystają z zalet systemu MEC. Aplikacje przetwarzające i generujące duże ilości danych np. monitoring wizyjny, mogą zostać w pełni przeniesione do MEC, co znacznie odciąży ruch w sieci szkieletowej.

System SyMEC jest platformą umożliwiającą uruchamianie i zarządzanie wieloma instancjami aplikacji przygotowanych do korzystania przez użytkowników systemu. Jednakże, każda z tych aplikacji musi zostać odpowiednio dostosowana do prawidłowego działania w systemie SyMEC. Typ oraz kategoria uruchamianej aplikacji nie mają znaczenia, jeżeli ta aplikacja jest zgodna z najnowszymi standardami ETSI [4]. Aktualnie istnieje bardzo duża baza aplikacji i część z nich można łatwo przystosować do działania w systemie SyMEC. Aplikacjami, które najłatwiej dostosować są aplikacje chmurowe, które nie potrzebują wielu zmian, aby prawidłowo działać również na platformie MEC.

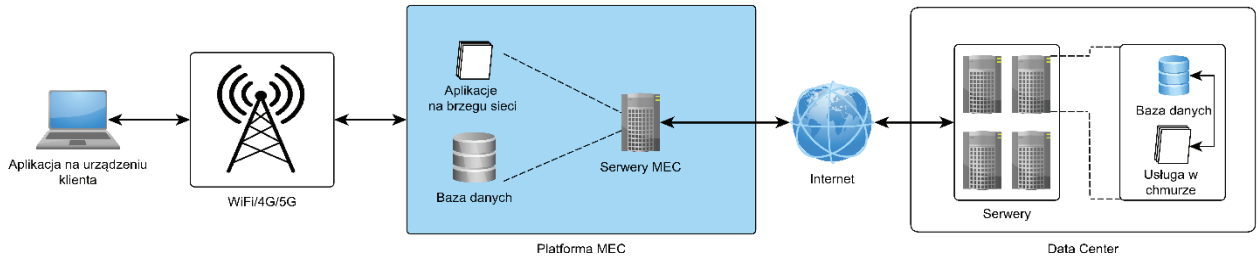
W systemie SyMEC aplikacje i usługi uruchamia się za pomocą portalu obsługi użytkowników systemu. Portal ten pozwala na dodanie pakietów aplikacji i uruchamianie ich instancji i stanowi on główny interfejs pomiędzy użytkownikiem a systemem.

W kolejnych rozdziałach przedstawiono strukturę i zalety aplikacji uruchamianych w systemie SyMEC oraz to jak prawidłowo dostosować je do uruchomienia w systemie. Ostatni rozdział opisuje budowę oraz funkcjonalności portal użytkownika.

2. Aplikacje w systemie SyMEC

Wykorzystując technologię MEC umożliwiającą budowanie nowych aplikacji, które wykorzystają przetwarzanie danych blisko użytkownika końcowego na brzegu sieci znacznie zmniejszamy opóźnienia spowodowane przesyłaniem danych oraz możliwe jest skorzystanie z dodatkowych parametrów, np. informacji o stanie sieci dostępowej.

Przykładowymi usługami świadczonymi w systemie MEC mogą być różne aplikacje z kategorii m.in. takich jak: narzędzia do rozwijania rozwiązań sztucznej inteligencji, aplikacje sztucznej inteligencji, bazy danych, strumieniowanie wideo, narzędzia do monitorowania,



Rys. 1. Schemat aplikacji MEC podzielonej na trzy części

aplikacje zwiększające bezpieczeństwo w sieci, gry, przechowywanie danych w chmurze itp.

Aplikacje MEC mogą zostać podzielone na trzy części: aplikację kliencką, aplikację na brzegu sieci oraz serwis w chmurze. Przykładowo: aplikacja sztucznej inteligencji może wysyłać tylko dane wymagające bardzo dużych zasobów obliczeniowych do centrum danych, a większość obliczeń i wstępna obróbka danych może być przeprowadzana na serwerze brzegowym. Schemat aplikacji podzielonej na trzy części przedstawiono na rysunku 1.

Taka architektura aplikacji w pełni korzysta z zalet zarówno systemu MEC, jak i rozwiązań chmurowych. Obliczenia krytyczne czasowo są wykonywane na brzegu sieci, a obliczenia wymagające dużej mocy obliczeniowej zlecane są do serwerów w chmurze. Dodatkowo każda z aplikacji może wykorzystać dedykowane interfejsy API, aby uzyskać informacje o liczbie użytkowników danej sieci bezprzewodowej, wykorzystaniu kanału, poziomie zakłóceń. Informacje te mogą zostać wykorzystane w procesie lokowania usługi lub szacowania jej jakości i dostosowywania do panujących w sieci warunków. Dodatkową zaletą aplikacji uruchomionej w systemie MEC jest możliwość zarządzania zasobami chmury brzegowej i optymalny dobór lokalizacji oraz parametrów serwera z perspektywy użytkownika końcowego.

3. Dostosowanie aplikacji do systemu SyMEC

System SyMEC wprowadza szereg wymagań, które muszą zostać spełnione przez aplikację. Aplikacja musi m.in. udostępniać interfejs do komunikacji sieciowej np. w postaci REST API. Dodatkowo musi istnieć w formie skonteneryzowanej na platformie Docker [3], ponieważ system SyMEC w tym formacie przechowuje wszystkie aplikacje. Standard ETSI wymaga od aplikacji również szeregu parametrów, które muszą zostać opisane w deskryptorze aplikacji.

Deskryptor aplikacji jest wymagany od każdego dostawcy aplikacji i powinien zawierać wszystkie ustandaryzowane informacje o aplikacji. Deskryptor, który jest używany przez system SyMEC do uruchomienia aplikacji musi być zapisany w formacie JSON (ang. *JavaScript Object Notation*). W deskryptorze powinny znaleźć się informacje m.in. o nazwie i dostawcy aplikacji, jej opis, wymagane zasoby obliczeniowe, porty do komunikacji, zmienne środowiskowe i informacje o zasobach dyskowych. W tabeli 1 przedstawiono opisy najważniejszych pól deskryptora.

Tab. 1. Opis pól deskryptora

Nazwa	*Opis*
appDId	Identyfikator deskryptora
appName	Nazwa aplikacji
appProvider	Dostawca aplikacji
appSoftVersion	Wersja aplikacji
appDVersion	Wersja deskryptora aplikacji
appInfoName	Czytelna nazwa aplikacji
appDescription	Opis aplikacji
virtualComputeDescriptor	Opis wymaganych zasobów obliczeniowych
swImageDescriptor	Opis obrazu aplikacji
appDNSRules	Wpisy DNS aplikacji
appClientFacingPorts	Porty aplikacji przeznaczone do komunikacji z urządzeniem użytkownika
appEnvironmentalVariables	Opis zmiennych środowiskowych wymaganych przez aplikację
appPersistantVolumes	Opis stałych wolumenów

Dostawca aplikacji powinien także w dość dokładny sposób określić ilość wymaganych zasobów obliczeniowych, ponieważ system SyMEC przydziela tylko określoną część zasobów dla danej aplikacji. Dlatego, dla każdej aplikacji powinny zostać przeprowadzone testy wydajnościowe, które określą

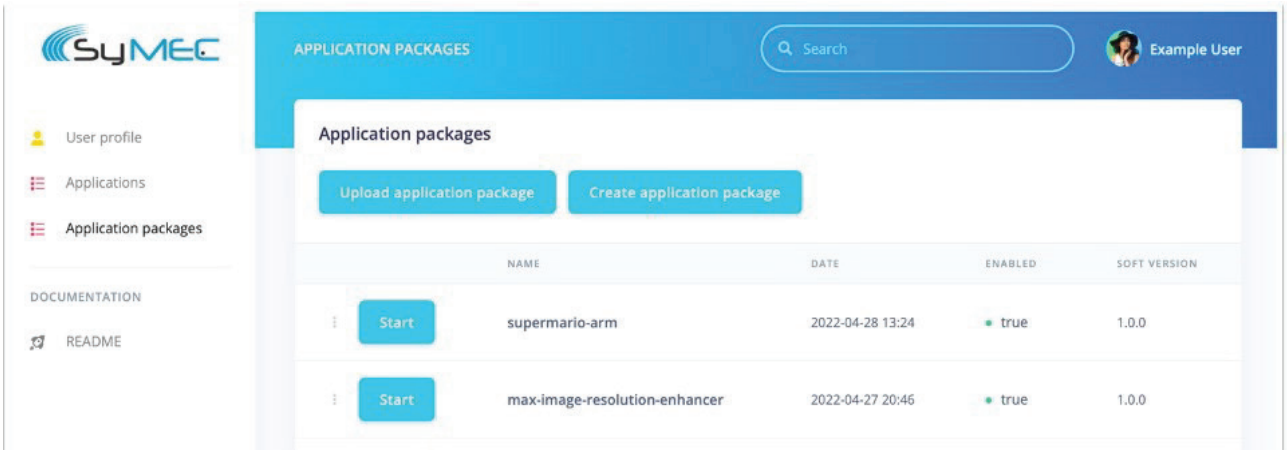
maksymalne zużycie CPU i RAM. Bez dokładnego określenia tych parametrów, w przypadku dużego obciążenia aplikacji, może ona nie działać prawidłowo.

W środowisku SyMEC współpracują serwery oparte o architekturę amd64 i arm64, z tego powodu w deskryptorze znajdują się także informacje o wspieranej przez aplikacje architekturze procesora. Określenie tego parametru wskazuje na jakim serwerze może zostać uruchomiona dana aplikacja. Jeżeli aplikacja wspiera arm64 będzie uruchamiana tylko na serwerach o tej samej architekturze. Żeby w pełni wykorzystać możliwości systemu SyMEC, dostawca aplikacji może przygotować wersje aplikacji dla obu platform. Dla każdej z nich należy przygotować osobny deskryptor.

4. Portal SyMEC

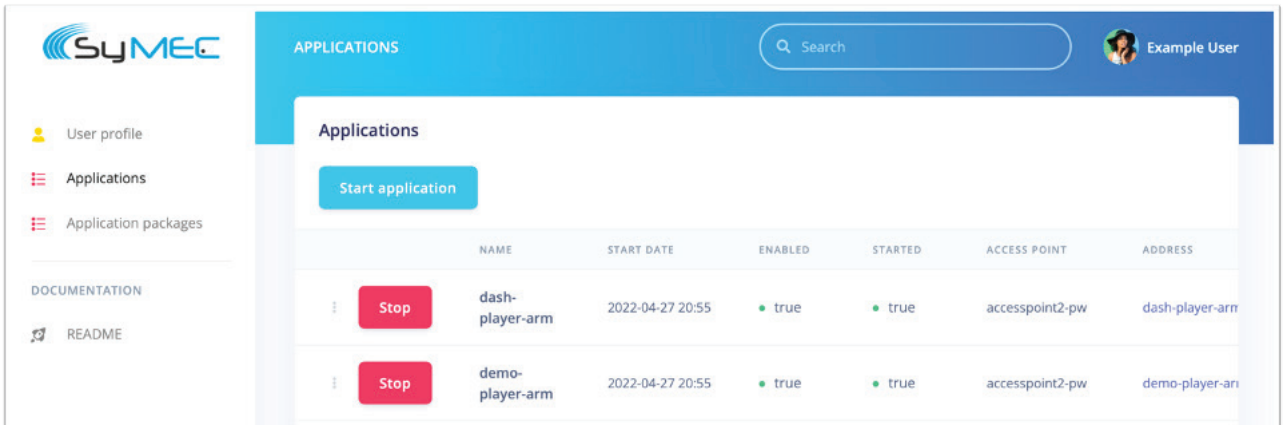
Aby móc w pełni zademonstrować system SyMEC, w projekcie zdecydowano się na implementację niezbędnego oraz znacząco uproszonego zestawu funkcjonalności systemu OSS/BSS (ang. *Operational Support Systems, Business Support Systems*), umożliwiającego prezentację usług MEC w portalu użytkownika oraz zarządzenie cyklem życia instancji tych usług. Portal jest aplikacją typu SPA (ang. *Single Page Application*), która umożliwia interakcję z systemem SyMEC w oparciu o graficzny interfejs użytkownika. Serwis ten udostępnia funkcjonalności systemu SyMEC odpowiedzialne za zarządzanie pakietami oraz instancjami aplikacji MEC. Fundamentalnymi narzędziami wykorzystanymi w procesie opracowywania portalu SyMEC były platformy Angular [1] oraz Bootstrap [2], a także zestaw bazujących na nich szablonów oraz wzorców dostępnych na licencji MIT.

Portal SyMEC wspiera zarówno ładowanie pakietu aplikacji MEC w oparciu o wczytanie deskryptora z pliku w formacie JSON lub YAML, jak i dodanie pakietu aplikacji MEC za pośrednictwem usługi ręcznego definiowania deskryptora. Użytkownik korzystający z aplikacji webowej może również wykorzystać funkcjonalność umożliwiającą usuwanie pakietów aplikacji MEC z systemu SyMEC. Ponadto interfejs graficzny użytkownika pozwala na wyświetlanie listy dostępnych pakietów aplikacji MEC, a także na uzyskanie szczegółowych informacji oraz deskryptorów dotyczących wybranych pakietów aplikacji MEC. Na rysunku 2 znajduje się widok pakietów aplikacji umożliwiający wykorzystanie wszystkich ww. funkcjonalności dotyczących obsługi procesów związanych z zarządzaniem pakietami aplikacji MEC w systemie SyMEC.



Rys. **1**. *Widok pakietów aplikacji*

Portal SyMEC wspiera zarówno uruchamianie instancji aplikacji MEC z domyślnie skonfigurowanymi parametrami deskryptora aplikacji, jak i uruchamianie instancji aplikacji MEC przy wykorzystaniu funkcjonalności umożliwiającej ręczne skonfigurowanie parametrów deskryptora aplikacji MEC. Użytkownik korzystający z aplikacji webowej może również wykorzystać funkcjonalność umożliwiającą usuwanie instancji aplikacji MEC z systemu SyMEC. Ponadto interfejs graficzny użytkownika pozwala na wyświetlanie listy dostępnych instancji aplikacji MEC, a także na uzyskanie szczegółowych informacji dotyczących wybranych instancji aplikacji MEC. Na rysunku 3 znajduje się widok instancji aplikacji umożliwiający wykorzystanie wszystkich ww. funkcjonalności dotyczących obsługi procesów związanych z zarządzaniem instancjami aplikacji MEC w systemie SyMEC.



Rys. **2**. Widok instancji aplikacji

5. Podsumowanie

Aktualnie na rynku znajduje się wiele aplikacji, lecz wciąż nie ma rozwiązań w pełni nadających się do uruchomienia w systemie MEC. Jednak część z dostępnych aplikacji można łatwo dostosować do działania w systemie SyMEC, ale mimo wszystko nadal nie wykorzystują one w pełni zalet systemu.

Z powodu małej ilości aplikacji w pełni dostosowanych do działania na brzegu sieci, trzeba zachęcać deweloperów do tworzenia takich aplikacji. Można to osiągnąć poprzez rozwój systemów MEC i systemów pozwalających w łatwy sposób zarządzać nimi. System SyMEC jest dużym krokiem w tym kierunku, ponieważ zapewnia kompletne środowisko do uruchamiania i zarządzania aplikacjami oraz usługami. W SyMEC możemy uruchomić każdą aplikację niezależnie od tego jakie funkcje oferuje. Wystarczy dostosować ją w odpowiedni sposób, według wytycznych. Wiele z aplikacji aktualnie działających w chmurze mogłoby być z łatwością przeniesionych do systemu MEC.

Przygotowany na potrzeby projektu portal SyMEC ułatwia zarządzanie instancjami aplikacji i jest łatwy w użyciu, pozwalając niezaznajomionemu z systemem użytkownikowi, korzystać z niego w przystępny sposób. Mając przygotowaną aplikację, wystarczy załadować jej deskryptor na portalu SyMEC i uruchomić ją po zakończeniu po zakończeniu pobierania obrazu aplikacji.

Badania, których wyniki są przedstawione w tym artykule, były prowadzone w ramach projektu POIR 04.01.02.00-00-0107/17,pt. „System MEC dla wspierania zaawansowanych aplikacji w środowisku sieci przewodowych i bezprzewodowych 3G/4G/5G”, współfinansowanego w ramach Programu Operacyjnego Innowacyjny Rozwój na lata 2014-2020.

LITERATURA

[1] Angular, https://angular.io [data uzyskania dostępu: 28.06.2022]
[2] Bootstrap, https://getbootstrap.com [data uzyskania dostępu: 28.06.2022]
[3] Docker, https://www.docker.com [data uzyskania dostępu: 28.06.2022]
[4] ETSI standard, „Multi-access Edge Computing (MEC); MEC Management; Part 2: Application lifecycle, rules and requirements management", ETSI GS MEC 0 010-2, v2.1.1,” November 2019.
[5] Multi-access Edge Computing (MEC), https://www.etsi.org/technologies/multi-access-edge-computing [data uzyskania dostępu: 27.06.2022]

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

AUTOMATYCZNY DOBÓR KĄTA POCHYLENIA W SYSTEMACH KOMÓRKOWYCH

AUTOMATIC ANTENNA TILT SELECTION IN CELLULAR NETWORKS

Małgorzata Michalska[1]; Adrian Kliks[2]

[1] Politechnika Poznańska, malgorzatamichalska97@wp.pl
[2] Politechnika Poznańska, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2022.4.54

Streszczenie: **W pracy zaproponowano algorytm automatycznego doboru kąta pochylenia anten stosowanych w systemach komórkowych oraz oceniono jego działanie. Wypracowano sposób doboru pochylenia anten z uwzględnieniem stanu środowiska, pokrycia sygnałem radiowym, pojemności sieci oraz interferencji międzykomórkowych. Zebrane wyniki pozwoliły na oszacowanie istotności ustawienia pochylenia anten, aby zapewnić odpowiednią jakość usług.**

Abstract: **The purpose of this thesis is to propose an algorithm for the automatic antenna tilt selection in cellular networks and evaluation of its performance. A method of adjusting the antenna inclination angle was developed, taking into account the state of the environment, coverage, system capacity, and inter-cell interference. Collected results allowed for conducting the importance of the proper setting of the antenna inclination angle to ensure appropriate quality of service.**

Słowa kluczowe: **automatyczny dobór kąta, pochylenie anten, uczenie maszynowe, uczenie ze wzmocnieniem, zarządzanie interferencją**

Keywords: **automatic antenna tilt selection, antenna tilt, interference management, machine learning, reinforcement learning**

1. WSTĘP

Telefonia komórkowa na przestrzeni około czterdziestu ostatnich lat stała się jednym z najczęściej wybieranych form komunikacji na świecie. Operatorzy sieci komórkowych dążą do ich efektywnego działania oraz zapewnienia odpowiednich wskaźników *QoE* (ang. *Quality of Experience* – jakość doświadczenia). Wraz z powstawaniem kolejnych generacji sieci komórkowe stawały się coraz bardziej złożone, a jednocześnie wymagania dotyczące jakości usług również wciąż rosną. W miarę rozwoju technologii sieci komórkowych oraz wymagań klientów, wydajność sieci powinna dać się odpowiednio kalibrować tak, aby dopasować ją do zadanych wymagań, uwzględniając ciągłe i dynamiczne zmiany w środowisku sieciowym i radiowym. Przykładami zmiennych czynników mogą być chwilowe lub okresowe fluktuacje obciążenia komórek ruchem sieciowym, ale także wspomniany już trend ciągłego, wykładniczy wzrost ruchu sieciowego na przestrzeni lat oraz przewidywany dalszy wzrost w kolejnych latach. Ponadto należy uwzględnić zmienne warunki propagacji radiowej, pojawiające się awarie, wyłączenia i integracje stacji bazowych, a także zakłócenia wywołane przez zewnętrzne niepożądane źródła.

Kluczowe elementy wpływające na efektywne działanie sieci to zasięg transmisji oraz uzyskiwana przepustowość kanału. Jednym z głównych czynników wpływających na te elementy jest kąt nachylenia wiązki nadawczej anteny. Dzięki odpowiedniemu pochyleniu wiązki stacji bazowej można uniknąć zakłóceń wytworzonych przez promieniowanie pochodzące z innych, sąsiednich komórek korzystających z tych samych pasm częstotliwości. Dodatkowo możliwa jest poprawa wydajności sieci poprzez dostrojenie pokrycia oraz pojemności.

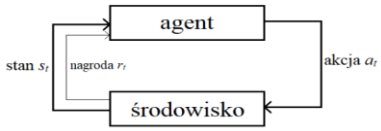
Dla operatorów sieci komórkowych ręczna optymalizacja wydajności jest bardzo złożonym, kosztownym i czasochłonnym zadaniem. Aby taka sieć stałą się bardziej autonomiczna, międzynarodowa organizacja standaryzacyjna *3GPP* (*3rd Generation Parentship Project*) wprowadziła koncepcję *SON* (*Self-Organizing Network* – sieć samoorganizująca się) [1], która jest technologią automatyzacji w celu łatwiejszej i szybszej konfiguracji, optymalizacji i naprawy sieci komórkowych. Kategoria optymalizacji obejmuje m.in. funkcje poprawy zasięgu oraz pojemności sieci *CCO (Coverage and Capacity Optimization* – optymalizacja pokrycia oraz pojemności) [2], [3]. Polegają one na zapewnieniu wystarczającego zasięgu na danym obszarze tak, aby osiągnąć ciągłe pokrycie sygnałem na tyle silnym, że obserwowana wartość *SINR* (ang. *Signal to Interference Plus Noise Ratio* – stosunek sygnału do zakłóceń plus szum) po stronie odbiorczej jest równa lub większa niż wymagane minimum. Dodatkowo funkcje powinny zapewnić wystarczającą przepustowość sieci na danym obszarze przy minimalnych zasobach radiowych. W zależności od aktualnego środowiska sieciowego, można uzyskać najlepszy kompromis między zasięgiem a przepustowością sieci, których priorytet ustala operator.

Jedną z możliwości uzyskania takiego kompromisu jest dostosowanie kąta pochylenia anten w zależności od sytuacji panującej w sieci. W ogólności można wyróżnić dwie główne klasy rozwiązań – mechaniczne (fizyczna zmiana pochylenia całego modułu antenowego) oraz elektryczne (opierające się na modyfikacji samego nadawanego sygnału) [4]. W artykule rozważamy ten drugi przypadek, pozwalający na dynamiczną zmianę kata nachylenia anteny. W szczególności, bazując na [5], w pracy rozważono optymalizację pokrycia oraz pojemności sieci mobilnej za pomocą elektrycznego sterowania kątem pochylenia wiązki nadawanej poszczególnych komórek. Skupiono się wyłącznie na łączu w dół, a jako główne narzędzie optymalizacyjne wykorzystano algorytmy sztucznej inteligencji uczenia się ze wzmocnieniem.

2. PROPOZYCJA WYKORZYSTANIA UCZENIA MASZYNOWEGO

2.1. Uczenie ze wzmocnieniem

Proponowane rozwiązanie korzysta z koncepcji uczenia się ze wzmocnieniem, które składa się z dwóch komponentów: środowiska oraz agenta. Podstawą uczenia się ze wzmocnieniem są dynamiczne akcje, które polegają na obserwacji przez agenta stanów środowiska w kolejnych krokach czasowych, na podstawie których wykonywane są odpowiednie instrukcje. Jeżeli agent wykona akcję a_t w kroku t, spowoduje to, że środowisko przejdzie w stan s_t. Następnie stan s_t zostanie odwzorowany w nagrodzie r_t. Takie działanie przedstawiono na rysunku 1.

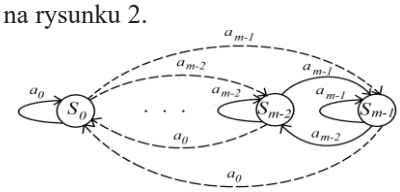


Rys. 1. Struktura działania akcji w kroku czasowym t

Dzięki komunikacji między stacjami bazowymi (np. poprzez interfejs X2) możliwa jest optymalizacja kątów anten nadawczych w klastrze komórek. Zakłada się, że komórki inne niż w aktualnym klastrze sieci są zbyt daleko, aby mogły powodować znaczne zakłócenia, a komórki w jednym klastrze na tyle blisko, że zmiana kąta nachylenia ich anten nadawczych znacząco wpływa na interferencje międzykomórkowe. Dlatego każda komórka w klastrze jest optymalizowana osobno, natomiast wszystkie klastry sieci mogą być optymalizowane w tym samym czasie. Założono, że agent to algorytm umieszczony na każdej komórce w klastrze sieci. Jego zadaniem jest obserwacja środowiska (sieci) poprzez gromadzone pomiary: rozmieszczenia *UE* (ang. *User's Equipment* – sprzęt użytkownika), kątów nachylenia anten nadawczych, mocy odbieranej przez *UE* ze wszystkich anten nadawczych oraz interferencje międzykomórkowe. Agent działa na środowisko, a następnie uczy się poprzez obserwację zmian zachodzących w środowisku. Z biegiem czasu, dzięki danym zebranym w macierzach wzmocnienia, agent nauczy się wykonywać właściwą akcję w określonym stanie środowiska.

2.2. Dobór kątów

Zasadę działania odbierania kąta pochylenia anten można przedstawić za pomocą decyzyjnych procesów Markova, *MDP* (*Markov Decision Process*). Bieżący stan jest mapowany do aktualnego ustawienia kątów w klastrze, a przejście między różnymi stanami oznacza, że konfiguracja kątów uległa zmianie. Przyjmijmy, że zbiór kątów, jakie można ustawić na danej antenie to: $\theta = \{\theta_0, \theta_1, \theta_2, \theta_3\}$. W związku z tym, stanami w zbiorze $S = \{s_0, \ldots, s_{m-2}, s_{m-1}\}$ jest m kombinacji z powtórzeniami zbioru θ. Każde przejście do odpowiedniego stanu zdefiniowane jest przez akcje $A = \{a_0, \ldots, a_{m-2}, a_{m-1}\}$. Takie działanie stan-akcja przedstawiono w postaci grafu – automatu przejść – na rysunku 2.



Rys. 2. Graf stan-akcja

Za każdą wykonaną funkcję agent otrzymuje nagrodę, która opisana jest ogólnym wzorem:

$$R = \mu \cdot S + \eta \cdot U \tag{1}$$

gdzie S – suma szybkości transmisji danych wszystkich użytkowników należących do komórki, znormalizowana względem maksymalnej możliwej sumy szybkości transmisji danych w tej komórce; μ – współczynnik, który określa priorytet przepustowości; U – liczba użytkowników należących do komórki, dla których moc odbierana jest wystarczająco wysoka, znormalizowana względem liczby wszystkich użytkowników należących do tej komórki; η – współczynnik, który określa priorytet zadowolenia z mocy odbieranej sygnału. Jeżeli $\mu > \eta$, to optymalizacja ma na celu osiągnięcie wyższej przepustowości dla mniejszej liczby użytkowników. Natomiast jeżeli $\mu < \eta$, to większa liczba użytkowników zostanie objęta przez komórkę, kosztem średniej przepustowości. Współczynniki μ oraz η ustala operator. Ponadto zmienną S można zdefiniować wzorem:

$$S = \frac{B \cdot \Delta f \cdot \sum_{i=1}^{N} \quad log_2(1+SINR_i)}{B \cdot \Delta f \cdot \sum_{i=1}^{N} \quad log_2(1+SNR_i)} \tag{2}$$

gdzie B – liczba podnośnych wykorzystanych przy transmisji danych; Δf - szerokość wykorzystywanego pasma jednej podnośnej; N – liczba użytkowników należących do komórki. $SINR_i$ oznacza stosunek sygnału pożądanego do zakłóceń plus szum, a SNR_i – stosunek sygnału pożądanego do szumu, zdefiniowane jako:

$$SINR_i = \frac{P_{ci}}{\sum_{j=1}^{M} \quad P_j + N_0} \tag{3}$$

$$SNR_i = \frac{P_{ci}}{N_0} \tag{4}$$

gdzie $N_0 = -174 + 10\log(\Delta f)$, M to liczba pozostałych komórek w klastrze; P_{ci} – moc sygnału docierająca do *UE* z bieżącej komórki; P_j – moc sygnału z pozostałych komórek klastra, który dociera do *UE* należącego do bieżącej komórki. Do obliczeń tłumienia ścieżki brano pod uwagę trójwymiarową charakterystykę wybranej anteny sektorowej zastosowaną w stacji bazowej oraz model propagacyjny – Okumura-Hata.

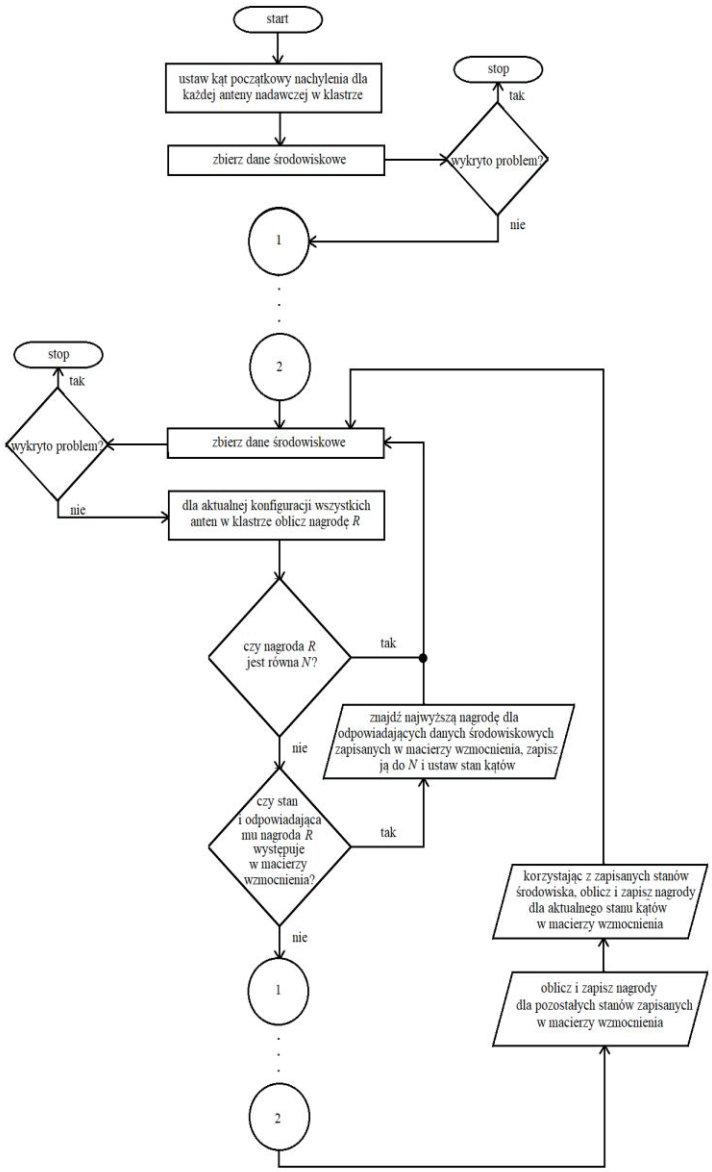
2.3. Działanie algorytmu

Całość działania zilustrowano na schematach – rysunki 3 oraz 4. Założono możliwość sterowania kątem anteny jedynie w sposób elektryczny. Każdy krok czasowy rozpoczyna się zebraniem danych środowiskowych – rozmieszczenie *UE* w pobliżu stacji bazowych, moce odbierane oraz *SINR*. Po dobraniu odpowiednich komórek sieci do klastra, można rozpocząć działanie algorytmu.

Pierwszym zadaniem algorytmu jest ustawienie kątów początkowych anten nadawczych w klastrze. Drugi etap wykonywany jest tylko raz. Rozpoczyna się zebraniem danych środowiskowych. Ich obecność jest niezbędna do obliczenia nagrody, dlatego, jeżeli wystąpi awaria systemu lub błąd podczas gromadzenia danych, należy zakończyć działanie algorytmu. Ważna jest znajomość wszystkich wartości pochylenia, jakie bieżąca antena może zaoferować. Z każdą iteracją d ustawiany jest inny kąt anteny o numerze b. Dla każdego stanu kątów anten w klastrze, wyliczana jest nagroda R i zapisywana w macierzy nagrody. W trakcie działania algorytmu nie można wykluczyć możliwości wystąpienia awarii toru antenowego lub innych komponentów stacji bazowej,

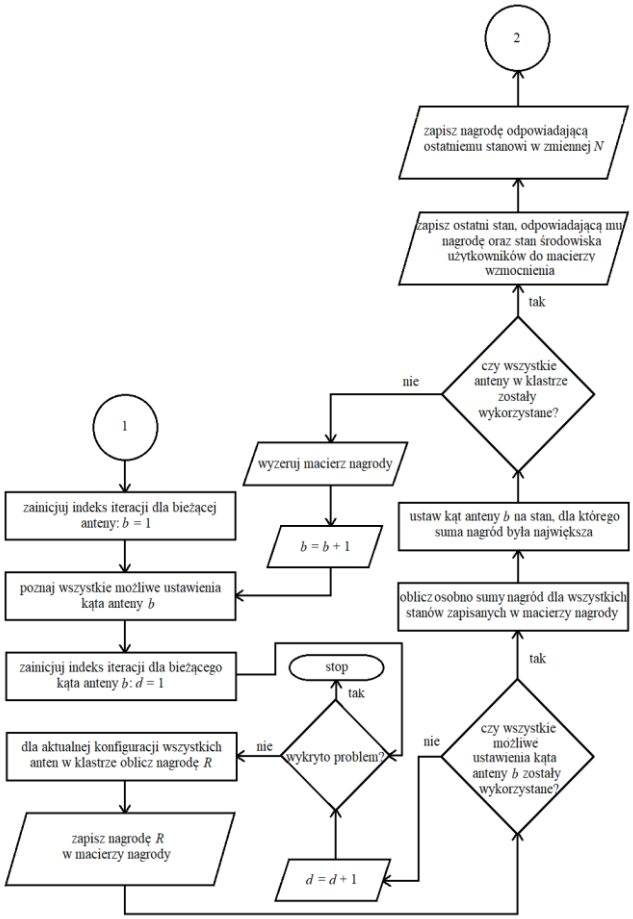
w związku z tym dalsza analiza agenta byłaby wątpliwa. Jeżeli taki problem wystąpi, należy zakończyć działanie algorytmu. Jeżeli wszystkie możliwe kąty zostały już sprawdzone, wartości nagrody wszystkich stanów zostały zapisane do macierzy nagrody. Celem przechowywania tych informacji jest odnalezienie najlepszej opcji pochyleń dla danego środowiska. Dla każdego stanu z macierzy obliczana jest suma nagród, a następnie wybierana najwyższa wartość z sum. Stan, który odpowiada największej sumie, powinien teraz zostać odzwierciedlony w fizycznych ustawieniach kątów anten w klastrze. Stan ten, wraz z odpowiadającą mu nagrodą oraz stan środowiska użytkowników, zapamiętywane są w macierzach wzmocnienia. Dodatkowo, po dobraniu kątów dla całego klastra, w zmiennej pomocniczej N zapisywana jest nagroda dla ostatniego najlepszego stanu.

W etapie trzecim, po pozytywnym zebraniu danych środowiskowych, nagroda R obliczana jest na podstawie nowych danych środowiskowych oraz aktualnego stanu kątów, a następnie porównywana ze zmienną N. W ten sposób agent ma możliwość szybkiego sprawdzenia, czy występuje konieczność korekty kątów w klastrze.



Rys. 3. Schemat algorytmu – część pierwsza

Jeżeli zmienne R i N różnią się, to kolejnym krokiem jest sprawdzenie, czy taka nagroda wraz z odpowiadającym mu stanem nie wystąpiły już w macierzach wzmocnienia. Jeśli tak, to oznacza to, że taki sam lub zbliżony stan środowiska już wystąpił podczas działania algorytmu i można skorzystać z zapisanych w macierzach wzmocnienia danych. W takim wypadku należy odnaleźć najwyższą nagrodę dla tego środowiska w macierzach wzmocnienia, ustawić odpowiadający mu stan kątów, podmienić zmienną N na znalezioną nagrodę i powrócić do początku etapu trzeciego. Jeżeli nie znaleziono szukanej nagrody w macierzach wzmocnienia, oznacza to, że bieżący stan środowiska wystąpił pierwszy raz podczas działania algorytmu, w związku z czym należy przejść do czwartego etapu. Wykorzystano w nim całą procedurę dobierania kątów opisaną w etapie drugim, z wyjątkiem początkowego zadania – pobrania danych środowiskowych. Etap kroku czwartego kończy się obliczeniem nagrody dla pozostałych stanów w macierzach wzmocnienia dla aktualnego stanu środowiska oraz dla zapisanych stanów środowiska z wcześniejszych etapów. Po wykonaniu tych instrukcji algorytm ponownie przechodzi do kroku trzeciego.



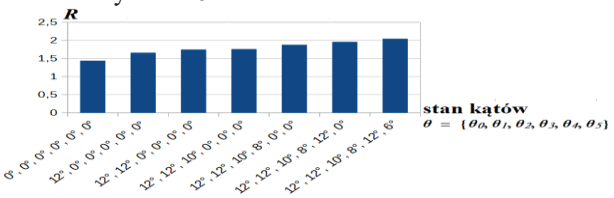
Rys. 4. Schemat algorytmu – część druga

3. WYBRANE SCENARIUSZE TESTOWE ORAZ WYNIKI

Zrealizowano scenariusze symulacyjne dobierania kątów anten nadawczych w sieci złożonej z trzech trzysektorowych stacji bazowych tworzących klaster. Zasto-

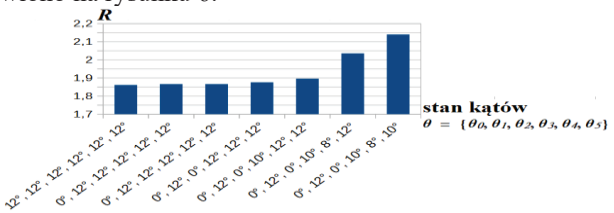
sowano powyższy algorytm doboru anten. Na jeden sektor przypada jedna antena nadawcza. Możliwości każdej anteny to ustawienie kąta o wartości ze zbioru: $K = \{0°, 2°, 4°, 6°, 8°, 10°, 12°\}$. Symulację przeprowadzono za pomocą środowiska Altair FEKO 2021 Student Edition na podstawie przykładowej mapy ukształtowania terenu, rozmieszczeń użytkowników oraz mocy sygnału odbieranego przez *UE*. Ustalono dwa środowiska S_1 i S_2, które różnią się rozmieszczeniem *UE* na tym samym wycinku mapy.

W stałym środowisku S_1 dla ustawień kątów początkowych: $\theta_p = \{\theta_0 = 0°, \theta_1 = 0°, \theta_2 = 0°, \theta_3 = 0°, \theta_4 = 0°, \theta_5 = 0°\}$, ustawienia końcowe kątów to: $\theta_k = \{\theta_0 = 12°, \theta_1 = 12°, \theta_2 = 10°, \theta_3 = 8°, \theta_4 = 12°, \theta_5 = 6°\}$. Wyniki sumy nagrody *R* dla całego klastra na poszczególnych etapach działania algorytmu przedstawiono na rysunku 5.



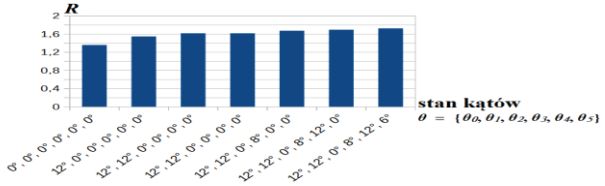
Rys. 5. Suma R na poszczególnych etapach działania algorytmu

W tym samym środowisku S_1 dla ustawień kątów początkowych: $\theta_p = \{\theta_0 = 12°, \theta_1 = 12°, \theta_2 = 12°, \theta_3 = 12°, \theta_4 = 12°, \theta_5 = 12°\}$, ustawienia końcowe kątów to: $\theta_k = \{\theta_0 = 0°, \theta_1 = 12°, \theta_2 = 0°, \theta_3 = 10°, \theta_4 = 8°, \theta_5 = 10°\}$. Wyniki sumy *R* dla całego klastra na poszczególnych etapach działania algorytmu przedstawiono na rysunku 6.



Rys. 6. Suma R na poszczególnych etapach działania algorytmu

W stałym środowisku S_2 dla ustawień kątów początkowych: $\theta_p = \{\theta_0 = 0°, \theta_1 = 0°, \theta_2 = 0°, \theta_3 = 0°, \theta_4 = 0°, \theta_5 = 0°\}$, ustawienia końcowe kątów to: $\theta_k = \{\theta_0 = 12°, \theta_1 = 12°, \theta_2 = 0°, \theta_3 = 8°, \theta_4 = 12°, \theta_5 = 6°\}$. Wyniki sumy *R* dla całego klastra na poszczególnych etapach działania algorytmu przedstawiono na rysunku 7.



Rys. 7. Suma R na poszczególnych etapach działania algorytmu

4. WNIOSKI

Stwierdzono, że kąt początkowy anteny oraz kolejność anten, z jaką obliczane są ich najlepsze ustawienia pochylenia mają wpływ na końcowe ustawienie algorytmu. Dla przeprowadzonych scenariuszy różnice w wynikach nagrody są niewielkie, natomiast ustawienia końcowe kątów znacznie się różnią. Wyniki zawarte na rysunkach 5-7 pokazują, że zmiany środowiskowe mają wpływ na wynik końcowy zarówno nagrody, jak i końcowych ustawień kątów. Z pokazanych przykładów wynika, że algorytm zawsze prowadzi do maksymalizacji pokrycia oraz pojemności. Możliwe jest powstawianie obszarów, w których moc sygnału odbieranego przez *UE* nie jest wystarczający do prawidłowego działania usług operatora sieci komórkowej. Można temu zapobiec, ustalając (na podstawie pomiarów mocy odbieranej) zbiór kombinacji kątów odpowiednich anten, dla których taka możliwość nie wystąpi.

Ponieważ złożoność obliczeniowa procedury dobrania najlepszych kątów jedynie za pomocą macierzy nagrody jest wyższa niż proste wyszukiwanie zapamiętanych danych w macierzach wzmocnienia, wykorzystano koncepcję uczenia się jako czynnik ulepszający i przyspieszający działanie algorytmu. Dzięki temu algorytm jest elastyczny w środowisku, w jakim pracuje. Może zostać wykorzystany w miejscach, w których występuje powtarzalność ruchu sieciowego w czasie, np. makrokomórkach miejskich. W związku z tym proces uczenia się będzie krótki, więc agent będzie mógł często sięgać do macierzy wzmocnienia, nie tracąc przy tym mocy obliczeniowej. Algorytm prawdopodobnie nie sprawdzi się w środowiskach, w których nie obserwuje się powtarzalności ruchu i rozmieszczenia *UE*, ze względu na szybkie przepełnienie macierzy wzmocnienia oraz wysoką złożoność obliczeniową.

Dzięki przedstawionemu algorytmowi doboru pochylenia anten można zapewnić lepsze parametry *QoE*. Działanie algorytmu pozwala na uzyskanie kompromisu między zasięgiem a przepustowością. Algorytm może być uznany jako baza lub komponent algorytmów *RL* wspierających samooptymalizację sieci z wykorzystaniem elektrycznego sterowania kątem pochylenia anten.

LITERATURA

[1] A. Imran, A. Zoha and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G" IEEE Network, tom 28, nr 6, str. 27-33, listopad-grudzień 2014

[2] J. Zhang, C. Sun, Y. Yi and H. Zhuang, "A hybrid framework for capacity and coverage optimization in self-organizing LTE networks," IEEE PIMRC, 2013, str. 2919-2923

[3] P. Zhao, L. Feng, P. Yu, W. Li and X. Qiu, "A fairness resource allocation algorithm for coverage and capacity optimization in wireless self-organized network," in China Communications, tom 15, nr 11, str. 10-24, listopad 2018

[4] J. Volakis, "Antenna Engineering Handbook", Mcgraw Hill Book, 2018

[5] N. Dandanov, H. Al-Shatri, A. Klein, V. Poulkov, „Dynamics Self-Optimization of the Antenna Tilt for Best Trade-off Between Coverage and Capacity in Mobile Networks". *Wireless Personal Communications*, Springer 2016, tom 192, str. 251-278.

ROZKŁADY PÓL ELEKTROMAGNETYCZNYCH W UKŁADZIE ANTENY TUBOWEJ DIAGONALNEJ Z SOCZEWKĄ

Agnieszka Piwowar; Paweł Bajurko

Politechnika Warszawska, Instytut Radioelektroniki i Technik Multimedialnych, Warszawa,
{agnieszka.piwowar.stud, pawel.bajurko}@pw.edu.pl

DOI: 10.15199/59.2022.4.55

Streszczenie**: W referacie przedstawiono analizę rozkładu pola elektromagnetycznego w układzie anteny tubowej diagonalnej z soczewką hiperboliczną na częstotliwościach subterahercowych. Układy takie są wykorzystywane m.in. w badaniach materiałowych. Analiza została przeprowadzona za pomocą hybrydowych symulacji elektromagnetycznych. Rozmiar próbki możliwej do badania w układzie oszacowano na 8 mm.**

Abstract**: The paper presents an analysis of the electromagnetic field distribution in a system of a diagonal horn antenna with a hyperbolic lens at subterahertz frequencies. Such systems are used, among others, in materials research. The analysis has been performed using hybrid electromagnetic simulations. The size of the testable sample in the system was estimated to be 8 mm.**

Słowa kluczowe**: Antena tubowa diagonalna, soczewka hiperboliczna, fale milimetrowe.**

Keywords**: Diagonal horn antenna, hyperbolic lens, millimeter waves.**

1. WSTĘP

Wykorzystanie w komunikacji bezprzewodowej coraz wyższych pasm radiowych wymaga rozwoju technologii wytwarzania układów elektronicznych, w tym technologii materiałowych. Wiąże się to z potrzebą określania właściwości materiałów metodami eksperymentalnymi na wysokich częstotliwościach. Powstało wiele metod służących do tego celu [5]. Wśród nich istotną grupę stanowią metody quasi-optyczne, polegające na transmisyjnych lub odbiciowych badaniach próbki materiału za pomocą zogniskowanej wiązki promieniowania. Nazwa tej grupy metod pochodzi od faktu, że na częstotliwościach subterahercowych i wyższych możliwe jest zastosowanie znanych z optyki elementów takich jak soczewki i zwierciadła do formowania wiązki promieniowania [2].

Rozkład pól w układzie quasi-optycznym zależy od użytych elementów i wykorzystanych materiałów. Wartości współczynnika załamania fali na częstotliwościach terahercowych się różnią się w zależności od wykorzystanego do produkcji soczewki materiału dielektrycznego. Także współczynnik absorpcji jest silnie zależny od częstotliwości i materiałów [11]. Na częstotliwościach subterahercowych korzystne właściwości posiada materiał PTFE (politetrafluoroetylen) – ma on niskie straty własne oraz stosunkowo niską wartość współczynnika załamania, co wpływa na niskie straty odbiciowe [3].

Przedmiotem niniejszej pracy jest określenie rozkładu pola w układzie antena-soczewka dielektryczna, złożonym z komponentów używanych w Laboratorium Anten i Techniki Subterahercowej na Politechnice Warszawskiej. Układ ten jest wykorzystywany w badaniach materiałowych na częstotliwościach subterahercowych [1, 2, 4, 5, 12, 13].

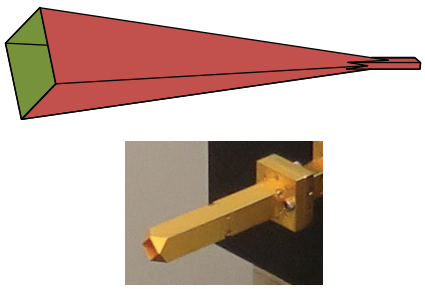
2. ELEMENTY UKŁADU

Rozpatrywany fragment układu quasi-optycznego składa się z dwóch elementów: anteny tubowej diagonalnej oraz soczewki.

2.1. Antena tubowa diagonalna

Antena jest zasilana z falowodu prostokątnego, który jest połączony z tubą zakończoną kwadratową aperturą. Apertura jest obrócona o 45° względem falowodu [6, 7]. Anteny diagonalne często znajdują zastosowanie w układach quasi-optycznych [8, 9], ze względu na fakt, że wiązka promieniowania ma stałą szerokość niezależnie od przekroju, w przeciwieństwie na przykład do standardowej anteny tubowej.

Budowa wewnętrzna oraz zdjęcie anteny zostały przedstawione na rys. 1. Antena będąca przedmiotem analizy zasilana jest falowodem WR-3.4 i przeznaczona jest do transmisji fal z zakresu 220–325 GHz. Długość całkowita anteny wynosi 56 mm, a bok apertury ma 5,6 mm.



Rys. 1. Antena tubowa diagonalna

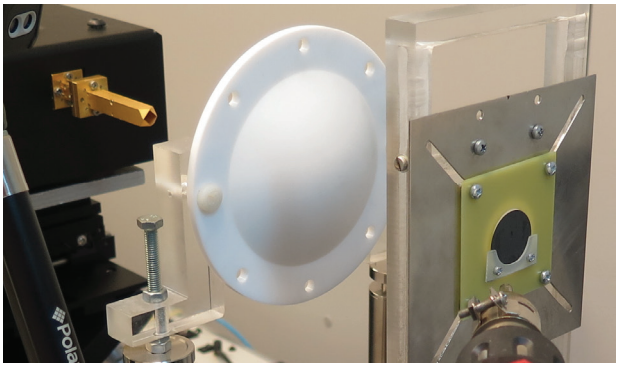
2.2. Soczewka

Drugim elementem układu jest dwuwypukła soczewka hiperboliczna [10] wykonana z materiału PTFE o względnej przenikalności elektrycznej ε = 2,055. Ogni-

skowa wynosi f = 100 mm (mierzona od szczytu soczewki). Powierzchnia hiperboliczna soczewki opisana jest równaniem

$$\frac{\left(x+\frac{f}{n+1}\right)^2}{\left(\frac{f}{n+1}\right)^2}-\frac{y^2}{f^2\frac{(n-1)}{n+1}}=1 \qquad (1)$$

gdzie $n=\sqrt{\varepsilon}$. Na rys. 2 przedstawiono soczewkę w układzie pomiarowym.



Rys. 2. Fragment quasi-optycznego układu pomiarowego. W jednym ognisku soczewki umieszczona jest antena, a w drugim badana próbka.

3. ROZKŁADY PÓL

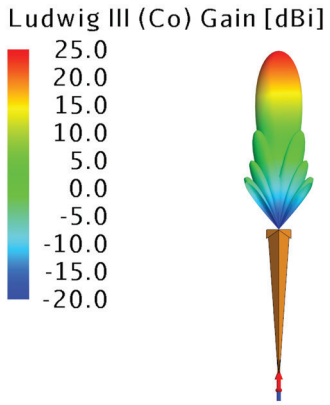
Modele obu elementów układu zostały przygotowane w programie Altair FEKO służącym do obliczeń elektromagnetycznych.

3.1. Charakterystyka promieniowania anteny

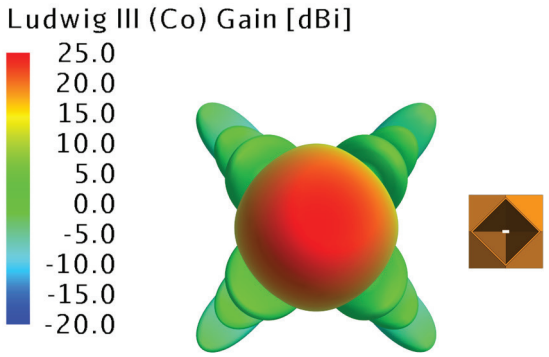
Model anteny posłużył do wyznaczenia charakterystyki promieniowania anteny oraz rozkładu pola bliskiego w pobliży apertury, potrzebnego do symulacji soczewki. Charakterystyki anteny obliczono dla częstotliwości 280 GHz metodą MLFMM (*multilevel fast multipole method*), która jest odmianą metody momentów. Rozwiązanie problemu elektromagnetycznego przez program wymagało znalezienia amplitud 309024 funkcji bazowych rozpisanych na strukturze anteny. Symulator potrzebował w tym celu 9,156 GB pamięci RAM. W celu ograniczenia złożoności obliczeniowej symulacji model anteny ograniczono do ścianek przewodzących leżących wewnątrz anteny oraz przylegających do apertury.

Na rys. 3 przedstawiono zysk na pożądanej polaryzacji. Na charakterystyce można zaobserwować wiązkę główną promieniowania anteny oraz listki boczne. Rys. 4 przedstawia widok charakterystyki promieniowania anteny z innej perspektywy, która pozwala zaobserwować, że listki boczne w charakterystyce powstają w płaszczyznach obróconych o 45° względem płaszczyzn symetrii falowodu zasilającego antenę. Jest to cecha charakterystyczna anten tubowych diagonalnych.

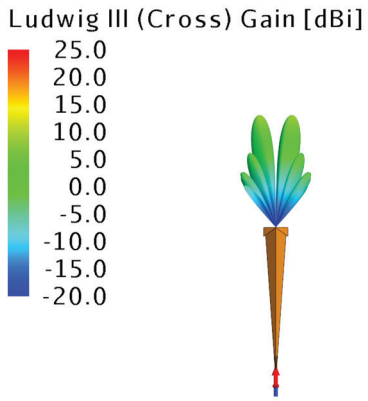
Na rys. 5 przedstawiono zysk na polaryzacji skrośnej. Całkowicie zniknął listek główny. Zysk na składowej skrośnej nie przekracza wartości 10 dBi (wobec 24,4 dBi zysku na składowej pożądanej). Maksimum wiązki dla składowej skrośnej jest odchylone o 15° od osi anteny.



Rys. 3. Zysk anteny na pożądanej polaryzacji

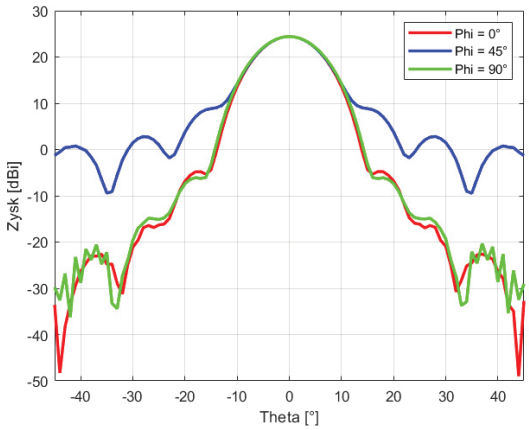


Rys. 4. Zyk anteny – widok z góry, obok charakterystyki pokazany jest widok anteny z tej samej perspektywy



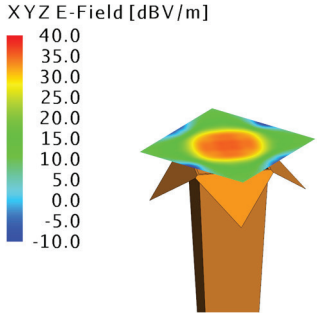
Rys. 5. Zysk anteny na polaryzacji skrośnej

Na rys. 6 przedstawiono zysk anteny dla trzech różnych płaszczyzn φ = 0° (płaszczyzna H), 45° i 90° (płaszczyzna E). Charakterystyki w płaszczyznach E i H są do siebie bardzo podobne. Jest to kolejna cecha charakterystyczna anten diagonalnych, odróżniająca je od standardowych anten tubowych, gdzie charakterystyki w płaszczyznach E i H różnią się zarówno pod względem szerokości wiązki głównej jak i poziomu listków bocznych. Jak można zaobserwować na rys. 6, listki boczne w antenie diagonalnej pojawiają się w przekroju 45°, a szerokość wiązki głównej pozostaje praktycznie stała w każdym przekroju i wynosi ok. 11°.



Rys. 6. Zysk anteny w przekrojach charakterystyki promieniowania

Wyznaczono także rozkład pola elektrycznego w strefie bliskiej promieniowania anteny w odległości 1 mm od apertury (rys. 7). Rozkład ten posłuży jako źródło przy symulacji soczewki.



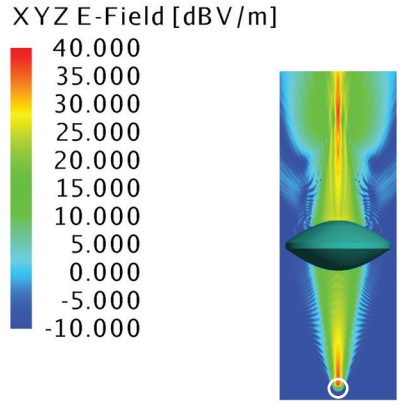
Rys. 7. Rozkład pola E w pobliży apertury anteny

3.2. Rozkład pola w układzie antena-soczewka

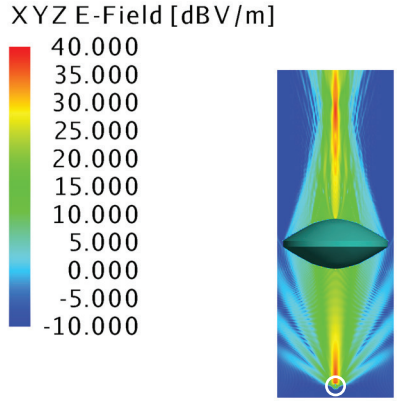
Rozkład pola elektromagnetycznego w układzie anteny z soczewką obliczono metodą hybrydową wykorzystującą wynik symulacji pola bliskiego w sąsiedztwie apertury anteny jako pobudzenie w modelu zawierającym soczewkę. Soczewkę symulowano metodą *ray-launching*. Symulacja uwzględniała zasady optyki geometrycznej z pominięciem zjawisk dyfrakcji.

Rys. 8 przedstawia rozkład pola w płaszczyźnie H, a rys. 9 w płaszczyźnie skośnej 45°. Antena jest umieszczona w jednym z ognisk soczewki (na dole rysunków 8 i 9). Fala po przejściu przez soczewkę jest skupiana w drugim ognisku po przeciwnej stronie soczewki. Na rys. 9 widoczne są wiązki promieniowania pochodzące od listków bocznych charakterystyki promieniowania anteny. Część z nich omija soczewkę.

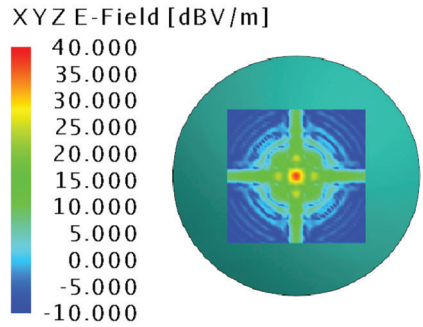
Rysunki 10 i 11 przedstawiają rozkład pola elektrycznego w płaszczyźnie poprzecznej w ognisku, odpowiednio dla polaryzacji pożądanej i skrośnej. Na rys. 10 widoczne jest skupienie pola E w jednym punkcie, jednak obecne są także „pasma" oraz lokalne maksima leżące w płaszczyznach E i H. Warto zwrócić uwagę, że w rozkładzie pola nie są widoczne żadne wyraźne wzrosty natężenia pola w płaszczyznach skośnych, mimo że antena, której użyto w układzie z soczewką, posiada listki boczne w płaszczyznach skośnych. Soczewka transformuje rozkład pola sprowadzając wzrosty natężenia z płaszczyzn skośnych z jednej strony soczewki do płaszczyzn E i H po stronie przeciwnej.



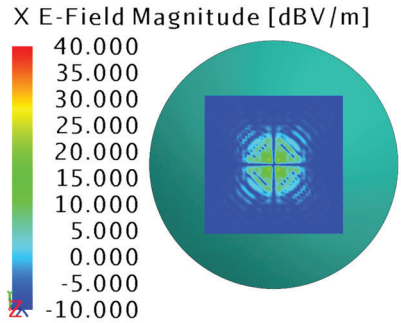
Rys. 8. Rozkład pola E w płaszczyźnie H. Kółkiem zaznaczono pozycję anteny.



Rys. 9. Rozkład pola E w płaszczyźnie skośnej 45°. Kółkiem zaznaczono pozycję anteny.



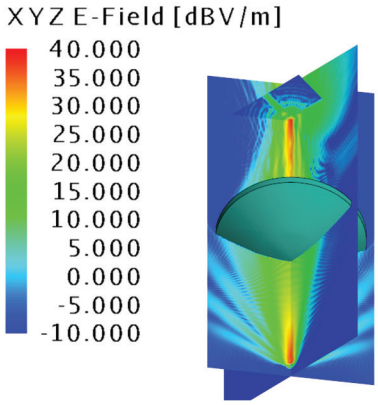
Rys. 10. Rozkład pola E w płaszczyźnie poprzecznej w ognisku



Rys. 11. Rozkład pola E w płaszczyźnie poprzecznej w ognisku – polaryzacja skrośna

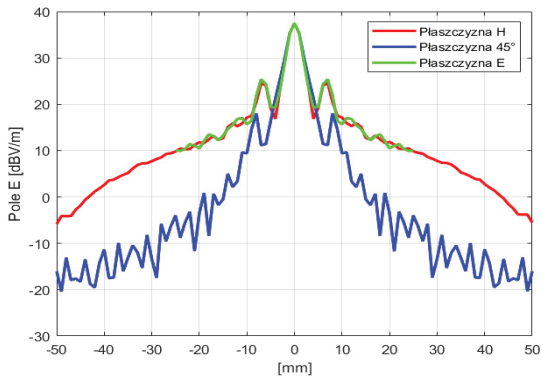
W rozkładzie przedstawionym na rys. 11 można zaobserwować, że pole skrośne występuje poza płaszczyznami E i H, a jego maksymalne natężenie jest o 15 dB słabsze od składowej pożądanej w maksimum.

Zestawienie wszystkich rozkładów z rys. 8–11 zostało przedstawione w widoku 3D na rys. 12. Widok ten pozwala zauważyć, że wspomniane wcześniej „pasma" nie znajdują się w tej samej płaszczyźnie co listki boczne charakterystyki promieniowania anteny.



Rys. 12. Rozkład pola E – widok 3D

Rozkład pola elektrycznego w ognisku dla różnych płaszczyzn we współrzędnych kartezjańskich został przedstawiony na rys. 13. Średnica plamki wynosi ok. 2,4 mm, lokalne maksima znajdują się w odległości 6,5 mm od osi optycznej. Z uzyskanych wyników można wywnioskować, że w układzie możliwe byłoby badanie próbek o minimalnej średnicy ok. 8 mm.



Rys. 13. Rozkład pola E w ognisku

4. PODSUMOWANIE

Zasymulowany rozkład pola elektrycznego w przedstawionym układzie wykazuje skupienie fali po przejściu przez soczewkę, co daje możliwość zbadania właściwości różnych próbek, w szczególności niewielkich płytek, ustawionych w odpowiedniej odległości za soczewką na częstotliwościach subterahercowych. Rozmiar próbki możliwej do badania w układzie oszacowano na 8 mm.

Praca ta jest częścią projektu finansowanego z programu Unii Europejskiej Horyzont 2020 na rzecz badań naukowych i innowacji w ramach umowy grantowej „Maria Skłodowska-Curie" nr 778156.

LITERATURA

[1] P. Bajurko, *et al.*, „Determination of $Bi_{12}SiO_{20}$ permittivity and loss tangent in the 220–325 GHz band and the influence of UV exposure on these parameters," *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, Lviv-Slavske, Ukraine, 2020, pp. 576-579.

[2] P. R. Bajurko, „Millimeter wave permittivity and loss tangent measurements of LTCC materials", 21st International Conference on Microwave, Radar and Wireless Communications (MIKON), 9-11 May 2016, Krakow, Poland, pp. 1-4.

[3] Burnett, A., *et al.* (2006, September). Complementary spectroscopic studies of materials of security interest. In *Optics and Photonics for Counterterrorism and Crime Fighting II* (Vol. 6402, p. 64020B). International Society for Optics and Photonics.

[4] K. Godziszewski, Y. Yashchyshyn, „Nowa metoda charakteryzacji dielektryków w paśmie częstotliwości subterahercowych", *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, nr 6, str. 501-504, czerwiec 2018.

[5] K. Godziszewski and Y. Yashchyshyn, „Broadband characterization of dielectrics in sub-THz range," *22nd International Microwave and Radar Conference (MIKON)*, Poznań, 2018, str. 576-579.

[6] Johansson Joakim F., Whyborn Nicholas „The Diagonal Horn as Sub-Milimeter Wave Antenna". *IEEE Transactions on Microwave Theory and Techniques*, 40 (5) : 795-800.

[7] Love, A. W. (1962). The diagonal horn antenna. *Microwave J.*, 117-122.

[8] J.Anthony Murphy, Stafford Withington, Alfred Egan, „Mode conversion at diffracting apertures in millimeter and submillimeter wave optical systems" *IEEE transactions on microwave theory and techniques*, 1993-10, Vol.41 (10), p.1700-1702

[9] J.Anthony Murphy, Neil Trappe, Stafford Withington „Gaussian beam mode analysis of partial reflections in simple quasi-optical systems fed by horn antennas" *Infrared physics & technology,* 2003, Vol.44 (4), p.289-297

[10] Piksa Petr, Zvanovec Stanislav, Cerny Petr „Elliptic and hyperbolic dielectric lens antennas in mm-waves" *Radioengineering*, 2011-04, Vol.20 (1), p.270-275

[11] Siemon Agnieszka, *et al.* „Paraffin Diffractive Lens for Subterahertz Range—Simple and Cost Efficient Solution". *IEEE transactions on terahertz science and technology,* 2021-07, Vol.11 (4), p.396-401

[12] B. Synkiewicz-Musialska, *et al.*, „Impact of additives and processing on microstructure and dielectric properties of willemite ceramics for LTCC terahertz applications," *Journal of the European Ceramic Society*, vol. 40, no. 2, February 2020, pp. 362-370.

[13] D. Szwagierczak, *et al.*, "Low temperature sintering of $Zn_4B_6O_{13}$ based substrates, their microstructure and dielectric properties up to the THz range", *Journal of Alloys and Compounds*, vol. 819, 5 April 2020, pp. 153025.

BADANIE POLA ELEKTRYCZNEGO DOJRZAŁEJ CHMURY BURZOWEJ PRZY POMOCY DŁUGICH ANTEN NAPOWIETRZNYCH

INVESTIGATION ON THE MATURE STORM CLOUD'S ELECTRIC FIELD USING LONG AIRBORNE ANTENNAS

Tomasz Aleksander Miś[1]

[1] Politechnika Warszawska, Warszawa; tomasz.a.mis@mailplus.pl

DOI: 10.15199/59.2022.4.56

Streszczenie: **Referat porusza kwestię wyznaczania wartości natężenia pola elektrycznego wewnątrz dojrzałej chmury typu *Nimbostratus* (ciepły front burzowy) przy pomocy danych doświadczalnych zebranych w trakcie balonowego lotu testowego napowietrznej długiej anteny liniowej. Na podstawie zarejestrowanych sygnałów i założonego modelu teoretycznego różnicowania ładunków elektrycznych na antenie wewnątrz chmury wyznaczono wartości pola elektrycznego zewnętrznego, generującego proces ładowania i powstawania wyładowań koronowych na przewodzie antenowym. Uzyskane dane porównano z danymi doświadczalnymi z lat poprzednich dla chmur burzowych typu *Nimbostratus*, ukazując nowe zastosowanie długich napowietrznych anten liniowych w bezpośrednim badaniu niebezpiecznych procesów zachodzących w atmosferze Ziemi.**

Abstract: **The paper describes an attempt to define the electric fields strength inside a mature *Nimbostratus* cloud (warm storm front) using experimental data collected during the flight of a long linear airborne antenna. Basing on the registered signals and a theoretical model of electrical interactions between the antenna and the cloud's interior the values of the external electric field strength have been calculated – the field that allows the build-up of the electric charge of the antenna and the excitation of corona. The calculations have been compared with experimental data on the electric field strength in *Nimbostratus* clouds, showing a new use of long airborne antennas in direct research on dangerous processes in the Earth's atmosphere.**

Słowa kluczowe: **VLF, chmury, loty stratosferyczne, pole elektryczne, remote sensing**

Keywords: **VLF, clouds, stratospheric flights, electric field, remote sensing**

1. WSTĘP. ANTENY LATAJĄCE A CHMURY

Wykorzystanie długich anten napowietrznych w celu emisji sygnałów radiowych poniżej 300 kHz oferuje możliwości doraźnego zastąpienia niskich acz skomplikowanych i kosztownych naziemnych struktur wsporczych przez struktury lżejsze, bez praktycznego ograniczenia długości anteny, oraz umieszczenia urządzenia nadawczego bliżej środka falowodu ziemia-warstwa D jonosfery [1] [10] – dającego potencjalnie wyższe uzyskiwalne zasięgi sygnału przy zachowaniu (względem przypadku naziemnego) polaryzacji pionowej [4]. Urządzenia tego typu, będące – w przypadku w pełni napowietrznym, tj. nieutwierdzonym (mechanicznie i galwanicznie) do ziemi – bardzo długimi (rzędu setek metrów długości) obiektami typu „floating earth conductor" [13], docierając do docelowego miejsca funkcjonowania (np. wysokości docelowej), oddziałują wzajemnie z otaczającymi warstwami atmosfery oraz zawartymi w nich obiektami (chmurami, mgłami). Oddziaływania te manifestują się w sposób bezpośredni w formie zmiany pola elektrycznego wokół latającego przewodu, różnicowania się ładunków elektrycznych między przewodem a jego otoczeniem oraz tworzeniem się wyładowań elektrycznych w formie korony oraz wyładowań dwukierunkowych [13]. Rejestracji tych zjawisk można dokonać na wiele sposobów, jednak część z nich – np. badania pola elektrycznego wokół przewodu / wewnątrz chmury przy pomocy próbników/sond kulowych [7] – nie da wyników niezniekształconych w związku z ładowaniem się struktur samych przyrządów pomiarowych [6]. Dane uzyskiwane w sposób pośredni, tj. metodami typowymi dla *remote sensing* – w postaci konkretnych sygnatur radiowych określonych mechanizmów – pozwalają uzyskać przybliżony opis obserwowanych zjawisk, dostarczający informacji kluczowych zarówno z perspektywy urządzenia nadawczego (szybkość ładowania się, możliwe przeciążenia nadajnika wskutek impulsów prądowych i napięciowych), jak i otoczenia (określenie struktury elektrycznej chmur, utrata bądź uzyskiwanie ładunku elektrycznego). W niniejszym referacie omówiono drugie zagadnienie, tj. próbę określenia struktury pola elektrycznego dojrzałej chmury burzowej na podstawie uzyskanych danych radiowych dotyczących zjawisk powstających wokół długiej anteny latającej – zagadnienie odwrotne w stosunku do analiz aerologicznych określających wymagania konstrukcyjne dla napowietrznych nadajników radiowych. Zastosowano tu nowy, autorski model ładowania się przewodu antenowego i utraty zgromadzonego na nim ładunku elektrycznego, łączący funkcje doświadczalne, teoretyczne dokładne i teoretyczne przybliżone; model oparty o ekstrapolowane zależności doświadczalne dla małych obiektów powierzchniowych przy wykorzystaniu danych eksperymentalnych dotyczących rzeczywistych natężeń naturalnych pól elektrycznych wewnątrz chmur i szybkości ładowania wewnątrz nich. Model ten może służyć jako weryfikator dla symulacji komputerowych, opartych o numeryczne rozwiązywania równań różniczkowych poszczególnych mechanizmów elektrycznych, wymagających uzupełnienia o rzeczywiste mechanizmy (i ich rzędy wielkości) zachodzące wewnątrz chmur burzo-

wych [2] – mechanizmy trudno definiowalne ze względu na warunki fizyczne panujące wewnątrz chmur (wysokie natężenia pól elektrycznych, wyładowania o charakterze niszczącym, ładowanie się całości struktur urządzeń zamiast dedykowanych elektrod pomiarowych itp).

2. POLA ELEKTRYCZNE A WYŁADOWANIA

Zasadniczym typem danych wejściowych w niniejszej analizie są przebiegi sygnału audio, odebranego przez odbiornik radiowy umieszczony na pokładzie gondoli balonowej, w bezpośredniej bliskości testowanej anteny (umocowanej poniżej gondoli) niewyposażonej w nadajnik. Sygnał ten, w przypadku wystąpienia gwałtownych interakcji elektrycznych anteny z otoczeniem, wykaże istnienie charakterystycznych impulsów, wprost proporcjonalnych do prądów wyładowań zarówno koronowych, jak i dwukierunkowych [9] [15].

W czasie testów długich anten napowietrznych przeprowadzonych w latach 2014-2015 [10], umieszczony na pokładzie gondoli balonowej anteny 200-metrowej odbiornik długofalowy uległ awarii i nie zarejestrował żadnych sygnałów radiowych (częstotliwość robocza: 225 kHz, A3E). Eksperyment powtórzono 12 czerwca 2021 roku z Ośrodka Badań Lotniczych i Kosmicznych Politechniki Warszawskiej, z anteną aluminiową wspartą na cienkiej warstwie tworzywa sztucznego, całość o długości 140 m. Długofalowy przenośny odbiornik radiowy typu abonenckiego firmy TENTO/KU-Zawod [12], umieszczony w gondoli głównej misji balonowej (powyżej punktu mocowania anteny) i pracujący na częstotliwości środkowej 144 kHz (A3E), w zakresie wysokości lotu 300-3300 m zarejestrował wyraźne impulsy sygnałów świadczące o występowaniu wyładowań koronowych w czasie przelotu przez dojrzałą chmurę typu *Nimbostratus*, będącą składnikiem ciepłego frontu burzowego [14]. Wspomniane sygnały zostały zarejestrowane jedynie we wskazanym zakresie wysokości lotu – nie zarejestrowano ich na ziemi, co wskazuje na sprowokowanie niniejszych zjawisk przez pionowy, wolno poruszający się obiekt o dużej powierzchni i dużym wydłużeniu.

Każdy zarejestrowany impuls posiadał przypisaną mu wartość unormowanej amplitudy A_i [-], wskazującej na intensywność zjawiska generującego impuls – prądu konkretnego wyładowania; samo wyładowanie zaś następuje za każdym razem powyżej przekroczenia określonej wartości natężenia pola elektrycznego, obliczeniowo wynoszącego (troposferycznie) 2.1 kV/m [8]. Upływ ładunku elektrycznego z anteny przy każdorazowym wyładowaniu może być określony doświadczalnie w przypadku optycznego zarejestrowania wyładowania [13], co ze względu na zbyt niską widzialność wewnątrz frontu burzowego nie było możliwe – wprowadzono założenie, iż natężenie pola elektrycznego po wyładowaniu spada do wartości granicznego natężenia pola elektrycznego $E_{gr.}$ [V/m] pomnożonej przez unormowaną amplitudę A_i dla danego wyładowania ($0<A_i<1$), a następnie wykazuje wzrost liniowy aż do kolejnego wyładowania przy $E_{gr.}$. Utrzymując założenie o równomierności występowania wyładowań koronowych na całej długości długich latających anten [5] oraz maksymalnej wartości ładunku elektrycznego [6], jaki zdoła się wyindukować w warunkach chmurowych na przewodzie o długości 2c = 140 m, grubości b = 10^{-4} m i przy szybkości ładowania w chmurze burzowej v_C = 0.3 nC/m^3/s, można sformułować równanie opisujące całkowite natężenie pola elektrycznego wokół poruszającej się i ładującej się w chmurze długiej anteny napowietrznej (indeksy i, i+1 oznaczają parametry dla sąsiadujących wyładowań):

$$E_{gr.}\left[A_i+\frac{1-A_i}{t_{i+1}-t_i}t\right]-E_{ext.}=\frac{3E_{ext.}}{4\pi r}\left[\log\frac{a+c}{a-c}\right]\left(1-e^{-\frac{bv_C}{3E_{ext.}\varepsilon_0}t}\right) \quad (1)$$

gdzie t [s] jest czasem wystąpienia wyładowania, $E_{ext.}$ [V/m] – natężeniem pola elektrycznego zewnętrznego (chmury wokół anteny) – szukaną zmienną, ε_0 – przenikalnością dielektryczną próżni, r [m] – odległością od miejsca wyładowania w chmurze (wartość przyjęta jako 1 m – ze względu na fizyczny brak dwóch elektrod w przestrzeni) i wielkość a [m] będąca pierwiastkiem z sumy kwadratów r [m] i c [m]. Wielkość logarytmowana określa wielkość natężenia pola elektrycznego według rozkładu dla długiego przewodu [3]. Można zauważyć, iż w niniejszym modelu osiągnięcie krytycznej wartości natężenia pola elektrycznego, przy którym następuje wyładowanie, uzależniono od złożenia dwóch funkcji – doświadczalnej (zewnętrznej) $E_{ext.}$ oraz przybliżonej dla założonego uprzednio wzrostu liniowego natężenia pola elektrycznego między sąsiednimi wyładowaniami; funkcje te przyrównano do założonych uprzednio mechanizmów ładowania przewodu antenowego o zadanej powierzchni i szybkości ładowania.

Rozwiązaniem jest funkcja:

$$E_{ext.}=\frac{\left[A_i+\frac{1-A_i}{t_{i+1}-t_i}t\right]\frac{bv_Ct}{3\varepsilon_0}}{\left[A_i+\frac{1-A_i}{t_{i+1}-t_i}t\right]W_k\left(-\frac{e^{-\frac{\left(\frac{3\log\left[\frac{a+c}{a-c}\right]}{4\pi r}+1\right)\frac{bv_Ct}{3\varepsilon_0}}{\left[A_i+\frac{1-A_i}{t_{i+1}-t_i}t\right]}}e^{\frac{3\log\left[\frac{a+c}{a-c}\right]}{4\pi r}\frac{bv_Ct}{3\varepsilon_0}}}{\left[A_i+\frac{1-A_i}{t_{i+1}-t_i}t\right]}\right)+\frac{3\log\left[\frac{a+c}{a-c}\right]}{4\pi r}\frac{bv_Ct}{3\varepsilon_0}+\frac{bv_Ct}{3\varepsilon_0}} \quad (2)$$

gdzie W_k jest funkcją W Lamberta wyznaczaną dla gałęzi k=0 (argument funkcji przyjmuje niewielkie wartości ujemne <<-1).

3. REZULTATY

Zależność (2) wykreślono na Rys. 1 dla argumentów $t=0.5\cdot(t_{i+1}-t_i)$, tj. środków przedziałów między sąsiednimi wyładowaniami, między wartościami $E_{gr.}A_i$ dla t_i i $E_{gr.}$ dla t_{i+1}. Na Rys. 2 przedstawiono porównanie wyznaczonego przebiegu natężenia pola elektrycznego z uśrednionym przebiegiem doświadczalnym, utworzonym na podstawie 10 zestawów danych z Taszkentu z lat 1960-1961 [6]. Zasadniczymi różnicami między obydwoma przebiegami jest ich początek,

naznaczających wysokość podstawy chmury (znacznie niższa w przypadku frontu burzowego znad OBLOT-u PW) oraz silna zmienność przebiegu wyznaczonego na podstawie danych z wyładowań koronowych wokół długiej anteny. Zmienność ta jest podyktowana faktem silnego wpływu tego rodzaju obiektu na parametry zewnętrznego pola elektrycznego w chmurze w swojej bezpośredniej bliskości, na mocy konieczności miejscowego zachowania warunków równowagi we wnętrzu chmury [6]. Kształt przebiegu natężenia pola elektrycznego w przypadku bez obecności przewodu (pola pierwotnego) może zostać przybliżony przy pomocy wielomianu czwartego stopnia (linia kropkowana na Rys. 2), wykazującego – podobnie, jak przebieg taszkencki – dwa maksima lokalne (choć mniej zaznaczone). Wartości tego przybliżonego przebiegu nie przekraczają ~1300 V/m.

4. PODSUMOWANIE

Na podstawie informacji dostarczonych przez przebiegi sygnału zawierającego informację o występujących wokół anteny napowietrznej wyładowaniach pozwalają na wyznaczenie przebiegu zewnętrznego natężenia pola elektrycznego w bezpośredniej bliskości anteny, dostarczając bezpośredniej informacji na temat struktury elektrycznej chmury burzowej w sposób niebezpośredni, tj. eliminujący wykorzystanie przyrządów pomiarowych podatnych na uszkodzenia i samoładowanie się (zniekształcających wskazania rzeczywiste) – dane te są możliwe do wykorzystania przy badaniu ewolucji chmur burzowych i dokładnym definiowaniu zależności ilościowych zachodzących w ich wnętrzu, co jest niezwykle pomocne przy weryfikacji i uzupełnianiu ich modeli numerycznych. Przyjęte liniowe wzrosty natężenia pola elektrycznego między wyładowaniami – wskutek zachodzących silnych mechanizmów elektryzacji i bardzo krótkich, rzędu do kilku sekund, interwałów między sąsiednimi wyładowaniami – są wystarczające do przybliżenia zmian natężenia pola elektrycznego w danym czasie (funkcja liniowa styczna do przebiegów mających charakter eksponencjalny przy braku wyładowań w długim okresie). Dodatkowe informacje pochodzące z innych źródeł a odnoszące się do rejestrowanych wyładowań – np. obserwacja wizualna – pozwolą dokładniej określić ilości traconego przez antenę ładunku elektrycznego, zwiększając dokładność wyznaczonego przebiegu rzeczywistego. Przyjęte przybliżenie spadku wartości natężenia pola elektrycznego po wyładowaniu ($E_{gr.}A_i$) okazało się jednak wystarczające do osiągnięcia poprawnego rzędu wielkości wskazań wyznaczonej funkcji.

LITERATURA

[1] Burrows Michael L. 1978. „ELF Communications Antennas". Peter Peregrinus Ltd., England.

[2] Doshi N. A., Agashe S. D. 2015. "Modelling of ions for seeding technique to electrify the atmosphere". *Journal of Electrostatics*, 75

[3] Durand E. 1964. „Électrostatique. Tome I. Les Distributions". Masson et Cie, France.

[4] Galejs Janis. 1970. „VLF fields of elevated sources". *Radio Science*, 5 (8, 9): 1163-1169.

[5] Hall Wayne C. 1947. „Electrostatic dischargers for aircraft*". Journal of Applied Physics*, 18: 759-765.

[6] Imianitow I. M., Czubarina E. W., Szwarc J. M. 1974. „Elektryczność chmur". Państwowe Wydawnictwo Naukowe, Warszawa.

[7] Jaworek Anatol. 1985. „Metody pomiarowe elektrostatyki". Instytut Maszyn Przepływowych PAN, Gdańsk.

[8] Masewicz Tadeusz, Wenda Stanisław. 1973. „Materiałoznawstwo radiotechniczne". Wydawnictwa Komunikacji i Łączności, Warszawa.

[9] Miki M., Rakov V. A., Shindo T., Diendorfer G., Mair M., Heidler F., Zischrank W., Uman M. A., Thottappillil R., Wang D. 2005. „Initial stage in lightning initiated from tall objects and in rocket-triggered lightning*". Journal of Geophysical Research: Atmospheres*, 110 (D2).

[10] Miś Tomasz. 2018. „The concept of an airborne VLF transmitter with vertical electric dipole antenna". IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting. doi:10.1109/APUSNCURSINRSM.2018.8608979.

[11] Miś Tomasz Aleksander, Józef Modelski. 2018. „Stratospheric VLF Vertical Electric Mono- And Dipole Antenna Tests in 2014-2015". 2018 Baltic URSI Symposium (URSI), CFP18N89-ART:566–70.

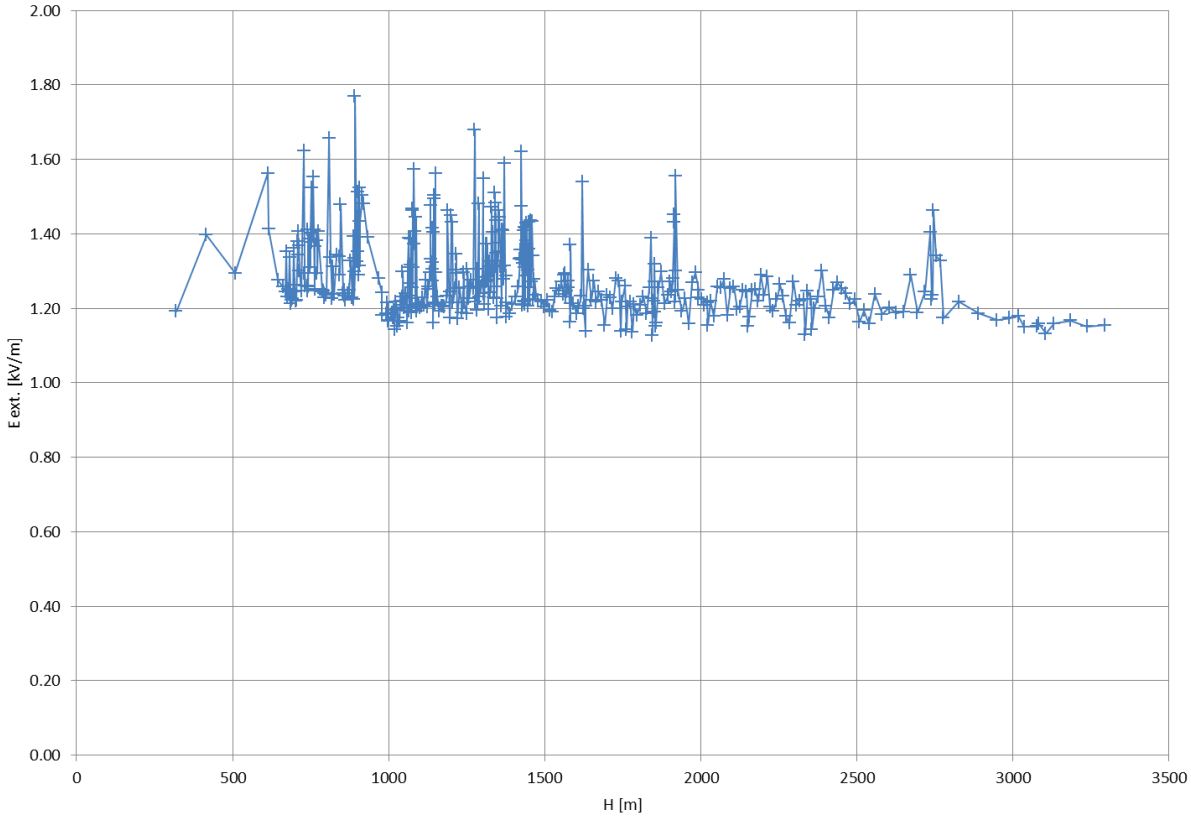
[12] Miś Tomasz Aleksander. 2020. „Częstotliwościowo-dystrybucyjna analiza szumów w radiofonicznym pasmie fal długich". Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji 2020, Krajowe Sympozjum Telekomunikacji i Teleinformatyki 2020

[13] Pavan C., Fontanes P., Urbani M., Nguyen N. C., Martinex-Sanchez M., Peraire J., Montanya J., Guerra-Garcia C. 2019. „Aircraft charging and its influence on triggering lightning". *Journal of Geophysical Research: Atmospheres*, 125 (1).

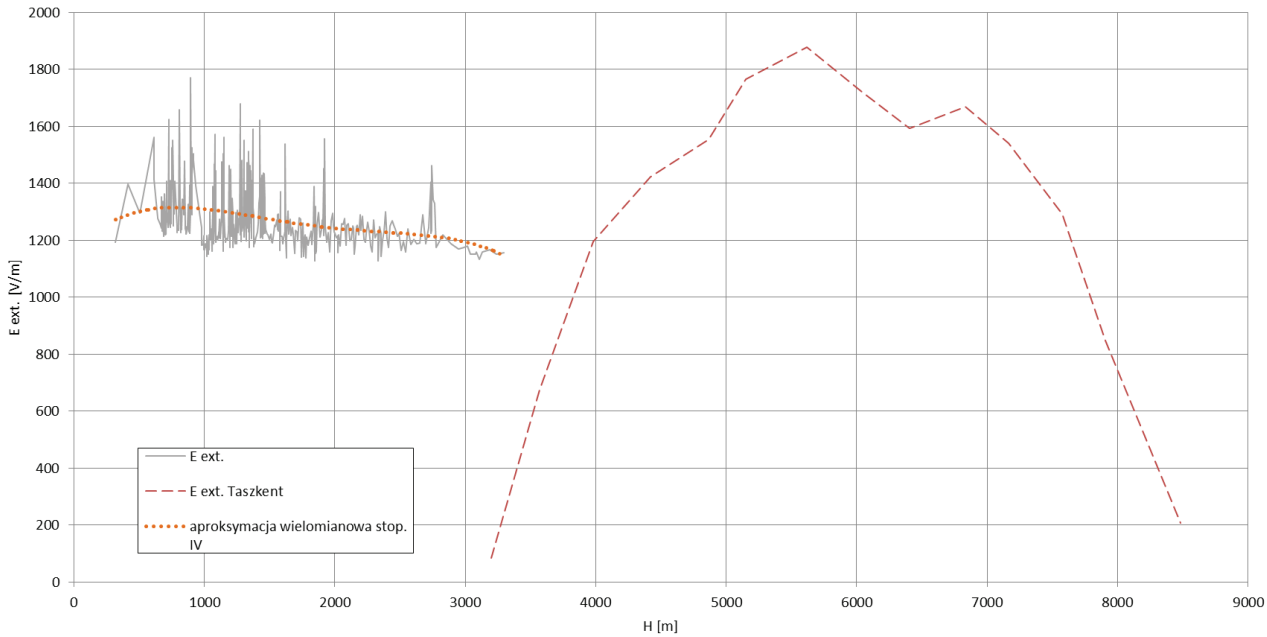
[14] Schmidt Marcin. 1975. „Meteorologia". Aeroklub Polskiej Rzeczypospolitej Ludowej. Wydawnictwo Komunikacji i Łączności, Warszawa.

[15] Sugimoto Toshiyuki, Kikuchi Hiroyuki, Higashiyama Yoshio. 2005. „Positive discharge from a grounded electrode toward negatively charged particles cloud". *Journal of Electrostatics*, 63: 609-614.

.



Rys. 1. Wyznaczona zależność natężenia zewnętrznego pola elektrycznego w chmurze od wysokości.



Rys. 2. Porównanie natężeń pól elektrycznych chmur burzowych Nimbostratus dla różnych przypadków doświadczalnych.

OPTYMALIZACJA SKUTECZNEJ CHARAKTERYSTYKI ANTENOWEJ I JEJ ZNACZENIE DLA OCENY EKSPOZYCJI NA PROMIENIOWANIE ELEKTROMAGNETYCZNE

Kamil Bechta

Nokia Solutions and Networks Sp. z o.o., Wrocław, kamil.bechta@nokia.com

DOI: 10.15199/59.2022.4.57

Streszczenie: **Zakładając, że charakterystyka promieniowania anteny jest filtrem przestrzennym dla propagującej się energii elektromagnetycznej, możliwa jest optymalizacja kształtu tego filtra, zgodnie z kątowym rozkładem mocy w kanale radiowym, w celu poprawy budżetu łącza radiowego. Przedstawiona została metoda wyznaczania skutecznej charakterystyki układu wieloantenowego oraz jej optymalizacji. Dodatkowo wskazano znaczenie skutecznego zysku antenowego dla oceny ekspozycji na promieniowanie elektromagnetyczne od stacji bazowej systemu 5G.**

Abstract: **Considering an antenna beam pattern as a spatial filter of electromagnetic energy propagated in the space, it should be possible to optimize the shape of this filter and match it to the power angular spectrum in a given radio channel. This paper presents a method for determination of effective and optimal pattern of an antenna array. In addition, it was indicated how the effective antenna gain impacts estimation of exposure on electromagnetic filed from the base stations of 5G system.**

Słowa kluczowe: **5G, ekspozycja na PEM, kątowe rozkład mocy, układ wieloantenowy**

Keywords: **5G, antenna array, EMF exposure, power angular spectrum**

1. WSTĘP

Obliczanie budżetu energetycznego łącza radiowego dla kanału z propagacją wielodrogową wymaga uwzględnienia wpływu zjawiska kątowego rozkładu mocy (PAS) na kształt skutecznej charakterystyki promieniowania układów wieloantenowych stosowanych w systemach komórkowych piątej generacji (5G).

Podczas projektowania anteny jednym z głównych celów jest uzyskanie założonej charakterystyki antenowej. W przypadku układów wieloantenowych pożądane jest uzyskanie wysokiego zysku na głównym kierunku promieniowania oraz niskich poziomów listków bocznych. Charakterystyka antenowa, która została zaprojektowana, a następnie zmierzona w komorze bezodbiciowej jest w tym artykule określana jako *charakterystyka nominalna*.

Wraz ze wzrostem liczby elementów antenowych w układzie, maksymalny nominalny zysk g_{max}^{Nom} układu wieloantenowego wzrasta, a szerokość jego charakterystyki antenowej w połowie mocy (HPBW) maleje. Relacje te opisuje równanie (1) [6]:

$$g_{max}^{nom} = \frac{2}{B_{ho} \cdot B_{vo_e}} = N \cdot G_e, \qquad (1)$$

gdzie B_{ho} i B_{vo} są nominalną średnią kwadratową (RMS) szerokości wiązki odpowiednio w płaszczyźnie poziomej i pionowej (w radianach), N jest liczbą elementów antenowych w układzie, a G_e jest zyskiem pojedynczego elementu antenowego.

W rzeczywistym środowisku z propagacją wielodrogową i rozproszeniami, które znacznie różni się od warunków propagacji w komorze bezodbiciowej, maksymalny możliwy do uzyskania zysk, i związany z nim parametr HPBW, dla układów wieloantenowych różnią się od ich wartości nominalnych i są określane jako skuteczne. Dlatego też charakterystyka anteny mierzona w kanale radiowym z rozproszeniami jest w dalszej części artykułu określana jako *skuteczna charakterystyka* anteny dla tego kanału.

Nominalna charakterystyka i nominalny zysk są specyficzne dla anteny, podczas gdy skuteczne charakterystyki i odpowiadające im skuteczne zyski zmieniają się w zależności od kanału radiowego. Różnica między nominalnymi i skutecznymi charakterystykami anten zależy od kątowego rozkładu mocy w danym kanale radiowym. Zgodnie z publikacją [6], równania (2) – (4) opisują, jak można analitycznie uzyskać skuteczne zyski anteny g^{Eff} znając zyski nominalne g^{Nom} oraz kątowy rozkład mocy p w założonym kanale radiowym.

$$g^{Eff}(\phi_0, \theta_0) = \int_{-180^\circ}^{180^\circ} \int_{-90^\circ}^{90^\circ} g^{Nom}(\phi_0 - \phi, \theta_0 - \theta) \cdot p(\phi, \theta) d\phi d\theta, \qquad (2)$$

$$g_{Az}^{Eff}(\phi_0) = \int_{-180^\circ}^{180^\circ} g_{Az}^{Nom}(\phi_0 - \phi) \cdot p_{Az}(\phi)\, d\phi, \qquad (3)$$

$$g_{Ele}^{Eff}(\theta_0) = \int_{-90^\circ}^{90^\circ} g_{Ele}^{Nom}(\theta_0 - \theta) \cdot p_{Ele}(\theta)\, d\theta, \qquad (4)$$

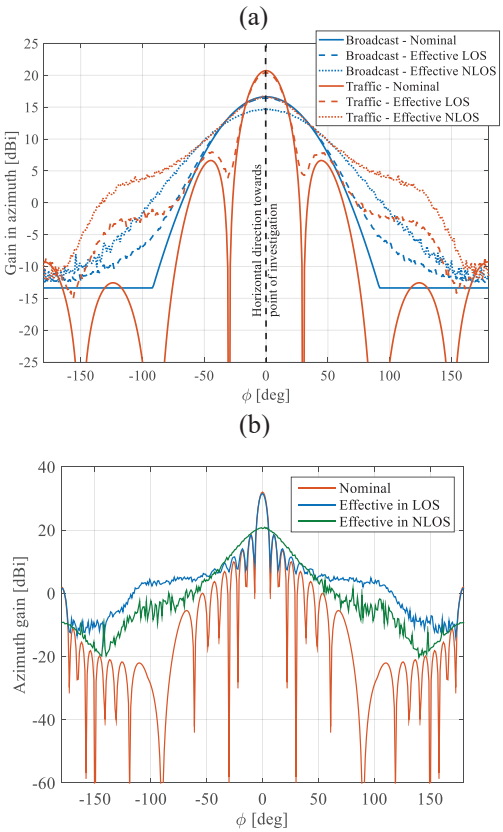
gdzie ϕ i θ definiują dziedzinę kątową odpowiednio w płaszczyźnie poziomej i pionowej, natomiast ϕ_0 i θ_0 wskazują kąty orientacji wiązki głównej odpowiednio dla tych płaszczyzn. Parametry g_{Az}^{Nom}, g_{Ele}^{Nom}, g_{Az}^{Eff} oraz g_{Ele}^{Eff} oznaczają nominalne i skuteczne zyski anteny, jeśli rozważana jest odpowiednio tylko płaszczyzna pozioma lub pionowa promieniowania, podczas gdy p_{Az} i p_{Ele} reprezentują realizacje PAS odpowiednio w płaszczyźnie poziomej i pionowej. Skuteczne zyski anteny g^{Eff} obliczone dla pełnych zakresów kątów orientacji wiązki, tj. $\phi_0 \in [-180^\circ; 180^\circ)$ oraz $\theta_0 \in [-90^\circ; 90^\circ]$, pozwalają

uzyskać skuteczną trójwymiarową (3D) charakterystykę anteny w danym kanale radiowym określonym przez daną realizację kątowego rozkład mocy p.

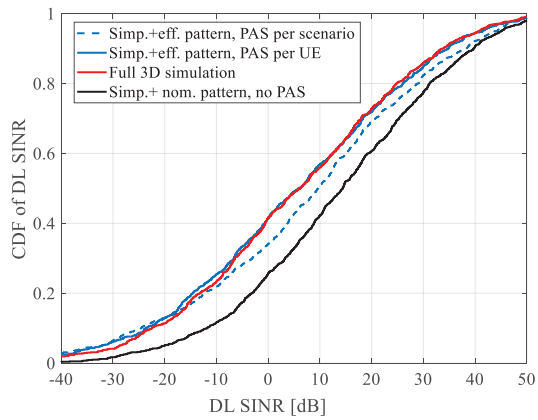
Dla zilustrowania, Rys. 1 (rysunek 3 z [8] i rysunek 1 z [2]) przedstawia przykładowe porównanie nominalnych i skutecznych charakterystyk antenowych w płaszczyźnie poziomej dla praktycznych układów wieloantenowych stacji bazowych 5G w środowisku typu urban macro (UMa) określonym przez standard 3GPP TR 38.901. Przedstawione charakterystyki uzyskano poprzez uśrednienie 1000 realizacji symulacyjnych równania (3).

Skuteczne charakterystyki antenowe przedstawione na Rys. 1 zostały określone na podstawie PAS dla kanałów radiowych opisanych przez 3GPP w TR 38.901. Jak można zauważyć, każda z tych ilustracji zawiera skuteczne charakterystyki osobno dla warunków propagacji z bezpośrednią widocznością (LOS) oraz bez bezpośredniej widoczności (NLOS). Takie podejście wynika z faktu, że kątowy rozkład mocy dla warunków LOS i NLOS różni się znacznie i prowadzi do różnych skutecznych charakterystyk. W przypadku LOS spadek skutecznego zysku anteny (w stosunku do nominalnego zysku anteny) jest zauważalnie mniejszy niż w przypadku NLOS. Jednak zwiększenie wartości HPBW (w odniesieniu do nominalnej charakterystyki anteny), a co za tym idzie wzrost poziomów listków bocznych, jest znaczące zarówno w LOS, jak i NLOS. Wszystkie wskazane odchylenia od kształtu nominalnej charakterystyki anteny i zysku zależą od PAS w danym scenariuszu implementacyjnym, a także od parametrów nominalnej charakterystyki anteny i powinny być brane pod uwagę podczas modelowania budżetu łącza radiowego zarówno dla użytecznych, jak i zakłócających sygnałów w komórce radiowej systemu 5G z formowaniem wiązki.

W związku z tym zbadanych zostało kilka scenariuszy symulacyjnych, które zostały przedstawione w publikacjach [1], [2] i [4]. Celem tych analiz było zobrazowanie błędu w modelowaniu stosunku mocy sygnału użytecznego do sumy mocy szumu i zakłóceń (SINR) dla łącza w dół (DL) w komórce 5G pracującej w zakresie fal milimetrowych (mmWave), jeśli podczas obliczania budżetu łącza radiowego nie bierze się pod uwagę skutecznej charakterystyki antenowej. W przypadku środowiska UMa pokazano na rysunkach 5 i 8 w publikacji [1] oraz rysunku 10 w publikacji [2], że przeszacowanie DL SINR może sięgać nawet 20 dB. Przeszacowanie DL SINR uzyskanego dla środowiska typu urban micro street canyon (UMi SC) jest niższe, ale nadal może osiągnąć więcej niż 10 dB, jak przedstawiono na rysunku 10 z publikacji [2], podczas gdy wyniki symulacji zawarte na rysunku 9 z publikacji [4] wskazują na przeszacowanie o więcej niż 5 dB w przypadku środowiska podmiejskiego (SubU). Wspomniane wyniki symulacji zakładają, że w transmisji w łączu w dół występują zarówno warunki LOS, jak i NLOS, jednak należy zauważyć, że znaczne przeszacowanie wartości SINR w łączu w dół wynika głównie z warunków NLOS. W obecności silnej ścieżki bezpośredniej w warunkach LOS zysk skuteczny anteny jest zbliżony do wartości nominalnej, natomiast w warunkach NLOS zysk skuteczny jest zauważalnie niższy od wartości nominalnej. Ta różnica w zyskach antenowych powoduje przeszacowanie wartości mocy sygnału użytecznego (DL S), jeśli wykorzystywana jest charakterystyka nominalny. Z drugiej strony kątowy rozkład mocy w płaszczyźnie poziomej powoduje zwiększone skuteczne zyski dla listków bocznych w stosunku do wartości nominalnych zarówno w warunkach LOS, jak i NLOS. Jest to przyczyną niedoszacowania mocy sygnału zakłócającego (DL I), ponieważ większość zakłóceń jest odbierana przez listki boczne. Dlatego też użycie charakterystyki nominalnej powoduje przeszacowanie wartości DL S i niedoszacowa-



Rys. 1. Przykładowe porównanie nominalnych i skutecznych charakterystyk antenowych w środowisku 3GPP UMa dla anteny 3.5 GHz Commscope "RRZZHHTTS4-65B-R7" (a) oraz układu wieloantenowego 16×16 dla pasma 28 GHz (b)



Rys. 2. Dystrybuanta wartości DL SINR uzyskana z symulacji dla scenariusza 3GPP UMi SC zaimplementowanego w zakresie fal milimetrowych (wyniki dla warunków LOS i NLOS)

nie wartości DL I, co prowadzi do znacznego przeszacowania DL SINR we wszystkich symulowanych scenariuszach implementacyjnych, jak opisano bardziej szczegółowo w publikacjach [2] i [6].

Wspomniane wyniki symulacji, częściowo przedstawione również na Rys. 2 (rysunek 5 z [6]), wyraźnie pokazują, że uproszczona metoda z wykorzystaniem charakterystyki nominalnej do szacowania budżetu łącza radiowego 5G, może dać błędny obraz wydajności łącza, nieosiągalnej w rzeczywistej implementacji.

2. OPTYMALIZACJA CHARAKTERYSTYKI ANTENOWEJ W CELU POPRAWY BUDŻETU ŁĄCZA RADIOWEGO

Antenę kierunkową można traktować jako filtr przestrzenny dla energii elektromagnetycznej w kanale radiowym, który można dopasać do kątowego rozkładu mocy w tym kanale w celu zmaksymalizowania skutecznego zysku. Aby udowodnić to założenie, przeprowadzono szereg symulacji i pomiarów. W wyniku przeprowadzonych badań wyprowadzono dwa szczegółowe rozwiązania dla:

1. Oszacowania kątowego rozkładu mocy w danym kanale radiowym w oparciu o pomiary mocy sygnału przy użyciu trzech lub więcej różnych konfiguracji układów wieloantenowych po stronie nadawczej stacji bazowej 5G, jak opisano szczegółowo w rozdziale III-B publikacji [5];
2. Optymalizacji geometrii układu wieloantenowego w celu maksymalizacji mocy wypromieniowanej do lub odebranej z przestrzeni przy określonych wartościach kątowego rozkładu mocy w danym kanale radiowym, jak opisano szczegółowo w rozdziale III-A publikacji [5].

Przy założeniu, że znany jest kątowy rozkład mocy w miejscu rozmieszczenia anteny stacji bazowej 5G (określony na podstawie statystycznych modeli kanału radiowego, np. 3GPP TR 38.901, lub oszacowany w oparciu o pierwszą z ww. metod), optymalizacja zysku skutecznego G tej anteny można otrzymać stosując drugą z ww. metod. Opis tej metody jest przedstawiony w publikacjach [4] i [6], natomiast poniższe równania przedstawiają jest końcowy rezultat, którym jest wymiar zoptymalizowanego układu wieloantenowego w płaszczyźnie pionowej K_1 i płaszczyźnie poziomej K_2 pozwalający na uzyskanie optymalnego zysku anteny G.

$$G(N, B_{ve}, B_{he}, \sigma_v, \sigma_h) = \frac{2}{\sqrt{\left(\frac{B_{ve}}{K_1}\right)^2 + \sigma_v^2}\sqrt{\left(\frac{B_{he}}{K_2}\right)^2 + \sigma_h^2}}, \quad (5)$$

$$K_1 = \sqrt{\frac{NB_{ve}\sigma_h}{B_{he}\sigma_v}}, K_2 = \sqrt{\frac{NB_{he}\sigma_v}{B_{ve}\sigma_h}}. \quad (6)$$

Poszczególne z parametrów oznaczają:

N - liczbę elementów antenowych w układzie;

B_{ve}- nominalną średniokwadratową szerokość wiązki pojedynczego elementu antenowego w płaszczyźnie pionowej;

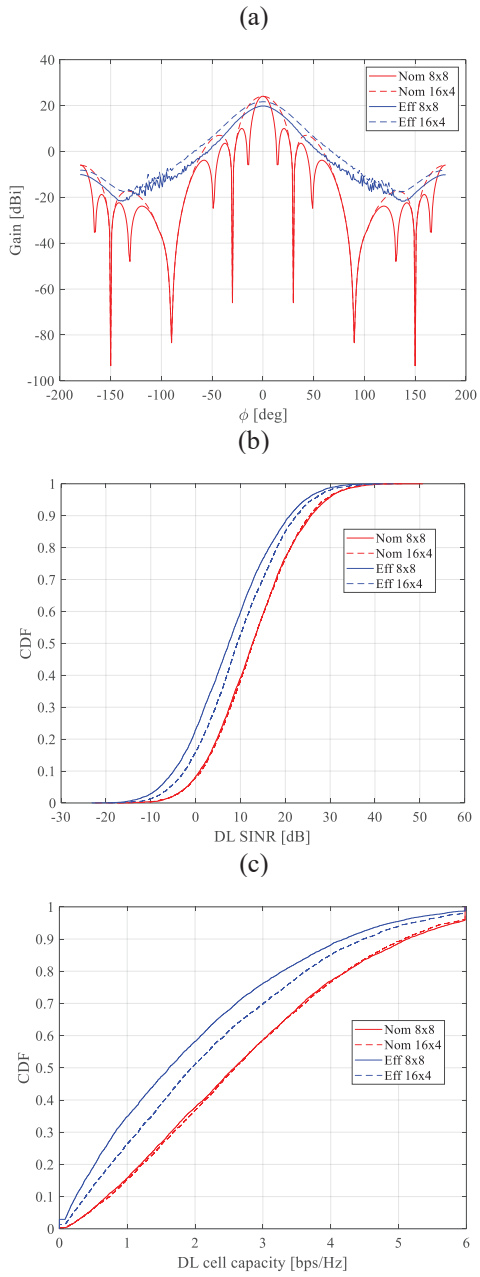
B_{he}- nominalną średniokwadratową szerokość wiązki pojedynczego elementu antenowego w płaszczyźnie poziomej;

σ_v - średniokwadratową wartość kątowego rozkładu mocy w płaszczyźnie pionowej dla anteny nadawczej,

σ_h - średniokwadratową wartość kątowego rozkładu mocy w płaszczyźnie poziomej dla anteny nadawczej.

Zasada działania przedstawionej metody optymalizacji geometrii układu wieloantenowego sprowadza się do założenia, że kształt jego charakterystyki (rozumiany jako filtr przestrzenny), jest dopasowany do kątowego rozkładu mocy w płaszczyźnie poziomej i pionowej. Metoda ta została zweryfikowana przez symulacje statystyczne i pomiary w warunkach rzeczywistych w celu określenia optymalnej geometrii układu wieloantenowego z analogowym formowaniem wiązki. Symulacje przeprowadzone dla środowiska UMi SC i przedstawione na rysunku 6 w publikacji [5] wskazują, że optymalizacja kształtu charakterystyki anteny, dzięki rekonfiguracji jednorodnej geometrii z konfiguracji 8×16 na 42×3, pozwala na zwiększenie DL SINR o około 6,6 dB. Optymalizacja geometrii z konfiguracji 8×8 na 16×4, jak przedstawiono na Rys. 3(a) (rysunek 4 z [4]) dla środowiska SubU, pozwala zwiększyć DL SINR o prawie 2 dB w scenariuszu z pojedynczym użytkownikiem (SU-MIMO), co jest przedstawione na Rys. 3(b) (rysunek 9 z [4]). Wyższa wartość DL SINR prowadzi do poprawy przepustowości w łączu w dół na granicy komórki o około 60%, jak przedstawiono na Rys. 3(c) (rysunek 10 z [4]).

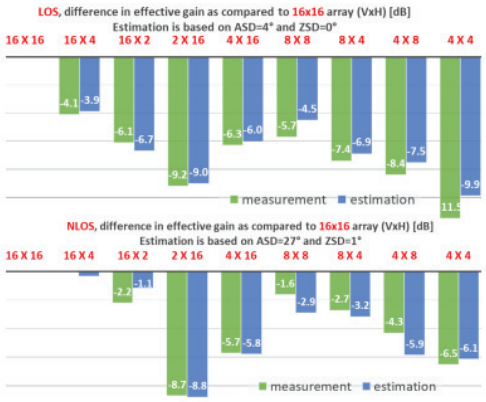
Aby zweryfikować skuteczność przedstawionej metody optymalizacji, przeprowadzono pomiary w rzeczywistym kanale radiowym z wykorzystaniem po stronie nadawczej komercyjnie dostępnego układu wieloantenowego o wymiarach 16×16 i pracującego na częstotliwości 28 GHz (Nokia AEUA AirScale Multi Antenna Array, 2Tx-2Rx , 512 elementów antenowych, 64 dBm maksymalnej równoważnej izotropowej mocy promieniowanej - EIRP). Po stronie odbiorczej została użyta antena tubowa o zysku 10 dBi. Szczegółowy opis procedury pomiarowej oraz obróbki końcowej wyników zawarto w rozdziale IV-C publikacji [5]. Pomiary wykonano w warunkach LOS i NLOS dla dziewięciu konfiguracji geometrii układu wieloantenowego (16×16, 16×4, 16×2, 2×16, 4×16, 8×8, 8×4, 4×8 i 4×4) . Rysunek 4 (rysunek 7 z [5]) ilustruje skonfigurowane stanowisko pomiarowe. Na podstawie wyników pomiarów mocy odebranej wyznaczono względne efektywne zyski antenowe dla wszystkich mierzonych konfiguracji w odniesieniu do konfiguracji wyjściowej 16×16. Równolegle oszacowano kątowy rozkład mocy w płaszczyźnie poziomej i pionowej w danych warunkach propagacji zgodnie z metodą przedstawioną w rozdziale III-B publikacji [5]. Oszacowane wartości kątowego rozkładu pozwoliły na obliczenie względnych skutecznych zysków anteny dla wszystkich założonych konfiguracji układu zgodnie z równaniem (5). Porównanie



Rys.3. Charakterystyki nominalne i skuteczne dla konfiguracji 8x8 oraz 16x4 anteny stacji bazowej 5G w warunkach NLOS (a), dystrybuanty wartości DL SINR (b) oraz dystrybuanty przepustowości łącza w dół (c) dla komórki systemu 5G pracującego w zakresie fal milimetrowych i w środowisku SubU

względnych skutecznych zysków uzyskanych z pomiarów i obliczeń wykazuje dobre dopasowanie, ale przede wszystkim wskazuje, że geometrię układu wieloantenowego można zoptymalizować w celu poprawy skutecznego zysku anteny, jak pokazano na Rys. 5 (rysunek 8 z [5]). Na przykład, dla warunków LOS konfiguracja 16×2 ma podobny zysk jak 8×8 przy użyciu 2 razy mniej elementów antenowych, co wskazuje, że efektywny zysk anteny „nasyca się" w przypadku analogowego formowaniu wiązki i przy danym kątowym rozkładzie mocy nie może być dalej zwiększony, nawet przez większą liczbę elementów antenowych w układzie. Dla warunków NLOS skuteczny zysk konfiguracji 16×2 jest tylko o 2,2 dB niższy niż dla 16×16, podczas gdy dla konfiguracji 2×16 jest

Rys. 4. Stanowisko pomiarowe dla warunków LOS (po lewej) i warunków NLOS (po prawej)



Rys. 5. Wyniki pomiarów i odpowiadające im oszacowane wartości skutecznego zysku w warunkach LOS (górny wykres) oraz NLOS (dolny wykres)

to już różnica aż 8,7 dB, co jasno pokazało potrzebę optymalizacji geometrii układu wieloantenowego.

3. ESTYMACJA EKSPOZYCJI NA POLE ELEKTROMAGNETYCZNE OD STACJI BAZOWEJ 5G W KANALE Z ROZPROSZENIAMI

Wraz z wprowadzeniem systemu 5G aktualizowane są metody oceny ekspozycji na pole elektromagnetyczne (PEM) o częstotliwości radiowej opracowane dla poprzednich systemów, aby uwzględnić rzeczywiste parametry transmisji, a przede wszystkim formowanie i sterowanie wiązką antenową. Międzynarodowa Komisja Elektrotechniczna (IEC) opracowuje i zatwierdza metody oceny ekspozycji na PEM od stacji bazowych w normie IEC 62232. Jeżeli ocena jest przeprowadzana na miejscu (in-situ), zaleca się ekstrapolację do maksymalnego poziomu ekspozycji (zwykle związanego z kanałem użytkownika) na podstawie pomiaru stabilnych sygnałów, takich jak kanał rozgłoszeniowy. Analiza porównawcza metody ekstrapolacji opartej na nominalnej i skutecznej charakterystyce antenowej stacji bazowej 5G została przedstawiona w publikacjach [7] i [8]. Analiza ta pokazuje, że metoda ekstrapolacji z nominalną charakterystyką antenową może prowadzić do znacznego przeszacowania maksymalnej ekspozycji w warunkach NLOS.

W badaniach przedstawionych w publikacjach [7] i [8] przyjęto charakterystykę komercyjnej anteny 5G dla pasma 3,5 GHz (Commscope „RRZZHHTTS4-65B-R7") i założono jej użycie w trzech różnych środowiskach propagacyjnych, zgodnych ze standardem 3GPP TR 38.901

– UMi SC, UMa oraz rural macro (RMa). Niniejsze badanie koncentruje się na estymacji współczynnika ekstrapolacji $F_{extBeam}$ występującego w równaniu (7) (równanie (1) w [8]).

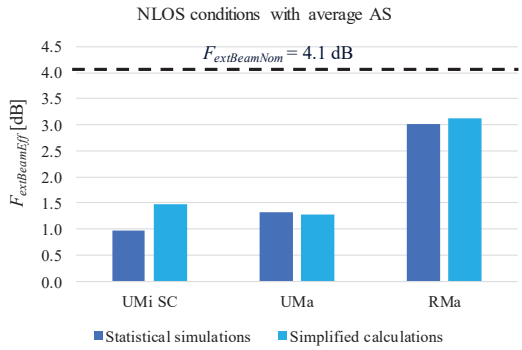
$$E_{asmt} = E_{broadcast} \cdot \sqrt{F_{extBeam} \cdot F_{BW} \cdot F_{PR} \cdot F_{TDC}}, \quad (7)$$

gdzie E_{asmt} i $E_{broadcast}$ to odpowiednio ekstrapolowane natężenie pola elektrycznego sygnału dla kanału użytkownika w V/m i oszacowane (zmierzone) natężenie pola elektrycznego sygnału dla kanału rozgłoszeniowego w V/m. $F_{extBeam}$ jest współczynnikiem ekstrapolacji odpowiadającym stosunkowi obwiedni EIRP ze wszystkich sygnałów dla kanałów użytkownika do obwiedni EIRP ze wszystkich sygnałów dla kanału rozgłoszeniowego w kierunku punktu pomiarowego. Współczynnik F_{BW} odpowiada stosunkowi szerokości pasma dla całego łącza w dół do odstępu w dziedzinie częstotliwości pomiędzy podnośnych sygnału rozgłoszeniowego. Z kolei współczynniki F_{PR} oraz F_{TDC} dotyczą redukcji mocy (jeśli stosuje się rzeczywiste maksymalne podejście) i maksymalnego stosunku czasu trwania transmisji w łączu w dół do długości trwania całego cyklu dla założonego typu dupleksu. Równanie (7) pochodzi się z normy IEC 62232, która zawiera dokładniejszy opis wszystkich powyższych parametrów.

Uzyskane wyniki symulacji, jak przedstawiono na rysunku 4 z publikacji [8], wskazują, że współczynnik ekstrapolacji może być przeszacowany w warunkach NLOS nawet o 4,4 dB, 5,4 dB i 4,1 dB odpowiednio dla założonych scenariuszy UMi SC, UMa i RMa, jeśli występują wysokie wartości kątowego rozkładu mocy. Bezpośrednią przyczyną tego przeszacowania są różnice między nominalnymi i skutecznymi charakterystykami antenowymi dla kanałów rozgłoszeniowych i kanałów użytkownika, jak przedstawiono na Rys.1 (a) (rysunek 3 z [8]). Konsekwencją tego przeszacowania mogą być niepotrzebnie ograniczone moce sygnałów w łączu w dół, lub zbyt duże strefy ochronne.

To potencjalne przeszacowanie można zredukować, stosując skuteczne charakterystyki w środowiskach z rozproszeniami, w celu lepszego odwzorowania rzeczywistych warunków propagacji, takich jak te występujące na obszarach miejskich lub gęsto zabudowanych. W takich rzeczywistych warunkach propagacji zaleca się wykonanie ekstrapolacji maksymalnej ekspozycji na PEM za pomocą odpowiednich modeli kanałów, zwłaszcza dla kątowego rozkładu mocy, ze względu na fakt, że zastosowanie kształtowania wiązki i filtrowania przestrzennego jest wrażliwe na zmienne w czasie warunki kanału radiowego. Prostą metodę poprawy dokładności ekstrapolacji ekspozycji na PEM na podstawie wyników pomiarów sygnałów dla kanałów rozgłoszeniowych przedstawiono w sekcji IV publikacji [8], wykorzystując wspólne modelowanie charakterystyki antenowej i kątowego rozkładu mocy.

Zastosowanie przedstawionej metody jest ograniczone do przypadków, w których maksymalna ekspozycja na PEM jest badana w częściach komórki 5G, gdzie skierowane są listki główne wiązek antenowych odpowiadających sygnałom dla kanałów rozgłoszeniowego i użytkownika, ponieważ pozwala ona jedynie na przybliżenie maksymalnego skutecznego zysku danej charakterystyki antenowej. Ta metoda sprawdza się dla technik formowania wiązek opartych na książce kodowej, gdzie charakterystyki i zyski wiązek rozgłoszeniowych i użytkownika są wstępnie zdefiniowane i znane przed wdrożeniem. W tym sensie proponowana metoda może być wykorzystana w przypadku implementacji formowania wiązki takich jak przemiatanie wiązek lub grid of beam (GoB), SU-MIMO, MU-MIMO i mMIMO. W bardziej zaawansowanych typach formowania wiązki, takich jak eigen-based Beamforming (EBB), gdzie współczynniki wagi wiązki (BWF) są określane w czasie rzeczywistym na podstawie informacji o stanie kanału (CSI), ogólna koncepcja ekstrapolacji maksymalnej ekspozycji oparta na porównaniu wiązek rozgłoszeniowych i użytkownika nie ma zastosowania, ponieważ charakterystyki i zyski tych wiązek nie są z góry znane.



Rys. 6. Porównanie wyników symulacji i obliczeń dla wartości skutecznego współczynnika ekstrapolacji w warunkach NLOS dla środowisk 3GPP UMi SC, UMa oraz RMa przy założonej średniej wartości kątowego rozkładu mocy

Na Rys. 6 (rysunek 6 z [8]) porównano symulowane wartości $F_{extBeamEff}$ uzyskane z symulacji 3D przy założeniu warunków NLOS dla scenariuszy UMi SC, UMa i RMa oraz odpowiadające im wyniki obliczeń według wspomnianej uproszczonej metody. Jak można zauważyć, obliczony współczynnik ekstrapolacji wykazuje dobrą zgodność z wynikami symulacji statystycznych, jeśli w obu metodach stosuje się te same wartości kątowego rozkładu mocy. Zaproponowana metoda obliczeniowa jest jednak mniej czasochłonna obliczeniowo niż pełne symulacje statystyczne. Jednocześnie metoda ta daje przegląd statystycznie reprezentatywnego zakresu współczynnika ekstrapolacji, jeśli do obliczeń wybrano właściwy rozkład statystyczny kątowego rozkładu mocy, tj. założono jego wartość średnią i odchylenie standardowe. Nawet jeśli dokładna wartość kątowego rozkładu mocy w punkcie szacowania maksymalnej ekspozycji na PEM nie jest znana, zakres obliczonego współczynnika ekstrapolacji pozwala wybrać wartość, która prowadzi do najbardziej bezpiecznego oszacowania ekspozycji, ale jedno-

cześnie nie jest nadmiernie zawyżona, ponieważ dla ekstrapolacji zakłada się skuteczne zyski wiązki rozgłoszeniowej i użytkownika zamiast odpowiednich zysków nominalnych. Wnioski te pozwoliły umieścić prezentowaną metodę w najnowszej wersji normy IEC 62232, w sekcji B.8. (*Extrapolation for massive MIMO and beamforming BS*).

4. WNIOSKI

Wynikiem wszystkich badań przeprowadzonych i opublikowanych w pracach [1] – [8] jest kilka wniosków, które wskazują, w jaki sposób pominięcie wzajemnego modelowania formowania wiązki i kątowego rozkładu mocy w środowisku w rozproszeniami może zniekształcać szacunki wydajności systemu 5G i negatywnie wpływać na proces planowania i optymalizacji całej sieci radiowej 5G. Wnioski te potwierdzają ogólną tezę stojącą za przeprowadzonymi badaniami, a która sprowadza się do stwierdzania, że *odwzorowanie wpływu zjawiska kątowego rozproszenia mocy na charakterystykę układu wieloantenowego znacząco poprawia dokładność szacowania budżetu łącza radiowego oraz wydajność procesu planowania i optymalizacji sieci radiowej 5G*. Zgodnie z tą tezą w wyniku przeprowadzonych badań wyciągnięte zostały poniższe wnioski:

- Możliwe jest wykorzystanie koncepcji skutecznej charakterystyki układu wieloantenowego do poprawy dokładności w obliczeniach uproszczonego budżetów łącza radiowego;
- Możliwa jest optymalizacja geometrii układu wieloantenowego w celu zwiększenia jego skutecznego zysku w określonych warunkach kątowego rozkładu mocy, wraz z metodą szacowania kątowego rozkładu mocy w oparciu o minimalną liczbę wyników pomiarów (Międzynarodowe Zgłoszenie Patentowe nr PCT/IB2019/053142);
- Metoda obliczania skutecznego współczynnika ekstrapolacji poprawia dokładność szacowania maksymalnej ekspozycji na PEM w środowiskach z rozproszeniami, przez co została wymieniona w normie IEC 62232.

LITERATURA

[1] Bechta Kamil, Rybakowski Marcin, Hsieh Frank, D. Chizhik Dmitry. 2018. "Modeling of radio link budget with beamforming antennas for evaluation of 5G systems". *2018 IEEE 5G World Forum (5GWF)*. DOI: 10.1109/5GWF.2018.8516969.

[2] Bechta Kamil, Rybakowski Marcin, Du Jinfeng. 2019. "Impact of Effective Antenna Pattern on Millimeter Wave System Performance in Real Propagation Environment". *European Conference on Antennas and Propagation (EuCAP)*.

[3] Bechta Kamil, Rybakowski Marcin, Du Jinfeng. 2019. "Efficiency of Antenna Array Tapering in Real Propagation Environment of Millimeter Wave System". *European Conference on Antennas and Propagation (EuCAP).*

[4] Bechta Kamil, Du Jinfeng, Rybakowski Marcin. 2019. "Optimized Antenna Array for Improving Performance of 5G mmWave Fixed Wireless Access in Suburban Environment", *IEEE 5G World Forum (5GWF)*. DOI: 10.1109/5GWF.2019.8911630.

[5] Du Jinfeng, Rybakowski Marcin, Bechta Kamil, Valenzuela Reinaldo. 2019. "Matching in the Air: Optimal Analog Beamforming under Angular Spread", *under review before planned submission to IEEE Transactions on Communications*. (arxiv:1910.11054)

[6] Bechta Kamil, Du Jinfeng, Rybakowski Marcin. 2020. "Rework the Radio Link Budget for 5G and Beyond," *IEEE Access*. vol. 8, 211585–211594. DOI: 10.1109/ACCESS.2020.3039423.

[7] Bechta Kamil, Grangeat Christophe, Du Jinfeng. 2020. "Impact of Effective Antenna Pattern on Radio Frequency Exposure Evaluation for 5G Base Station with Directional Antennas". *XXXIIIrd General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS 2020)*. DOI: 10.23919/URSIGASS49373.2020.9232371.

[8] Bechta Kamil, Grangeat Christophe, Du Jinfeng, Rybakowski Marcin. 2022. „Analysis of 5G Base Station RF EMF Exposure Evaluation Methods in Scattering Environments". *IEEE Access*. vol. 10, 7196 - 7206. DOI: 10.1109/ACCESS.2022.3142673

METODY WYZNACZANIA GĘSTOŚCI MOCY W POLU BLISKIM I DALEKIM WOKÓŁ ANTEN PARABOLICZNYCH LINII RADIOWYCH

METHODS FOR DETERMINING POWER DENSITY IN THE NEAR FIELD AND FAR FIELD AROUND RADIO LINK PARABOLIC ANTENNAS

Waldemar Karcz
Emitel S.A., Szczecin, waldemar.karcz@emitel.pl

DOI: 10.15199/59.2022.4.58

Streszczenie: **Zmiany dopuszczalnych poziomów PEM w środowisku miały ułatwić realizację inwestycji w Polsce. Efekt ten nie został jednak osiągnięty, ponieważ w przepisach wykonawczych nadal jako kryterium instalacji mogących negatywnie oddziaływać na środowisko obowiązuje aksjomat „moc EIRP niemniejsza niż 15 W". Również kryteria zaliczenia instalacji do mogących zawsze znacząco oddziaływać na środowisko wymagają aktualizacji. Ocena oddziaływania instalacji na środowisko musi być oparta o sprawdzenie dotrzymywania dopuszczalnych limitów PEM.**

Abstract: **Changes in the allowed environmental EMF levels were supposed to investment projects in Poland. However, this goal has not been achieved since executory provisions still include "EIRP of not less than 15 W" as criterion for potential negative environmental impact of installations. The criteria for installations that can always a negative impact on the environment also require updating. Environmental impact assessment for installations must be based on verifying whether they meet EMF limits.**

Słowa kluczowe: **dopuszczalne poziomy PEM, gęstość mocy, sonda pomiarowa.**

Keywords: **acceptable levels of EMF, power density, measuring probe.**

1. WSTĘP

Zmiany dopuszczalnych poziomów pól elektromagnetycznych w środowisku zostały wprowadzone Rozporządzeniem Ministra Zdrowia z dnia 17 grudnia 2019 r. [6] i weszły w życie 1 stycznia 2020 r. Jednocześnie z dniem 1 stycznia 2020 r. przestało obowiązywać Rozporządzenie Ministra Środowiska z dnia 30.10.2003 r. Rozporządzenie to zostało zastąpione Rozporządzeniem Ministra Klimatu z dnia 17 lutego 2020 r. [5] w sprawie sposobów dotrzymywania dopuszczalnych poziomów pól elektromagnetycznych w środowisku. Pozytywna zmiana dopuszczalnych poziomów pól w środowisku wprowadziła nowe limity PEM dla poszczególnych zakresów częstotliwości zgodne z wytycznymi ICNIRP. Niestety pozostałe rozporządzenia nie zostały zaktualizowane i dodatkowo w ramach tak zwanego „sprawnego i szybkiego procesu legislacyjnego" utrzymano w POŚ artykuł 122a oraz utrzymano w mocy obowiązek zgłaszania instalacji radiokomunikacyjnych, których anteny emitują moc EIRP ≤15W. Pomimo zwiększenia, dla zakresu częstotliwości powyżej 2 GHz, dopuszczalnego poziomu gęstości mocy z 0,1 W/m^2 do 10 W/m^2 ustawodawca w ustawie [8] POŚ w § 338 narzuca obowiązek wykonania pomiarów PEM w środowisku przez akredytowane laboratoria bez względu na fakt, że poziom PEM pochodzący od radiolinii jest kilkaset razy mniejszy od czułości sond pomiarowych wykorzystywanych do pomiarów przez te laboratoria. Art. 338a „Kto, będąc obowiązany do wykonywania pomiarów pól elektromagnetycznych w środowisku, na podstawie art. 122a, nie wykonuje tych pomiarów, podlega karze aresztu albo ograniczenia wolności albo karze grzywny".

2. KRYTERIUM MOCY EIRP ≥15W

2.1. Geneza określenia instalacji mogących negatywnie oddziaływać na środowisko ze względu na moc EIRP ≥15 W

Określenie „ moc EIRP niemniejsza niż 15 W" pojawia się po raz pierwszy w Ustawie Prawo ochrony środowiska [8] jako kryterium kwalifikujące instalację do obowiązku uzyskania pozwolenia na emisję. Art. 234 miał następujące brzmienie: „Pozwolenie na emitowanie pól elektromagnetycznych jest wymagane dla instalacji radiokomunikacyjnych, radionawigacyjnych i radiolokacyjnych, których równoważna moc promieniowana izotropowo jest równa 15 W lub wyższa, emitujących pola elektromagnetyczne o częstotliwościach od 0,03 MHz do 300 000 MHz". Następnie w 2005 roku wprowadzone zostały zmiany w POŚ [8] polegające na rezygnacji z obowiązku uzyskiwania pozwolenia na emisję. Został usunięty artykuł 234, a do rozdziału 6 został dodany kontrowersyjny art. 122a zobowiązujący prowadzących instalację oraz użytkowników urządzeń emitujących pola elektromagnetyczne o częstotliwościach od 30 kHz do 300 GHz, których równoważna moc promieniowana izotropowo wynosi nie mniej niż 15 W, do wykonania pomiarów poziomów pól elektromagnetycznych w środowisku.

2.2. Instalacje mogące negatywnie oddziaływać na środowisko a moc EIRP

Obecnie kryterium mocy EIRP ≥ 15 W występuje w rozporządzeniach wykonawczych jako swoisty dogmat nie podlegający sprawdzeniu. Analizując stan faktyczny należy zauważyć, że parametrem określającym narażenie ludzi na promieniowanie jest wartość skuteczna natężenia pola elektrycznego lub wartość średnia gęstości mocy. Gęstość mocy w dowolnym punkcie obserwacji wyznacza się z następujących zależności:

$$S = \frac{P \cdot G \cdot F_{f(\varphi,\theta)}}{4\pi \cdot d^2}; E = \sqrt{\frac{30 \cdot P \cdot G \cdot F_{f(\varphi,\theta)}}{d^2}} \qquad (1)$$

gdzie:

S – gęstość mocy [W/m²]

E – natężenie pola [V/m]
P – moc na wejściu anteny [W]
G – zysk anteny w odniesieniu do anteny izotropowej
$F_{f(\varphi,\theta)}$ – funkcja tłumienia gęstości mocy
d – odległość od środka elektrycznego anteny do punktu, w którym wyznaczana jest gęstość mocy

2.3. Ochrona przed polami elektromagnetycznymi

Ochrona przed polami elektromagnetycznymi polega na zapewnieniu jak najlepszego stanu środowiska poprzez utrzymanie poziomów pól elektromagnetycznych poniżej dopuszczalnych wartości określonych w Rozporządzeniu Ministra Zdrowia [6]. Ten cel osiąga się poprzez:

a) instalowanie anten na wysokich podporach
b) dobór anten o kształtowanej odpowiednio charakterystyce w płaszczyźnie pionowej i poziomej
c) dobór optymalnej mocy EIRP
d) zachowanie odpowiedniej odległości od miejsc dostępnych dla ludności

Parametrami charakteryzującymi oddziaływanie na środowisko są gęstość mocy lub natężenie pola. Znajomość wartość mocy EIRP dotyczącej danej instalacji jest warunkiem koniecznym, ale niewystarczającym aby przesądzać o wartości gęstości mocy lub o wartości natężenia pola w miejscach dostępnych dla ludności. Miejsca dostępne dla ludności znajdują się w obszarach pola dalekiego, gdzie prawdziwa jest następująca zależność:

$$S = |E| \cdot |H| = \frac{E^2}{120 \cdot \pi} = 120 \cdot \pi \cdot H^2 \qquad (2)$$

gdzie:

E – wartość skuteczna natężenia pola elektrycznego
H – wartość skuteczna natężenia pola magnetycznego

3. ROZKŁAD GĘSTOŚCI STRUMIENIA MOCY W POLU BLISKIM I DALEKIM

3.1. Anteny paraboliczne

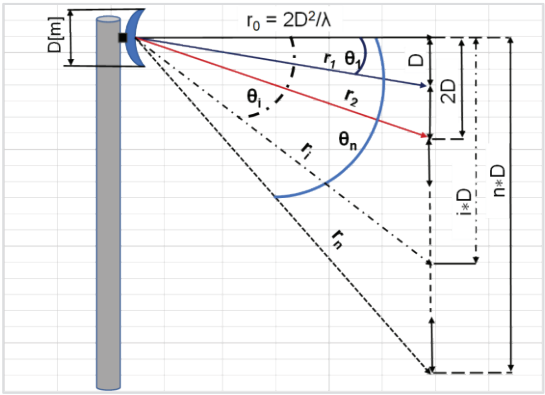
Linie radiowe pracują z mocą nadajników w zakresie od 5 mW do 1 W. Najważniejsze cechy anten parabolicznych to duża kierunkowość i niskie promieniowanie poza kierunkiem wiązki głównej. Do obliczenia maksymalnej gęstości strumienia mocy z tych anten potrzebne jest tylko kilka parametrów takich jak częstotliwość, wydajność apertury i zysk Gi (od 30 do 46 dBi).

3.2. Rozkład gęstości strumienia mocy w osi głównej i w miejscach dostępnych dla ludności.

Przestrzeń wokół anteny jest dzielona na obszar bliski pola indukcyjnego, obszar bliski pola promieniowania i obszar pola dalekiego. W obszarze pola indukcyjnego wiązki głównej gęstość strumienia mocy może osiągać maksimum zanim zacznie się zmniejszać wraz z odległością. W obszarze bliskim pola promieniowania gęstość strumienia mocy zmienia się odwrotnie proporcjonalnie do odległości. Gęstość mocy w polu dalekim maleje odwrotnie proporcjonalnie do kwadratu odległości. Zgodnie z załącznikiem do Rozporządzeniem Ministerstwa Klimatu z dnia 17 lutego 2020 r. pomiary pól elektromagnetycznych w zakresach częstotliwości od 10 MHz do 300 GHz wykonuje się w szczególności w tych miejscach , w których, na podstawie uprzednio przeprowadzonych obliczeń, stwierdzono występowanie w danych zakresach częstotliwości pól elektromagnetycznych o poziomach zbliżonych do poziomów dopuszczalnych, określonych w przepisach wydanych na podstawie art. 122 ust. 1 ustawy z dnia 27 kwietnia 2001 r. [8]

3.3 Metoda wyznaczenia rozkładu gęstości strumienia mocy w pionach pomiarowych

Miejsca dostępne dla ogółu ludności znajdują się w obszarze pola dalekiego. Pomiary pól elektromagnetycznych wykonuje się w pionach pomiarowych. Pionem pomiarowym jest linia pionowa, wzdłuż której przemieszczana jest sonda miernika w celu wyznaczenia maksymalnych poziomów pól elektromagnetycznych w środowisku. Obliczenia, o których mowa w załączniku do Rozporządzenia [4], zostały przeprowadzone w pionach pomiarowych odległych od anteny o $0{,}6D^2/\lambda$, D^2/λ i $2D^2/\lambda$. Zależności punktów pomiarowych od kątów θ_i w pionie pomiarowym w odległości $2D^2/\lambda$ przedstawione zostały na Rys. 1.



Rys. 1. Pion w odległości $2D^2/\lambda$ i punkty pomiarowe w odległości od osi głównej unormowanej przez D

Zastosowanie modelu sferycznego do obliczeń gęstości mocy w odległości $0{,}6D^2/\lambda$ i w odległości D^2/λ jest dopuszczalne, ponieważ w ten sposób otrzymujemy w tych odległościach poziomy PEM wyższe o około 20% od wyznaczonych metodami stosownymi w polu bliskim.

Poziom gęstości strumienia mocy w polu dalekim Di wyznacza się z następującej zależności:

$$S_{i(2)} = \frac{P \cdot G \cdot F(\theta_i)_{(2)}}{4 \cdot \pi \cdot \left(i^2 \cdot D^2 + r_{0(2)}^2\right)} \qquad (3)$$

gdzie:

$S_{i(2)}$ – gęstość mocy w [W/m²] w punkcie D_i, w pionie pomiarowym w odległości $2D^2/\lambda$ od anteny
P – moc na wejściu anteny [W]
G – zysk anteny [W/W]
$F(\theta_i)$ – współczynnik tłumienia mocy przy kącie θ_i
D – średnica anteny [m]
i – liczba z przedziału od 0 do n
$r_{0(2)}$ – odległość w osi głównej do pionu pomiarowego $2D^2/\lambda$
$r_{0(06)}$ – odległość pionu pomiarowego $0{,}6D^2/\lambda$
$r_{0(1)}$ – odległość pionu pomiarowego $1D^2/\lambda$

Ponieważ podstawowe dane takie jak moc, zysk anteny i współczynnik tłumienia mocy są podawane w mierze decybelowej, to gęstość mocy można wyznaczyć z następującej zależności:

$$S_{i(2)} = P + G + F_{i(2)} - 10\log 4\pi - 10\log r_{i(2)}^2 \quad (4)$$

gdzie:

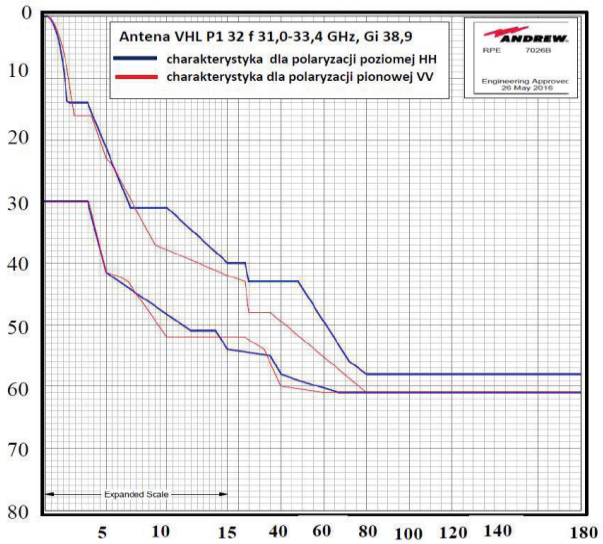
$S_{i(2)}$ – gęstość mocy w punkcie D_i, [dB] w pionie pomiarowym w odległości $2D^2/\lambda$
P – moc na wejściu anteny [dBW]
G – zysk anteny [dB]
$F(\theta_i)$ – tłumienie mocy [dB] przy kącie θ_i

Długość fali λ można wyznaczyć dzieląc prędkość rozchodzenia się fali elektromagnetycznej w próżni przez jej częstotliwość.

$$\lambda[m] = \frac{0,3}{f[GHz]} \quad (5)$$

$$r_{i(2)}^2 = i^2 \cdot D^2 + \left[\frac{2 \cdot D^2}{0,3} \cdot f[GHz]\right]^2 \quad (6)$$

Wartość tłumienia mocy w zależności od kąta elewacji θ_i wyznacza się korzystając z charakterystyki anteny, która jest dostępny na stronach producenta. (Rys. 2.)



Rys. 2. Charakterystyki anteny VHL P1 32 dla polaryzacji poziomej HH,VH i pionowej VV.VH

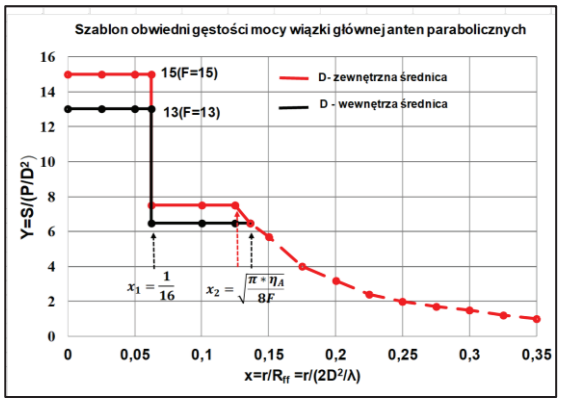
W celu odczytania wartości tłumienia mocy $F(\theta_i)$ należy wyznaczyć tangens kąta θ_i jako stosunek odcinka i ·D do odległości od anteny w osi głównej do pionu pomiarowego odpowiednio równej $06D^2/\lambda$, D^2/λ i $2D^2/\lambda$. Następnie za pomocą funkcji arctg wyznaczony zostanie kąt θ_i. Na przykład jeżeli θ_i wyniesie 10°, to $F(\theta_i)$ przy polaryzacji H wyniesie -31 dB, a przy polaryzacji V wyniesie -38 dB. W tabeli 1 zestawione zostały wyniki rozkładu gęstości strumienia mocy w wiązce głównej oraz w pobliżu tej wiązki i w miejscach dostępnych dla ludności na poziomie 29,6 m i 30 m poniżej osi głównej. Obliczenia dotyczyły radiolinii pracującej w paśmie 32 GHz z mocą doprowadzoną do anteny 24 dBm (0,25 W). W przypadku zwiększenia mocy doprowadzonej do anteny o 6 dB poziomy gęstości mocy zestawione tabeli 1 wzrosną czterokrotnie, a przy zmniejszeniu mocy do 21 dBm poziomy gęstości mocy przedstawione w tabeli zmniejszą się dwukrotnie.

Tab. 1. Rozkład gęstości mocy w pionach pomiarowych i w miejscach dostępnych dla ludności

h=i·D	*Punkty*	*Piony pomiarowe*			*Jednostka*
[m]	*i*	*0,6D²/λ*	*D²/λ*	*2D²λ*	
oś gł.	0	4,6	1,7	0,4	W/m²
0,3	1	0,1	0,3	0,26	W/m²
0,6	2	0,01	0,04	0,06	W/m²
0,9	3	0,001	0,007	0,001	W/m²
1,2	4	0,0006	0,002	0,001	W/m²
29,4	98	0,15	0,26	0,5	μW/m²
30	100	0,14	0,23	0,46	μW/m²

3.4 Metody sprawdzania zgodności poziomów PEM z wytycznymi ICNIRP

Gęstość mocy w polu dalekim jest odwrotnie proporcjonalna do kwadratu odległości od anteny, zatem gęstość mocy maleje bardzo szybko wraz z odległością, a najwyższe poziomy mocy występują w polu bliskim (w strefie Fresnela). Gęstość mocy w polu bliskim oscyluje w zależności od odległości pomiędzy maksymami i minimami w obszarze średnicy anteny [2]. W związku z potrzebą ustalania granicy zgodności z limitami opracowany został szablon dla gęstości mocy (Rys. 3). Szablon ten [1] wyznacza granicę większą niż otrzymane wyniki z przeprowadzonych symulacji i z wykonanych pomiarów zgodnych normą EN 50383.



Rys. 3. Szablon obwiedni dla gęstości mocy w głównej wiązce anten parabolicznych

Do wyznaczenia odległości zgodności z limitami dla ogółu społeczeństwa musimy znać wartość mocy dostarczonej do anteny, zysk anteny oraz średnicę anteny. Na przykład dla anteny o średnicy 0,6 m zasilanej mocą 24 dBm (0,25 W), o zysk anteny 36 dBi, f=13GHz wyznacza się ze wzoru (7)

$$S_{max} = \gamma \cdot F \cdot \frac{P}{D^2} = 0,8 \cdot 15 \cdot \frac{0,25}{0,36} = 8,33\ W/m^2 \quad (7)$$

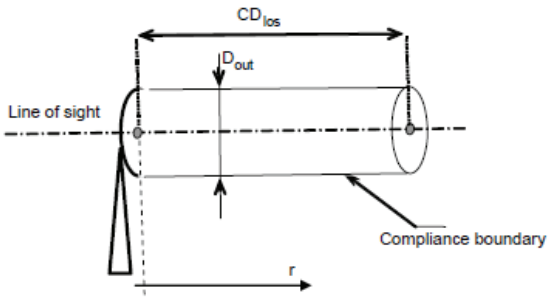
Jeżeli $S_{max} < S_{limit}$, to $CD_{los} = 0$ Oznacza to, że taka instalacja jest w pełni zgodna i nie wymaga dalszego sprawdzania.

Jeżeli zwiększymy moc z 24 do 27 dBm, to

$$S_{max} = \gamma \cdot F \cdot \frac{P}{D^2} = 0{,}8 \cdot 15 \cdot \frac{0{,}5}{0{,}36} = 16{,}7\ W/m^2 \quad (8)$$

Ponieważ $0{,}5 \cdot S_{max} < S_{limit}$, to $CD_{los} = R_{ff}/16 = 0{,}98$ m
W takim przypadku to instalacja jest zgodna z limitami w odległości o 1 m w osi głównej.
Jeżeli zwiększymy moc do 30 dB, to ponieważ $0{,}5 \cdot S_{max} > S_{limit}$, to odległość od anteny wyznacza się

$$CD_{los} = \sqrt{\frac{P \cdot 10^{\frac{G}{10}}}{4\pi \cdot S_{limit}}} = 5{,}6\ m \quad (9)$$

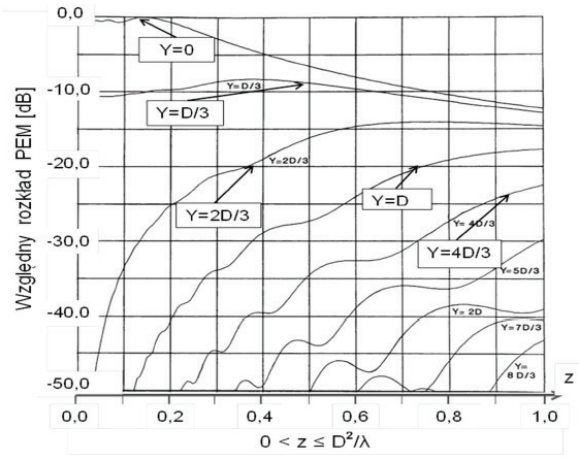


Rys. 4. Granica zgodności radiolinii z dopuszczalnymi limitami

Metody sprawdzania zgodności z wytycznymi ICNIRP ograniczają się do sprawdzania zgodności z limitami w polu bliskim, gdzie można się spodziewać najwyższych poziomów gęstości mocy (Rys. 4).

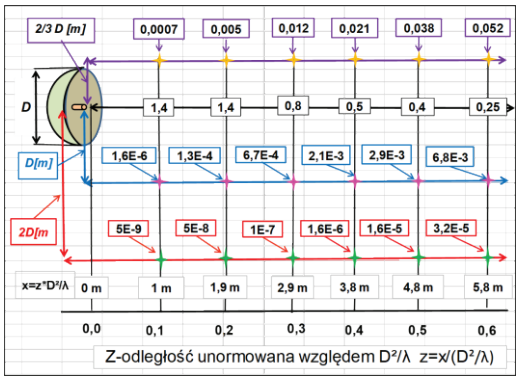
3.5 Rozkład gęstości strumienia mocy w polu bliskim anten radioliniowych instalowanych na dachach

Metoda opracowana przez R. Lewis i A. Newell [3] pozwala w prosty sposób wyznaczyć gęstość mocy w dowolnym punkcie obserwacji w strefie bliskiej pola anteny parabolicznej.



Rys. 5. Względny rozkład gęstości mocy wzdłuż linii równoległych do osi głównej oddalonych od siebie o D/3

Posługując się wykresem względnego rozkładu gęstości mocy dla anten o D>30λ (Rys. 5) zostały wyznaczone wskaźniki S/S_{limit} dla radiolinii 32 GHz, pracującej z mocą P= 23 dBm, z anteną o średnicy D= 0,3 m i zysku, G= 38 dBi , S_{limit} = 10 W/m^2.



Rys. 6. Wskaźniki S/S_{limit} w osi głównej i w osiach równoległych oddalonych o 2/3 D, D, i 2D

Największy poziom gęstości mocy występuje w osi głównej anteny w odległości do 2 m (Rys, 6). W punktach na liniach równoległych do osi głównej odległych o 2/3 D, o D i 2D wskaźnik S/S_{limit} jest wielokrotnie mniejszy.

4. POMIARY PEM W ŚRODOWISKU

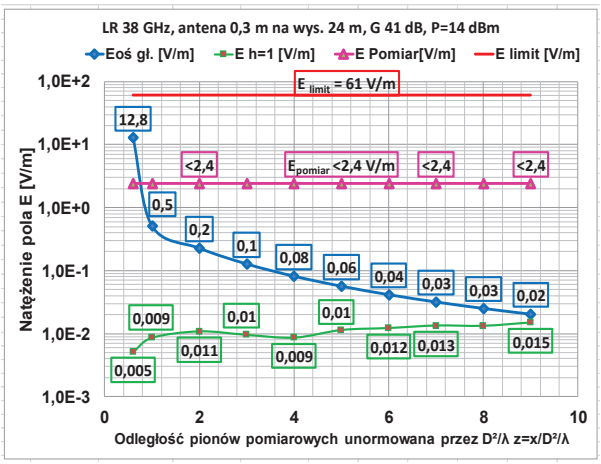
4.1 Pomiary stan prawny

Z dniem wejścia w życie Rozporządzenia Ministra Zdrowia w sprawie dopuszczalnych poziomów pól elektromagnetycznych w środowisku [6] przestało obowiązywać Rozporządzenie Ministra Środowiska z dnia 30 października 2003 r. Od 18 lutego obowiązuje Rozporządzenie Ministra Klimatu [5] w sprawie sposobów sprawdzania dotrzymywania dopuszczalnych poziomów pól elektromagnetycznych w środowisku. W związku z wprowadzeniem zróżnicowanych dopuszczanych poziomów pól w zależności od zakresu częstotliwości pomiary PEM w środowisku stały się bardzo złożone i pracochłonne. Sposoby sprawdzania dotrzymywania dopuszczalnych poziomów pól w środowisku są szczegółowo opisane w załączniku do rozporządzenia [5]. Pomiary poziomów pól elektromagnetycznych w otoczeniu instalacji radiokomunikacyjnych wykonuje się wzdłuż przyjętych pionów pomiarowych, w punktach na wysokościach od 0,3 m do 2 m, a za wynik pomiaru przyjmuje się największą wartość zmierzonej gęstości mocy lub natężenia pola. Celem pomiarów jest wyszukanie największych wartości zmierzonych z całego obszaru, na którym prowadzone były pomiary. W sprawozdaniu z pomiarów dokonuje się porównania największych wartości zmierzonych z danych obszarów i z danych zakresów częstotliwości z dopuszczalnymi poziomami pól zawartymi w załączniku do Rozporządzenia Ministra Zdrowia [6]. Zachowany został w załączniku do rozporządzenia [5] ust. 5 pkt. 2 mówiący, że pomiary wykonuje się w szczególności w tych miejscach, w których, na podstawie uprzednio przeprowadzonych obliczeń, stwierdzono poziomy pól o wartościach zbliżonych do dopuszczalnych. Przeprowadzenie obliczeń przed przystąpieniem do pomiarów powinno pozwolić na uniknięcie wykonywania pomiarów w pionach i punktach, o których z góry wiadomo, że tam nie wystąpią poziomy natężenia pola o wartościach istotnych dla ochrony środowiska. Niestety w załączniku do rozporządzenia [5] znalazł się w ust.1 pkt 1 zapis o obowiązującym art. 122a z ustawy [8]. Na skutek wprowadzenia tego zapisu

powstaje obowiązek zlecania pomiarów PEM akredytowanym laboratoriom dla instalacji, od których poziom gęstości mocy jest wielokrotnie niższy od czułości sond pomiarowych wykorzystywanych przez te laboratoria. Pomimo braku zgodności takiego postępowania z opisaną powyżej definicją mezurandu oraz braku zgodności z podstawami metrologii, pomiary PEM dla takich instalacji są w praktyce powszechnie zlecane. Nie jest to wynikiem siły logicznych argumentów, to wynika z argumentu siły zawartego w Art. 338a „Kto, będąc obowiązany do wykonywania pomiarów pól elektromagnetycznych w środowisku, na podstawie art. 122a, nie wykonuje tych pomiarów, podlega karze aresztu albo ograniczenia wolności albo karze grzywny". W załączniku do rozporządzenia [5] jest szczegółowy opis jak należy wykonywać pomiary dla instalacji długofalowych, krótkofalowych instalacji UKF i TV, instalacji radiolokacyjnych, instalacji służb mobilnych i amatorskich, ale nie ma wytycznych do wykonywania pomiarów PEM wokół instalacji radioliniowych. Ten sam problem występował podczas obowiązywania Rozporządzenia z października 2003 roku. W okresie ostatnich 40 lat nie udało się zmierzyć poziomów PEM od radiolinii przekraczających poziom 0,1 W/m^2. W artykule Pani Haliny Aniołczyk opublikowanym KKRRiT z 2002 r. w podsumowaniu z pomiarów PEM jest jednoznaczna informacja: „Nie stwierdza się mierzalnych wartości PEM na poziomie przebywania ludzi od anten linii radiowych horyzontalnych , wytwarzających, w ściśle określonych kierunkach, wąskie wiązki promieniowania elektromagnetycznego w zakresie częstotliwości od 2 GHZ do18 GHz z łączną mocą wyjściową nadajników od 5 mW do 20 W".

4.2 Zestawienie wyników pomiarów z wykonanymi obliczeniami

1. Linia radiowa 38 GHz z anteną 0,3 m na wysokości 24 m n.p.t.



Rys. 7. Zestawienie wyników pomiaru natężenia pola z obliczeniami natężenia pola na wys. 1 m n.p.t.

Rys. 7 zawiera zestawienie wyników pomiarów natężenia pola elektromagnetycznego pochodzącego od linii radiowej 38 GHz, z anteną 0,3 m zainstalowanej na elewacji budynku na wysokości 24 m n.p.t. Moc doprowadzona do anteny 14 dBm. Pomiary wykonane zostały przez akredytowane laboratorium miernikiem MEH 25 z sondą AS-3. (zakres pomiarowy 2,4-70 [V/m]). Natężenie pola we wszystkich pionach i punktach pomiarowych było mniejsze od 2,4 V/m. Wykonane obliczenia zgodnie z punktem 5, podpunkt 2 Rozporządzenia [5] zostały przedstawione na Rys.7. Najniżej położona krzywa przedstawia rozkład natężenie pola w miejscach dostępnych dla ludności. Największa wartość obliczona na poziomie h =1m n.p.t. wynosi 15 mV/m. Zostały wykonane obliczenia również przy zwiększeniu mocy z 14 do 30 dBm. Przy zwiększeniu mocy do 30 dBm poziomy natężenia pola w punktach i w pionach zwiększą się 6,3 razy tj. 15 mV/m * 6,3 = 94,5 mV/m.

2. Linia radiowa 32 GHz z anteną 0,3 m zainstalowaną na wysokości 300 m n.p.t. Moc doprowadzona do anteny wynosi 20,5 dBm (0,112 W). Akredytowane laboratorium wykonało pomiary natężenia pola w 16 pionach pomiarowych (Tab. 2). Pierwszy pion znajdował się 350 m od podstawy masztu (tj. 1 m od ogrodzenia obiektu), a dziesiąty pion w odległości 650 m i 15 w odległości 750 m od podstawy masztu. W sprawozdaniu z pomiarów we wszystkich pionach pomiarowych zmierzone natężenie pola wynosiło mniej niż 2V/m, czyli było poniżej dolnej granicy aparatury pomiarowej użytej przez akredytowane laboratorium.

Tab. 2. Zestawienie wartości wielkości obliczonych i pomierzonych

	Wartości obliczone			Pomiar
Pion	*E[mV/m]*	*S[nW/m^2]*	*S/S_{limit}*	*E[V/m]*
1	1,1	3,4	$3{,}4*10^{-10}$	< 2
3	1,1	3,7	$3{,}7*10^{-10}$	< 2
10	1,4	5,4	$5{,}4*10^{-10}$	< 2
15	1,5	5,5	$5{,}5*10^{-10}$	< 2

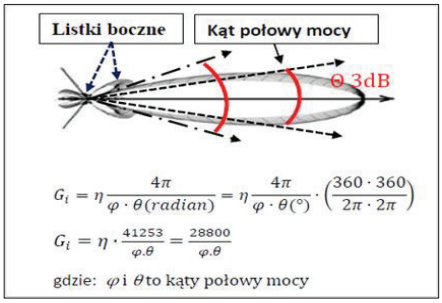
Piony pomiarowe są oddalone od podstawy masztu o: pion 1 -350 m, pion 3 – 450 m, pion 10- 650 m i pion 15 750 m.

Sondy i przyrządy pomiarowe używane przez akredytowane laboratoria są przystosowane do sprawdzania, czy dana instalacja nie przekracza dopuszczalnych limitów, ale nie można ich stosować do pomiaru znikomo niskich poziomów PEM pochodzących od radiolinii w miejscach dostępnych dla ludności.

5. PODSUMOWANIE

1. Kwalifikowanie instalacji do mogących znacząco oddziaływać na środowisko powinno odbywać się na podstawie prognozowanego rozkładu gęstości mocy lub natężenia pola z uwzględnieniem dopuszczalnych poziomów pól w danych zakresach częstotliwości.
2. Instalacje radioliniowe powinny być jednoznacznie wyłączone z obowiązku wykonywania pomiarów pól elektromagnetycznych w środowisku oraz nie powinny podlegać zgłoszeniu, ponieważ radiolinie nie mogą negatywnie oddziaływać na środowisko. W Rozporządzeniu Rady Ministrów [7] w sprawie przedsięwzięć mogących znacząco oddziaływać na środowisko instalacje radioliniowe nie zostały zaliczone do przedsięwzięć mogących

negatywnie oddziaływać na środowisko. Anteny radioliniowe są zasilane mocą nie większą niż 14-30 dBm (0,025W-1W), a kąt połowy mocy w płaszczyźnie poziomej i pionowej wynosi od 0,9 do 1,8 stopnia. Moc w kierunku miejsc dostępnych dla ludności jest od 30dB do 50 dB mniejsza od mocy w kierunku osi głównej anteny. Znając kąt połowy mocy w płaszczyźnie pionowej i poziomej można wyznaczyć z dość dobrą dokładnością zysk kierunkowy takiej anteny. Wzór znajduje się na Rys. 8.



Rys. 8. Zależność zysku anteny parabolicznej od kąta połowy mocy

3. Linie radiowe emitują pole stacjonarne o ściśle określonym rozkładzie przestrzennym i o ściśle określonej mocy doprowadzonej do anteny. Postępując zgodnie z punktem 5, podpunkt 2 Rozporządzenia [5] można wykonać obliczenia poziomu gęstości mocy i natężenia pola pochodzącego od radiolinii w miejscach dostępnych dla ludności. Wskaźnik S/S_{limit} będzie wielokrotnie niższy od wartości 1 i dużo mniejszy od czułości sond pomiarowych używanych do pomiarów przez akredytowane laboratoria.

LITERATURA

[1] ETSI TR 102 457 v2.1.0 (2018-09) *Fixed Radio Systems;Evaluation of the ElektroMagnetic Field(EMF) radiated by Line-of-Sight(LoS) fixed radio stations using parabolic dish directional antennas*

[2] ICNIRP *Guidelines for limiting exposure to timevarying electric, magnetic and electromagnetic fields (up to 300 GHz),* Health Physics, vol. no. 4. April 1998

[3] Lewis Richard L., Newell Alaen C. 1985. An Efficient and Accurate Method for Calculating and Representing Power Density in the Near-Zone of Microwave Antennas

[4] Obwieszczenie Ministra Środowiska z dnia 22 lipca 2019 r., w sprawie ogłoszenia jednolitego tekstu rozporządzenia Ministra Środowiska w sprawie rodzajów instalacji, których eksploatacja wymaga zgłoszenia Dz. U. 2019 poz.1510

[5] Rozporządzenie Ministra Klimatu z dnia 17 lutego 2020 r., *w sprawie sposobów sprawdzania dotrzymywania dopuszczalnych poziomów pól elektromagnetycznych w środowisku*, Dz. U. Nr 2020 poz. 258

[6] Rozporządzenie Ministra Zdrowia z dnia 17 grudnia 2019 r., *w sprawie dopuszczalnych poziomów pól elektromagnetycznych w środowisku*

[7] Rozporządzenie Rady Ministrów z dnia10 września 2019 r., *w sprawie przedsięwzięć mogących znacząco oddziaływać na środowisko*, Dz. U. Nr 2019 poz.1839

[8] Ustawa Prawo Ochrony Środowiska Dz. U. Nr 2001 Nr 62 poz. 627, z późniejszymi zmianami Dz. U. z 2021 r. poz. 1973, 2127, 2269, z 2022 r. poz. 1079

WYBÓR MODULACJI CYFROWEJ W NANOSATELITARNYM SYSTEMIE TRANSMISYJNYM

Krzysztof Wesołowski

Instytut Radiokomunikacji, Politechnika Poznańska
krzysztof.wesolowski@put.poznan.pl

DOI: 10.15199/59.2022.4.59

Streszczenie: **W pracy przeanalizowano wybór modulacji cyfrowych w systemie łączności nanosatelitarnej aktualnie budowanym w Instytucie Radiokomunikacji Politechniki Poznańskiej. Przedstawiono podstawowe wymagania transmisyjne projektowanego systemu oraz przeanalizowano zalety i wady modulacji cyfrowych rekomendowanych przez Komitet Konsultacyjny dla Kosmicznych Systemów Danych (CCSDS). Zwrócono szczególną uwagę na możliwości generowania tzw. miękkich sygnałów wyjściowych demodulatora, które są niezbędne dla dekoderów zastosowanego kodu kanałowego. Uzasadniono decyzję wyboru zastosowanej w systemie modulacji OQPSK.**

Abstract: **In this paper, selection of digital modulation applied in a nanosatellite system currently constructed in the Institute of Radiocommunications at Poznan University of Technology has been analyzed. Basic transmission requirements for the designed system has been presented and pros and cons of the digital modulations recommended by the CCSDS – Consultative Committee for Space Data Systems have been analyzed. Particular attention has been paid to possibilities of soft output signal generation on the output of the demodulator. These soft decisions are needed by the soft input channel code decoder. The final selection of the digiral modulation has been justified,**

Słowa kluczowe: **komunikacja nanosatelitarna. modulacje cyfrowe, GMSK. OQPSK, dekodowanie miękkodecyzyjne.**

Keywords: **nanosatellite communications, digital modulations, GMSK, OQPSK, soft decision decoding.**

1. WSTĘP

W końcu lat dziewięćdziesiątych w wyniku współpracy ośrodków uniwersyteckich w USA, agencji DARPA i Aerospace Corporation powstał standard nanosatelitarny zwany CubeSat. Podstawowa jednostka, z których składają się nanosatelity ma wymiar sześcianu o boku 10 cm. Często nanosatelita składa się z kilku połączonych jednostek, np. trzech. Bardzo małe wymiary i waga tego rodzaju satelitów sprawiają, że misje satelitarne stają się dostępne dla krajów i instytucji, które dotychczas nie brały udziału w misjach satelitarnych z powodów finansowych i technologicznych.

Aktualnie w Instytucie Radiokomunikacji Politechniki Poznańskiej jest konstruowane łącze telekomunikacyjne zapewniające łączność pomiędzy nanosatelitą krążącym nad Ziemią na wysokości rzędu 600 km. Prace te są wykonywane w ramach współpracy z Politechniką Kijowską oraz w ramach projektu NCBiR realizowanego wraz z firmą SatRevolution Sp. z o.o.

Oczekuje się, że na pokładzie nanosatelity zostaną umieszczone urządzenia, np. kamery cyfrowe, które będą generować dane wielkości megabajtowej. Implikuje to konieczność zapewnienia łączności z szybkością megabitów na sekundę o odpowiedniej jakości mierzonej pakietową i binarną stopą błędów.

Niniejszy artykuł jest poświęcony jedynie wybranej części przeprowadzonych prac badawczo-konstrukcyjnych, a konkretnie doborowi modulacji cyfrowych i skojarzonych z nimi demodulatorów, które są w stanie wygenerować sygnały miękkodecyzyjne. Konieczność generacji takich sygnałów wynika z zastosowania w projektowanym systemie zaawansowanego kodowania kanałowego.

2. PODSTAWOWE WYMAGANIA NA SYSTEM ŁĄCZNOŚCI

Przyjęto, że łącze od nanosatelity do stacji naziemnej będzie działało w paśmie 5830 - 5850 MHz (paśmie C), które zostało przydzielone transmisji w tym kierunku dla systemów o charakterze amatorskim przez Międzynarodową Unię Radia Amatorskiego (*International Amateur Radio Union*) [1], zaś systemy sterowania nanosatelitą i telemetryczne (TT&C) będą działać w paśmie VHF/UHF. Skoncentrujemy naszą uwagę na łączu w dół w paśmie 5830-5850 MHz. Uwzględniając wysokość orbity nanosatelity, rodzaj i parametry zastosowanej anteny odbiorczej w stacji naziemnej umieszczonej w Politechnice Poznańskiej, w ramach projektu obliczono bilans mocy łącza, w którego wyniku otrzymano oczekiwane wartości stosunku mocy sygnału do mocy szumu (SNR) w funkcji kąta elewacji nanosatelity. Wartości SNR zależą w znacznym stopniu od szerokości pasma sygnału oraz kąta elewacji. Postanowiono więc, aby system miał możliwość wyboru nie tylko parametrów modulacji i kodowania, ale również dokonania selekcji szerokości pasma, aby zapewnić możliwość transmisji dla zróżnicowanych wartości SNR. I tak, dla pasma sygnału o szerokości 20 MHz SNR zmienia się od około −7 dB dla kąta elewacji 10° do około 4 dB dla kąta elewacji równego 90°. Z kolei dla sygnałów o szerokości pasma 1 MHz wartości te wynoszą odpowiednio 5 i 17 dB. Niskie wartości SNR stanowią więc poważne wyzwanie projektowe. Szerokości pasma sygnału nadawa-

nego mogą być dobierane w zależności od maksymalnego kąta elewacji, które będzie obserwowane w danym oknie widzialności nanosatelity.

Wymagania transmisyjne łącza streszczono w tab. 1.

Tab. 1. Podstawowe wymagania transmisyjne łącza od nanosatelity do stacji naziemnej

Parametr	*Wartość*
Minimalna szybkość transmisji netto	10 Mb/s
Pakietowa stopa błędów	$< 10^{-2}$
Maksymalna moc sygnału nadawanego	33 dBm (2 W)
Zakres używanych częstotliwości	5830 – 5850 MHz
Wybór szerokości pasma	1,25 do 20 MHz
Odporność na przesunięcie Dopplera	≥ ± 150 kHz

Transmisja obywa się w trybie, w którym bloki danych są przeplatane odpowiednio dobranymi blokami znanych w odbiorniku danych, które są wykorzystywane do synchronizacji ramkowej i symbolowej oraz do estymacji przesunięcia częstotliwości Dopplera. Struktura ramki została opisana w pracy [2].

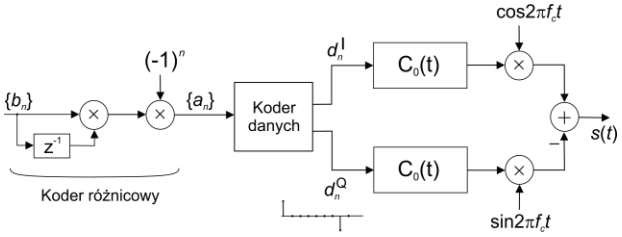
3. WYBÓR MODULACJI CYFROWEJ

W procesie wyboru modulacji zastosowanej w łączu w dół systemu nanosatelitarnego należy wziąć pod uwagę takie czynniki jak: efektywność energetyczną, własności widmowe zmodulowanego sygnału, stosunek mocy szczytowej do mocy średniej (PAPR – *Peak-to-Average Power Ratio*), wymagania obliczeniowe nadajnika i odbiornika i oferowane szybkości danych dla przyjętej szerokości pasma sygnału. W swoich dokumentach [3], [4] komitet CCSDS rekomenduje możliwość zastosowania dwóch typów modulacji: GMSK (*Gaussian Minimum Shift Keying*) oraz OQPSK (*Offset Quadrature Phase Shift Keying*).

W ramach projektu realizowanego w Instytucie Radiokomunikacji Politechniki Poznańskiej przeanalizowano możliwości zastosowania obu typów modulacji. Jednym z czynników, który miał istotny wpływ na ocenę potencjalnego zastosowania określonej modulacji była łatwość generowania „miękkich" sygnałów na wyjściu demodulatora w formie logarytmicznego stosunku wiarygodności (LLR – *Log-Likelihood Ratio*) lub jego przybliżenia dla każdego transmitowanego bitu danych. Nowoczesne dekodery kodów kanałowych korzystają bowiem z takich sygnałów w algorytmach dekodowania. Wykorzystanie jedynie wartości binarnych („twardych decyzji") przez dekoder skutkowałoby stratą E_b/N_0 nawet do dwóch decybeli.

3.1. Analiza modulacji GMSK

Jak wiadomo, modulacja GMSK została zastosowana w systemie GSM z powodu stałości swojej obwiedni i dzięki doskonałym własnościom widmowym uzyskanym przez kształtowanie impulsu częstotliwościowego modulacji FSK z indeksem modulacji h=1/2 za pomocą filtru o charakterystyce gaussowskiej. Stałość obwiedni implikuje wartość PAPR = 0 dB. W celu ułatwienia realizacji modulatora i demodulatora GMSK analizowano zlinearyzowaną aproksymację tej modulacji zaproponowaną przez Laurenta [5]. Tak więc schemat modulatora wygląda tak, jak przedstawiono go na rys. 1.



Rys. 1. Schemat zlinearyzowanego modulatora GMSK

Strumień danych bipolarnych $\{b_n\}$ podlega wstępnemu kodowaniu różnicowemu kompensując tym samym różnicowy charakter modulacji GMSK (przesunięcie fazy w danym odstępie modulacji odbywa się względem fazy uzyskanej w poprzednim odstępie modulacji [6]). W wyniku działania w nadajniku wstępnego kodera różnicowego, w odbiorniku próbki odpowiadające symbolom danych pojawiają się w postaci bezpośredniej naprzemiennie jako część rzeczywista albo urojona zespolonej próbki danych na wyjściu filtru odbiorczego. W wyniku działania wstępnego kodera różnicowego uzyskujemy bipolarny strumień danych $\{a_n\}$, który w koderze danych jest zamieniany na pary (d_n^I, d_n^Q), które stanowią impulsy wejściowe dla filtrów nadawczych o odpowiedzi impulsowej $C_0(t)$. Para impulsów danych może być reprezentowana w postaci zespolonej jako $d_n = d_n^I + jd_n^Q$. Odpowiedź impulsowa $C_0(t)$ jest obliczona numerycznie zgodnie z algorytmem Laurenta [5]. Filtry te, dla typowych wartości iloczynu 3-decybelowej szerokości pasma B i odstępu modulacji T, mają odpowiedź impulsową o efektywnej długości kilku odstępów modulacji. Wizualnie przypomina ona kształtem krzywą Gaussa. Filtry te wprowadzają więc interferencję międzysymbolową. Należy zatem zastosować odbiornik, który ją uwzględnia. Ze względu na łatwość implementacji rozważono odbiornik realizujący zasadę maksimum wiarygodności sekwencji danych zaproponowany przez Ungerboecka [7]. Odbiornik ten składa się z filtru dopasowanego do impulsów odbieranych oraz detektora sekwencyjnego, który znajduje sekwencję symboli danych stosując algorytm Viterbiego obliczający metryki przejścia λ_n^{jk} ze stanu j w chwili n–1 do stanu k w chwili n za pomocą wzoru:

$$\lambda_n^{jk} = \mathrm{Re}\left\{-\left(d_n^{jk}\right)^*\left(2y_n - d_n^{jk} g_0 - 2\sum_{i=1}^{L} g_i d_{n-i}\right)\right\} \qquad (1)$$

zaś d_n^{jk} są symbolami danych skojarzonymi z przejściem na wykresie kratowym ze stanu j-tego do k-tego, natomiast g_i (i=0,1,…,L) są próbkami odpowiedzi impulsowej widzianej na wyjściu filtru dopasowanego. W przypadku, gdy kanał może być rozważany jako kanał z addytywnym, białym szumem gaussowskim, współczynniki te mogą być obliczone na podstawie zależności:

$$g_i = \sum_{k=i}^{L} C_0(kT)C_0^*((k-i)T) \qquad (2)$$

W standardowym rozwiązaniu detektor wybiera taką sekwencję danych ($d_1,d_2,\ldots,d_n$), dla której metryka dojścia do stanu końcowego będąca sumą metryk przejść pomiędzy stanami jest maksymalna. Zatem detektor generuje „twarde decyzje". Jak wspomniano, dekoder kodu kanałowego powinien mieć do dyspozycji decyzje miękkie. Bardzo popularnym rozwiązaniem tego problemu jest zastosowanie metody zaproponowanej przez Hagenauera i Hoehera. Jej dokładny opis można znaleźć w [8]. W ogólności metoda ta, zwana SOVA (*Soft Output Viterbi Algorithm*), polega na przeskalowaniu twardej decyzji przez czynnik, którego wielkość zależy od różnicy metryk dojścia do danego stanu w momencie podejmowania decyzji dotyczącej wyboru, z którego stanu z chwili poprzedniej algorytm dochodzi do stanu aktualnego. Jeśli metryki dojścia po obu drogach są bardzo bliskie sobie, wtedy twarda decyzja ma niską wiarygodność. W trakcie eksperymentów symulacyjnych okazało się, że jakość algorytmu, choć jego stosowanie było bardzo często relacjonowane, jest niska. W konsekwencji rozważono zastosowanie innej metody wyznaczania miękkich decyzji. Została ona opisana w patencie USA [9]. Autorzy założyli, że aby ocenić miękkie wyjście dla symbolu danych d_k., symbole poprzedzające i następujące po nim są poprawne. Założenie można uznać za uzasadnione, gdy prawdopodobieństwo błędu symbolu jest na poziomie 10^{-2}, co umożliwia uzyskanie bardzo niskiej stopu błędów na wyjściu dekodera kodu kanałowego. Interesujący nas symbol d_k ma wkład do wartości sekwencji następujących próbek odebranych

$$\mathbf{y} = (y_{k-2L},\ldots,y_{k-1},y_k,y_{k+1},\ldots,y_{k+2L})$$

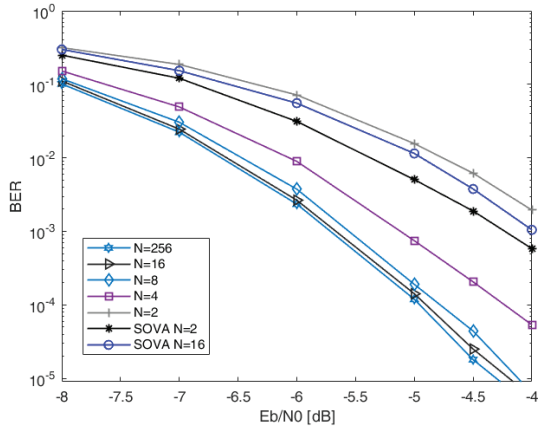
Z tego wynika, że zakładając biały gaussowski szum addytywny możemy napisać warunkowy rozkład prawdopodobieństwa w postaci:

$$p(y_{k+L},\ldots,y_{k+1},y_k,y_{k-1},\ldots,y_{k-L} \mid d_{k-2L},\ldots,d_{k-1},d_k,d_{k+1},\ldots,d_{k+2L})$$
$$= \left(\frac{1}{2\pi\sigma^2}\right)^L \prod_{i=-L}^{L} \exp\left\{-\frac{\left|y_{k+i} - \sum_{j=-L}^{L} g_j d_{k+i-j}\right|^2}{2\sigma^2}\right\} \qquad (3)$$

Zatem, po prostych obliczeniach otrzymujemy wzór na miękką decyzję (LLR) dla symbolu d_k:

$$\Lambda(d_k) = \ln\frac{p(\mathbf{y} \mid d_k = A)}{p(\mathbf{y} \mid d_k = -A)} = \qquad (4)$$
$$= \frac{1}{2\sigma^2}\sum_{i=-L}^{L}\left\{-\left|y_{k+i} - \sum_{j=-L,j\neq i}^{L} g_j d_{k+i-j} - g_i A\right|^2 + \left|y_{k+i} - \sum_{j=-L,j\neq i}^{L} g_j d_{k+i-j} + g_i A\right|^2\right\}$$

gdzie A=1 lub A=j w zależności od tego, czy indeks k ma wartość parzystą czy nieparzystą (co wynika z naprzemiennego korzystania ze składowych synfazowych i kwadraturowych z powodu wstępnego kodowania różnicowego w modulatorze GMSK). To samo dotyczy kolejnych symboli d_{k+i-j}. Na rys. 2 przedstawiono wyniki estymacji binarnej stopy błędów BER na wyjściu dekodera kodu kanałowego, który w eksperymentach doboru modulacji był popularnym kodem splotowym [133,171]. Zastosowano oba algorytmy generowania miękkich decyzji, tj. SOVA oraz algorytm (4) z N poziomami kwantyzacji na wejściu dekodera. Jak widać, zastosowanie algorytmu (4) skutkuje niższymi wartościami BER niż w przypadku algorytmu SOVA. Można również pokazać, że algorytm (4) jest znacznie prostszy implementacyjnie niż algorytm SOVA.



Rys. 2. Binarna stopa błędów (BER) na wyjściu dekodera kodu splotowego przy zastosowaniu algorytmu (4) i SOVA z różną liczbą poziomów kwantyzacji N na wejściu dekodera

Bardzo istotnym czynnikiem w przypadku doboru modulacji, który należy brać pod uwagę jest osiągalna szybkość binarna dla danej szerokości pasma. Estymowano gęstość widmową mocy sygnału GMSK i założono, że w pasmie kanału musi mieścić się znacząca część tego widma do wartości niższej o 30 dB w stosunku do jego wartości maksymalnej. W tab. 2 przedstawiono osiągalne szybkości binarne brutto dla różnych szerokości pasma W, którymi ma dysponować system.

Tabela 2. Szybkości binarne [Mb/s] sygnału GMSK dla różnych wartości iloczynu BT.

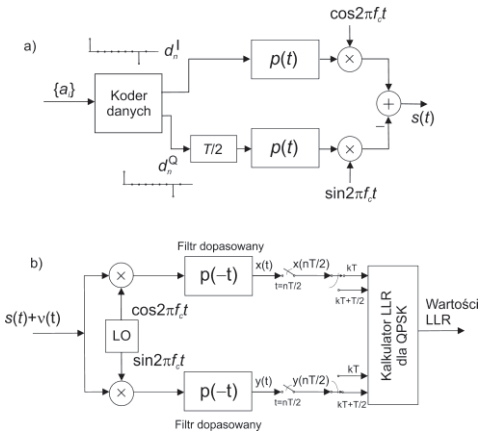
	BT = 0.2	BT = 0.25	BT = 0.3	BT = 0.35
F	0.56	0.62	**0.65**	0.67
W=1.25 MHz	1.12	1.01	**1.01**	0.93
W=5 MHz	4.46	4.03	**3.85**	3.73
W=10 MHz	8.93	8.06	**7.69**	7.46
W=20 MHz	17.86	16.13	**15.38**	14.93
PAPR [dB]	1.20	0.73	**0.55**	0.41

F jest częstotliwością znormalizowaną względem szybkości modulacji ($F = f/R$) zaś $R = W/2F$. Wybór typowej wartości BT=0.3 skutkuje uzyskaniem w paśmie W=20 MHz szybkości brutto równej 15,38 Mb/s, co nie wystarcza, aby spełnić wymagania stawiane systemowi.

3.2. Analiza modulacji OQPSK

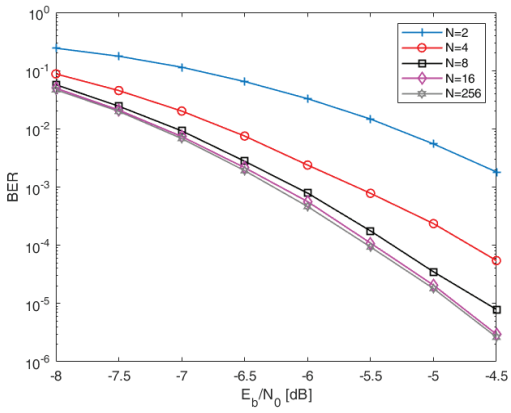
Modulacja OQPSK została wynaleziona z myślą o systemach satelitarnych i cechuje się niską, choć niezerową (w dB) wartością PAPR. Jej charakterystyczną cechą jest generowanie sygnału kwadraturowego z przesunięciem o połowę odstępu modulacji T w porównaniu z sygnałem synfazowym. Z kolei w demodulatorze wyjście filtru w jego gałęzi kwadraturowej jest przesunięte o T/2, dzięki czemu jest możliwa analiza konstelacji takiej,

jak w przypadku standardowej konstelacji QPSK. Modulator i demodulator OQPSK pokazano na rys. 3.



Rys. 3. Schemat modulatora (a) i demodulatora (b) sygnałów OQPSK.

Przyjęto, że filtr nadawczy $p(t)$ ma charakterystykę częstotliwościową o kształcie pierwiastka z podniesionego kosinusa o współczynniku opadania α. Każdy symbol danych przenosi dwa bity ($b_{1,n}, b_{2,n}$). Przy odpowiednim odwzorowaniu bloków binarnych punktom konstelacji, obliczenia przybliżonych wartości LLR dla każdego z dwóch bitów może być bardzo proste. W tab. 3 pokazano możliwe przepływności z zastosowaniem modulacji OQPSK dla różnych szerokości pasma i współczynnika opadania α. W projekcie wybrano transmisję ze współczynnikiem opadania filtru nadawczego α=0.35.



Rys. 4 Wykresy BER na wyjściu dekodera kodu kanałowego w funkcji E_b/N_0 dla różnej liczby poziomów kwantyzacji na wejściu dekodera

Tablica 3. Szybkości binarne [Mb/s] dla modulacji OQPSK z filtrem nadawczym o różnych współczynnikach opadania α.

	$\alpha = 0.2$	$\alpha = 0.25$	$\alpha = 0.3$	$\alpha = \mathbf{0.35}$
F	0.60	0.63	0.65	**0.67**
W=1.25 MHz	2,08	1.98	1.92	**1.87**
W=5 MHz	8.33	7.94	7.69	**7.46**
W=10 MHz	16.67	15.87	15.38	**14.93**
W= 20 MHz	33.33	31.74	30.77	**29.85**
PAPR [dB]	4.0	3.76	3.52	**3.28**

Jak można wywnioskować z rys. 4, zastosowanie już N=8 poziomów kwantyzacji skutkuje znaczącą poprawą jakości dekodowania w porównaniu z dekodowaniem twardodecyzyjnym (N=2). Z kolei z tab. 3 wnioskujemy, że przy założeniu sprawności kodowania R=1/2 i umiarkowanym nadmiarze z powodu zastosowania struktury ramkowej i sekwencji synchronizacyjnych jest możliwe w sprzyjającej konfiguracji osiągnięcie założonej szybkości transmisji rzędu 10 Mb/s. Ze względów konstrukcyjnych okazało się również, że gorsza wartość PAPR w porównaniu z modulacja GMSK nie ma istotnego znaczenia dla osiągniętych parametrów transmisji.

4. WNIOSKI

Przeprowadzone badania wskazują, że znacznie korzystniejsza pod względem parametrów i możliwości realizacyjnych w konstruowanym systemie nanosatelitarnym jest modulacja OQPSK w porównaniu z możliwą do zastosowania modulacją GMSK.

*Niniejsza praca była sfinansowana w ramach projektu NCBR Nr POIR.*01.01.01-00-1211/19-00.

LITERATURA

[1] Amateur and Amateur-satellite Service Spectrum, Administrative Council International Amateur Radio Union, Revised January 2020.

[2] Krenz Rafał, Sroka Paweł, Sybis Michał, Wesołowski Krzysztof. 2022. „A Low-Cost High-Throughput C-Band Communication System for CubeSats," artykuł wysłany na IEEE GLOBECOM 2022.

[3] The Consultative Committee for Space Data Systems. 2018. „Bandwidth-efficient modulations — summary of definition, implementation and performance." Informational Report CCSDS 413.0-G-3, Green Book, NASA, Washington, DC, USA, Feb. 2018

[4] The Consultative Committee for Space Data Systems. 2021. „Radio frequency and modulation systems — Part 1, Earth stations and spacecraft." Recommended Standard CCSDS 401.0-B-31, Blue Book, NASA, Washington, DC, USA, Feb. 2021

[5] Laurent, P. A. 1986. „Exact and approximate construction of digital phase modulations by superposition of amplitude modulated pulses," IEEE Trans. Commun. Vol. COM-34, Nr 2, Feb. 1986, str. 150-160

[6] Lui, G. L. 1998. „Threshold detection performance of GMSK signal with BT=0.5," Proc. of IEEE Military Communication Conference (MILCOM), Boston, 19-21 Oct. 1998, str. 515-519

[7] Ungerboeck, G. 1974. „Adaptive maximum-likelihood receiver for carrier-modulated data-transmission systems," IEEE Trans. Commun., vol. COM-22, No. 5, May 1974, str. 624-636

[8] Hagenauer, J., Hoeher, P. 1989. „A Viterbi algorithm with soft-decision outputs and its applications," Proc. of IEEE Global Communications Conference (GLOBECOM), Dallas, 1989, str. 1680-1686

[9] Yakhnich, E., Reshef, E. 2004. „Soft decision output generator," Patent No. US 6,731,700 B1, May 4, 2004

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

WYKORZYSTANIE PASMA C DO ŁĄCZNOŚCI Z NANO-SATELITAMI CUBESAT

APPLICATION OF C-BAND TO CUBESAT NANO-SATELITE COMMUNICATION

Rafał Krenz [1]; Michał Sybis [2]; Ilia Zainutdinov [3]

[1] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, rafal.krenz@put.poznan.pl
[2] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, michal.sybis@put.poznan.pl
[3] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, ilia.zainutdinov@doctorate.put.poznan.pl

DOI: 10.15199/59.2022.4.60

Streszczenie: **Rosnąca popularność satelitów standardu Cubesat oraz ich nowe zastosowania, wymagające przesyłania coraz większych ilości danych spowodowały konieczność wykorzystania do komunikacji nowych zakresów częstotliwości. W referacie przedstawiono system łączności pracujący w pasmie C, który umożliwia transmisję danych z szybkością wynoszącą nawet 10+ Mb/s, w zależności od wzajemnego położenia satelity i stacji naziemnej oraz innych zjawisk propagacyjnych.**

Abstract: **The growing popularity of Cubesat satellites, together with their new applications, requiring transmission of increasing data volumes may be supported by assigning new frequency bands for their communication. This paper presents a communication system operating in C-band, capable of transmitting 10+ Mbit/s which adapts to varying propagation conditions, depending on the relative position of the satellite and the ground station as well as other propagation phenomena.**

Słowa kluczowe: **Cubesat, łączność satelitarna, radio definiowane programowo**

Keywords: **Cubesat, satellite communication, software defined radio**

1. WSTĘP

Satelity standardu Cubesat już od ponad dekady zyskują coraz większą popularność, nie tylko w środowiskach akademickich, ale i w zastosowaniach profesjonalnych. Wynika to niewątpliwie z ich stosunkowo prostej budowy oraz niedużych kosztów umieszczenia na orbicie okołoziemskiej w porównaniu z satelitami „pełnowymiarowymi".

W najprostszych konstrukcjach do łączności z satelitami Cubesat wykorzystywane są pasma amatorskie. Najczęściej jest to zakres VHF 144 MHz (2 m) do przesyłania sygnałów sterujących ze stacji naziemnej oraz zakres UHF 435 MHz (70 cm) do przesyłania sygnałów telemetrycznych z pokładu satelity. Ze względu na ograniczoną szerokość kanału radiowego oraz zastosowanie prostych modulacji cyfrowych w pasmie akustycznym (np. AFSK) szybkości transmisji ograniczone są do kilku-kilkunastu kbit/s. Bardziej zaawansowane konstrukcje wykorzystują zazwyczaj pasma S oraz X, co wiąże się ze znacznie większymi kosztami urządzeń nadawczo-odbiorczych oraz niejednokrotnie z koniecznością uzyskiwania zezwoleń na wykorzystanie pasm licencjonowanych.

W niniejszej pracy opisano system szybkiej transmisji danych pracujący w pasmie C (5830-5850 MHz), przeznaczonym do satelitarnej łączności amatorskiej na łączu satelita -> stacja naziemna. Zakres ten jest obecnie wykorzystywany przez nielicznych operatorów satelitów Cubesat, którzy nie publikują szczegółowych informacji dotyczących konstrukcji systemu łączności ani jego parametrów [1]. Brakuje również na rynku gotowych rozwiązań modułów satelitarnych i urządzeń dla stacji naziemnych, w przeciwieństwie do wcześniej wspomnianych pasm VHF/UHF/S/X.

Zakres pracy jest następujący. W Sekcji 2 przedstawiono podstawowe parametry zaprojektowanego systemu oraz opisano zastosowane techniki transmisyjne. Sekcja 3 prezentuje implementację systemu, ze szczególnym uwzględnieniem modułu satelitarnego. W sekcji 4 zaprezentowano budowę segmentu naziemnego systemu łączności satelitarnej, natomiast sekcja 5 stanowi podsumowanie i wskazuje dalszy rozwój systemu.

2. BUDOWA SYSTEMU

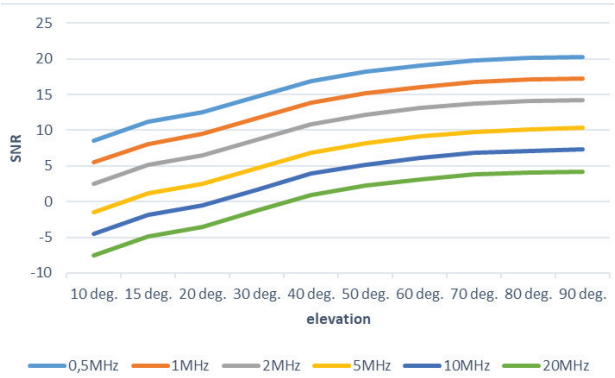
2.1. Podstawowe wymagania i parametry

Celem projektu, realizowanego wspólnie z Politechniką Kijowską oraz firmą SatRevolution S.A., było zaprojektowanie i budowa systemu transmisji danych o dużej przepływności przeznaczonego dla satelitów Cubesat. Cechują go następujące parametry:

- częstotliwość nośna – 5830 - 5850 MHz
- szerokość pasma kanału radiowego – 1 MHz, 1,25 MHz, 5 MHz, 10 MHz, 20 MHz
- moc sygnału transmitowanego – 2W (+33dBm)
- modulacja – OQPSK z kształtowaniem widma filtrem RRC, $\alpha = 0{,}35$
- kodowanie kanałowe – turbokod o sprawności kodowania z przedziału 0.19 - 0.91
- przepływność – 200 kbit/s - 13,5 Mbit/s

Ponieważ satelity Cubesat krążą na niskich orbitach (LEO), zazwyczaj na wysokości 400-600 km, są one widoczne w konkretnej lokalizacji przez okres 5-15 minut, kilkukrotnie w ciągu doby. Wraz ze zmianą kąta elewacji zmienia się również odległość pomiędzy satelitą a stacją naziemną, a co za tym idzie, również moc sygnału odbieranego. Przyjmując określone parametry odbiornika wyzna-

czono spodziewane wartości stosunku sygnału do szumu (SNR) w projektowanym systemie, patrz Rys. 1.



Rys. 1. Spodziewane wartości SNR w funkcji kąta elewacji dla wybranych szerokości pasma kanału, współczynnik szumów F = 3 dB

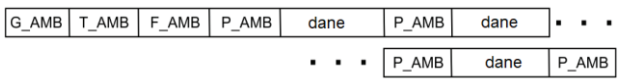
Ze względu na dużą zmienność wartości stosunku sygnału do szumu, parametry systemu (w tym przypadku szerokość pasma kanału oraz sprawność kodowania) powinny zmieniać się wraz z położeniem satelity w stosunku do stacji naziemnej, co umożliwi maksymalizację wolumenu przesyłanych danych.

2.2. Struktura ramki fizycznej

System komunikacyjny opracowywany z wykorzystaniem satelitów CubeSat musi działać przy bardzo niskich wartościach SNR. Jest to związane z ograniczoną moc nadawania oraz ograniczonym rozmiarem anteny umieszonej na satelicie. Dodatkowym czynnikiem, który utrudnia komunikację jest również, charakterystyczne dla komunikacji z satelitami na orbicie LEO, znaczne przesunięcie Dopplera wynikające ze względnego ruchu między satelitą (nadajnikiem) a stacją naziemną (odbiornikiem). Dla projektowanego systemu może ono osiągnąć wartość ±150 kHz (przy maksymalnej szybkości zmiany wynoszącej blisko 1 kHz/s). Dlatego niezbędne jest zastosowanie efektywnych procedur synchronizacji i kodowania.

Na Rys. 2 przedstawiono proponowaną strukturę ramki fizycznej. W rozważanym systemie zdefiniowano cztery różne typy sekwencji treningowych, aby spełnić wszystkie wymagania jakościowe projektowanego systemu:

- Preambuła G_AMB – służy do wytrenowania układu automatycznej kontroli wzmocnienia (ang. *Automatic Gain Control, AGC*). Preambuła ta jest opracowana w oparciu o właściwości zintegrowanego transceivera AD9364.
- Preambuła T_AMB – używana do synchronizacji czasowej. Służy ona do wykrycia początku ramki radiowej (fizycznej) z dokładnością do pojedynczej próbki. Ponieważ transmisja rozpoczyna się na początku okna widzialności satelity (tj. przy bardzo małym kącie elewacji) wiąże się to z bardzo niskimi wartościami SNR, co stawia wysokie wymagania algorytmom synchronizacji czasu.
- Preambuła F_AMB – niezbędna do oszacowania dokładnego przesunięcia częstotliwości nośnej (ang. Carrier Frequency Offset, CFO). CFO jest wstępnie kompensowane z wykorzystaniem wartości obliczonych na podstawie parametrów orbitalnych i aktualnej chwili czasowej, co jest typowym rozwiązaniem używanym w komunikacji w pasmie VHF/UHF. Parametry orbitalne można uzyskać dzięki systemowi TLE (ang. Two-Line Elements, TLE), którego komunikaty są publikowane przez North American Aerospace Defence Command (NORAD).
- P_AMB midamble – przesunięcie Dopplera zmienia się w oknie widoczności satelity. Dlatego po wstępnym oszacowaniu CFO, zmiana fazy musi być śledzona i kompensowana. Midambuły P_AMB są przeplatane danymi użytkownika (tj. fragmentami słów kodowych). Dodatkowo, różne sekwencje midambul są wykorzystywane do wskazania aktualnego współczynnika kodowania (wykorzystywan w kodowaniu adaptacyjnym).



Rys. 2. Schemat ramki fizycznej

Czas trwania ramki radiowej wynosi 1 s, co stanowi kompromis pomiędzy wydajnością warstwy fizycznej a typowym czasem trwania okna widoczności satelity. Rodzaj i długość sekwencji wykorzystywanych jako preambuły i środkowe są określone w rozdziale 2.3. Liczba słów kodowych w ramce zmienia się wraz z szerokością pasma sygnału i szybkością kodowania, np. dla pasma sygnału 5 MHz i sprawności kodowania równej R = 0,57 pojedyncza ramka radiowa zawiera 140 słów kodowych o długości 5280 symboli OQPSK (przyjęto długość bloku transportowego równą 6016 bitów), każde podzielone na 16 fragmentów o długości 330 symboli.

2.3. Synchronizacja

Synchronizacja jest kluczowym elementem łańcucha odbiorczego w systemie komunikacyjnym. Znajomość synchronizacji symboli, przesunięcia fazy nośnej (ang. *Carrier Phase Offset, CPO*) i przesunięcia częstotliwości nośnej są niezbędne do przeprowadzenia poprawnej demodulacji. Efektywność synchronizacji ma kluczowe znaczenie zwłaszcza w transmisji w trybie „burst", w której czas akwizycji musi być możliwie jak najkrótszy, pomimo pracy w warunkach niskiego SNR i przy dużym przesunięciu częstotliwości ze względu na efekt Dopplera. W związku z krótkim czasem akwizycji dla mechanizmu synchronizacji zaproponowano strukturę sprzężenia zwrotnego wspomaganego danymi (*ang. Data Aided, DA*), w której procedura synchronizacji podzielona jest na następujące etapy:

- Zgrubna estymacja i korekcja CFO na podstawie estymacji parametrów orbitalnych satelity [2]. Przesunięcie Dopplera spowodowane ruchem satelity można oszacować za pomocą procedury

przedstawionej w [3]. Celem tego etapu jest znaczne zmniejszenie CFO, co pozwala na poprawę dokładności synchronizacji ramkowej.

- Synchronizacja ramkowa (synchronizacja czasu) w celu oszacowania początku paczki danych.
- Dokładna synchronizacja częstotliwości mająca na celu dokładne oszacowanie i korekcję pozostałego CFO.
- Synchronizacja fazy wraz ze śledzeniem częstotliwości, monitorowanie CPO spowodowanego przez nieusunięte przesunięcie częstotliwości ze względu na zmiany w efekcie Dopplera.

Z wyjątkiem początkowej zgrubnej korekcji CFO, każdy etap synchronizacji wymaga użycia dedykowanej sekwencji do wykonania estymacji.

W przypadku synchronizacji ramki (synchronizacji czasu) głównym celem jest wykrywanie początku ramki radiowej z dokładnością na poziomie pojedynczej próbki. Ze względu na oczekiwaną niską wartość SNR, wymagana jest wysoka odporność preambuły na zniekształcenia wynikające z kanału radiowego. W związku z tym, zaproponowano użycie sekwencji x_k, która jest rozszerzona sekwencja Zadoff-Chu (ZC) (k oznacza indeks sekwencji), posiadającą bardzo dobre własności korelacyjne nawet w bardzo zaszumionym środowisku [4]. Działanie bloku synchronizacji czasu (wykrywania ramek) będzie polegało na znalezieniu indeksu próbki odpowiadającego maksimum korelacji krzyżowej odebranego sygnału $r(n)$ z zapisaną sekwencją ZC.

Na potrzeby synchronizacji przyjęto sekwencję ZC o $k = 1$ i długości $N_{ZC} = 107$. Sekwencja ta została rozszerzona do $N_{PT} = 128$ symboli, aby uwzględnić wpływ filtrowania w części nadawczej i odbiorczej. Ostateczna postać wykorzystanej sekwencji jest, zatem następująca:

$$\boldsymbol{p_T(n)} = \begin{cases} \boldsymbol{x(N_{ZC} - 12 + n)} & \textbf{dla } \boldsymbol{n \le 10} \\ \boldsymbol{x(n - 11)} & \textbf{dla } \boldsymbol{10 < n \le N_{ZC} + 10} \\ \boldsymbol{x(n - N_{ZC} - 10)} & \boldsymbol{n > N_{ZC} + 10} \end{cases} \tag{1}$$

W celu poprawy jakości synchronizacji ramkowej (preambuła T_AMB) zastosowano dwukrotne powtórzenie przyjętej sekwencji. Zatem, metryka korelacji $m(n)$ jest otrzymywana jako średnia wyników korelacji krzyżowej dwóch sekwencji z referencyjną sekwencją ZC. Następnie, moment zakończenia preambuły jest znajdowany jako maksimum wartości bezwzględnej metryki $m(n)$.

Poza synchronizacją ramki, omawiana sekwencja może być również wykorzystana do wstępnego oszacowania przesunięcie fazy początkowej (estymowana jako kąt wyniku korelacji krzyżowej, dla maksimum wykrytej metryki $m(n)$). Pozwala to na wstępną kompensację przesunięcia fazy przed właściwą synchronizacją częstotliwości.

Ponieważ znaczna część CFO spowodowana przesunięciem Dopplera może zostać usunięta za pomocą bloku wstępnej korekcji CFO, celem algorytmu dokładnej estymacji CFO jest wykrycie niedopasowania spowodowanego różnicami oscylatorów i pozostałego CFO wynikającego z przesunięcia Dopplera.

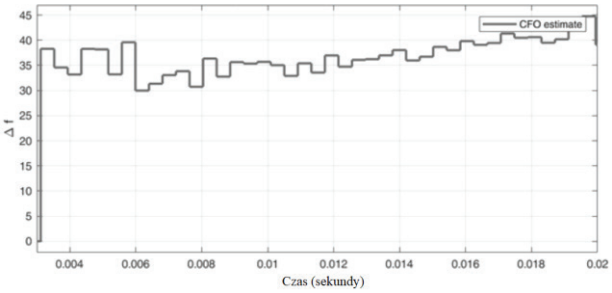
W rozważanej metodzie synchronizacji wymagana jest odpowiednio zaprojektowana i wystarczająco długa (oznaczona jako L_0) preambuła synchronizacji częstotliwości, pozwalająca na estymację CFO z dokładnością bliską teoretycznym limitom (dolna granica Cramera-Rao). W celu dokładnego oszacowania rozważana jest preambuła w postaci złożonego sygnału sinusoidalnego, który można wyrazić jako:

$$\boldsymbol{p_F(n)} = \boldsymbol{exp}\left(-\boldsymbol{j}\frac{\boldsymbol{\pi n}}{\boldsymbol{4}}\right), \text{ n = 0, 1, …, } L_0\text{-1} \tag{2}$$

Zaletą tak przygotowanej preambuły jest jej odporność na wszelkie zmiany spowodowane np. filtrowaniem sygnału nadawanego lub odbieranego. Odebrany sygnał odpowiadający preambule synchronizacji częstotliwości $r_F(n)$ jest demodulowany przez pomnożenie ze sprzężoną sinusoidą o tych samych parametrach. Stąd, wynikowy sygnał $s(n)$ jest sinusoidą o częstotliwości odpowiadającej CFO odbieranego sygnału. Zatem estymacja CFO w tym przypadku polega na znalezieniu częstości tej sinusoidy, co można zrealizować stosując podejście maksimum wiarygodności w postaci maksymalizacji periodogramu [5].

Wadą metod opartych na DFT jest to, że są one dokładne z rozdzielczością zależną od rozmiaru DFT [4]. Tak więc, przy wysokich wartościach częstotliwości próbkowania Fs, wymagane byłoby użycie bardzo dużego rozmiar N_{DFT} aby uzyskać dokładną estymatę CFO. Jako alternatywę można rozważyć metody interpolacji, takie jak metoda oparta na iteracyjnym przetwarzaniu tzw. ułamkowych współczynników Fouriera (ang. *Fractional Fourier Coeffcients, FFC*), zaproponowaną w [6].

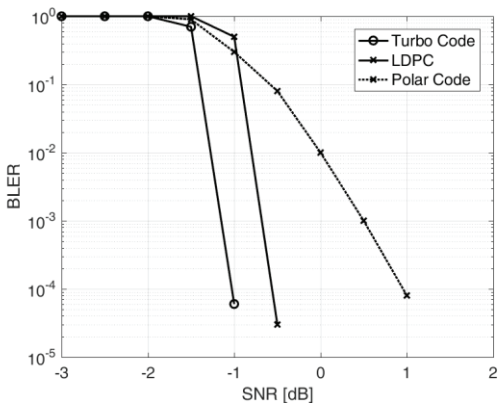
Ostatnim elementem systemu synchronizacji jest blok synchronizacji fazy i śledzenia częstotliwości, który jest stosowany po synchronizacji ramki i kompensacji CFO. Jego celem jest skompensowanie pozostałego małego błędu CFO spowodowanego zmiennością przesunięcia Dopplera, gdy satelita zmienia swoją pozycję. Przesunięcie fazowe, wstępnie skompensowane przy użyciu wcześniej opisanego podejścia, może zmieniać się w obrębie długiej ramki. W związku z tym, w ramce radiowej zawarte są krótkie, dedykowane midambuły, przeplatane blokami danych, które służą do ciągłego monitorowania zmian przesunięcia fazowego. Ponieważ wykorzystywane midambuły powinny być relatywnie krótkie, w porównaniu do długości bloku danych (midambule o długości 64 symboli towarzyszy blok danych o długości 330 symboli), zastosowano rozszerzoną sekwencję ZC, z $k = 1$ i długości $N_{ZC} = 47$, z dołączonymi 8 dodatkowymi próbkami z przodu oraz 9 próbkami ogona, podobnie jak w równaniu (1). Po stronie odbiornika, śledzenie przesunięcia fazy jest wykonywane przy użyciu filtru dopasowanego do środkowej sekwencji ZC. Następnie obliczana jest różnica w CPO dla kolejnych midambuł, tak, aby znaleźć estymatę pozostałego CFO. W celu zmniejszyć wpływu szumu, kolejne oszacowania są uśredniane przy użyciu prostego, wykładniczego filtru wygładzającego. Przykładowy wynik śledzenia CFO za pomocą proponowanej metody pokazano na Rys. 3.



Rys. 3. Przykład śledzenia CFO dla systemu o szerokości pasma 1,25 MHz, SNR = 5 dB i współczynnik filtru wygładzającego α = 0,1

2.4. Kodowanie kanałowe

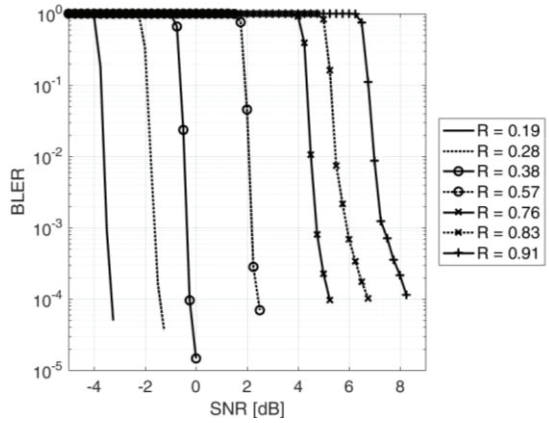
Innym ważnym aspektem projektowania systemu jest wybór schematu kodowania kanałowego. Wybrane podejście powinno z jednej strony pozwolić na uzyskanie wysokiej wydajności. Z drugiej strony należy brać pod uwagę tylko rozwiązania, które są możliwe do wdrożenia. Dlatego w tej sekcji pokrótce przedstawiono najbardziej obiecujące schematy kodowania kanałowego oraz oceniono ich przydatność dla łącza satelitarnego. W związku z tym dyskusja ogranicza się do najbardziej obiecujących kodów, tj. Turbokodów, kodów LDPC (ang. *Low-Density ParityCheck, LDPC*) i kodow polaryzacyjnych (ang. *Polar Codes, PC*) [7].



Rys. 4. Porównanie blokowej stopy błędów BLER dla wiadomości o długości K= 4096 i sprawności kodowania R= 1/3, modulacja QPSK

Porównanie jakości dekodowania badanych sposobów kodowania kanałowgo przedstawiono na Rys. 4 (wszystkie badane schematy mają jednakową złożoność dekodowania [7]). Jak widać na powyższym rysunku, jakość dekodowania turbokodów z algorytmem dekodowania max-log-MAP i kodów LDPC z algorytmem dekodowania OMSA (*ang. Offfset Min-Sum Algorithm*) jest zbliżona. Różnica między tymi dwoma podejściami wynosi około 0,5 dB przy BLER = 10^{-4}. Najsłabszą jakość dekodowania obserwuje się dla kodów polaryzacyjnych wykorzystujących algorytm SCL (*ang. Successive Cancellation List*), $L = 50$.

Ponieważ różnica wydajności między dwoma najlepszymi schematami pozostaje w granicach 0,5 dB, jakość dekodowania nie może być jedynym kryterium stosowanym do wyboru schematu kodowania.



Rys. 5. Porównanie jakości dekodowania BLER przyjętych BCS

Dlatego zdecydowano się na porównanie złożoności procesu kodowania obu rozwiązań (zużycie energii przez satelitę powinno być minimalizowane). Na tej podstawie wybrano turbokody jako schemat kodowania dla opracowywanego systemu.

Ze względu na charakter łącza satelitarnego, osiągana wartość SNR jest silnie powiązana z kątem elewacji. W związku z tym wydajność systemu można zoptymalizować za pomocą adaptacji łącza. Aby pokryć zakres operacyjny osiągalnych wartości SNR, zastosowano zestaw siedmiu współczynników kodowania połączonych z czterema przepustowościami kanałów. Procedura jest określana skrótem BCS (*ang. Bandwidth and Coding Scheme*). Jakość dekodowania BLER dla przyjętego zbioru BCS jest przedstawiona na Rys. 5.

Aby zapewnić jak największą przepustowość systemu, parametry transmisji takie jak szerokość pasma sygnału oraz sprawność kodowania będą zmieniały się podczas sesji komunikacyjnych, planowanych w kolejnych „oknach" widoczności. Ze względu na ograniczenia kanału zwrotnego (znaczne opóźnienia lub niedostępność) nie ma możliwości zastosowania tradycyjnych metod adaptacji łącza, znanych np. systemów komórkowych. Dlatego parametry transmisji będą planowane ze znacznym wyprzedzeniem i przesyłane do satelity w postaci tzw. planów sesji. Do ich przygotowania zastosowane zostaną algorytmy bazujące na technikach uczenia maszynowego, które są aktualnie opracowywane w ramach prowadzonych badań.

3. IMPLEMENTACJA SYSTEMU

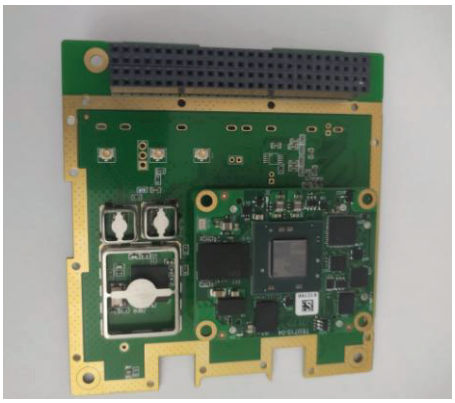
Zarówno moduł nadawczy umieszczony na satelicie, jak i odbiornik w stacji naziemnej, zostały zbudowane z wykorzystaniem techniki radia programowalnego (SDR). Sercem modułu nadawczego jest scalony transceiver AD9364 produkowany przez firmę Analog Devices [8], który współpracuje z układem programowalnym SoC typu Zynq7030 firmy Xilinx [9]. Ten ostatni zawiera w swojej strukturze m.in. macierz programowalną FPGA oraz dwu rdzeniowy procesor ARM Cortex A9.

Wszystkie procedury pasma podstawowego, takie jak kodowanie kanałowe, modulacja, formowanie ramki fizycznej (dołączanie preambuł i midambuł) oraz filtracja sygnału zastały zaimplementowane w macierzy FPGA.

Procesor jest z kolei odpowiedzialny za realizację protokołów wyższych warstw (np. podział przesyłanych plików na fragmenty i ich numerowanie) oraz zarządzanie modułem i komunikację z komputerem pokładowym satelity.

Do implementacji systemu wykorzystano model opracowany w środowisku Matlab/Simulink. Umożliwił on symulację zaprojektowanego systemu w wersji zmiennoprzecinkowej, dobór reprezentacji liczb dla wersji stałoprzecinkowej oraz generowanie kodu VHDL dla własnych modułów IP-core. Ostatecznej integracji systemu dokonano w środowisku Xilinx Vivado Design Suite.

Moduł nadawczy został zbudowany zgodnie z wymaganiami standardu Cubesat i bazuje na standardzie PC/104 [10]. Główną część modułu stanowi 8-warstwowa płyta PCB (Rys. 6), przykryta radiatorem, w którym umieszczono wzmacniacz nadawczy.



Rys. 6. Płyta PCB modułu nadawczego

4. SEGMENT NAZIEMNY

Odbiór sygnałów realizowany jest przez segment naziemny, w którego skład wchodzą system antenowy oraz odbiornik wraz z układem sterowania anteną.

Antena paraboliczna o średnicy 3 m, wyposażona w promiennik typu „septum feed" dla pasma 5,8 GHz, została umieszczona na rotorze BIG-RAS HR [11], który umożliwia śledzenie satelitów LEO poruszających się ze znaczną szybkością względem Ziemi. Dokładność śledzenia wynosi ok. 0,1° a dzięki umieszczeniu anteny na dachu budynku o wysokości ok. 35 m zapewniona jest niczym niezakłócona widoczność całego widnokręgu. Bezpośrednio przy antenie umieszczono wzmacniacz niskoszumny, z którego sygnał przesyłany jest nisko stratnym kablem koncentrycznym do odbiornika wewnątrz budynku. Jak wspomniano już wcześniej, odbiornik został zbudowany również z wykorzystaniem techniki radia programowalnego, jednak ze względu na znacznie większą złożoność algorytmów po stronie odbiorczej wykorzystano moduł SDR wyposażony w układ Zynq7100, dysponujący kilkukrotnie większymi zasobami w macierzy FPGA. Przesyła on zdekodowane pakiety danych do komputera PC, odpowiedzialnego za realizację wyższych warstw protokołu komunikacyjnego.

Planowana jest również budowa uproszczonej wersji odbiornika, bazująca na module SDR typu USRP B210 [12], współpracującym z odbiornikiem zrealizowanym programowo na platformie GnuRadio [13] i przetwarzającym odbierane sygnały w trybie „off-line".

Istotnym elementem segmentu naziemnego jest sterownik rotora antenowego, którego zadaniem jest śledzenie satelity podczas sesji komunikacyjnej, współpracujący z oprogramowaniem *HamRadio Deluxe*. Korzysta on z plików TLE umożliwiających wyznaczanie pozycji orbitalnej dowolnego satelity.

5. PODSUMOWANIE

Zaprojektowany system szybkiej transmisji danych pracujący w pasmie C będzie wykorzystany w misjach satelitów Cubesat planowanych przez firmę SatRevolution oraz Politechnikę Kijowską. Opracowany system obejmuje projekt zarówno części nadawczej (umieszczonej na orbicie) jak i odbiorczej. Pierwsze praktyczne próby powinny rozpocząć się w 2023 r. podczas misji satelity PW6U.

Warto dodać, że w Politechnice Poznańskiej realizowany jest obecnie projekt budowy profesjonalnej naziemnej stacji łączności satelitarnej, zlokalizowanej w Kąkolewie k. Grodziska Wlkp. Zostanie ona wyposażona w dwie anteny paraboliczne o średnicy 4,5 m, jedna umieszczona pod kopułą zapewniającą ochronę przed niekorzystnymi warunkami atmosferycznymi i pracująca w pasmach S/X oraz druga, przeznaczona do pracy w pasmie C i współpracująca z antenami Yagi dla pasm VHF/UHF. W ich pobliżu znajdą się kontenery z aparaturą nadawczo-odbiorczą oraz centrum przetwarzania danych satelitarnych z serwerami GPU (łącznie 80000 rdzeni CUDA).

Praca została sfinansowana przez Narodowe Centrum Badań i Rozwoju w ramach grantu nr POIR.01.01.01-00-1211/19-00 a udział w konferencji z projektu 0312/SBAD/8162.

LITERATURA

[1] Amateur and Amateur-satellite Service Spectrum, Administrative Council International Amateur Radio Union, Revised Jan 2020
[2] H. Rouzegar, M. Ghanbarisabagh, "Estimation of Doppler curve for LEO satellites", Wireless Personal Communications, No. 108, 2019, pp. 2195–2212
[3] S. Amiri and M. Mehdipour, "Accurate Doppler frequency shift estimation for any satellite orbit," 2007 3rd International Conference on Recent Advances in Space Technologies, 14-16 June 2007, Istanbul, Turkey. pp. 602-607
[4] M. M. U. Gul, X. Ma and S. Lee, "Timing and frequency synchronization for OFDM downlink transmissions using Zadoff-Chu sequences," IEEE Trans. Wireless Commun, vol.14, No. 3, March 2015, pp. 1716-1729
[5] E. Hosseini and E. Perrins, "Timing, carrier, and frame synchronization of burst-mode CPM," IEEE Trans. Commun., vol. 61, No. 12, Dec. 2013, pp. 5125-5138,
[6] E. Aboutanios, "Frequency estimation for low Earth orbit satellites", PhD thesis, 2002 [online, available at: https://opus.lib.uts.edu.au/bitstream/10453/43211/2/02whole.pdf, [dostęp 12.04.2021]
[7] M. Sybis, i in. "Channel coding for ultra-reliable low-latency communication in 5G systems," 2016 IEEE 84th Vehicular Technology Conference (VTC), 2016, pp. 1-5
[8] AD9364 Data Sheet, Analog Devices Inc., 2013
[9] ds190 Zynq7000 Overview, Xilinx Inc., 2014
[10] Pumpkin Cubesat Kit User Manual, Pumpkin Inc., 2003
[11] http://spid.net.pl/pl/big-ras-hr/ [dostęp, 05.05.2022]
[12] https://www.ettus.com/all-products/ub210-kit/ [dostęp 05.05.2022]
[13] https://www.gnuradio.org/ [dostęp 05.05.2022]

MODELE ENERGETYCZNE DLA SIECI BEZPRZEWODOWYCH Z MOBILNYMI WĘZŁAMI DOSTĘPOWYMI*

ENERGY MODELS FOR WIRELESS NETWORKS WITH MOBILE ACCESS NODES

Adam Samorzewski[1]; Adrian Kliks[2]

[1] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, adam.samorzewski@doctorate.put.poznan.pl
[2] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2022.4.61

Streszczenie: **W artykule przedstawiono modelowanie cyklu energetycznego dla mobilnych węzłów dostępowych w heterogenicznych systemach bezprzewodowych. Za punkty dostępowe przyjmuje się urządzenia typu dron, które cechuje zdolność do pozyskiwania energii z Odnawialnych Źródeł Energii (OZE), a konkretniej z promieniowania słonecznego. Badania symulacyjne przeprowadzono w zakresie sprawdzenia ilości wygenerowanej energii przez urządzenie typu dron w stosunku do jego zapotrzebowania energetycznego o danej porze dnia i roku. W celu zdiagnozowania wpływu warunków pogodowych na ilość pozyskanej energii uwzględniono również absorpcję promieniowania słonecznego dla różnych typów zachmurzenia nieba.**

Abstract: **The paper considers modeling the energy cycle for mobile access nodes in heterogeneous wireless systems. Access points are drone-type devices that are characterized by the ability to obtain energy from Renewable Energy Sources (RES), and more specifically from solar radiation. The simulation tests were conducted in order to check the amount of energy generated by a drone-type device in relation to its energy demand at a given time of the day and year. In order to diagnose the impact of weather conditions on the amount of obtained energy, the absorption of solar radiation for various types of sky cloudiness was also taken into account.**

Słowa kluczowe: **heterogeniczne sieci komórkowe, modele energetyczne, Odnawialne Źródła Energii, UAV.**

Keywords: **energy models, heterogeneous cellular networks, Renewable Energy Sources, UAV.**

1. WSTĘP

W ciągu ostatnich lat pozyskiwanie energii z promieniowania słonecznego stało się bardzo powszechnym rozwiązaniem odnajdującym zastosowanie nie tylko w przemyśle, ale i w wielu różnych obszarach życia codziennego. Proces pozyskiwania energii słonecznej przeprowadzany jest za pomocą generatorów w postaci tzw. paneli fotowoltaicznych (PV – ang. Photovoltaics). Teraźniejsza atrakcyjność rynkowa fotowoltaiki wynika nie tylko z ekologicznego podejścia do pozyskiwania zasobów energetycznych, lecz także z cen modułów solarnych, które od ostatnich kilku lat wykazują tendencję spadkową [1]. Zatem nic dziwnego, że panele fotowoltaiczne obecnie odnajdują swoje zastosowanie również w systemach telekomunikacyjnych. Szczególne korzyści mogą być dostrzeżone z perspektywy heterogenicznych systemów bezprzewodowych jak np. sieci komórkowe. Z uwagi na wzrastające wymogi transmisyjne (usługowe) sieci ICT (ang. Information and Communication Technology), z każdym kolejnym rokiem wykazują one coraz większe zapotrzebowanie energetyczne względem lat poprzednich. Wyposażenie punktów dostępowych sieci bezprzewodowych w panele fotowoltaiczne mogłoby zredukować lub całkowicie wyeliminować zapotrzebowanie na energię elektryczną ze źródeł konwencjonalnych. To z kolei pozwoliłoby nie tylko zmniejszyć emisję CO_2 do atmosfery, lecz spowolniłoby również zużycie surowców nieodnawialnych [2]. W dodatku jednym z głównych paradygmatów w teraźniejszych i przyszłych systemach telekomunikacyjnych jest dostarczanie usług na terenach trudno-dostępnych. Obszary takie cechują niekorzystne warunki dla rozbudowy infrastruktury telekomunikacyjnej, które wynikają chociażby z ukształtowania terenu, czy braku możliwości podłączenia się do istniejących sieci telekomunikacyjnych i energetycznych. W sytuacji, gdy ten pierwszy problem można by rozwiązać stosując mobilne węzły dostępowe (np. drony) i angażując łączność bezprzewodową dla komunikacji dosyłowej, tak drugiemu przypadkowi niezwykle ciężko jest sprostać [3]. Jednakże rozważając podejście pozasieciowe (ang. off-grid) w kontekście zasilania poszczególnych komponentów systemu, generatory OZE (np. panele fotowoltaiczne, turbiny wiatrowe) w połączeniu z autonomicznym układem bateryjnym mogłyby być w stanie zaspokoić potrzeby energetyczne sieci RAN (ang. Radio Access Network) [4]. Wymaga to jednak odpowiedniego zamodelowania cyklu energetycznego systemu, tj. przewidzenia rezultatów produkcji oraz wykorzystania zasobów energii w czasie w taki sposób, aby zapewnić ich adekwatną ilość dla każdego rodzaju obciążenia łączy sieciowych [5].

W niniejszym artykule zaproponowane zostały modele energetyczne dla bezprzewodowych sieci dostępowych, których węzły są urządzeniami mobilnymi typu dron (UAV – ang. Unmanned Arieal Vehicle). Komunikacja między węzłami na łączu dosyłowym jest z kolei realizowana zgodnie z technologią FSO (ang. Free Space Optics). Artykuł został podzielony następująco. W sekcji drugiej przybliżono zarówno rozpatrywany scenariusz symulacyjny systemu bezprzewodowego, jak i zaimplemen-

* Praca sfinansowana ze środków statutowych 0312/SBAD/8161

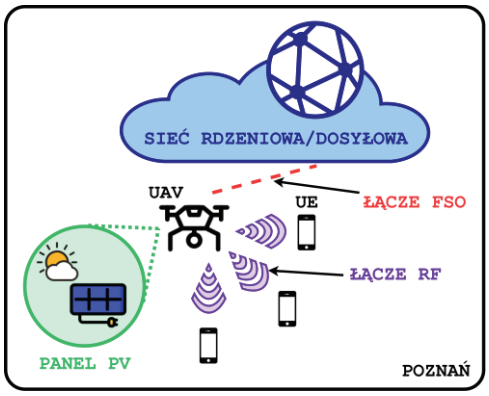
towane modele energetyczne dotyczące generowania i zużywania zasobów przez komponenty systemu. W sekcji trzeciej opisano konfigurację przebiegu symulacyjnego oraz wyniki otrzymane po jego zakończeniu. Na końcu, w sekcji czwartej, przedstawiono konkluzje wyciągnięte na podstawie otrzymanych rezultatów.

2. SYSTEM I MODELE ENERGETYCZNE

W celu oszacowania efektywności energetycznej rozważanego systemu bezprzewodowego niezwykle istotną krokiem jest dokładne skonfigurowanie scenariusza badanego systemu bezprzewodowego oraz zamodelowanie cyklu energetycznego dla jego mobilnych stacji bazowych. W związku z tym, poniżej opisano analizowane środowisko oraz formuły matematyczne użyte do symulacyjnej predykcji zarówno użycia, jak i produkcji zasobów energii w czasie.

2.1. Opis systemu

Na Rys. 1 zaprezentowany został schemat systemu opisanego w niniejszej publikacji. Rozważany w artykule scenariusz symulacyjny zakłada wykorzystanie urządzenia UAV jako stacji bazowej, które jest wyposażone w panele fotowoltaiczne do pozyskiwania energii ze Słońca. Węzeł UAV jest w stanie utrzymać połączenie ze stacją nadrzędną na łączu dosyłowym zrealizowanym zgodnie z technologią FSO oraz na łączu dostępowym w paśmie częstotliwości radiowych. Stacja UAV unosi się nad ziemią w wolnej przestrzeni na terenie miasta. W pracy rozpatrywany jest przypadek statyczny, w którym drony nie zmieniają swojego położenia – optymalizacja położenia stacji UAV oraz jego wpływ na funkcjonowanie sieci są tematem równoległych działań. W związku z tym charakterystyka poboru mocy uwzględnia zużycie związane z unoszeniem się stacji UAV i świadczeniem przez nią usług na rozważanym terenie. Ruch telekomunikacyjny w sieci generowany jest na podstawie dobowego modelu bazującego na pracy [6], który zaprezentowano na Rys. 2.

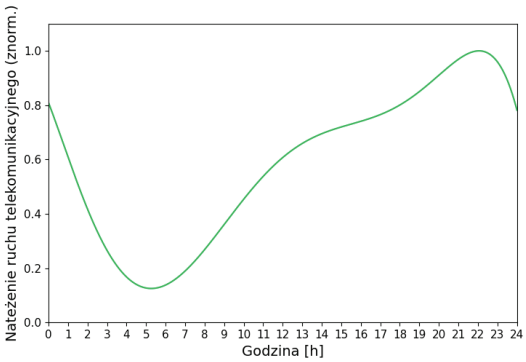


Rys. 1. Schemat scenariusza systemowego

Natomiast w kontekście pozyskiwania zasobów energetycznych z OZE, proces generacyjny został przeanalizowany zarówno dla różnych pór dnia i roku, jak i dla różnych typów zachmurzenia, tj. różnych grubości chmur.

2.2. Model zużycia energetycznego

Zatem zgodnie z [7] i [8], dla każdej stacji bazowej n, gdzie $n = 1, \ldots, N$, energia zużyta w czasie obserwacji t do obsługi ruchu telekomunikacyjnego oraz utrzymania urządzenia nadawczego UAV w powietrzu jest opisana jako:



Rys. 2. Charakterystyka natężenia ruchu sieciowego.

$$E_{\mathrm{cons},n}(t) = \left(P_{\mathrm{hov},n} + \frac{\sum_{i=\tau}^{t+\tau} P_{\mathrm{oper},n}(i)}{l_{\mathrm{i}}}\right) \cdot t, \qquad (1)$$

gdzie przez $P_{\mathrm{oper},n}$ oznaczono pobór mocy przez komponenty mobilnej stacji bazowej odpowiedzialne za utrzymanie połączenia radiowego z użytkownikami w danym momencie i. Symbol τ reprezentuje obecną chwilę czasową, tj. krok symulacyjny, a l_{i} definiuje liczbę próbek pomiarowych mocy $P_{\mathrm{oper},n}$ w czasie obserwacji t. Z kolei $P_{\mathrm{hov},n}$ jest to pobór mocy zainicjowany przez komponenty urządzenia typu dron realizujące efekt jego unoszenia się w powietrzu. Ostatni z tych parametrów, zgodnie z [7], może zostać opisany przez następujące równanie:

$$P_{\mathrm{hov},n} = \sqrt{\frac{(m_n \cdot g)^3}{2\pi \cdot r_{\mathrm{p},n}^2 \cdot l_{\mathrm{p},n} \cdot \rho}}, \qquad (2)$$

gdzie m_n, l_n i r_n reprezentują odpowiednio masę stacji UAV (z uwzględnieniem masy panelu fotowoltaicznego), liczbę jej wirników oraz promień pojedynczego wirnika tej stacji. Kolejno g jest to przyspieszenie ziemskie, a poprzez ρ zdefiniowano gęstość powietrza. Następnie, pobór mocy przez n-ty węzeł dostępowy UAV podczas świadczenia usług w rozpatrywanej sieci bezprzewodowej w chwili τ może zostać oszacowany przy użyciu wzoru [8]:

$$P_{\mathrm{oper},n}(\tau) = \frac{P_{\mathrm{tx},n}}{\mu_{\mathrm{amp},n}} \sigma(\tau) + P_{\mathrm{over},n} + l_{\mathrm{back},n} P_{\mathrm{back},n}, \qquad (3)$$

gdzie $P_{\mathrm{tx},n}$ oznacza maksymalną moc nadawania sygnału RF, $\mu_{\mathrm{amp},n}$ reprezentuje efektywność zastosowanego wzmacniacza mocy sygnału RF, a σ to natężenie ruchu telekomunikacyjnego (znormalizowane względem szczytowego natężenia ruchu w badanym okresie [6]) w chwili τ. Z kolei $P_{\mathrm{over},n}$ jest to stały pobór mocy przez komponenty stacji UAV. Zmienne $l_{\mathrm{back},n}$ i $P_{\mathrm{back},n}$ oznaczają odpowiednio liczbę aktywnych łączy dosyłowych FSO oraz stały pobór mocy dla pojedynczego łącza.

2.3. Model pozyskiwania zasobów energii

W celu oceny produktywności paneli fotowoltaicznych zastosowanych w stacjach UAV, nawiązując do [9] i [10], opracowano model energetyczny aproksymujący ilość pozyskanej energii w czasie obserwacji t (zakładając, że kąt nachylenia powierzchni kolektora słonecznego jest równy 0°):

$$E_{\mathrm{prod},n}(t) = T_{\mathrm{atm}} \cdot I_0 \cdot \mu_{\mathrm{PV},n} \cdot S_{\mathrm{PV},n} \cdot \sin(\alpha) \cdot \varphi_n \cdot t, \quad (4)$$

gdzie T_{atm} i I_0 oznaczają kolejno transmitancję atmosferyczną oraz średnie promieniowanie słoneczne na zewnątrz atmosfery. Za pomocą parametrów $\mu_{\mathrm{PV},n}$ oraz $S_{\mathrm{PV},n}$ zdefiniowano odpowiednio efektywność i powierzchnię paneli fotowoltaicznych, w które wyposażono mobilne stacje bazowe UAV. Kolejno α opisuje kąt między wiązkami promieniowania słonecznego a płaszczyzną równoległą do ziemi (wyrażony w stopniach), który może zostać określony dla dowolnej lokalizacji oraz pory dnia i roku zgodnie z formułami zawartymi w [9] i [11]. Kąt ten określa również wspomnianą transmitancję T_{atm} uzależnioną dodatkowo od lokalnego ciśnienia atmosferycznego (wyrażonego w [Pa]), które z kolei może zostać oszacowane zgodnie ze wzorem zawartym w [12]. Parametr oznaczony jako φ odpowiedzialny jest za modelowanie wpływu zachmurzenia na aktualną moc promieniowania słonecznego. Zgodnie z [10], wartość φ może zostać przybliżona przez następujące sformułowanie:

$$\varphi_n = \begin{cases} e^{-\beta(L_{\mathrm{u}}-L_{\mathrm{d}})}, & \text{jeżeli } h_n < L_{\mathrm{d}} \\ e^{-\beta(L_{\mathrm{u}}-h_n)}, & \text{jeżeli } L_{\mathrm{d}} \le h_n < L_{\mathrm{u}} \\ 1, & \text{jeżeli } L_{\mathrm{u}} \le h_n \end{cases} \quad (5)$$

gdzie h_n, L_{d} i L_{u} to odpowiednio wysokość unoszenia się n-tego drona oraz dolna i górna granica chmur na niebie. Natomiast β (≥ 0) jest to współczynnik absorpcji modelujący charakterystykę optyczną chmury.

3. SYMULACJA

3.1. Konfiguracja przebiegu symulacyjnego

Badania symulacyjne zakładają rozważenie funkcjonowania pojedynczego węzła dostępowego typu dron zasilanego przez panele PV i będącego częścią bezprzewodowego systemu telekomunikacyjnego. Jako lokalizację umiejscowienia mobilnej stacji bazowej UAV przyjęto punkt o ściśle zdefiniowanych współrzędnych, znajdujący się w obrębie miasta Poznań. W przeprowadzonej analizie symulacyjnej porównano własności cyklu energetycznego urządzenia typu dron dla różnych konfiguracji klimatycznych otaczającego go środowiska, tj. dla kilku różnych momentów w roku. W tym celu przebiegi symulacyjne zostały wygenerowane dla czterech pór roku. W szczególności rozważono równonoc wiosenna (20.03.2022), przesilenie letnie (21.06.2022), równonoc jesienna (23.09.2022) i przesilenie zimowe (22.12.2022).

Tab. 1. Wartości parametrów dla stacji bazowej UAV

Parametr	*Wartość*	*Jednostka*
Masa (m_n)	4 [10]	[kg]
Promień wirnika (r_n)	0,25	[m]
Liczba wirników (l_n)	4	
Wysokość latania (h_n)	20	[m]
Maks. moc nad. RF ($P_{\mathrm{tx},n}$)	15 [10]	[W]
Stały pobór mocy ($P_{\mathrm{over},n}$)	5 [10]	[W]
Pobór mocy ł. FSO ($P_{\mathrm{back},n}$)	0,2 [13]	[W]
Liczba łączy FSO ($l_{\mathrm{back},n}$)	1	
Sprawność PV ($\mu_{\mathrm{PV},n}$)	0,2	
Powierzchnia PV ($S_{\mathrm{PV},n}$)	1 [10]	[m^2]

Symulowany dzień dla każdej daty został podzielony na trzy pory, tj. na godziny: 8:30, 12:00 i 15:30, które względem siebie różnią się wielkościami natężenia ruchu sieciowego oraz nasłonecznienia terenu.

W Tab. 1 zamieszono wartości parametrów definiujących specyfikację podzespołów mobilnej stacji UAV, a tym samym określających zarówno stopień utylizacji, jak i efektywność procesu generowania zasobów energetycznych przez drony w danym czasie obserwacji.

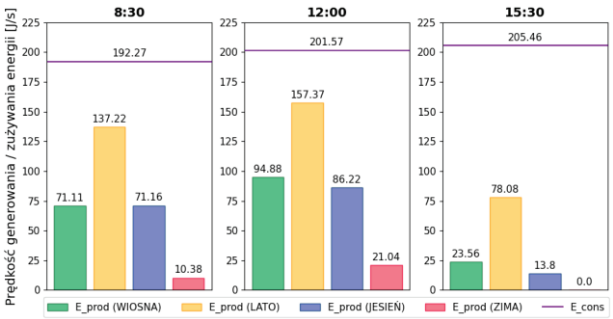
Tab. 2. Wartości parametrów środowiskowych

Parametr	*Wartość*	*Jednostka*
Długość geograficzna	16,95571	[°E]
Szerokość geograficzna	52,40035	[°N]
Czas obserwacji (t)	1	[s]
Przyspieszenie ziemskie (g)	9,8 [10]	[m/s^2]
Gęstość powietrza (ρ)	1,225[10]	[kg/m^3]
Śr. prom. sł. poza atmo. (I_0)	1353 [9]	[W/m^2]
Współ. absorpcji chmur (β)	0,01 [10]	
Grubość chmur ($L_{\mathrm{u}} - L_{\mathrm{d}}$)	0-1000	[m]

Natomiast w Tab. 2 pokazano wartości parametrów odnoszących się do warunków środowiskowych rozważanej sieci bezprzewodowej, takich jak współrzędne geograficzne lokalizacji, przyspieszenie ziemskie, gęstość powietrza, promieniowanie słoneczne oraz jego absorbcja przez chmury na obecne niebie.

3.2. Wyniki symulacyjne

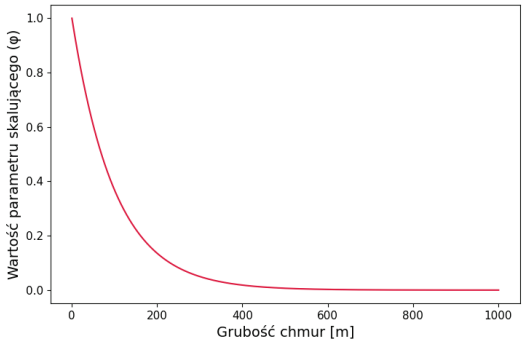
Na poniższych rysunkach zaprezentowane zostały wyniki przebiegów symulacyjnych dla rozpatrywanego scenariusza badawczego. Wykres na Rys. 3 reprezentuje charakterystykę pozyskiwania zasobów energetycznych przez panele fotowoltaiczne dronów w ciągu sekundy dla czterech pór roku (wiosna, lato, jesień, zima) i trzech pór dnia (rano, wczesne południe, późne południe). Ponadto wykres oznaczony linią ciągłą (kolor fioletowy) określa ilość energii zużytej przez stację UAV o danej porze dnia w ciągu 1 s. Zakładając stały model ruchu telekomunikacyjnego przez cały rok (Rys. 2), można stwierdzić, iż pobór mocy przy zachowaniu niezmiennych wartości parametrów stacji UAV pozostaje ten sam przez cały rok. Uzyskane rezultaty przedstawione za pomocą Rys. 3 obrazują stan zgodny z oczekiwaniami. Z perspektywy pory dnia największy odnotowany przyrost energetyczny osiągnięto w południe (godzina 12:00) dla każdego sezonu. Na miejscu drugim plasuje się charakterystyka generacji energii o godzinie 8:30, a na ostatniej pozycji ta o godzinie 15:30. Odwołując się natomiast do pór roku, okres let-



Rys. 3. Charakterystyka produkcji energii (brak chmur)

ni prezentuje się jako najbardziej produktywny. Kolejne pozycje w klasyfikacji zajmują odpowiednio: wiosna, jesień i zima.

Analizując wpływ grubości chmur na otrzymane rezultaty (Rys. 4), można zauważyć redukcję ilości pozyskiwanych zasobów energetycznych o ok. 37% na każde dodatkowe 100 m warstwy zachmurzenia na niebie. Taka charakterystyka tłumienia promieniowania słonecznego reprezentuje spadek w skali hiperbolicznej odwrotnie proporcjonalny do grubości występujących chmur.



Rys. 4. Charakterystyka tłumienia chmur

Warto również nadmienić, iż dla każdego rozważanego przypadku symulacyjnego, tj. dla każdej pory roku i dnia, przy założonych wartościach parametrów dla mobilnej stacji bazowej ilość generowanej energii w ciągu sekundy nie jest w stanie zaspokoić zapotrzebowania na nią w ciągu tego samego okresu. Jednakże z reguły w celu zaspokojenia potrzeb energetycznych urządzenia typu dron wyposażane są w systemy bateryjne. Panele fotowoltaiczne pozyskując dodatkową energię ze Słońca odciążają wykorzystanie energii zakumulowanej w ogniwach baterii stacji UAV. To z kolei przekłada się na rzadszą potrzebę ładowania tych ogniw, a więc w konsekwencji na zredukowanie zużycia zasobów uzyskanych z sieci energetycznej zasilanej w sposób konwencjonalny.

Co więcej, charakterystykę produkowania energii można również zmienić poprzez odpowiednie dostosowywanie parametrów systemowych. Jednakże, aby zachować rzetelność rezultatów należy pamiętać o konsekwencjach, jakie niosą za sobą wprowadzone zmiany (przykładowo wykorzystanie paneli PV o większej powierzchni zwiększa masę drona i może mieć niekorzystny wpływ na jego poruszanie się, a co za tym idzie na świadczenie usług w sieci). Zatem należy wówczas znaleźć optymalne rozwiązanie problemu, które pozwoli na maksymalizację efektywności cyklu energetycznego urządzenia UAV.

4. PODSUMOWANIE

W ramach niniejszego artykułu opracowano oraz zaprezentowano model cyklu energetycznego dla urządzeń typu dron pełniących funkcję punktów dostępowych w bezprzewodowych sieciach telekomunikacyjnych oraz wyposażonych w generatory OZE. Przytoczone wyniki symulacyjne charakteryzują zużycie i pozyskiwanie energii przez stacje UAV. W artykule ukazano również wpływ zachmurzenia oraz pory dnia i roku (w kontekście obciążenia punktu dostępowego ruchem sieciowym, jak i promieniowania słonecznego) na stosunek energii wygenerowanej do zużytej w cyklu energetycznym mobilnej stacji bazowej. Uzyskany w ten sposób model energetyczny ułatwia planowanie oraz predykcję zysku energii pozyskanej z OZE dla wybranych zastosowań sieciowych. W dodatku model ten pozwala określić zarówno optymalny rozmiar, jak i rodzaj paneli fotowoltaicznych do użycia w danej lokalizacji.

LITERATURA

[1] Kozlov V., Sałabun W. 2021. „Challenges in Reliable Solar Panel Selection Using MCDA Methods". *Procedia Computer Science*, 192: 4913–23. Elsevier.

[2] Jahid A., Monju M. K. H., Hossain M. E., Hossain M. F. 2018. „Renewable Energy Assisted Cost Aware Sustainable Off-Grid Base Stations with Energy Cooperation". *IEEE Access,* 6: 60900–920.

[3] Chaoub A., Giordani M., Lall B., Bhatia V., Kliks A., et al. 2022. „6G for Bridging the Digital Divide: Wireless Connectivity to Remote Areas". *IEEE Wireless Communications*, 29 (1): 160–68.

[4] Hossain M. S., Jahid A., Islam K. Z., Alsharif M. H., et al. 2020. „Towards Energy Efficient Load Balancing for Sustainable Green Wireless Networks under Optimal Power Supply". *IEEE Access,* 8: 200635–54.

[5] Arnold O., Richter F., Fettweis G. P., Blume O. 2010. „Power Consumption Modeling of Different Base Station Types in Heterogeneous Cellular Networks". *2010 Future Network & Mobile Summit*. IEEE.

[6] Trinh H. D., Bui N., Widmer J., Giupponi L., Dini P. 2017. „Analysis and Modeling of Mobile Traffic Using Real Traces". *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 1–6. IEEE.

[7] Sekander S., Tabassum H., Hossain E. 2021. „Statistical Performance Modeling of Solar and Wind-Powered UAV Communications". *IEEE Transactions on Mobile Computing,* 20 (8): 2686–2700.

[8] Guo Weisi, Wang Siyi, O'Farrell Tim, Fletcher Simon. 2013. „Energy Consumption of 4G Cellular Networks: A London Case Study". *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*,1–5.

[9] Dimitriou S., Kapsalis S., Panagiotou P., Yakinthos K. 2020. „Energy Efficiency Improvement Potential of a Tactical BWB UAV Using Renewable Energy Sources". *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, 1384–91. IEEE.

[10] Yan S., Xu D., Ng D. W. K., Dai L., Schober R. 2019. „Optimal 3D-Trajectory Design and Resource Allocation for Solar-Powered UAV Communication Systems". *IEEE Transac. on Comm.,* 67 (6): 4281–98.

[11] Khanduri N., Kukreti A., Shah N. 2017. „Implementation of Solar Time Based Sun Tracking Systems for Mobile Platforms and Smart Cities". *2017 IEEE Region 10 Symposium (TENSYMP)*, 1–5. IEEE.

[12] Iribarne J. V., Godson W. L. 1981. „Atmospheric Thermodynamics". Springer Netherlands.

[13] Alzenad M., Shakir M. Z., Yanikomeroglu H., Alouini M. S. 2018. „FSO-Based Vertical Backhaul /Fronthaul Framework for 5G+ Wireless Networks". *IEEE Communications Magazine*, 56 (1): 218–24.

PROCEDURA ROZMIESZCZENIA MOBILNYCH PRZEKAŹNIKÓW DOSYŁOWYCH FSO ZASILANYCH Z OZE*

PROCEDURE FOR THE DEPLOYMENT OF FSO MOBILE BACKHAUL RELAYS POWERED BY RES

Salim Janji[1]; Adam Samorzewski[2]; Adrian Kliks[3]

[1] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, salim.janji@doctorate.put.poznan.pl
[2] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, adam.samorzewski@doctorate.put.poznan.pl
[3] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2022.4.62

Streszczenie: **W artykule rozważono zastosowanie komunikacji optycznej w wolnej przestrzeni (ang. Free Space Optics, FSO) dla łączności dosyłowej w sieci małych komórek umieszczonych na dronach. Założono, że drony są wyposażone w aparaturę nadawczo-odbiorczą FSO, przez co mogą służyć jako węzły przekaźnikowe. Dodatkowo założono, że drony wyposażone są w panele słoneczne pozwalające zwiększyć czas ich pracy. W pracy przedstawiono algorytm rozmieszczenia takich stacji przekaźnikowych w środowisku miejskim o wysokiej zabudowie. Zapewnia on wzajemną widoczność stacji dla poprawnej transmisji danych FSO, preferując wybór miejsc o dużym nasłonecznieniu.**

Abstract: **The paper considers the use of Free Space Optics (FSO) for backhaul communication in a network of drone small cells. It was assumed that drones are equipped with FSO transceivers, thus they can serve as relays. In addition, it was assumed that the drones are equipped with solar panels to increase their operating time. The paper presents an algorithm for the distribution of such relay stations in an urban environment with high buildings. It ensures mutual visibility of stations for correct FSO data transmission, preferring the choice of places with high sun exposure.**

Słowa kluczowe: **FSO, heterogeniczne sieci komórkowe, małe komórki, Odnawialne Źródła Energii, UAV.**

Keywords: **FSO, heterogeneous cellular networks, Renewable Energy Sources, Small Cells, UAV.**

1. WSTĘP I PRZEGLĄD LITERATURY

Współczesne sieci komórkowe są coraz silniej heterogeniczne z perspektywy różnorodności rodzajów komórek w nich występujących. W szczególności coraz częściej wykorzystywane są tzw. małe komórki (ang. Small Cells, SCs) w ramach strategii zagęszczania i zmniejszania pokrycia stacji bazowych. Pozwala to spełnić coraz trudniejsze wymagania dotyczące rosnącej szybkości transmisji danych i niskich opóźnień [1]. Bardzo ciekawym rozwiązaniem jest propozycja wykorzystania bezzałogowych platform powietrznych BPP (dronów – UAV, ang. Unmanned Arieal Vehicles) jako stacji bazowych. Stacje takie mogą dostosowywać swoje położenie do lokalizacji użytkowników i do warunków środowiskowych. Równocześnie rozmieszczenie wielu stacji bazowych na dronach wymaga zapewnienia niezawodnej łączności dosyłowej (ang. backhaul) [2]. Badania pokazują, że wykorzystanie komunikacji optycznej w wolnej przestrzeni (ang. Free Space Optics, FSO) zapewnia bardzo dobrą wydajność, także w porównaniu z łączami o częstotliwości radiowej (ang. Radio Frequency, RF) [3]. Podejście takie rozważano w kilku pracach badawczych, np. w [4] autorzy rozważyli przypadek, w którym urządzenia BPP służą jako przekaźniki (ang. Drone Relay Station, DRS) sygnału FSO pomiędzy węzłami stacjonarnymi. Co więcej, czas lotu BPP jest ograniczony żywotnością baterii pokładowej. Aby poradzić sobie z ograniczoną ilością energii, w [5] zademonstrowano prototyp BPP wyposażony w panele fotowoltaiczne (ang. Photovoltaics, PV). W tym artykule rozważono problem rozmieszczania przekaźników DRS zasilanych przez Odnawialne Źródła Energii (OZE) na badanym terenie w taki sposób, aby zapewnić łączność dosyłową pomiędzy nimi. Jako generatory energii z OZE przyjęto panele fotowoltaiczne montowane na górnej obudowie urządzeń BPP. W pracy rozważono środowisko miejskie o wysokiej zabudowie, zakładając, że z racji wysokości drony nie mogą być umiejscowione ponad budynkami.

Artykuł składa się z sześciu rozdziałów. W drugiej części opisano rozważany scenariusz wraz z modelem zużycia, gromadzenia i magazynowania energii. W rozdziale trzecim przedstawiono proponowane rozwiązanie znajdowania lokalizacji DRS wykorzystujące tzw. graf widoczności. Następnie opisano rozszerzony mechanizm wyznaczania najkrótszej ścieżki propagacyjnej uwzględniający priorytet w wyborze słonecznych lokalizacji. W rozdziale piątym pokazano wyniki symulacji, a sekcja szósta zawiera podsumowanie.

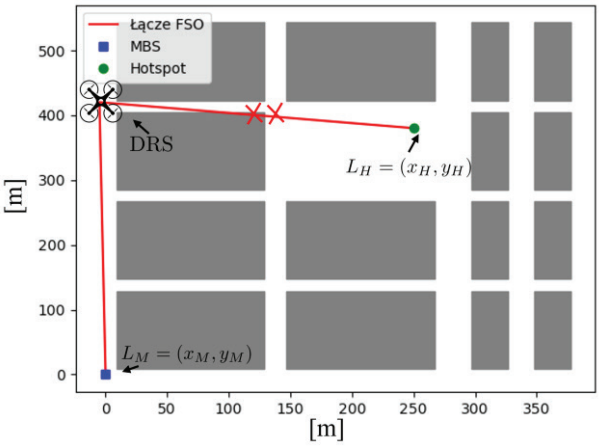
2. MODEL SYSTEMU

2.1. Opis scenariusza

W artykule rozważono obszar miejski o wysokiej zabudowie z jedną stacją bazową typu makro (MBS, ang. Macro Base Station) zlokalizowaną w punkcie $L_M = (x_M, y_M)$. Założono istnienie obszaru o dużej gęstości użytkowników (tzw. hotspot) umieszczonego w punkcie $L_H = (x_H, y_H)$. Stacja bazowa obsługująca ten

* Praca sfinansowana ze środków statutowych 0312/SBAD/8161

obszar jest połączona z MBS za pośrednictwem optycznych łączy dosyłowych FSO umieszczonych na stacjach przekaźnikowych DRS. Z tego powodu kolejne przeskoki między węzłami nie mogą napotykać żadnych przeszkód fizycznych blokujących tor optyczny między nimi. Rys. 1 prezentuje przykładowy schemat systemu, w którym stacja DRS działa jako przekaźnik pomiędzy hotspotem a stacją MBS. Ponadto w pracy założono, że stacje DRS są wyposażone w panele słoneczne, które dostarczają energię do akumulatorów pokładowych, gdy przekaźniki znajdują się w nasłonecznionych miejscach.



Rys. 1. Wysoka zaobudowa miejska z MBS i hotspotem

2.2. Model zużycia energii

Zgodnie z [6], dla każdego n-tego przekaźnika DRS, $(n = 1, \dots, N)$, energia aerodynamiczna zużyta w czasie obserwacji t_o jest opisana jako $E_{\mathrm{UAV},n}(t_o) = P_{\mathrm{hov},n} \cdot t_o$, gdzie $P_{\mathrm{hov},n}$ to pobór mocy potrzebny stacji UAV do unoszenia się na ziemią. Wartość tego parametru dla n-tego DRS może zostać określona następująco:

$$P_{\mathrm{hov},n} = \sqrt{\frac{(m_n g)^3}{2\pi r_{\mathrm{p},n}^2 l_{\mathrm{p},n} \rho}}, \tag{1}$$

gdzie m_n to masa drona, a g to przyspieszenie grawitacyjne. Wyrażenie $2\pi r_{\mathrm{p},n}^2 l_{\mathrm{p},n}$ opisuje całkowitą powierzchnię wirników stacji UAV, gdzie $l_{\mathrm{p},n}$ to ich liczba, a $r_{\mathrm{p},n}$ to promień pojedynczego wirnika. Z kolei ρ to gęstość powietrza. Następnie energia zużyta do podtrzymania połączenia FSO jest opisana przez $E_{\mathrm{back},n}(t_o) = P_{\mathrm{back},n} \cdot t_o$, gdzie $E_{\mathrm{back},n}$ oznacza ilość energii w [Wh] przeznaczoną na utrzymanie łącza dosyłowego, gdzie $P_{\mathrm{back},n}$ definiuje pobór mocy na łączu FSO i jest to wartość stała.

2.3. Model pozyskiwania energii

Aby ocenić wydajność paneli fotowoltaicznych obsługujących drony, zastosowano następujący model pozyskiwania energii: $E_{\mathrm{PV},n}(t_o) = P_{\mathrm{PV},n} \cdot t_o$, gdzie $P_{\mathrm{PV},n}$ opisuje moc generowaną przez pojedynczy panel słoneczny zamontowany na obudowie n-tego DRS. Zakładając, że kąt nachylenia powierzchni kolektora słonecznego jest równy 0°, to zgodnie z [7]

$$P_{\mathrm{PV},n} = T_{\mathrm{atm}} \cdot I_0 \cdot \delta \cdot \eta_{\mathrm{PV},n} \cdot S_{\mathrm{PV},n} \cdot \sin(\alpha), \tag{2}$$

gdzie α to kąt między wiązkami promieniowania słonecznego a płaszczyzną równoległą do ziemi podany w stopniach. Z kolei $\eta_{\mathrm{PV},n}$ i $S_{\mathrm{PV},n}$ to odpowiednio współczynnik sprawności i powierzchnia panelu fotowoltaicznego. Następnie I_0 oznacza średnią powierzchniową gęstość mocy promieniowania słonecznego na zewnątrz atmosfery, a T_{atm} określa transmitancję atmosferyczną. Ponadto, aby rozważyć wpływ chmur na pozyskiwanie energii z promieniowania słonecznego, wykorzystany został parametr skalujący δ, który jest liczbą pseudolosową generowaną z rozkładu równomiernego o określonym zakresie zgodnie z symulowanymi warunkami pogodowymi, które mogą być zdefiniowane jako dobre, umiarkowane lub złe.

2.4. System składowania energii

Zakładając idealną wydajność ładowania, magazynowanie energii w czasie obserwacji $t_o = t_s - t_{s-1}$ zgodnie z [8] jest sformułowane jako:

$$E_{\mathrm{batt},n}(t_s) = \tag{2}$$
$$E_{\mathrm{batt},n}(t_{s-1}) + E_{\mathrm{PV},n}(t_o) - E_{\mathrm{UAV},n}(t_o) - E_{\mathrm{back},n}(t_o)$$

gdzie $E_{\mathrm{batt},n}(t_s)$ to energia zgromadzona w baterii n-tego drona w chwili t_s.

3. TRANSMISJA DOSYŁOWA FSO

3.1. Definicja problemu

Celem algorytmu jest znalezienie ścieżki komunikacyjnej od L_M do L_H, która „nie przechodzi" przez żaden budynek (Rys. 1). Biorąc pod uwagę, że najkrótsza droga między przeszkodami to zbiór prostych linii, które łączą źródło z celem poprzez możliwie najmniej złożoną sekwencję wierzchołków przeszkód (budynków) [9], rozważany problem można opisać jako tworzenie tzw. grafu widoczności. W tym przypadku graf wejściowy G opisuje podstawowy zbiór geometrycznych obiektów (np. wielokątów wypukłych zdefiniowanych jako zbiór wierzchołków i krawędzi łączących je), a wynikowym rozwiązaniem problemu jest graf widoczności G_v, który zawiera wszystkie możliwe krawędzie pomiędzy dowolnymi dwoma wierzchołkami w G tak, że te dwa wierzchołki nie są blokowane przez żadną przeszkodę. Ostatecznie stosując algorytm najkrótszej ścieżki (np. Dijkstry) dla wynikowego grafu można wybrać optymalne lokalizacje dla stacji DRS. W pracy użyto algorytm Lee opisany w [10].

3.2. Algorytm rozmieszczenia DRS

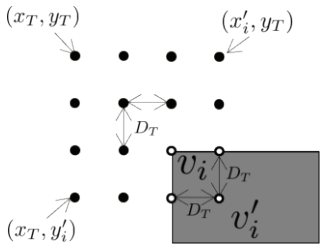
Mając mapę przeszkód, można wyodrębnić graf $G(V, E)$ zawierający wszystkie wierzchołki V i krawędzie E, które definiują budynki (tj. narożniki i krawędzie budynków w widoku z góry). Do zbioru wierzchołków, V, należy dodać dwa punkty definiujące położenie początkowe i końcowe (podane odpowiednio przez L_M i L_H), aby obliczyć również ich ścieżkę widoczności. Następnie należy wdrożyć algorytm Lee dla każdego wierzchołka $v_i \in$ V, tak aby uzyskać kompletny wykres widoczności G_v w sposób przedstawiony w [11].

4. LOKOWANIE DRS ZASILANYCH Z PV

Aby poprawić efektywność energetyczną, w artykule rozważono wykorzystanie stacji DRS zasilanych energią słoneczną oraz sposób ich umieszczania na nasłonecznionym obszarze z występującymi zacieniowanymi miejscami, które z kolei wynikają z przysłaniania światła słonecznego przez budynki.

4.1. Algorytm rozmieszczenia DRS

Ponieważ DRS powinny być umieszczane w miejscach nasłonecznionych, pierwszym krokiem jest znalezienie zbioru wszystkich możliwych punktów o danej godzinie dnia, które są słoneczne (tj. nie są zasłonięte przez żaden budynek). Dobrym punktem wyjścia, który prowadzi do znalezienia najkrótszej ścieżki, jest zbiór wierzchołków budynków V, który został użyty w poprzedniej sekcji. W ten sposób, mając na uwadze kąt między wiązkami promieniowania słonecznego i płaszczyzną poziomą, kąt azymutu oraz zbiór przeszkód zdefiniowany przez graf $G(V,E)$, można sprawdzić dowolny punkt w rozpatrywanym obszarze, aby stwierdzić, czy jest on zacieniony, czy nie. Aby znaleźć zbiór punktów słonecznych V_s należy wykonać następujące kroki dla każdego wierzchołka $v_i = (x_i, y_i) \in V$. Niech $v_i' = (x_i', y_i')$ będzie punktem we wnętrzu wielokąta, którego wierzchołek v_i jest taki, że v_i' jest w równej odległości D_T od dwóch krawędzi połączonych z v_i (Rys. 2). Najpierw tworzona jest siatka punktów testowych wokół wierzchołka v_i oznaczona $V_T = \{(x_j, y_j) \mid x_j = x_i' \pm nD_T,\ y_j = y_i' \pm nD_T$, gdzie $n = 1, 2, \ldots, C_T$, a $(x_j, y_j) \notin \{(x_i', y_i'), (x_i', y_i), (x_i, y_i)\}\}$, gdzie odejmowanie wzdłuż osi x następuje, gdy $x_T < x_i'$, w przeciwnym razie należy zastosować dodawanie. Podobnie wzdłuż osi y. C_T i D_T określają odpowiednio liczbę punktów testowych i odstępy między nimi. Punkt początkowy nie jest brany pod uwagę, ponieważ znajduje się w budynku. Podobnie punkty znajdujące się na krawędziach należy odrzucić w celu uwzględnienia nierówności powierzchni budynku, które mogłyby zasłaniać światło (np. balkony, banery reklamowe itp.). Punkty ze zdefiniowanego zbioru są kolejno testowane, zaczynając od punktów znajdujących się najbliżej budynku, w taki sposób, że gdy punkt zostanie znaleziony w nasłonecznionym miejscu, ten punkt zostanie dodany do V_s, a wyszukiwanie zostanie zatrzymane.



Rys. 2. Punkty testowe do szukania nasłonecznionych miejsc ($C_T = 4$)

4.2. Lokowanie DRS w słonecznych miejscach

Po uzyskaniu zbioru V_s słonecznych lokalizacji można teraz zastosować algorytm Lee (patrz podrozdział 3.2), jednak tym razem ze zbiorem wierzchołków (i punktów) zdefiniowanym jako $V' = \{V \cup V_s \cup \{L_H, L_M\}\}$. Spowoduje to również uzyskanie grafów widoczności dla nowych punktów V_s. Po określeniu ścieżki widoczności ponownie należy wykonać algorytm Dijkstry z funkcją kosztu zdefiniowaną dla każdej krawędzi $E_k = \{v_i, v_j\}$ w następujący sposób:

$$D - \begin{cases} \frac{d(v_i,v_j)}{d_{\max}} + 1, & \textit{jeżeli } v_j \in V_s, \\ \frac{d(v_i,v_j)}{d_{\max}} + \frac{100}{\delta_{\text{avg}}} & \textit{w przeciwnym razie.} \end{cases} \tag{5}$$

gdzie $d(v_i, v_j)$ to odległość euklidesowa między dwoma punktami końcowymi, a $d_{\max}$ jest współczynnikiem normalizującym, który ogranicza koszt odległości do przedziału $[0, 1]$ i może być ustawiony na maksymalną odległość między dwoma dowolnymi punktami na rozpatrywanym obszarze. Zależność (5) ma na celu zwiększenie przez algorytm kosztu za umieszczenie stacji DRS w zacienionym miejscu o wartość $\frac{100}{\delta_{\text{avg}}}$, gdzie δ_{avg} definiuje oczekiwane warunki pogodowe. Dla małych wartości δ_{avg} koszt jest maksymalizowany, ponieważ oczekuje się dobrych wyników energetycznych w związku z pozyskaniem zasobów z paneli słonecznych. Natomiast zwiększenie wartości parametru δ_{avg} zmniejsza koszt za wybranie zacienionego miejsca.

5. SYMULACJA I WYNIKI

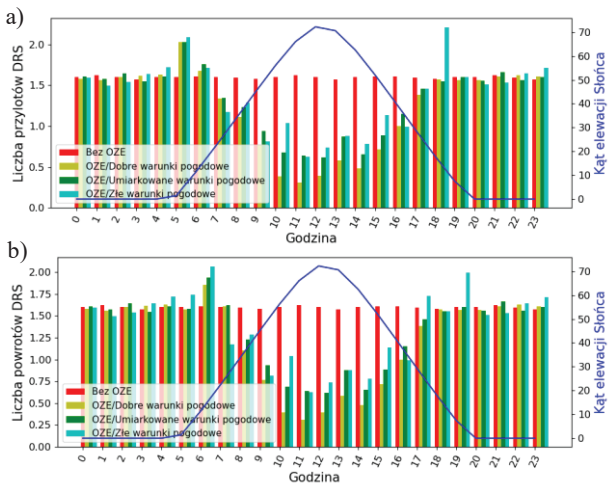
W przebiegach symulacyjnych do przeprowadzenia obliczeń wykorzystano charakterystykę promieniowania słonecznego w ciągu doby dla miasta Madryt w Hiszpanii. Ponadto, przyjęto model środowiska pokazany na Rys. 1. Parametry symulacji zostały wymienione w Tab. 1.

Tab. 1. Przyjęte wartości parametrów symulacyjnych

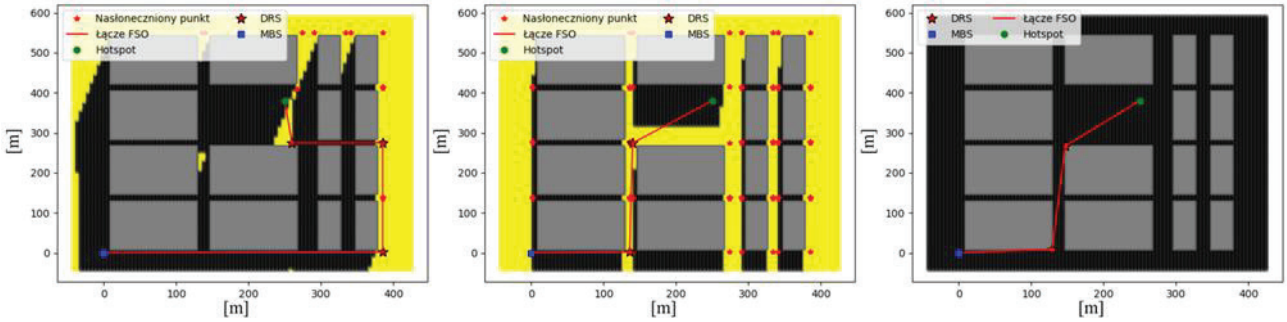
Parametr	*Wartość*	*Jednostka*
Długość geograficzna	3,68892	[°W]
Szerokość geograficzna	40,42218	[°N]
Wysokość lotu UAV (h)	25	[m]
Śr. prom. sł. poza atmo. (I_0)	1353 [7]	[W/m^2]
Wsp. efektywności PV (η_{PV})	0,2	
Powierzchnia PV (S_{PV})	1 [12]	[m^2]
Masa UAV (m)	4,5 [12]	[kg]
Promień wirnika UAV (r_p)	0,25	[m]
Liczba wirników UAV (l_p)	4	
Pojemność baterii UAV	222 [12]	[Wh]
Pobór mocy na ł. FSO (P_{back})	200 [13]	[mW]
Wsp. normalizujący (d_{max})	650	[m]
Liczba pkt. testowych (C_T)	4	
Odstęp między pkt. test. (D_T)	8	[m]

5.1. Umiejscowienie DRS dla różnych pór dnia

Rys. 3 przedstawia mapę rozmieszczenia stacji DRS w wyniku użycia algorytmu, z uwzględnieniem obszarów nasłonecznionych oraz zacienionych dla 3 różnych pór



Rys. 4. Śr. l. przylotów (a) i powrotów (b) DRS na godz.



(a) (b) (c)

Rys. 3 Liczba wymaganych DRS o godzinie 7:00 (a), 8:00 (b) oraz w nocy (c)

dnia (o godzinie 7:00, 8:00 oraz w nocy). Warto zauważyć, iż algorytm jest w stanie znaleźć dla każdej pory dnia najkrótszą ścieżkę przechodzącą przez słoneczne lokalizacje bez wybierania miejsc, w których znajduje się cień. Jednakże, gdy słoneczne miejsca nie występują (noc), wówczas to kryterium jest pomijane przez algorytm i wybierana jest ścieżka o najmniejszej liczbie przeskoków.

5.2. Zrównoważenie zużycia zasobów sieci

Na Rys. 4a można zaobserwować liczbę nowych przylotów stacji DRS z centrum ładowania na godzinę, a liczba lotów powrotnych DRS na godzinę została przedstawiona na Rys. 4b. Przyloty i powroty mobilnych przekaźników mogą wystąpić w trzech przypadkach: w celu doładowania, gdy potrzebny jest nowy DRS lub gdy stał się on zbędny. Obie figury przedstawiają wyniki z użyciem i bez zastosowania OZE, oraz dla 3 różnych warunków pogodowych (bezchmurne, średnio zachmurzone i całkowicie zachmurzone niebo) zdefiniowanych przez zakres parametru skalowania δ. Wartości liczb przybyć i powrotów są znacznie zmniejszone w godzinach, w których kąt elewacji słońca jest powyżej zera, a przekaźniki korzystają z paneli fotowoltaicznych. Ilość energii odnawialnej uzyskanej z paneli słonecznych przy wspomnianych stanach pogody, tj. dobrym (bezchmurne niebo), umiarkowanym (średnio zachmurzone niebo) i złym (całkowicie zachmurzone niebo), wynoszą odpowiednio 4477 Wh, 3717 Wh i 3070 Wh.

6. WNIOSKI

W artykule zaproponowano rozwiązanie problemu świadczenia usług na łączu dosyłowym FSO za pomocą grupy DRS zasilanych z paneli fotowoltaicznych oraz wdrażanych z uwzględnieniem istniejących przeszkód za pomocą grafu widoczności. Wyniki symulacji wykazały znaczny spadek wymaganej liczby lotów potrzebnych do naładowania stacji DRS. W związku z tym dla rozpatrywanych warunków pogodowych zaobserwowano duży spadek wykorzystania energii nieodnawialnej.

LITERATURA

[1] Andrews J. G., Buzzi S., Choi W., Hanly S. V., et al. 2014. „What Will 5G Be?". *IEEE Journal on Selected Areas in Communications*, 32 (6): 1065–82.

[2] Siddique U., Tabassum H., Hossain E., Kim D. I. 2015. „Wireless backhauling of 5G small cells: challenges and solution approaches". *IEEE Wireless Communications*, 22 (5): 22–31.

[3] Chaudhry A. U., Yanikomeroglu H. 2021. „Free Space Optics for Next-Generation Satellite Networks". *IEEE Con. Elec. Magazine*, 10 (6): 21–31.

[4] Fawaz W., Abou-Rjeily C., Assi C. 2018. „UAV-Aided Cooperation for FSO Communication Systems". *IEEE Comm. Magazine*, 56 (1): 70–75.

[5] Morton S., D'Sa R., Papanikolopoulos N. 2015. „Solar powered UAV: Design and experiments". *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2460–66. IEEE.

[6] Sekander S., Tabassum H., Hossain E. 2021. „Statistical Performance Modeling of Solar and Wind-Powered UAV Communications". *IEEE Transactions on Mobile Computing*, 20 (8): 2686–2700.

[7] Dimitriou S., Kapsalis S., Panagiotou P., Yakinthos K. 2020. „Energy Efficiency Improvement Potential of a Tactical BWB UAV Using Renewable Energy Sources". *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, 1384–91. IEEE.

[8] Jahid A., Monju M. K. H., Hossain M. E., Hossain M. F. 2018. „Renewable Energy Assisted Cost Aware Sustainable Off-Grid Base Stations with Energy Cooperation". *IEEE Access,* 6: 60900–920.

[9] Lozano-Pérez T., Wesley M. A. 1979. „An Algorithm for Planning Collision-Free Paths among Polyhedral Obstacles". *Comm. ACM*, 22 (10): 560–570.

[10] Lee D. T. 1978. „Proximity and Reachability in the Plane".

[11] Coleman D. 2012. „Lee's O (n2 log n) Visibility Graph Algorithm Implementation and Analysis".

[12] Yan S., Xu D., Ng D. W. K., Dai L., Schober R. 2019. „Optimal 3D-Trajectory Design and Resource Allocation for Solar-Powered UAV Communication Systems". *IEEE Tran. on Comm.,* 67 (6): 4281–98.

[13] Alzenad M., Shakir M. Z., Yanikomeroglu H., Alouini M. S. 2018. „FSO-Based Vertical Backhaul /Fronthaul Framework for 5G+ Wireless Networks". *IEEE Communications Magazine*, 56 (1): 218–24.

AUTONOMICZNY BSP DO ROZPOZNANIA I ZAKŁÓCANIA ŁĄCZNOŚCI SIECI KOMÓRKOWYCH

UAV FOR CELLULAR NETWORKS SIGNAL ANALYSIS AND JAMMING

Jan Dułowicz[1]; Paweł Skokowski[2]

[1]Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, jan.dulowicz@student.wat.edu.pl
[2]Wojskowa Akademia Techniczna, Instytut Systemów Łączności, Warszawa, pawel.skokowski@wat.edu.pl

DOI: 10.15199/59.2022.4.63

Streszczenie: **W artykule przedstawiono koncepcję systemu do rozpoznania i zakłócania łączności radiowej, opartej na systemach komórkowych, umieszczonego na bezzałogowym statku powietrznym (BSP). Zaproponowana koncepcja, jej wstępna implementacja, wyniki z przeprowadzonych badań oraz przedstawione wnioski składają się na treść opracowania.**

Abstract: **The article presents the concept of a system for recognizing and jamming radio communication, based on cellular systems, located on an unmanned aerial vehicle (UAV). The proposed concept, its initial implementation, the results of the conducted research and the presented conclusions constitute the content of the study.**

Słowa kluczowe: **rozpoznanie radiowe, zakłócanie, BSP, systemy komórkowe**

Keywords: **spectrum sensing, jamming, UAV, cellular systems**

1. WSTĘP

Proponowane rozwiązanie jest systemem w kategorii operacyjno-rozpoznawczej. Jako zasadnicze warunki pracy systemu należy rozumieć misje poza granicami kraju, gdzie zakłócenie kluczowej stacji bazowej może ułatwić realizacje zadań przydzielonych wojskom specjalnym. Należy przy tym zaznaczyć, że zakłócanie winno być prowadzone z możliwie najniższą mocą promieniowaną za względu na możliwe wykrycie sygnału zakłócającego przez systemu rozpoznania radioelektronicznego EW (ang. Electronic Warfare) przeciwnika.

W konwencjonalnym podejściu systemy zakłócające generują duże moce w znacznej odległości od źródła sygnału zakłócanego. W proponowanym rozwiązaniu, ze względu na możliwą niewielką do uzyskania odległość bezzałogowego statku powietrznego UAV (ang. Unmanned Aerial Vehicle) od stacji bazowej (dron podlatuje do źródła promieniowania/odbioru sygnału możliwie blisko) wymagana moc promieniowana do uniemożliwienia łączności jest znacznie mniejsza niż w klasycznych rozwiązaniach. Warto podkreślić, że takie podejście umożliwia również dłuższą pracę UAV ze względu na oszczędności pokładowego źródła zasilania. Wpływa to również na utrudnione rozpoznanie źródła emisji sygnału zakłócającego, ponieważ sygnał zakłócający będzie wysyłany wyłącznie w przypadku wykrycia transmisji. Dodatkowo system zamontowany na statku powietrznym będzie łatwy w transporcie oraz jego obsługa będzie mogła pracować z dala od strefy potencjalnego zagrożenia ze strony wroga.

Głównym elementem systemu jest wielokanałowe programowalne radio (ang. SDR – software-defined radio)[3], które w sposób autonomiczny wykrywa emisje na wybranych częstotliwościach na pierwszym torze radiowym oraz w oparciu o wyniki analizy częstotliwościowej – wysyła w rejonie pracy stacji bazowej sygnał zakłócający na drugim torze radiowym.

2. OPIS PROBLEMU I PROPOZYCJA ROZWIĄZANIA

2.1. Opis problemu

Problematyka projektu głównie skupia się na detekcji sygnału [10][4][11][12][13][2] ze stacji bazowej, analizie częstotliwościowej odebranego sygnału[1][6][7][8] i na tej podstawie generacji sygnału zakłócającego [15][9][5] przez radio SDR do stacji bazowej. Jako istoty element założenia należy rozważyć kwestie wymaganej minimalnej mocy odebranego sygnału, która będzie tożsama z progiem wykrycia emisji [14][16]. Jako kluczowe parametry wymagane przy analizie można wymienić:

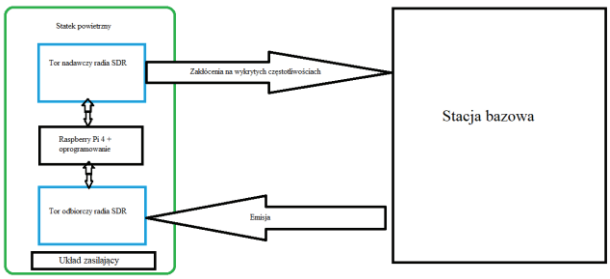
poziom mocy sygnału emitowanego przez stacje bazowe,
poziom mocy sygnału, przy którym dochodzi do wykrycia emisji,
poziom sygnału zakłócającego emitowanego przez radio SDR,
zbiór częstotliwości, na których prowadzone są emisje od stacji bazowych,
zbiór częstotliwości, na których prowadzane będą emisje zakłócające do stacji bazowych.

2.2. Proponowane rozwiązanie

Koncepcja proponowanego autonomicznego BSP do rozpoznania łączności sieci komórkowych i zakłócania stacji bazowych opiera się na kilku połączonych ze sobą komponentach, rozbudowanych o autorski program do zapewnienia autonomii całego systemu (patrz Rys.1).

Na komercyjnym statku powietrznym (DJI Mavic 3) zamontowane jest programowalne radio SDR USRP B200 mini pracujące w paśmie od 70 MHz do 6 GHz. SDR jest podłączone do komputera pokładowego, zbudowanego w oparciu o mikrokomputer Raspberry Pi 4. Mikrokomputer ma w sobie zaimplementowany program,

który w sposób autonomiczny zmienia zakres częstotliwości pracy radia. Platforma radia programowalnego jest 2 kanałowa. Na jednym z nich odbierane są sygnały radiowe i przekształcane wstępnie do reprezentacji sygnału w postaci próbek IQ (ang. In-phase, Quadraturate). Następnie odbierane sygnały są automatycznie przekształcane z wykorzystaniem szybkiej transformaty Fouriera (FFT) na postać częstotliwościową. Jeśli poziom mocy na częstotliwości osiągnie wartość progową wówczas drugi tor radiowy radia SDR jest przestrajany na wykrytą aktywność radiową (zakłócanie odzewowe) i rozpoczyna emisje sygnału zakłócającego do momentu, aż sygnał wykryty nie zaniknie.



Rys. 1. Koncepcja inteligentnego systemu zakłóceń

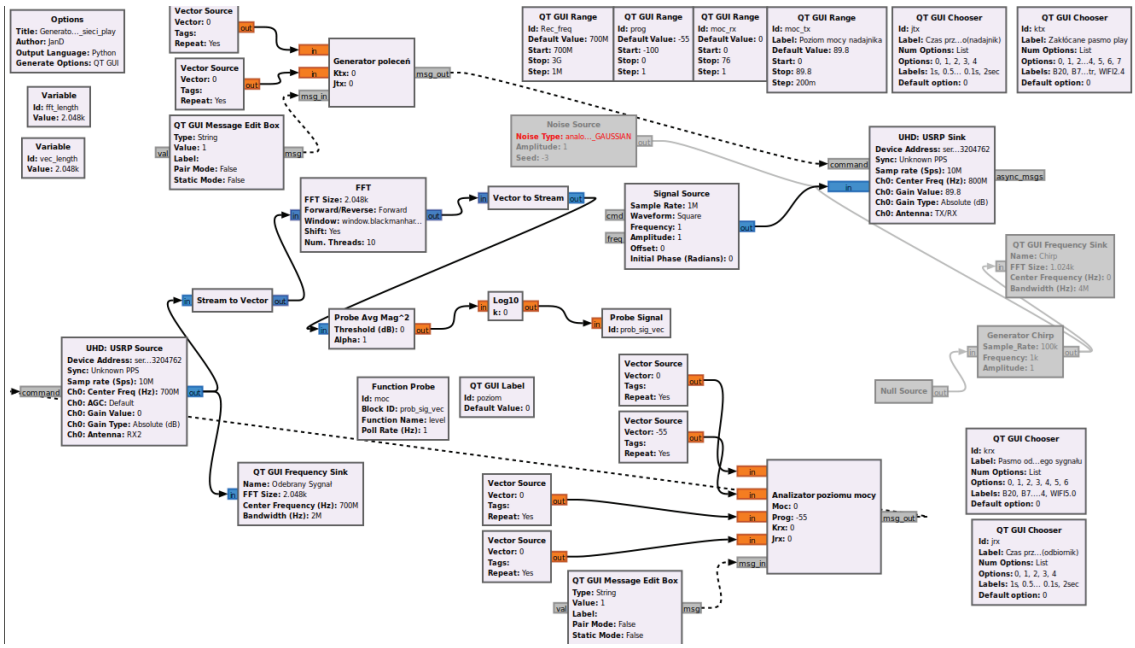
Tor nadawczy oraz odbiorczy są dodatkowo wyposażone w zestaw cyfrowych filtrów pasmowo-przepustowych w formie elementu składowego programu komputerowego. To rozwiązanie umożliwia ograniczenie zakresu skanowanych sygnałów do tych, na których pracują stacje bazowe.

Proponowane rozwiązanie monitorowania widma częstotliwościowego na wybranych zakresach pozwala zawęzić zakres pomiarów do wymaganych częstotliwości, ograniczyć emisję sygnału zakłócającego wyłącznie po wykryciu sygnału. Dzięki temu układ w znacznym stopniu może zaoszczędzić wykorzystanie mocy oraz wydłużyć czas pracy systemu. System w zależności od konfiguracji ustawionej w module mikrokomputera pokładowego ma w założeniu umożliwia zakłócanie wybranych rodzajów łączności telefonii komórkowej takich jak GSM, UTMS, LTE lub 5G.

Warto podkreślić, że funkcjonalności proponowanego BSP są podobne do komercyjnych rozwiązań stosowanych w pododdziałach walki elektronicznej obcych armii. Zastosowanie małych UAV do przenoszenia systemów zakłócających staje się coraz bardziej powszechne. Jako przykład można wskazać rosyjskie UAV ORLAN-10 pracujące w ramach systemu LEER 3

3. IMPLEMENTACJA

Funkcjonalność systemu zakłócającego jest realizowana przez programowalne radio SDR B200 mini zamontowane na platformie DJI Mavic 3. Algorytm działania radia jest realizowany przez program komputerowy w oparciu o środowisko GNURadio, w którym zaimplementowano wybrane bloki funkcjonalne realizujące wybrane zadania radia SDR (patrz Rys.2). Program w momencie uruchomienia grafu przepływu generuje próbki IQ bezpośrednio do radia SDR, na których podstawie są realizowane funkcje kontroli i obsługi toru odbiorczego i nadawczego radia SDR.



Rys. 2. Graf przepływu sterujący torem nadawczym i odbiorczym w GNURadio

Do uruchomienia programu i generacji próbek IQ wykorzystano moduł mikrokomputera Raspberry PI4 , ze

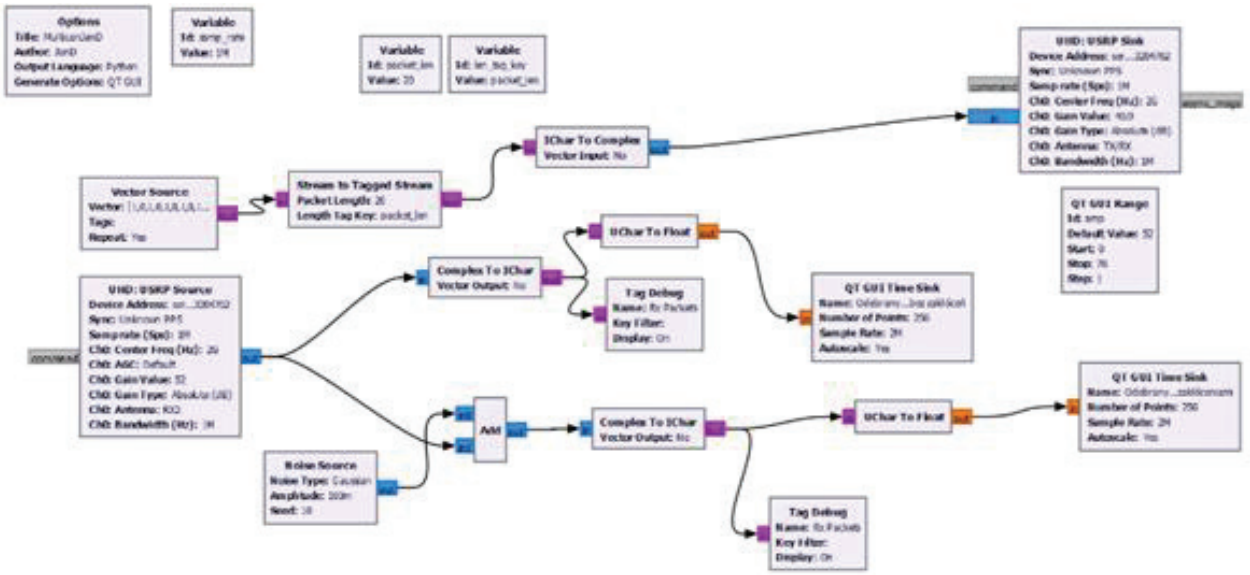
względu na jego rozmiary i niską masę. Środowisko GNURadio obsługuje bloki napisane w języku programowania Python, za pomocą których można realizować niestandardowe zadania takie jak automatyczna zmiana częstotliwości pracy wybranego toru radiowego. Do zakłócenia sygnałów telefonii wykorzystano sygnał szumu białego.

4. STANOWISKO BADAWCZE I SCENARIUSZ TESTOWY

Do wykonania stanowiska badawczego wykorzystano radio SDR B200mini, wyposażone w tor nadawczo-odbiorczy (TX/RX) oraz tor odbiorczy (RX2).

Na tor nadawczo-odbiorczy (TX/RX) wysłany został wektor bitów symulujący transmisje danych od telefonu komórkowego (transmisja uplink). Transmitowane dane zostały odebrane przez tor odbiorczy (RX2) symulujący stację bazową, gdzie zostały zmieszane z sygnałem zakłócającym (szum biały). Wyniki badania zostały przedstawione jako dwie charakterystyki amplitudowo-częstotliwościowe, gdzie pierwsza reprezentuje sygnał telefonu komórkowego odebrany przez stację bazową, a druga – sygnał telefonu komórkowego odebrany przez stację bazową z dodaniem zakłóceń.

Moc sygnału transmisji komórkowej odbierana przez stację bazową jest stosunkowo niska, ponieważ maksymalna wartość emitowana przez telefon komórkowy nie przekracza wartości 2 [W]. Sygnał zakłócający emitowany przez radio SDR zamontowane na bezzałogowym statku powietrznym jest emitowany w bliskiej okolicy anteny odbiorczej stacji bazowej. Można więc założyć, że moc sygnału zakłócanego będzie wyższa lub co najmniej równa mocy sygnału od telefonu komórkowego odbieranego przez staję bazową. Ponieważ zakłócany jest sygnał uplink nie ma potrzeby stosowania nadajnika zakłóceń z mocą wyższą niż 100mW. Dla zakłócania sygnałów downlink wymagana moc jest znacznie większa.



Rys. 3. Graf przepływu dla radia SDR realizujący scenariusz testowy

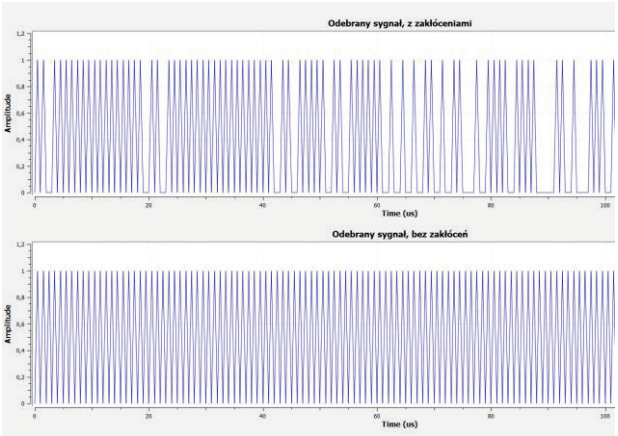
5. WYNIKI BADAŃ

Wykorzystując środowisko GNURadio oraz radio programowalne B200mini udało się odtworzyć scenariusz odbierania sygnału telefonii komórkowej przez stację bazową. Tor nadawczy radia B200mini posłużył jako źródło sygnału uplink (emitowanego z telefonu komórkowego do stacji bazowej). Tor odbiorczy radia B200 mini posłużył jako symulacja stacji odbiorczej w dwóch scenariuszach:

odbiór sygnału przez stację bazową po wprowadzeniu zakłóceń,

odbiór sygnału przez stację bazową bez wprowadzenia zakłóceń.

Porównanie obu scenariuszy zostało zobrazowane na Rys. 4.



Rys.4 Porównanie odbioru sygnału przez stację bazową po wprowadzeniu zakłóceń i bez zakłóceń

Na podstawie przeprowadzonych badań można zaobserwować, że wprowadzenie sygnału zakłócającego do stacji bazowej w znacznym stopniu zmienia kształt charakterystyki amplitudy sygnału w funkcji czasu. Nie wymaga to wygenerowania sygnału dużej mocy i zakres mocy radia B200 mini (do 100 mW) w zupełności wystarcza do zniekształcenia obieranego sygnału. Powoduje to brak otrzymywania poprawnych danych przez stację bazową z minimalizacją szansy na wykrycie źródła emisji zakłóceń.

WNIOSKI

W ramach zaproponowanego rozwiązania przedstawiono autonomiczny BSP do rozpoznania łączności sieci komórkowej i zakłócania stacji bazowych. Do niepodważalnych zalet opracowanego systemu należy zaliczyć:

mobilność,

zdolność do umieszczenia blisko źródła, które ma być zakłócone,

skalowalność rozumiana jako zdolność do rozbudowy lub też adaptacji w zależności od potrzeb operacyjnych,

łatwość portowania ze względu modułowość można przenosić poszczególne komponenty na różne platformy nośne,

możliwość uaktualnienia lub dodania nowych funkcjonalności ze względu wykorzystanie platformy radia programowalnego,

trudność detekcji emisji zakłócającej ze względu możliwą minimalną moc promieniowania.

W ramach dalszych prac planowane są badania poligonowe zbudowanego systemu z wykorzystaniem działającego komercyjnego sprzętu telefonii komórkowej systemów GSM, 3G, 4G i 5G.

LITERATURA

[1] Arslan H, Yucek T. 2009. „A survey of spectrum sensing algorithms for cognitive radio applications". *IEEE Communications Surveys & Tutorials*, 11 (1) : 116–130.

[2] Chęciński R., Kaszuba-Chęcińska A, Kryk M., Łopatka J., Malon K., Skokowski P., Performance Analysis of the Cognitive Radio Network with Opportunistic Spectrum Access, International Journal of Electronics and Telecommunications (IJET) vol. 64 no 1, 2018.

[3] ITU-R Report SM.2152, "Definitions of Software Defined Radio (SDR) and Cognitive Radio System (CRS)", SM Series, Spectrum management, 2009.

[4] Jakubiak A., „Probabilistyczne metody detekcji sygnałów na tle zakłóceń", Oficyna Wydawnicza Politechniki Warszawskiej, 2013.

[5] Jameel, S. Wyne, G. Kaddoum, and T. Q. Duong, "A comprehensive survey on cooperative relaying and jamming strategies for physical layer security," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 3, pp. 2734–2771, 2019, doi: 10.1109/COMST.2018.2865607.

[6] Kandeepan, S., Giorgetti A., Chiani M., "Periodic Spectrum Sensing Performance and Requirements for Detecting Legacy Users with Temporal and Noise Statistics in Cognitive Radios," P-roceedings of IEEE Globecom/Workshop, Hawaii, 2009.

[7] Kostylev, V. I., Energy Detection of a Signal with Random Amplitude", Proc. IEEE Int. Conf. On Commun. (ICC02), New York, ss. 1606–1610, 2002.

[8] Malon K., Skokowski P., Kelner J. M., Łopatka J., Koncepcja nadzorcy w symulatorze radia kognitywnego CORASMA, Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, ISSN: 1230-3496, 2014.

[9] R. A. Poisel, Modern communications jamming. Principles and techniques, 2nd ed. Boston, MA, USA: Artech House Publishers, 2011.

[10] Skokowski P., 2021. „Budowanie świadomości sytuacji elektromagnetycznej w sieciach doraźnych z węzłami kognitywnymi". Warszawa: Wojskowa Akademia Techniczna.

[11] Skokowski P., Łopatka J., *Metody wykrywania i analizy sygnałów DSSS*, VII Konferencja Naukowo Techniczna Systemy Rozpoznania i Walki Elektronicznej KNTWE 2008, ISBN: 978-83-61486-04-6.

[12] Skokowski P., Łopatka, "Efficient Method of 3G Signal Detection", PIERS Online, Vol. 7, No. 1, ss. 51-54, 2011.

[13] Skokowski P., Malon K., Łopatka J., "Properties of centralized cooperative sensing in cognitive radio networks," Proc. SPIE 10418, XI Conference on Reconnaissance and Electronic Warfare Systems, 1041807 (20 April 2017); https://doi.org/10.1117/12.2269996

[14] Sun H., Nallanathan A., Wang C., Chen Y., "Wideband spectrum sensing for cognitive radio networks: A survey", IEEE Wireless Communications, vol. 20, ss. 74–81, 2013.

[15] Y. Arjoune and S. Faruque, "Smart jamming attacks in 5G New Radio: A review," in 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, Jan. 2020, pp. 1010–1015. doi: 10.1109/CCWC47524.2020.9031175.

[16] Zeng, Y., and Y. C. Liang, "Spectrum-Sensing Algorithms for Cognitive Radio Based on Statistical Covariances," IEEE Transactions On Vehicular Technology, Vol. 58, ss. 1804–1815, No. 4, 2009.

PARAMETRYZACJA GEOMETRII SCENY DLA ESTYMACJI GŁĘBI W DEKODERZE WIZJI WSZECHOGARNIAJĄCEJ

PARAMETRIZATION OF SCENE GEOMETRY FOR DECODER SIDE DEPTH ESTIMATION IN IMMERSIVE VIDEO

Błażej Szydełko, Dominika Klóska, Adrian Dziembowski
Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
szydelkob@hotmail.com, dominikakkloska@gmail.com, adrian.dziembowski@put.poznan.pl

DOI: 10.15199/59.2022.4.64

Streszczenie: **Technika MPEG Immersive Video (MIV) jest rozwijana w celu efektywnego dostarczania wizji w systemie swobodnej nawigacji. W technice tej przewidziano brak przesyłania informacji o geometrii sceny, a rekonstruowanie jej po stronie dekodera w czasochłonnym procesie estymacji map głębi. Sposobem usprawnienia procesu estymacji jest wykorzystanie dodatkowej informacji o geometrii sceny przesyłanej jako sparametryzowane cechy map głębi. W artykule skupiono się na analizie skuteczności istniejącego rozwiązania, rozszerzając je o rekurencyjną ekstrakcję cech.**

Abstract: **The MPEG Immersive Video (MIV) technique is being developed for efficient delivery of immersive video. In one scheme, instead of transmitting geometry information, the technique aims to reconstruct it at the decoder side in a time-consuming depth map estimation process. The way to improve the estimation process is to use additional information about the scene geometry transmitted as parametrized features of the depth maps. This paper focuses on the efficiency analysis of an existing solution, extending it with recursive feature extraction.**

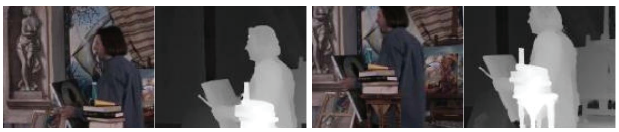
Słowa kluczowe: **estymacja głębi, wizja wszechogarniająca, rzeczywistość wirtualna, system swobodnej nawigacji**

Keywords: **depth estimation, immersive video, virtual reality, free navigation system**

1. WSTĘP

Ostatnimi czasy, wizja wszechogarniająca (immersyjna) stała się znaczącym kierunkiem rozwoju cyfrowych systemów multimedialnych [20]. W odróżnieniu od klasycznej wizji, gdzie widz zależny jest od woli reżysera bądź producenta, wizja immersyjna umożliwia zmianę punktu widzenia lub nawet ruch w obrębie sceny trójwymiarowej [17]. Doświadczenie immersji nie byłoby możliwe bez rozpowszechnienia się wyświetlaczy nagłownych (ang. Head-Mounted Display – HMD), które są w stanie w pewnym stopniu przekonać zmysły widza o przebywaniu w przedstawionej treści. Oprócz urządzeń HMD, oglądanie treści immersyjnych umożliwiają również standardowe wyświetlacze (w tym urządzenia mobilne) przy pomocy dowolnego urządzenia sterującego, takiego jak mysz komputerowa, joystick czy ekran dotykowy [16].

Zapewnienie użytkownikom swobody ruchu wymaga rejestracji sceny z kilku perspektyw oraz informacji o jej geometrii, by w procesie syntezy generować widoki pośrednie [4], [3]. Obecnie do reprezentacji sceny wykorzystuje się sekwencje wielowidokowe w formacie MVD (ang. Multiview Video + Depth) [15], składające się na zarejestrowane widoki i odpowiadające im mapy głębi (Rys. 1), które powstają w procesie estymacji lub są brane wprost z narzędzi do modelowania 3D, np. Blender (dla scen generowanych komputerowo).



Rys. 1. Sekwencja wielowidokowa (1 ramka, 2 widoki)

Ilość danych generowanych przez system immersyjny stawia wyzwanie wobec istniejących metod strumieniowania wizji. Zastosowanie standardowego kodera wizyjnego do niezależnego kodowania każdego z widoków oraz map głębi byłoby bardzo nieefektywne i niepraktyczne. Te niedogodności oraz rozwój technologii zapewniających immersję zintensyfikowały wysiłki standaryzacyjne w tej dziedzinie. Obecnie rozwijany standard MPEG Immersive Video (MIV) [1] ma na celu znormalizowanie efektywnego kodowania sekwencji wielowidokowych. Podczas gdy jego główny profil koncentruje się na usuwaniu redundancji między widokami i upakowaniu pozostałych danych [1], profil „Geometry Absent" (MIV GA) realizuje ideę estymacji geometrii sceny po stronie dekodera (ang. decoder-side depth estimation – DSDE) [5], [11]. W MIV GA strumień bitowy zawiera wyłącznie zestaw widoków wejściowych i ich parametrów. Pomimo rezygnacji z przesyłania informacji o geometrii sceny, jest ona wciąż dostępna po stronie kodera. Umożliwia to koderowi ekstrakcję zestawu parametrów i przesłanie ich w celu ułatwienia estymacji głębi po stronie dekodera [6], [5].

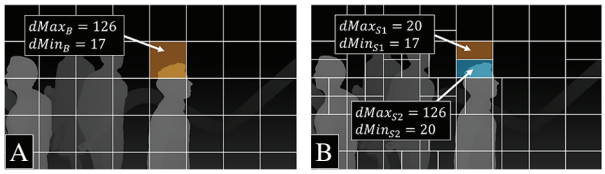
W niniejszym artykule przedstawiono propozycję rozszerzenia metody wspieranej w standardzie MIV o rekurencyjny algorytm ekstrakcji parametrów [18] wraz z efektywnym sposobem kompresji parametrów map głębi opartej na algorytmie kodowania kontekstowego CABAC (ang. Context-based Adaptive Binary Arithmetic Coding) [10]. Propozycja ta pozwala na lepsze dopasowanie parametrów do struktury map głębi, a tym samym na poprawę jakości i zmniejszenie czasu obliczeniowego estymacji głębi.

2. PARAMETRYZACJA GEOMETRII

Wykorzystanie parametrów pochodzących z map głębi, które są dostępne w koderze pomaga rozwiązać

podstawowe problemy związane z estymacją głębi po stronie dekodera. Po pierwsze, taka estymacja ma duży wpływ na złożoność procesu dekodowania, gdyż stanowi jego integralną część. Nawet najszybsze metody są kosztowne obliczeniowo i nie są w stanie zapewnić wysokiej jakości map głębi w czasie rzeczywistym [9]. Ponadto, jako że mapy głębi są estymowane przy użyciu skompresowanych widoków, zwykle obserwuje się pewne pogorszenie ich jakości [13]. Tę degradację spowodowaną kompresją można zmniejszyć, stosując metody redukcji artefaktów kompresji [7] lub poprawy jakości map głębi [12]. Jednakże zastosowanie dodatkowych etapów w przetwarzaniu końcowym wydłuża czas od momentu pozyskania strumienia bitów do zaprezentowania widoku wirtualnego końcowemu odbiorcy.

Podejście estymacji głębi zaprezentowane w [5] zakłada uwzględnienie dwóch kluczowych typów parametrów w zakodowanym strumieniu bitów. Parametry te są ekstrahowane bezpośrednio z map głębi i organizowane w postaci siatki bloków. Pierwszym typem jest zakres głębi w każdym bloku wejściowych map głębi (Rys. 2A). Przesłanie tej informacji zwiększa jakość estymowanej głębi poprzez redukcję możliwych błędów estymacji (które mogą być efektem kompresji widoków).



Rys. 2. Podstawowa idea ekstrakcji zakresu głębi; (A) siatka bloków głównych; (B) idea podziału bloków: blok B został podzielony na podbloki $S1$ i $S2$

Drugim typem parametru jest flaga pomijania estymacji, która informuje estymator, że dany blok głębi nieznacznie różni się od bloku w poprzedniej ramce i w związku z tym można skopiować blok z poprzedniej ramki. Należy nadmienić, że rozmiar bloków nie jest stały i istnieje możliwość podziału ich na cztery kwadratowe podbloki [5] lub na dwa prostokątne, zaproponowane przez autorów niniejszego artykułu w [19]. Wszystkie możliwe typy podziału bloków przedstawiono w Tab. 1.

Tab. 1. Możliwe typy bloków i odpowiadające im słowa kodowe [8].Oznaczenia typów: Q – czwórkowy (ang. quad), R – prostokątny (ang. rectangular)

Flaga \ Typ		Q	R					
split	0	1	1	1	1	1	1	1
split_quad	-	1	0	0	0	0	0	0
split_orientation	-	-	0	0	0	1	1	1
split_symmetry	-	-	1	0	0	1	0	0
split_first_bigger	-	-	-	0	1	-	0	1

3. REKURENCYJNY PODZIAŁ BLOKÓW

3.1. Ekstrakcja parametrów

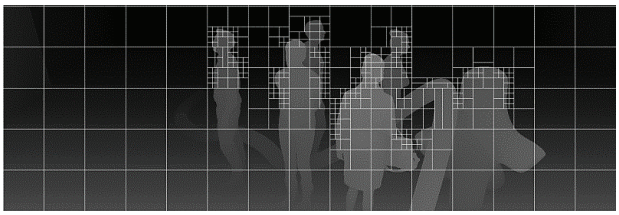
Obecne rozwiązania [5], [2] pozwalają podzielić blok tylko jeden raz (Rys. 2B). Z tego powodu nie ma możliwości dopasowania siatki bloków do bardziej skomplikowanej struktury mapy głębi. Aby temu zaradzić, proponujemy rozwiązanie rekurencyjne gdzie każdy blok może być wielokrotnie podzielony w celu dopasowania struktury bloków do struktury krawędzi mapy głębi.

Argument stojący za dokładniejszą parametryzacją geometrii wynika z podejścia stosowanego w estymacji głębi. Estymator nie sprawdza wszystkich możliwych wartości głębi (rozbieżności), a jedynie ogranicza się do zadanego zakresu ($dMin$, $dMax$). W związku z tym, jeżeli w bloku zawarte są punkty mapy głębi reprezentujące pojedynczy obiekt (np. fragment ściany), to zakres badanej rozbieżności jest bardzo wąski, co skutkuje redukcją czasu estymacji głębi oraz mniejszym prawdopodobieństwem wyznaczenia nieprawidłowej wartości głębi. W najkorzystniejszym przypadku, jeśli $dMin$ i $dMax$ dla bloku są równe, nie ma potrzeby estymacji głębi dla tego bloku.

Przed rozpoczęciem podziału rekurencyjnego, cała ramka głębi jest wstępnie dzielona na siatkę bloków o zadanym rozmiarze (Rys. 2A). Dla dalszego przetwarzania wymagane jest znalezienie zakresu głębi w każdym bloku. Kiedy różnica $dMax - dMin$ przekracza zadany próg podziału (ang. split threshold, domyślnie równy 2562 dla 16 bitowych map głębi), blok może zostać podzielony. Pozwala to na redukcję sumarycznej objętości bloku a tym samym rozłożenie procesu estymacji na mniejsze zakresy głębi. Jak wspomniano, obecne rozwiązanie oferuje 7 typów podziału bloku (Tab. 1). Wymagane jest więc wybranie podziału, który najefektywniej redukuje objętość bloku, a więc minimalizuje koszt (ang. cost volume – CV) dany wzorem:

$$CV_B = \sum_{S \in B} \left((dMax_S - dMin_S) + 1 \right) \cdot w_S \cdot h_S \,, \quad (1)$$

gdzie w_S i h_S są odpowiednio szerokością i wysokością podbloku S w bloku B. Opisany proces podziału jest powtarzany rekurencyjnie dla bloków utworzonych w poprzednich podziałach. Kiedy blok nie spełnia warunku do podzielenia lub jeden z wymiarów podbloku staje się mniejszy od ustalonego minimalnego rozmiaru, rekurencja kończy się. Rezultat podziału rekurencyjnego przedstawia Rys. 3.



Rys. 3. Mapa głębi podzielona rekurencyjnie

Ostatnim krokiem jest wykorzystanie informacji o spójności czasowej z map głębi. W tym celu odpowiadające sobie bloki są porównywane z blokami z poprzedniej ramki. Czasowe zróżnicowanie bloków jest wyrażone wzorem:

$$Z = \frac{\sum_{i=0}^{w_B-1} \sum_{j=0}^{h_B-1} \left| B_f(i,j) - B_{f-1}(i,j) \right|}{w_B \cdot h_B} \,, \quad (2)$$

gdzie f jest numerem ramki. Oczywiście zróżnicowanie Z nie jest liczone dla pierwszej ramki sekwencji.

Jeśli zróżnicowanie Z jest mniejsze od ustalonego progu pomijania (ang. skip threshold, domyślnie równy 2%), wysyłana jest flaga pominięcia estymacji głębi dla tego bloku. W celu zmniejszenia rozmiaru metadanych, flaga pominięcia wysyłana jest dla bloku nadrzędnego,

którego wszystkie podbloki są pomijane. Dodatkowo wartości $dMax$ i $dMin$ podlegają kwantyzacji.

3.2. Kompresja metadanych

Na metadane składają się wartości graniczne głębi oraz flagi sygnalizujące pomijanie estymacji i informujące o sposobie podziału bloku (Tab. 1). Jak można zauważyć, najprostszy rodzaj bloku jest sygnalizowany przez jeden bit o wartości równej zero, natomiast bloki pozwalające lepiej dostosować parametry do bardziej skomplikowanych struktur map głębi wymagają nawet do pięciu bitów sygnalizacyjnych. Podobieństwo zakresów głębi w sąsiednich blokach oraz zdefiniowane słowa kodowe stwarzają dogodne warunki do zakodowania metadanych przy użyciu kodowania kontekstowego CABAC [10], co w konsekwencji znacznie zmniejszy ich rozmiar. Dodanie nowych typów bloków oraz propozycja podziału rekurencyjnego wymusza aktualizację kodera metadanych o obsługę nowych słów kodowych i kontekstów.

Parametry geometrii są zorganizowane w formie drzewa czwórkowego, gdzie główny blok pełni funkcję korzenia, a ostatnie bloki, zawierające wartości $dMin$ i $dMax$ lub flagę pominięcia estymacji, stanowią liście. Każda z flag ma przypisany zbiór kontekstów reprezentujących prawdopodobieństwa wystąpienia symboli. W przypadku kodowania flagi pominięcia estymacji, kontekst jest wybierany na podstawie sumy wartości tychże flag z sąsiednich bloków (lewy i górny) i bloku z poprzedniej ramki, które są liśćmi drzewa (flaga pomięcia występuje tylko tam). Natomiast dla flag informujących o sposobie podziału bloku (a więc o strukturze drzewa) kontekst zależy tylko od wartości flag w bloku z poprzedniej ramki (znajdowany jest pierwszy blok o rozmiarze mniejszym lub równym kodowanemu). Na koniec, wartości $dMin$ i $dMax$ są kodowane w sposób predykcyjny, wykorzystując już zakodowane wartości w sąsiednim lewym i/lub górnym bloku.

4. BADANIA EKSPERYMENTALNE

4.1. Metodyka badań

Eksperymenty przeprowadzono według ustalonych warunków testowych (ang. Common Test Conditions) standardu MIV [21] przy wykorzystaniu potoku przetwarzania zdefiniowanego dla profilu MIV GA. Zbiór testowy zawierał 15 sekwencji w formacie MVD (sekwencja liczy 17 ramek). Z dostępnych w koderze map głębi wyekstrahowano parametry w 12 konfiguracjach, które następnie wraz z widokami przesłano do dekodera. Zdekodowane widoki i parametry geometrii wykorzystano do estymacji głębi metodą IVDE – Immersive Video Depth Estimation [14], która jest powszechnie wykorzystywana w eksperymentach związanych z MIV. Odtworzone mapy głębi i widoki wykorzystano do syntezy widoków wirtualnych, których jakość została oceniona miarą PSNR w odniesieniu do widoków oryginalnych.

W badaniach wykorzystano parametry dla 12 konfiguracji podziału bloków reprezentowanych w formacie $T - MAX - MIN$, gdzie MAX i MIN oznaczają odpowiednio maksymalny i minimalny rozmiar bloku, a T to jeden z dwóch rozważanych typów podziału bloków: Q przy podziale czwórkowym, a A przy wykorzystaniu wszystkich dostępnych podziałów ($Q + R$, R – podział prostokątny) (Tab. 1). Na przykład, $A - 64 - 8$ oznacza parametryzację z wykorzystaniem wszystkich typów podziału bloków, rozpoczynając od bloku o rozmiarze 64×64, a na 8×8 kończąc.

Dwa główne cele estymacji głębi po stronie dekodera z wykorzystaniem parametrów obejmują zwiększenie jakości syntezowanych widoków i skrócenie czasu dekodowania (w tym czasu potrzebnego na estymację głębi). W związku z powyższym, do oceny testowanych konfiguracji wzięto pod uwagę jakość syntezowanych widoków, prędkość bitową i czas estymacji głębi. Zaprezentowane w tabelach wyniki są średnimi wartościami wyznaczonymi po wszystkich zdekodowanych widokach testowanych sekwencji w pięciu różnych przepływnościach.

4.2. Wyniki

Analizując Tab. 2, można zauważyć, że mniejsze (lepiej dopasowane) bloki pozwalają na skrócenie czasu estymacji głębi. Jednakże, w przypadku użycia bardzo małych prostokątnych bloków czas estymacji może nieznacznie wzrosnąć, co ma związek z częstotliwością występowania flagi pomijania estymacji, której sygnalizacja następuje tylko dla odpowiadających sobie bloków w kolejnych ramkach. W przypadku podziału czwórkowego, pomimo niestabilności czasowej głębi, struktura bloków w sąsiednich ramkach może pozostawać niezmieniona, a w związku z tym, prawdopodobieństwo sygnalizacji pominięcia estymacji wzrasta przyczyniając się do skrócenia czasu estymacji map głębi.

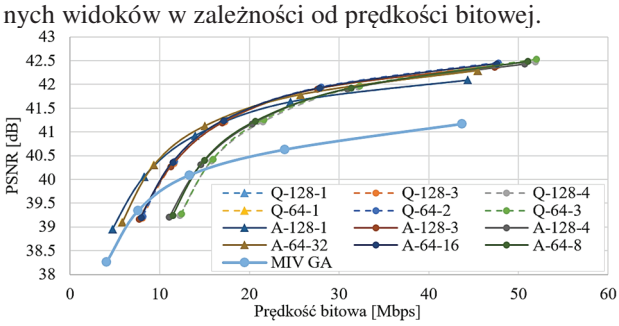
Tab. 2. Średni czas estymacji głębi w dekoderze [s]

Rozmiar bloku		**Typ podziału bloku *T***	
MAX	*MIN*	*Q*	*A(Q + R)*
128	64[1]	4471,1	3731,3
	16	3939,5	3222,2
	8	3219,2	3885,9
64	32[1]	3880,4	3429,4
	16	3648,9	3873,7
	8	3558,7	3265,1
MIV GA[1]		12417,2	

[1] rozwiązanie wspierane w standardzie MIV [8]

W ostatnim wierszu Tab. 2 zaprezentowano wyniki uzyskane dla profilu MIV GA bez użycia parametryzacji głębi. Jak pokazano, przesłanie dodatkowych parametrów umożliwia znaczące przyspieszenie estymacji głębi, a więc całego procesu dekodowania.

Na Rys. 4 przedstawiono uśrednione krzywe RD uzyskane dla wszystkich testowanych konfiguracji kodeka MIV, które przedstawiają średnią jakość syntezowanych widoków w zależności od prędkości bitowej.



Rys. 4. Krzywe RD testowanych konfiguracji

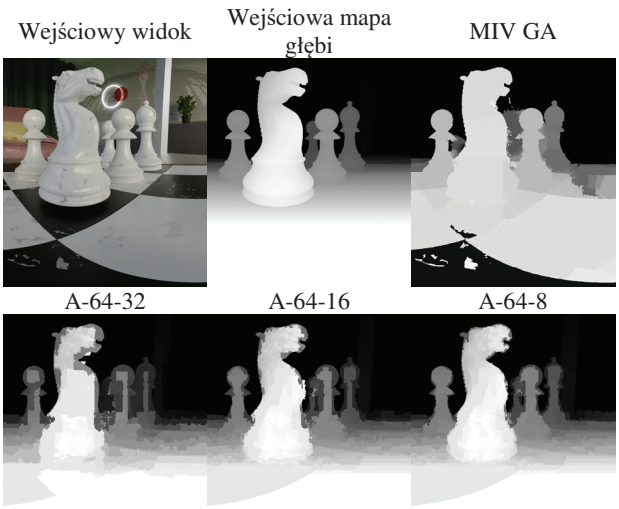
Wykorzystanie dowolnej konfiguracji do estymacji głębi znacznie poprawia jakość syntezowanych widoków w stosunku do MIV GA. Dokładniejsza parametryzacja powoduje wzrost prędkości bitowej metadanych (Tab. 3), znacznie przewyższając wartość dla MIV GA, jednakże zysk jakości oraz skrócenie czasu estymacji pozytywnie przeważają na korzyść prezentowanej metody.

Tab. 3. Średnie prędkości bitowe zakodowanych metadanych z parametrami [Mbps], wraz z ich procentowym udziałem w całym strumieniu

Rozmiar bloku		**Typ podziału bloku *T***	
MAX	*MIN*	Q	$A(Q+R)$
128	64[1]	0,656 (1,5%)	0,674 (1,5%)
	16	3,971 (8,3%)	3,651 (7,7%)
	8	8,171 (15,8%)	6,978 (13,7%)
64	32[1]	1,735 (3,8%)	1,775 (3,9%)
	16	4,040 (8,5%)	3,863 (8,1%)
	8	8,331 (16%)	7,375 (14,4%)
MIV GA[1]		0,0085 (0,02%)	

[1] rozwiązanie wspierane w standardzie MIV [8]

Jak można zauważyć na Rys. 5, wraz z dokładniejszą parametryzacją, szczegółowość estymowanej mapy głębi zwiększa się, a w związku z tym precyzyjniej odzwierciedla detale wejściowej mapy głębi i kształt obiektów w scenie. Ponadto, wykorzystanie parametrów pozwala wierniej odtworzyć wejściową mapę w stosunku do podejścia MIV GA, zwłaszcza w przestrzeniach pomiędzy obiektami i na płaskich powierzchniach.



Rys. 5. Fragment map głębi dla porównywanych konfiguracji i odpowiadający im widok wejściowy

5. PODSUMOWANIE

Zaprezentowana w niniejszym artykule metoda dokładniejszej parametryzacji geometrii sceny znacznie zwiększa wydajność estymacji głębi po stronie dekodera. Przeprowadzone eksperymenty pokazały, że pomimo zwiększenia ilości metadanych przesyłanych do dekodera, lepsze dopasowanie parametrów do struktury mapy głębi skutkuje zwiększeniem jakości wyznaczanej w dekoderze geometrii przy jednoczesnym skróceniu czasu estymacji. W związku z tym metoda ta idealnie nadaje się do zastosowań praktycznych, ponieważ dzięki niej można uzyskać wizję wszechogarniającą o lepszej jakości w porównaniu z metodą odniesienia.

PODZIĘKOWANIA

Praca finansowana ze środków przyznanych przez Ministerstwo Edukacji i Nauki.

LITERATURA

[1] Boyce J. M. *i in.* 2021. „MPEG Immersive Video Coding Standard". *Proc. IEEE,* 109 (9) : 1521–1536.
[2] Clare G. *i in*. 2021. „[MIV] Combination of m56626 and m56335 for Geometry Assistance SEI message". ISO/IEC JTC1/SC29/WG4 MPEG2021/M56950.
[3] Dziembowski A. *i in.* 2019. „Virtual View Synthesis for 3DoF+ Video". *Picture Coding Symposium*, 1–5.
[4] Fachada S. *i in*. 2018. „Depth image based view synthesis with multiple reference views for virtual reality". *3DTV-Conference.*
[5] Garus P. *i in*. 2021. „Immersive Video Coding: Should Geometry Information be Transmitted as Depth Maps?". *IEEE T. on Circ. & Sys. for Vid. Technology.*
[6] Garus P. *i in*. 2019. „Bypassing Depth Maps Transmission For Immersive Video Coding". *PCS Conference*.
[7] He X. *i in*. 2020. „MV-GNN: Multi-View Graph Neural Network for Compression Artifacts Reduction". *IEEE Trans. on Image Processing*, 29 : 6829–6840.
[8] ISO/IEC 23090-12. „Information technology - Coded representation of immersive media - Part 12: MPEG Immersive video". ISO/IEC JTC 1/SC 29.
[9] Laga H. *i in*. 2020. „A Survey on Deep Learning Techniques for Stereo-based Depth Estimation", *IEEE Trans. on Pat. An. & Mach. Intell.* 44 (4) : 1738-1764.
[10] Marpe D. *i in*. 2003. „Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard". *IEEE Transactions on Circuits and Systems for Video Technology,* 13 (7) : 620–636.
[11] Mieloch D. *i in*. 2022. „Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video''. *IEEE T. on Circ. & Sys. for Vid. Technology.*
[12] Mieloch D. *i in*. 2021. „Depth Map Refinement for Immersive Video". *IEEE Access,* 9 : 10778–10788.
[13] Mieloch D. *i in*. 2021. „Point-to-Block Matching in Depth Estimation". *WSCG Conference*.
[14] Mieloch D. *i in*. 2020. „Depth Map Estimation for Free-Viewpoint Television and Virtual Navigation". *IEEE Access,* 8: 5760–5776.
[15] Mueller K. *i in*. 2011. „3-D Video Representation Using Depth Maps". *Proc. of the IEEE,* 99 : 643–656.
[16] Stankiewicz O. *i in.* 2018. „A Free-Viewpoint Television System for Horizontal Virtual Navigation". *IEEE Transactions on Multimedia,* 20 (8) : 2182–2195.
[17] Stankiewicz O. *i in.* 2018. „Nowe media immersyjne". *Przegląd Telek. + Wiadomości Telekomunikacyjne,* 6.
[18] Szydełko B. *i in.* 2022. „Recursive block splitting in feature-driven decoder-side depth estimation". *ETRI Journal,* 44.
[19] Szydełko B. *i in.* 2021. „Rectangular blocks in encoder-derived features for decoder-side depth estimation". ISO/IEC JTC1/SC29/WG4 MPEG2021/M5635.
[20] Wien M. *i in.* 2019. „Standardization Status of Immersive Video Coding". *IEEE Journal on Emerging & Selected Topics in Circuits & Systems,* 9 (1).
[21] „Common Test Conditions for MPEG Immersive Video".ISO/IEC JTC1/SC29/WG4 MPEG2021/N005.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

WYKORZYSTANIE WEJŚCIOWYCH MAP GŁĘBI W ESTYMACJI GŁĘBI PO STRONIE DEKODERA

THE USE OF INPUT DEPTH MAPS IN DECODER-SIDE DEPTH ESTIMATION

Dominika Klóska, Błażej Szydełko, Dawid Mieloch
Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
dominikakkloska@gmail.com, szydelkob@hotmail.com, dawid.mieloch@put.poznan.pl

DOI: 10.15199/59.2022.4.65

Streszczenie: **Artykuł opisuje usprawnienia procesu estymacji map głębi po stronie dekodera wizji wszechogarniającej. W prezentowanej propozycji mapy głębi są przesyłane tylko dla części widoków, umożliwiając znaczące zmniejszenie czasu wymaganego na estymację geometrii przy jednoczesnym zwiększeniu jej jakości. Rozwiązanie obejmuje modyfikację oprogramowania ISO/IEC MPEG służącego do wyznaczania geometrii sceny oraz implementacji kodera MPEG Immersive Video. Wyniki testów obejmują porównanie czasu dekodowania i uzyskiwanych prędkości bitowych.**

Abstract: **This paper presents a proposal to improve the depth estimation process performed in decoder-side depth estimation. In the proposed solution, depth maps are sent only for a part of the views, allowing a significant reduction in the time required for geometry estimation and a simultaneous increase of its quality. The solution includes the modification of the ISO/IEC MPEG Video Coding software for depth estimation and the implementation of the MPEG Immersive Video encoder. The test results consist of the comparison of decoding times and required bitrates.**

Słowa kluczowe: **estymacja po stronie dekodera, wizja wszechogarniająca, wejściowe mapy głębi**

Keywords: **decoder-side depth estimation, immersive vision, input depth maps.**

1. WSTĘP

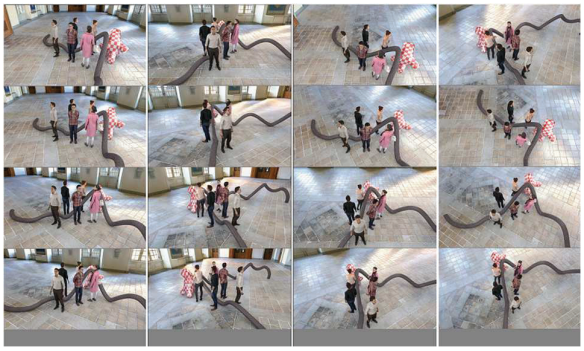
Obecnie istnieje wiele systemów multimedialnych, które dają użytkownikom dostęp do trójwymiarowej wizji z dodatkową możliwością swobodnego nawigowania polem widzenia w wyświetlanej scenie [3], [9]. Wizja generowana przez te systemy ma charakteryzować się wysokim poziomem realizmu, co wiąże się z koniecznością posiadania szczegółowej informacji na temat geometrii sceny. Aby uzyskać tego typu informację należy m.in. wykorzystać wysokiej jakości widoki wejściowe w procesie estymacji map głębi [4]. Sekwencje potrzebne do stworzenia wizji wszechogarniającej są rejestrowane przez system wielokamerowy, w którym zarówno rodzaj kamer (np. dookólne) oraz ich ułożenie (np. macierz, łuk) mogą być dowolne. Istnieje zależność między liczbą kamer użytych do rejestrowania sceny trójwymiarowej oraz jakością estymowanych map głębi – im więcej kamer użyto do rejestracji sceny tym lepszą jakość ma wyznaczona głębia [8].

Jak opisano powyżej, do stworzenia wizji wszechogarniającej wymagana jest duża ilość danych, czego następstwem są wysokie prędkości bitowe wymagane do zakodowania strumienia danych. Z tego powodu, aby umożliwić przesyłanie tych danych, należy zakodować widoki i odpowiadające im mapy głębi używając metod kodowania o dużej wydajności.

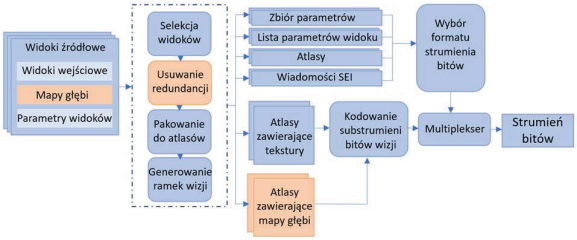
Do tej pory w celu kodowania sekwencji wielowidokowych wykorzystywano kodery MV-HEVC oraz 3D-HEVC [11], które jednak nie są w stanie do końca sprostać wymaganiom, jakie stawia przed nimi nowoczesna wizja wszechogarniająca. Nie są one dostosowane do kodowania sekwencji nagrywanych systemem wielokamerowym, w którym użyte kamery były dookólne lub ich ułożenie było nieliniowe.

Przedstawione powyżej problemy oraz wzrost popularności wizji wszechogarniającej wśród społeczności specjalizującej się w przetwarzaniu wizji i obrazu sprawiły, że zaczęły pojawiać się nowe propozycje standaryzacji procesu kodowania oraz dekodowania tego typu wizji [12]. Najnowszym ze standardów kodowania stworzonym dla takich celów jest MIV – MPEG Immersive Video [1]. Jest on formą przetwarzania wstępnego, w której wybierane są takie widoki lub fragmenty widoków, które są niezbędne do wygenerowania wizji wszechogarniającej. Te dane są następnie pakowane do atlasów, czyli obrazów zawierających jeden lub więcej widoków wejściowych lub fragmenty tych widoków. Po przetwarzaniu wstępnym, MIV wykorzystuje dowolny koder wizyjny (np. HEVC [10], VVC [2]) do zakodowania stworzonych atlasów. Główny profil standardu, MIV Main, pozwala na kompresję dokonywaną poprzez usunięcie międzywidokowej informacji nadmiarowej, ale przesyłane, zakodowane mapy głębi mogą zajmować nawet ponad 50% strumienia bitów [7].

W niniejszej pracy skupiono się na profilu MIV DSDE (ang. *decoder-side depth estimation*, czyli estymacja głębi po stronie dekodera), ponieważ pozwala on znacznie zredukować prędkość bitową w porównaniu do profilu głównego, w związku z tym, że mapy głębi nie są przesyłane w strumieniu. Atlasy profilu MIV DSDE zostały przedstawione na Rys. 1, gdzie każda kolumna przedstawia osobny atlas. Jak opisano wyżej, profil ten nie przesyła głębi do dekodera, a więc nie ma atlasów zawierających geometrie. Dzięki temu zwiększona zostaje wydajność typowych koderów wizyjnych, które nie zostały stworzone do kodowania geometrii. Schemat kodera MIV został przedstawiony na Rys. 2.



Rys. 1. Atlasy wygenerowane przez profil MIV DSDE. Każda kolumna widoków odpowiada jednemu atlasowi.



Rys. 2. Schemat kodera profilu MIV Main. Na pomarańczowo zaznaczono te elementy, których nie wykorzystuje koder profilu MIV DSDE.

Estymacja głębi w procesie dekodowania w MIV nie została ujęta w standardzie [6], a więc można ją realizować dowolną metodą. Immersive Video Depth Estimation (IVDE) [8] jest oprogramowaniem referencyjnym grupy ISO/IEC MPEG do wyznaczania geometrii sceny, które zostało dostosowane do standardu MIV. Dzięki temu odpowiada wymaganiom narzucanym przez MIV na proces estymacji głębi i dlatego to właśnie to oprogramowanie zostało wykorzystane na potrzeby niniejszego artykułu. Pomimo wspomnianych zalet, wciąż istnieją pewne aspekty procesu wyznaczania geometrii sceny, które wymagają szczególnej uwagi. Jednym z nich jest estymacja głębi powierzchni płaskich znajdujących się w scenie. Gradienty tych powierzchni często są nieciągłe w wyznaczonych mapach głębi (Rys. 3), co negatywnie wpływa na subiektywną jakość wizji wszechogarniającej.

Rys. 3. Przykładowe mapy głębi z problemem braku gradientów dla powierzchni gładkich dla sekwencji Classroom (A), Kitchen (B) oraz Group (C).

Ten artykuł opisuje propozycję rozwiązania problemu estymacji głębi powierzchni gładkich w oprogramowaniu IVDE oraz wykorzystanie wprowadzonej zmiany w zmodyfikowanej implementacji kodera zrealizowanej w oprogramowaniu TMIV (Test Model for MPEG Immersive Video [1]). W prezentowanej propozycji mapy głębi są przesyłane tylko dla części widoków, umożliwiając znaczące zmniejszenie czasu wymaganego na estymację głębi przy jednoczesnym zwiększeniu jej jakości. Wyniki badań obejmują porównanie prędkości bitowych wymaganych do zakodowania sceny oraz czasy dekodowania zmodyfikowanego oprogramowania TMIV z oprogramowaniem referencyjnym. Zaproponowane zmiany pozwoliły na znaczne skrócenie czasu dekodowania oraz zmniejszyły problem estymacji powierzchni gładkich nie zwiększając przy tym prędkości bitowych użytych na potrzeby przesyłania danych.

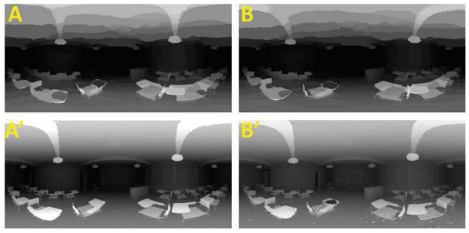
2. WEJŚCIOWE MAPY GŁĘBI

2.1. Modyfikacja oprogramowania IVDE

IVDE jest oprogramowaniem referencyjnym grupy ISO/IEC MPEG służącym do wyznaczania geometrii sceny na podstawie widoków zarejestrowanych przez system wielokamerowy. Głębia punktu może być wyznaczona jeżeli został on zarejestrowany przez dwie lub więcej kamer systemu. Dla każdego punktu z widoku wejściowego w sąsiednich widokach poszukiwany jest jego odpowiednik, czyli punkt o największym podobieństwie. Jak można zauważyć, proces poszukiwania odpowiadających sobie punktów tworzy ścisłą relację między mapami głębi sąsiadujących widoków.

Sekwencje wielowidokowe posiadają wysokiej jakości mapy głębi pozyskane wprost z programu graficznego, w którym zostały stworzone (w przypadku sekwencji wygenerowanych komputerowo), lub wyestymowane przy użyciu ręcznie dobranych parametrów. W proponowanym podejściu, zmodyfikowano oprogramowanie referencyjne tak, aby umożliwić wykorzystanie tych map głębi.

Wprowadzone zmiany zakładały, że danymi podawanymi na wejście oprogramowania IVDE będą nie tylko widoki wejściowe, ale również odpowiadające im wysokiej jakości mapy głębi. Wartości próbek mapy głębi podanej na wejście IVDE są kopiowane do odpowiadającej jej mapy wyjściowej. Ze względu na to, iż w procesie estymacji wszystkie mapy głębi są ze sobą powiązane, przesłanie jednej lub kilku wejściowych map głębi wpływa pozytywnie na wszystkie głębie wyjściowe. Dodatkowo, proces estymacji głębi w IVDE jest oparty na małych segmentach, a nie na punktowej estymacji [8], więc umożliwia poprawę błędów wynikających z kompresji. Z tego powodu wejściowe mapy głębi mogą być silnie skompresowane.



Rys. 4. Mapy głębi dla sekwencji Classroom Video: (A) i (B) – metoda referencyjna, (A') i (B') – modyfikacja.

Na Rys. 4 pokazano jak wprowadzone zmiany wpłynęły na obecność gradientów powierzchni gładkich w wyznaczonej geometrii sceny (na wejście została podana mapa głębi odpowiadająca widokowi 0).

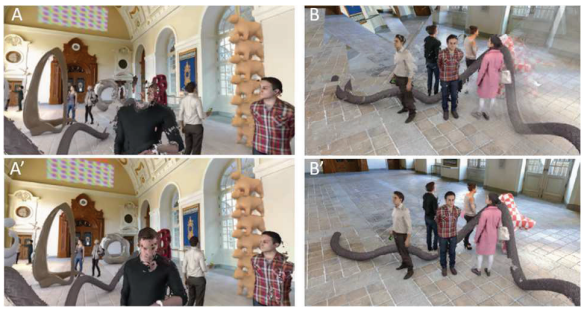
2.2. Modyfikacja oprogramowania TMIV

Wizja wszechogarniająca w dużej mierze składa się z widoków wirtualnych prezentujących te części sceny, które nie zostały zarejestrowane przez żadną z kamer systemu wielokamerowego. Są one syntezowane na podstawie wyznaczonej głębi, a więc istnieje ścisła zależność

między jakością wizji wszechogarniającej, a jakością wyznaczonej geometrii. Z tego powodu problem opisany w rozdziale 1 wpływa negatywnie na jakość wizji wszechogarniającej. Zważywszy na to, iż modyfikacja opisana w podrozdziale 2.1 poprawiła jakość głębi, wprowadzono kilka zmian w oprogramowaniu TMIV by móc ją wykorzystać.

Jako iż modyfikacja dotyczy profilu MIV DSDE, wykorzystywane są w niej te same widoki wejściowe, które są pakowane do atlasów w MIV DSDE. Różnica polega na kolejności, w jakiej widoki te są układane w atlasach metody zmodyfikowanej. W pierwszym atlasie w metodzie zmodyfikowanej przesyłane są widoki zawierające jak najwięcej informacji o scenie (np. w przypadku gdy kamery były ustawione na macierzy byłyby to widoki usytuowane najdalej od siebie, mające najmniej informacji wspólnej). W drugim atlasie przesyłane są wysokiej jakości głębie wejściowe odpowiadające widokom z atlasu pierwszego, które zostaną wykorzystane przez zmodyfikowane oprogramowanie IVDE po stronie dekodera. Geometria wyestymowana na podstawie wejściowych map głębi przesłanych w drugim atlasie będzie miała lepszą jakość (co pokazano w podrozdziale 2.1). W dwóch pozostałych atlasach przesyłane są widoki wejściowe, które pozostały po dokonaniu selekcji do atlasu pierwszego. Zważywszy na dodanie dodatkowego atlasu map głębi, całkowita liczba przesłanych widoków zostaje zmniejszona o liczbę widoków mieszczących się w jednym atlasie. Rys. 5 przedstawia jak wyglądają atlasy po wprowadzonych zmianach. Na Rys. 6 zaprezentowano jak wprowadzone zmiany wpłynęły na jakość zsyntezowanego w dekoderze widoku wirtualnego.

Rys. 5. Atlasy MIV DSDE po wprowadzonej modyfikacji. Każda kolumna widoków odpowiada jednemu atlasowi.



Rys. 6. Widoki wirtualne stworzone przy pomocy oprogramowania referencyjnego MIV DSDE: (A) i (B), oraz przy wprowadzonej modyfikacji: (A') i (B').

3. EKSPERYMENTY

3.1. Metodyka

Badania opisane w tym rozdziale zostały przedstawione w odniesieniu do wyników uzyskanych za pomocą MIV DSDE. Atlasy zawierające widoki wejściowe oraz widoki wejściowe z mapami głębi (w przypadku wprowadzonej modyfikacji) zostały zakodowane za pomocą jednej z wydajnych implementacji kodera VVC – Vvenc. Dane te zostały przesłane do dekodera, a następnie odzyskane ze zdekodowanych strumieni bitów VVC. Następnie, zostały one wykorzystane do estymacji głębi przy użyciu oprogramowania IVDE. Testy zostały przeprowadzone na 15 różnych sekwencjach testowych, przy czym 9 tych sekwencji było wygenerowane komputerowo, a 6 z nich zawierało sceny naturalne. Sekwencje różniły się między sobą ułożeniem kamer oraz ich rodzajem (5 sekwencji pozyskanych kamerami wszechkierunkowymi, pozostałe 10 pozyskano zwykłymi kamerami perspektywicznymi). Sprawdzono w jaki sposób wprowadzona modyfikacja wpływa na prędkości bitowe oraz w jaki sposób zmieniły się czasy dekodowania.

3.2. Wyniki

Wyniki eksperymentów zostały podzielone na trzy podrozdziały. W podrozdziale 3.2.1. zaprezentowano jak zmienił się czas dekodowania dla każdej z sekwencji, w 3.2.2. pokazano zmianę przepływności przy zachowaniu stałej jakości, natomiast podrozdział 3.2.3. prezentuje ile średnio strumienia bitowego zajmuje przesyłana głębia w zmodyfikowanym TMIV.

3.2.1. Porównanie szybkości przetwarzania procesu kodowania i dekodowania

Wyniki pokazane w Tab. 1 ukazują czas działania metody wykorzystującej wejściowe mapy głębi wyrażony jako procent czasu działania metody referencyjnej. Kodowanie TMIV to proces selekcji widoków wejściowych oraz pakowania ich do atlasów, a kodowanie Vvenc to proces kompresji wizyjnej któremu poddawane są atlasy stworzone przez TMIV. Zwiększenie czasu kodowania jest efektem spodziewanym, jako iż zwiększono ilość danych wejściowych i wprowadzono dodatkowy krok identyfikowania widoków zawierających najwięcej informacji.

Tab. 1. Procentowa zmiana czasu obliczeń dla proponowanej metody w porównaniu do metody referencyjnej.

Sekwencja testowa	Kodowanie TMIV [%]	Kodowanie Vvecn [%]	Dekodowanie i synteza [%]
ClassroomVideo	1468,3	88,6	44,3
Museum	1901,6	122,1	65,6
Fan	367,9	114,0	63,5
Kitchen	804,2	101,5	59,1
Painter	358,6	113,9	59,4
Frog	285,4	101,1	43,2
Carpark	214,4	115,2	47,2
Chess	412,1	106,8	75,9
Group	558,3	85,6	58,9
Fencing	360,5	94,9	52,1
Hall	278,5	103,9	61,7
Street	439,4	130,7	57,7
ChessPieces	435,1	115,4	70,5
Hijack	785,7	96,1	56,6
Mirror	228,1	113,9	66,0
Cadillac	479,6	137,8	52,4

Warto jednak zaznaczyć, że udało się zmniejszyć czas dekodowania o połowę w porównaniu z metodą referencyjną. Jest to bardzo wskazane w przypadku MIV DSDE, jako że złożoność dekodera dla tego profilu jest bardzo duża, co jest wynikiem przeprowadzania w nim estymacji głębi.

3.2.2. Zmiana przepływności dla stałej jakości

W Tab. 2 pokazano wpływ zmiany przepływności dla stałej jakości sekwencji mierzonej IV-PSNR [5]. Na zielono zaznaczono te miejsca, w których przy zachowaniu tej samej jakości można zmniejszyć strumień skompresowanych danych (brak określenia procentowego oznacza, że krzywe jakości w zależności od przepływności nie nakładały się na siebie, dlatego niemożliwe było określenie procentowej zmiany przepływności). Jak można zauważyć, dla większości sekwencji następuje zmniejszenie przepływności, więc proponowane modyfikacje mogą zapewnić lepszą jakość wizji wszechogarniającej dostarczanej widzowi jeżeli zdecydujemy się na zachowanie tej samej przepływności.

Tab. 2. Zmiana przepływności w proponowanej metodzie w porównaniu z metodą referencyjną dla tej samej jakości mierzonej za pomocą IV-PSNR [5].

Sekwencja testowa	Niskie przepływności [%]	Wysokie przepływności [%]
ClassroomVideo	-43,2	-29,0
Museum	3,8	-4,0
Fan	-1,1	13,9
Kitchen	-14,3	-15,2
Painter	-37,4	-26,0
Frog	11,5	9,8
Carpark	-5,5	0,6
Chess	---	---
Group	---	---
Fencing	-24,8	-9,0
Hall	---	-82,4
Street	-6,5	-7,8
ChessPieces	---	---
Hijack	---	---
Mirror	7,2	-1,7
Cadillac	-19,8	-11,7

3.2.3. Średnia liczba bitów poświęcanych na głębię

Wyniki zaprezentowane w Tab. 3 pokazują, że w zmodyfikowanej metodzie głębia zajmuje średnio jedynie 8% strumienia bitowego. Oznacza to, że jakość skompresowanych widoków wejściowych została jedynie nieznacznie zmniejszona, a jakość map głębi (jak pokazano na Rys. 4) została zauważalnie zwiększona. Daje to końcowy wzrost jakości widoku wirtualnego, co potwierdzają wyniki zaprezentowane w poprzednim podrozdziale.

Tab. 3. Podsumowanie całkowitej przepływności oraz zawartości strumienia bitowego.

Średnia przepływność [Mbps]		Część strumienia [%]					
MIV DSDE	Modyfikacja	MIV DSDE			Modyfikacja		
Całość strumienia (widoki + głębie + metadane)		Widoki	Metadane	Głębie	Widoki	Metadane	Głębie
36,870	35,751	100,0	0,0	0,0	95,8	0,0	4,2
21,332	20,620	99,9	0,1	0,0	94,2	0,1	5,7
11,644	11,209	99,9	0,1	0,0	91,9	0,1	8,0
5,970	5,815	99,8	0,2	0,0	89,4	0,2	10,4
3,039	2,900	99,6	0,4	0,0	86,8	0,4	12,8

4. PODSUMOWANIE

W tym artykule opisano usprawnienie procesu estymacji map głębi po stronie dekodera wizji wszechogarniającej. W metodzie referencyjnej do dekodera przesyłane są tylko atlasy, w których znajdują się widoki zawierające najwięcej informacji o scenie. Estymacja głębi dokonywana jest po stronie dekodera tylko na podstawie widoków wejściowych.

W proponowanym podejściu do pierwszego z czterech atlasów zapisywane są widoki zawierające najwięcej informacji o scenie, a do drugiego atlasu pakowane są wysokiej jakości mapy głębi odpowiadające widokom z atlasu pierwszego. W pozostałych dwóch atlasach przesyłane są pozostałe widoki wejściowe.

Przeprowadzono wszechstronne eksperymenty by zaprezentować zalety wprowadzonych modyfikacji. Zbadano jak modyfikacje zmieniły przepływność bitową, w jaki sposób zmienia się przepływność przy zachowaniu stałej jakości obiektywnej wizji wszechogarniającej, oraz jaki procent strumienia bitowego stanowiła skompresowana wejściowa głębia. Wyniki eksperymentów ukazały, że udało się osiągnąć pożądaną poprawę przy zachowaniu takiej samej całościowej przepływności bitowej. Daje to możliwość uzyskania lepszej jakości przy zachowaniu stałej przepływności. Dodatkowo, udało się skrócić czas dekodowania (zawierający czas estymacji głębi) o średnio 50% w porównaniu do metody referencyjnej, znacząco zwiększając konkurencyjność metody kodowania wizji wszechogarniającej opartej na estymacji map głębi po stronie dekodera.

PODZIĘKOWANIA

Praca finansowana ze środków przyznanych przez Ministerstwo Edukacji i Nauki.

LITERATURA

[1] Boyce J. M. i in. 2021. „MPEG Immersive Video Coding Standard''. *Proc. of the IEEE.*

[2] Bross B. i in. 2021. „Overview of the Versatile Video Coding (VVC) standard and its applications''. *IEEE T Circ. & Sys. Vid. Tech.*

[3] Domański M. i in. 2017. „Immersive visual media – MPEG-I: 360 video, virtual navigation and beyond''. *Proc. Int. Conf. Syst. Signals Image Process.*

[4] Doumanoglou A. 2018. „Quality of experience for 3-D immersive media streaming''. *IEEE T Br 64.*

[5] Dziembowski A. 2020. „Software manual of IV-PSNR software''.

[6] ISO/IEC 23090-12. „Information technology - Coded representation of immersive media - Part 12: MPEG Immersive video''. ISO/IEC JTC 1/SC 29.

[7] Mieloch D. i in. 2022. „Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video''. *IEEE T Circ. & Sys. Vid. Tech.*

[8] Mieloch D. i in. 2020. „Depth Map Estimation for FreeViewpoint Television and Virtual Navigation''. *IEEE Access.*

[9] Senoh T. i in. 2018. „Depth estimation and view synthesis for immersive media''. *IC3D conference.*

[10] Sullivan G. i in. 2012. „Overview of the High Efficiency Video Coding (HEVC) standard''. *IEEE T Circ. & Syst. Vid. Tech.*

[11] Tech G. i in. 2016. „Overview of the Multiview and 3D Extensions of High Efficiency Video Coding''. *IEEE T Circ. & Sys. Vid. Tech.*

[12] Wien M. i in. 2019. „Standardization Status of Immersive Video Coding''. *IEEE J Em. & Sel. Top. Circ. & Sys.*

PORÓWNANIE RÓŻNYCH METOD DETEKCJI I ROZPOZNAWANIA ZWIERZĄT NA OBRAZACH Z KAMERY TERMOWIZYJNEJ

COMPARISON OF DIFFERENT METHODS OF ANIMAL DETECTION AND RECOGNITION ON THERMAL CAMERA IMAGES

Łukasz Popek[1]; Rafał Perz[2]; Grzegorz Galiński[3]

[1] Politechnika Warszawska, Warszawa, lukasz.popek2.stud@pw.edu.pl
[2] Politechnika Warszawska, Wydział Mechaniki Energetyki i Lotnictwa, Warszawa, rafal.perz@pw.edu.pl
[3] Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Warszawa, grzegorz.galinski@pw.edu.pl

DOI: 10.15199/59.2022.4.66

Streszczenie: **Dla większości dziko żyjących zwierząt szczyt aktywności przypada na godziny nocne – ich obserwacja możliwa jest tylko przy użyciu specjalistycznych urządzeń. W niniejszej pracy zostało przeprowadzone porównanie różnych metod detekcji zwierząt na zdjęciach z kamery termowizyjnej: klasycznych (HOG/SVM) oraz opartych na głębokich sieciach neuronowych. Podczas testów na zbiorze danych zawierającym dwie rodziny zwierząt (*Cervidae* i *Suidae*) dla sieci YOLOv3 otrzymano wyniki mAP powyżej 90% dla IoU>50%.**

Abstract: **For most wild animals, peak of activity takes place during the night hours - their observation is possible only with the use of specialized equipment. In this study, a comparison of different methods for animal detection in thermal camera images was performed: classical (HOG/SVM) and based on deep neural networks. When tested on a dataset containing two animal families (*Cervidae* and *Suidae*) for the YOLOv3 network, obtained mAP was above 90% for IoU>50%.**

Słowa kluczowe: **detekcja, głębokie sieci neuronowe, termowizja, obserwacja dzikiej przyrody**

Keywords: **detection, deep neural networks, thermal imaging, wildlife surveillance**

1. WSTĘP

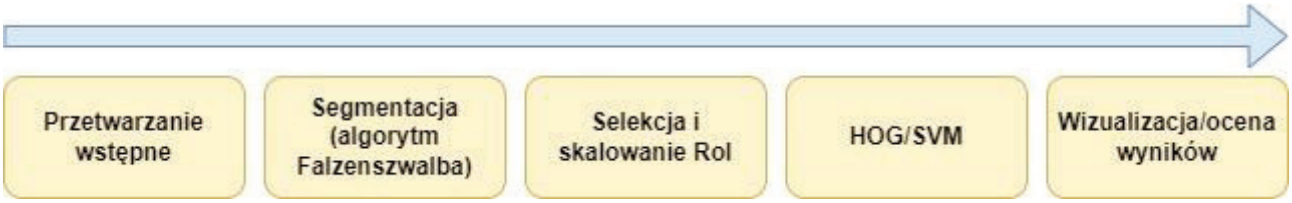
W dzisiejszych czasach, postępująca urbanizacja i przekształcanie naturalnych ekosystemów powoduje częste kontakty między człowiekiem oraz dzikimi zwierzętami. Rozwój technologii termowizyjnej oraz jej szeroka dostępności pozwoliły ludziom na badania i obserwacje zwierząt po zmroku. Większość obecnie prowadzonych badań dotyczy taksacji zwierzyny przy użyciu dronów, które pozwalają na zgrubne oszacowanie populacji danego gatunku. Jednak przy szacowaniu zasięgu występowania oraz badania rozprzestrzeniania się gatunków inwazyjnych takie podejście jest zbyt niedokładne – obserwacja powinna być prowadzona z powierzchni ziemi. Co więcej, proces rozpoznawania obiektów na zdjęciach powinien być możliwie jak najbardziej autonomiczny, żeby zmniejszyć czas pracy ekspertów. W niniejszej pracy podejmowane jest zagadnienie automatycznego znajdowania zwierząt na obrazach. Głównym celem jest porównanie istniejących metod, w celu wybrania najlepszej do dalszego badania wyżej wymienionego zagadnienia.

W automatyzacji detekcji obiektów występują dwa podstawowe problemy: lokalizacja regionów potencjalnie będących poszukiwanym obiektem (ang. *Region of Intrest*, RoI) oraz przetworzenie wielowymiarowego tensora do postaci wektora cech, który może być sklasyfikowany. W niemniejszej pracy rozpatrzono dwa podejścia. Jednym z najbardziej popularnych konceptów związanych z wyszukiwaniem potencjalnych obszarów zawierających obiekty jest podejście oparte na przesuwnym oknie względem obrazu. Jest to metoda jednak bardzo czasochłonna i kosztowna obliczeniowo. Zdjęcia z kamery termowizyjnej są monochromatyczne, co znacznie upraszcza ich analizę semantyczną w oparciu o bardziej wyrafinowane metody algorytmiczne. Po analizie literatury oraz testach do wyszukiwania regionów zainteresowania zdecydowano się wykorzystać algorytm Falzenszwalba [5]. Następnie znalezione obszary poddano operacji wyznaczenia histogramu zorientowanych gradientów (ang. *Histogram of Oriented Gradients*, HOG) [9], co w końcowym etapie pozwoliło przypisać klasy za pomocą wytrenowanej maszyny wektorów nośnych (ang. *Support Vector Machine*, SVM) [8].

Równolegle przetestowano rezultaty otrzymywane przy wykorzystaniu popularnych architektur splotowych sieci neuronowych. W tym kontekście zastosowano podejście *transfer learning* - tzn. dla wcześniej wytrenowanych modeli YOLOv3 [6] oraz Faster R-CNN [7] rozpoczęto proces dostrajania, aby przystosować parametry sieci do konkretnego problemu. Zbiór danych, na podstawie którego były ewaluowane wyniki, został stworzony od podstaw. Rezultaty obydwu zaproponowanych podejść porównano w oparciu o średnią precyzję (ang. *mean Avarage Precision*, mAP) dla różnych progów pokrycia (ang. *Intersection over Union*, IoU) oraz czułość (ang. *Recall*).

2. ZBIÓR DANYCH

Należy podkreślić, że głównym problemem w budowie skutecznych modeli do detekcji jest posiadanie poprawnie oznaczonego zbioru danych (ang. *dataset*). W przypadku niniejszego problemu nie istniały publicznie dostępne materiały spełniające wymagania.



Rys. 1 Schemat blokowy procesu detekcji opartego na HOG/SVM.

Z tego powodu został stworzony zbiór danych zawierający zdjęcia termowizyjne dwóch rodzin zwierząt: jeleniowatych (*Cervidae*), w skład której wchodziły zdjęcia: jelenia szlachetnego (*Cervus elaphus*), sarny europejskiej (*Capreolus capreolus*) i daniela (*Dama dama*) oznaczanych dalej jako klasa „*deer*", oraz świniowatych, w skład której wchodziły głównie zdjęcia dzika europejskiego (*Sus scrofa*), oznaczanych dalej jako klasa „*wild_boar*".

Rys. 2 Przykład dwóch rozróżnianych obiektów – po lewej klasa „deer", po prawej klasa „wild_boar.

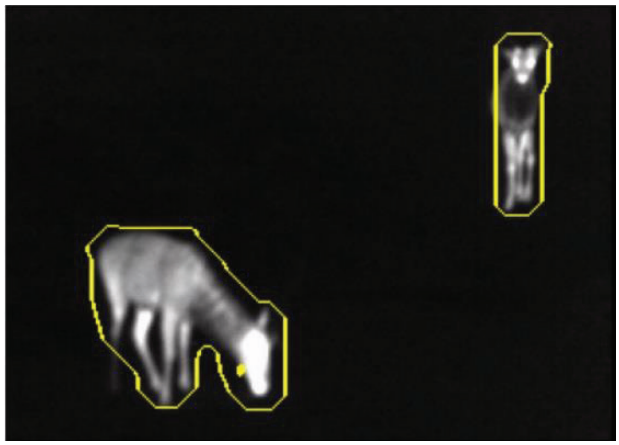
Do stworzenia zbioru danych została wykorzystana kamera Pulsar Helion 2 XP50 PRO. Część danych została uzyskana z prywatnych zbiorów dzięki uprzejmości polskich myśliwych. Do oznaczania obiektów na zdjęciach wykorzystany został program LabelIMG, generujący pliki .xml z parametrami obwiedni (ang. *boundingbox*). Wyniki przedstawione w niniejszej pracy uzyskano na zbiorze danych liczącym 400 zdjęć monochromatycznych – 250 zawierających obiekty z klasy „*deer*" i 150 z klasy „*wild_boar*".

3. ZASTOSOWANE METODY

3.1. HOG/SVM

Algorytm przedstawiający proces detekcji przy użyciu klasycznych metod został przedstawionych na schemacie blokowym na Rys 2. Pierwszym krokiem był przetwarzanie wstępne zdjęć, polegające na usunięciu artefaktów oraz symboli, które różne kamery zostawiają na zdjęciach. Kolejnym krokiem w zaproponowanym algorytmie było znalezienie regionów zainteresowania z potencjalnymi obiektami. Ze względu na to, że zdjęcia z kamery termowizyjnej są monochromatyczne, zdecydowano się na segmentację opartą na algorytmie Falzenszwalba, który grupuje piksele o podobnej jasności.

Tak wyznaczone segmenty zostały następnie przekształcone w prostokąty na podstawie skrajnych pikseli. Następnie był wykonywany proces pierwszej selekcji – odrzucano elementy zbyt małe (liczące poniżej 100 pikseli) oraz te, których stosunek długości do szerokości znacząco odbiegał od wartości 1 (zwierzęta na zdjęciach mieszczą się w obwiedni, która jest zbliżona do kwadratu). Tak wyselekcjonowane obiekty były najpierw przekształcane, aby długość i szerokość była jednakowa, a następnie skalowane do wymiarów 128x128 pikseli. Ponieważ charakter obrazów pozwala na uchwycenie krawędzi obiektów w sposób jednoznaczny, zdecydowano się na wyliczanie wektora cech do późniejszej klasyfikacji w oparciu o histogram zorientowanych gradientów. Technika ta zlicza występowanie konkretnej orientacji gradientu jasności w zlokalizowanym fragmencie obrazu. W efekcie dany obszar obrazu można przedstawić w postaci jednowymiarowego wektora, co umożliwia zastosowanie maszyny wektorów nośnych jako klasyfikatora. Do trenowania klasyfikatora zostało specjalnie przygotowane po 30 przykładów z każdej klasy. Do implementacji algorytmu Falzenszwalba, HOG i SVM została wykorzystana biblioteka Scikit-Image [4]. Ostatnim etapem była ewaluacja i wizualizacja wyników w celu oceny pracy całej metody.



Rys. 3 Przykłady regionów wyselekcjonowanych algorytmem Falzenszwalba.

3.2. Sieci neuronowe

W ostatnich latach podejście oparte na głębokich sieciach neuronowych zyskało szeroką popularność w kontekście analizy obiektów na obrazach cyfrowych. Większość rozwiązań obecnie stosowanych opiera się na zasadzie dostrajania parametrów sieci (ang. *finetuning*), wstępnie wytrenowanych na bardzo dużych zbiorach danych. Takie podejście pozwala na osiąganie dobrych rezultatów przy jednocześnie małej liczbie przykładów trenujących. Należy jednak podkreślić, że modele najczęściej trenowane są na zwykłych zdjęciach RGB, co oznacza, że wykorzystanie ich dla zagadnień o innej charakterystyce zdjęć, może skończyć się porażką. W niniejszej pracy wykorzystano dwie popularne architektury – Faster R-CNN oraz YOLOv3.

Faster R-CNN jest siecią, której działanie opiera się na rozbiciu obrazu na wiele warstw w poszukiwaniu cech charakterystycznych, następnie za pomocą RPN (ang. *Region Proposal Network*) znalezieniu obszarów potencjal-

nie będących obiektami, aż w końcu ich klasyfikacji i wyznaczeniu obwiedni. Podczas testów wykorzystano implementacje dla Faster R-CNN z biblioteki Detectron2 Model Zoo [3] z backbonem R_50_FPN.

YOLOv3 pomija wyszukiwanie map regionów interesujących z perspektywy problemu. Dla obrazu wejściowego generowana jest siatka, a następnie dopasowywane są obwiednie. Takie podejście, w przeciwieństwie do przesuwającego się okna po mapach wyekstrahowanych cech, pozwala modelowi YOLO na unikanie pomyłek fałszywie pozytywnego rezultatu dla tła, jak i na spojrzenie całościowe na obraz. Do testów w niniejszej pracy użyto wcześniej wytrenowanej architektury YOLO z biblioteki Darknet [2].

Ponieważ trenowanie sieci jest bardzo kosztowne obliczeniowo, wykorzystano aplikację Google Colab [1]. Pozwala ona na darmowe wykorzystanie wirtualnych maszyn zawierających GPU i umożliwiających szybki trening implementowanych rozwiązań.

4. WYNIKI

Testy przeprowadzono dla różnych konfiguracji hiperparametrów dostrajania sieci. Głównymi zmiennymi był współczynnik uczenia się (ang. *learning rate*, Lr), który odpowiada za regulowanie szybkości zmian współczynników wewnątrz sieci, oraz liczba iteracji, która odpowiada za czas nauki sieci. Początkowe poziomy wartości hiperparametrów zostały wyznaczone w sposób eksperymentalny.

Do oceny dokładności detekcji wykorzystano metryki mierzące precyzję oraz czułość modelu. Wzór ogólny na uśrednioną między klasami precyzję:

$$mAP_t = \frac{1}{n}\sum_{1}^{c=n}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{1}$$

gdzie t – minimalny próg pokrycia obwiedni wyznaczonej przez model z obwiednią odniesienia, aby uznać obiekt za poprawnie znaleziony, n - liczba klas, TP – liczba poprawnie wyznaczonych obiektów, FP – liczba fałszywie dodatnich rezultatów. W przypadku czułości wzór można przedstawić w postaci:

$$Recall = \frac{1}{n}\sum_{1}^{c=n}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{2}$$

z tą różnicą, że FN oznacza liczbę fałszywie negatywnych przypadków.

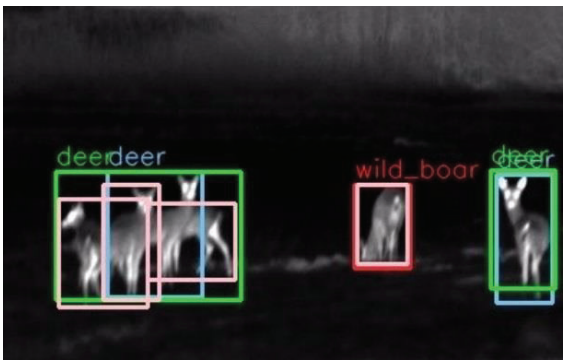
Najlepsze rezultaty dla poszczególnych detektorów zostały zebrane w Tabeli 1.

Tab. 1. Uzyskane wyniki

Algorytm	*Recall*	*mAP_{50}*	*mAP_{75}*	*Czas trenowania*
HOG/SVM	49%	40%	-	<1min
Yolov3	70%	94%	85%	~50min
Faster R-CNN	68%	87%	77%	~60min

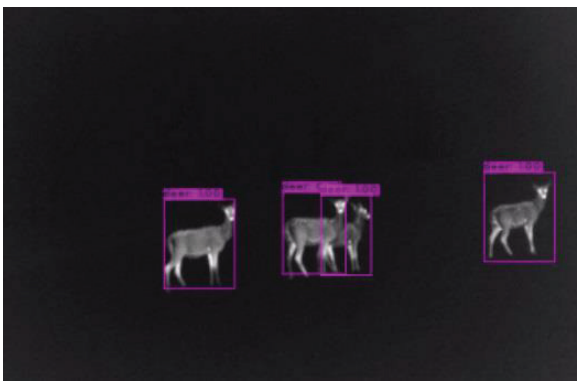
Jak widać klasyczne podejście, pomimo łatwości w trenowaniu SVM widocznym w czasie trwania, zdecydowanie odbiega od wyników metryk określających poprawność klasyfikacji uzyskiwanych dla obu sieci neuronowych.

Głównym problemem zaobserwowanym poprzez wizualizacje rezultatu detekcji był brak jednorodności jasności sylwetek zwierząt. Charakterystyka zdjęć oraz obiektów jakimi są zwierzęta powoduje występowanie niejednorodności jasności całego ciała na zdjęciu. Kolejnym problem jaki został zauważony wiązał się z detekcją zwierząt znajdujących się w zwartej grupie, przez co ich sylwetki nakładały się na siebie. Należy pamiętać że segmentacja w oparciu o algorytm Falzenszwalba bazuje na grupowaniu pikseli po barwie, zatem jej rezultat będzie zbliżony do segmentacji semantycznej, nie na poszczególne instancje. W przypadku dwóch zwierząt stojących koło siebie na obrazie, oznaczenie granicy sylwetek staje się dla algorytmu niemożliwe.



Rys. 4 Przykład rezultatu detekcji dla HOG/SVM.

Sieci neuronowe poprawnie poradziły sobie z zadaniem znajdowania dwóch zadanych klas obiektów. Głównym źródłem błędów obniżającym końcową wartość metryk były błędy przy rozdzieleniu wielu sylwetek zwierząt, natomiast błędy w klasyfikacji praktycznie się nie zdarzały tak samo jak występowanie fałszywie pozytywnych wyników czy zaznaczanie fragmentu sylwetki, który ma większą jasność. Jedyną wadą pozostaje czas – proces trenowania trwał około godziny, podczas której wykonano około 1000 iteracji trenowania sieci. Należy też podkreślić, że klasa „*deer*" miała większą precyzję i czułość detekcji w stosunku do klasy „*wild_boar*".



Rys. 5 Rezultat detekcji dla YOLOv3 z określonym prawdopodobieństwem zadanej klasy.

Rys. 6 Rezultat detekcji dla Faster R-CNN z określonym prawdopodobieństwem zadanej klasy.

5. PODSUMOWANIE

W toku prac zrealizowano podstawowe założenie jakim było wybranie optymalnej metody do dalszego rozwoju służącej do detekcji zwierząt na obrazach z kamery termowizyjnej. Sieci neuronowe poprawnie radzą sobie z problemem klasyfikacji i znajdowaniu obiektów na zdjęciach o takim charakterze, czego nie można stwierdzić o metodzie opartej na HOG/SVM.

W dalszej perspektywie należy rozszerzyć problem o kolejne klasy. Na pewno poważnym problemem może być uchwycenie różnic pomiędzy zwierzętami z tej samej rodziny, ale innych gatunków (np. sarna czy jeleń). Różnice w postaci cech morfologicznych (inne ubarwienie) nie są możliwe do uchwycenie kamerą termowizyjną, podczas gdy rozmiar w przypadku braku porównania może nie być wystarczającym do poprawnego rozróżnienia gatunkowego. Kolejną drogą rozwojową dla tak zdefiniowanego problemu jest udoskonalenie detekcji do postaci, pozwalającej na analizę filmów np. do analizy materiałów z fotopułapek umieszczonych przy żerowiskach. W celu zwiększania bezpieczeństwa lub w charakterze edukacyjnym, np. dla nocnych aktywności ludzkich takich jak praca służb mundurowych nocą czy łowiectwo, możliwa też będzie implementacja modułu, który w czasie rzeczywistym rozpoznawałby obiekty w takiej kamerze.

LITERATURA

[1] Bisong, E., 2019, Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA.

[2] Dokumentacja Darknet, 2022, https://github.com/pjreddie/darknet

[3] Dokumentacja Detectron2, 2022, https://github.com/facebookresearch/detectron2

[4] Dokumentacja Scikit-Image, wersja 0.19.2 – Luty 2022

[5] Felzenszwalb Pedro, Huttenlocher Daniel, 2004. „ Efficient Graph-Based Image Segmentation". International Journal of Computer Vision volume 59, strony 167–181.

[6] Redmon Joseph, Farhadi Ali, 2018, "An Incremental Improvement", arXiv, Cornel University

[7] Ren Shaoqing, He Kaiming, Girshick Ross, Sun Jian 2015, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv, Cornel University

[8] Russel Stuart, Peter Norwig, 2016, "Artificial Intelligence: A Modern Approach, Global Edition", Pearson, wydanie 3, rozdział 18, strony 745-748

[9] Russel Stuart, Peter Norwig, 2016, "Artificial Intelligence: A Modern Approach, Global Edition", Pearson, wydanie 3, rozdział 24, strona 946

[10] Tzutalin. LabelImg. Git code, 2015. https://github.com/tzutalin/labelImg

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

CZY KALIBRACJA OKULOGRAFU W INTERFEJSIE CZŁOWIEK-KOMPUTER ZAWSZE JEST NIEZBĘDNA?

IS AN EYETRACKER CALIBRATION IN HUMAN-COMPUTER INTERFACE ALWAYS NECESSARY?

Karolina Gabor-Siatkowska[1], Artur Janicki[2]

[1,2] Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Instytut Telekomunikacji, Zakład Cyberbezpieczeństwa, Warszawa
{karolina.gabor-siatkowska.dokt, artur.janicki}@pw.edu.pl

DOI: 10.15199/59.2022.4.67

Streszczenie**: Okulograf to narzędzie do pomiaru m.in. parametrów ruchów gałki ocznej, ostatnio także wykorzystywane w interfejsie człowiek-komputer. Służy wtedy jako źródło danych np. o stanie psychologicznym człowieka. Producenci okulografów zalecają ich kalibrację, co bywa kłopotliwe. Celem badania było ustalenie, czy ta kalibracja jest konieczna. W badaniu porównano dane uzyskane z okulografu z kalibracją oraz bez niej – mierzono m.in. wielkość średnicy źrenic. Wstępna analiza wykazała, że nie ma znaczącej różnicy między wynikami tych pomiarów.**

Abstract**: Eyetrackers are used to measure, e.g., eye movements, recently are also used in the human-computer interface. They serve as a source of data e.g. of the psychological state of a person. Manufacturers of eyetrackers recommend their calibration, which may be troublesome. The purpose of this study was to determine if this calibration is necessary. The study compared data obtained from an eyetracker with and without calibration, e.g., pupil diameter was measured. Preliminary analysis showed that there was no significant difference between the results of these measurements.**

Słowa kluczowe**: agent dialogowy, interfejs człowiek – komputer, kalibracja, okulografia**

Keywords**: calibration, dialog agent, eyetracking, human – computer interface**

1. WSTĘP

Okulograf (ang. *eyetracker*) to urządzenie umożliwiające śledzenie ruchów gałek ocznych wykorzystujące w tym celu widmo światła z zakresu bliskiej podczerwieni. Do najważniejszych uzyskiwanych parametrów można zaliczyć: średnicę źrenic oczu oraz parametry dotyczące ruchu gałek ocznych. Wybór konkretnego parametru do analizy zależy od celu przeprowadzonego badania. Śledzenie ruchów gałek ocznych oraz samej źrenicy wykorzystywane jest w wielu dziedzinach, takich jak: marketing, psychologia, jak również medycyna. Wykorzystuje się go również w badaniach rynku.

Od kilkunastu lat okulografy są wykorzystywane również do badania interakcji człowiek-komputer. Dane wynikowe mogą służyć np. algorytmom jako dane wejściowe w algorytmach wykorzystywanych we wspomaganiu wczesnej diagnostyki medycznej lub jako element interfejsu człowiek-komputer (ang. Human Computer Interface).

Przykładem w tej dziedzinie są agenty konwersacyjne (ang. Embodied Conversatonal Agents, ECA), którym poświęca się ostatnio wiele uwagi. Prace te wykorzystują zachowania niewerbalne do nawiązania kontaktu z ludzkim użytkownikiem [2]. Aby taki dialog mógł stać się bardziej wiarygodny, agenty te powinny być wyposażone w zdolności komunikacyjne i ekspresyjne podobne do tych, które znamy z kontaktów międzyludzkich (mowa, gesty, mimika twarzy, spojrzenie itp.).

Dzięki badaniom z zastosowaniem okulografu możliwe jest uzyskanie szeregu parametrów charakteryzujących np. stan emocjonalny człowieka [4], koncentrację [5]. Coraz częściej stosuje się badania okulograficzne jako dodatkowe źródło informacji podczas rozmów użytkowników z agentem dialogowym ([1], [2]). Dzięki temu uzyskiwane dane są następnie wykorzystywane do wspomagania albo sterowania agentami konwersacyjnymi. Prowadzony dialog człowiek-komputer jest dzięki temu bardziej naturalny [1], a co za tym idzie, agenty dialogowe są w stanie wejść w bardziej realistyczną interakcję z ludźmi.

Niniejszy artykuł ma następującą strukturę: najpierw w rozdziale 2. przedstawiono cel badań. Następnie w rozdziale 3. opisano szczegóły dotyczące przeprowadzonych eksperymentów. W 4. rozdziale dokonano analizy uzyskanych danych. W kolejnym, ostatnim rozdziale umieszczono podsumowanie oraz sugestie na temat dalszych prac.

2. CEL BADAŃ

Celem niniejszych badań jest sprawdzenie, czy kalibracja okulografu w interfejsie człowiek-komputer jest niezbędna. Jeżeli przeprowadzone badania i analiza wykażą, że kalibrację można pominąć, wówczas możliwe będzie zastosowanie okulografu do dalszych badań przeprowadzonych z udziałem pacjentów. Badania te będą polegały na wsparciu pacjentów podczas terapii z wykorzystaniem agenta dialogowego [8]. Okulograf będzie stanowił dodatkowe źródło informacji od pacjenta, w celu lepszego dostosowania dialogu awatara do indywidualnej rozmowy. U pacjentów kluczową rolę odgrywa uczucie

bezpieczeństwa i komfortu podczas przeprowadzenia terapii. Istotnym punktem jest również bezinwazyjność badania okulografem.

W czasie kalibracji urządzenia, zbierane są dane o skrajnych położeniach środka źrenicy prawego i lewego oka. Podczas ruchu gałki ocznej tworzy się obszar, w którym porusza się oko. Obszar ten zostaje odwzorowany na odpowiadający mu rejon ekranu monitora.

W przypadku okulografu firmy Gazepoint GP3 kalibracja polega na wodzeniu wzrokiem za okręgiem przemieszczającym się kolejno powoli w rogach monitora oraz na koniec centralnie w jego środku. Jeżeli kalibracja przebiegła pomyślnie, tzn. wzrok osób badanych został wykryty w odpowiednich okręgach i występowało minimum 4 na 5 punktów zgodności, wówczas można było przejść do przeprowadzenia badania. Natomiast jeżeli kalibracja wykazała zgodność poniżej 4 na 5 punktów, kalibrację u tej osoby powtarzano, aż otrzymano pozytywny komunikat. Taka procedura wymaga dodatkowego czasu oraz skupienia ze strony użytkownika. Niestety jest to niewygodne w przypadku użycia okulografu dla pacjentów.

Przeprowadzony eksperyment ma odpowiedzieć na pytanie, czy kalibracja okulografu firmy Gazepoint GP3 jest konieczna. Jeżeli dalsza analiza potwierdzi, że nie ma statystycznej różnicy pomiędzy danymi pochodzącymi z eksperymentu z kalibracją oraz bez kalibracji można by ominąć kalibrację i poprzez to zniwelować uczucie stresu czy dyskomfortu podczas tego etapu. W artykule zostaną przedstawione wstępne statystyki uzyskanych wyników z przeprowadzonego eksperymentu.

3. METODA BADAWCZA

3.1. Grupa badawcza

W badaniu wzięły udział 32 osoby w wieku 22-30 lat, w tym 16 kobiet oraz 16 mężczyzn, studentów i studentek Politechniki Warszawskiej i Uniwersytetu Warszawskiego. Trzynaście z nich nosi na co dzień okulary korekcyjne lub soczewki. W wywiadzie okazało się, że nie mają oni żadnych chorób okulistycznych czy neurologicznych. Biorąc pod uwagę przyszłe zastosowanie okulografu u pacjentów, celowo nie proszono osoby badane o zdjęcie okularów – miało to na celu przeprowadzenie badań o warunkach jak najbardziej zbliżonych do warunków w szpitalu.

3.2. Stanowisko badawcze

Eksperyment został przeprowadzony w jednym z laboratoriów Politechniki Warszawskiej, na Wydziale Elektroniki i Technik Informacyjnych. W pomieszczeniu zostało zapewnione naturalne oświetlenie światłem dziennym, równomiernie rozproszone, wpadające przez okno – zostało ono dopasowane przy pomocy żaluzji. Podczas badań każdy uczestnik siedział na regulowanym fotelu – ruch nie był skrępowany i osoba badana czuła się możliwie komfortowo. Fotel był przesunięty niedaleko stołu, na którym umieszczony był laptop wraz z niezbędnym oprogramowaniem i okulografem. Głowa osoby badanej była za każdym razem w odległości ok. 40-50 cm od monitora i okulografu. Stanowisko badawcze składało się z okulografu firmy Gazepoint GP3, który był podłączony do laptopa z ekranem o przekątnej 15,3' z zainstalowanym oprogramowaniem tej samej firmy (programy 'Gazepoint Control' oraz 'Gazepoint Analysis'). Okulograf ten jest lekki i przenośny, dzięki czemu może zostać postawiony w dowolnym miejscu. Podczas tego badania został umieszczony pod monitorem laptopa i za klawiaturą w taki sposób, aby możliwa była detekcja twarzy, a w szczególności oczu uczestników. Częstotliwość próbkowania urządzenia wynosiła 150 Hz, zaś dokładność wynosi 0,5 – 1° (według producenta, w warunkach idealnych).

Rysunek 1 – Osoba badana przy stanowisku badawczym

3.3. Opis eksperymentu

Każda z badanych osób miała za zadanie oglądanie obrazu przedstawionego na ekranie monitora przez 10s. W tym czasie okulograf monitorował ruch gałek ocznych uczestnika oraz rozpoznawał jego źrenicę.

Oglądany obraz był kolorowy, przedstawiający możliwie neutralną scenerię. Wybrano krajobraz – zdjęcie Dwunastu Apostołów, grupy naturalnie powstałych kolumn z wapienia stojących w morzu w pobliżu Australii.

W eksperymentach przeprowadzanych bez kalibracji dbano jedynie o odpowiednią odległość dookulografu za pomocą oprogramowania Gazepoint Control. Dzięki stałemu podglądowi oczu na żywo w tym programie można było ustalić odpowiednią odległość umieszczenia głowy i oczu osoby badanej do samego okulografu. Jeżeli program dawał pozytywną informację zwrotną o odległości oczu od urządzenia, wówczas można było przejść do wykonania badania.

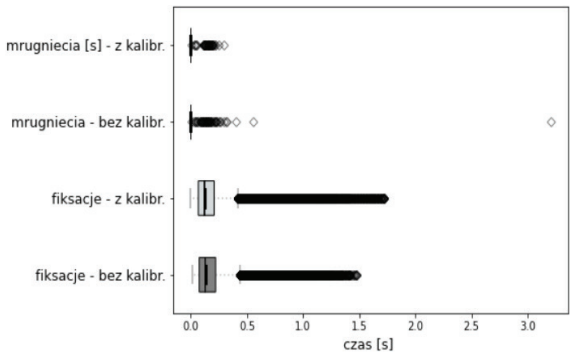
Z kolei w eksperymentach przeprowadzanych z kalibracją badanie było poprzedzone odpowiednio zdefiniowaną przez producenta okulografu procedurą kalibracji. Przed wykonaniem badania udzielano użytkownikowi krótkich wskazówek na temat tego, w jaki sposób przebiega badanie. Ważne było, aby badanie przebiegało w warunkach komfortowych, jak najbardziej zbliżonych do warunków, które będą na przykład w szpitalu. Oznaczało to, że osoba badana mogła mrugać, swobodnie ruszać głową, nie musiała uważać na żadne ograniczające ją reguły. W praktyce okazało się, że wszystkie osoby zachowywały się naturalnie, tzn. siedziały wyprostowane i spokojnie patrzyły się na ekran laptopa, nie wykonując żadnych nadmiernych ruchów głową.

3.4. Uzyskane dane

Podczas trwania badania rejestrowano szereg różnych danych, jednak do dalszej analizy wykorzystano parametry (które będą wykorzystywane w interfejsie człowiek-komputer): wielkość średnicy źrenicy prawej i lewej [mm], czas trwania mrugnięcia [s] oraz czas trwania fiksacji [s]. Fiksacja to moment, w którym gałka oczna znajduje się w fazie bezruchu. Czas trwania fiksacji wynosi zazwyczaj ok. 0,3-0.4 s. W tym czasie pobierana jest informacja wzrokowa z otoczenia i trafia ona do mózgu w celu jej dalszego przetwarzania [6]. Przejścia pomiędzy jedną fiksacją a kolejną nazywa się sakkadami. Ostatecznie do dalszej analizy użyto danych od 31 osób, ponieważ w jednym przypadku nastąpił błąd podczas zapisu badania.

4. ANALIZA UZYSKANYCH DANYCH

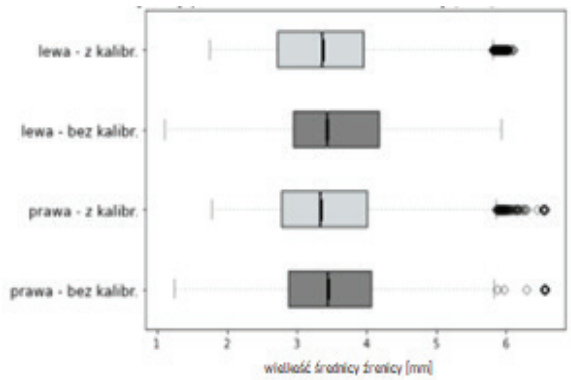
Po wykonaniu badania dane wyeksportowano do plików „csv". Do analizy użyto języka Python wraz z niezbędnymi pakietami służącymi do analizy danych i rysowania wykresów, m.in. Seaborn, Pyplot, Numpy, Pandas. Zazwyczaj w badaniach dotyczących analizy wzroku, np. [3], przy analizie danych dotyczących średnicy źrenicy pomija się pierwsze sekundy czy minuty (w zależności od czasu trwania całego eksperymentu). Uzasadnieniem takiego postępowania jest fakt, że w tym czasie następuje akomodacja oka do zaistniałych warunków, tzn. do poziomu jasności pomieszczenia oraz oświetlenia padającego z ekranu komputera. W niniejszym badaniu, które trwało 10 s usunięto pierwsze 3 s badania u każdej z osób. W pierwszej kolejności dokonano analizy czasu trwania mrugnięć oraz fiksacji wszystkich uczestników (Rysunek 2).



Rysunek 2 – Wykresy pudełkowe dla wartości fiksacji i mrugnięć [s]

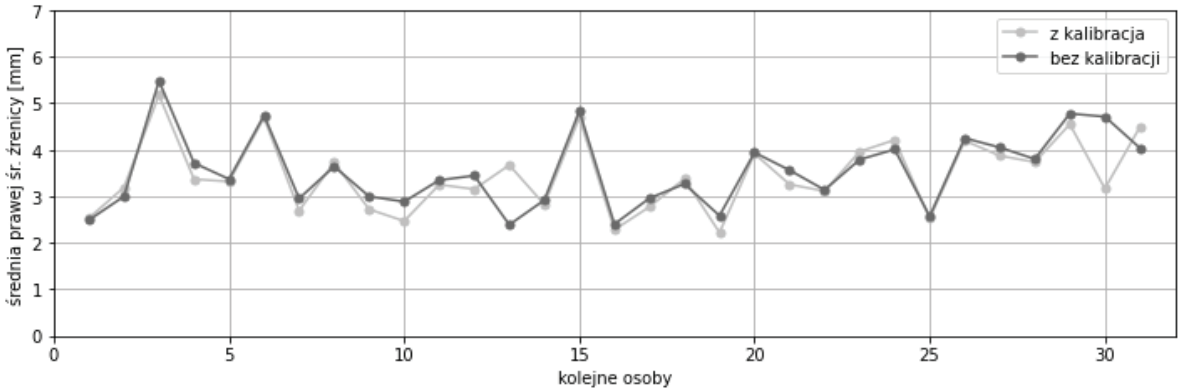
Analizując powyższe wykresy, można zaobserwować, że zarówno wartości fiksacji jak i wartości mrugnięć, niezależnie od przeprowadzenia kalibracji, przyjmowały niemal identyczny zakres. W każdym wariancie eksperymentu występowały wartości odstające. Można również zauważyć bardzo dużą wartość mrugnięcia wynoszącą powyżej 3 s – nie można jednoznacznie powiedzieć, czy typ eksperymentu miał tutaj jakiekolwiek znaczenie, być może osoba badana zamknęła oczy na ok. 3 s.

Następnie wykonano analizę porównawczą uzyskanych wartości średnicy źrenicy oka prawego oraz oka lewego (Rysunek 3). Pionowa linia wewnątrz poszczególnych pudełek wskazuje na medianę.



Rysunek 3 – Wykresy pudełkowe dla wartości średnicy źrenicy lewej oraz prawej [mm]

Analizując powyższy wykres, można ponownie powiedzieć, że zakres wartości lewej i prawej średnicy źrenicy osób badanych jest niemal identyczny w przypadku obu typów badań. Wartości odstające (zobrazowane rombami) występują zarówno w przypadku badania z kalibracją, jak i w przypadku badania bez kalibracji. Ponadto należy zauważyć, że dane pochodzące z obu eksperymentów przyjmują typowe wartości średnicy źrenicy (w zależności od zaistniałych warunków między 2-8 mm) [7]. Można zauważyć, że tylko w przypadku eksperymentu bez kalibracji odnotowywano wartości minimalne średnicy źrenicy w okolicy 1 mm – takie dane raczej nie występują. Prawdopodobnie są to dane, które nie zostały zakwalifikowane do mrugnięcia podczas zamykania oczu, ponieważ analiza i wszystkie wykresy zostały przeprowadzone na surowych danych. Zatem detekcja mrugnięć odbywała się bez dodatkowej analizy czy przetwarzania, bazując tylko na standardowym rozpoznaniu mrugnięć przez oprogramowanie Gazepoint. Aby lepiej uwidocznić ewentualne różnice pomiędzy uzyskanymi danymi, przeanalizowano średnie wartości wielkości średnicy źrenicy dla każdej osoby biorącej udział w eksperymencie (Rysunek 4).



Rysunek 4 – Wykres przedstawiający średnie wartości prawej średnicy źrenicy (z kalibracją oraz bez) dla poszczególnych osób.

Rysunek nr 4 przedstawia średnie z całego badania dla każdej osoby – z kalibracją oraz bez kalibracji. Z wykresów można wywnioskować, że u większości różnica między średnią wartością średnicy źrenicy mierzoną zarówno w jednym jak i w drugim przypadku nie przekracza 0,5 mm. Tylko w jednym przypadku (u przedostatniej osoby) odnotowano dużą różnicę wynoszącą ok. 1,25 mm. Ponieważ ta osoba niosła okulary, przypuszczamy, że jest to spowodowane błędnym rozpoznaniem odblasku okularów jako źrenicy podczas badania. W dalszych etapach prac planowana jest bardziej szczegółowa analiza statystyczna osiągniętych wyników.

5. REZULTATY I SUGESTIE

Przeprowadzone badania umożliwiły rozeznanie, czy kalibracja okulografu stosowanego np. w interfejsie człowiek-komputer jest zawsze konieczna. Z danych wstępnie wynika, że nie ma znaczącej różnicy pomiędzy średnimi występującymi dla poszczególnych osób w eksperymencie poprzedzonym kalibracją oraz bez niej. Uzyskany rozrzut danych dotyczących prawej i lewej średnicy źrenicy sugeruje, że kalibracja nie ma znaczącego wpływu na rejestrowane wartości. Mediana oraz drugi i trzeci kwartyl danych pochodzących z dwóch eksperymentów wydają się przyjmować niemal identyczne wartości. Również w przypadku danych dotyczących mrugnięć oraz fiksacji nie można zauważyć znaczącej różnicy. Obserwacje te sugerują, że hipoteza o braku konieczności przeprowadzania kalibracji jest bardzo prawdopodobna, co warto by jeszcze potwierdzić odpowiednimi testami statystycznymi. Z wysokim prawdopodobieństwem można jednak stwierdzić, że nawet bez kalibracji okulograf może służyć np. do wspomagania dialogów z agentami dialogowymi w interfejsie człowiek-komputer.

Badania były finansowane przez Centrum Badawcze Priorytetowego Obszaru Badawczego Sztuczna Inteligencja Robotyka ze środków Politechniki Warszawskiej w ramach Programu Inicjatywa Doskonałości – Uczelnia Badawcza (IDUB).

LITERATURA

[1] Bailly Gérard, Elisei Frédéric, Raidt Stephan, Casari Alix, and Picot Antoine. 2006. "Embodied Conversational Agents: Computing and Rendering Realistic Gaze Patterns.". *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 4261 LNCS.

[2] Bee Nikolaus, André Elisabeth, and Tober Susanne. 2009. "Breaking the Ice in Human-Agent Communication: Eye-Gaze Based Initiation of Contact with an Embodied Conversational Agent.". *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 5773 LNAI.

[3] Bentivoglio Anna Rita, Bressman Susan B., Cassetta Emanuele, Carretta Donatella, Tonali Pietro and Albanese Alberto. 1997. "Analysis of Blink Rate Patterns in Normal Subjects." *Movement Disorders* 12 (6).

[4] Bradley Margaret M., Miccoli Laura, Escrig Miguel A. and Lang Peter J. 2008. "The Pupil as a Measure of Emotional Arousal and Autonomic Activation." *Psychophysiology* 45 (4).

[5] Chang Kang Ming, and Wu Chueh Miao Tien. 2019. "Using Eye Tracking to Assess Gaze Concentration in Meditation.". *Sensors (Switzerland)* 19 (7).

[6] Sielicka-Różyńska Maria. 2021. "Wykorzystanie Badań Eye Trackingowych Do Oceny Postrzegania Wzrokowego Znakowania Produktów Spożywczych.". *PRZEMYSŁ SPOŻYWCZY* 1 (2): 18–23.

[7] Wickremasinghe Sanj S., Smith Guy T., and Stevens Julian D. 2005. "Comparison of Dynamic Digital Pupillometry and Static Measurements of Pupil Size in Determining Scotopic Pupil Size before Refractive Surgery.". *Journal of Cataract and Refractive Surgery* 31 (6): 1171–76.

[8] Zygadło Artur, Kozłowski Marek, and Janicki Artur. 2021. "Text-Based Emotion Recognition in English and Polish for Therapeutic Chatbot.". *Applied Sciences (Switzerland)* 11 (21).

WYKRYWANIE KONFLIKTÓW DECYZJI W INTELIGENTNYCH STEROWNIKACH RADIOWYCH W OTWARTYCH SIECIACH DOSTĘPU RADIOWEGO O-RAN

CONFLICT DETECTION FRAMEWORK FOR CONFLICT MITIGATION
IN OPEN RADIO ACCESS NETWORKS

Cezary Adamczyk[1]; Adrian Kliks[2]

[1] Politechnika Poznańska, Poznań, cezary.adamczyk@doctorate.put.poznan.pl
[2] Politechnika Poznańska, Poznań, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2022.4.68

Streszczenie: **Koncepcja otwartych sieci dostępu radiowego (O-RAN) uwzględnia aplikacje xApp, które sterują konfiguracją sieci. Mimo pojawiania się nowych zastosowań xApp, problem konfliktów między ich decyzjami nie został zgłębiony. Artykuł zawiera opis struktury systemu osadzonego w architekturze sieci O-RAN, który pozwala na wykrywanie wszystkich rodzajów konfliktów między decyzjami xApp. Wnioskuje się, że określenie metod wykrycia konfliktów jest pierwszym krokiem do zapewnienia ustandaryzowanej metody rozwiązywania konfliktów w sieciach O-RAN.**

Abstract: **Evolution of the Open RAN concept brings light to xApps and what are their potential use-cases in the O-RAN-compliant networks. The issue of detecting conflicts between decisions made by various xApps has not yet been addressed in detail. This paper defines a Conflict Detection Framework built into the existing O-RAN architecture to reliably detect all conflict types defined in the O-RAN Technical Specifications. It is concluded that defining a unified framework for all conflict types is the first step to provide a standard method of conflict mitigation in O-RAN.**

Słowa kluczowe: **otwarte sieci radiowe, rozwiązywanie konfliktów, wykrywanie konfliktów, xApp**

Keywords: **conflict detection, conflict mitigation, O-RAN, xApp**

1. WSTĘP

Koncepcja Otwartych Sieci Dostępu Radiowego (ang. *Open Radio Access Network*, O-*RAN*) jest rozwijana od wielu lat przez konsorcja składające się z operatorów sieci mobilnych, firm telekomunikacyjnych i placówek badawczych. Jednym z tych konsorcjów jest O-RAN Alliance – spółka, która aktywnie dąży do rozwoju sieci O-RAN poprzez opracowywanie i publikację specyfikacji technicznych.

Priorytety poszczególnych zagadnień technicznych we wszystkich obszarach sieci O-RAN są ustalane na podstawie planu Minimum Viable Plan (MVP) w kierunku komercjalizacji koncepcji [1], który określa minimalny zbiór cech sieci O-RAN sprawiający, że będzie ona wnosiła rzeczywistą wartość biznesową dla operatorów sieci mobilnych. Priorytety poszczególnych rozwiązań są wybierane przez operatorów sieci mobilnych będącymi członkami O-RAN Alliance. Od czerwca 2021 r. priorytetem są funkcje sieci O-RAN mające na celu m. in. sterowanie ruchem, optymalizację jakości usług (QoS) i doświadczenia (QoE) oraz optymalizacja transmisji w technice MIMO (Many Inputs Many Outputs).

Otwartość sieci O-RAN pozwala na nowatorskie podejście do dostosowania działania sieci do różnych warunków. Interfejsy pomiędzy logicznymi komponentami w architekturze O-RAN są ustandaryzowane i otwarte, co daje możliwość stosowania elementów systemu różnych dostawców bez ograniczania funkcjonalności sieci. Jest to poważna odmiana względem tradycyjnych sieci dostępu radiowego, w których sprzęt i oprogramowanie są zwykle dostarczane przez jednego producenta.

Kolejnym unikalnym aspektem sieci O-RAN jest wprowadzenie dwóch nowych funkcji logicznych, które mogą wpływać na działanie sieci radiowej, tj. Inteligentnych Sterowników Radiowych (ang. *Radio Intelligent Controller*, *RIC*) [2]: RIC działającego w czasie bliskim rzeczywistemu (ang. *Near-Real Time RIC*, *nRT-RIC*) oraz RIC działającego nie w czasie rzeczywistym (ang. *Non-Real Time RIC*, *NRT-RIC*). Pierwszy z nich działa w pętlach sterowania o czasie trwania od 10 milisekund do 1 sekundy, a drugi w pętlach sterowania o czasie trwania powyżej 1 sekundy. Działanie obu komponentów ma na celu optymalizację pracy sieci radiowej; nRT-RIC reaguje na krótkoterminowe trendy w pracy sieci decydując o jej konfiguracji i przydziale zasobów, natomiast NRT-RIC monitoruje długoterminowe trendy w pracy sieci i dostarcza nRT-RIC aktualizacje polityki zarządzania siecią.

Zastosowanie sterowników RIC pozwala na uruchamianie aplikacji xApp (w nRT-RIC) i rApp (w NRT-RIC), które wspomagają sterowanie siecią. Zarówno aplikacje xApp, jak i rApp, mogą być opracowywane przez niezależne firmy, a następnie wdrażane we wszystkich implementacjach sterowników RIC zgodnych ze standardem O-RAN, niezależnie od dostawcy sprzętu i/lub oprogramowania w sieci radiowej. Przykładem wykorzystania aplikacji rApp jest dostosowanie alokacji zasobów radiowych w celu zmniejszenia opóźnień i poprawy wydajności w dynamicznym przełączaniu w scenariuszach V2X [3]. Przykładem zastosowania aplikacji xApp jest zarządzanie zasobami radiowymi w sposób maksymalizujący QoS dla dedykowanej grupy użytkowników [4].

2. KONFLIKTY MIĘDZY DECYZJAMI APLIKACJI XAPP

2.1. Rodzaje konfliktu

Elastyczność sieci O-RAN i nowatorskie komponenty inteligentnego sterowania, chociaż pozwalają na

stosunkowo prostą modyfikację działania RAN, tworzą nowe wyzwania w sterowaniu siecią. Ponieważ dodatkowe komponenty mogą decydować o parametrach działania sieci radiowej (aplikacje xApp i rApp), istnieje pewne ryzyko wystąpienia konfliktów między decyzjami w środowisku O-RAN. Konflikty te mogą potencjalnie neutralizować lub zmniejszać wpływ sterowania na wydajność sieci; w określonych scenariuszach konflikt może nawet pogorszyć wydajność w porównaniu do stanu sieci przed konfliktującymi zmianami konfiguracji sieci [5]. Konflikty decyzji w sieci O-RAN zostaną omówione w tym artykule w kontekście aplikacji xApp w sterowniku nRT-RIC.

Specyfikacje techniczne O-RAN rozróżniają trzy rodzaje konfliktów, które mogą wystąpić podczas działania sterownika nRT-RIC [6]. Pierwszy z nich to konflikt bezpośredni, który dotyczy sprzecznych decyzji następujących jedna po drugiej. Przykładem takiego konfliktu może być sytuacja, w której xApp #1 decyduje się przypisać użytkownika do komórki, a zaraz po tym xApp #2 decyduje się przypisać tego samego użytkownika do innej komórki. W tym konkretnym scenariuszu rezultatem sprzecznych działań będzie to, że tylko decyzja xApp #2 będzie miała rzeczywisty skutek. Jeśli konflikt nie zostałby wykryty, może to sprawić, że xApp #1 wyciągnie błędne wnioski na temat skutków swojej decyzji.

Kolejnym rodzajem konfliktu w nRT-RIC jest konflikt pośredni. Odnosi się do sytuacji, w której decyzje są ze sobą sprzeczne, modyfikując różne parametry wpływające na te same obszary działania sieci. Na przykład, może on wystąpić między aplikacjami xApp modyfikującymi parametry wpływające na granice przekazywania użytkowników między komórkami, np. gdy xApp #1 kontroluje nachylenie elektryczne anteny w komórce, a xApp #2 zmienia offset pomiarów w tej samej komórce. Może to prowadzić do sytuacji, w których efektywna granica przekazywania użytkowników między komórkami zmienia się chaotycznie, ponieważ obie aplikacje reagują na wzajemnie sprzeczne decyzje o konfiguracji sieci.

Ostatnim i najbardziej problematycznym rodzajem konfliktu jest konflikt niejawny. Dotyczy on sytuacji, w których sprzeczne wpływy różnych decyzji są trudne do zidentyfikowania. Konflikt niejawny może wystąpić, gdy wiele aplikacji xApp jednocześnie optymalizuje działanie sieci z oddzielnymi celami modyfikując różne parametry sieci, np. w sytuacji, gdy xApp #1 dąży do maksymalizacji QoS dla priorytetyzowanej grupy użytkowników, podczas gdy xApp #2 dąży do zminimalizowania liczby przełączeń z sąsiednimi komórkami.

2.2. Rozwiązywanie konfliktów

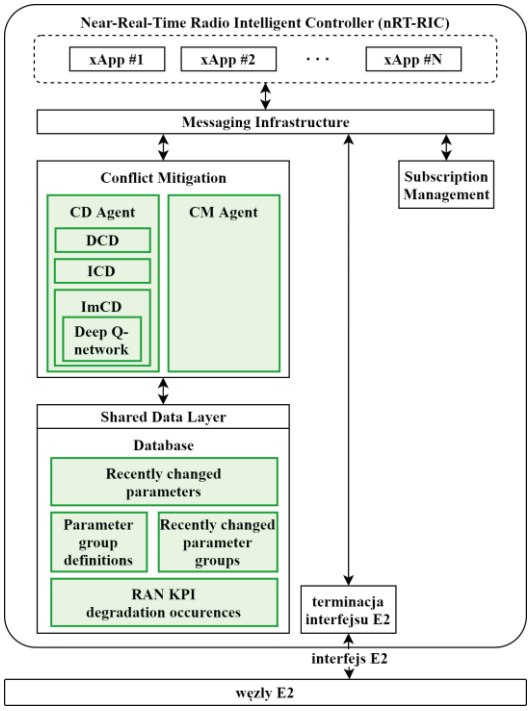
Aby osiągnąć przewidywalną wydajność i zachowanie sieci O-RAN, wymagane jest minimalizowanie negatywnych skutków konfliktów między decyzjami aplikacji xApp. Sterownik nRT-RIC musi być w stanie łagodzić wszelkie konflikty, co w pierwszej kolejności wymaga wykrycia ich.

Temat wykrywania i łagodzenia konfliktów nie został jeszcze w znaczący sposób omówiony w artykułach naukowych, a także w specyfikacjach technicznych O-RAN (będą one obiektem przyszłych badań [6]). Jedyny artykuł na ten temat został opublikowany przez Zhang i in. (2022) [7]. Autorzy zaproponowali w nim algorytm uczenia zespołowego oparty na głębokim uczeniu maszynowym Deep Q wymagający dwóch aplikacji xApp współpracujących w środowisku O-RAN: jednej przydzielającej moc do bloków zasobów fizycznych stacji bazowej i drugiej przydzielającej bloki zasobów fizycznych użytkownikom sieci. Wyniki działania algorytmu wykazały, że gdy obie aplikacje biorą pod uwagę swoje decyzje, ich działanie jest znacznie korzystniejsze niż gdy aplikacje działają niezależnie. Współpraca aplikacji prowadzi do zwiększenia przepustowości systemu i zmniejszenia prawdopodobieństwa utraty pakietów. Podejście to jest jednak trudne do skalowania do większej liczby aplikacji, gdyż logika decyzyjna we wszystkich uczestniczących aplikacjach musi dostosować się do liczby współpracujących aplikacji i być w stanie właściwie interpretować ich decyzje; uczenie zespołowe może być stosowane tylko wtedy, gdy aplikacjesą zaprojektowane z myślą o nim.

3. CONFLICT DETECTION FRAMEWORK

3.1. Struktura systemu CDF

W tym artykule autorzy opisują strukturę systemu do wykrywania konfliktów (odtąd określaną jako *Conflict Detection Framework* lub *CDF*) w sieciach O-RAN. Proponowany system jest przewidziany jako składowa komponentu łagodzenia konfliktów (ang. *Conflict Mitigation*, *CM*) w nRT-RIC. CDF ma na celu zapewnienie znormalizowanej metody wykrywania wszystkich rodzajów konfliktów zdefiniowanych w specyfikacjach technicznych O-RAN, bez specjalnych wymagań wobec implementacji aplikacji xApp. Strukturę systemu CDF w kontekście architektury nRT-RIC pokazano na rysunku 1.



Rys. 1. Struktura systemu CDF w kontekście architektury nRT-RIC (komponenty CDF w kolorze zielonym)

Struktura wykrywania konfliktów przewiduje komponenty: agenta do wykrywania konfliktów (ang. *Conflict Detection Agent*, *CDA*) i agenta do łagodzenia konfliktów (ang. *CM Agent*, *CMA*) jako jednostki w komponencie CM sterownika nRT-RIC, które sygnalizują wszelkie wykryte konflikty i łagodzą ich skutki. W CDA wyróżnia się trzy oddzielne komponenty logiczne, które opisano w kolejnych podrozdziałach poświęconych metodom wykrywania poszczególnych rodzajów konfliktów pomiędzy decyzjami aplikacji xApp. Zakłada się, że CMA jest w stanie złagodzić skutki wszystkich konfliktów wykrytych przez CDA; logika jego działania wykracza poza zakres tego artykułu.

Dane przechowywane przez CDA są przechowywane w bazie danych (ang. *Database*), do której ma dostęp za pomocą warstwy współdzielonych danych (ang. *Shared Data Layer*, *SDL*). Zakłada się, że dane związane z wykrywaniem konfliktów mogą być odczytywane przez dowolny komponent w ramach sterownika nRT-RIC, natomiast zapis może być wykonywany tylko przez CDA.

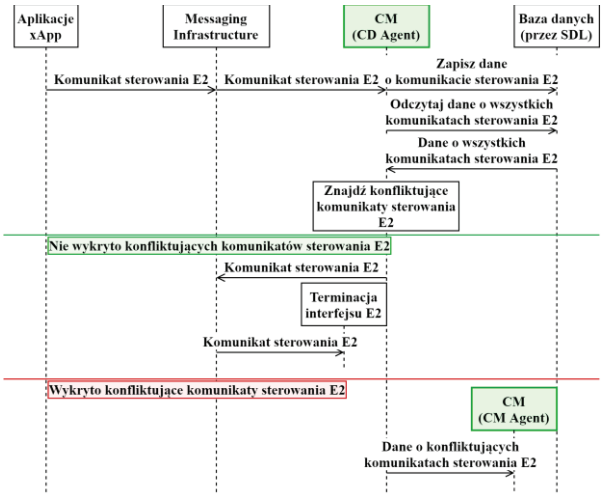
System CDF działa przy założeniu, że wszystkie komunikaty sterowania E2 dostarczane przez aplikacje xApp uruchomione w nRT-RIC są przekazywane do CM. Dlatego infrastruktura komunikatów (ang. *Messaging Infrastructure*) w nRT-RIC musi być skonfigurowana w taki sposób, aby przekierowywać wszystkie komunikaty sterowania E2 z aplikacji xApp do CM. W zależności od wyników działań CDA i CMA, komunikaty sterowania E2 mogą być przekazywane do węzłów E2 przez terminacje interfejsu E2 (z modyfikacjami lub bez, w zależności od rozwiązania w CMA) lub blokowane przed sterowaniem węzłem E2. Należy podkreślić, że dodanie dodatkowego kroku w procesie przetwarzania komunikatów sterowania E2 zwiększa opóźnienie między dostarczeniem przez aplikację decyzji o sterowaniu pracą RAN a zadziałaniem tej decyzji w sieci. Ponieważ nRT-RIC ma na celu kontrolę działania sieci w pętlach, które nie są dłuższe niż 1 sekunda, przetwarzanie komunikatów sterowania E2 wykonywane w CM musi być stosunkowo szybkie.

Ponadto, wymogiem CDF dla architektury sterownika nRT-RIC jest to, że CDA musi subskrybować dane o wydajności sieci dostarczane przez węzły E2. Jest to wymagane do wykrywania niejawnych konfliktów, które zostaną szczegółowo opisane w jednym z kolejnych rozdziałów.

3.2. Wykrywanie konfliktów bezpośrednich

Bezpośrednie konflikty mogą być obserwowane bezpośrednio przez komponent Direct Conflict Detection (DCD) w CDA. DCD wykrywa konflikty przed podjęciem działań (tj. zanim jakakolwiek decyzja aplikacji będzie miała wpływ na sieć) na podstawie zarejestrowania zmiany ostatnio zmienionego parametru związanego z danym użytkownikiem, komórką lub nośnikiem połączenia (ang. *bearer*). Wszystkie komunikaty sterowania E2 są rejestrowane przez CDA w bazie danych („Recently changed parameters" na rysunku 1) wraz z sygnaturą czasową, identyfikatorem źródłowej aplikacji xApp, celem komunikatu sterowania E2, parametrami modyfikowanymi przez komunikat i okresem sterowania komunikatu (tj. wartością, która określa czas ważności zmiany w sieci oczekiwanej przez dany komunikat sterowania E2).

Logika wykrywania konfliktów w DCD jest prosta: każdy nowy komunikat sterowania E2 jest zapisywany w bazie danych, a następnie porównywany ze wszystkimi innymi zapisanymi i nadal ważnymi komunikatami kontrolnymi dla danego węzła E2. Jeśli jakiekolwiek dwie ważne decyzje modyfikują ten sam parametr, dane o konflikcie są dostarczane do CMA. Przepływ komunikatów w DCD przedstawiono na rysunku 2.



Rys. 2. Przepływ danych związany z wykrywaniem konfliktów bezpośrednich za pomocą komponentu DCD

3.3. Wykrywanie konfliktów pośrednich

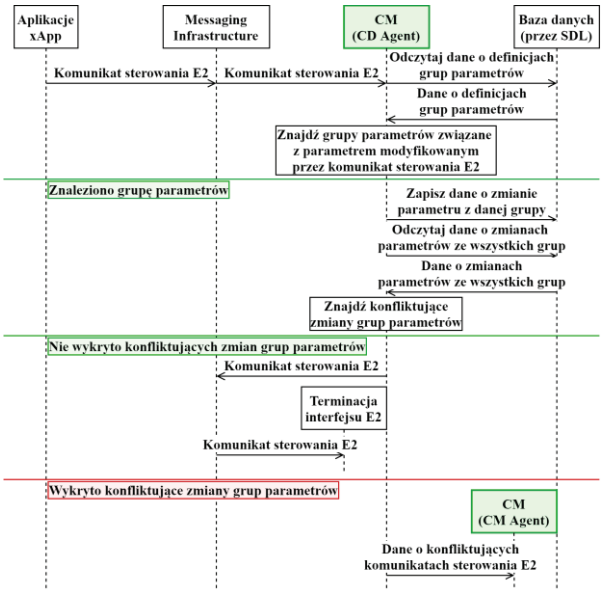
Konflikty pośrednie nie mogą być obserwowane bezpośrednio przez komponent Indirect Conflict Detection (ICD), jednak mogą one być przewidywane znając grupy parametrów, które wpływają na ten sam obszar działania sieci O-RAN. Dlatego wszelkie relacje między parametrami muszą być wstępnie zdefiniowane i znane przez CDA, aby możliwe było wykrycie takich konfliktów. Relacje te są przechowywane w bazie danych jako definicje powiązanych grup parametrów („Parameter group definitions" na rysunku 1). Dodatkowo baza danych przechowuje informacje o komunikatach sterowania E2 modyfikujących parametry z tych grup („Recently changed parameter groups" na rysunku 1), z zakresem danych takim samym jak w „Recently changed parameters" dotyczącym DCD, jednak odnoszącym się do grup parametrów zamiast do pojedynczych parametrów.

Wykrywanie konfliktów pośrednich to niewielka modyfikacja DCD; również rejestruje każdy nowy komunikat sterowania E2, ale najpierw mapuje modyfikowany parametr na uprzednio zdefiniowaną grupę parametrów. Po zarejestrowaniu komunikatu dotyczącego danej grupy parametrów, ICD sprawdza wpisy w bazie danych, aby znaleźć wszelkie ważne sprzeczne decyzje modyfikujące parametry z tej samej grupy; jeśli zostanie znalezione jakiekolwiek dopasowanie, dane o konflikcie są dostarczane do CMA. Przepływ komunikatów w ICD przedstawiono na rysunku 3.

3.4. Wykrywanie konfliktów niejawnych

Niejawnych konfliktów nie można zaobserwować bezpośrednio, a wzajemne oddziaływanie między

sprzecznymi decyzjami nie jest łatwe do wykrycia. Dlatego podejście jak dla DCD i ICD, tj. wykrywanie konfliktu przed przesłaniem komunikatu sterowania E2, nie



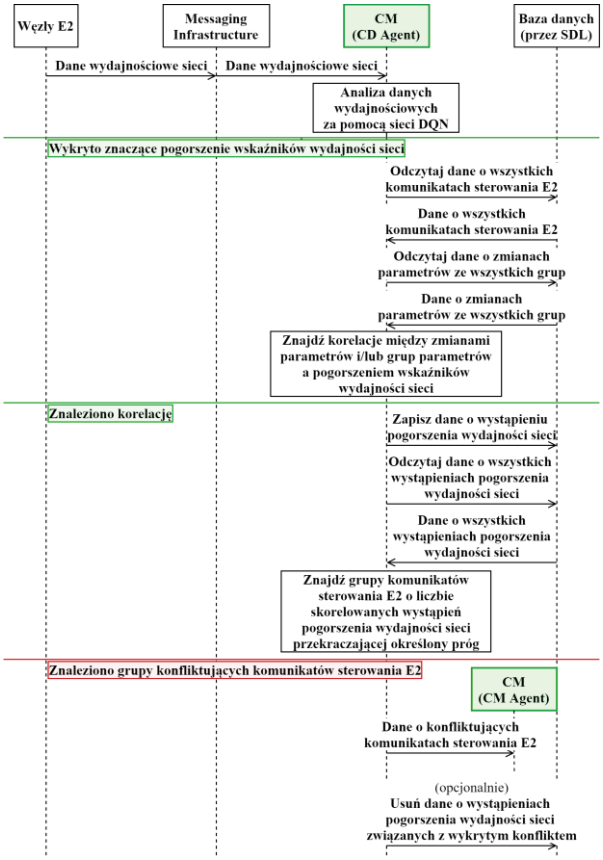
Rys. 3. Przepływ danych związany z wykrywaniem konfliktów pośrednich za pomocą komponentu ICD

jest możliwe – komponent Implicit Conflict Detection (ImCD) wykrywa konflikty już po zadziałaniu decyzji w sieci. W ramach ImCD przewiduje się komponent do analizowania danych wydajnościowych sieci raportowanych przez węzły E2 i wykrywania przypadków pogorszenia wydajności sieci; rolę tę spełnia sieć neuronowa Deep Q-network (DQN). Zakłada się, że sieć DQN jest trenowana w taki sposób, aby skutecznie wykrywać znaczące pogorszenie wskaźników wydajności sieci, ale szczegóły jej implementacji są poza zakresem tego artykułu. ImCD zapisuje dane o wystąpieniach degradacji wydajności sieci i skorelowanych decyzji aplikacji xApp („RAN KPI degradation occurences" na rysunku 1).

Działanie ImCD nie jest wyzwalane przez żaden komunikat sterowania E2, ale przez sieć DQN sygnalizującą wykrycie degradacji wydajności sieci. ImCD analizuje wtedy które parametry i/lub grupy parametrów zostały ostatnio zmodyfikowane przez wiele komunikatów sterowania E2; wykorzystuje w tym celu dane zarejestrowane przez DCD i ICD, więc nie musi monitorować przychodzących komunikatów. Jeśli liczba zaobserwowanych korelacji między decyzjami aplikacji xApp i pogorszeniem wydajności sieci, osiągnie określony próg, CDA dostarcza dane o konflikcie do CMA. Zakłada się, że na koniec procedury CDA może usunąć z bazy danych dane o wystąpieniach degradacji wydajności sieci związane ze zgłoszonym konfliktem. Przepływ komunikatów w ImCD przedstawiono na rysunku 4.

4. PODSUMOWANIE

W tym artykule zdefiniowano strukturę systemu wykrywania konfliktów CDF w sterowniku nRT-RIC sieci O-RAN. Opisany system jest w stanie wykrywać wszystkie trzy typy konfliktów opisane w specyfikacjach technicznych opublikowanych przez O-RAN Alliance, czyli konflikty bezpośrednie, pośrednie i niejawne. Proponowana struktura systemu działa poprzez śledzenie wszystkich komunikatów kontrolnych E2 wysyłanych przez aplikacje xApp uruchomionych w ramach inteligentnego sterownika radiowego nRT-RIC i monitorowanie danych dotyczących wydajności raportowanych przez węzły E2. Koncepcja CDF jest łatwo skalowalna i nie nakłada żadnych ograniczeń na sposób łagodzenia konfliktów. Zdefiniowanie struktury systemu wykrywania konfliktów w sieciach O-RAN jest postrzegane przez autorów artykułu jako pierwszy krok do opracowania znormalizowanej metody łagodzenia konfliktów w sterowniku nRT-RIC.



Rys. 4. Przepływ danych związany z wykrywaniem konfliktów niejawnych za pomocą komponentu ImCD

LITERATURA

[1] O-RAN Alliance. 2021. „Minimum Viable Plan and Acceleration towards Commercialization".

[2] O-RAN Working Group 1. 2022. „O-RAN Architecture Description v06.00".

[3] O-RAN Working Group 2. 2022. „Non-RT RIC & A1 Interface: Use Cases and Requirements v05.00".

[4] O-RAN Working Group 3. 2021. „Near-Real-time RIC: Use Cases and Requirements v01.00".

[5] Polese M., Bonati L., D'Oro S., Basagni S., Melodia T. 2022. „Understanding O-RAN".

[6] O-RAN Working Group 3. 2022. „Near-Real-time RIC Architecture v02.01".

[7] Zhang Han, Zhou Hao, Erol-Kantarci Melike. 2022. „Team Learning-Based Resource Allocation for Open Radio Access Network (O-RAN)".

ZASTOSOWANIE UCZENIA MASZYNOWEGO DO STEROWANIA SZYBKOŚCIĄ TRANSMISJI W SIECIACH STANDARDU IEEE 802.11

APPLICATION OF MACHINE LEARNING FOR RATE ADAPTATION IN IEEE 802.11 NETWORKS

Aleksandra Brytańczyk[1]; Katarzyna Kosek-Szott[2]

[1] AGH Akademia Górniczo-Hutnicza, Kraków, abrytanczyk@student.agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza, Kraków, katarzyna.kosek-szott@agh.edu.pl

DOI: 10.15199/59.2022.4.69

Streszczenie: **Niniejszy artykuł przedstawia sposób adaptacji szybkości transmisji danych do warunków panujących w sieci z użyciem uczenia maszynowego. Zaproponowane rozwiązanie rozszerza działanie istniejącego algorytmu dla przypadku, kiedy stacja odbiorcza znajduje się poza zasięgiem stacji nadawczej. Wartości przepustowości uzyskiwane z użyciem zaproponowanego rozwiązania są porównywalne z wynikami uzyskiwanymi dla tradycyjnych algorytmów Minstrel i CARA.**

Abstract: **This paper describes how to adapt data transmission rates to the varying network conditions using machine learning. The proposed algorithm is based on an earlier state-of-the-art solution and extends its operation for the case when the receiver is outside the range of the transmitter. The throughput values obtained with the use of the proposed algorithm are comparable to the results obtained with the traditional Minstrel and CARA algorithms.**

Słowa kluczowe: **IEEE 802.11, sterowanie szybkością transmisji, uczenie maszynowe, uczenie ze wzmocnieniem**

Keywords: **IEEE 802.11, machine learning, rate adaptation, reinforcement learning**

1. WSTĘP

Sterowanie szybkością transmisji w sieciach IEEE 802.11 może bazować na różnego rodzaju pomiarach, w celu dostosowania szybkości transmisji danych do zmiennych warunków sieciowych. Algorytmy dostępne w literaturze korzystają najczęściej z informacji dotyczących przepustowości, prawdopodobieństwa udanej transmisji, bitowej (ang. *Bit Error Rate*, BER) oraz ramkowej stopy błędów, stosunku sygnału do szumu (ang. *Signal to Noise Ratio*, SNR) [3,4,6-9]. Do tradycyjnych metod sterowania szybkością transmisji należą, m.in., algorytmy Minstrel [5], ARF (*Automatic Rate Fallback*) [10] i CARA (*Collision-Aware Rate Adaptation*) [7]. Minstrel opiera swoje działanie na pomiarze przepustowości; ARF bazuje na liczbie bezpośrednio następujących po sobie udanych i nieudanych transmisji; CARA wprowadza do algorytmu ARF sprawdzenie czy powodem nieudanej transmisji była kolizja czy błąd będący skutkiem zaszumienia kanału lub efektu wielodrogowości, w przypadku kolizji, licznik nieudanych transmisji (który wpływa na wybór szybkości transmisji danych) nie jest zwiększany. Oprócz rozwiązań tradycyjnych, w literaturze pojawiły się ostatnio rozwiązania oparte na uczeniu maszynowym, realizowanym w czasie rzeczywistym [3,4] oraz w oparciu o zbiór danych historycznych [9]. Większość z nich opiera się na uczeniu ze wzmocnieniem [3,4]. W niniejszym artykule skupiono się na poprawie działania algorytmu zaproponowanego w [4], który bazuje na informacjach dostępnych lokalnie na urządzeniu i nie wprowadza dodatkowych narzutów. Zauważono, że w przypadku, gdy stacja nadawcza znajduje się poza zasięgiem stacji odbiorczej, algorytm ten nie przerywa prób transmisji [4]. W zaproponowanym rozwiązaniu (por. rozdział 4), nadawca uczy się rozpoznawać ten stan i na niego reagować (por. rozdział 5).

2. PRZEGLĄD LITERATURY

W literaturze występują rozwiązania do sterowania szybkością transmisji stosujące uczenie maszynowe. Algorytm *Dynamic* ARF [9] wprowadza zmienne wartości graniczne dla liczników udanych i nieudanych transmisji algorytmu ARF. Dodatkowo, korzysta z modelu MLP (*Multilayer Perceptron*) w celu wyznaczenia zależności pomiędzy aktualnymi warunkami sieciowymi (BER oraz intensywnością ruchu), a wartością graniczną tych liczników. Algorytm EDRA (*Experience Driven Rate Adaptation*) [3] bazuje na głębokim uczeniu ze wzmocnieniem. Jako stan, oprócz typowych cech wpływających na szybkość transmisji, wybrano informacje dotyczące: współczynnika utraty pod-ramek, wskaźnika mocy odbieranego sygnału i czas obsługi ramek. Uczenie algorytmu polega na poruszaniu się w trójwymiarowym układzie współrzędnych reprezentowanym przez: poziom MCS (*Modulation and Coding Scheme*), tryb MIMO (*Multiple Input Multiple Output*), szerokość pasma. Dla każdej z wybranych akcji mierzona jest przepustowość. Algorytm zaproponowany w [4] (nazywany dalej algorytmem Cho) opiera się na algorytmie *Q-learning* [2]. Każda ze stacji implementuje własnego agenta, który uczy się w czasie rzeczywistym. Głównym parametrem, na którym bazuje algorytm jest wartość okna współzawodnictwa (ang. *Contention Window*, CW). Obserwowany stan jest obliczany z aktualnej wartości CW zgodnie ze wzorem:

$$s = \log_2(CW + 1) - 4 \tag{1}$$

W przypadku niepowodzenia transmisji, wartość CW jest zwiększana, zgodnie z działaniem algorytmu *backoff* [1] standardu IEEE 802.11. Nieudana transmisja może być spowodowana zarówno przez wystąpienie kolizji jak i przez wybranie niedostosowanego poziomu MCS.

W przypadku transmisji pomiędzy jedną parą nadawca-odbiorca, zakładaną w [4], kolizje nie występują. Zwiększanie CW wynika jedynie ze zmiennej odległości pomiędzy nadawcą, a odbiorcą. Wyższa wartość CW, oznacza pogorszenie się wartości SNR i odpowiada wyższej wartości stanu *s*. Taka obserwacja powinna spowodować zmniejszenie wybieranego poziomu MCS. Natomiast sytuacja odwrotna – jego zwiększenie. Liczba otrzymanych potwierdzeń ramek z danymi stanowi nagrodę w algorytmie. Natomiast akcją jest fizyczna szybkość transmisji korespondująca z ustawionym poziomem MCS. Szczegółowe działanie algorytmu Cho oraz zaproponowane modyfikacje są przedstawione w rozdziale 4.

3. PROPONOWANE MODYFIKACJE ALGORYTMU CHO

Zaobserwowano, że działanie algorytmu Cho jest niezadawalające w porównaniu do algorytmów Minstrel oraz CARA dla odległości bliskiej maksymalnemu zasięgowi transmisji danych [4]. Dodatkowo, po jej przekroczeniu, algorytm Cho wybiera głównie trzeci i czwarty poziom MCS, zamiast najniższy (równy 0). Problem ten można rozwiązać poprzez rozszerzenie stanu o dodatkową informację dotyczącą możliwości poprawnego wysłania ramki $s' \epsilon \{0,1\}$. W zmodyfikowanym algorytmie Cho (nazywanym dalej algorytmem Cho*), $s' = 1$ oznacza, że stacja nadawcza jest w zasięgu odbiorcy i $s' = 0$ oznacza, że jest poza jego zasięgiem. W przypadku, gdy $s' = 0$, algorytm ustawia najniższy poziom MCS i przestaje się douczać. Informację o pozostawianiu w zasięgu punktu dostępowego (odbiorcy), stacja określa za pomocą obserwacji ramek *beacon*, które są wysyłane cyklicznie z najniższym poziomem MCS. Dodatkową modyfikacją algorytmu Cho jest mechanizm całkowitego zaprzestania transmisji. Jeżeli dana stacja wykryje, że znajduje się poza zasięgiem punktu dostępowego przez czas β (odpowiadający np. trzykrotnemu interwałowi pomiędzy kolejnymi ramkami *beacon*), stacja ta zaprzestaje generowania nowych ramek danych oraz kolejka urządzenia zostaje opróżniona. W momencie ponownego wykrycia ramki *beacon*, możliwość transmisji zostaje wznowiona (ramki danych ponownie są generowane) i algorytm ponownie doucza się.

4. ZMODYFIKOWANY ALGORYTM CHO

Na początku działania algorytmu, wartości w tablicy *Q-table*, w której stan (*s*) reprezentowany jest tylko przez komponent dany wzorem (1), zostają wypełnione zerami, a współczynnik eksploracji ϵ przyjmuje wartość 1. Następnie powtarzane są kolejne czynności. Sprawdzana jest informacja o możliwości poprawnego wysłania ramki. Jeżeli wynosi ona 0, wybierana jest najniższa akcja i kolejka urządzenia zostaje opróżniona. W takim przypadku dalsze działania algorytmu dla tego kroku nie są wykonywane. Natomiast jeżeli wynosi 1 to kolejne kroki są takie same jak w klasycznym algorytmie Cho. Stosowana jest polityka ϵ-*greedy* [2], która pozwala na szybszą eksplorację środowiska na początku działania algorytmu oraz pomaga uniknąć problemu osiągnięcia lokalnego minimum, poprzez wybór losowej akcji lub obliczonej ze wzoru:

$$a = argmax_a(Q_{krok}(s,a)) \tag{2}$$

gdzie $argmax_a$ jest funkcją, która zwraca akcję (a), w przypadku której dla aktualnego stanu (s) wartość w tablicy *Q-table* w danym kroku algorytmu jest największa.

Po zastosowaniu wybranej akcji, pobierana jest nagroda (r) i następny stan (s_n). Następnie obliczana jest nowa wartość w tablicy *Q-table* według wzoru:

$$Q_{krok+1}(s,a) = (1-\alpha)Q_{krok}(s,a) + \\ +\alpha[r + \gamma max_{a_n} Q_{krok}(s_n, a_n)] \tag{3}$$

gdzie max_{a_n} zwraca największą wartość *Q-table* dla podanego stanu s_n, natomiast a_n jest akcją, dla której ta wartość jest przyjmowana. Parametr α jest szybkością uczenia się, natomiast γ współczynnikiem dyskontowym.

Ostatnim działaniem, wykonywanym w danym kroku algorytmu, jest pomniejszenie wartości parametru ϵ według wzoru (4), o ile jest on większy od wartości minimalnej (ϵ_{min}).

$$\epsilon = \epsilon * \epsilon_{decay} \tag{4}$$

gdzie ϵ_{decay} jest współczynnikiem eksploracji.

*Tab. 1. Stan, nagroda, akcja dla algorytmu Cho**

Stan	[logarytm z CW (*s*), informacja dotycząca możliwości poprawnego wysłania ramki (*s'*)]
Akcja	Poziom MCS
Nagroda	Liczba otrzymanych potwierdzeń

Tab. 2. Parametry modelu uczenia maszynowego

Symbol	*Parametr*	*Wartość*
α	szybkość uczenia	0,75
γ	współczynnik dyskontowy	0,95
ϵ_{decay}	współczynnik eksploracji	0,9999
ϵ_{min}	minimalna wartość współczynnika ϵ	0,01
β	czas od ostatniej ramki *beacon* [s]	0,3

W tabeli 1 zebrano definicje stanu, nagrody i akcji w algorytmie Cho*. Tabela 2 zawiera wartości hiper-parametrów modelu. Wartości te były dobrane w oparciu o [4] oraz wyniki dodatkowych badań. Przy wartości współczynnika eksploracji (0,9999), pomimo gorszych wyników dla pierwszego epizodu (spowodowanych większą eksploracją środowiska), ostateczne wyniki były porównywalne lub lepsze z wynikami dla mniejszych wartości tego współczynnika (0,99 i 0,999). Zbadano też wpływ wartości parametrów α i γ (0,75, 0,85 i 0,95), jednak otrzymane wyniki były porównywalne.

5. WYNIKI BADAŃ

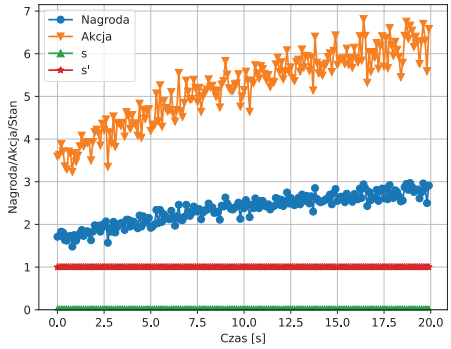
Badania przeprowadzono w symulatorze ns-3.29 [8] z rozszerzeniem ns3-gym [6] dla sieci składającej się z punktu dostępowego i stacji działającej w standardzie IEEE 802.11a. W celu sprawiedliwego porównania działania algorytmów Cho i Cho*, wzorując się na [4], ustawiono model propagacji *TwoRayGround*, wysokość anten równą 1,5 m, pasmo 5 GHz, założono brak zewnętrznych interferencji i ruch UDP o intensywności 60 Mb/s w kierunku punktu dostępowego. W ramach badań rozważono trzy scenariusze: stacjonarny oraz dwa niestacjonarne.

Scenariusz stacjonarny i niestacjonarny 1 są odtworzone z [4]. Natomiast scenariusz niestacjonarny 2 pozwala na zbadanie wpływu wprowadzonych modyfikacji, gdy stacja powraca do zasięgu punktu dostępowego.

5.1. Scenariusz stacjonarny

W tym scenariuszu odległość pomiędzy punktem dostępowym i stacją jest stała i wynosi 10 m, a czas symulacji (czyli jednego epizodu uczenia się algorytmu) wynosi 20 s. Ten prosty przypadek pozwala na weryfikację podstawowych założeń algorytmu.

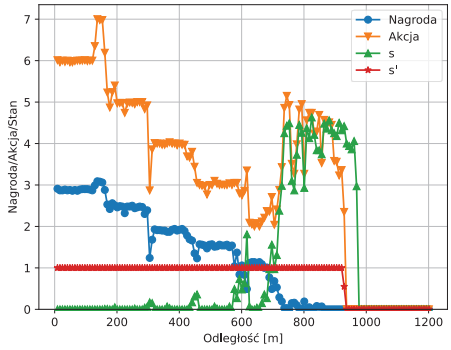
Jeżeli tylko jedno urządzenie nadaje, a odległość do odbiorcy pozwala na transmisję z każdym poziomem MCS, wszystkie pakiety powinny być poprawnie dostarczone. CW powinno ciągle utrzymywać minimalną wartość, czyli wartość stanu s powinna być równa zeru. Otrzymane wyniki prawidłowo zweryfikowały te założenia (por. rysunek 1) i były analogiczne do tych z [4].



Rys. 1 Nagroda, akcja i stan dla pierwszego epizodu algorytmu Cho dla scenariusza stacjonarnego.*

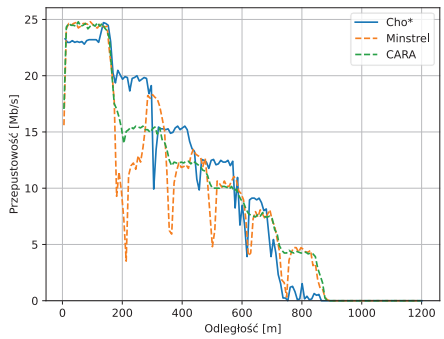
5.2. Scenariusz niestacjonarny 1

W tym scenariuszu wprowadzono mobilność urządzeń – stacja oddala się od punktu dostępowego z prędkością 80 m/s (podobnie jak w [4]), a jedna symulacja (epizod) trwa 15 s. Z powodu zadanej wysokości anteny (1,5 m) maksymalny zasięg transmisji jest równy około 900 m. Zdecydowano się pozostać przy tej wartości w celu przeprowadzenia analizy dla takich samych warunków jak przedstawionych w artykule [4].



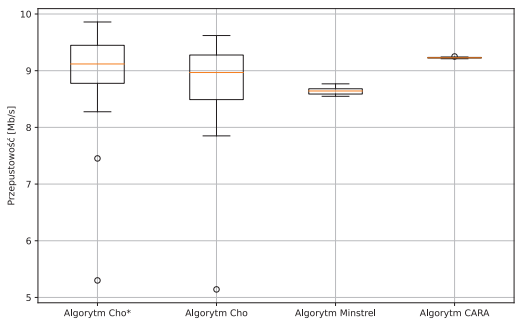
Rys. 2 Nagroda, akcja i stan dla 49 epizodu uczenia się dla algorytmu Cho. Scenariusz niestacjonarny 1.*

Na rysunku 2 przedstawiono wartości nagrody, akcji i stanu uzyskane przez algorytm Cho* w 49. epizodzie. Punkty przedstawione na wykresie odpowiadają średnim ze 100 kroków działania algorytmu. Można zauważyć charakterystyczne „schodki" występujące dla kolejno najkorzystniejszych poziomów MCS dla danej odległości. Niestety początkowo zamiast najwyższej akcji (MCS=7) wybierana jest akcja szósta (MCS=6). Jednak porównując przepustowości uzyskane w nowym algorytmie z wynikami dla algorytmów Minstrel [5] oraz CARA [7] (por. rysunek 3) widać, że uzyskiwane wartości przepustowości dla algorytmu Cho* nie są znacząco gorsze. Dodatkowo, obserwując kolejne akcje widać, że dla odległości większej od około 930 m wybierany jest poziom MCS=0 (ponieważ stacja wykrywa brak ramek *beacon*). Jednak CW utrzymuje niezerowe wartości aż do odległości około 970 m. Jest to najprawdopodobniej spowodowane próbami retransmisji ramek. Dla odległości do ok 700 m algorytm Cho* uzyskuje wyższe przepustowości niż CARA i Minstrel. Dodatkowo, przejścia pomiędzy kolejnymi wartościami przepustowości są mniej gwałtowne niż dla Minstrel i bardziej gwałtowne niż dla CARA (por. rysunek 3). Dla odległości wynoszącej ponad 700 m, przepustowości Minstrel oraz CARA są wyższe, co wskazuje na konieczność dalszego ulepszania algorytmu Cho.



Rys. 3 Wartości przepustowości uzyskane dla 49 epizodu uczenia się algorytmu Cho w porównaniu z algorytmami Minstrel i CARA. Scenariusz niestacjonarny 1.*

Rysunek 4 przedstawia wykres pudełkowy przepustowości średniej obliczonej z 20 symulacji (epizodów) dla algorytmów Cho*, Cho, Minstrel i CARA. Do średniej wliczane są przepustowości z całości trwania symulacji, również z chwil, gdy stacja jest już poza zasięgiem transmisji (tj., gdy ma zerową przepustowość).



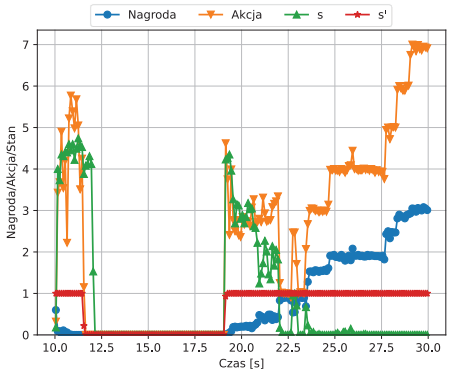
Rys. 4 Wykres pudełkowy średnich przepustowości dla algorytmów Cho, Cho, Minstrel i CARA. Scenariusz niestacjonarny 1.*

Algorytm Cho* charakteryzuje się większą zmiennością niż Minstrel i CARA ponieważ zebrane wyniki uwzględniają cały proces uczenia się algorytmu (tj. bez wiedzy *a priori*). Dodatkowo polityka ϵ-*greedy* zapewnia

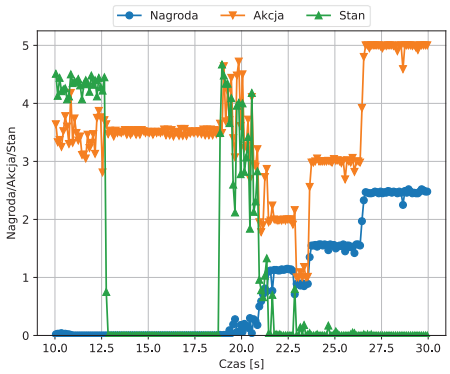
eksplorację środowiska, która również może powodować wybranie niekorzystnej akcji w danym kroku. Mediana uzyskiwanych wartości przepustowości dla algorytmów Cho oraz Cho* jest wyższa niż dla algorytmu Minstrel, i nieznacznie niższa niż dla algorytmu CARA. Wprowadzone modyfikacje pozytywnie wpłynęły na działanie algorytmu Cho. Algorytm Cho* charakteryzuje się mniejszym rozproszeniem danych oraz wyższą medianą niż oryginalny algorytm Cho.

5.3. Scenariusz niestacjonarny 2

Ten scenariusz zakłada wydłużenie czasu trwania symulacji do 30 s. Obejmuje on zatrzymanie oddalania się stacji po 15 s symulacji i ponowne jej przybliżanie do punktu dostępowego z tą samą prędkością (80 m/s) przez kolejne 15 sekund. Na rysunku 5 widać, że gdy stacja znajduje się ponownie w zasięgu punktu dostępowego (wartość *s'*=1), transmisja jest wznawiana i algorytm ponownie wybiera najlepszą, wg. jego obliczeń, akcję.



Rys. 5. Nagroda, akcja i stan dla algorytmu Cho. Scenariusz niestacjonarny 2.*



Rys. 6. Nagroda, akcja, stan dla algorytmu Cho. Scenariusz niestacjonarny 2.

Dla algorytmu Cho* obserwujemy niższe wartości CW niż dla algorytmu Cho (por. *s* na rysunkach 5 i 6), czyli więcej ramek jest poprawnie wysłanych bez potrzeby retransmisji. Dodatkowo, uzyskiwane wartości nagród i akcji są wyższe dla algorytmu Cho* niż dla algorytmu Cho, co wskazuje na jego lepsze działanie.

6. WNIOSKI

W artykule zaproponowano rozszerzenie algorytmu Cho [4]. Działanie zmodyfikowanego algorytmu Cho* jest porównywalne do działania standardowych algorytmów sterowania szybkością transmisji Minstrel i CARA. Zauważono jednak, że dobór szybkości transmisji wspomagany przez uczenie maszynowe cechuje się większą zmiennością niż w przypadku rozwiązań bez uczenia maszynowego. Trudno zatem jednoznacznie stwierdzić przewagę nad rozwiązaniami tradycyjnymi. Jako dalszy krok badań należy rozważyć bardziej skomplikowane scenariusze oraz dodanie kolejnych stanów (np. rozróżniające straty ramek wynikające z kolizji i błędów transmisyjnych). Dodatkowo należy rozszerzyć algorytm Cho* o możliwość wyboru kolejnych parametrów konfiguracyjnych, np. szybkości transmisji dostępnych w najnowszych rozszerzeniach standardu IEEE 802.11.

PODZIĘKOWANIA

Praca została sfinansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji naukowej (nr 16.16.230.434) dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH w Krakowie

LITERATURA

[1] Anderton William C., Trisha Chakraborty, Maxwell Young. 2021. „Windowed backoff algorithms for WiFi: theory and performance under." *Distributed Computing*, 34(5):367-393.

[2] Baeldung. „Epsilon-Greedy Q-learning". Dostęp 23.06.2022. https://www.baeldung.com/cs/epsilon-greedy-q-learning.

[3] Chen Syuan-Cheng, Chi-Yu Li, Chui-Hao Chiu. 2021. „An Experience Driven Design for IEEE 802.11 ac Rate Adaptation based on Reinforcement Learning." *IEEE INFOCOM 2021*.

[4] Cho Soohyun. 2020. „Rate adaptation with Q-learning in CSMA/CA wireless networks." *Journal of Information Processing Systems*, 16(5):1048-1063.

[5] Dokumentacja algorytmu Minstrel. Dostęp 23.06.2022. https://wireless.wiki.kernel.org/en/developers/Documentation/mac80211/RateControl/minstrel.

[6] Gawłowicz Piotr, Anatolij Zubow. 2019. „Ns-3 meets OpenAI Gym: The Playground for Machine Learning in Networking Research". *ACM MSWiM*. 113-120.

[7] Kim Jong-Seok, Seong-Kwan Kim, Sung-Hyun Choi, Daji Qiao. 2006. „CARA: Collision-Aware Rate Adaptation for IEEE 802.11 WLANs". *IEEE INFOCOM 2006*.

[8] Symulator sieciowy ns-3. Dostęp 23.06.2022. https://www.nsnam.org/.

[9] Wang Chiapin. 2013. „Dynamic ARF for throughput improvement in 802.11 WLAN via a machine-learning approach". *Journal of Network and Computer Applications*, 36(2):667– 676.

[10] Yin Wei, Peizhao Hu, Jadwiga Indulska, Marius Portmann, Ying Mao. 2020. „MAC-layer rate control for 802.11 networks: A survey". *Wireless Networks*, 26(5):3793–3830.

KLASYFIKATOR SVM W ZASTOSOWANIU DO SYNCHRONIZACJI SYGNAŁU OFDM ZNIEKSZTAŁCONEGO PRZEZ KANAŁ WIELODROGOWY

SVM CLASSIFIER IN APPLICATION TO SYNCHRONIZATION OF THE OFDM SIGNAL DISTORTED BY THE MULTIPATH CHANNEL

Marek Blok[1]; Karol Gmiński[2]; Patryk Miklaszewski[3]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, marek.blok@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, s171856@student.pg.edu.pl
[3] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, s172010@student.pg.edu.pl

DOI: 10.15199/59.2022.4.70

Streszczenie: **W pracy przedstawiono analizę przydatności klasyfikatora SVM bazującego na uczeniu maszynowym do estymacji przesunięcia czasowego odebranego symbolu OFDM. Przedstawione wyniki wykazują, że ten klasyfikator potrafi zapewnić synchronizację dla różnych kanałów wielodrogowych o wysokim poziomie szumu. Eksperymenty przeprowadzone w Matlabie z użyciem modeli modulatora i demodulatora wykazały, że w większości przypadków klasyfikator prawidłowo oceniał przesunięcie czasowe symbolu OFDM, nawet gdy był trenowany dla innego modelu kanału.**
Abstract: **The paper presents an analysis of capabilities of machine learning based SVM classifier to estimate the OFDM symbol time offset at the receiver. The presented results demonstrate classifier's ability to provide synchronization for various multipath channel models with high noise levels. The research was conducted using OFDM modulator and demodulator models implemented in Matlab. It has been shown that in most cases the classifier was able to determine the time shift of the OFDM symbol with high confidence even if it was trained for a different channel model.**

Słowa kluczowe: **klasyfikator SVM, OFDM, prefiks cykliczny, synchronizacja symbolowa.**

Keywords: **cyclic prefix, OFDM, SVM classifier, symbol synchronization.**

1. WSTĘP

OFDM (Orthogonal Frequency-Division Multiplexing) jest bardzo popularną metodą multipleksacji częstotliwościowej, wykorzystywaną między innymi w technologii dostępowej ADSL, sieciach bezprzewodowych 802.11, telefonii mobilnej 4G/LTE oraz 5G jak i w systemach telewizji cyfrowej DVB-T oraz DVB-T2 [3, 5]. Zastosowanie techniki OFDM upraszcza uzyskanie wysokich przepływności strumienia danych poprzez rozdzielenie danych pomiędzy wiele mniejszych strumieni przesyłanych z wykorzystaniem ortogonalnych podnośnych. Użycie w OFDM ortogonalnych podnośnych powoduje zwiększenie efektywności widmowej transmisji i uzyskanie wyższych szybkości transmisji. Dodatkowo podejście to pozwala na znaczące zredukowanie wpływu zniekształceń wynikających z propagacji wielodrogowej i odbiór sygnału transmitowanego przez kanał cechujący się wysokim poziomem ISI (*Intersymbol Interference*) [1, 5].

Prawidłowy odbiór sygnału OFDM wymaga poprawnej synchronizacji symbolowej, która istotna jest również w odniesieniu do minimalizacji wpływu ISI. Synchronizacja symbolu OFDM może być uzyskana na różne sposoby. Ze względu na złożoną postać symbolu OFDM zwykle wykorzystuje się do tego dodatkowe elementy takie jak sygnały pilotowe, czy prefiks cykliczny [1, 5]. Kanał wielodrogowy sprawia, że odebrany sygnał jest sumą wielu sygnałów nadanych różniących się opóźnieniem, co skutkuje rozmyciem sygnału obserwowanym jako ISI, co utrudnia prawidłową ocenę przesunięcia czasowego odebranego symbolu OFDM. W związku z tym autorzy postanowili do tego zadania zastosować klasyfikator bazujący na sieci neuronowej i zweryfikować jego przydatność do tego zastosowania. Jak udało się wykazać w ramach przeprowadzonych symulacji, zastosowany klasyfikator pozwalana na skuteczne oszacowanie opóźnienia symbolu OFDM nawet dla silnych interferencji międzysymbolowych, jednak użycie sieci neuronowej utrudnia ocenę czy podejmowana decyzja bazuje na cyklicznym prefiksie dodawanym do symboli OFDM, czy też na innych cechach przetwarzanego sygnału.

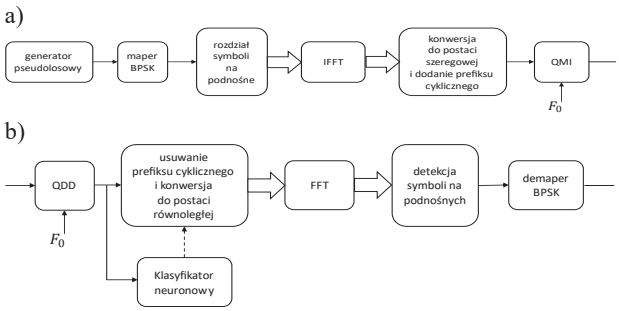
2. PROPONOWANE ROZWIĄZANIE

Zaproponowane rozwiązanie synchronizacji symbolowej zakłada uzupełnienie odbiornika OFDM [1, 4, 5] o klasyfikator wykorzystujący uczenie maszynowe. W symulowanym systemie zakładamy, że odebrany sygnał jest przekształcany z użyciem demodulatora kwadraturowego do postaci zespolonego sygnału podstawowo-pasmowego. Następnie klasyfikator SVM (*Support Vector Machines*) [2] dostępny w środowisku Matlab jest wykorzystywany do oceny przesunięcia czasowego bieżącego segmentu odebranego sygnału, co pozwala na realizację synchronizacji symbolowej nawet, jeżeli odbierany sygnał jest zniekształcony przez zaszumiony kanał wielodrogowy. Na potrzeby treningu klasyfikatora oraz jego testów konieczne było również zrealizowanie generatora podstawowo pasmowego sygnału OFDM.

2.1. Generacja i demodulacja sygnału OFDM

Na rys. 1 przedstawiono ogólny schemat nadajnika (a) oraz odbiornika (b) sygnału OFDM [1, 5]. Istotną cechą OFDM jest, to że generacja oraz demodulacja sygnału

OFDM może być zrealizowana z użyciem wydajnego numerycznie algorytmu IFFT/FFT. Poza tym elementem odpowiedzialnym za umieszczanie/wydobywanie symboli na/z podnośnych, po stronie nadawczej (rys. 1a) mamy konwersję ciągu binarnego na symbole konstelacji podstawowej modulacji cyfrowej używanej w podkanałach (tutaj dla uproszczenia zastosowano BPSK), uzupełnianie symbolu OFDM o prefiks cykliczny oraz ulokowanie tak wygenerowanego sygnału podstawowo-pasmowego na zakładanej częstotliwości nośnej F_0 (QMI – *quadrature modulation with interpolation*). Z kolei, po stronie odbiorczej odwracamy ten proces zaczynając od demodulacji do pasma podstawowego (QDD – *quadrature decimation with demodulation*) z filtracją decymacyjną usuwającą jednocześnie zakłócenia pozapasmowe. Następnie konieczna jest identyfikacja początku symbolu OFDM pozwalająca na usunięcie prefiksu cyklicznego a następnie wydobycie symboli z użyciem FFT.



Rys. 1. Nadajnik (a) i odbiornik (b) OFDM

2.2. Synchronizacja symbolowa

W zaproponowanym rozwiązaniu zakładamy, tak jak to pokazano na rys. 1, że symbole generowanego sygnału OFDM są uzupełniane o prefiks cykliczny. Dodanie prefiksu cyklicznego jest z jednej strony wykorzystywane do ograniczenia wpływu ISI, ułatwiając wydobycie po stronie odbiorczej segmentu symbolu o ograniczonym wpływie ISI na ortogonalność w obrębie symbolu. Z drugiej strony prefiks cykliczny przenosi dodatkową informację o przesunięciu czasowym symbolu OFDM, powiązaną z tym, że początek każdego przedłużonego symbolu jest kopią jego końcowej części [1, 5].

Można przypuszczać, że zastosowany klasyfikator uczy się przede wszystkim identyfikować przesunięcie czasowe prefiksu cyklicznego w obserwowanym segmencie sygnału. Ponieważ długość segmentu sygnału podawanego na wejście klasyfikatora może być różna, w ramach badań sprawdzono również wpływ jego długości na dokładność klasyfikacji. Aby zagwarantować, że niezależnie od opóźnienia w analizowanym fragmencie będzie zawarty przynajmniej jeden kompletny symbol OFDM zawierający prefiks cykliczny i odpowiadający mu oryginalny fragment symbolu, przyjęto, że najmniejsza długość okna równoważną dwóm okresom symbolowym OFDM.

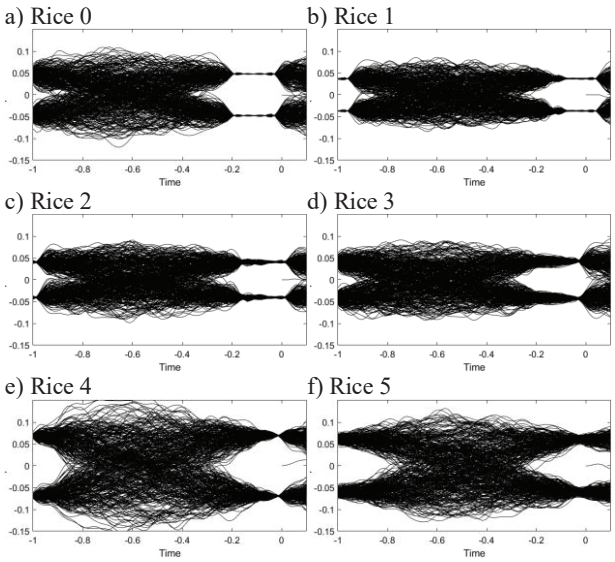
2.3. Model kanału

Na potrzeby treningu klasyfikatora oraz jego testów przyjęto kanał Rice'a, który jest wielodrogowym modelem kanału z zanikami, z addytywnym gaussowskim szumem białym. Na potrzeby testów założono, że sygnał dociera ścieżkami o pięciu różnych opóźnieniach i średnich tłumieniach. Poza referencyjnym kanałem z pojedynczą ścieżką (rys. 2a, dalej oznaczany jako Rice 0), przyjęto pięć różnych modeli kanałów.

Na rys. 2 pokazano wykresy oczkowe zastosowanych modeli kanałów dla wysokiego SNR w kanale. W efekcie zniekształceń wielodrogowych transmitowany sygnał ulega rozproszeniu czasowemu co skutkuje zamykaniem się oka (*eye diagram*) i wymaga dodatkowej korekty w ramach synchronizacji symbolowej. Dla modeli Rice 1 (rys. 2b), Rice 2 (rys. 2c) oraz Rice 3 (rys. 2d) przyjęto następujące średnie tłumienia ścieżek: 0, -10, -20, -25 oraz -30dB. W tab. 1 zestawiono wielkość opóźnień ścieżek modelu podanych jako procent czasu trwania prefiksu cyklicznego. Tylko dla modelu Rice 1 opóźnienia te nie przekraczają czasu trwania prefiksu, ale ze względu na duże tłumienie kolejnych ścieżek możliwy jest odbiór praktycznie pozbawiony ISI (rys. 2b, c i d). Dwa kolejne modele Rice 4 (rys. 2e) oraz Rice 5 (rys. 2f) cechują nieduże średnie tłumienia ścieżek: 0, -2, -4, -5 oraz -6dB, stąd też rozmycie wykresów oczkowych jest znacząco wyraźniejsze. Model Rice 4 cechują stosunkowo nieduże opóźnienia, stąd i w tym przypadku możliwy jest odbiór pozbawiony ISI (rys. 2e). Z kolei, dwie ostatnie ścieżki modelu Rice 5 cechuje opóźnienie przekraczające wielkość prefiksu cyklicznego i chociaż w przypadku braku addytywnego szumu w kanale możliwy jest prawidłowy odbiór, to jednak zniekształcenia ISI są nieuniknione (rys. 2f). W tym przypadku zadaniem synchronizacji symbolowej jest minimalizacji tych zniekształceń.

Tab. 1. Opóźnienia ścieżek modeli kanałów; % CP

Model	*1.*	*2.*	*3.*	*4.*	*5.*
Rice 1	0	24	48	72	98
Rice 2	0	35	69	115	172
Rice 3	0	104	111	118	125
Rice 4	0	52	66	82	104
Rice 5	0	78	99	123	156



Rys. 2. Wykresy oczkowe obserwowane dla demodulacji podkanału OFDM dla zastosowanych modeli kanałów

2.4. Dane treningowe i klasyfikator neuronowy

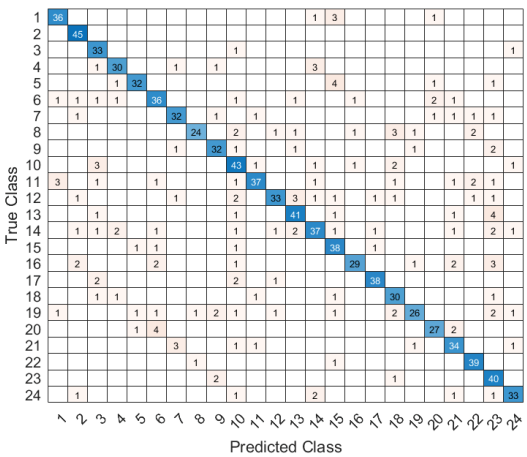
Jak wspomniano wcześniej, do rozwiązania problemu synchronizacji symbolowej sygnału OFDM proponujemy zastosować klasyfikator neuronowy. Przygotowanie, trening i walidację klasyfikatora pozwalającego na

ocenę przesunięcia czasowego symbolu OFDM wykonano z wykorzystaniem narzędzi dostępnych w środowisku Matlab. W ramach wstępnych analiz do dalszych badań wytypowano nieliniowy klasyfikator SVM.

W ramach eksperymentów przyjęto, że generowany sygnał OFDM przenosi 20 podkanałów oraz 20% prefiks cykliczny, co daje surowy symbol OFDM o długości 24 próbek. Przyjęto, że na wejście klasyfikatora podajemy część rzeczywistą obwiedni zespolonej zdemodulowanego sygnału, czyli sygnału wyjściowego QDD (rys. 1b).

Dane treningowe oraz walidacyjne przygotowano dla losowo generowanego strumienia binarnego podawanego na wejście modulatora OFDM (rys. 1a), sygnał wyjściowy był przepuszczany przez model kanału (p.2.3) a po dodaniu szumu białego dla zadanego SNR przepuszczany w demodulatorze przez QDD (rys. 1b). Tak pozyskany sygnał był cięty na segmenty o długości równej czasowi trwania 2, 5 lub 9 symboli OFDM. Początek każdego kolejnego wycinanego segmentu dobierano losowo, a wielkość tego przesunięcia była zachowywana jako etykieta na potrzeby treningu oraz walidacji.

Założono tu, że zadaniem klasyfikatora będzie ocena przesunięcia czasowego symbolu OFDM z dokładnością do jednej próbki, co przekłada się na 24 klasy wyjściowe klasyfikatora. Na rys. 3 przedstawiono przykładową macierz błędu (*Confusion Matrix*) uzyskaną dla treningu klasyfikatora na podstawie sygnału przesyłanego przez model kanału Rice 3. Jak widać przesunięcie czasowe dla większości ramek z sygnałem OFDM zostało w tym przypadku prawidłowo zaklasyfikowane (segmenty wyróżnione niebieskim kolorem), pozostałe pola sygnalizują błędną klasyfikację przesunięcia czasowego sygnału.



Rys. 3. Macierz błędu dla klasyfikatora trenowanego na podstawie sygnału przesyłanego przez kanał Rice 3.

3. WYNIKI

W punkcie tym przedstawiono wyniki eksperymentów służących ocenie przydatności klasyfikatora SVM do synchronizacji symbolowej w OFDM.

3.1. Wyniki treningu klasyfikatora zależnie od długości analizowanego segmentu sygnału

W ramach eksperymentów zweryfikowano dokładność klasyfikacji trenowanych klasyfikatorów (procent analizowanych segmentów, dla których w ramach walidacji dokonano bezbłędnej klasyfikacji przesunięcia symbolu OFDM wewnątrz analizowanego segmentu) zależnie do liczby symboli OFDM (2, 5 oraz 9) znajdujących się w segmencie analizowanym przez klasyfikator dla różnych wartości SNR (od 5 do 30dB). Dodatkowo zbadano jak dokładność klasyfikacji zależy od liczby segmentów zawartych w zestawie treningowym (od 5 do 80 tysięcy segmentów). W tab. 2-5 przedstawiono wyniki uzyskane dla kanału Rice 3, przy czym dla pozostałych modeli zaobserwowano podobne efekty.

Tab. 2. Dokładność klasyfikacji [%] kanału Rice 3 i użycia 2 symboli OFDM na segment

liczba segmentów	SNR [dB]				
	5	10	15	20	30
5k	4,3	7,5	22,3	52,4	76,8
10k	4,5	10,8	32,8	71,5	91,3
20k	5,2	11,7	43,6	83,0	97,0
40k	5,8	15,1	51,0	89,3	98,7
60k	6,0	16,8	53,7	91,3	99,3
80k	6,1	16,9	56,5	92,3	99,5

Tab. 3. Dokładność klasyfikacji [%] kanału Rice 3 i użycia 5 symboli OFDM na segment

liczba segmentów	SNR [dB]				
	5	10	15	20	30
5k	4,0	5,3	13,2	33,0	56,0
10k	5,2	8,3	32,5	73,5	92,7
20k	5,2	12,8	60,3	95,5	99,6
40k	5,7	21,2	80,5	99,1	99,9
60k	6,5	25,5	86,1	99,7	100,0
80k	7,0	27,8	88,1	99,8	100,0

Tab. 4. Dokładność klasyfikacji [%] kanału Rice 3 i użycia 9 symboli OFDM na segment

liczba segmentów	SNR [dB]				
	5	10	15	20	30
5k	4,0	5,8	7,9	17,3	28,5
10k	4,8	5,5	16,3	42,1	72,7
20k	4,5	9,9	45,3	90,7	99,1
40k	5,1	19,5	84,7	99,8	100,0
60k	6,3	28,0	93,4	100,0	100,0
80k	6,7	33,9	96,4	100,0	100,0

Zgodnie z oczekiwaniem zwiększanie liczby symboli w segmencie pozwala na uzyskanie lepszych wyników, co jest przede wszystkim istotne przy większym poziomie szumu w kanale. Dla dwóch symboli OFDM w analizowanym segmencie sieć niezależnie od długości treningu nie potrafi przekroczyć istotnie 50% trafnych decyzji dla SNR równego 15dB. Zwiększenie liczby symboli OFDM w segmencie do 9-ciu pozwala na uzyskanie w tym przypadku 96% dokładności, a przy SNR równego 10dB uzyskać 33% dokładność. Uzyskanie takiej dokład-

ności wymaga jednocześnie zwiększania wielkości zestawu treningowego. Warto tutaj zauważyć, że jeżeli średnie opóźnienie kanału nie ulega gwałtownym zmianom, to pracę klasyfikatora można uzupełnić o mechanizmy śledzenia bieżącego opóźnienia sygnału, co może pozwolić na skuteczną synchronizację symbolową nawet przy mniejszej dokładności klasyfikacji.

3.2. Walidacja dla niedopasowanego kanału

W kolejnym eksperymencie zweryfikowano, na ile skutecznie klasyfikator wytrenowany na wybranym modelu kanału potrafi klasyfikować przesunięcie segmentu sygnału, który odebrano na wyjściu innego modelu kanału. Tab. 5-9 przedstawia wyniki uzyskane dla weryfikacji modeli wytrenowanych dla wszystkich pięciu przyjętych modeli kanałów (trenowanych dla SNR = 20dB). We wszystkich przypadkach walidację przeprowadzono na prostym kanale AWGN (Rice 0) oraz wszystkich pozostałych modelach dla różnych poziomów SNR.

Tab. 5. Dokładność [%] klasyfikatora trenowanego dla kanału Rice 1 i użycia 5 symboli OFDM na segment

Kanał	SNR [dB]					
	5	10	5	20	5	30
Rice 0	13,0	43,2	88,0	98,9	99,8	99,9
Rice 1	12,4	36,9	84,8	98,4	99,0	99,8
Rice 2	12,7	42,2	87,2	98,7	99,5	99,9
Rice 3	11,0	31,6	74,8	95,5	98,8	99,3
Rice 4	2,5	3,0	2,3	1,5	1,3	1,2
Rice 5	6,6	11,2	18,9	30,0	36,7	38,6

Tab. 6. Dokładność [%] klasyfikatora trenowanego dla kanału Rice 2 i użycia 5 symboli OFDM na segment

Kanał	SNR [dB]					
	5	10	15	20	25	30
Rice 0	12,0	38,3	86,8	98,0	99,4	99,7
Rice 1	11,1	36,8	79,5	97,0	99,3	99,4
Rice 2	14,3	45,7	90,2	98,8	99,8	100,0
Rice 3	9,2	30,4	75,7	93,2	98,0	98,6
Rice 4	3,7	1,2	1,1	0,5	0,3	0,1
Rice 5	5,9	8,6	15,9	21,3	29,8	28,0

Tab. 7. Dokładność [%] klasyfikatora trenowanego dla kanału Rice 3 i użycia 5 symboli OFDM na segment

Kanał	SNR [dB]					
	5	10	15	20	25	30
Rice 0	13,0	44,0	89,6	98,9	99,9	100,0
Rice 1	11,2	33,1	79,4	98,5	99,5	99,5
Rice 2	14,4	40,3	86,1	98,6	99,7	99,7
Rice 3	12,9	32,7	83,9	98,0	99,6	99,8
Rice 4	5,6	3,9	2,0	1,9	1,8	0,8
Rice 5	7,0	10,4	19,1	31,4	38,4	39,3

Jak można zauważyć, klasyfikatory trenowane na sygnałach pochodzących z mniej zniekształconych kanałów (Rice 1-3) dobrze sobie radzą, jeżeli są zastosowane dla innych kanałów podobnego typu. Najbardziej obiecujący wydaje się w tym przypadku klasyfikator trenowany na modelu Rice 3, który dla SNR nie mniejszego od 15dB dla kanałów Rice 0-3 osiąga dokładność rzędu 80% lub lepszą. Z drugiej strony kanały, w których ISI nie można skompensować zastosowanym CP dobre efekty można uzyskać jedynie dla klasyfikatora trenowanego dla danego modelu.

Tab. 8. Dokładność [%] klasyfikatora trenowanego dla kanału Rice 4 i użycia 5 symboli OFDM na segment

Kanał	SNR [dB]					
	5	10	15	20	25	30
Rice 0	8,2	14,9	31,7	47,0	51,8	56,4
Rice 1	6,9	10,0	22,0	36,0	39,2	40,5
Rice 2	6,6	10,4	15,9	24,2	29,7	31,1
Rice 3	6,0	11,0	21,6	31,7	37,3	39,9
Rice 4	3,9	4,5	3,8	3,6	1,8	1,8
Rice 5	12,1	34,6	80,8	97,3	98,7	99,6

Tab. 9. Dokładność [%] klasyfikatora trenowanego dla kanału Rice 5 i użycia 5 symboli OFDM na segment

Kanał	SNR [dB]					
	5	10	15	20	25	30
Rice 0	3,9	3,4	2,8	2,4	1,8	2,5
Rice 1	4,3	2,9	1,7	1,3	1,1	1,3
Rice 2	2,7	2,7	1,4	0,7	0,4	0,2
Rice 3	2,9	3,5	3,2	1,8	2,1	1,9
Rice 4	13,9	43,2	82,4	96,2	98,3	99,4
Rice 5	5,6	5,7	7,3	5,7	4,8	4,2

4. PODSUMOWANIE

W pracy przedstawiono wyniki badań możliwości zastosowania klasyfikatora SVM bazującego na uczeniu maszynowym do synchronizacji symbolowej w demodulatorze OFDM w obecności szumu i zniekształceń wielodrogowych. Wyniki pokazały, że dobre efekty można uzyskać dla kanałów, dla których ISI są ograniczane przez użyty prefiks cykliczny (CP). W takim przypadku wytrenowany klasyfikator można zastosować również dla odmiennego modelu kanału. Dla kanałów, dla których CP niewystarczająco eliminuje ISI klasyfikator SVM również dobrze sobie radził, ale w tym przypadku konieczne jest użycie klasyfikatora dedykowanego dla konkretnego modelu kanału.

LITERATURA

[1] Ćwikowski Łukasz, Blok Marek. 2009. Educational model of the OFDM modulator and demodulator. *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments*, SPIE, vol. 7502, pp. 485-490.

[2] Garcia M. G., Rojo-Álvarez J. L., Alonso-Atienza F., Martínez-Ramón M. 2006. „Support vector machines for robust channel estimation in OFDM. IEEE Signal Processing Letters, 13(7), 397-400.

[3] Khaizuran Abdullah, Hussain Zahir M. 2007. „Performance of Fourier-based and wavelet-based OFDM for DVB-T systems." *IEEE Australasian Telecommunication Networks and Applications Conference*, pp. 475-479.

[4] Kung Te-Lung, Parhi Keshab K. 2011. "Frequency Domain Symbol Synchronization for OFDM Systems", *IEEE International Conference on Electro/Information Technology*, pp. 1-5.

[5] Langton Charan. 2004. Orthogonal Frequency Division Multiplexing (OFDM) Tutorial, *Intuitive Guide to Principles of Communications*.

DYNAMICZNE WYGASZANIE PUNKTÓW TRANSMISYJNYCH W WIELOANTENOWEJ SIECI PIĄTEJ GENERACJI Z UŻYCIEM GŁĘBOKIEGO UCZENIA ZE WZMOCNIENIEM

DYNAMIC POINT BLANKING IN 5G MASSIVE MIMO NETWORK DRIVEN BY DEEP REINFORCEMENT LEARNING

Marcin Hoffmann [1]

DOI: 10.15199/59.2022.4.71

[1] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, marcin.ro.hoffmann@doctorate.put.poznan.pl

Streszczenie: **Dynamiczne wygaszanie punktów transmisyjnych to jedna z technik skoordynowanej transmisji z wielu punktów transmisyjnych, w której niektóre stacje bazowe mogą być tymczasowo wyciszone, np. w celu poprawy przepustowości użytkowników na obrzeżach komórki. W niniejszym artykule zaproponowano wykorzystanie głębokiego uczenia ze wzmocnieniem do wyboru schematu wygaszania, który zwiększa przepływność użytkowników znajdujących się na skraju komórki. Zaproponowany algorytm wykorzystuje sieć neuronową do wyboru schematu wygaszania na podstawie lokalizacji użytkowników. Badania symulacyjne wykazały, że dzięki zaproponowanemu rozwiązaniu przepływność użytkownika na krawędzi komórki możne wzrosnąć około 14,14 razy.**[1]

Abstract: **Dynamic Point Blanking (DPB) is one of the Coordinated MultiPoint (CoMP) techniques, where some Base Stations (BSs) can be temporarily muted, e.g., to improve the cell-edge users throughput. In this paper, it is proposed to obtain the muting pattern that improves cell-edge users throughput with the use of Deep Reinforcement Learning technique. The proposed algorithm utilizes deep neural network to select muting pattern on the basis of user locations. Simulation studies have shown that cell-edge user throughput can be improved by the ratio of about 14.14, while using the proposed algorithm.**

Słowa kluczowe: **5G, Dynamiczne Wygaszanie Punktów Transmisyjnych, Głębokie Uczenie ze Wzmocnieniem, Systemy Wieloantenowe**

Keywords: **5G, Deep Reinforcement Learning, Dynamic Point Blanking, Massive MIMO**

1. WSTĘP

W sieciach mobilnych Czwartej Generacji (4G) wprowadzone zostały techniki skoordynowanej transmisji z wielu punktów transmisyjnych CoMP (*ang. Coordinated Multi-Point*) [1]. Są one niezwykle istotne również dla nowoczesnych sieci Piątej Generacji (5G). Główną ideą stojącą za techniką CoMP jest inteligentna współpraca punktów transmisyjnych (tutaj stacji bazowych), w celu poprawy przepływności osiąganych przez użytkowników. W literaturze wyróżnionych zostało kilka schematów CoMP, np. wspólna transmisja JT (*ang. Joint Transmission*), lub dynamiczny wybór punktu transmisyjnego DPS (*ang. Dynamic Point Selection*) [2]. Schemat JT polega na jednoczesnej obsłudze użytkownika przez kilka stacji bazowych, natomiast w przypadku DPS, użytkownik jest obsługiwany przez pojedynczą stację bazową, ale może zostać błyskawicznie przepięty do innej. Dużo mniej zbadanym, schematem CoMP jest tzw. dynamiczne wygaszanie punktów transmisyjnych DPB (*ang. Dynamic Point Blanking*) [3]. DPB polega na chwilowym wygaszeniu wybranych stacji bazowych w całym pasmie lub na wybranej częstotliwości, np. w celu zmniejszenia interferencji i poprawy przepływności osiąganych przez użytkowników znajdujących się na skraju komórki [4][5]. Głównym wyzwaniem wynikającym z zastosowania DPB jest odpowiednie dobranie tzw. schematu wygaszania (*ang. Mutting Pattern*) do aktualnego stanu sieci, np. rozmieszczenia użytkowników na obszarze komórki. Odpowiedni wybór schematu wygaszania jest szczególnie trudny w sieci wykorzystującej transmisje wieloantenową M-MIMO (*ang. Massive Multiple-Input Multiple-Output*), z uwagi na interferencje pomiędzy użytkownikami zależne od korelacji przestrzennej ich kanałów radiowych [6].

W literaturze zaproponowane zostało kilka algorytmów wykorzystujących DPB w celu zmniejszenia interferencji pomiędzy komórkami [4][5]. W tych pracach schemat wygaszania ustalany jest na podstawie średniego poziomu mocy odbieranej w całym pasmie przez urządzenie końcowe od danej stacji bazowej. Takie podejście sprawdza się w rozważanych przez autorów tych prac scenariuszach, gdzie stacje bazowe wyposażone są w pojedyncze anteny. Nie pozwala jednak właściwie oszacować interferencji w przypadku systemów M-MIMO, w których kluczową role odgrywa korelacja przestrzenna pomiędzy kanałami radiowymi poszczególnych użytkowników. Dodatkowo przepływności osiągane przez użytkowników w rzeczywistych sieciach mogą być w dużym stopniu zależne od skomplikowanych algorytmów przydziału zasobów radiowych, oraz sterowania ruchem. Z tego względu zależność pomiędzy średnim poziomem mocy odbieranej przez urządzenie końcowe a osiąganą przez nie przepływnością jest skomplikowana i trudna w matematycznym opisie.

W związku z powyższymi niedoskonałościami standardowych algorytmów DPB w tej pracy zaproponowano wykorzystanie metod uczenia maszynowego, a w

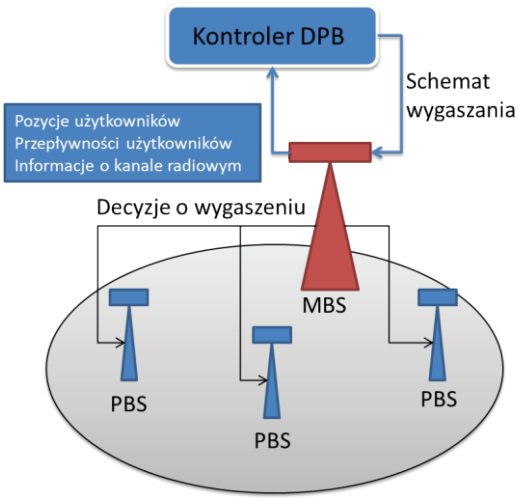
[1] Praca finansowana ze środków przyznanych przez Ministerstwo Edukacji i Nauki w ramach subwencji badawczej 0312/SBAD/8159

szczególności uczenia ze wzmocnieniem (*ang. Reinforcement Learning* [7]) w celu inteligentnego dostosowania schematu wygaszania do aktualnego stanu sieci. Wybrany schemat wygaszania powinien polepszyć przepływności użytkowników znajdujących się na skraju komórki. Podobnie jak w poprzednich pracach zaproponowany algorytm DPB będzie korzystał z informacji zależnej od położenia użytkowników w komórce, zapisywanej i przetwarzanej w tzw. Mapie Środowiska Radiowego REM (*ang. Radio Environment Map*) [8]. Nowością w tej pracy jest wykorzystanie tzw. Głębokiego Uczenia ze Wzmocnieniem (*ang. Deep Reinforcement Learning*) [9]. REM będzie przechowywał, trenował i wykorzystywał sieć neuronową głębokiego uczenia DNN (*ang. Deep Neural Network*) w celu estymacji przepływności osiąganych przez użytkowników znajdujących się na skraju komórki, i związanych z różnymi schematami wygaszania i przestrzennymi rozkładami użytkowników w komórce. Zaproponowane rozwiązanie zostanie zweryfikowane i porównane ze standardowymi algorytmami DPB za pomocą symulacji komputerowych wykorzystujących zaawansowany model kanału radiowego, bazujący na Trójwymiarowym Śledzeniu Promieni (*ang. 3D-Ray-Tracing*) i dostosowany do symulacji systemów M-MIMO [10].

Pozostała część artykułu zorganizowana jest w następujący: W Sekcji 2 opisany został model rozważanego systemu. W Sekcji 3 przedstawiono zaproponowany algorytm DPB. Sekcja 4 zawiera opis scenariusza symulacyjnego. Dyskusja nad wynikami symulacji znajduje się w Sekcji 5. Wnioski zostały sformułowane i przedstawione w Sekcji 6.

2. Model Rozważanego Systemu

W tej pracy rozważane jest zastosowanie DPB w pojedynczej komórce heterogenicznej sieci M-MIMO, składającej się z pojedynczej stacji bazowej makro MBS (*ang. Macro Base Station*) i N_b stacji bazowych piko PBS (*ang. Pico Base Station*). MBS zapewnia stałe pokrycie komórki sygnałem i jej wygaszanie nie jest rozważane. Zainstalowanie PBS ma na celu zwiększenie pojemności systemu, i zapewnienie obsługi użytkownikom będącym poza zasięgiem MBS. PBS mogą być w zależności od potrzeb wygaszone. Wybór schematu wygaszania jest scentralizowany tzn. MBS jest wyposażona w kontroler DPB pozwalający wygasić dowolną kombinacje PBS, co skutkuje 2^{N_b} możliwymi schematami wygaszania. Kontroler DPB wybiera schemat wygaszania na podstawie informacji raportowanych przez sieć np. przepływności użytkowników, położenie użytkowników. W związku z założeniem braku opóźnień związanych z przesyłaniem informacji sygnalizacyjnych przez siec szkieletową każda PBS może być wygaszona w sposób natychmiastowy, tj. bez opóźnień. Rozważana procedura DPB została przedstawiona na *Rys. 1*. Celem zastosowania tu DPB jest zwiększenie przepływności użytkowników znajdujących się na skraju komórki. Podobnie jak zostało to przyjęte w innych pracach za przepływność użytkowników znajdujących się na skraju komórki będziemy uważali kwantyl rzędu 10 z rozkładu przepływności użytkowników [11].



Rys. 1 Schemat zastosowania DPB w pojedynczej komórce heterogenicznej sieci M-MIMO

3. ALGORYTM DQN-DPB

W sieciach M-MIMO przepływności użytkowników wynikają m.in. ze skomplikowanych zależności korelacji przestrzennych pomiędzy ich kanałami radiowymi. Dodatkowo estymację przepływności użytkowników komplikują algorytmy przydziału zasobów i przypisywania użytkowników do stacji bazowych. Z tej perspektywy rozsądne wydaje się wykorzystanie metod uczenia maszynowego, które mogą na podstawie historycznych danych nauczyć się wskazywać kontrolerowi DPB odpowiednie schematy wygaszania. W tej pracy zaproponowano wykorzystanie uczenia ze wzmocnieniem, w którym proces uczenia przebiega poprzez cykliczną interakcję agenta ze środowiskiem: agent obserwuje stan środowiska, następnie podejmuje akcję, po czym obserwuje nagrodę i aktualny stan środowiska [7]. W przypadku DPB zaproponowano następujące definicje (przedstawione również na *Rys. 2*):

- **Środowisko:** pojedyncza komórka heterogenicznej sieci M-MIMO w ramach której działa kontroler DPB.

- **Agent:** kontroler DPB, którego zadaniem jest podejmowania akcji, tzn. wybór schematu wygaszania, który zapewni wzrost kwantyla rzędu 10 z rozkładu przepływności użytkowników.

- **Stan:** może być zdefiniowany jako zbiór pozycji użytkowników, jednak w ten sposób będzie wrażliwy na liczbę użytkowników znajdujących się w komórce N_U oraz uszeregowanie kolejnych pozycji użytkowników. W ten sposób ograniczone byłoby użycie DNN, które wymagają stałego rozmiaru wektora wejściowego. Z tego względu proponujemy przetwarzanie wstępne, tzn. podzielenie użytkowników na $K \leq N_U$ grup za pomocą algorytmu centroidów (*ang. K-Means*) [12]. W ten sposób stan będzie reprezentowany przez wektor o rozmiarze $K \times 3$

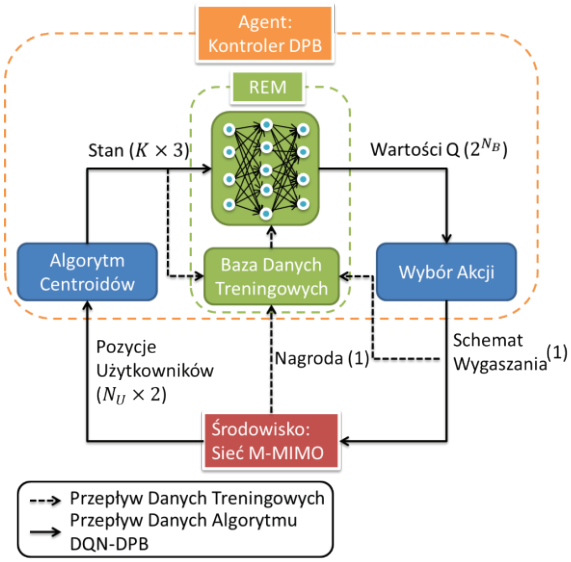
składający się z dwuwymiarowych współrzędnych kartezjańskich centroidów, oraz liczby użytkowników przypisanych do każdej z grup.

- **Akcja:** schemat wygaszania, jedna z 2^{N_b} możliwych konfiguracji wygaszonych PBS.

- **Wybór akcji:** zastosowana została strategia ε-zachłanna, tzn. z prawdopodobieństwem równym ε wybrana zostanie akcja związana z najwyższą spodziewana nagrodą, z prawdopodobieństwem równym $1-\varepsilon$ wybrana zostanie losowa akcja [7].

- **Nagroda:** przepływność użytkowników znajdujących się na skraju komórki, tj. kwantyl rzędu 10 z rozkładu przepływności użytkowników.

W podstawowym podejściu uczenie ze wzmocnieniem wykorzystuje tabelę, w której zapisane są stany, akcje oraz odpowiadające im wartości Q (*ang. Q-Values*), zależne od nagrody, tj. tym większa jest wartość Q danej akcji w danym stanie im większa nagrodę pozwoli ona osiągnąć. Taka tabela może być zrealizowana za pomocą inteligentnej bazy REM, w której pojedynczy wpis zawiera etykietę w postaci stanu (pozycji użytkowników), oraz informacje o wartościach Q związanych z każdą akcja możliwą do wybrania w danym stanie [13]. Łatwo jednak zauważyć, że może istnieć wiele stanów, w których wartości Q dla poszczególnych akcji są podobne. W podejściu z użyciem tabeli te podobieństwa nie zostaną uwzględnione, co po pierwsze zwiększa objętość bazy danych, po drugie stwarza konieczność uczenia się przez agenta wybierania akcji we wszystkich tych stanach niezależnie. Ten problem może być z powodzeniem rozwiązany, kiedy zamiast tabeli użyta zostanie DNN. Wejściem takiej sieci jest Stan, a wyjściem aproksymacja wartości Q dla wszystkich możliwych akcji, jest to tzw. głęboka sieć Q DQN (*ang. Deep Q Network*) [9]. W tym przypadku REM będzie składał się z jednej strony z Bazy Danych Treningowych potrzebnych do trenowania agenta, a z drugiej strony z sieci DQN, której wyjście jest używane do wyboru schematu wygaszania. Zastosowanie DQN do wyboru schematu wygaszania przez kontroler DPB nazwane zostało algorytmem DQN-DPB. Algorytm DQN-DPB i związany z nim cyklem uczenia ze wzmocnieniem został przedstawiony na *Rys. 2*.



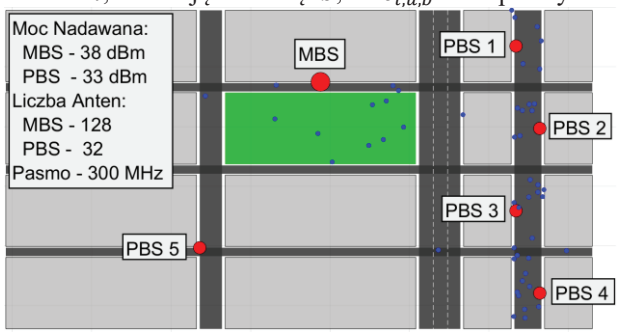
Rys. 2 Algorytm DQN-DPB i związany z nim cykl uczenia ze wzmocnieniem, (~) - rozmiar danych

4. SCENARIUSZ SYMULACYJNY

Rozmieszenie stacji bazowych, wraz z przykładowym rozmieszczeniem użytkowników w rozważanej komórce heterogenicznej sieci M-MIMO jest przedstawione na *Rys. 3*. W komórce znajduje się jedna MBS, której przyporządkowany jest indeks 0 i $N_B = 5$ PBS, którym przyporządkowane są indeksy od 1 do N_B. MBS jest wyposażona w macierz antenową o 128 elementach ($M_0 = 128$), natomiast na każdej PBS zainstalowana jest macierz antenowa o 32 elementach ($M_1 = M_2 \ldots = M_{N_B} = 32$). Stacje bazowe nadają ze stałą mocą, która wynosi $P_0 = 38$ dBm dla MBS, a dla wszystkich PBS $P_1 = P_2 \ldots = P_{N_B} = 33$ dBm. Pasmo systemu wynosi $B = 300$ MHz i jest wspólne zarówno dla MBS jak i wszystkich PBS. Zakładamy, że użytkownicy są przypisywani do stacji bazowej, która oferuje najwyższą moc odbieraną, z wyłączeniem wygaszonych stacji bazowych. W ramach tej samej stacji bazowej wszyscy użytkownicy są obsługiwani równocześnie, z użyciem tych samych zasobów czasowo-częstotliwościowych (multipleksacja przestrzenna, *ang. Space Division Multiple Access*), dzięki prekodowaniu MRT (*ang. Maximum Ratio Transmission* [14]). Przepływność i-tego użytkownika podłączonego do b-tej stacji bazowej dana jest wzorem [14]:

$$r_{i,b} = B\log_2\left(1+\frac{M_b\beta_{i,b}\frac{P_b}{K_b}}{\sum_{l=0}^{N_B}\sum_{u=1,u\neq i}^{N_U} x_{u,l}c_{i,u,l}^2\beta_{i,l}M_l\frac{P_l}{K_l}+\sigma^2}\right), \quad (1)$$

gdzie K_b oznacza liczbę użytkowników obsługiwanych przez stacje bazową o indeksie b, $x_{i,b} = 1$ w przypadku, gdy użytkownik i jest obsługiwany przez stacje bazową b, w przeciwnym wypadku $x_{i,b} = 0$. $\beta_{i,b}$ oznacza średnie tłumienie kanału radiowego pomiędzy użytkownikiem i, a stacją bazową b, a $c_{i,u,b}$ to współczynnik



Rys. 3 Rozmieszenie stacji bazowych (czerwone punkty), wraz z przykładowym rozmieszczeniem użytkowników (niebieskie punkty)

korelacji przestrzennej pomiędzy kanałami radiowymi użytkowników i, oraz u, a stacją bazową b. Moc białego szumu gaussowskiego jest oznaczona jako σ^2 i wynosi -174 dBm/Hz. Na potrzeby symulacji wartości $\beta_{i,b}$ oraz $c_{i,u,b}$ są generowane z użyciem modelu kanału radiowego dedykowanego dla systemów M-MIMO, i wykorzystującego Trójwymiarowe Śledzenie Promieni [10]. Model uwzględnia odbicia śledzonych promieni od budynków, oraz od innych obiektów np. pieszych nieobsługiwanych przez sieć, a znajdujących się w jej obrębie. Warto zaznaczyć, że powyższy opis dotyczy modelu środowiska wykorzystywanego na potrzeby przeprowadzonych symulacji. Zaproponowany algorytm DQN-DPB może być użyty również w innych środowiskach.

Algorytm DQN-DPB wymaga zdefiniowania architektury sieci neuronowej, której będzie używał w celu aproksymacji wartości Q związanych z poszczególnymi schematami wygaszania. Liczba wejść do sieci neuronowej zależy bezpośrednio od liczby grup K, na które algorytm centroidów dzieli użytkowników. Liczba grup została ustalona na $K = 10$, stąd pierwsza warstwa sieci ma 30 neuronów. Kolejne dwie warstwy ukryte mają odpowiednio 128 i 64 neurony, i oddzielone są warstwami aktywacji tzw. ReLU (*ang. Rectified Linear Unit*). Warstwa wyjściowa ma rozmiar zależny od liczby możliwych schematów wygaszania, która dla rozważanej liczby PBS $N_B = 5$ wynosi $2^5 = 32$. Baza danych treningowych przechowuje ostatnie 1000 krotek (*ang. touple*) składających się z 3 elementów: stanu, akcji, oraz nagrody. Przyjęto, że pojedyncze uaktualnienie wag sieci neuronowej odbywa się na podstawie 32 krotek wylosowanych z Bazy Danych Treningowych. Za kryterium optymalizacyjne dla wag sieci neuronowej przyjęto błąd średniokwadratowy pomiędzy nagrodą, a wartością Q estymowana przez sieć neuronową.

Symulacje składają się z tzw. kroków (*ang. step*). Każdy krok odpowiada pojedynczemu cyklowi uczenia ze wzmocnieniem i składa się na niego:

1. Wylosowanie pozycji $N_U = 50$ użytkowników, i wygenerowanie parametrów ich kanałów radiowych. 70% użytkowników znajduje się w alei z prawej strony komórki, 25% znajduje się w parku, niedaleko MBS, a 5% użytkowników w pozostałych obszarach komórki.

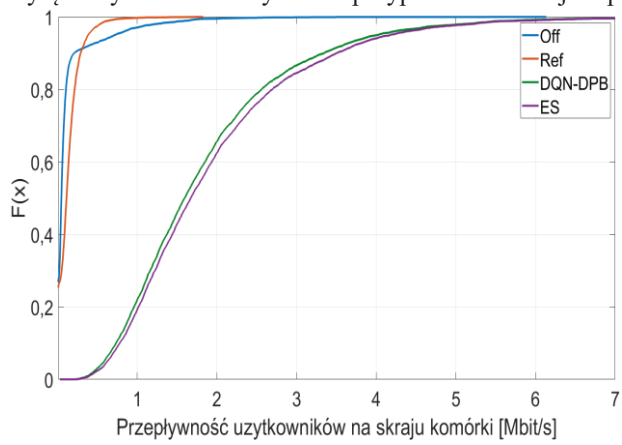
2. Kontroler DPB wybiera schemat wygaszania. Rozważane będą algorytmy bez DPB (*Off*), algorytm referencyjny (*Ref*) wybierający schemat wygaszania na podstawie średniej mocy sygnału odbieranego przez użytkowników do poszczególnych stacji bazowych [4], oraz zaproponowany algorytm DQN-DPB. Dodatkowo możliwe jest znalezienie optymalnego schematu wygaszania poprzez wyszukiwanie wyczerpujące (*ang. Exhaustive Search ES*).

3. Użytkownicy zostają przypisani do aktywnych stacji bazowych na podstawie odbieranej mocy sygnału.

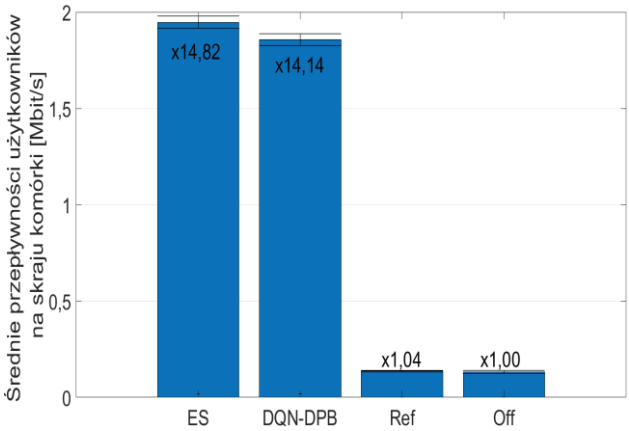
4. Przepływność dla każdego użytkownika zostaje obliczona na podstawie (1). Następnie wyznaczona jest przepływność użytkowników znajdujących się na skraju komórki, tj. kwantyl rzędu 10 z rozkładu przepływności użytkowników.

5. WYNIKI

Zaproponowany algorytm DQN-DPB został zweryfikowany z wykorzystaniem scenariusza symulacyjnego i parametrów opisanych w Sekcji 4. Pierwszym etapem było wytrenowanie algorytmu DPB-DQN, tak aby był w stanie właściwie dobierać schemat wygaszania do aktualnych położeń użytkowników. Wykonano 32000 kroków symulacji dla $\varepsilon = 0$, tzn. za każdym razem wybierana była losowa akcja (schemat wygaszania). Krotki (stan, akcja, nagroda) były zapisywane w Bazie Danych Treningowych i służyły do aktualizacji wag sieci, która odbywała się po każdym kroku symulacyjnym. Łącznie dokonano 32000 aktualizacji wag sieci neuronowej. Po fazie treningowej porównane zostało działanie Kontrolera DPB stosującego różne algorytmy wyboru schematu wygaszania: Algorytm bez DPB (*off*), algorytm referencyjny (*Ref*) [4], zaproponowany algorytm DQN-DPB, oraz przeszukiwanie wyczerpujące (*ES*). Dla każdego z algorytmów wykonano 5000 kroków symulacji, rozpoczynając od tego samego ziarna generatora liczb losowych (innego niż ziarno użyte podczas trenowania algorytmu DQN-DPB), co zapewniło zbadanie algorytmów dla identycznych położeń użytkowników w komórce. Na *Rys. 4* przedstawiono dystrybuantę rozkładu przepływności użytkowników znajdujących się na skraju komórki (kwantyl rzędu 10 z rozkładu przepływności użytkowników) otrzymanych w 5000 kroków symulacyjnych. Można zaobserwować, że algorytm referencyjny w większości przypadków pozwala nieznacznie poprawić przepływności użytkowników znajdujących się na skraju komórki, względem algorytmu bez DPB. Z drugiej strony, algorytm referencyjny w każdym przypadku jest znacznie gorszy od algorytmów DQN-DPB i *ES*. Jest to związane ze złą estymacją interferencji, odpowiednią dla systemów o pojedynczej antenie, i w rezultacie wyborem schematu wygaszania, w którym zbyt wiele PBS jest wyłączonych. We wszystkich przypadkach istnieje op



Rys. 4 Dystrybuanta rozkładu przepływności użytkowników znajdujących się na skraju komórki dla 5000 kroków symulacyjnych



Rys. 5 Przepływności użytkowników znajdujących się na skraju komórki uśrednione po 5000 kroków symulacyjnych razem z 95% przedziałami ufności

tymalny schemat wygaszania, który pozwala osiągnąć znaczne zyski w przepływności użytkowników znajdujących się na skraju komórki. Widać też, że algorytm DQN-DPB jest w stanie z powodzeniem wybierać schematy wygaszania bliskie optymalnym. Przepływności użytkowników znajdujących się na skraju komórki uśrednione po 5000 przebiegów symulacyjnych zostały przedstawione na *Rys. 5* wraz z 95% przedziałami ufności. Optymalny schemat wygaszania poprawia średnią przepływność użytkowników znajdujących się na skraju komórki około 14,82 razy w stosunku do scenariusza bez DPB. Zaproponowany algorytm DQN-DPB charakteryzuje się 95% skutecznością w stosunku do rozwiązania optymalnego i skutkuje średnio 14,14-krotnym zyskiem względem scenariusza bez DPB. W końcu algorytm referencyjny poprzez niewłaściwe wybory schematu wygaszania pozwala na poprawienie przepływność użytkowników znajdujących się na skraju komórki średnio jedynie o 4%.

6. WNIOSKI

W pracy zaproponowane zostało wykorzystanie głębokiego uczenia ze wzmocnieniem do wyboru schematu wygaszania na podstawie położeń użytkowników w komórce, tj. algorytm DQN-DPB. Na podstawie wyników symulacji można stwierdzić, że w skomplikowanym scenariuszu sieci M-MIMO algorytmy dobierające schemat wygaszania na podstawie prostej estymacji interferencji nie sprawdzają się, tj. poprawiają przepływność użytkowników znajdujących się na skraju komórki jedynie o 4%. Sprawdza się natomiast zaproponowany algorytm DQN-DPB, który poprzez interakcję ze środowiskiem w procesie uczenia, i obserwację rzeczywistej nagrody jest w stanie właściwie dobrać schemat wygaszania do danego zbioru położeń użytkowników w komórce. Pozwala to na 14,14-krotne zwiększenie przepływności użytkowników znajdujących się skraju komórki względem scenariusza bez DPB. Ponadto zaproponowany algorytm DQN-DPB osiąga 95% zysków otrzymanych metodą wyszukiwania wyczerpującego.

LITERATURA

[1] Dahlman, Erik, Parkvall, Stefan, Sköld, Johan. 2018. “Chapter 4 - LTE—An Overview, 5G NR: the Next Generation Wireless Access Technology”. *Academic Press,* :39-55.

[2] Bassoy, S., Farooq, H., Imran, M. A., Imran, A. 2017. "Coordinated Multi-Point Clustering Schemes: A Survey". *IEEE Communications Surveys & Tutorials,* 19 (2): 743-764.

[3] Li, Mingju, Yun, Xiang, Nagata S., Chen, Len. 2013. "Power allocation of Dynamic Point Blanking for downlink CoMP in LTE-Advanced," *International Conference on Wireless Communications and Signal Processing*, :1-5.

[4] Wang, H., Tao, X., Zhang, P. 2015. "Adaptive Modulation for Dynamic Point Selection/Dynamic Point Blanking". *IEEE Communications Letters*, vol. 19 (3): 343-346.

[5] Xiong, Z., Zhang, M., Helmers, H., Baker, M., Godin, P., Li, D. 2015. "Centralized Dynamic Point Blanking in LTE-Advanced Network for Inter-Cell Interference Mitigation". *IEEE 81st Vehicular Technology Conference (VTC Spring),* :1-5.

[6] Sanguinetti, L., Björnson, E., Hoydis, J. 2020. "Toward Massive MIMO 2.0: Understanding Spatial Correlation, Interference Suppression, and Pilot Contamination". *IEEE Transactions on Communications,* 68 (1): 232-257.

[7] Richard S. Sutton and Andrew G. Barto. 2018. “Reinforcement Learning: An Introduction”. *A Bradford Book*.

[8] Hoffmann, M., Kryszkiewicz, P. 2021. "Reinforcement Learning for Energy-Efficient 5G Massive MIMO: Intelligent Antenna Switching". *IEEE Access,* 9:130329-130339.

[9] Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. "Playing atari with deep reinforcement learning." *arXiv preprint arXiv:1312.5602*.

[10] Qamcom/QCM, Mar. 2021, [online] Available: https://github.om/qamcom/QCM

[11] Fehske, Albrecht J., Fred Richter, and Gerhard P. Fettweis. 2009. "Energy efficiency improvements through micro sites in cellular mobile radio networks." *2009 IEEE Globecom Workshops*. :1-5.

[12] Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. 2003. "The global k-means clustering algorithm." *Pattern recognition* 36 (2): 451-461.

[13] Hoffmann, Marcin, Paweł Kryszkiewicz, and Adrian Kliks. 2020. "Increasing energy efficiency of massive-mimo network via base stations switching using reinforcement learning and radio environment maps." *Computer Communications* 169: 232-242.

[14] Björnson, Emil, Jakob Hoydis, and Luca Sanguinetti. 2017. "Massive MIMO networks: Spectral, energy, and hardware efficiency." *Foundations and Trends in Signal Processing* 11 (3-4): 154-655.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

PORÓWNANIE METOD DETEKCJI ZAJĘTOŚCI WIDMA RADIOWEGO Z WYKORZYSTANIEM UCZENIA FEDERACYJNEGO Z ORAZ BEZ WĘZŁA CENTRALNEGO

COMPARISON OF SPECTRUM OCCUPANCY DETECTION METHODS WITH THE USE OF FEDERATION LEARNING WITH AND WITHOUT A CENTRAL NODE

Łukasz Kułacz[1]

[1] Politechnika Poznańska, Poznań, lukasz.kulacz@put.poznan.pl

DOI: 10.15199/59.2022.4.72

Streszczenie: **Systemy dynamicznego dostępu do widma w celu podjęcia decyzji o przydziale widma dla nowego urządzenia prze-ważnie wymagają informacji o zajętości widma, a tym samym o obecności innych użytkowników. Proste metody detekcji zajętości widma są często dalekie od niezawodnych, stąd często i z powodzeniem stosowane są algorytmy detekcji zajętości widma wspierane uczeniem maszynowym czy też sztuczną inteligencją. W celu ochrony prywatności danych użytkowników i redukcji ilości danych kontrolnych przekazywanych w systemie interesującym podejściem okazuje się użycie federacyjnego uczenia maszynowego. W tej pracy porównane zostały dwa podejścia do projektowania systemu wykorzystujące federacyjne uczenie maszynowe: z wykorzystaniem węzła centralnego oraz bez wykorzystania węzła centralnego.**

Abstract: **Dynamic spectrum access systems typically require information about the spectrum occupancy and thus the presence of other users in order to make a spectrum allocation decision for a new device. Simple methods of spectrum occupancy detection are often far from reliable, hence spectrum occupancy detection algorithms supported by machine learning or artificial intelligence are often and successfully used. To protect the privacy of user data and to reduce the amount of control data, an interesting approach is to use federated machine learning. This paper compares two approaches to system design using federated machine learning: with and without a central node.**

Słowa kluczowe: **detekcja zajętości widma, federacyjne uczenie maszynowe, sieci bezprzewodowe.**

Keywords: **federated machine learning, spectrum occupancy detection, wireless networks.**

1. WSTĘP

Jednym z głównych wyzwań stojących przed bezprzewodowymi sieciami telekomunikacyjnymi jest efektywne wykorzystanie dostępnych zasobów częstotliwościowych. Pomiary przeprowadzone na przełomie wieków w różnych regionach świata wykazały dużą nieefektywność przydziału częstotliwości statycznej do danej usługi lub zbioru usług. Ta obserwacja była jedną z przyczyn powstania koncepcji radia kognitywnego, przedstawionej w rozprawie doktorskiej J. Mitoli III z 1999 roku, w której dostęp do widma radiowego realizowany jest dynamicznie i adaptacyjnie. W ciągu ostatnich 20 lat na całym świecie przeprowadzono liczne badania mające na celu opracowanie skutecznych metod określania zajętości widma (tzw. sensing), zarządzania widmem i elastycznego dostępu do niego. Główne wnioski z przeprowadzonych analiz to duża zawodność i niedokładność metod detekcyjnych zaimplementowanych tylko w jednym węźle sieci i wynikająca z tego konieczność stosowania kooperacyjnych rozwiązań detekcyjnych lub dodatkowych informacji zapisanych w specjalnych bazach danych. Bardzo obiecującym kierunkiem badań w kontekście wyznaczania zajętości widma jest wykorzystanie różnych metod uczenia maszynowego, zarówno w podejściu jednowęzłowym, jak i kooperacyjnym. W ostatnich latach pojawiło się wiele propozycji wykorzystania technik uczenia maszynowego w kontekście radia kognitywnego, w tym wykrywania zajętości widma. I choć rozwiązania te mogą być bardzo skuteczne, to często wiążą się z dość dużą złożonością obliczeniową, zwłaszcza w kontekście tworzenia, uczenia i aktualizowania modelu wykorzystywanego w algorytmie uczenia maszynowego. Co więcej, bardzo często algorytmy te wymagają przesyłania wielu dodatkowych danych, co pociąga za sobą pogorszenie efektywności widmowej i energetycznej całego systemu. Ponadto istotną kwestią może być zapewnienie prywatności przesyłanych danych, zwłaszcza w sytuacji, gdy w procesie detekcji będą wykorzystywane smartfony użytkowników. W tym kontekście bardzo ciekawym i popularnym rozwiązaniem jest federacyjne uczenie maszynowe, którego główną ideą jest cykliczne przekazywanie modeli zamiast bezpośrednio danych pomiarowych. Takie podejście pozwala nie tylko zapewnić ochronę prywatności danych, ale także inteligentnie uczyć się w poszczególnych węzłach, aby móc podejmować lepsze decyzje lokalnie, korzystając z wiedzy całej sieci. Algorytm ten ogranicza również ilość dodatkowych danych przesyłanych pomiędzy urządzeniami w sieci, a co więcej zapewnia znacznie większe bezpieczeństwo danych użytkownika.

W wyniku zastosowania rozproszonego i/lub scentralizowanego uczenia federacyjnego w procesie wykrywania zajętości widma radiowego można znacznie poprawić efektywność wykorzystania zasobów radiowych w sieciach telekomunikacyjnych. W tej pracy porównane zostały metody rozproszonego uczenia federacyjnego, gdzie modele są przenoszone pomiędzy sąsiednimi węzłami bez udziału węzła koordynującego, jak i wersji scentralizowanej z węzłem centralnym.

2. OPIS PROBLEMU BADAWCZEGO

Zakłada się, że dla ochrony obecnych transmisji należy zapewnić wydajny algorytm wykorzystujący informacje o środowisku, uzyskiwane poprzez wykrywanie widma

[10] lub z dedykowanej bazy danych [1][6]. Szczególną uwagę zyskały dwa podejścia do elastycznego współdzielenia widma, głównie licencjonowany współdzielony dostęp (ang. *Licensed Shared Access*, LSA) i obywatelska usługa radiowa szerokopasmowa (ang. *Citizens Broadband Radio Service*, CBRS) z powiązanym systemem dostępu do widma (ang. *Spectrum Access System*, SAS). Pierwsze podejście jest bardziej popularne w Europie (wprowadzone przez Komisję Europejską [8]), drugie natomiast jest standaryzowane i wdrażane w Stanach Zjednoczonych. W tym modelu wprowadzani są dodatkowi licencjonowani użytkownicy. Mogą oni wykorzystywać fragmenty widma używane przez inne zastane systemy, o ile są zgodne z licencją na współdzielenie widma.

W wyniku prac CEPT i ETSI, określone zostały ramy współdzielenia dla LSA, które zawierają zharmonizowane warunki techniczne, koordynację transgraniczną itp. Z perspektywy architektury, wdrożenie koncepcji LSA zakłada obecność dwóch komponentów na szczycie istniejącej architektury komórkowej, głównie, repozytorium LSA i kontroler LSA. Ewolucja LSA jest również rozwijana w celu zapewnienia priorytetów użytkowników i bardziej dynamicznego ogólnego podejścia do dostępu do współdzielonych zasobów widma, niż może zapewnić system LSA opracowany dla pasma 2,3-2,4 GHz [5][7]. Drugi z najpopularniejszych schematów współdzielenia widma, CBRS, to trójwarstwowy model współdzielenia wprowadzony przez FCC w Stanach Zjednoczonych dla pasma 3550-3700 MHz [9]. W przeciwieństwie do LSA, umożliwia on dodatkowe wykorzystanie zarówno licencjonowanego, jak i nielicencjonowanego widma, o ile prawa operatora zastałego są chronione. Wprowadzane są tutaj dwa rodzaje wykorzystania widma, licencje dostępu priorytetowego (ang. *Priority Access Licensees*, PAL), czyli licencjonowani użytkownicy działający na zasadach podobnych do tych w koncepcji LSA, oraz zwolnione z licencji, czyli ogólny autoryzowany dostęp (ang. *General Authorized Access*, GAA). Dedykowany system dostępu do widma (SAS) kontroluje dostęp do widma zarówno dla użytkowników PAL, jak i GAA, jednak tylko użytkownicy PAL są chronieni przed zakłóceniami. Rozważano jednak nowe podejścia do zarządzania zakłóceniami między użytkownikami GAA, głównie CBRS Alliance omawia potrzebę wprowadzenia dedykowanego menedżera współistnienia. Schemat takiego systemu można zastosować do użytkowników tylko wtedy, gdy istnieją niezawodne metody wykrywania obecności operatora zastałego. W takim przypadku szeroko stosowane jest uczenie maszynowe.

Uczenie maszynowe jest wykorzystywane w procesie wykrywania zajętości widma zarówno w wersji jednowęzłowej, jak i wielowęzłowej [2]. Wyzwaniem jest jednak współpraca pomiędzy poszczególnymi węzłami. Model tworzony lokalnie i generowana decyzja są wystarczające, gdy pojedynczy węzeł sieci odpowiada za detekcję zajętości widma. Jeżeli kilka sąsiednich węzłów wykonuje detekcję zajętości widma, możliwa jest wymiana uzyskanych danych między tymi węzłami w celu optymalizacji podejmowanych decyzji. W jaki sposób te informacje powinny być wymieniane i jak powinny wpływać na poszczególne węzły jest istotnym problemem badawczym. Potencjalnym rozwiązaniem tego problemu jest wykorzystanie federacyjnego uczenia maszynowego. Podejmowane są pierwsze próby wykorzystania tego narzędzia w procesie detekcji zajętości widma i pojawiają się pierwsze publikacje na ten temat [3][4]. Uczenie federacyjne wydaje się być obiecującym rozwiązaniem w tym kontekście, ale nie zostało jeszcze dokładnie zbadane. W ramach tej pracy zakładamy, że dobór odpowiedniej metody uczenia maszynowego w połączeniu z odpowiednią metodą wymiany parametrów modelu powinien skutkować poprawą ogólnej skuteczności detekcji widma. Sam pomysł można zrealizować na dwa sposoby. Najpierw należy sprawdzić skuteczność różnych metod uczenia maszynowego podczas wykrywania zajętości widma w systemie rozproszonym, gdzie parametry modelu wymieniane są między sąsiednimi węzłami. Po drugie, należy sprawdzić różne metody tworzenia grup i wyboru węzłów centralnych, a dalej odpowiednie metody zbierania danych z węzłów, ich przetwarzania i doboru odpowiednich modeli dla węzłów w ramach grupy. Najprostszy koncepcyjnie algorytm federacyjnego uczenia maszynowego jest federacyjne uśrednianie (ang. *Federated Averaging*, FedAvg), który może w tym przypadku nie wystarczyć i należy przetestować inne podejścia. Ogólnym celem jest zaproponowanie algorytmu federacyjnego uczenia w wersji rozproszonej i scentralizowanej, gdzie najpierw wybierana jest odpowiednia metoda uczenia maszynowego, a następnie odpowiednia metoda wyboru i wymiany modeli. Wykorzystanie sieci neuronowych w tym kontekście wydaje się być dobrym rozwiązaniem, jednak ze względu na dużą złożoność obliczeniową może być nieakceptowalna w pojedynczych węzłach sieci.

Warto zaznaczyć, że przedstawione w pracy pomysły są innowacyjne, gdyż badania nad wykorzystaniem federacyjnego uczenia maszynowego w procesie detekcji zajętości widma są na wczesnym etapie i pokazują potencjał tego rozwiązania. Znaczenie potencjalnych wyników uzyskanych w ramach realizacji tych pomysłów to przede wszystkim poprawa skuteczności wykrywania zajętości widma w sieciach bezprzewodowych. Detekcja zajętości widma oraz nierozłączny z nią termin radia kognitywnego są od dawna przedmiotem badań. Pojawia się w normalizacji dotyczącej wykrywania widma w 3GPP. Ponadto standard CBRS w dużej mierze opiera się na wykrywaniu użytkowników pierwotnych w systemie i jest wykorzystywany komercyjnie w Stanach Zjednoczonych. W Europie, w Niemczech, jest duże zainteresowanie przemysłu wykorzystaniem pasma 3,5 GHz na podobnych zasadach jak w systemie CBRS. Z tego powodu znalezienie niezawodnych i wydajnych metod określania dostępności widma radiowego jest bardzo ważnym tematem.

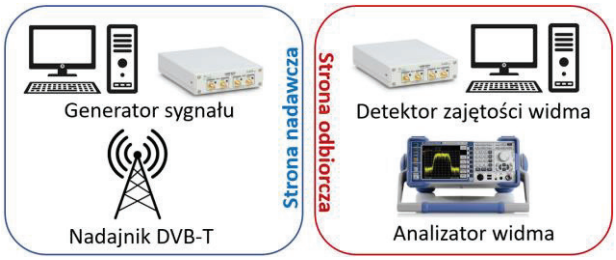
3. DETEKCJA ZAJĘTOŚCI WIDMA

W tej sekcji przedstawione zostały dwie koncepcje opracowania algorytmu federacyjnego uczenia maszynowego do poprawy detekcji zajętości widma radiowego: wykorzystująca węzeł centralny oraz w pełni rozproszona. W obu przypadkach przedstawiony został opis symulacji i analizowanego systemu, który może być wykorzystany jako podstawa do weryfikacji opracowanych algorytmów.

3.1. Federacyjne uczenie maszynowe bez węzła centralnego

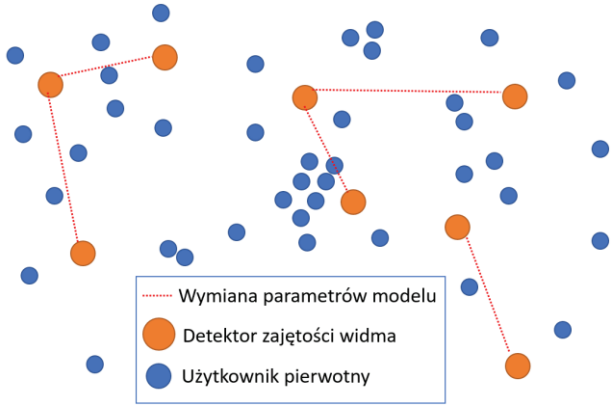
Kluczowym aspektem podczas opracowywania algorytmu federacyjnego uczenia maszynowego w celu zwiększenia efektywności wykrywania zajętości widma jest opracowanie mechanizmu wymiany i aktualizacji parametrów modelu pomiędzy najbliższymi urządzeniami w sieci. Kwestia ta dotyczy nie tylko wyboru jakie parametry wybranego modelu powinny być wymieniane między węzłami, ale również zakresu wymiany parametrów (tj. w jakim promieniu i ile urządzeń w bezpośrednim sąsiedztwie ma wymieniać informacje o modelu), a także częstotliwość ich wymiany. Dodatkowo, z punktu widzenia detekcji zajętości widma i podejmowania lokalnych decyzji, należy zaproponować sposób wykorzystania parametrów wymienionych pomiędzy węzłami (np. czy wyznaczać ich średnią wartość, czy należy ważyć parametry modelu uzyskane z różnych węzłów proporcjonalnie do odległości lub liczby próbek użytych do trenowania modelu itp.). W pierwszym kroku zalecana jest weryfikacja proponowanych mechanizmów i algorytmów za pomocą specjalnie stworzonej symulacji komputerowej. Warto rozważyć również wykorzystanie sieci neuronowych jako algorytmu odniesienia. Kluczowym jest sprawdzenie skuteczności detekcji oraz złożoności obliczeniowej poszczególnych algorytmów. Złożoność tą obliczyć można poprzez pomiar czasu obliczeń, który przekłada się na energię zużytą do przetwarzania oraz ilość pamięci potrzebnej do zbudowania modelu.

W tym celu przeprowadzenia weryfikacji przez symulację komputerową jako odniesienie warto rozważyć wygenerowanie kilku sygnałów referencyjnych za pomocą uniwersalnego radia programowalnego (ang. *Universal Software Radio Peripheral*, USRP). Za pośrednictwem innego urządzenia USRP odebrane próbki sygnału dla różnych parametrów nadawanego sygnału (modulacja sygnału, moc nadawania, długość transmisji danych, długość przerwy w transmisji itp.) powinny być zgromadzone i zapisane w plikach. Ponadto w ten sam sposób mierzony powinien być sygnał telewizyjny, którego dokładne parametry (lokalizacja nadajnika, moc nadajnika, nachylenie anteny itp.) są dostępne na stronie internetowej. Za pomocą sprzętu laboratoryjnego możliwe jest również wygenerowanie lokalnego sygnału telewizyjnego. Utworzone pliki wykorzystać można do trenowania i weryfikacji modeli uczenia maszynowego. Schemat ideowy systemu do zebrania zestawu referencyjnego próbek sygnału przedstawiono na Rys. 1.



Rys. 1. Schemat zestawu do zbierania danych referencyjnych

Uproszczony schemat sieci, analizowanej w opisywanym przypadku przedstawiono na Rys. 2.

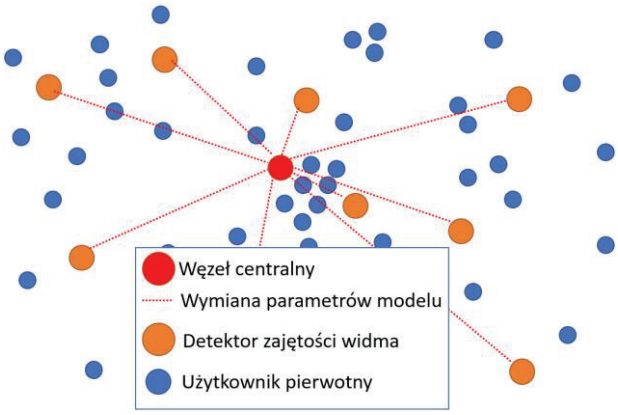


Rys. 2. Uproszczony schemat sieci bez węzła centralnego

W ramach symulacji na określonej powierzchni umieszczone może być kilkanaście urządzeń do wykrywania zajętości widma. Wtedy, użytkownicy pierwotni o pozycji wybranej losowo w obrębie tego samego obszaru są umieszczani w systemie wraz z modelowaniem ich transmisji, czyli czy dany użytkownik przesyła dane w danym momencie. Przedstawiona symulacja pozwoliłaby sprawdzić możliwość wykorzystania federacyjnego uczenia maszynowego do wykrywania zajętości widma w celu poprawy efektywności dynamicznego wykorzystania zasobów radiowych bez obecności węzła centralnego.

3.2. Federacyjne uczenie maszynowe z węzłem centralnym

Drugim rozważanym podejściem jest opracowanie algorytmu federacyjnego uczenia maszynowego z węzłem centralnym, a w szczególności opracowanie mechanizmu wymiany parametrów modelu pomiędzy urządzeniami w sieci a węzłem centralnym. Rozważany problem badawczy, ponownie, koncentruje się na określeniu częstotliwości wymiany danych między węzłami, rodzaju wymienianych danych oraz sposobu ich przetwarzania w węźle centralnym. W celu weryfikacji opracowanych mechanizmów i algorytmów, podobnie jak poprzednie, powinien zostać zaimplementowany specjalny symulator komputerowy. Warto porównać uzyskane wyniki z tej symulacji z wynikami uzyskanymi w drodze symulacji algorytmu federacyjnego uczenia maszynowego bez wykorzystania węzła centralnego. Tutaj podobnie powinna być wybrana odpowiednia metoda uczenia maszynowego, natomiast dodatkowym wyzwaniem jest zarządzanie węzłem centralnym - zbieranie i dystrybucja parametrów modelu. Podczas analizy obu algorytmów zalecane jest wykorzystanie tych samych próbek sygnału, opisanych wcześniej. Podobnie jak poprzednio, opracowany algorytm powinien zostać w drodze symulacji poprzez ocenę złożoności obliczeń, ilość danych przesyłanych do węzła centralnego z pojedynczego węzła sieci i odwrotnie. Dodatkowo istotnym aspektem jest sprawdzanie złożoności obliczeniowej algorytmu działającego w węźle centralnym. Bardzo uproszczony schemat sieci analizowanej w tym zadaniu przedstawiono na Rys. 3.



Rys. 3. Uproszczony schemat sieci z węzłem centralnym

W uproszczeniu sieć symulowana powinna być generowana identycznie jak w poprzedniej symulacji. Ważną różnicą jest obecność jednego dodatkowego węzła centralnego. Ma to ogromny wpływ na sposób wymiany informacji między węzłami przechodzącymi przez węzeł centralny, zamiast wymiany między sąsiednimi urządzeniami. Najbardziej złożonym aspektem w tym przypadku będzie sposób wyznaczenia współczynników modelu dla poszczególnych urządzeń. Przedstawiona symulacja pozwoliłaby na sprawdzenie możliwości wykorzystania sfederowanego uczenia maszynowego do wykrywania zajętości widma w celu poprawy efektywności dynamicznego wykorzystania zasobów radiowych w podejściu z węzłem centralnym.

Wstępna analiza rozważanych algorytmów wskazuje na kilka istotnych różnic wynikających z zastosowania węzła centralnego. Przede wszystkim węzeł centralny zwiększa ilość informacji kontrolnych przesyłanych w sieci, jednocześnie powodując duże zagęszczenie takich transmisji w okolicy węzła centralnego. Potencjalnie, może to skutkować dużym poziomem interferencji. Podsumowanie tej analizy zostało przedstawione w Tab. 1.

Tab. 1. Porównanie algorytmów federacyjnego uczenia maszynowego

Aspekt	*Użycie węzła centralnego*	*Brak węzła centralnego*
Konieczność komunikacji z okolicznymi węzłami	opcjonalna	wymagana
Elastyczność względem topologii sieci	ograniczona pozycją węzła centralnego	znaczna
Ilość danych przesyłanych w jedno lub z jednego miejsca	znacząca; zwłaszcza w węźle centralnym	zrównoważona
Złożoność obliczeniowa algorytmu tworzącego wspólny model	duża; szczególnie w samym węźle centralnym	średnia; zależna od zagęszczenia węzłów
Spodziewana skuteczność detekcji zajętości widma	duża; ogromna ilości danych	średnia; ograniczona ilość danych

4. WNIOSKI

W tej pracy przybliżono dwa podejścia do opracowania algorytmu federacyjnego uczenia maszynowego do zwiększania skuteczności detekcji zajętości widma: z użyciem i bez użycia węzła centralnego. Przedstawiona została również metoda symulacji pozwalająca na weryfikację opracowanych rozwiązań. Wstępne porównanie obu podejść może posłużyć do wyboru odpowiedniej metody do konkretnej analizowanej sieci. Wykorzystanie federacyjnego uczenia maszynowego w tym kontekście wygląda obiecująco, jednakże wymaga dalszych, obszernych badań, weryfikacji symulacyjnej oraz sprzętowej.

5. PODZIĘKOWANIE

Praca powstała w ramach projektu typu PRELUDIUM finansowanego przez Narodowe Centrum Nauki (nr 2021/41/N/ST7/01298).

LITERATURA

[1] Astaneh Saeed, Gazor Saeed. 2013. „Relay-assisted spectrum sensing", *Communications IET*, 8 (1) : 11–18.

[2] Bkassiny Mario, Li Yang. 2013. „A survey on machine-learning techniques in cognitive radios", *IEEE Communications Surveys & Tutorials*, 15 (3) : 1136–1159.

[3] Gao Yujia, Liu Liang. 2021. „Federated sensing: edge-cloud elastic collaborative learning for intelligent sensing", *IEEE Internet of Things Journal*, 8 (14) : 11100–11111.

[4] Jeon Yo-Seb, Amiri Mohammad. 2021. „A compressive sensing approach for federated learning over massive MIMO communication systems", *IEEE Transactions on Wireless Communications*, 20 (3) : 1990–2004.

[5] Jokela Tero, Kokkinen Heikki. 2018. „Trial of spectrum sharing in 2.3GHz band for two types of PMSE equipment and mobile network", *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB),* 1–5.

[6] Kliks Adrian, Goratti Leonardo, Chen Tao. 2016. „REM: revisiting a cognitive tool for virtualized 5G networks", *23rd International Conference on Telecommunications (ICT)*, 1–5.

[7] Mustonen Miia, Matinmikko Marja. 2014. „Evaluation of recent spectrum sharing models from the regulatory point of view", *1st International Conference on 5G for Ubiquitous Connectivity*, 11–16.

[8] Radio Spectrum Policy Group. 2011. „Report on collective use of spectrum and other spectrum sharing approaches", 11–392.

[9] Sohul Munawwar, Yao Miao. 2015. „Spectrum access system for the citizen broadband radio service", *IEEE Communications Magazine*, 53 (7) : 18–25.

[10] Yücek Tevfik, Arslan Hüseyin. 2009. „A survey of spectrum sensing algorithms for cognitive radio applications". *IEEE Communications Surveys & Tutorials*, 11 (4) : 116–130.

ALGORYTM DETEKCJI WĄSKOPASMOWYCH INTERFERENCJI MIĘDZYSYSTEMOWYCH UKRYTYCH W PAŚMIE PRACY SYSTEMU LTE

DETECTION ALGORITHM OF HIDDEN NARROWBAND INTERSYSTEM INTERFERENCE IN LTE SYSTEM

Robert Burczyk[1]; Jarosław Sadowski[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, robert.burczyk@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jarsadow1@pg.edu.pl

DOI: 10.15199/59.2022.4.73

Streszczenie**: W artykule zaprezentowano algorytm detekcji wąskopasmowych interferencji międzysystemowych w sygnale w łączu w dół sieci LTE. Zaproponowany algorytm detekcji interferencji bazuje na analizie zmienności faz konstelacji symboli OFDM w czasie. Artykuł zawiera także porównanie efektywności detekcji interferencji z metodą znaną z literatury.**

Abstract**: In the article, authors present the detection algorithm of narrowband intersystem interference in LTE system. The proposed interference detection algorithm is based on the analysis of received OFDM symbols phase variation in time. The article also contains the comparison of the interference detection effectiveness with the method known from the literature.**

Słowa kluczowe**: LTE, OFDM, interferencje wąskopasmowe.**

Keywords**: LTE, narrowband interferences, OFDM.**

1. WSTĘP

Wąskopasmowe interferencje międzysystemowe (WIM) można zdefiniować jako obecność w sygnale użytecznym danego systemu, zakłócającego sygnału pochodzącego od innego systemu, który charakteryzuje się znacząco węższym pasmem transmisji. Niepożądana obecność WIM w sygnale użytecznym LTE (*Long Term Evolution*) może spowodować degradację parametrów jakościowych transmisji takich jak BER (*Bit Error Rate*), przepływność, opóźnienie [1-4] lub może negatywnie wpłynąć na synchronizację czasową lub częstotliwościową [5], gdy poziom mocy sygnału interferencji jest wysoki. Źródłem interferencji wąskopasmowych może być zarówno emisja niepożądana z pobliskich urządzeń radiokomunikacyjnych jak też emisja sygnału w celu realizacji transmisji ukrytej w widmie sygnału szerokopasmowego [6], natomiast detekcja takich interferencji przy użyciu metod bazujących na analizie mocy w dziedzinie częstotliwości [7] może okazać się mało efektywna. W związku z powyższym opracowano algorytm detekcji WIM w sygnale użytecznym z sieci LTE bazujący na przetwarzaniu sygnałów w łączu w dół w dziedzinie czasu.

2. ALGORYTM DETEKCJI WIM

Opracowany algorytm opiera się na przetwarzaniu w dziedzinie czasu sygnału uprzednio skorygowanego o estymowane parametry kanału radiowego w łączu w dół (po synchronizacji czasowej i częstotliwościowej). Przyjęto zatem założenie, że wprowadzono WIM (węższe pasmo transmisji) do pasma pojedynczej podnośnej OFDM (szersze pasmo transmisji). Mimo, że obecność zakłóceń wąskopasmowych niezsynchronizowanych z sygnałem OFDM będzie mogła wpływać negatywnie na jakość odbioru na kilku podnośnych, detekcję obecności interferencji wykonuje się oddzielnie dla każdej podnośnej.

Zdefiniowano algorytm w założeniu przetwarzania sygnału w dziedzinie czasu symboli techniki OFDM, która ma zastosowanie w obecnie w najnowocześniejszych scentralizowanych systemach radiokomunikacyjnych. Algorytm detekcji WIM przedstawiono dla trybu pracy interfejsu LTE z 14 symbolami w podramce czasowej:

Krok 1. Oblicz wartość symboli referencyjnych modulacji QPSK (*Quadrature Phase Shift Keying*), 16QAM (*Quadrature Amplitude Modulation*), 64QAM, sygnałów synchronizacyjnych PSS (*Primary Synchronization Signal*) i SSS (*Secondary Synchronization Signal*) korzystając z mediany kwadratów modułów symboli sygnału referencyjnego LTE S_{rs} (1) dla ramki czasowej LTE.

$$P = Me(|S_{rs1}|^2, |S_{rs2}|^2, \dots, |S_{rsN_2}|^2) \tag{1}$$

Wartość N_2 jest zależna od szerokości pasma LTE. Dodatkowym symbolem referencyjnym jest $S_0 = 0$ jako przypadek braku transmisji sygnału na danej podnośnej.

Krok 2. Wykonaj detekcję symboli odebranych.

Krok 3. Oblicz wektor błędu $\vec{e}$ dla każdego symbolu odebranego w podramce czasowej LTE jako różnicę symbolu odebranego i symbolu referencyjnego określonego w kroku nr 1 i 2. Wynikiem działania jest ciąg wektorów błędów: $\overrightarrow{e_1}, \overrightarrow{e_2}, \dots \overrightarrow{e_{14}}$.

Krok 4. Oblicz różnicę fazy $\phi_{\overrightarrow{e_n}}$ pomiędzy $\overrightarrow{e_n}$ a $\overrightarrow{e_{n+1}}$ wektorem błędu. Ostatni wektor w podramce należy porównać z pierwszym wektorem kolejnej podramki. Wynikiem działania jest ciąg różnic faz pomiędzy kolejnymi wektorami błędów $\phi_{\overrightarrow{e_1}}, \phi_{\overrightarrow{e_2}}, \dots, \phi_{\overrightarrow{e_{14}}}$.

Krok 5. Oblicz wariancję ciągu różnic faz z poprzedniego kroku według wzoru (2).

$$\sigma_s^2 = \frac{(\phi_{\overrightarrow{e_1}} - \frac{\sum_{i=1}^{14}\phi_{\overrightarrow{e_i}}}{14})^2 + (\phi_{\overrightarrow{e_2}} - \frac{\sum_{i=1}^{14}\phi_{\overrightarrow{e_i}}}{14})^2 + \dots + (\phi_{\overrightarrow{e_{14}}} - \frac{\sum_{i=1}^{14}\phi_{\overrightarrow{e_i}}}{14})^2}{14} \tag{2}$$

Krok 6. Oblicz średnią wariancję dla L podramek czasowych według wzoru (3).

$$\sigma_{avg}^2 = \frac{\sigma_{s1}^2+\sigma_{s2}^2+\cdots+\sigma_{sL}^2}{L} \quad (3)$$

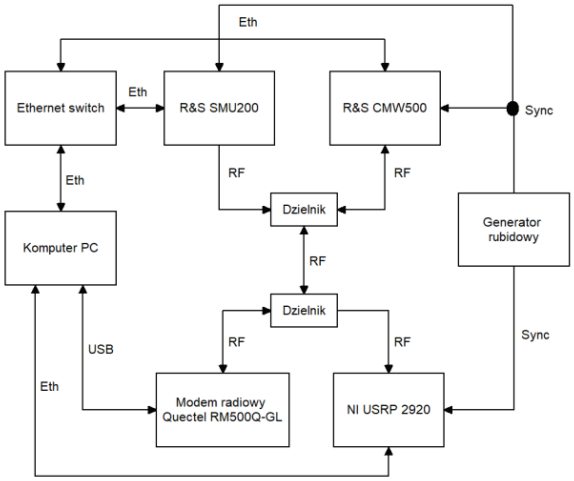
Krok 7. Porównaj wartość obliczoną w kroku 6. z parametrem progowym α. Jeśli $\sigma_{avg}^2 < \alpha$, podejmij decyzję o wystąpieniu WIM, w innym przypadku podejmij decyzją o braku obecności WIM w paśmie pracy systemu LTE.

Położenie symboli sygnałów synchronizacyjnych w konstelacji, w kroku nr 1, określono zgodnie ze specyfikacją ETSI (*European Telecommunications Standards Institute*) LTE [8]. Parametr liczby podramek czasowych uśredniania L we wzorze (3) jest zależny od czasu trwania sygnału interferencyjnego. Parametr progowy α, w kroku nr 7, jest zmienny, zależny od liczby podramek czasowych uśredniania L, modulacji, rodzaju kanału radiowego.

3. WERYFIKACJA POMIAROWA

3.1. Stanowisko pomiarowe

W celu w celu zbadania efektywności detekcji obecności interferencji wąskopasmowych wykonano stanowisko pomiarowe, zaprezentowane na rys. 1.



Rys. 1. Stanowisko pomiarowe do badań WIM

Generatorem sygnału użytecznego jest tester radiokomunikacyjny CMW500 producenta Rohde&Schwarz umożliwiający generowanie sygnału sieci komórkowych o różnych parametrach, zgodnego ze standardami ETSI, zaś źródłem sygnału interferencyjnego jest wektorowy generator sygnałów SMU200A tego samego producenta. Zsumowany sygnał z generatorów trafia na wejścia dwóch urządzeń: NI USRP 2920 producenta National Instruments i modemu radiowego RM500Q-GL producenta Quectel. Pierwsze z urządzeń służy do rejestracji próbek sygnału LTE w paśmie podstawowym w celu jego dalszego przetwarzania, natomiast drugie z urządzeń, czyli modem LTE, służy do wygenerowania transmisji danych zgodnie z protokołem FTP (*File Transfer Protocol*). Zarówno generatory sygnałów jak i platforma USRP (*Universal Software Radio Peripheral*) były synchronizowane generatorem rubidowym w celu uzyskania pożądanej odchyłki częstotliwości między WIM a sygnałem użytecznym. Na komputerze klasy PC zostało uruchomione autorskie oprogramowanie do modyfikacji parametrów nastaw generatorów przy użyciu komend SCPI (*Standard Commands for Programmable Instruments*) oraz obsługi USRP w celu rejestracji próbek.

3.2. Scenariusz pomiarowy

Konfiguracja parametrów sygnału użytecznego, tj. sygnału szerokopasmowego podlegającego zakłócaniu, została przedstawiona w tabeli 1.

Tab. 1. Parametry sygnału LTE

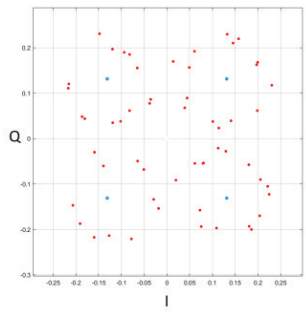
parametr	*wartość*
częstotliwość środkowa	2140 MHz
szerokość pasma	1,4 MHz
modulacja	QPSK, 16QAM, 64QAM
profil kanału	AWGN, EPA 5Hz
SNR	0 – 30 dB
moc szumu	-113 dBm

Zmiennymi były takie parametry jak: SNR, profil kanału AWGN (*Additive White Gaussian Noise*) lub z zanikami o modelu EPA (*Extended Pedestrian A model*), typowego dla ruchu pieszych o maksymalnej odchyłce częstotliwości Dopplera 5 Hz i różnych rodzajów modulacji, możliwych do wygenerowania w testerze.

Tab. 2. Parametry sygnału interferencyjnego

parametr	*wartość*
modulacja	BPSK
S/I	-20 – 20 dB
szerokość pasma	400 Hz – 15 kHz
odchyłka częstotliwości	-6 kHz
profil kanału	AWGN, EPA 5Hz

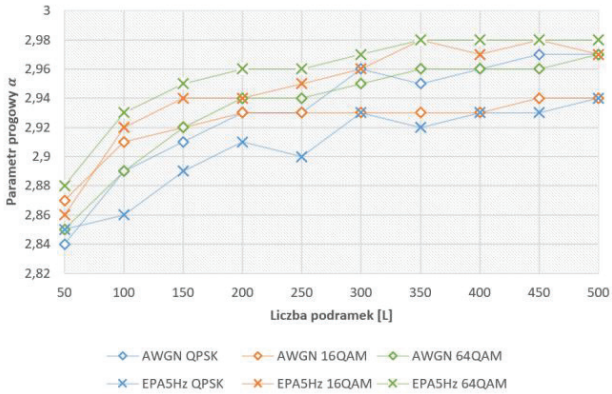
W tabeli nr 2. zostały umieszczone parametry sygnału interferencyjnego. Została wybrana modulacja BPSK (*Binary Phase Shift Keying*) ze względu na jej największy poziom odporności na zakłócenia, co może mieć znaczenie w przypadku traktowania sygnału wąskopasmowego jako transmisji ukrytej w widmie sygnału o szerszym paśmie. Szeroki zakres stosunku mocy sygnału podnośnej (S) do mocy sygnału interferencyjnego (I) umożliwiły badania zarówno w warunkach realizacji transmisji ukrytej jak i w przypadku wprowadzenia interferencji pogarszających parametry jakościowe sygnału użytecznego. Odchyłkę częstotliwości zdefiniowano jako różnicę częstotliwości środkowej wybranej podnośnej sygnału użytecznego i sygnału interferencyjnego. Czas trwania rejestracji próbek ustalono na 100 s. Na rys. 2. zaprezentowano przykładową konstelację sygnału odebranego na zakłócanej podnośnej, w obecności interferencji wąskopasmowych. Różnica między częstotliwością środkową podnośnej a częstotliwością środkową WIM powoduje charakterystyczny wygląd konstelacji, w której symbole odebrane układają się w okręgi naokoło punktów nominalnych konstelacji (symboli referencyjnych).



Rys. 2. Konstelacja sygnału użytecznego z WIM dla 50 symboli odebranych

3.3. Analiza efektywności zaproponowanego algorytmu

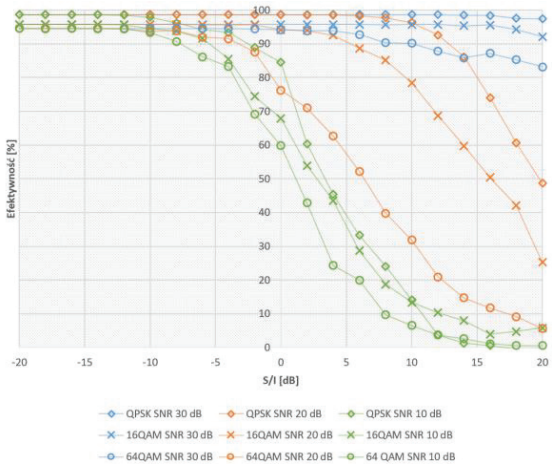
W ostatnim kroku algorytmu, opisanego w rozdziale 2. należy porównać obliczoną, średnią wartość wariancji z parametrem progowym α. Wartość parametru progowego α dobierano eksperymentalnie na podstawie serii pomiarów maksymalizując średnią efektywność detekcji interferencji.



Rys. 3. Wartość parametru progowego alfa w funkcji liczby podramek L

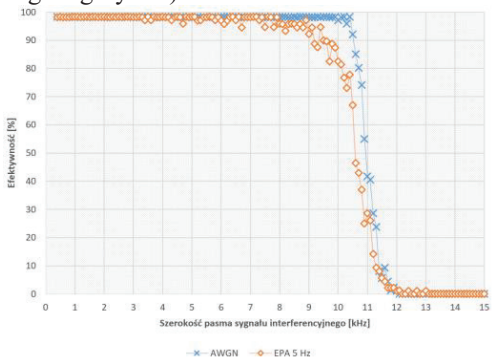
Na rys. 3. zaprezentowano wykres dobranych wartości parametru progowego alfa w funkcji liczby podramek *L* dla różnego typu modulacji i kanału radiowego. Wykonane pomiary wykazały, że wartość parametru α powinna być zwiększana przy wzroście liczby podramek analizy sygnału i przy wzroście wartościowości modulacji. Ponadto, wartość progowa α powinna być większa w przypadku transmisji sygnału w kanale z zanikami, w porównaniu do transmisji bez zaników (AWGN).

Miarą jakości pracy algorytmu detekcji interferencji jest wartość parametru nazwanego efektywnością, obliczanego jako różnica pomiędzy odsetkiem poprawnych detekcji (sytuacja w której algorytm podjął decyzję o wystąpieniu WIM w przypadku jej obecności w sygnale użytecznym) a odsetkiem fałszywych alarmów (sytuacja w której algorytm podjął decyzję o wystąpieniu WIM w przypadku braku jej obecności w sygnale użytecznym). Na rys. 4. zaprezentowano wykres efektywności detekcji WIM w funkcji S/I dla różnych typów modulacji i stosunku sygnału do szumu dla kanału radiowego o profilu EPA 5 Hz dla *L*= 100.



Rys. 4. Efektywność detekcji w funkcji S/I dla różnych typów modulacji i SNR w kanale o profilu zaników EPA 5Hz dla L=100

Zauważono spadek efektywności algorytmu detekcji wraz ze wzrostem S/I oraz wartościowości modulacji. W przypadku wzrostu SNR można zauważyć wzrost efektywności, co jest spowodowane większą stabilnością wartości wariancji różnic faz wektora błędów (krok nr 5 zaproponowanego algorytmu).



Rys. 5. Efektywność detekcji w funkcji szerokości pasma sygnału interferencyjnego dla różnego typu kanału, modulacji QPSK, SNR 30 dB

Na rys. 5. zaprezentowano wykres efektywności w funkcji szerokości pasma sygnału interferencyjnego dla kanału AWGN i EPA 5 Hz. Powyżej wartości granicznej 9 kHz następuje gwałtowny spadek efektywności algorytmu, co jest spowodowane tym, że czas trwania symbolu interferencyjnego i symbolu LTE jest porównywalny. Dla warunków z mniejszym SNR i większą wartościowością modulacji, należy spodziewać się spadku efektywności dla mniejszej wartości szerokości pasma.

4. PORÓWNANIE AUTORSKIEGO ALGORYTMU Z ALGORYTMEM Z LITERATURY

Efektywność pracy autorskiego algorytmu detekcji interferencji (AA) porównano z algorytmem opisanym w literaturze (AL) [7], który bazuje na porównaniu mocy sygnału danej podnośnej ze średnią mocą w całym paśmie transmisji LTE pow_{mid}. Zdecydowano o wyborze algorytmu [7] w celu porównania algorytmów przetwarzania sygnałów w czasie (AA) i częstotliwości (AL). Autorzy AL wskazali na odpowiednie dobranie parametru α_{lit}, koniecznego do wyznaczenia progu detekcji pow_{thre} według zależności (4).

$$pow_{thre} = \alpha_{lit} * pow_{mid} \tag{4}$$

gdzie:

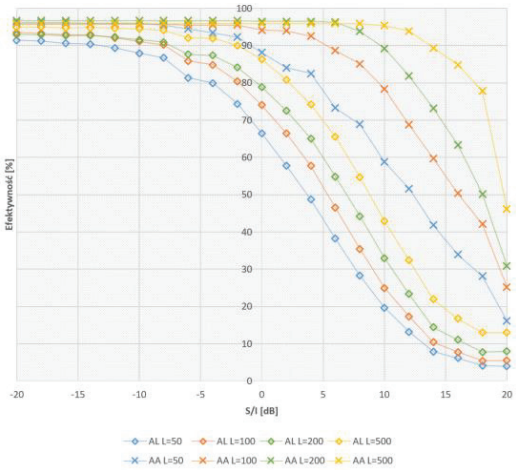
pow_{thre} – próg detekcji algorytmu z literatury,

pow_{mid} – średnia wartość mocy sygnału w całym paśmie LTE.

Tab. 3. Dobór parametru α_{lit}

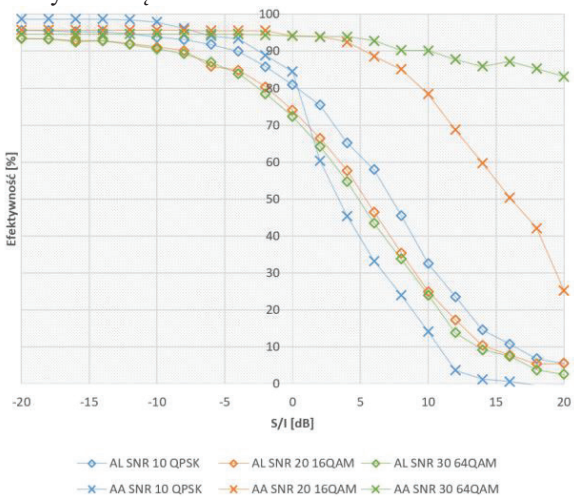
Liczba podramek L	Modulacja	SNR	Wartość α_{lit}
50	16QAM	20	1,12
100	16QAM	20	1,09
200	16QAM	20	1,06
500	16QAM	20	1,04
100	QPSK	10	1,06
100	64-QAM	30	1,05

Wartość parametru progowego α_{lit} (tab. nr 3) dobierano eksperymentalnie na podstawie serii pomiarów maksymalizując średnią efektywność detekcji interferencji w szerokim zakresie parametru S/I. Dobór parametru progowego α_{lit} dla metody AL był wykonany niezależnie od treści artykułu [7] w celu ujednolicenia kryterium oceny obu algorytmów. Należy zwrócić uwagę, że w odróżnieniu do AA, dla AL nie można było wykonać optymalizacji parametru progowego dla średniej efektywności w szerokim zakresie SNR ze względu na duże rozbieżności uzyskiwanej wartości parametru progowego. Dobór parametru α_{lit} z tab. 3. wykorzystano do obliczenia efektywności AL w funkcji S/I dla różnej wartości parametru L (rys. nr 6). Można wywnioskować, że efektywność AA dla profilu zaniku EPA 5Hz, modulacji 16 QAM i SNR 20 dB w każdym przypadku jest lepsza niż efektywność AL, nawet porównując różną liczbę podramek czasowych L, co świadczy o mniejszej bezwładności AA.



Rys. 6. Efektywność autorskiego algorytmu (AA) i algorytmu z literatury (AL) w funkcji S/I dla różnej liczby L w kanale o profilu EPA 5Hz, modulacji 16QAM,SNR 20 dB

Na rys. 7. zaprezentowano wykres porównania AA i AL dla różnego typu modulacji i SNR. Dla przypadku kombinacji parametrów modulacji QPSK i SNR 10 dB efektywność AA jest gorsza niż AL, co wynika z dużej wrażliwości AA na zmianę różnic faz, spowodowaną obecnością wysokiego poziomu szumu w odbieranym sygnale. Dla pozostałych przypadków AA cechuje się wyższą efektywnością niż AL.



Rys. 7. Efektywność autorskiego algorytmu (AA) i algorytmu z literatury (AL) w funkcji S/I dla różnej wartości SNR i typu modulacji w kanale o profilu EPA 5Hz

5. PODSUMOWANIE

Dla zdecydowanej większości różnorodnych warunków transmisji, zaprezentowanych w niniejszym artykule algorytm autorski cechuje się większą efektywnością detekcji interferencji wąskopasmowych w porównaniu z przykładowym algorytmem opisanym w literaturze. W przyszłości, autorzy planują rozszerzyć badania dla większej liczby i zakresu zmienności parametrów transmisji, takich jak odchyłka częstotliwości, profil kanału radiowego. Możliwe jest przebadanie negatywnych skutków oddziaływania wąskopasmowych interferencji międzysystemowych na parametry jakości odbioru sygnału z sieci LTE, odbieranego w łączu w dół, synchronizację czasową, częstotliwościową oraz estymację kanału radiowego.

LITERATURA

[1] Vogeler S., L. Broetje, K.-D. Kammeyer, R. Rueckriem, S. Fechtel. 2003. „Blind Bluetooth Interference Detection and Suppression for OFDM Transmission". The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers.

[2] Ovtcharov M., V. Poulkov, G. Iliev, Z. Nikolova. 2009. "Narrowband Interference Suppression for IEEE UWB Channels". Fourth International Conference on Digital Telecommunications, s. 43-47.

[3] Baumgartner S., G. Hirtz. 2014. "Narrowband interference mitigation for SDR based DAB/DAB+ receivers". IWSSIP 2014 Proceedings. s. 151-154.

[4] Tekovic A., G. Šimac, K. Šakić. 2012. "LTE downlink system performance measurement with intersystem interference caused by DVB-T signal". Proceedings ELMAR-2012. s. 255-258.

[5] Sun P., L. Zhang. 2010. "Narrowband Interference Effect on Timing Synchronization for OFDM-based Spectrum Sharing System". 6th International Conference on Wireless and Mobile Communications. S 274-278.

[6] Grzesiak K., Z. Piotrowski, J. Kelner. 2021. „A Wireless Covert Channel Based on Dirty Constellation with Phase Drift". Electronics. 10. 647. 10.3390/electronics10060647.

[7] Fang F., Z. Yang. 2016. "Time domain suppression scheme of narrowband interference based on narrowband OFDM power line communications", International Conference on Communication and Electronics Systems (ICCES). s. 1-6.

[8] „Evolved Universal Terrestrial Radio Access (E-UTRA) Physical channels and modulation". 2017. 3GPP TS 36.211 version 13.7.1 Release 13

ESTYMACJA PARAMETRÓW KANAŁU NB-IOT W ŁĄCZU W DÓŁ
NB-IOT DOWNLINK CHANNEL ESTIMATION

Jarosław Magiera[1]; Piotr Rajchowski [1]; Krzysztof Cwalina [1]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk
jaroslaw.magiera@pg.edu.pl; piorajch@pg.edu.pl; kkcwalina@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.74

Streszczenie: **Artykuł dotyczy doboru metody estymacji parametrów kanału, odpowiedniej do zastosowania w terminalu wąskopasmowego Internetu Rzeczy. Przedstawiono strukturę sygnału pilotowego NRS, używanego do wyznaczenia współczynników kanału w LTE-NB. Zaproponowano cztery warianty metody wyznaczenia współczynników kanału dla całej ramki transmisyjnej. Dokonano oceny tych wariantów w oparciu o dwa wskaźniki jakościowe skojarzone z błędem odtworzonych symboli względem konstelacji odniesienia. Na tej podstawie wskazano wariant optymalny.**
Abstract: **The paper treats about the selection of channel estimation method suitable for narrowband internet-of-things user equipment. The structure of NRS reference signal, used for channel coefficients calculation in LTE-NB, is presented. Authors propose four variants of estimating channel coefficients for resource elements in whole transmission frame. These variants were evaluated using two quality indicators related to error between recovered symbols and reference complex constellation. As a result, optimal variant was selected.**

Słowa kluczowe: **Estymacja kanału, Internet Rzeczy, NB-IoT.**

Keywords: **Channel estimation, Internet of Things, NB-IoT.**

1. WSTĘP

Standard NB-IoT (Narrowband Internet of Things), inaczej LTE Cat NB, zakłada wykorzystanie energooszczędnych terminali bezprzewodowych, umożliwiających sporadyczne transmisje niewielkich ilości informacji za pośrednictwem istniejącej infrastruktury sieci komórkowej LTE/5G. Zgodnie z ideą Internetu Rzeczy, informacje są przesyłane od lub do urządzeń, działających zwykle w sposób autonomiczny. Przyjmuje się, że urządzenia zgodne ze standardem NB-IoT powinny działać bezobsługowo przez okres co najmniej kilku lat [1].

Istotną cechą odróżniającą wąskopasmowy interfejs radiowy LTE Cat NB od szerokopasmowego LTE jest możliwość realizacji komunikacji pomiędzy stacją bazową eNodeB a nieruchomymi terminalami, umieszczonymi w lokalizacjach o trudnych warunkach propagacji sygnału, jak np. piwnice, garaże podziemne, tunele itp. Jest to możliwe m.in. dzięki zastosowaniu redundancji czasowej, tzn. wielokrotnych powtórzeń transmisji sygnałów przenoszących tę samą informację.

Biorąc pod uwagę wymóg odbioru przez terminal sygnału o niskiej jakości, należy zwrócić szczególną uwagę na kwestię estymacji parametrów kanału radiowego w łączu w dół. Mniejszy błąd estymacji kanału może umożliwić odbiór przy niższej wartości stosunku sygnał-szum. Niemniej, bardziej zaawansowane algorytmy estymacji kanału charakteryzują się większą złożonością obliczeniową, stąd ich zastosowanie nie jest wskazane z uwagi na konieczność redukcji zużycia energii przez urządzenie.

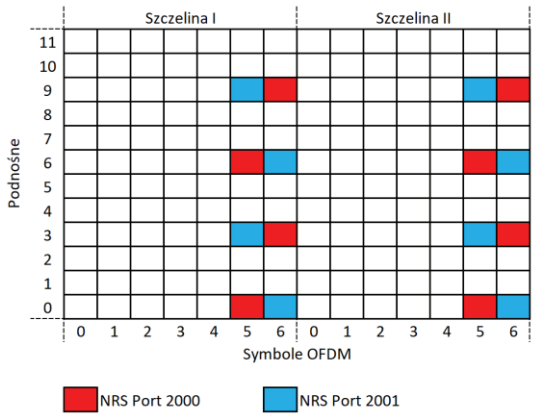
W niniejszym artykule porównano cztery warianty metody estymacji kanału NB-IoT o niskiej złożoności obliczeniowej. Warianty te bazują na interpolacji lub uśrednianiu współczynników kanału wyznaczonych na podstawie sygnału referencyjnego.

2. SYGNAŁ REFERENCYJNY W NB-IOT

Standard komunikacji LTE Cat NB1 został wprowadzony w wydaniu 13 specyfikacji 3GPP [3]. Obejmuje on m.in. definicję warstwy fizycznej, gdzie dla transmisji NB-IoT jest przydzielony pojedynczy blok zasobów fizycznych, obejmujący 12 podnośnych OFDM (Orthogonal Frequency-Division Multiplexing) w pasmie 180 kHz. W dziedzinie czasu transmisja jest zorganizowana w ramki składające się z dziesięciu podramek o czasie trwania 1 ms. Z kolei każda podramka składa się z dwóch szczelin po 7 symboli ODFM.

Estymacja parametrów kanału jest realizowana w oparciu o sygnał pilotowy NRS (Narrowband Reference Signal), przesyłany we wszystkich podramkach NB-IoT, poza tymi przenoszącymi sygnały synchronizacyjne NPSS i NSSS (Narrowband Primary/Secondary Synchronization Signal). Dla transmisji sygnału NRS w szczelinie przydzielono po cztery elementy zasobów dla każdego aktywnego portu antenowego w eNodeB. Stacja bazowa może nadawać sygnał NB-IoT przez co najwyżej dwie anteny, stosując technikę dywersyfikacji przestrzennej. Na rys. 1 przedstawiono przykładowe rozmieszczenie elementów zasobów NRS w pojedynczej podramce. Zasoby te występują zawsze w dwóch ostatnich symbolach, a ich przyporządkowanie do podnośnych jest uzależnione od wartości identyfikatora komórki NB cell ID.

Sygnał NRS ma charakter sekwencji liczb zespolonych, odpowiadających symbolom modulacji QPSK (Quadrature Phase Shift Keying). Wartości rzeczywiste i urojone tych symboli są wyznaczane w oparciu o odpowiednio parzyste i nieparzyste bity ciągu Golda o długości $2^{31} - 1$.



Rys. 1. Przykładowa alokacja zasobów ODFM dla sygnału NRS w jednej podramce.

3. PROCEDURA ESTYMACJI KANAŁU

Na wstępie należy zaznaczyć, że procedura estymacji parametrów kanału nie podlega standaryzacji, tzn. nie jest ona zdefiniowana w dokumentacji 3GPP i pozostaje ona w gestii projektantów terminali NB-IoT. W związku z tym, poniższy opis należy traktować jako koncepcję zaproponowaną przez autorów.

Estymacja parametrów kanału jest realizowana każdorazowo po wyznaczeniu siatki zasobów czasowo-częstotliwościowych OFDM, odpowiadających jednej ramce transmisyjnej. Procedura rozpoczyna się od wyznaczenia współczynników kanału dla zasobów zajętych przez transmisję sygnału pilotowego (NRS). Współczynnik kanału h jest liczbą zespoloną, reprezentującą zmianę amplitudy i fazy sygnału odebranego y względem sygnału nadanego x w określonym elemencie zasobów:

$$y(n,k) = h(n,k) \cdot x(n,k) + \eta(n,k) \qquad (1)$$

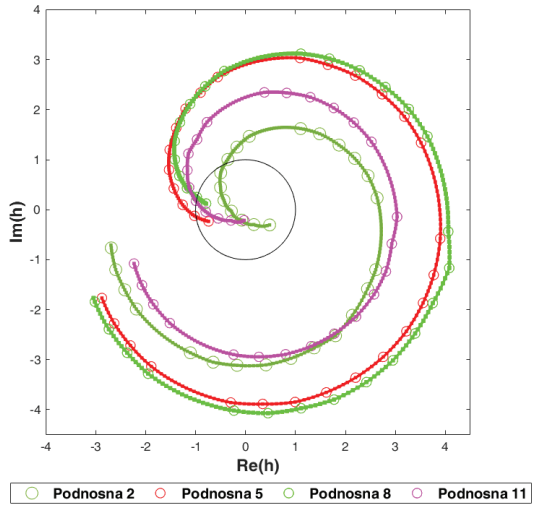
gdzie n, k stanowią indeksy symbolu i podnośnej, a η reprezentuje zakłócenia addytywne.

Dla potrzeb estymacji pilotów powszechnie stosuje się metodę najmniejszych kwadratów LS (least squares), zgodnie z którą współczynnik kanału $\hat{h}_{LS}$ jest obliczany jako iloraz zespolony odebranego i nadanego elementu sygnału pilota [2,5]:

$$\hat{h}_{LS}(n,k) = [x(n,k)]^{-1} \cdot y(n,k) \qquad (2)$$

Istotą procedury estymacji kanału jest wyznaczenie wartości współczynników kanału dla zasobów innych niż zasoby sygnału NRS. Rozszerzenie estymacji kanału na całą siatkę zasobów czasowo-częstotliwościowych może być realizowane wieloma sposobami, różniącymi się między sobą dokładnością i złożonością obliczeniową. W proponowanej koncepcji, za bazowe rozwiązanie uznaje się dokonanie interpolacji liniowej współczynników kanału wyznaczonych z sygnału NRS. Interpolacja jest wykonywana dwuetapowo – najpierw w dziedzinie czasu dla czterech podnośnych przenoszących NRS, a następnie w dziedzinie częstotliwości, uzupełniając współczynniki dla pozostałych ośmiu podnośnych.

Jako że współczynniki kanału są zespolone, zastosowano interpolację liniową odrębnie ich modułów i argumentów. Komentarza wymaga również kwestia estymacji współczynników dla skrajnych podnośnych, nieleżących pomiędzy dwiema podnośnymi NRS. Zastosowanie ekstrapolacji liniowej modułu liczby zespolonej może prowadzić do uzyskania nieprawidłowych rezultatów w postaci wartości ujemnych. W związku z tym, zamiast ekstrapolacji liniowej, jako wartości współczynników kanału na skrajnych podnośnych powielono wartości wyznaczone na najbliższej im podnośnej NRS.



Rys. 2. Interpolacja współczynników kanału w dziedzinie czasu, na czterech podnośnych przenoszących NRS.

Estymacja parametrów kanału odbywa się w odstępach 10 ms, odpowiadających czasowi trwania jednej ramki transmisyjnej. Aby zachować ciągłość interpolacji czasowej współczynników kanału pomiędzy kolejnymi ramkami, jednorazowo przetwarzany jest segment sygnału obejmujący całą bieżącą ramkę oraz pierwszą szczelinę ramki następnej – łącznie 147 symboli OFDM. Na rys. 2 przedstawiono przykładowy wynik interpolacji czasowej współczynników kanału z zanikami o profilu EPA (Extended Pedestrian A [4]) przy transmisji sygnału bez szumu. Są to wyniki dla czterech podnośnych przenoszących sygnał NRS w dwóch kolejnych ramkach transmisyjnych. Symbole okręgów oznaczają współczynniki wyznaczone z zasobów NRS metodą najmniejszych kwadratów, a punkty pomiędzy nimi są wynikiem interpolacji. Widoczna jest ciągłość współczynników w dziedzinie czasu.

W celu uzyskania dalszej redukcji złożoności obliczeniowej estymacji kanału, zaproponowano możliwość zastąpienia interpolacji linowej uśrednianiem współczynników. W dziedzinie czasu uśrednianie jest dokonywane w obrębie jednej ramki z zasobów NRS na czterech podnośnych, a w dziedzinie częstotliwości uśrednia się współczynniki z podnośnych NRS w pojedynczym symbolu OFDM. Należy zwrócić uwagę, że uśrednianie, podobnie jak interpolacja, musi być wykonane w odpowiedniej kolejności, tzn. najpierw w dziedzinie czasu, a następnie po podnośnych.

Biorąc pod uwagę możliwość zastosowania różnych kombinacji interpolacji liniowej i uśredniania, wyróżniono cztery warianty zaproponowanej metody wyznaczania współczynników kanału:

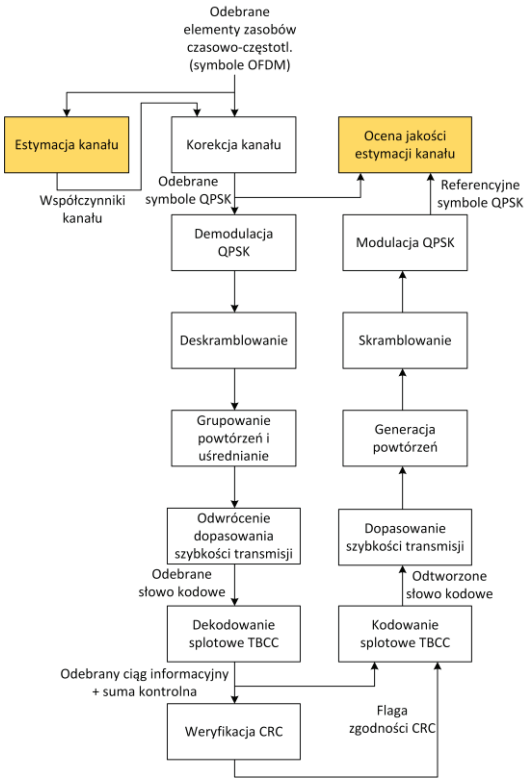
1. Interpolacja w czasie i interpolacja po podnośnych.
2. Interpolacja w czasie i uśrednianie po podnośnych.
3. Uśrednianie w czasie i interpolacja po podnośnych.
4. Uśrednianie w czasie i uśrednianie po podnośnych.

Każdy z powyższych wariantów poddano ocenie jakościowej zgodnie z metodyką przedstawioną w kolejnym rozdziale.

4. OCENA JAKOŚCI ESTYMACJI KANAŁU

4.1. Metodyka oceny

Ocena jakości estymacji parametrów kanału jest dokonywana poprzez porównanie położenia punktów konstelacji na płaszczyźnie zespolonej, odpowiadających symbolom odebranym i nadanym. Odebrane punkty konstelacji uzyskuje się w wyniku korekcji kanału, która wykorzystuje współczynniki wyznaczone w procesie estymacji parametrów kanału.



Rys. 3. Odtwarzanie referencyjnych symboli QPSK dla potrzeb oceny jakości estymacji kanału.

W łączu w dół LTE Cat NB stosowana jest wyłącznie modulacja QPSK, co oznacza że konstelacja składa się z czterech punktów o jednakowej amplitudzie i przesuniętych w fazie o wielokrotność 90 stopni. Aby określić błąd odtworzonych punktów konstelacji należy uprzednio określić odpowiadające im punkty konstelacji QPSK po stronie nadawczej. Można tego dokonać realizując pełne dekodowanie wiadomości zawartej w odebranym sygnale. Jeśli wiadomość jest odebrana poprawnie, o czym świadczy weryfikacja sumy kontrolnej CRC, jej bity są następnie poddawane takim samym operacjom, jakie są realizowane po stronie nadawczej, do etapu przypisania parom bitów odpowiednich punktów konstelacji QPSK. Schemat tej procedury został przedstawiony na rys. 3.

Dysponując zbiorem odtworzonych i referencyjnych symboli konstelacji QPSK, można dokonać oceny dokładności estymacji kanału poprzez obliczenie wskaźników jakościowych. Przyjęto tutaj dwa następujące parametry:

1. Błąd RMS fazy odtworzonych symboli, obliczony według wzoru:

$$\Delta\theta_{RMS} = \sqrt{\frac{1}{N_{symb}}\sum_{i=1}^{N_{symb}}\left(\theta_i - \theta_{i,ref}\right)^2}\,[rad] \qquad (3)$$

gdzie θ reprezentuje fazę a N_{symb} jest liczbą analizowanych symboli.

2. Korelacja symboli odtworzonych i referencyjnych, obliczona w oparciu o składowe synfazowe (I) i kwadraturowe (Q) symboli, według wzoru:

$$\rho_{IQ} = \frac{1}{\alpha}\cdot\sum_{i=1}^{N_{symb}}\left(I_i \cdot I_{i,ref} + Q_i \cdot Q_{i,ref}\right) \qquad (4)$$

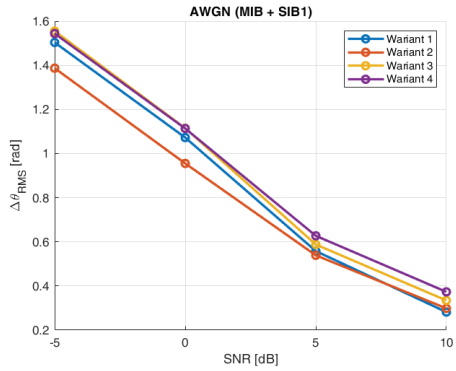
gdzie α jest współczynnikiem normalizacyjnym, ograniczającym wartości ρ_{IQ} do zakresu od -1 do 1.

4.2. Wyniki

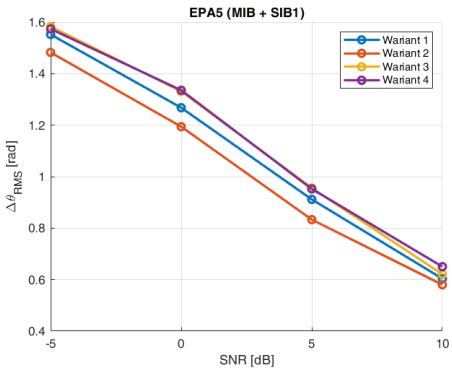
Jakość estymacji kanału została zweryfikowana poprzez przetworzenie sygnałów zarejestrowanych w obecności szumu gaussowskiego, przy SNR o wartościach -5 dB, 0 dB, 5 dB i 10 dB. Analizowano transmisję w kanale bez zaników (AWGN) oraz z zanikami o profilu EPA z maksymalną odchyłką dopplerowską 5 Hz. Dla każdej kombinacji SNR i profilu kanału podjęto 1000 prób zdekodowania bloków informacyjnych, kolejno Master Information Block-NB oraz System Information Block 1-NB. W tab. 1 podano liczbę prób zakończonych powodzeniem, tzn. zakończonych pomyślną weryfikacją CRC.

Tab. 1. Liczba zdekodowanych par bloków informacyjnych MIB-NB i SIB1-NB na 1000 prób

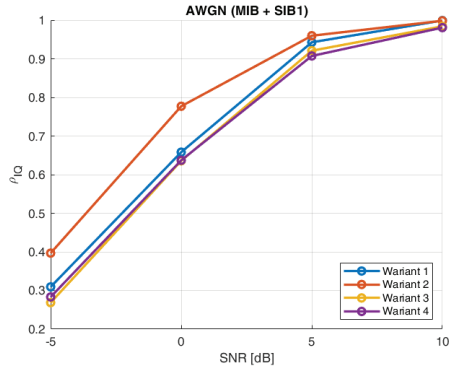
SNR [dB]	*AWGN*	*EPA5*
-5	952	509
0	937	801
5	941	919
10	995	962



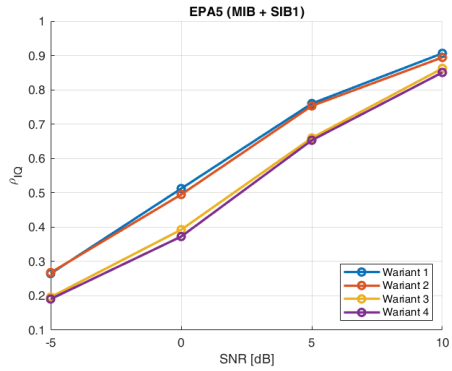
Rys. 4. Błąd RMS fazy – kanał bez zaników



Rys. 5. Błąd RMS fazy – kanał z zanikami



Rys. 6. Korelacja symboli – kanał bez zaników



Rys. 7. Korelacja symboli – kanał z zanikami

W oparciu o odebrane bloki informacyjne dokonano odtworzenia referencyjnych symboli modulacji QPSK i obliczono wartości parametrów zdefiniowanych w sekcji 4.1. Wyniki zostały zaprezentowane na rys. 4-7. Numeracja wariantów odpowiada ich kolejności wymienionej w rozdziale 3.

Analizując błąd RMS (Root Mean Square) fazy symboli, można zauważyć jego niemal liniowy spadek w funkcji stosunku sygnał-szum. W przypadku kanału AWGN spadek ten jest większy niż w kanale z zanikami. Zarazem widać, że najmniejsze wartości tego błędu są odnotowane dla wariantu 2, czyli interpolacji współczynników kanału w dziedzinie czasu, połączonej z ich uśrednianiem po podnośnych.

Wskaźnik korelacji symboli, przy transmisji w kanale z zanikami (EPA 5 Hz), wykazuje wyraźną przewagę wariantów z interpolacją w dziedzinie czasu, w porównaniu z wariantami stosującymi uśrednianie po czasie. Zauważalny jest również niemal liniowy wzrost korelacji ze zmianą SNR. Z kolei w przypadku kanału AWGN, dla wysokich wartości SNR widać nasycenie – wskaźnik korelacji jest bliski jedności, co oznacza dużą dokładność odtwarzania symboli. Widoczna jest także ponownie przewaga jakościowa wariantu 2, w szczególności dla sygnałów o niskim SNR.

Analizowano również wyniki wyznaczone odrębnie dla dekodowania tylko jednego z bloków informacyjnych. W wypadku wariantów stosujących uśrednianie po czasie (3 i 4) wyniki dla dekodowania MIB-NB były wyraźnie gorsze niż dla dekodowania SIB1-NB. Wynika to z faktu, że MIB-NB jest przesyłany w podramce 0, a SIB1-NB w podramce 4. Uśredniając współczynniki kanału po całej ramce, estymacja jest dokładniejsza dla zasobów występujących blisko środka ramki niż jej początku lub końca.

5. PODSUMOWANIE

W artykule przedstawiono cztery warianty metody estymacji parametrów kanału, których niska złożoność obliczeniowa umożliwia ich zastosowanie w energooszczędnych terminalach NB-IoT. Analiza wskaźników jakościowych wskazuje, że warianty stosujące uśrednianie współczynników w dziedzinie czasu, charakteryzują się większym błędem i nie powinny być brane pod uwagę. Najlepsze wyniki uzyskano dla wariantu łączącego interpolację liniową w dziedzinie czasu z uśrednianiem w dziedzinie częstotliwości. Wskazuje to, że kanał o szerokości 180 kHz można uznać za płaski, przy założeniu przyjętego profilu zaników EPA 5Hz. Należy także zaznaczyć, że stosowanie interpolacji współczynników w dziedzinie czasu uniezależnia błąd estymacji od numeru podramki, w której są zlokalizowane zasoby.

***

Praca została zrealizowana w ramach projektu pt. „Definiowany programowo, uniwersalny interfejs radiowy inteligentnych urządzeń Internetu Rzeczy", współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego, umowa nr POIR.01.01.01-00-1025/19-00 z dnia 28.05.2020 r.

LITERATURA

[1] Fattah Hossam. *5G LTE Narrowband Internet of Things (NB-IoT)*. CRC Press, 2018.

[2] Khlifi Abdelhakim, Bouallegue Ridha. *Performance analysis of LS and LMMSE channel estimation techniques for LTE downlink systems.* arXiv preprint arXiv:1111.1666 (2011).

[3] Specyfikacja techniczna 3GPP TS 36.211, *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation*, wydanie 13, 2017.

[4] Specyfikacja techniczna 3GPP TS 36.116, *Evolved Universal Terrestrial Radio Access (E-UTRA); Relay radio transmission and reception*, wydanie 14, 2017.

[5] Van De Beek J-J. et al. *On channel estimation in OFDM systems.* 1995 IEEE 45th Vehicular Technology Conference. Countdown to the Wireless Twenty-First Century. Vol. 2. IEEE, 1995.

ANALIZA METOD SYNCHRONIZACJI W SIECI NB-IOT

Synchronisation methods analysis in NB-IoT

Olga Błaszkiewicz[1]; Piotr Rajchowski[2]; Jarosław Sadowski[3]

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk,
[1] olga.blaszkiewicz@pg.edu.pl, [2] piotr.rajchowski@eti.pg.edu.pl, [3] jaroslaw.sadowski@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.75

Streszczenie: **W artykule zaprezentowano przegląd i analizę metod synchronizacji odbiornika w sieci NB-IoT. Zaimplementowano trzy metody synchronizacji, a ich porównania dokonano na bazie rzeczywistych sygnałów NB-IoT wygenerowanych przy użyciu testera radiokomunikacyjnego w wyizolowanym i emulowanym środowisku propagacyjnym.**

Abstract: **In the article a survey of synchronization of reception methods and analysis were made. Three different methods of synchronization were implemented. Analysis and comparation of the results were based on real NB-IoT signals generated by radiocommunication tester in confined and emulated propagation environment.**

Słowa kluczowe: **NB-IoT, NPSS, synchronizacja odbiornika**

Keywords: **NB-IoT, NPSS, receiver synchronization**

1. WSTĘP

Wzrost liczby urządzeń komunikujących się ze sobą, ze szczególnym uwzględnieniem urządzeń pomiarowych i rozwiązań typu „smart" [2][3] doprowadził do opracowania nowego standardu technologii radiowej, tj. NB-IoT (Narrowband Internet of Thing). Komunikacja IoT może definiować różne wymagania na łącza radiowe. Specyficzną częścią IoT są urządzenia niewielkiej mocy wymagające obsługi komunikacji o niewielkich przepływnościach, które instalowane są w miejscach o trudnych warunkach propagacyjnych, tj. piwnicach czy pod ziemią. Mechanizmem użytym w interfejsie NB-IoT do realizacji takiej komunikacji jest wielokrotne powtarzanie transmisji, co wiąże się ze wzrostem zużycia energii [1].

Wysokie zużycie energii spowodowane retransmisjami powinno zatem zostać skompensowane, np. poprzez skrócenie czasu energochłonnego procesu synchronizacji odbiornika, który jest niezbędny do uzyskania przez urządzenie zdolności do komunikacji z siecią NB-IoT [10]. Dlatego też zagadnienie synchronizacji odbiornika stanowi przedmiot tego artykułu.

Niniejszy artykuł został podzielony na pięć rozdziałów. W rozdziale drugim znajdują się informacje ogólne dotyczące NB-IoT, w rozdziale trzecim zaprezentowano metody synchronizacji odbiornika zaprezentowane w literaturze przedmiotu, w rozdziale czwartym – stanowisko pomiarowe, sposób akwizycji danych oraz wyniki przeprowadzonych analiz a w rozdziale piątym podsumowanie i przyszłe prace.

2. INTERFEJS NB-IoT

Interfejs NB-IoT bazuje na rozwiązaniach znanych dotychczas ze standardu LTE (Long-Term Evolution), jednak nie jest w pełni kompatybilny wstecz z urządzeniami pracującymi w tym systemie [5].

Przyjęcie koncepcji bazowania na istniejącej infrastrukturze LTE wpłynęło na czas związany z wdrożeniem tego rozwiązania na potrzeby mMTC (Massive Machine-Type Communications). System działa w licencjonowanych pasmach pracy, ale niewykluczone, że rosnące zapotrzebowanie związane z zapewnieniem łączności bezprzewodowej sprawi, że na potrzeby komunikacji będą używane także pasma nielicencjonowane [8].

Ogólna koncepcja systemu NB-IoT zakłada m.in. łatwe wdrożenie - rozumiane jako integrację z istniejącymi systemami komórkowymi, niski koszt urządzeń, zapewnienie dostępu do usług na rozległym obszarze, także w środowisku trudnym, oraz niski pobór energii. Lista potencjalnych zastosowań NB-IoT jest prawie nieograniczona, ale charakteryzować się one powinny niewielkimi wymaganiami przepływnościowymi, przy wartościach średnich, rzędu pojedynczych bitów czy kilobitów na sekundę.

2.1. Tryby pracy NB-IoT

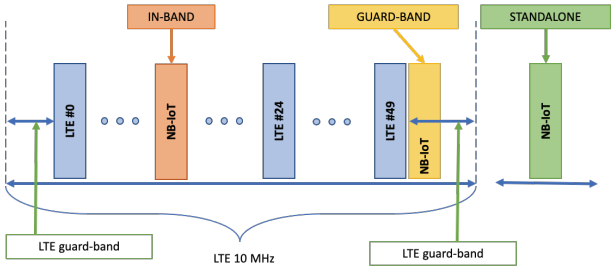
W związku z koniecznością współistnienia z innymi systemami komórkowymi w NB-IoT wyróżnić można trzy tryby pracy:

- standalone (zasoby NB-IoT są ulokowane w paśmie przeznaczonym dla pracy systemu GSM (Global System for Mobile Communication),

- in-band (zasoby NB-IoT są ulokowane jako blok zasobów PRB (Physical Resource Block) systemu LTE,

- guard-band (zasoby NB-IoT są ulokowane w paśmie ochronnym systemu LTE).

Wybór trybu pracy nie wpływa na sposób przeprowadzania procesu synchronizacji odbiornika, gdyż modem NB-IoT nie posiada informacji a priori o tym, w jakim trybie pracuje najbliższa stacja bazowa eNodeB, ale - jak wspomniano - determinuje sposób ulokowania zasobów NB-IoT, co przedstawiono na rys. 1.

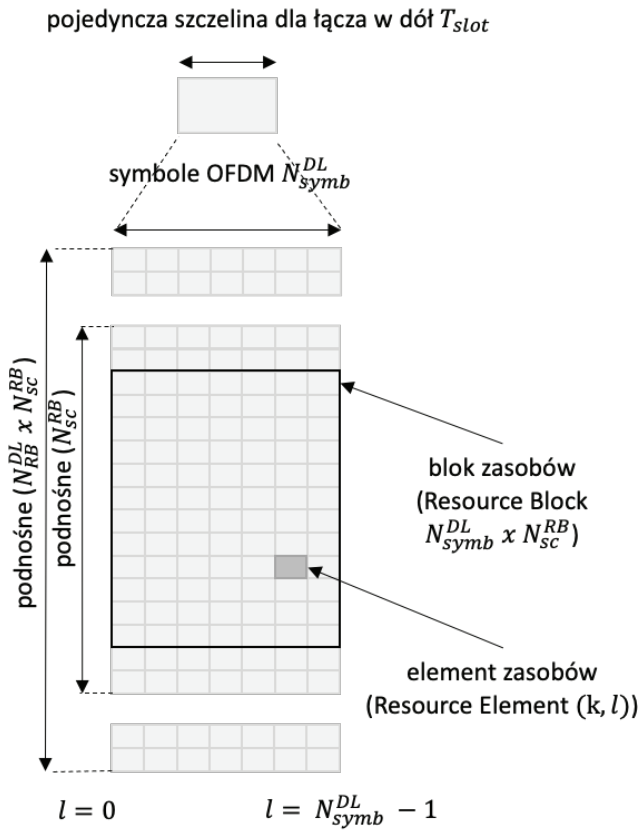


Rys. 1. Graficzna prezentacja sposobu ulokowania zasobów NB-IoT w zależności od trybu pracy [10]

Sygnały NB-IoT muszą być poszukiwane w całym paśmie pracy urządzenia z krokiem co 100 kHz, a im więcej zakresów częstotliwości obsługuje dane urządzenie, tym dłuższy czas poszukiwania, co przekłada się na większe zużycie energii.

2.2. Struktura ramkowa

Dokładnie tak jak w standardzie LTE, w przypadku systemu NB-IoT ramka składa się z dziesięciu podramek a każda z podramek z dwóch szczelin czasowych. Czas trwania szczeliny wynosi 0,5 ms. Wynika stąd, że czas trwania ramki to 10 ms.

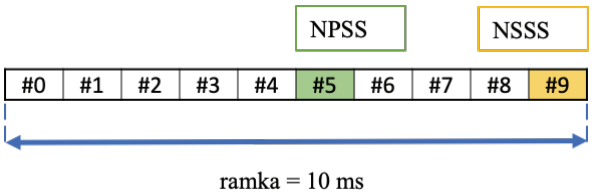


Rys. 2. Siatka zasobów dla łącza w dół [9]

Pojedynczą szczelinę w interfejsie NB-IoT tworzą symbole OFDM (Orthogonal Frequency-Division Multiplexing). Natomiast najmniejszą, niepodzielną część warstwy fizycznej stanowi element zasobów RE (Resource Element). Elementy zasobów pogrupowane są w bloki zasobów RB (Resource Block) a te w siatkę zasobów RG (Resource Grid) [9]. Zależności pomiędzy opisanymi strukturami zaprezentowano na rys. 2.

2.3. Sygnały synchronizacyjne

W systemie NB-IoT występują dwa rodzaje sygnałów synchonizacyjnych, tj. NPSS (Narrowband Primary Synchronization Signal) oraz NSSS (Narrowband Secondary Synchronization Signal). Są one transmitowane przez stację bazową cyklicznie co 10 ms (NPSS) oraz co 20 ms (NSSS). Sygnał NPSS transmitowany jest w #5 podramce a NSSS w podramce #9 [9]. Graficznie przedstawiono to na rys. 3.



Rys. 3. Sposób ulokowania sygnałów NPSS i NSSS w ramce

Sekwencje wzorcowe sygnału NB-IoT generowane są na podstawie sekwencji Zadoff-Chu. W przypadku sekwencji NPSS liczba jej symboli wynosi 121, a dla NSSS liczba ta jest równa 132 [9]. W systemie NB-IoT zdefiniowana została jedna postać sygnału NPSS i wiele postaci sygnału NSSS zależnego od numeru ramki, w której jest nadawany oraz identyfikatora komórki (Cell ID).

3. METODY SYNCHRONIZACJI ODBIORNIKA ZAPREZENTOWANE W LITERATURZE PRZEDMIOTU

W ogólności metody synchronizacji odbiornika można podzielić na trzy grupy:

- metody autokorelacyjne,
- metody korelacyjne,
- metody mieszane.

Korzystając z faktu, że sekwencja NPSS składa się z 11 symboli OFDM i pamiętając o dobrych właściwościach autokorelacyjnych samej sekwencji, zamiast korelacji skrośnej można zastosować autokorelację. Zaletą tego rozwiązania jest brak konieczności generowania sekwencji wzorcowej po stronie odbiorczej, a dodatkowo metoda autokorelacyjna cechuje się dużą tolerancją na odchyłkę częstotliwości.

Metoda korelacyjna opiera się na obliczeniu wartości korelacji wzajemnej pomiędzy sekwencją wzorcową wygenerowaną po stronie odbiorczej a sekwencją odebranych próbek.

Ostatnim z typów metod są metody mieszane, które łączą w sobie cechy dwóch poprzednich rozwiązań. Te metody cechują się zwykle parametrami będącymi wypadkową pomiędzy złożonością implementacyjną i obliczeniową a efektywnością działania w określonym środowisku.

Pierwsza z przebadanych metod synchronizacji odbiornika, oznaczona dalej jako M1, została opisana w artykule [7] i obejmuje obliczanie autokorelacji fragmentów odbieranego sygnału radiowego, z uwzględnieniem odwracania polaryzacji symboli NPSS.

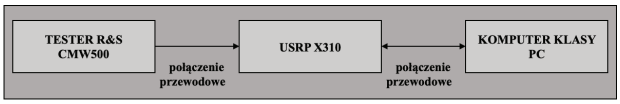
Druga z metod, M2, korzysta z prostego poszukiwania maksimum globalnego funkcji korelacji skrośnej pomiędzy sygnałem odebranym a sygnałem NPSS o pełnej długości i zerowej odchyłce częstotliwości.

Ostatnia z zaimplementowanych metod, M3, opisana w [4] w swoim działaniu łączy cechy opisanych rozwiązań korelacyjnych i autokorelacyjnych.

4. BADANIA POMIAROWE I ICH ANALIZA

Porównanie wyników działania metod synchronizacji odbiornika (M1-M3) zostało dokonane na bazie badań rzeczywistych sygnałów NB-IoT wygenerowanych przy użyciu testera radiokomunikacyjnego. Stanowisko do rejestracji sygnałów NB-IoT składało się z: szerokopasmowego testera radiokomunikacyjnego CMW500, radia programowalnego USRP-X310 i komputera klasy PC.

Wszystkie z wymienionych urządzeń były ze sobą połączone przewodowo w celu zapewniania wyizolowanego środowiska pomiarowego, co zostało przedstawione w postaci schematu blokowego na rys 4.



Rys. 4. Schemat blokowy stanowiska pomiarowego

Sekwencja próbek sygnału odbieranego przez urządzenie USRP - X310 była zapisywana do plików w celu późniejszej analizy.

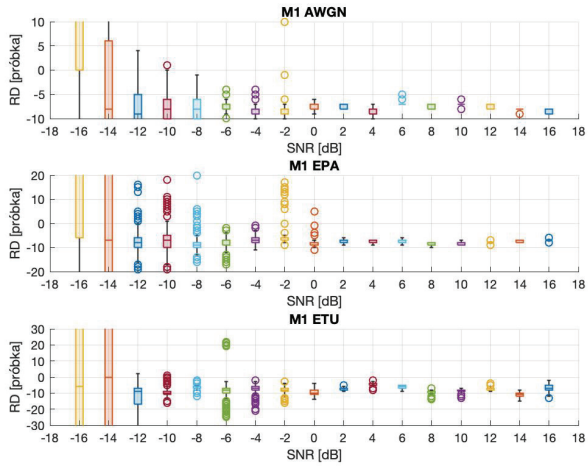
Czas rejestracji pojedynczego pliku wyniósł 30 sekund, ale należy pamiętać, że analizy dotyczące detekcji sygnału synchronizacyjnego przeprowadza się na fragmentach 10-milisekundowych, zgodnie ze strukturą ramkową opisaną w rozdziale drugim. W czasie rejestracji próbek stosowano zmienny poziom SNR (Signal Noise Ratio): na początku rejestrowano fragment sygnału z wysokim poziomem SNR w celu określenia wartości referencyjnej czasu wystąpienia sygnału NPSS, po czym parametr SNR był zmieniany na wartość docelową do badań odchyłki wskazań czasu wystąpienia sygnałów NPSS w zmiennych warunkach odbioru.

Wszystkie parametry konfiguracyjne niezbędne do rejestracji sygnałów zostały zaprezentowane w tab. 1.

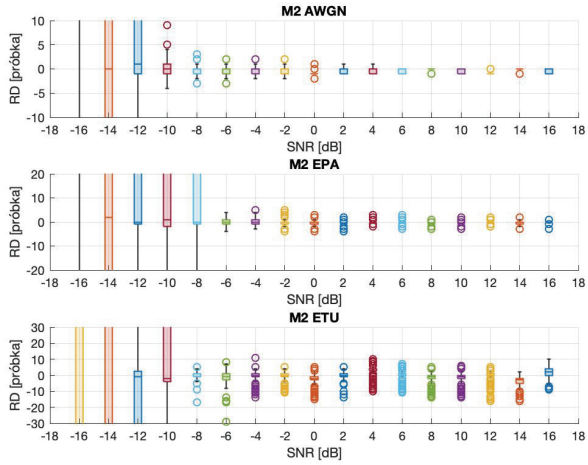
Tab. 1. Parametry konfiguracyjne

Parametr	*Wartość*	*Jednostka*
profil kanału	AWGN, EPA5, ETU1	-
SNR	od -16 do 16 z krokiem co 2	[dB]
NB-IoT Cell ID	100	-
tryb pracy	standalone	-
czas rejestracji	30	[s]
częstotliwość próbkowania	1,92	[MHz]
częstotliwość rejestracji sygnałów	465,5	[MHz]
widmowa gęstość mocy sygnału	-60/15	[dBm/kHz]

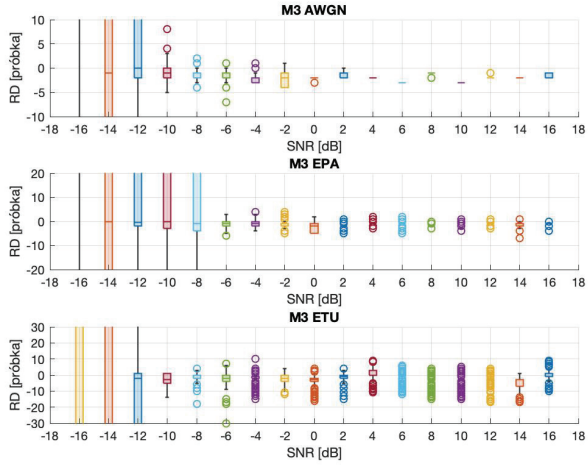
Każda z metod została sprawdzona dla plików zarejestrowanych zgodnie z zaprezentowanymi w tab. 1 parametrami konfiguracyjnymi. Dla każdej z wartości SNR przetestowano tysiąc fragmentów o czasie trwania 10 ms każdy.



Rys. 5. Rozrzut detekcji sygnału NPSS dla metody M1



Rys. 6. Rozrzut detekcji sygnału NPSS dla metody M2



Rys. 7. Rozrzut detekcji sygnału NPSS dla metody M3

Różnica pomiędzy czasem odbioru sygnału NPSS będącego wynikiem działania metody synchronizacji odbiornika a wartością referencyjną wyznaczoną dla fragmentu o wysokim SNR została przedstawiona w postaci wykresu pudełkowego i wyrażona w liczbie próbek na rys. 5-7, gdzie RD oznacza różnicę detekcji sygnału NPSS.

Na podstawie analizy rys. 5-7 można zauważyć, że skuteczność działania metod zależy m.in. od wartości SNR – im niższa wartość SNR, tym większy rozrzut wartości detekcji sygnału NPSS, natomiast rozrzut wartości jest najmniejszy dla profilu AWGN a największy dla profilu ETU.

Dla metody bazującej na autokorelacji (M1) zauważono, że uzyskiwane wartości rozrzutów są do siebie zbliżone pomimo zmieniających się w szerokim zakresie warunków w kanale radiokomunikacyjnym (w przypadku przeprowadzanych analiz jest to wyizolowane środowisko propagacyjne). Dla metody M2 i M3 uzyskane wyniki są prawie bliźniacze, a wynika to z podobieństwa algorytmicznego tych metod.

Ponadto, dla wszystkich zaprezentowanych metod wyraźnie widać próg wartości SNR, poniżej którego występuje znaczny rozrzut czasu detekcji sygnału NPSS. Zjawisko to jest szczególnie widoczne dla wartości SNR < -12 dB, ale jak wspomniano we wstępie, mechanizmem pozwalającym na pracę w warunkach trudnych jest wielokrotna transmisja, bez oczekiwania na żądanie powtórzenia, co jednak wymaga rozszerzenia badań o analizę więcej niż jednego okna 10-milisekundowego.

Przez wzgląd na wymagania dotyczące pracy systemu NB-IoT, metoda synchronizacji odbiornika powinna estymować miejsce wystąpienia sygnału NPSS w sposób precyzyjny (kolejne wartości detekcji sygnału NPSS powinny być ze sobą tożsame), gdyż wszystkie sygnały synchronizacyjne pojawiają się cyklicznie, poza tym pożądana jest również dokładność rozumiana jako stopień zgodności wartości wystąpienia NPSS w kolejnych fragmentach 10-milisekundowych z wartością referencyjną. Dopuszczalna tolerancja synchronizacji docelowej powinna wynosić w przybliżeniu połowę czasu trwania przedrostka cyklicznego.

Pomimo zawężenia przeprowadzonych badań do analizy wyłącznie jednego z sygnałów synchronizacyjnych - NPSS - warto podkreślić, że proces synchronizacji czasowo-częstotliwościowej jest niezbędnym krokiem pozwalającym na dalsze etapy przetwarzania odbieranego sygnału skutkującego odbiorem danych użytecznych.

5. PODSUMOWANIE I PRZYSZŁE PRACE

Przedstawione w rozdziale czwartym wyniki przeprowadzonych dotychczas analiz wykazały, że zaprezentowane w literaturze przedmiotu metody synchronizacji odbiornika (opisane w rozdziale trzecim) różnią się od siebie m.in. pod względem powtarzalności uzyskiwanych wyników detekcji sygnałów NPSS.

Przyszłe prace będą się skupiały na przeprowadzeniu analogicznych rozważań dla innych trybów pracy, tj. in-band i guard-band, rozszerzeniu analiz o badanie tolerancji metod na odchyłki częstotliwości oraz próbie opracowania algorytmu synchronizacji pozwalającego na poprawę prawdopodobieństwa poprawnej detekcji sygnału synchronizacyjnego przy mniejszej złożoności obliczeniowej.

***

Praca została zrealizowana w ramach projektu pt. „Definiowany programowo, uniwersalny interfejs radiowy inteligentnych urządzeń Internetu Rzeczy", współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego, umowa nr POIR.01.01.01-00-1025/19-00 z dnia 2020-05-28.

LITERATURA

[1] Feltrin Luca, Galini Tsoukaneri, Massimo Condoluci, Chiara Buratti, Toktam Mahmoodi, Mischa Dohler, Roberto Verdone. 2019."Narrowband IoT: A Survey on Downlink and Uplink Perspectives." *IEEE Wireless Communications*, 26 (1): 78–86.

[2] GSMA. 2018. "Mobile IoT in the 5G Future NB-IoT and LTE-M in the Context of 5G".

[3] Huawei. 2017. "NB-IoT Commercial Premier Use Case Library Issue 1.0".

[4] Isaacs Omri, Zeev Kaplan. 2018. "NB-IoT System and Method for Radio Cell Synchronization White Paper".

[5] Kanj Matthieu, Vincent Savaux, Mathieu le Guen. 2020. "A Tutorial on NB-IoT Physical Layer Design". *IEEE Communications Surveys & Tutorials,* 22(4): 2408-2446.

[6] Martinez, Borja, Ferran Adelantado, Andrea Bartoli, and Xavier Vilajosana. 2019. "Exploring the Performance Boundaries of NB-IoT". *IEEE Internet of Things Journal*, 6(3): 5702-5712.

[7] Qualcomm Incorporated. 2016. "NB-PSS and NB-SSS Design (Revised)".

[8] Sun Rongrong, Salvatore Talarico, Wenting Chang, Huaning Niu, Hongwen Yang. 2019. "Enabling NB-IoT on Unlicensed Spectrum". *30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*.

[9] Third Generation Partnership Project. 2020."TS 36.211 Version 14.14.0 Release 14".

[10] Wang Eric Y.-P., Xingqpin Lin, Ansuman Adhikary, Asbjorn Grovlen, Yutao Sui, Yufei Blankenship, Johan Bergman, Hazhir S. Razaghi. 2017. "A Primer on 3GPP Narrowband Internet of Things (NB-IoT)." *IEEE Communications Magazine*, 55 (3): 117-123.

LOKALIZOWANIE TERMINALA UŻYTKOWNIKA Z UŻYCIEM INTERFEJSU NB-IoT

LOCALISATION OF THE USER TERMINAL BY USING THE NB-IoT RADIO INTERFACE

Piotr Rajchowski[1]; Krzysztof K. Cwalina[2]; Jarosław Sadowski[3]

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk,
[1]piorajch@eti.pg.edu.pl, [2]kkcwalina@eti.pg.edu.pl, [3]Jaroslaw.Sadowski@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.76

Streszczenie: **W artykule przedstawiono metodę estymacji położenia terminala użytkownika korzystając z sygnałów okazjonalnych nadawanych w łączu w dół w interfejsie NB-IoT. Zweryfikowano praktyczną zdolność estymacji położenia na testowym obszarze korzystając z sygnałów wzorcowych o zmiennej wartości stosunku sygnału do szumu oraz z różnym emulowanym profilem kanału radiowego. Dodatkowo przedstawiono metodę umożliwiającą zwiększenie dokładności estymacji położenia terminala bez konieczności zwiększania częstotliwości próbkowania sygnału radiowego.**

Abstract: **In the article the method of user terminal position estimation using the occasional signals transmitted in the downlink of the NB-IoT interface was presented. The practical ability to estimate the position in the test area was verified using the reference signals with a variable signal-to-noise ratio and different radio channel profiles. Additionally, the sample method is presented that allows to increase the precision of the terminal position estimation without the need of increasing the sampling frequency of the radio signal.**

Słowa kluczowe: **NB-IoT, Radiolokalizacja, Sygnały okazjonalne.**

Keywords: **NB-IoT, Radiolocalization, Signals of Opportunity.**

1. INTERFEJS NB-IoT ORAZ JEGO ZASTOSOWANIA

Rozwój technologii 4G-LTE (Long Term Evolution) oraz 5G-NR (New Radio) jest zauważalnie ukierunkowany na zwiększenie pojemności sieci oraz umożliwienie realizowania szerokiego spektrum usług. Widoczne jest, iż wysoki nacisk kładziony jest na umożliwienie przesyłania danych z dużą szybkością oraz z niskim opóźnieniem. Niemniej jednak, wraz z publikacją specyfikacji 3GPP (3rd Generation Partnership Project) w wersji 13, zaproponowano nowy interfejs radiowy NB-IoT (Narrowband Internet of Things) dedykowany do realizacji usług skrajnie odmiennych, tj. w warunkach stacjonarnych oraz w trudnych warunkach propagacyjnych, z przepływnością na poziomie 250 kb/s w łączu w górę. Wprowadzone modyfikacje względem 4G-LTE, zarówno w warstwie fizycznej jak i warstwach wyższych, umożliwiają realizację usług w warunkach o mocno utrudnionej propagacji sygnałów radiowych pozwalając uzyskać parametr maximum coupling loss (MCL) na poziomie 164 dB [1,4,7].

Lokalizowanie terminala z użyciem interfejsu NB-IoT wydaje się być procesem naturalnym, zwłaszcza w kontekście interfejsu LTE, lecz niezwykle trudnym mając na względzie jego założenia techniczne – np. pasmo pracy wynoszące 180 kHz oraz częstotliwość próbkowania sygnału radiowego równą 1,92 MHz. Aktualne prace badawcze skupiają się na przedstawieniu sposobu lokalizowania terminala użytkownika oraz wskazaniu stopnia degradacji dokładności estymacji położenia względem interfejsu LTE. W pełni uzasadnione wydaje się przeprowadzenie dalszych badań w zakresie analizy dokładności estymacji położenia terminala użytkownika biorąc pod uwagę sygnały rzeczywiste oraz mechanizm przetwarzania danych w torze odbiorczym terminala NB-IoT.

2. LOKALIZOWANIE TERMINALA Z UŻYCIEM INTERFEJSU NB-IoT

2.1. Metody lokalizowania terminala użytkownika z użyciem interfejsu NB-IoT

Specyfikacja techniczna 3GPP interfejsu NB-IoT definiuje w sygnale dla łącza w dół trzy sygnały referencyjne: NPSS (Narrowband Primary Synchronization Signal), NSSS (Narrowband Secondary Synchronization Signal) oraz NRS (Narrowband Reference Signal), używane odpowiednio do synchronizacji czasu i częstotliwości, określania identyfikatora komórki NB_{Cell_ID} oraz do estymacji stanu kanału radiowego. Możliwe jest także umieszczenie w podramce radiowej sygnału dedykowanego do procesu lokalizowania terminala użytkownika, tj. sygnału NPRS (Narrowband Positioning Reference Signal), który to przyjmuje postać zespolonych symboli umieszczanych w odpowiednich miejscach siatki zasobów [1,4,10]. Niestety, obserwując działania operatorów w interfejsie LTE, np. na terenie kampusu Politechniki Gdańskiej, emisja sygnałów dodatkowych często nie jest realizowana w celu przeznaczenia elementów zasobów na potrzeby kanału PDSCH (Physical Downlink Shared Channel). Podobna sytuacja ma miejsce w interfejsie NB-IoT, gdzie emisja sygnału NPRS jest niekorzystna z biznesowego punktu widzenia, gdyż specyfikacja techniczna narzuca tworzenie tzw. ramek wyciszonych w których poza emisją sygnałów synchronizacyjnych nie następuje emisja w kanale NPDSCH (Narrowband Physical Downlink Shared Channel). Mając niniejsze na uwadze zdecydowano się zweryfikować możliwość estymacji położenia terminala użytkownika bazując na sygnałach okazjonalnych, w tym przypadku nadawanym periodycznie w każdej ramce radiowej sygnale NPSS [9].

2.2. Analizowana metoda lokalizowania terminala użytkownika

Lokalizowanie terminala użytkownika z użyciem sygnału NPRS polega na przeprowadzeniu estymacji czasu jego wystąpienia w sygnale odebranym, np. z użyciem korelacji skrośnej. Zatem proces ten jest koncepcyjne tożsamy z operacją synchronizacji czasu i częstotliwości bazującą na sygnale referencyjnym NPSS. Zdecydowano się, że czas odbioru sygnału radiowego z różnych stacji eNodeB (Evolved Node B) będzie estymowany na podstawie wykrytego położenia sygnału NPSS z każdej ze stacji eNodeB, a dodatkowo będzie on powiązany z obligatoryjną operacją synchronizacji częstotliwości.

Lokalnie wytworzony referencyjny sygnał NPSS korelowano z odebranym sygnałem radiowym jednocześnie wprowadzając hipotezy występującej odchyłki częstotliwości, tj. z krokiem 100 kHz oraz z krokiem 10 Hz w zakresie +/- 15 kHz. Proces ten umożliwił określenie estymaty występującej odchyłki częstotliwości odbiornika względem eNodeB, jej skompensowanie oraz wyznaczenie estymaty czasu odbioru sygnału radiowego [1,4].

Wyznaczone estymaty czasów odebrania sygnału radiowego dla każdej z eNodeB stanowiły wejściowe dane pozycyjne dla algorytmu estymacji położenia bazującego na metodzie TDOA (Time Difference of Arrival) oraz algorytmie Foya [5,6].

Z powodu wąskiego pasma sygnału radiowego oraz niskiej częstotliwości próbkowania dokładność estymacji położenia terminala użytkownika w interfejsie NB-IoT jest z góry ograniczona. Zdecydowano się przeanalizować możliwość zwiększenia dokładności estymacji położenia stosując metodę estymacji maksimum funkcji korelacji skrośnej sygnału odebranego z referencyjnym sygnałem NPSS. Niniejszą aproksymację zrealizowano stosując przybliżenie kształtu fragmentu maksimum korelacji funkcją kwadratową $\widehat{K}(\tau) = a\tau^2 + b\tau + c$, gdzie współczynniki funkcji $[a, b, c]$ wyznaczono korzystając z równania (1) i (2)

$$T = \begin{bmatrix} \tau_1^2 & \tau_1 & 1 \\ \tau_2^2 & \tau_2 & 1 \\ \tau_3^2 & \tau_3 & 1 \end{bmatrix}, R = \begin{bmatrix} K(\tau_1) \\ K(\tau_2) \\ K(\tau_3) \end{bmatrix} \quad (1)$$

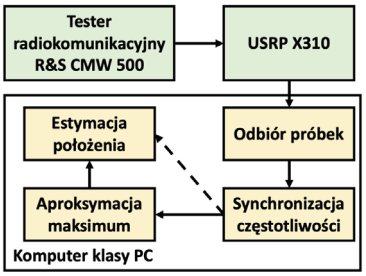
$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = T^{-1} \cdot R, \quad (2)$$

gdzie $K(\tau_1), K(\tau_2), K(\tau_3)$ są sąsiednimi wartościami funkcji korelacji, a wartość $K(\tau_2)$ jest wartością największą w uzyskanym pomiarze. Wynikiem aproksymacji jest estymata $\widehat{K}(\tau)$ czasu odbioru sygnału NPSS, dla której funkcja kwadratowa o współczynnikach (2) przyjmuje wartość maksymalną [2,5].

3. KAMPANIA POMIAROWA

3.1. Stanowisko laboratoryjne

Na potrzeby prowadzonych badań wymagane było posiadanie próbek sygnałów radiowych z kontrolowanym wpływem warunków środowiskowych tak aby możliwe było wiarygodne ocenienie procesu detekcji sygnałów referencyjnych używanych w procesie estymacji położenia terminala. Zdecydowano się wygenerować zestaw sygnałów testowych za pomocą testera radiokomunikacyjnego Rohde&Schwarz CMW500. Używając modułu emulacji interfejsu radiowego NB-IoT wygenerowano sygnał radiowy dla trybu pracy Standalone, jednocześnie modelując trzy profile kanału, EPA-5 (Extended Pedestrian A model), ETU-1 (Extended Typical Urban) oraz AWGN (Additive White Gaussian Noise). Dla każdego z profili kanału dokonywano zmiany wartości SNR (Signal to Noise Ratio), gdzie zmniejszenie wartości parametru SNR oznaczało dodanie filtrowanego szumu gaussowskiego do sygnału radiowego o widmowej gęstości mocy -60 dBm/15 kHz. Wytworzony sygnał radiowy w paśmie wysokich częstotliwości rejestrowano za pomocą utworzonego laboratoryjnego stanowiska pomiarowego, którego schemat przedstawiono na Rys. 1.



Rys. 1. Schemat blokowy stanowiska pomiarowego

Sygnał radiowy w postaci zestawu cyfrowych próbek pozyskiwano za pośrednictwem urządzenia typu USRP (Universal Software Radio Peripheral) Ettus-Research-X310 stosując natywną częstotliwość próbkowania dla interfejsu NB-IoT tj. 1,92 MHz. Do rejestracji oraz analizy próbek użyto opracowanego oprogramowania, zarówno w zakresie synchronizacji czasu i częstotliwości jak i estymacji położenia terminala.

3.2. Scenariusze testowe

Za cel prowadzonych badań przyjęto zweryfikowanie dokładności estymacji położenia terminala użytkownika na podstawie sygnałów NPSS nieprzeznaczonych do celów lokalizacyjnych, gdzie rolę terminala w rzeczywistych rozwiązaniach pełni stacjonarne urządzenie telemetryczne. Wykonano zatem rejestrację sygnałów NB-IoT wygenerowanych w taki sposób, by emulowany był obszar testowy w kształcie prostokąta o długości boków 1500 m oraz 3000 m. Na tymże obszarze rozmieszczono siatkę punktów kontrolnych (PK) odległych od siebie o 100 m, a wokół obszaru rozmieszczono trzy stacje eNodeB. Kształt obszaru testowego wraz z oznaczonym położeniem stacji eNodeB przedstawiono na rys. 2 w sekcji 4.1.

Dokładność estymacji położenia terminala użytkownika weryfikowano na podstawie danych pomiarowych przypisanych do każdego z PK. Danymi tymi były czasy propagacji sygnału radiowego od każdej z eNodeB do PK (sygnały testowe dla trzech eNodeB w każdym PK były innymi fragmentami zarejestrowanego sygnału radiowego) zaokrąglone do liczby całkowitej będącej wielokrotnością okresu próbkowania t_{pr} = 1/30,72 MHz. Odzwierciedla to liczbę próbek n o którą zostały opóźnione sygnały testowe odpowiadające jednej ramce radiowej sy-

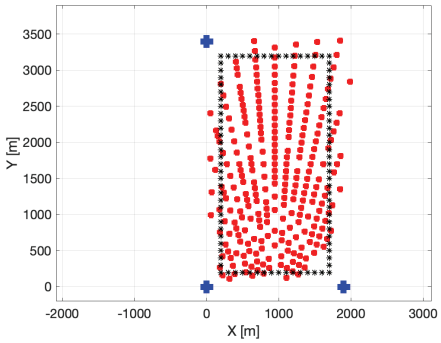
gnału LTE (n + 19200), które to stanowiły fragmenty sygnału radiowego wygenerowanego przez tester radiokomunikacyjny. Warto dodać, iż sposób generowania sygnałów NB-IoT w rzeczywistych eNodeB nie będzie bowiem pozwalał w wielu konstrukcjach urządzeń na kontrolę zależności czasowych w interfejsie radiowym z większą rozdzielczością niż przyjęta w prezentowanych badaniach [4].

4. ANALIZA UZYSKANYCH WYNIKÓW

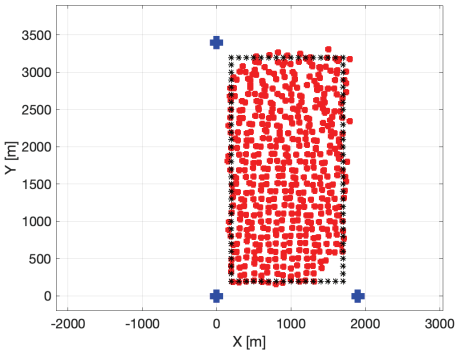
Analizę uzyskanych wyników podzielono na określenie dokładności estymacji położenia interpretowanej jako wierność wyznaczonych estymat położenia względem siatki PK oraz numeryczną ocenę dokładności estymat położenia względem koordynat PK.

4.1. Analiza dokładności estymacji położenia

Na rys. 2 oraz 3 przedstawiono wyznaczone estymaty położenia, jednocześnie obrazując granice obszaru testowego (punkty koloru czarnego) oraz stacje eNodeB (punkty koloru niebieskiego). Przy braku zastosowania aproksymacji maksimum funkcji korelacji widoczne jest, iż zbiory wyznaczonych estymat układają się wzdłuż hiperbol pozycyjnych zdefiniowanych skończoną rozdzielczością pomiaru zależności czasowych między sygnałami (rys. 2). Dla lepszego zobrazowania tego zjawiska wyrysowano estymaty obrazując wynik procesu lokalizowania z użyciem sygnału wejściowego o SNR równym 16 dB i profilu kanału typu AWGN.



Rys. 2. Wyznaczone estymaty położenia bez aproksymacji maksimum korelacji

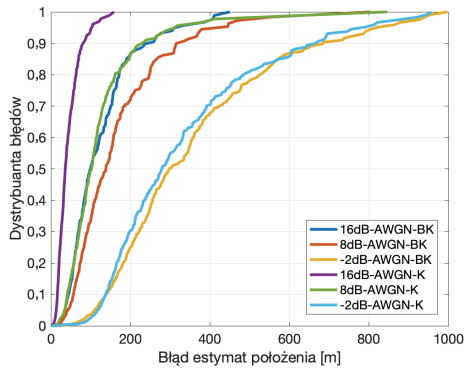


Rys. 3. Wyznaczone estymaty położenia z zastosowaniem aproksymacji maksimum korelacji

Jak zauważyć można na rys. 3, zastosowanie aproksymacji opisanej w podrozdziale 2.2 spowodowało, iż uzyskano wierniejsze odwzorowanie sitaki PK, co także spowodowało zmniejszenie błędu wyznaczonych estymat położenia.

4.1. Numeryczna ocena dokładności estymacji położenia

Numeryczną analizę dokładności estymacji położenia przeprowadzono dla sygnałów wejściowych o różnym stosunku SNR oraz różnym modelowanym wpływie kanału radiowego (AWGN, ETU-1, EPA-5). Na rys. 4 przedstawiono dystrybuanty błędów wyznaczonych estymat położenia dla profilu AWGN oraz trzech wybranych wartości parametru SNR – [16 dB, 8 dB, -2 dB]. Widoczne jest, że dla sygnału o dobrej jakości (SNR 16 dB) zastosowanie dodatkowej aproksymacji (oznaczenie K, bez aproksymacji: BK) umożliwia zmniejszenie błędu estymacji położenia z 258 m do 81 m dla 90% przypadków, i odchyleniu standardowym σ z 61 m do 21 m. Dla sygnałów o niskiej jakości (np. SNR -2 dB) widoczna jest znaczna degradacja dokładności estymacji położenia (687 m dla 90% przypadków, σ = 159) oraz zauważalne jest występowanie przypadków, gdy algorytm Foya nie uzyskuje zbieżności. Dla przedstawionego przypadku w czterech procentach analizowanych danych pomiarowych wyznaczone estymaty położenia były odległe od rzeczywistej lokalizacji PK o ponad 1000 metrów.

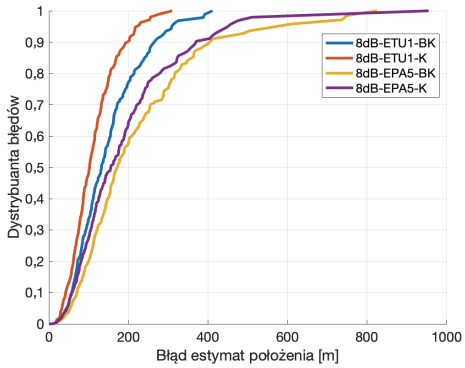


Rys. 4. Dystrybuanta błędów wyznaczonych estymat położenia dla profilu kanału typu AWGN

W kolejnej fazie badań zweryfikowano dokładność estymacji położenia, kiedy to do zamodelowania wpływu kanału użyto modelu typu ETU-1 oraz EPA-5. Na rys. 5 przedstawiono dystrybuanty błędów dla sygnału o SNR równym 8 dB. W tym przypadku uzyskano zwiększenie dokładności estymacji położenia z 267 m (σ = 69,03 m) do 202 m (przy σ = 50,46 m) i 405 m (σ = 112,18 m) do 370 m (σ = 98,44 m) dla 90% przypadków i odpowiednio profilu ETU-1 i EPA-5.

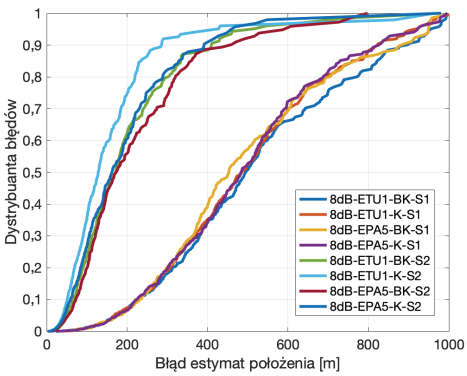
W ostatniej fazie badań zweryfikowano stopień degradacji dokładności estymacji położenia wprowadzając nieidealność synchronizacji czasowej stacji eNodeB. Stopień niezsynchronizowania każdej eNodeB zamodelowano rozkładem normalnym z wartością średnią $\mu = 0$ i $\sigma^2 = 2{,}25 \cdot 10^{-12}$ co odpowiada stopniowi zsynchronizowania stacji eNodeB na poziomie 1,5 µs (przypadek S1) oraz rozkładem normalnym o parametrach $\mu = 0$ i $\sigma^2 = 1 \cdot 10^{-14}$ co odpowiada stopniowi zsynchronizowania stacji eNodeB na poziomie 0,1 µs (przypadek S2). Stopień zsynchronizowania stacji eNodeB na poziomie

1,5 µs jest wartością oczekiwaną dla komórek o promieniu do 3 km, zaś przyjęta wartość 0,1 µs jest oczekiwana dla sieci LTE-A (LTE-Advanced) [3].



Rys. 5. Dystrybuanta błędów wyznaczonych estymat położenia dla profilu kanału typu ETU-1 oraz EPA-5

Dystrybuanty błędów estymat położenia dla analizy dwóch przypadków różnego stopnia zsynchronizowania stacji eNodeB przedstawiono na rys. 6. Widoczne jest, iż przy wyższym stopniu zsynchronizowania eNodeB dodatkowa aproksymacja maksimum funkcji korelacji umożliwia zmniejszenie błędu estymacji położenia, lecz różnica pomiędzy przypadkiem S1 a S2 wynosi odpowiednio 822 m oraz 286 m dla 90% przypadków i profilu kanału typu ETU-1. Warto jednak wskazać, że rzeczywiste eNodeB mają często wysoki stopień powtarzalności odchyłek synchronizacji czasu, co stwarza możliwość do wprowadzania estymowanych korekt do czasów detekcji sygnałów z poszczególnych stacji.



Rys. 6. Dystrybuanta błędów wyznaczonych estymat położenia dla profilu kanału typu ETU-1 oraz EPA-5 z wpływem nieidealnej synchronizacji stacji eNodeB

1. PODSUMOWANIE

W niniejszym artykule przedstawiono proces lokalizowania terminala użytkownika z użyciem interfejsu NB-IoT bazują na tzw. sygnałach okazjonalnych, w tym przypadkach sygnale synchronizacyjnym NPSS. Przedstawiono, że możliwe jest wyznaczenia estymat położenia terminala użytkownika na przykładowym obszarze miejskim uzyskując błąd estymat położenia poniżej 370 m dla 90% przypadków oraz 202 m dla 90% przypadków, odpowiednio dla profilu kanału EPA-5 oraz ETU-1 przy SNR sygnału odbieranego na poziomie 8 dB. Wykazano, że możliwe jest zwiększenie dokładności estymacji położenia o 65 m, 90% przypadków ETU-1, SNR 8 dB, bez konieczności zwiększania częstotliwości próbkowania, w zamian dokonując aproksymacji maksimum funkcji korelacji skrośnej sygnału odbieranego.

Wyniki przedstawione w niniejszym rozdziale wskazują, iż zasadne jest prowadzenie dalszych badań w zakresie możliwości zwiększenia dokładności estymacji położenia terminala użytkownika, zwłaszcza w warunkach o utrudnionej propagacji sygnałów radiowych i np. obecnych interferencjach zewnątrz i wewnątrzsystemowych. Dalsze prace badawcze będą także obejmowały zweryfikowanie funkcjonowania mechanizmów długoterminowego śledzenia dryftu częstotliwości terminala oraz stacji eNodeB, bezpośrednio przekładających się na dokładność estymacji położenia.

***

Praca została zrealizowana w ramach projektu pt. „Definiowany programowo, uniwersalny interfejs radiowy inteligentnych urządzeń Internetu Rzeczy", współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego, umowa nr POIR.01.01.01-00-1025/19-00 z dnia 2020-05-28.

LITERATURA

[1] 3GPP. 2020. „3rd Generation Partnership Project; LTE, Evolved Universal Terrestrial Radio Access (E-UTRA), Physical channels and modulation". 3GPP TS 36.211, V14.14.0.
[2] Bao-Yen T. J. 2000. Fundamentals of Global Positioning System Receivers: A Software Approach. John Wiley & Sons.
[3] Goetz I. 2016. „Timing; The Key Unlocking the Benefits of LTE-A, Timing & Security Considerations for Evolved IP Backhaul". Juniper white paper.
[4] Liberg. O i inni. 2020. Cellular Internet of Things, From Massive Deployments to Critical 5G Applications. Academic Press.
[5] Sadowski. J. 2018. Badanie i analiza systemów radiolokalizacyjnych do zastosowań specjalnych. Politechnika Gdańska.
[6] Sha H. L., Xuhong L., Rusek F., 2019. „On Time of Arrival Etimation in NB-IoT Systems". IEEE Wireless Communications and Networking Conference. 1-6.
[7] Rodriguez I., H. C. Nguyen, I. Z. Kovacs, T. B. Sørensen, P. Mogensen. 2016. „An Empirical Outdoor-to-Indoor Path Loss Model From Below 6 GHz to cm-Wave Frequency Bands". IEEE antennas and wireless propagation letters. vol.16 : 1329-1332.
[8] Tong F. ,H. Shibo. 2019. „On Positioning Performance for the Narrowband Internet of Things: How Participating eENBs Impact?". IEEE Transactions on Industrial Informatics. vol. 15, 423-433.
[9] Xingquin L. i inni. 2017. „Positioning for the Internet of Things: A 3GPP Perspective". IEEE Communications Magazine. vol.55 : 179-185.
[10] Yuxiang L. i inni. 2019. „An Efficient Synchronization Algorithm in OFDM-Based NB-IoT Systems". IEEE International Symposium on Dependable, Autonomic and Secure Computing (DASC). 459-464.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

ANALIZA EFEKTYWNOŚCI METOD GŁĘBOKIEGO UCZENIA W ODBIORZE SYGNAŁÓW GMSK

EFFICIENCY ANALYSIS OF THE DEEP LEARNING-BASED MODELS FOR GMSK SIGNAL DETECTION

Alicja Olejniczak[1]; Krzysztof K. Cwalina[2]; Jarosław Sadowski[3]; Jacek Stefański[4]

[1] Politechnika Gdańska, Wydział Elektroniki Telekomunikacji i Informatyki, Gdańsk, alicja.olejniczak@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki Telekomunikacji i Informatyki, Gdańsk, kkcwalina@eti.pg.edu.pl
[3] Politechnika Gdańska, Wydział Elektroniki Telekomunikacji i Informatyki, Gdańsk, jaroslaw.sadowski@eti.pg.edu.pl
[4] Politechnika Gdańska, Wydział Elektroniki Telekomunikacji i Informatyki, Gdańsk, jacek.stefanski1@pg.edu.pl

DOI: 10.15199/59.2022.4.77

Streszczenie: **Sztuczna inteligencja odnajduje coraz szersze zastosowanie we współczesnej radiokomunikacji, choć głównie w ujęciu badawczym. Niniejszy artykuł przedstawia przegląd i ewaluację metod z obszaru głębokiego uczenia umożliwiających detekcję sygnałów z modulacją GMSK (ang. *Gaussian Minimum Shift Keying*) w kanale AWGN. Badane modele porównane zostały z optymalnym detektorem pracującym zgodnie z regułą największej wiarygodności MLSE (ang. *Maximum Likelihood Sequence Estimation*), a miarą oceny efektywności była bitowa stopa błędu.**

Abstract: **Artificial intelligence dynamically enters the realm of radiocommunication, however, mainly in the scientific context. The paper presents evaluation of the deep learning (DL) models for Gaussian Minimum Shift Keying (GMSK) signal demodulation in AWGN channel. The proposed DL methods were compared with optimal Maximum Likelihood Sequence Estimation (MLSE) detector based on Bit Error Rate (BER) metric.**

Słowa kluczowe: **EC-GSM-IoT, głębokie uczenie, GMSK, sztuczna inteligencja**

Keywords: **EC-GSM-IoT, deep learning, GMSK, artificial intelligence**

1. WSTĘP

Zastosowanie modeli uczenia maszynowego w radiokomunikacji jest zjawiskiem coraz powszechniejszym. W kontekście rozwoju nowych technologii, jak choćby IoT (ang. *Internet of Things*) czy 5G, wprowadzanie rozwiązań umożliwiających przetwarzanie sygnałów w czasie rzeczywistym, w trudnych i/lub zmiennych warunkach propagacyjnych, przy jednoczesnym zwiększaniu przepływności i zachowaniu jakości usług staje się zadaniem tyleż wymagającym, co kluczowym. Stąd też zainteresowanie gałęzią informatyki, jaką jest głębokie uczenie DL (ang. *deep learning*), pozwalające na analizę zagadnień o charakterze nieliniowym, które to potwierdziło wysoką skuteczność w wielu zastosowaniach praktycznych, m.in. związanych z analizą obrazu [7].

W literaturze odnaleźć można liczne prace badające możliwość zastosowania uczenia maszynowego w różnych obszarach radiokomunikacji [19]: rozpoznawanie modulacji w radiu kognitywnym [3][4], rozwiązania lokalizacyjne [8], zarządzanie zasobami sieci [12]. Z uwagi na tematykę niniejszego artykułu, szczególnie istotne są jednak badania dotyczące DL w zakresie odbioru radiowego.

W [17] przedstawiono model 5-warstwowej sieci FNN (ang. *Feedforward Neural Network*), która realizowała estymację kanału i detekcję sygnałów OFDM z modulacją QPSK (ang. *Quadrature Phase Shift Keying*). Zaproponowane rozwiązanie przetestowane zostało dla danych symulacyjnych, a porównanie z klasycznymi metodami LS (ang. *Least Squares*) i MMSE (ang. *Minimum Mean-Square Error*) zrealizowano na podstawie uzyskanej bitowej stopy błędów BER (ang. *Bit Error Rate*).

Problem poprawy odbioru radiowego ujmowany jest w badaniach również jako tzw. system end-to-end, tj. system w którym modele DL zastępują cały blok nadawczo-odbiorczy, realizując kolejno funkcje nadajnika, modelowania i korekcji kanału oraz odbiornika. W [18] przeanalizowano 7-warstwową strukturę w koncepcji end-to-end dla transmisji z modulacją 16 QAM (ang. *Quadrature Amplitude Modulation*) dla kanału AWGN i kanału z zanikami Rayleigha. Z uwagi na ograniczoną długość analizowanych fragmentów transmisji (7 bitów), problem odbioru uproszczono do problemu klasyfikacji – 2^7 możliwych klas. W [14] zaproponowano nadajnik-odbiornik przetwarzający równolegle pojedyncze symbole, z dodatkowym określeniem synchronizacji na podstawie dodanych sygnałów pilotowych. Autorzy uwzględnili w modelu rotację fazy analizowanych sygnałów. Rozwiązanie o 14-warstwowej architekturze przebadane zostało w środowisku pomiarowym wykorzystującym radio programowalne SDR (ang. *Software Defined Radio*).

W literaturze odnaleźć można jeszcze wiele artykułów podejmujących zagadnienie odbioru i demodulacji sygnałów radiowych przy pomocy zarówno sieci FNN, jak i splotowych sieci neuronowych CNN (ang. *Convolutional Neural Network*), jednak najczęściej obejmują one technikę OFDM i modulacje fazy [2] [5][11] [16] [20] lub częstotliwości [10][13].

W przeciwieństwie do badań obejmujących identyfikację modulacji [3][4], w dostępnej literaturze brak analizy demodulacji sygnałów GMSK (ang. *Gaussian Minimum Shift Keying*). Mając na uwadze rozwój IoT, który stał się przyczynkiem do rozszerzenia przez 3GPP standardu GSM (Release 13) o technologię EC-GSM-IoT

(ang. *Extended Coverage GSM IoT*) [1][9], zasadnym zdaje się przebadanie efektywności metod DL w odbiorze sygnałów GMSK – modulacji stosowanej we wspomnianym standardzie.

W dalszej części niniejszego artykułu opisana zostanie metodyka doboru i sposobu generowania danych testowych GMSK oraz podstawowe zagadnienia z obszaru głębokiego uczenia. Następnie, zaprezentowane zostaną modele DL (sieci FNN i CNN) wybrane do ewaluacji i charakterystyki BER uzyskane dla proponowanych rozwiązań. Wyniki porównane zostaną z optymalnym detektorem pracującym zgodnie z regułą największej wiarygodności MLSE (ang. *Maximum Likelihood Sequence Estimation*). W ostatniej części znajdzie się podsumowanie przedstawionych badań oraz wnioski.

2. STRUKTURA PAKIETÓW GMSK

Jak już zostało wspomniane we wstępie, do analizy skuteczności modeli DL w odbiorze sygnałów radiowych wybrane zostały sygnały GMSK. Postać sygnału opisać można następująco [15]:

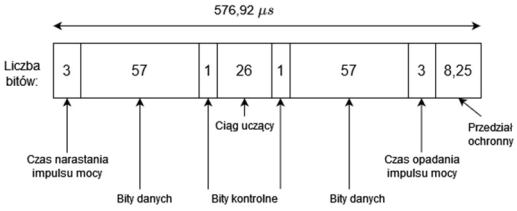
$$s(t) = \sum_{k=0}^{K-1} s_k(t) \tag{1}$$

gdzie $s_k(t)$ stanowi k-ty zmodulowany segment K segmentów:

$$s_k(t) = \begin{cases} A_0 \sin \Phi_k(t) \text{ dla } (k-\frac{1}{2})T_b \le t \le (k+\frac{1}{2})T_b \\ 0 \quad \text{dla pozostałych } t \end{cases} \tag{2}$$

gdzie T_b stanowi czas trwania segmentu zmodulowanego, A_0 reprezentuję amplitudę sygnału, a $\Phi_k(t)$ fazę chwilową sygnału (opisaną funkcją gaussowską), w której składowa stała dobierana jest tak, by zapewnić ciągłość fazy na stykach kolejnych przetwarzanych segmentów.

Z uwagi na zastosowanie modulacji GMSK w standardzie GSM, dane symulacyjne wygenerowano zgodnie z wytycznymi dla kanału TCH/F (ang. *Traffic Channel/Fullrate*) [15]. Ogólną strukturę pojedynczego pakietu zaprezentowano na Rys. 1.



Rys. 1. Struktura pakietu w kanale TCH/F [15]

W ramach zbioru treningowego dla modeli DL przyjęto trzy różne ciągi uczące i dla każdego z nich wygenerowano po 34 068 pakietów. W toku ewaluacji modeli, dla niższych wartości E_b/N_0 zbiory zwiększono do 54 108 pakietów, tj. ponad 162 tysięcy pakietów w pełnym zbiorze uczącym.

W następnym kroku, do wszystkich pakietów dodany został szum AWGN, dla przyjętego zakresu E_b/N_0 1-12 dB. Na tym etapie badań pominięte zostało zagadnienie synchronizacji, wobec czego jako dane wejściowe do sieci neuronowej przygotowane zostały 630 próbki sygnału zmodulowanego GMSK, po filtracji filtrem przeddetekcyjnym Butterwortha 4-rzędu [15] i wstępnej synchronizacji czasowej.

3. GŁĘBOKIE UCZENIE

W najprostszej postaci, wyjście pojedynczego neuronu klasycznej sieci FNN można opisać jako:

$$y_1 = f(\sum_{i=1}^{n} x_i w_i + b_0) \tag{3}$$

gdzie x_i reprezentuje i-tą wartość wejściową, a w jest zbiorem n współczynników wagowych dla danej wartości wejściowej. Do iloczynu $x_i w_i$ dodawany jest współczynnik b_0, (ang. *bias*). Suma wszystkich wartości wejściowych (wraz z korespondującymi współczynnikami wagowymi) jest następnie przetwarzana przez funkcję aktywacji f, w ogólności nieliniową.

Wartości wyjściowe z neuronów w ostatniej warstwie sieci są w procesie uczenia porównywane z wartościami wzorcowymi a minimalizacja zadanej funkcji straty dla obu wartości umożliwia aktualizację współczynników wagowych sieci w kolejnej iteracji. Liczba neuronów, warstw sieci, iteracji, funkcje aktywacji czy krok modyfikacji współczynników wagowych to parametry, których dobór stanowi ważny etap projektowania sieci i odbywa się tak by dostosować model do podjętego problemu badawczego i wyodrębnić architekturę optymalną [6].

Oprócz modelu FNN, w niniejszym artykule badana będzie również sieć splotowa CNN, która umożliwia analizę danych wielowymiarowych. Ten typ sieci intuicyjnie wydaje się być odpowiedni do przetwarzania sygnałów radiowych, w których zależności między próbkami powodowane interferencjami międzysymbolowymi są niezmienne w czasie trwania pakietu.

Oprócz klasycznego sposobu przekształcania danych, sieć CNN przemiata płaszczyzny danych wejściowych i dokonuje operacji splotu. Rozmiar obszaru przemiatania zwany jest filtrem i wraz z krokiem przesuwu filtra stanowi parametry dobieralne modelu [6].

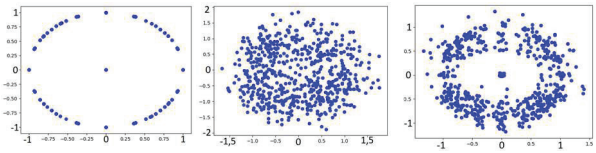
4. EWALUACJA MODELI DL

4.1. Redukcja szumu sygnału GMSK

W ramach niniejszych badań, w pierwszym etapie podjęto próbę redukcji szumu w sygnale odebranym. W tym celu, na wejście modeli DL przekazywany był zaszumiony sygnał GMSK, a wyjście porównywane było ze zmodulowanym sygnałem oryginalnym (bez szumu). Jako metrykę różnicy pomiędzy predykcją modelu $\hat{s}$ a pakietem referencyjnym s wybrano MSE (ang. *Mean Squared Error*) [6]:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(s_i - \hat{s}_i)^2 \tag{4}$$

Przykładowe konstelacje dla modelu CNN zaprezentowano na Rys. 2.



Rys. 2. Przykładowe konstelacje pakietów: oryginalnego, zaszumionego i po redukcji szumu siecią CNN

Niestety, pomimo wizualnej poprawy konstelacji, żaden z badanych modeli nie był w stanie odwzorować zależności fazowych sygnału, co skutkowało brakiem poprawy w odniesieniu do bitowej stopy błędów (dla E_b/N_0=12 dB BER rzędu 10^{-2}). Zdecydowano więc o redukcji obszaru predykcji, poprzez redefiniowanie jej do postaci detekcji sygnału. Zmiana postaci informacji estymowanej przez sieć oraz pominięcie ciągu uczącego umożliwiło zmniejszenie wymiaru wyjściowego modeli ponad dwukrotnie.

4.2. Detekcja sygnału GMSK

W ramach detekcji pakietów TCH/F przyjęto rozmiary danych opisane w rozdziale 2. Dokonywano estymacji wartości binarnych, tj. 114 bitów, z pominięciem stałych sekwencji bitowych (Rys.1). Z uwagi na dość duży rozmiar wektora wyjściowego, zrezygnowano ze sprowadzania niniejszego zagadnienia do klasyfikacji pełnego pakietu. Jako metrykę porównującą bity nadane i odebrane przez modele DL przyjęto funkcję BCEloss (ang. *Binary Cross Entropy Loss*):

$$BCE = -\frac{1}{m}\sum_{i=1}^{m}(y_i \log\widehat{y_i} + (1-y_i)\cdot\log(1-\widehat{y_i})), \qquad (5)$$

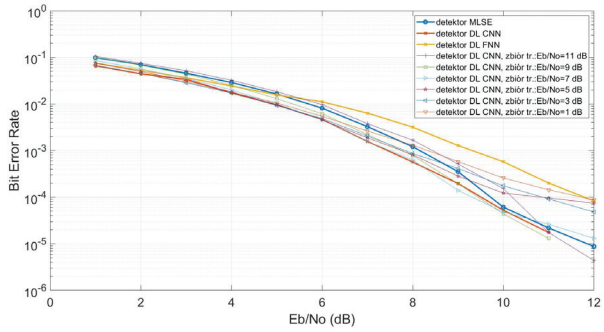
przy czym wartość estymacji $\widehat{y_j}$ należy traktować jako miarę prawdopodobieństwa przynależności danego wyjścia do wartości {0,1}. Z tego powodu, na końcu każdego modelu zastosowano warstwę z sigmoidalną funkcją aktywacji, skalującą wyjście do zakresu [0,1].

Poszukiwanie modelu o najwyższej skuteczności detekcji zrealizowano metodą siatkową (ang. *grid search*). Podstawowe parametry sieci (liczba warstw, funkcje aktywacji, etc.) testowano dla E_b/N_0=12 dB, a detekcję dla niższych wartości sprawdzono już dla najlepszej wyodrębnionej architektury. Podsumowanie przyjętych parametrów sieci FNN i CNN usystematyzowano w Tab. 1.

Tab. 1. Parametry końcowe modeli DL

Sieć FNN	***Sieć CNN***
warstwa skalująca [1260,1]	warstwa CNN [630, 2], Sigmoid, kernel {5,7,9}
warstwa ReLU [630,1]	warstwa CNN [630, 2], Sigmoid, kernel {5,7,9}
warstwa ReLU [315,1]	[1260,1] i przejście do FNN
warstwa Tanh [157,1]	warstwa Tanh [630,1]
warstwa skalująca Sigmoid [114,1]	warstwa Tanh [315,1]
	warstwa skalująca Sigmoid [114,1]
wsp. uczenia: 0,0001	wsp. uczenia: 0,00006
l. iteracji: 800, rozmiar zestawu danych (ang. *batch*): 100	

Charakterystyki BER badanych modeli DL i metody MLSE uzyskane dla wspólnego zbioru testowego (2004 pakiety) zaprezentowano na Rys. 3.



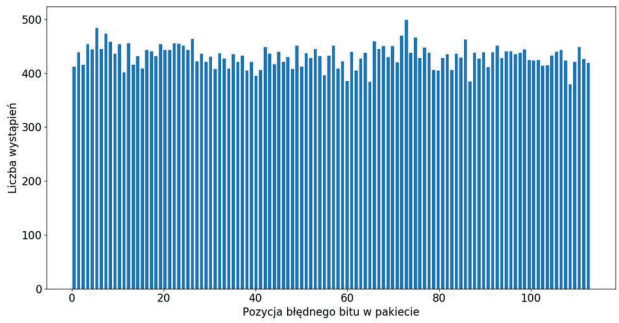
Rys. 3. Charakterystyki BER detekcji sygnałów GMSK

Analiza wyników pozwala stwierdzić, że sieć DL CNN prezentuje skuteczność porównywalną do optymalnego detektora MLSE. Poprawa jakości detekcji dla modeli DL względem optymalnego detektora MLSE może być spowodowana ograniczoną licznością zbioru testowego (2004 pakiety zawierające 114 bitów).

Obniżoną efektywność modelu FNN można wyjaśnić jednowymiarowym sposobem przekształcania próbek sygnału, który determinuje wspólne rozmieszczenie składowych zespolonych jako sąsiadujących wartości wejściowych typu rzeczywistego.

Należy zwrócić uwagę, że charakterystyka BER detektora DL CNN (Rys. 3) jest zbiorem wyników uzyskanych dla najlepiej wytrenowanych modeli, przy odrębnych wartości E_b/N_0. W celu oceny generalizacji sieci, na wykresie zamieszczono również stopy błędów detekcji dla modeli uczonych danymi przy ustalonym poziomie E_b/N_0. Można zauważyć, że modele wytrenowane dla zbiorów uczących E_b/N_0<7 dB prezentują nieznaczną degradację BER powyżej 9 dB co sugeruje, że zbyt mocno zniekształcony sygnał obniża skuteczność przebiegu procesu uczenia modelu.

Analiza pozycji błędnie odebranych bitów w pakiecie TCH/F pozwoliła wykluczyć możliwość występowania niepożądanych błędów powtarzalnych, wynikających np. z błędnej synchronizacji pakietów. Przykładowy rozkład błędów dla bloku kodowego 2004 pakietów, przedstawiono na Rys. 4.



Rys. 4. Liczba wystąpień błędów na kolejnych pozycjach pakietu TCH/F

5. PODSUMOWANIE

Otrzymane charakterystyki bitowej stopy błędów potwierdziły, że głębokie uczenie może stanowić obiecującą alternatywę dla klasycznych metod odbioru sygnałów radiowych z modulacją GMSK. Należy podkreślić, że w przypadku kanału AWGN, modele DL nie tylko zapewniły efektywność odbioru porównywalną z optymalnym detektorem MLSE, ale również zachowały wysoką generalizację predykcji – sieć CNN wytrenowana zbiorem treningowym dla zadanego E_b/N_0 umożliwiła także detekcję sygnałów przy innych poziomach E_b/N_0, bez istotnych zmian w jakości detekcji.

Mając na uwadze wielopłaszczyznową weryfikację prezentowanej metody, niezbędna jest dalsza analiza efektywności modeli DL, choćby wobec sygnałów w kanale z zanikami. Ponadto, bieżące zbiory danych uczących powinny zostać rozszerzone o rzeczywiste dane pomiarowe. Niezwykle ważnym aspektem, który został pominięty w niniejszym opracowaniu jest również złożoność obliczeniowa proponowanego rozwiązania. Wspomniane kwestie nie umknęły uwadze autorów, pozostają jednak w obszarze badań planowanych w przyszłości.

LITERATURA

[1] 3GPP, TS 143 064. 2020. "General Packet Radio Service (GPRS); Overall Description of the GPRS Radio Interface."

[2] Anil, Pooja P, and Renu Jose. 2019. "Signal Detection Scheme Based on Deep Learning in OFDM Systems." In *2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies*., 1226–30. Kannur.

[3] Chandhar, Prabhu. 2021. "Experimental Analysis of Deep Learning-Based Modulation Classification in Indoor Environments." In *18th India Council International Conference*. IEEE.

[4] Fengyuan, Shi, Hu Zeming, Yue Chunsheng, and Shen Zhichong. 2022. "Combining Neural Networks for Modulation Recognition." *Digital Signal Processing*. 120: 103264.

[5] Ganesh, Sriram, V. Sayee Sunder, and Arpita Thakre. 2018. "Performance Improvement in Rayleigh Faded Channel Using Deep Learning." In *International Conference on Advances in Computing, Communications and Informatics, ICACCI*., 1307–12. Bangalore: IEEE.

[6] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. London: MIT Press.

[7] Jiao, Licheng, and Jin Zhao. 2019. "A Survey on the New Generation of Deep Learning in Image Processing." *IEEE Access*. 7: 172231–63.

[8] Labinghisa, Boney, and Dong Myung Lee. 2021. "A Deep Learning Based Scene Recognition Algorithm for Indoor Localization." In *International Conference on Artificial Intelligence in Information and Communication*., 167–70. Jeju Island.

[9] Liberg, Olof, Marten Sundberg, Y.-P. Eric Wang, Johan Bergman, Joachim Sachs, and Gustav Wikstrom. 2020. *Cellular Internet of Things*. London: Academic Press.

[10] Lin, Xuming, Ruifang Liu, Wenmei Hu, Yameng Li, Xin Zhou, and XiaoXin He. 2017. "A Deep Convolutional Network Demodulator for Mixed Signals with Different Modulation Types." In *IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing*., 24–27. Orlando.

[11] Liu, Christopher Y, and Ric A Romero. 2021. "Deep Neural Network Detection for Pulsed Radar-Embedded M-PSK Communications." In *17th European Radar Conference*., 238–41.

[12] Ly, Amanda, and Yu-dong Yao. 2021. "A Review of Deep Learning in 5G Research : Channel Coding , Massive MIMO , Multiple Access , Resource Allocation , and Network Security." *IEEE Open Journal of the Communications Society*. 2: 396–408.

[13] Mohammad, Ahmad Saeed, Narsi Reddy, Fathima James, and Cory Beard. 2018. "Demodulation of Faded Wireless Signals Using Deep Convolutional Neural Networks." In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018*., 969–75.

[14] Schmitz, Johannes, Caspar Von Lengerke, Nikita Airee, Arash Behboodi, and Rudolf Mathar. 2019. "A Deep Learning Wireless Transceiver with Fully Learned Modulation and Synchronization." In *IEEE International Conference on Communications Workshops*. Shanghai: IEEE.

[15] Stefanski, Jacek. 2000. "Efektywność Odbioru Adaptacyjnego w Systemie Łączności Komórkowej GSM." Politechnika Gdanska.

[16] Xu, Tongyang, and Izzat Darwazeh. 2020. "Deep Learning for Over-the-Air Non-Orthogonal Signal Classification." In *IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. Antwerp.

[17] Ye, Hao, Geoffrey Ye Li, and Biing Hwang Fred Juang. 2017. "Power of Deep Learning for Channel Estimation and Signal Detection in OFDM Systems." *IEEE Wireless Communications Letters*. 7 (1): 114–17.

[18] Ye, Hao, Geoffrey Ye Li, Biing Hwang Fred Juang, and Kathiravetpillai Sivanesan. 2018. "Channel Agnostic End-to-End Learning Based Communication Systems with Conditional GAN." In *IEEE Globecom Workshops (GC Wkshps)*. Abu Dhabi: IEEE.

[19] Zhang, Chaoyun, Paul Patras, and Hamed Haddadi. 2019. "Deep Learning in Mobile and Wireless Networking: A Survey." *IEEE Communications Surveys and Tutorials*. 21 (3): 2224–87.

[20] Zhao, Zhongyuan, Mehmet Can Vuran, Fujuan Guo, and Stephen D. Scott. 2021. "Deep-Waveform: A Learned OFDM Receiver Based on Deep Complex-Valued Convolutional Networks." *IEEE Journal on Selected Areas in Communications*. 39 (8): 2407–20.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

PROJEKTOWANIE STACJI BAZOWEJ LTE-A NA PASMO 450 MHZ PRZY UŻYCIU UNIWERSALNEJ PLATFORMY SDR

DEVELOPMENT OF LTE-A BASE STATION FOR 450 MHZ BAND BASED ON VERSATILE SDR PLATFORM

Dawid W. Rosołowski[1]; Przemysław Korpas[2]; Michał Kajczuk[3]; Daniel Gryglewski[4]; Wojciech Wojtasiak[5]

[1,2,4,5] Politechnika Warszawska, Instytut Radioelektroniki i Technik Multimedialnych, Warszawa,
[1] dawid.rosolowski@pw.edu.pl, [2] przemyslaw.korpas@pw.edu.pl, [4] daniel.gryglewski@pw.edu.pl,
[5] wojciech.wojtasiak@pw.edu.pl
[3] IT Partners Telco Sp. z o.o., Warszawa, Michal.Kajczuk@itpartners.com.pl

DOI: 10.15199/59.2022.4.78

Streszczenie: **W komunikacie przedstawiono wybrane aspekty projektowania stacji bazowej LTE – 450 MHz z wykorzystaniem opracowanej uniwersalnej platformy radia programowalnego. Urządzenie integruje cztery niezależne tory nadawczo-odbiorcze pracujące z dupleksem częstotliwościowym w zakresach 452.5 MHz ÷ 457.5 MHz i 462.5 MHz ÷ 467.5 MHz odpowiednio dla odbioru i nadawania ze średnią mocą 30 dBm.**

Abstract: **This paper presents selected aspects of designing a 450 MHz LTE base station using a previously developed universal SDR platform. The device integrates four independent transmit and receive paths operating with frequency duplexing in the bands 452.5 MHz ÷ 457.5 MHz and 462.5 MHz ÷ 467.5 MHz for receiving and transmitting, respectively, with an average output power of 30 dBm.**

Słowa kluczowe: **LTE – 450 MHz, platformy rozwojowe SDR, prototypowanie systemów radiowych**

Keywords: **LTE – 450 MHz, prototyping of radio systems, SDR development platform**

1. WSTĘP

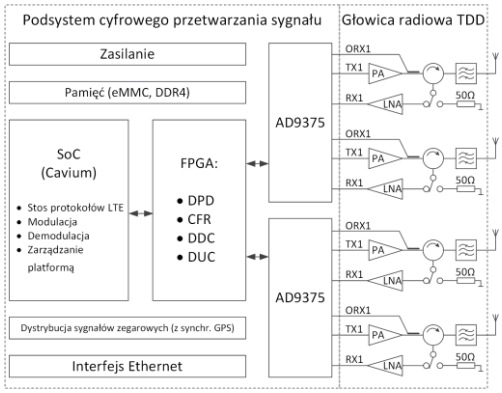
Rozwój systemów bezprzewodowych obserwuje się również w radiowych sieciach krytycznych. Zapotrzebowanie na świadczone usługi jest zbliżone do tego jakie znamy z telefonii komórkowej – oprócz transmisji głosu konieczna jest transmisja stowarzyszonych danych i treści wideo. Dodatkowo wymagany jest odpowiednio większy zasięg w wolnej przestrzeni i pokrycie w budynkach, a nawet pod ziemią tak, aby możliwe było zastąpienie starszej generacji systemów. Zauważalne w Europie zmiany w infrastrukturze radiowej używanej przez profesjonalne służby wiążą się z wdrażaniem technologii LTE (*Long Term Evolution*) poniżej 500 MHz (np. 450 MHz) [1]. Odpowiednio niskie pasmo zapewnia większe zasięgi, a relatywnie wąski kanał transmisyjny do 5 MHz w dupleksie częstotliwościowym z powodzeniem wystarczy do zapewnienia tego typu łączności.

Niezbędna infrastruktura sieciowa nie zawsze jest dostępna na rynku, w związku z tym pojawia się potrzeba opracowania własnych rozwiązań. Istotną przeszkodą jest jednak wysoki poziom zaawansowania technologicznego standardu LTE-A. Dlatego szczególnie cenne są możliwości wykorzystania zdobytych doświadczeń i opracowanych rozwiązań w trakcie już zrealizowanych projektów. Z tego powodu podczas opracowywania stacji bazowej (*eNodeB*) na pasmo 450 MHz zaproponowano wykorzystanie uniwersalnej platformy rozwojowej radia programowalnego (*Software Defined Radio*) opracowanej wcześniej, dla realizacji stacji bazowej LTE-A na pasmo 3.6÷3.8 GHz [2]. Intencją autorów było wyeksponowanie uniwersalności platformy SDR, pokazanie zasadniczych podobieństw i różnic w jej implementacji w stacjach bazowych z dupleksem częstotliwościowym (*FDD*) i czasowym (*TDD*) odpowiednio w pasmach 450 MHz i 3.6 GHz. Wskazano także możliwe kierunki rozwoju platformy w przypadku następnych projektów stacji bazowych.

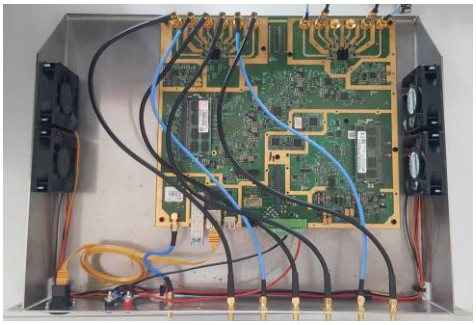
W rozdziale 2 przedstawiono zarys koncepcji modułowej platformy rozwojowej SDR i modyfikacje do pracy w paśmie 450 MHz. Opis dedykowanej głowicy radiowej oraz postulowanych parametrów zawarto w punkcie 3. Wyniki wstępnych pomiarów głównych podzespołów głowicy radiowej zaprezentowano w rozdziale 4. Podsumowanie wyników wykonanych prac oraz plany rozwojowe przedstawiono w ostatnim rozdziale.

2. ARCHITEKTURA BTS LTE – 450 MHz

W projekcie stacji bazowej (LTE – 450 MHz) wykorzystano platformę rozwojową (UPR-SDR), którą opracowano uprzednio w celu prototypowania systemów łączności pracujących w pasmie 3.6 ÷ 3.8 GHz z rozdziałem czasowym. Schemat blokowy bazowego rozwiązania pokazano na rys. 1, a fotografię płyty cyfrowego przetwarzania sygnału na rys. 2.



Rys. 1. Architektura uniwersalnej platformy SDR dla systemów z dupleksem czasowym (FDD).



Rys. 2. Fotografia płyty cyfrowego przetwarzania sygnału uniwersalnej platformy SDR.

O sprzętowej uniwersalności platformy decyduje zastosowany dwukanałowy zintegrowany transceiver AD9375 [3] z bezpośrednią przemianą częstotliwości z zakresu i do zakresu częstotliwości 300 MHz ÷ 6 GHz o jednoczesnym paśmie pracy o szerokości do 100 MHz. Transceiver wyposażono w mechanizm monitorowania wyjścia nadajnika przy pomocy dodatkowego odbiornika obserwacyjnego (ORX) i funkcję linearyzacji toru nadawczego przez predystorcję cyfrową (*Digital Predistortion*).

Od strony programowej uniwersalność platformy w zakresie dostosowania jej do różnych wariantów standardu LTE zapewniona jest zarówno na poziomie transceivera (wstępna filtracja i dostosowanie szybkości próbkowania) i procesorze pasma podstawowego, jak i w dodatkowym procesorze wspomagającym zrealizowanym w logice programowalnej. Procesor pasma podstawowego to układ SoC (MARVELL OCTEON Fusion-M CNF73xx) [4], który odpowiada za przygotowanie stosu protokołów stacji bazowej LTE (eNodeB), a w zakresie interfejsu radiowego (Uu) za generowanie i demodulację sygnałów standardu LTE (poszczególnych komponentów). Układ FPGA umożliwia realizację dodatkowych algorytmów takich jak: cyfrowa przemiana częstotliwości w górę lub w dół (DUC, DDC), redukcja szczytów (CFR), czy cyfrowa predystorcja DPD. Dzięki swej uniwersalności, platforma SDR może być zaadaptowana do nowej wersji standardu w znacznym stopniu na drodze modyfikacji programowych.

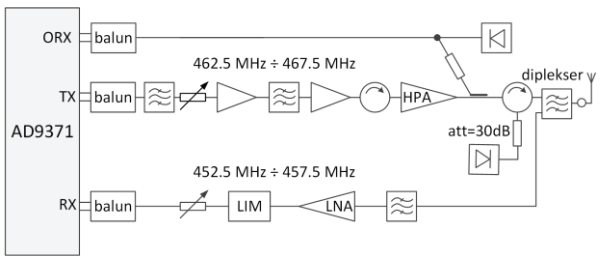
Jednak dostosowanie w pełni platformy SDR do konkretnego standardu wymaga dołączenia głowicy radiowej (*RF Front-end*), pracującej w zadanym paśmie. Ze względu na to, że przemiana częstotliwości realizowana jest w zintegrowanym transceiverze, funkcja głowicy sprowadza się do dostarczenia, możliwie bez zniekształceń, sygnałów o amplitudach optymalnych dla transceivera oraz zabezpieczenia wrażliwych elementów urządzenia platformy, a także realizacji zadanego trybu dupleksu.

Dostosowanie bazowej platformy SDR do pracy w paśmie 450 MHz wymagało następujących modyfikacji:

- optymalizacji obwodów dopasowujących w nowym paśmie w torach nadawczych TX, odbiorczych RX oraz monitorujących ORX transceivera,
- zaprojektowanie toru odbiorczego z założoną wartością współczynnika szumów i poziomem wzmocnienia,
- projekt toru nadawczego z wysokosprawnym wzmacniaczem i właściwym zakresem regulacji wzmocnienia,
- opracowanie mechanizmów zabezpieczających dla toru odbiorczego w nowym pasmie przed zbyt dużym sygnałem wejściowym,
- zapewnienia mechanizmów ochrony toru nadawczego przed: niewłaściwym, długotrwałym wysterowaniem wzmacniacza mocy zbyt dużym sygnałem oraz przed pracą z uszkodzonym systemem antenowym.

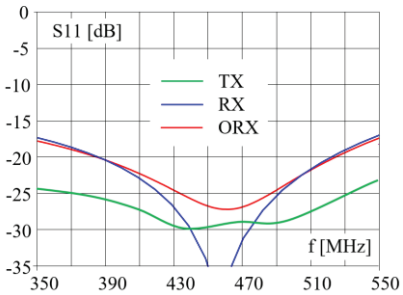
3. GŁOWICA RADIOWA

W przypadku projektowanego systemu z uniwersalną platformą SDR przez głowicę rozumie się moduł znajdujący się pomiędzy transceiverem a systemem antenowym, czyli RF front-end. Zgodnie z wytycznymi stacja bazowa LTE-A 450 MHz zawiera cztery niezależne tory nadawczo-odbiorcze pracujące z rozdziałem częstotliwościowym (FDD) w zakresach częstotliwości 452.5 MHz ÷ 457.5 MHz i 462.5 MHz ÷ 467.5 MHz odpowiednio dla odbiornika (*link w dół*) i nadajnika (*link w górę*). Schemat blokowy jednego z czterech torów głowicy pokazano na rys. 3.



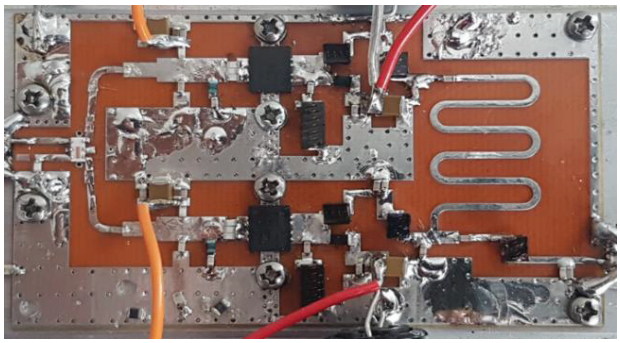
Rys. 3. Schemat blokowy pojedynczego toru radiowego dla dupleksu częstotliwościowego.

Adaptacja do nowego pasma wymagała optymalizacji obwodów dopasowujących transceivera. Co prawda uniwersalna platforma bazowa projektowana była z myślą o szerokopasmowym pokryciu częstotliwości w pełnym zakresie pracy układu AD9375 (do 6 GHz), to uznano, że straty odbicia -7 dB w paśmie 450 MHz są za duże. Znaczącą poprawę dopasowania uzyskano w wyniku przeprojektowania obwodów dopasowujących na wejściu transceivera – dobrano wartości elementy skupionych obwodów dopasowujących bez zmiany struktury. Wyniki symulacji, przedstawione na rys. 4, pokazują, że dla torów odbiorczych RX i ORX otrzymano straty odbicia na poziomie RL < - 27 dB, a dla torów TX, RL < -28 dB.

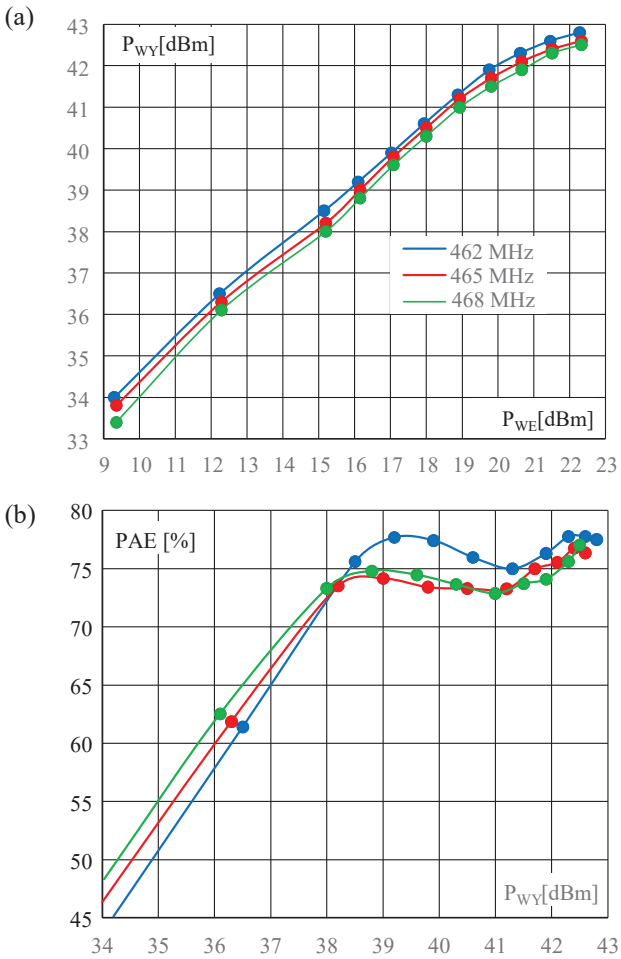


Rys. 4. Straty odbicia w funkcji częstotliwości na wejściu transceivera po modyfikacji obwodów dopasowujących.

Dla osiągnięcia wysokiej sprawności w szerokim zakresie zmian amplitud sygnału wejściowego, końcowy stopień mocy zrealizowany w konfiguracji Doherty'ego [5] z tranzystorami LD MOSFET BLP10H610 (Ampleon). Widok wzmacniacza Doherty'ego pokazano na rys. 5. Zmierzone charakterystyki mocy wyjściowej w funkcji mocy wejściowej $P_{WY}(P_{WE})$ i sprawności dodanej w zależności od mocy wyjściowej $PAE(P_{WY})$ dla trzech częstotliwości: 462 MHz, 465 MHz i 468 MHz sygnału sinusoidalnego impulsowanego o wypełnieniu 20 % przedstawiono odpowiednio na rys. 6a i 6b.



Rys. 5. Zdjęcie wzmacniacza Doherty'ego na pasmo 462 MHz ÷ 468 MHz

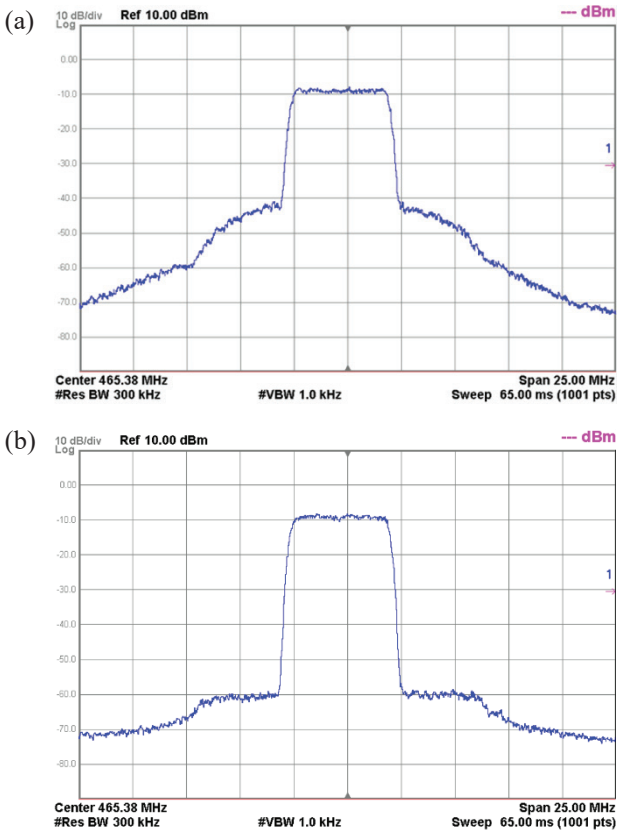


Rys. 6. Charakterystyki $P_{WY}=f(P_{WE})$ (a) i $PAE=f(P_{WY})$ (b) wzmacniacza Doherty'ego dla impulsowego pobudzenia tonem sinusoidalnym o wypełnieniu 20%.

Zastosowanie wzmacniacza mocy w konfiguracji Doherty'ego pozwoliło na uzyskanie w zadanym paśmie sprawności dodanej PAE > 60 % w szerokim zakresu mocy wyjściowych 36 ÷ 42 dBm w warunkach pobudzenia sygnałem sinusoidalnym. Oprócz stopnia Doherty'ego, tor nadawczy zawiera przedwzmacniacz SPF5043 (Qorvo) i driver ADL5611 (Analog Devices). Maksymalne wzmocnienie całego toru wyniosło 49 dB z możliwością regulacji w zakresie 0 ÷ 30 dB.

Przykład zmierzonego widma sygnału na wyjściu toru nadawczego z wyłączonym i włączonym algorytmem kompensacji zniekształceń nieliniowych pokazano na rys. 7. Pomiary przeprowadzono w warunkach pobudzenia sygnałem LTE o szerokości pasma 5 MHz dla średniej mocy wyjściowej $P_{ŚR}$ = 32.4 dBm i współczynniku PAPR = 9.5 dB. Uzyskano moc szczytową P_{MAX} = 41.9 dBm i całkowitą sprawność toru PAE = 35 % podczas, gdy sprawność wzmacniacza Doherty'ego w identycznych warunkach przekroczyła PAE = 41 %.

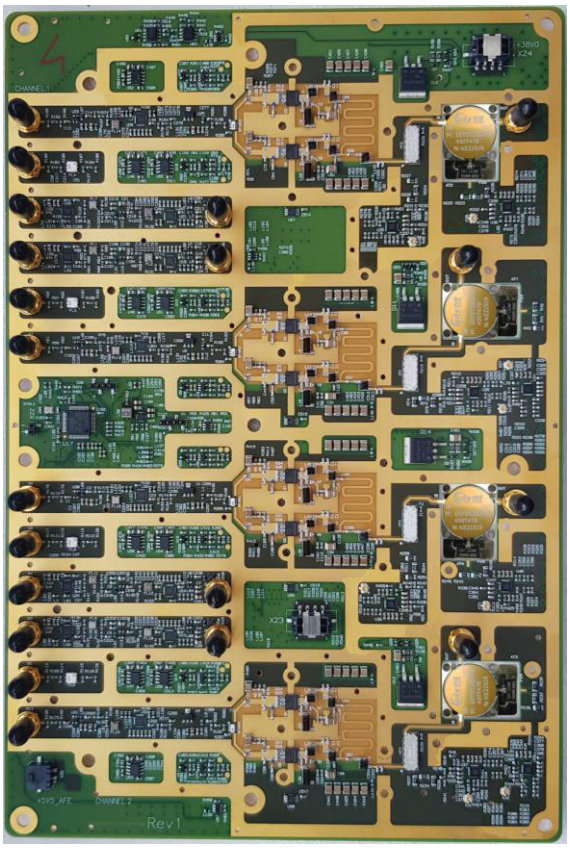
Tor odbiorczy, od strony anteny, zawiera ogranicznik, niskoszumny wzmacniacz (*LNA*) z funkcją regulacji wzmocnienia w zakresie 31 dB od wartości maksymalnej oraz filtry pasmowe. LNA składa się z dwóch układów MMIC SPF5043 połączonych kaskadowo umożliwiając uzyskanie wzmocnienia toru odbiorczego o wartości 37 dB i współczynnik szumów poniżej 1 dB.



Rys. 7. Widmo sygnału na wyjściu toru nadawczego bez (a) i z (b) korekcją DPD dla sygnału LTE 5MHz, $P_{ŚR}$=32.4 dBm, PAPR = 9.5 dB.

Fotografię płyty głowicy radiowej pokazano na rys. 8. Widoczne na niej szerokie ścieżki masy w połączeniu z wielogrodziowym systemem ekranowania stanowią istotne zabezpieczenie przed przesłuchami ze względu na

wymaganą równoczesność pracy torów nadawczych i odbiorczych.



Rys. 8. Fotografia płyty głowicy radiowej LTE – 450 MHz (wymiary płyty: 22 cm x 33 cm).

4. PARAMETRY

Prototyp stacji bazowej LTE – 450 MHz jest obecnie poddawany wstępnym pomiarom torów radiowych dla weryfikacji przyjętych rozwiązań układowych. Osiągnięte parametry urządzenia zebrano w tab. 1.

Tab.1 Parametry stacji bazowej LTE – 450 MHz.

Parametr	Wartość
Zakres częstotliwości [MHz]	*TX: 462.5 ÷ 467.5* *RX: 452.5 ÷ 457.5*
Tryb dupleksu	*częstotliwościowy*
Liczba torów w.cz.	*4*
Max. moc średnia nadajnika	*30 dBm @ PAPR 9.5, CCDF=0.01%*
Wzmocnienie małosygnałowe toru nadawczego (TX)	*19 ÷ 49 dB*
ACLR	*< -50 dBc z włączonym DPD w AD9375*
Sprawność wzmacniacza mocy	*42 % (LTE 5MHz)*
Minimalny współczynnik szumów	*1dB (RX)*
Wzmocnienie małosygnałowe toru odbiorczego (RX)	*7 ÷ 37 dB*
Max. użyteczny poziom sygnału wejściowego (RX)	*-7 dBm*

5. PODSUMOWANIE

Dysponowanie uniwersalną platformą SDR znacznie przyśpiesza prace badawczo-rozwojowe nad implementacją nowego standardu m.in. obniżając koszty i skracając czas opracowania. W realizowanym projekcie istniejącą platformę SDR, opracowaną pierwotnie dla badań głowicy radiowej systemu pracującego w standardzie LTE-A TDD na pasmo 3.6 ÷ 3.8 GHz, zaadaptowano do systemu LTE – 450 MHz FDD. Dokonane modyfikacje sprzętowe polegały przede wszystkim na wymianie głowicy radiowej i zmianie wartości elementów skupionych w wejściowych i wyjściowych obwodach dopasowujących transceivera (AD9375). Od strony programowej zmianie uległy głównie operacje realizowane przez koprocesor FPGA oraz konfiguracja wykorzystywanego sygnału LTE. W podobny sposób można przeprowadzić adaptację platformy SDR do standardów LTE w innych pasmach w zakresie częstotliwości od 300 MHz do 6 GHz.

PODZIĘKOWANIA

Praca została sfinansowana przez IT Partners Telco Sp. z o.o. oraz częściowo w ramach projektu MAZOWSZE/0126/19. Autorzy referatu pragną podziękować wszystkim osobom zaangażowanym w realizację projektu, zarówno w zakresie rozwoju sprzętu, jak i oprogramowania.

LITERATURA

[1] Kukliński Sławomir. 2019. Wymagania w zakresie łączności radiowej dla sektora energii w dobie transformacji cyfrowej" - *Opracowanie zespołowe*.

[2] Korpas Przemysław, Rosołowski Dawid, Michał Kajczuk i inni. 2022. “Prototyping a LTE-A Base Station with a Universal Hardware Platform”. Komunikat zgłoszony na konf. MIKON 2022.

[3] Analog Devices. 2017. “AD9371/AD9375 System Development User Guide”.

[4] ”MARVELL OCTEON Fusion-M CNF73xx: Next Generation Integrated Baseband Processors – Product Brief”. 2019.

[5] Staudinger J. 2016. "Doherty history, principles and key developments". IEEE MTT-S International Microwave Symposium (IMS) pp. 1-4.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

PROJEKTOWANIE MIKROFALOWEGO WZMACNIACZA DOHERTY'EGO W PRAKTYCE

Daniel Gryglewski [1]; Wojciech Wojtasiak [2]

[1] Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Warszawa, daniel.gryglewski@pw.edu.pl
[2] Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Warszawa, wojciech.wojtasiak@pw.edu.pl

DOI: 10.15199/59.2022.4.79

Streszczenie: **W referacie opisano nowe, bardziej efektywne podejście do projektowania mikrofalowych wzmacniaczy mocy w konfiguracji Doherty'ego. Istota pomysłu polega na przekształceniu standardowej struktury wzmacniacza do postaci wygodniejszej dla numerycznych symulacji. Jako przykład wykorzystania zaproponowanego podejścia przedstawiono wyniki pomiarów wzmacniacza Doherty'ego o mocy wyjściowej PEP ≥ 400 W opracowanego dla stacji bazowej łączności krytycznej LTE-A 450 MHz.**

Abstract: **The paper presents a more practical approach to designing Doherty amplifiers. The idea relies on transforming the Doherty amplifier structure to a more convenient form for numerical simulation. The 400 W (PEP) Doherty amplifier was developed as an example of using the proposed concept LTE-A 450 MHz base station.**

Słowa kluczowe: **LTE, PAE, wzmacniacz Doherty'ego.**

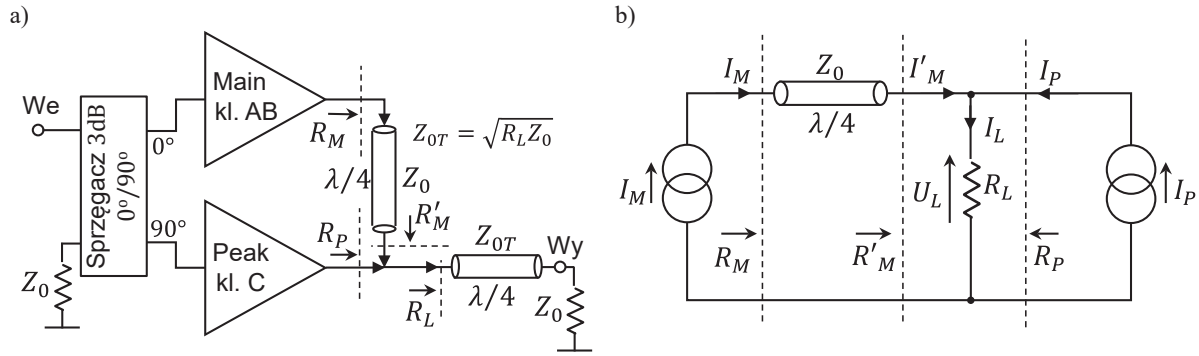
Keywords: **Doherty amplifier, LTE, PAE.**

1. WSTĘP

Jedną z charakterystycznych cech transmitowanych sygnałów w najnowszych systemach łączności bezprzewodowej takich jak np. LTE, 5G, jest duży odstęp (PAPR – *peak-to-avarege power ratio*) między maksymalną (PEP – *peak envelope power*) a średnią (P_{AV} – *average power*) mocą dostarczaną do anteny. Dla przykładu standard LTE-A wymaga odstępu PAPR na poziomie od 7 dB do 10 dB podczas, gdy 5G nawet ponad 13 dB. Szerokopasmowość nadawanych sygnałów (np. 100 MHz w pasmie 3.4 GHz ÷ 3.8 GHz, LTE-A) i duża wartość PAPR powodują znaczny spadek sprawności nadajnika. W rzeczywistości o sprawności decyduje wzmacniacz mocy, a dokładniej, końcowy stopień dużej mocy, który ma istotny udział w całym bilansie energetycznym stacji bazowej (BTS). Dla przykładu, BTS-LTE może potrzebować od 150 W do 350 W mocy DC, żeby nadawać sygnały o mocy średniej z zakresu 5 W ÷ 40 W [1]. Konsekwencją niskiej sprawności są poważne problemy z zarządzaniem ciepłem zwykle w niewielkiej obudowie BTS bez wymuszonego chłodzenia, zamontowanej na maszcie blisko anteny – wystawionej na działanie czynników klimatycznych. Sprawność obok liniowości jest, więc kluczowym parametrem wzmacniaczy mocy używanych we współczesnych systemach radiokomunikacyjnych. Idealnie byłoby uzyskanie wysokiej sprawności wzmacniacza niezależnie od poziomu sygnału wejściowego przy zachowaniu odpowiednio dużej liniowości. Niestety nie jest to możliwe. We wzmacniaczach klasy A, AB, B i C sprawność jest monotonicznie rosnącą funkcją mocy wejściowej osiągając maksimum dla mocy wyjściowej powyżej punktu 1dB kompresji wzmocnienia. Np. układ w klasie B osiąga maksymalną sprawność 78.5% dla PEP, 39% dla PEP/2 podczas, gdy dla sygnału 4G o PAPR = 10 dB niespełna 20% [2]. Dlatego sięgnięto do rozwiązania z roku 1936, które zaproponował W. H. Doherty [3]. Wzmacniacz Doherty'ego (DA) charakteryzuje się prawie stałą i dość wysoką sprawnością w szerokim zakresie wahań amplitudy sygnału wejściowego i liniowości typowej dla stopni klasy AB. Obecnie DA powszechnie stosuje się w nadajnikach jako końcowy stopień mocy. Zarys konstrukcji i zasady działania układu Doherty'go podano w następnym punkcie. Punkt 3 zawiera opis proponowanego przekształcenia struktury DA do postaci ułatwiającą optymalizację wzmacniacza w środowisku symulatora. Realizację DA o mocy $P_{AV} \geq 40$ W (PEP ≥ 400 W) dla BST-LTE-A 450 MHz przedstawiono w p.4. Podsumowanie stanowi przedmiot ostatniego 5 punktu.

2. WZMACNIACZ DOHERTY'GO

Ze względu na zalety konfiguracja Doherty'ego jest powszechnie stosowana konstrukcjach wzmacniaczy mocy dla nadajników współczesnych systemów łączności bezprzewodowej. Dlatego literatura przedmiotu jest bardzo obszerna – liczy z pewnością kilkaset pozycji. Dodatkowo każdy z wiodących producentów tranzystorów RF i mikrofalowych publikuje noty aplikacyjne swoich produktów z wzmacniaczami Doherty'ego. Warto wskazać tylko cztery publikacje [4-7], które są reprezentatywne dla staniu wiedzy i przyjazne czytelnikowi. Z tych powodów, jedynie w zarysie przedstawiono pomysł Doherty'ego dla uzyskania możliwie wysokiej i stałej wartości sprawności w szerokim zakresie zmian mocy wejściowej. Podstawową strukturę wzmacniacza Doherty'ego (DA) pokazano rys.1a, Na wejściu DA znajduje się 3 dB dzielnik kwadraturowy, który dostarcza sygnały o identycznej mocy, przesunięte w fazie o 90°, do wzmacniaczy: głównego (Main) i pomocniczego (Auxiliary) nazywanego także szczytowym (Peaking). Inwerter, ćwierć-falowy transformator, umieszczony na wyjściu stopnia Main stanowi element układu modulacji impedancji i kompensuje przesunięcia fazy wprowadzone przez dzielnik dla uzyskania synfazowego sumowania mocy. Tranzystory we wzmacniaczach pracują w zakresie źródła prądowego, przy czym Main w zwykle klasie AB, a Peak w klasie C.



Rys. 1. Schemat blokowy (a) i reprezentacja obwodowa (b) wzmacniacza Doherty'go

Zagadnienie modulacji impedancji łatwo wyjaśnić przy pomocy schematu zastępczego DA z rys.1b [5]. W ogólnym przypadku R'_M wyraża się następująco:

$$R'_M = \frac{U_L}{I'_M} = \frac{U_L}{I_L - I_P} = \frac{U_L}{I_L} \cdot \frac{I'_M + I_P}{I'_M} = R_L\left(1 + \frac{I_P}{I'_M}\right) \quad (1)$$

Po uwzględnieniu inwertera, rezystancja obciążenia stopnia Main wynosi:

$$R_M = \frac{Z_0^2}{R'^2_M} = \frac{Z_0^2}{R_L\left(1 + \frac{I_P}{I'_M}\right)} \quad (2)$$

Wzmacniacz Peak widzi rezystancję R_P równą:

$$R_P = R_L\left(1 + \frac{I'_M}{I_P}\right)$$

W typowych warunkach:

$$Z_0 = 50\Omega \rightarrow R_L = 25\Omega$$

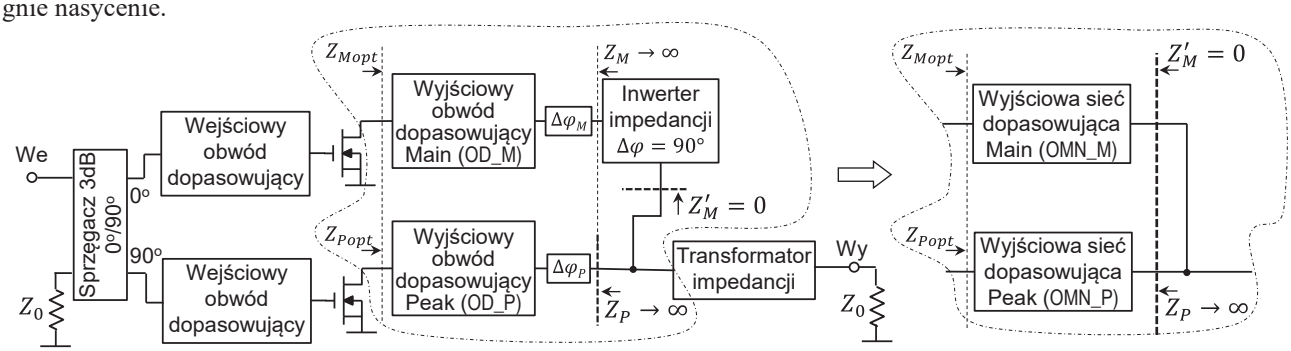
gdy $\quad I_P = 0 \rightarrow R_M = \frac{Z_0^2}{2R_L} = 100\Omega$ i $R_P \rightarrow \infty$

dla $\quad I_P = I'_M \rightarrow R_M = \frac{Z_0^2}{R_L} = 50\Omega$ i $R_P \rightarrow 50\Omega$

Przykład ten w uproszczeniu pokazuje efekt modulacji impedancji obciążenia stopnia Main. Podsumowując, gdy Peak jest wyłączony rezystancja obciążenia Main jest największą i maleje w trakcie włączania wzmacniacza Peak i osiągnie wartość 2-krotnie mniejszą niż maksymalna w stanie nasycenia wzmacniacza. Natomiast rezystancja obciążenia R_P stopnia Peak maleje od rozwarcia do wartości R_M, gdy oba stopnie pracują w nasyceniu. Taka zdefiniowana współpraca stopni Main i Peak prowadzi do poprawy liniowości dla dużych amplitud i sprawności w szerszym zakresie amplitud niż przypadku wzmacniacza z jednym tranzystorem. Charakterystyki $P_{wy}(P_{we})$ i $PAE(P_{we})$ DA składają się z dwóch zakresów odpowiadających pracy wyłącznie stopnia Main i wspólnego działania Main i Peak, który włącza się wtedy, gdy Main osiągnie nasycenie.

3. NOWA KONCEPCJA SYMULACJI DA

Mikrofalowe wzmacniacze mocy projektuje się obecnie wyłącznie z wykorzystaniem zaawansowanych wielofunkcyjnego oprogramowania – symulatorów takich jak Advanced Design System (ADS) firmy Keysight lub Cadence Microwave Office (daw. AWR MWO) Dla skutecznego projektowania oprócz symulatora wyposażonego w narzędzia analizy nieliniowej np. Harmonic Balance (HB), niezbędne jest dysponowanie wielkosygnałowym modelem tranzystora o dużej dokładności. Większość producentów tranzystorów mikrofalowych – Wolfspeed, Qorvo, Ampleon, NXP – oferuje nieodpłatnie modele swoich produktów. Mimo dużych możliwości symulatorów i dostępności modeli, nadal, być może najważniejsza jest wiedza i umiejętności projektanta. Zamiarem autorów było zaprezentowanie nowego podejścia do efektywnego projektowania wzmacniaczy Doherty'ego przy pomocy symulatora i modelu tranzystora. Zarys przekształcenia struktury DA pokazano na rys.2. Zaproponowana koncepcja polega na wchłonięciu do obwodu wyjściowego elementów, które w klasycznej strukturze DA znajdują się na lewo od węzła sumowania. Utworzona sieci dopasowująca na wyjściu OMN_M zawiera dotychczasowy obwód dopasowujący OD_M, przesuwnik fazy $\Delta\varphi_M$ i inwerter impedancji. Sieć OMN_P obejmuje obwód dopasowujące OD_P i przesuwnik fazy $\Delta\varphi_P$. W wyniku przekształcenia obwodu wyjściowego DA nie zmieniły się warunki – zwarcie i rozwarcie odpowiednio kierunku Main i Peak. Poszukiwanie struktury sieci dopasowujących OMN_M i OMN_P nie jest trudniejsze niż standardowym układzie Doherty'go, ponieważ należy spełnić



Rys. 2. Ilustracja przekształcenia obwodu wyjściowego DA do bardziej praktycznej symulacyjnie postaci.

identyczne warunki tj. zrealizować Z_{Mopt}, Z_{Popt} w płaszczyźnie wrót wyjściowych tranzystora i $Z_{Popt} \to \infty$ i $Z'_M = 0$, przy zachowaniu zgodnych faz sygnałów w węźle sumowania. Takie podejście umożliwia efektywniej optymalizować sieci wyjściowe, zmniejszyć ich wymiary i w ten sposób poszerzyć pasmo pracy.

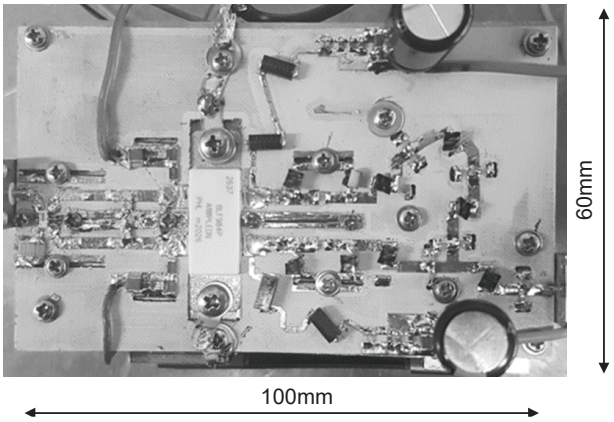
4. EKSPERYMENT

Dla weryfikacji zaproponowanego podejścia zaprojektowano wzmacniacz o mocy wyjściowej PEP ≥ 400 W i średniej $P_{AV} \geq 40$ W na pasmo 460 MHz ÷ 470 MHz dla stacji bazowej LTE o szerokości pasma 5 MHz. Dodatkowym wymaganiem była minimalizacja wymiarów stopnia mocy. Do realizacji wzmacniacza wybrano podwójny tranzystor LD MOSFET BLF984P firmy Ampleon. Podstawowe dane tranzystora BLF984P podano tabeli 1.

Tab. 1. Wybrane parametry LD MOSFET BLF984P

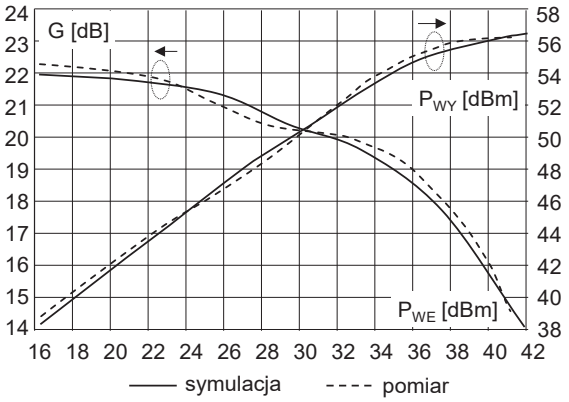
Parametr	*Wartość*	*Jednostka*
Częstotliwość	700	[MHz]
Moc wyjściowa	450 impuls CW	[W]
Zasilanie U_{DS}, I_{DQ}	50, 2x325	[V, mA]
Rezystancja termiczna	0.23	[K/W]
Wzmocnienie	23	[dB]
Sprawność drenowa	67	[%]
Dane dla impulsów CW, 1 ton dla 700 MHz		

Symulacje wzmacniacza wykonano przy pomocy ADS z wykorzystaniem nieliniowego modelu tranzystora, który udostępnił producent. Uproszczony schemat symulacyjny wzmacniacza Doherty'go w środowisku ADS pokazano na rys.3. Zdjęcie zrealizowanego wzmacniacza przedstawiono na rys.4.

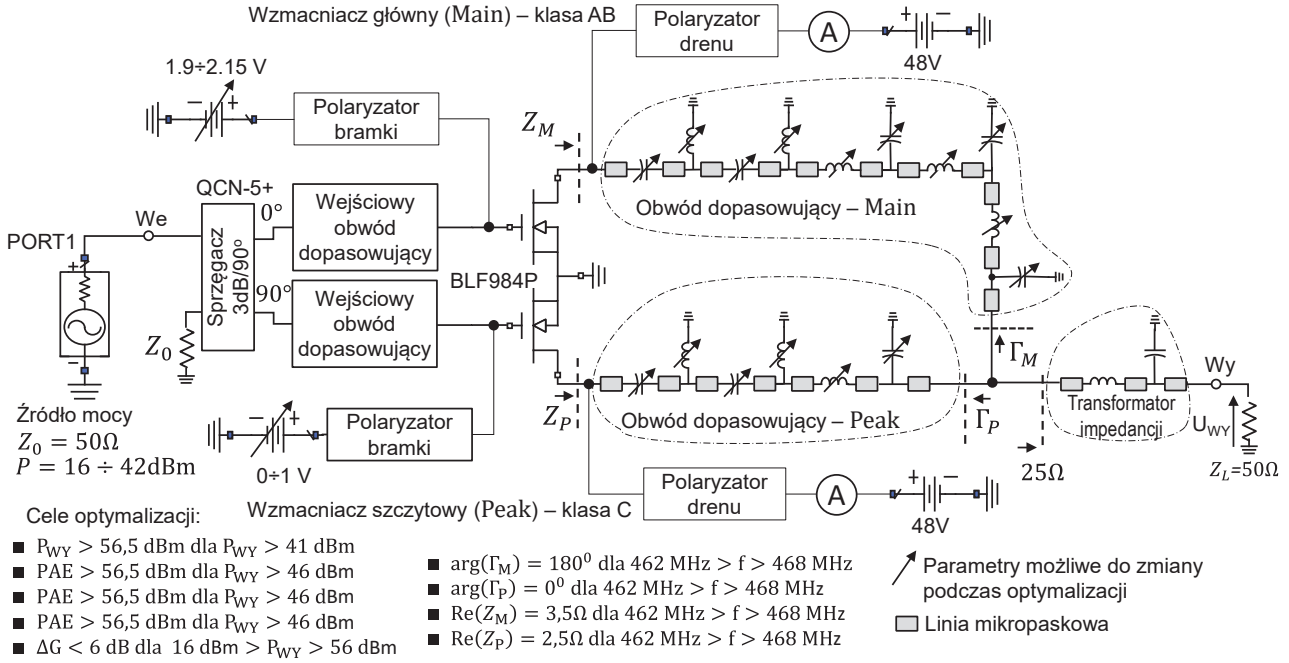


Rys. 4. Zdjęcie wzmacniacza Doherty'go z BLF984P

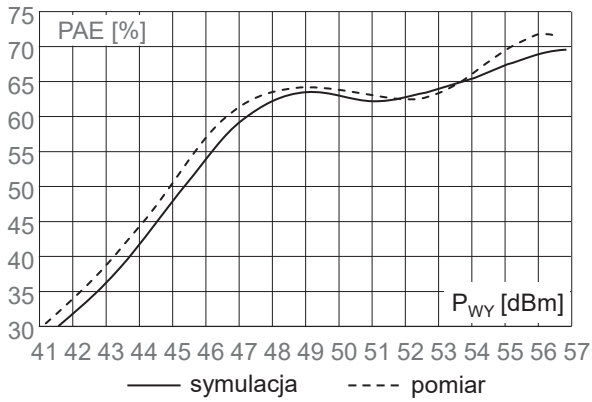
Obliczone i zmierzone charakterystyki DA tj. wzmocnienie $G = f(P_{WE})$, moc wyjściowa $P_{WY} = f(P_{WE})$ w funkcji mocy wejściowej pokazano na rys.5. Wykres sprawności dodanej w zależności od mocy wyjściowej $PAE = f(P_{WY})$ pokazano na rys.6. Pomiary przeprowadzono dla pobudzenia impulsowego tonem sinusoidalnym o wypełnieniu $\eta = 20\%$.



Rys. 5. $G = f(P_{WE})$ i $P_{WY} = f(P_{WE})$ DA dla f=465 MHz

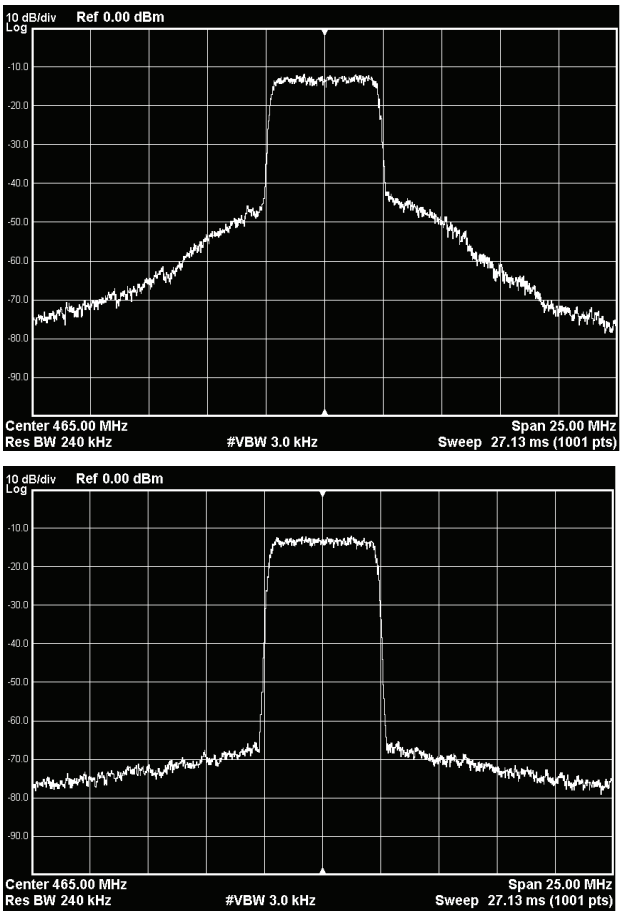


Rys. 3. Schemat symulacyjny w programie ADS



Rys. 6. Charakterystyka PAE= f(P_{WY}) dla f=465 MHz

Wzmacniacz zmierzono również dla sygnału w standardzie LTE i szerokości pasma 5MHz. Widmo sygnału wyjściowego bez i z korekcją DPD dla mocy szczytowej $PEP = 501$ W (57 dBm) przy $PAPR = 9.5$ dB przedstawiono na rys.7. W tych warunkach średnia moc wyjściowa była równa $P_{AV} = 56.2$ W (47.5 dBm), a średnia sprawność dodana przekroczyła $PAE \geq 50\%$.



Rys. 7. Widmo sygnału stopnia Doherty'ego dla sygnału LTE 5MHz bez i z korekcją DPD w warunkach: $PEP = 501\,W$, $PAPR = 9.5\,dB$

5. PODSUMOWANIE

Motywacją autorów było opracowanie efektywnego środowiska symulacyjnego w takich programach użytkowych jak ADS dla modelowania wzmacniaczy w konfiguracji Doherty'ego. W tym celu zaproponowano przekształcenie typowej struktury Doherty'ego poprzez włączenie elementów dotychczas peryferyjnych do wyjściowych sieci dopasowujących stopni Main i Peak. W rezultacie możliwe jest nie tyko optymalizowanie parametrów wzmacniacza, ale także jego rozmiarów. Uzyskane wyniki badań zaprojektowanego wzmacniacza potwierdziły skuteczność opisanego w referacie pomysłu. Należy podkreślić, że wymiary układu 100 mm x 60 mm są małe w porównaniu z długością fali ($\lambda = 304$ *mm*, $f = 465$ *MHz*) w linii paskowej na podłożu Taconic o stałej dielektrycznej $DK = 6.34$ grubości $h = 1.54$ *mm*.

PODZIĘKOWANIA

Referat powstał w ramach projektu NCBiR pt. „Stacja bazowa LTE 450MHz do zastosowań krytycznych", nr MAZOWSZE/0126/19 (lider: IT Partners Telco Sp. z o.o.).

LITERATURA

[1] Artesyn Embedded Technologies. 2017. "28V LDMOS Application Solution Brief". www.artesyn.com/assets/28v_ldmos_application_solution1490028979_ASB.pdf

[2] Wu Yu-Ting David. 2013. "A Doherty Power Amplifier with Extended Bandwidth and Reconfigurable Back-off Level". *PhD thesis*. University of Waterloo, Ontario, Canada.

[3] Doherty H. William. 1936. "A new high-efficiency power amplifier for modulated waves," *Proceedings of the IRE*, vol. 24, 1163–1182.

[4] Cripps, C. Steve. 2006. "RF Power Amplifiers for Wireless Communications", *Artech House*.

[5] Kazimierczuk K. Marian. 2015. „RF Power Amplifiers". Second Edition, *Wiley*.

[6] Colantonio P., Giannini. F., Giofrè R., Piazzon L. 2009. "The AB-C Doherty power amplifier. Part I: Theory". *Inter. Journal of RF and Microwave Computer-Aided Engineering*, 19 (3), 293-306

[7] Colantonio P., Giannini. F., Giofrè R., Piazzon L. 2009. "The AB-C Doherty power amplifier. Part II: Validation". *Inter. Journal of RF and Microwave Computer-Aided Engineering*, 19 (3), 307-316

WYBÓR DANYCH UCZĄCYCH DLA ALGORYTMÓW LINEARYZACJI WZMACNIACZY MOCY

SELECTION OF LEARNING DATA FOR POWER AMPLIFIER LINEARIZATION ALGORITHMS

Dawid Wilhelm Rosołowski1

[1] Politechnika Warszawska, Instytut Radioelektroniki i Technik Multimedialnych, Warszawa, dawid.rosolowski@pw.edu.pl

DOI: 10.15199/59.2022.4.80

Streszczenie: **W komunikacie omówiono metodę selekcji danych uczących wykorzystywanych w procesie linearyzacji wzmacniaczy radiowych. Badania weryfikacyjne przeprowadzono w rzeczywistym systemie pobudzanym sygnałem LTE20 MHz dla korygującego odkształcenia (predystorcji) modelowanego wielomianem pamięciowym w architekturze pośredniego uczenia. Zaproponowana metoda, polegająca na budowaniu wektorów uczących ze zbioru sekwencji próbek zbieranych wokół globalnych szczytów, pozwala na co najmniej dwukrotną redukcję rozmiaru wektora.**

Abstract: **In this paper the method for selecting of training dataset used in linearization process of RF amplifiers is presented. Verification studies have been carried out with LTE20 signal and linearized in indirect learning architecture with the predistorter based on memory polynomials. The proposed method, which involves building learning vectors with sequences of samples around the global peaks, allows at least a twofold reduction in vector size.**

Słowa kluczowe: **cyfrowa predystorcja, modelowanie wzmacniaczy, wielomiany pamięciowe**

Keywords: **DPD (Digital Predistiortion), memory polynomial, PA (Powe Amplifier) modeling**

1. WSTĘP

Nieodzownym elementem współczesnych nadajników, w których dąży się do zachowania kompromisu pomiędzy sprawnością i liniowością, jest mechanizm cyfrowej linearyzacji [1][2]. W typowych rozwiązaniach opartych na technologii radia programowalnego (SDR) z architekturą z bezpośrednią przemianą częstotliwości (ZIF), wbudowanym dodatkowym odbiornikiem do monitorowania sygnału na wyjściu wzmacniacza i przetwarzaniem cyfrowym w pasmie podstawowym, realizuje go się poprzez dodatkowy algorytm odkształcenia sygnału IQ, tzw. cyfrowej predystorcji (DPD), w taki sposób, aby skompensować nieliniową charakterystykę przejściową wzmacniacza [1][3]. Odkształcenie modelowane jest za pomocą wielomianów pamięciowych (na bazie wielomianów Volterry) lub sieci neuronowych. W obu przypadkach konieczne jest wyznaczenie współczynników definiujących predystorcję na podstawie zbioru danych zawierających pary ekwiwalentnych sygnałów w pasmie podstawowym: pobudzenia wzmacniacza oraz odpowiedzi wzmacniacza. O ile metody wyznaczania współczynników dla danego typu modelu są znane, np. minimalizacja średniokwadratowa lub propagacja wsteczna w przypadku modeli neuronowych, to efektywny sposób doboru w/w par nie jest jasno określony. Ponadto, na przykład dla sygnałów LTE, najistotniejsze próbki z punku widzenia linearyzacji i sprawności pojawiają się bardzo rzadko. W literaturze [3] spotyka się zdawkowe stwierdzenie, że do nauki algorytmów DPD, czyli wyznaczania współczynników predystorcji, stosuje się kilkadziesiąt tysięcy par próbek w kwadraturze.

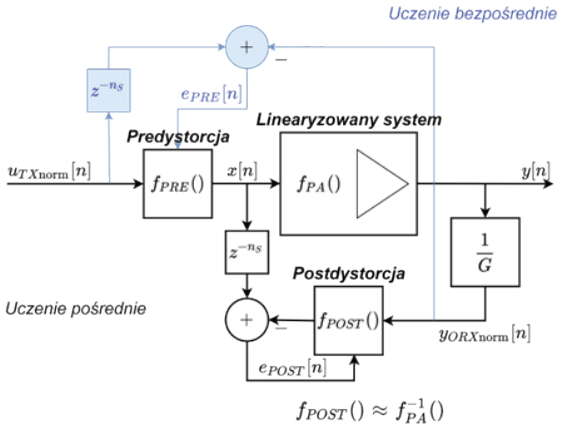
W niniejszym artykule podjęto próbę określenia wpływu rozmiaru danych uczących na dokładność modelu odwrotnego wzmacniacza z wielomianem pamięciowym oraz jakość linearyzacji przy pobudzeniu sygnałem LTE20. Zaproponowano metodę wyboru reprezentatywnych par próbek, która pozwala zredukować rozmiar danych uczących oraz poprawia modelownie w otoczeniu punktu kompresji - globalnych szczytów.

W rozdziale 2 przedstawiono algorytmy pozwalające na wyznaczanie modelu wzmacniacza i modelu do niego odwrotnego oraz omówiono użyty algorytm predystorcji przez postdystorcję. Następnie, w rozdziałach 3 i 4, opisano procedury zbierania par użytych do modelowania. Wyniki symulacji i pomiarów dla zredukowanego wektora uczącego zawarto w rozdziale 5. W zakończeniu podsumowano wyniki analiz oraz wskazano dalsze kierunki badań.

2. MODELOWANIE I PREDYSTORCJA

Algorytmy linearyzacji toru nadawczego bazują na nieliniowych modelach opisujących relacje pomiędzy sygnałami na jego wejściu (u, x) i wyjściu (y) ze szczególnym uwzględnieniem przetwarzania sygnałów przez bloki predystorcji, postdystorcji i linearyzowanego wzmacniacza. W zależności od metody zastosowanej do linearyzacji rozważany jest model w przód, w którym $y = f_{PA}(f_{PRE}(u))$ dla uczenia bezpośredniego lub model odwrotny połączony z modelem w przód $y = f_{PA}(f_{PA}^{-1}(u))$ dla uczenia pośredniego. Nazwy metod, pokazanych schematycznie na rys.1, wiążą się ściśle ze sposobem obliczania błędu oraz wyznaczania funkcji odkształcenia. W uczeniu bezpośrednim monitorowany jest wpływ predystorcji na sygnał na wyjściu wzmacniacza, podczas gdy w uczeniu pośrednim ma miejsce predystorcja przez postdystorcję, czyli korekta charakterystyki na wyjściu wzmacniacza na podstawie, znormalizowanej,

odwrotnej odpowiedzi wzmacniacza, która jest następnie przyjmowana jako funkcja odkształcenia.



Rys. 1. Schemat blokowy algorytmów linearyzacji cyfrowej z uczeniem pośrednim i bezpośrednim

Niezależnie od metody, do linearyzacji stosowane są te same, SDR'owe platformy sprzętowe z dodatkowym odbiornikiem monitorującym sygnał wyjściowy wzmacniacza (ORX, odbiornik obserwacyjny), a po stronie algorytmicznej najczęściej wielomiany pamięciowe pozwalające na modelowanie nieliniowe. Postać wielomianu pamięciowego modelującego postdystorcję dla metody z uczeniem pośrednim opisuje wzór:

$$\hat{x}[n] = \sum_{k=1}^{K} \sum_{m=0}^{M} a_{km} z|z|^{k-1} \qquad (1)$$

gdzie: $z = y_{ORXnorm}[n-m]$, K, M, a_{km} są parametrami modelu i oznaczają odpowiednio największą potęgę wielomianu, głębokość efektu pamięciowego oraz współczynniki modelu. Obie metody wykorzystują taki sam zbiór par danych uczących. Metoda z pośrednim uczeniem może być użyta niezależnie lub jako algorytm do wyznaczenia początkowych współczynników dla metody bezpośredniej. Wyznaczenie współczynników postdystorcji realizowane jest zwykle na drodze rozwiązywania silnie nadokreślonego układu równań zapisanego w postaci macierzowej w następujący sposób:

$$\boldsymbol{Z}\,\hat{\boldsymbol{a}} = \boldsymbol{x} \qquad (2)$$

Rozmiar macierzy $\boldsymbol{Z}$ wynosi $(M+1)KN$, gdzie N jest długość wektora uczącego. W typowych zastosowaniach wartość współczynników M i K nie przekracza odpowiednio wartości 3 i 5.

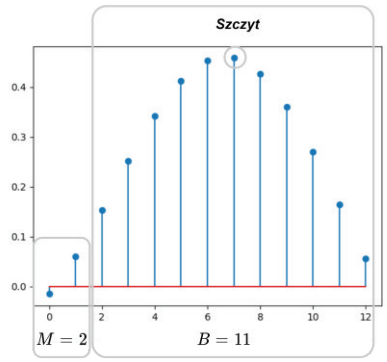
3. DANE DO MODELOWANIA

Współczesne platformy SDR pozwalają na efektywną weryfikację algorytmów cyfrowego przetwarzania sygnałów radiowych. Umożliwiają łatwe pozyskanie próbek do budowy wektora uczącego a następnie implementację różnych algorytmów w tym linearyzacji (DPD). Zgodnie z zasadą działania radiowych systemów z architekturą ZIF, przetwarzanie realizowane jest na dolnopasmowym ekwiwalencie sygnałów radiowych, w związku z tym na rys.1 nie ma jawnie wskazanych bloków przemiany częstotliwości.

Pozyskanie wektora danych uczących, który jest zbiorem zsynchronizowanych par próbek $\{\boldsymbol{u_{TXnorm}}[\boldsymbol{n_k} - \boldsymbol{n_S}], \boldsymbol{y_{ORXnorm}}[\boldsymbol{n}]\}$ wymaga przeprowadzenia procedury synchronizacji czasowej pomiędzy sygnałem z odbiornika obserwacyjnego a sygnałem nadawanym, w wyniku której wyznaczane jest opóźnienie $\boldsymbol{n_S}$. Konieczna jest również normalizacja obu sygnałów, tak by wyznaczana funkjca predystorcji nie zmieniała wzmocnienia w zakresie liniowego przetwarzania, a jedynie korygowała nieliniowości.

4. SELEKCJA DANYCH UCZĄCYCH

Globalne szczyty, które są w głównej mierze odpowiedzialne za generację produktów nieliniowych, występują w sygnale LTE bardzo rzadko. Powszechnie stosowane normy używane podczas pomiarów wzmacniaczy LTE definiują szczyty za pomoca współczynnika PAPR (stosunek mocy szczytowej do średniej) dla prawdopodobieństwa wystąpienia 0.01 %, przy typowych wartościach PAPR > 9 dB. W związku z tym losowo zbierane, krótkie wektory uczące zawierające pary następujących po sobie próbek, mogą okazać się niereprezentatywne i prowadzą do modelu nieoptymalnego dla kluczowego zakresu pracy wzmacniacza. Lepszym rozwiązaniem wydaje się selektywny wybór możliwie najmniejszej liczby par, które tworzą wektor uczący ze szczególnym uwzględnieniem próbek opisujących zachowanie systemu dla najwyższych amplitud sygnału pobudzającego. Mechanizm zbierania właściwych par może być łatwo wprowadzony do istniejącego systemu jako uzupełnienie bloku redukcji wartości szczytowych (CFR), w którym to w sposób ciąły monitoruje się wartość modułu obwiedni. Ze względu na rolę efektu pamięciowego w modelowaniu nieliniowości, istotne wydaje się zbieranie pewnych sekwencji następujących po sobie próbek. W związku z tym zaproponowano użycie wektorów uczących składających się z L-sekwencji zawierających po B próbek w kwadraturze. Każda z sekwencji składa się z pary reprezentujacej szczyt spełniający kryterium amplitudy oraz $B-1$ par okalających szczyt. W celu zapewnienia danych dla modelowania efektu pamięciowego o głębokości M, konieczne jest uzupełnienie sekwencji o dodatkowe M par po lewej stronie w stosunku do sekwencji pierwotnej. Pojedynczą sekwencję próbek o rozmarze $B = 11$ i $M = 2$ dla składowej synfazowej pokazano na rys. 2.



Rys. 2. Sekwencja próbek o rozmiarze B = 11 powiększona o próbki konieczne dla uwzględnienia efektu pamięciowego (M=2).

5. ANALIZA WPŁYWU ZAWAROŚCI DANYCH NA JAKOŚĆ MODELU

Opisaną powyżej metodę konstruowania bądź suplementacji wektora uczącego weryfikowano na poziomie symulacji i pomiarów z użyciem zmodyfikowanej platformy SDR [4]. W celu oceny zgodności modelu użyto miary unormowanego błędu średniokwadratowego zdefiniowanego w następujacy sposób:

$$\boldsymbol{NMSE = 10\log_{10}\left(\frac{\sum_k^N |z_{mod}[k]-z_{ref}[k]|^2}{\sum_k^N |z_{ref}[k]|^2}\right)} \qquad (3)$$

Skuteczności linearyzacji określono natomiast za pomocą pomiaru współczynników rozrostu widma w kanałach sąsiednich – ACLR, podając wartości dla lewo- i prawostronnego sąsiedztwa. Pary próbek IQ próbkowanych z częstotliwością 122.88 MS/s zebrano w sytemie złożonym z platformy SDR z dołączonym wzmacniaczem SKY66292-11 o mocy szczytowej 31 dBm, pracującym w pasmie 2.4 GHz i pobudzanym sygnałem testowym ETM1.1 LTE20 MHz zgodnym z zaleceniami [5]. Synchronizację próbek przeprowadzono dla 16-krotnie intepolowanego sygnału z odbiornika obserwacyjnego (ORX). Konieczne algorytmy przetwarzania sygnałów zaimplementowano za pomocą jęzka Python. Do linearyzacji użyto metody pośredniego uczenia z wielomianem pamięciowym o paramterach $\boldsymbol{M = 3}$ i $\boldsymbol{K = 5}$.

5.1. Wektor bez suplementacji

Eksperymenty poprzedzono wyznaczeniem błędów średniokwadratowych dla modelu odwrotnego (postdystorcja) i współczynników ACLR dla różnych długości N ciągłego wektora uczącego. Do wyliczenia błędu użyto zbioru danych dziesięciokrotnie dłuższego niż wektor uczący. Wyniki zebrano w tab.1.

Tab. 1. Wpływ długości ciągłego wektora uczącego na błąd modelu postdystorcji oraz jakość linearyzacji.

N	256	512	1024	**2048**	**10k**	20k
NMSE dla N	-36.4	-36.2	-36.3	-36.4	-36.6	-36.6
NMSE dla 10N	-34.9	-26.2	-26.2	-29.1	-36.1	-36.2
$ACLR_L$[dBc] $ACLR_P$	-46.6 -45.8	-48.2 -47.7	-50.5 -50.8	-51.8 -51.7	-53.2 -53.5	-53.6 -53.8
$ACLR_L$/$ACLR_P$ bez DPD [dBc]: -44.2 / -41.5						

Jak widać do wyznaczenia dobrego modelu predystorcji w linearyzowanym systemie wystarczy około 2 k par próbek, podczas gdy bardzo dobry model wymaga co najmniej 10 k par. Stały błąd modelu, niezależny od ilości danych użytych do jego wyznaczenia, sugeruje, że użyty wielomianowy model z pamięcią jest wystarczający dla zamodelowania użytego wzmacniacza. Wysoce prawdopodobne jest jednak to, że zgodność modelu zachodzi dla przeważającej liczby próbek o typowych wartościach i nie uwzględnia szczytów.

5.2. Wektor jako zbiór wybranych sekwencji

Badania nad selektywnym doborem par próbek rozpoczęto od konstruowania wektora uczącego z L-sekwencjami o różnej długości B, zbieranymi wokół szczytów globalnych. Błędy modeli wyznaczano dla całej transmitowanej ramki, czyli 1228800 par próbek. Próg detekcji szczytów globalnych przyjęto na poziomie 0.4 amplitudy maksymalnego szczytu wykrytego w obwiedni testowanego sygnału. Zebrane dane zestawiono w tab.2.

Tab. 2. Zależność błędu modelu DPD oraz ACLR od długości B i ilości L sekwencji zbieranych wokół szczytów.

L	**10**	**50**	**100**	**200**	**400**
B=5					
NMSE	-14.0	-25.9	-22.3	-34.4	-34.3
ACLR L/P	-36.9 -38.3	-47.8 -41.6	-44.6 -44.2	-51.2 -43.1	-48.7 -48.3
N	80	400	800	1600	3200
B=7					
NMSE	-15.8	-28.1	-31.5	-35.1	-35.1
ACLR L/P	-40.7 -45.0	-48.9 -49.9	-50.9 -48.9	-51.4, -50.6	-52.4 -50.8
N	100	500	1000	2000	4000
B=11					
NMSE	-20.9	-30.7	-35.7	-35.6	-35.8
ACLR L/P	-46.2 -27.28	-51.0 -50.9	-53.1 -52.5	-52.8 -52.9	-52.5 -53.0
N	140	**700**	**1400**	**2800**	**5600**
B=15					
NMSE	-12.2	-30.3	-35.6	-35.7	
ACLR L/P	-44.5 -46.5	-52.3 -51.6	-53.4, -52.4	-53.2 -53.2	
N	180	**900**	**1800**	**3600**	
B=21					
NMSE	-17.9	-32.5	-35.7	35.7	
ACLR L/P	-46.6 -45.22	-51.4 -52.3	-53.1 -52.7	-52.9 -53.3	
N	240	**1200**	**2400**	**4800**	
$ACLR_L$/$ACLR_P$ bez DPD [dBc]: -44.2 / -41.5					

Wyniki potwierdzają, że jakość linearyzacji nie zależy wprost od ilości zebranych par, ale od zawartości reprezentatywnych przebiegów sygnału. Stąd lepsze rezultaty dają sekwencje dłuższe, przy czym nadmierna długość prowadzi do zmniejszenia liczności zróżnicowanych sekwencji. W skrajnym przypadku zbyt krótkie lub źle dobrane sekwencje skutkują jeszcze większym zniekształceniem sygnału. Ponadto wysoka zgodność modelu nawet na przestrzeni całej ramki nie gwarantuje jakości linearyzacji, co może wynikać z metody wyznaczania błędu, która nie bierze pod uwagę statystyk występowania poszczególnych amplitud – błąd odzwierciedla globalne dopasowanie modelu do najczęściej pojawiających się wartości.

Eksperyment powtórzono uzupełniając dodatkowo wektory uczące sekwencjami zawierającymi szczyty lokalne wokół wartości średniej, w przedziale o szerokości 0.1 w stosunku do zakresu dynamiki przetwarzania. Przeanalizowano skuteczność uczenia za pomocą wektorów o różnej zawartości sekwencji ze szczytami globalnymi i szczytami lokalnymi. Na podstawie wyników z poprzedniej tabeli przyjęto arbitralnie długość sekwencji $\boldsymbol{B = 11}$. Dane zebrane w tab.3 pokazują, że dodatkowa suplementacja pogarsza jakość linearyzacji.

Tab. 3. Wpływ dodatkowej suplementacji sekwencjami zebranymi wokół wartości średniej sygnału.

Lg/Ll, B=11	**50/50**	**75/25**	**90/10**
NMSE	-28.0	-31.4	-36.1
	-51.9 / -51.7	-52.8 /-51.3	-53.1 / -52.9
N=1400			
Lg/Ll, B=11	**100/100**	**150/50**	**50/150**
NMSE	-36.2	-36.3	-32.2
$ACLR_L$/$ACLR_P$	-53.5 / -53.4	-53.4 / -53.2	-52.7 / -51.6
N=2800			
$ACLR_L$/$ACLR_P$ bez DPD [dBc]: -44.2 / -41.5			

Przedstawione powyżej dane wskazują, że kilkukrotny zysk wynikający z świadomego konstruowania wektora uczącego zauważalny jest dopiero w przypadku potrzeby

uzyskania możliwie najlepszej linearyzacji. Wówczas konieczne jest zebranie wektora złożonego z ponad 10 tys. par następujących po sobie próbek, w porównaniu do 2 lub 3 tysięcy odpowiednio spreparowanych sekwencji konstruowanych na podstawie globalnych szczytów wykrywanych w sygnale.

5.3. Dane ciągłe i suplementacja

W typowych implementacjach czasu rzeczywistego dąży się do redukcji zasobów koniecznych do implementacji algorytmów. Istotne jest więc, jaka jest optymalna ilość próbek wystarczająca do efektywnej linearyzacji. W tab.4 zebrano wyniki eksperymentu polegającego na suplementacji typowego zbioru ciągłych danych L-sekwencjami par o długości $B = 11$ zebranymi wokół szczytów globalnych (procedura jak w przypadku pozyskania wyników do tab.2).

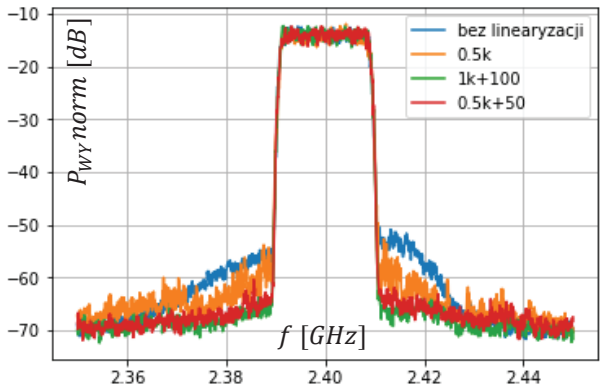
Tab. 4. Wpływ suplementacji na redukcję rozmiaru wektorów uczących dobieranych w celu linearyzacji przy ograniczonych zasobach.

L		**50**	**100**	**200**
N	**1k**	**1k+L**	**1k+L**	**1k+L**
	1024	1724	**2424**	3824
NMSE	-27.8	-31.9	-35.7	-35.8
ACLR L/P	-50.1 -50.7	-51.9 -52.8	-53.6 -52.8	-52.9 -52.6
N	**.5k**	**.5k+L**	**.5k+L**	**.5k+L**
	512	**1212**	1912	3312
NMSE	-16.8	-30.2	-35.8	-36.0
ACLR L/P	-47.6 -47.3	-52.1 -51.4	-53.1 -52.9	-52.7 -52.9
N	**.25k**	**.25k+L**	**.25k+L**	**.25k+L**
	256	956	1656	3056
NMSE	-16.1	-30.3	-35.5	-35.1
ACLR L/P	-47.6 -45.2	-50.8 -51.1	-53.2 -52.0	-52.7 -51.6
$ACLR_L/ACLR_P$ bez DPD [dBc]: -44.2 / -41.5				

Podobnie, jak w przypadku suplementacji sewencjami zbieranymi wokół szczytów lokalnych, sekwencja ciągłych danych pogarsza jakość uzyskanej linearyzacji w porównaniu do wektora preparowanego o porównywalnym rozmiarze. Jego generacja w systemie z dynamiczną liearyzacją jest jednak o wiele prostsza, bowiem można budować go z aktualnnie transmitowanych próbek uzupełnianych rzadziej wykrywanymi sekwencjami, gromadzonymi w postaci kolejki FIFO. Dla badanego systemu optymalnym rozmiarem wektora uczącego wydaje się około 2 k par próbek, z czego połowa to wyłapywane i gromadzone sekwencje wokół szczytów. Na rys. 3 pokazano widma sygnału na wyjściu wzmacniacza bez zastosowania algorytmu DPD oraz dla hybrydoywch wektorów uczących zaznaczonych w tab.4.

1. PODSUMOWANIE

W artykule przedstawiono wstępne wyniki badań metod służących do redukcji rozmiaru wektora uczącego używanego w algorytmach DPD.



Rys. 3. Widmo sygnału na wyjściu linearyzowanego wzmacniacza dla różnych danych treningowych.

Dla linearyzowanego systemu z sygnałem testowym LTE 20 MHz określono konieczną ilość par próbek w standardowym wektorze treningowym ETM 1.1 oraz wskazano zależność błędu modelu postdystorcji oraz jakości linearyzacji wyrażonej poprzez parametr ACLR. Zaproponowno metodę konstruowania relatywnie krótszych wektorów uczących na bazie sekwencji próbek zbieranych wokół wykrywanych szczytów sygnału. Przedstawione wyniki pokazują, że możliwa jest redukcja rozmiaru wektora co najmniej dwa razy przy zachowaniu takiego samego poziomu linearyzacji. Pokazano również, że w typowym rzeczywitym systemie możliwe jest używanie hybrydowego wektora uczącego złożonego z części ciągłego sygnału transmitowanego w systemie uzupełnionego sekwencjami pozwalającymi na poprawne zamodelowanie szczytów, tak istotnych w procesie linearyzacji wzmacniacza.

Uogólnienie przedstawionych wyników wymaga dalszych badań polegajcych na liearyzacji różnych wzmacniaczy mocy oraz poszerzeniu bazy sygnałów testowych, w szczególności o sygnały wielokomponentowe, a także zmodyfikowania błędu używanego dla oceny wiarygodności modelu w szczytach.

LITERATURA

[1] Ghannouchi, F. M., Hammi, O., & Helaoui, M. (2015). *Behavioral modeling and predistortion of wideband wireless transmitters*. John Wiley & Sons.

[2] Cripps, S. C. (2002). *Advanced techniques in RF power amplifier design.* Artech House.

[3] Wood, J. (2014). *Behavioral modeling and linearization of RF power amplifiers*. Artech House.

[4] Korpas Przemysław, Rosołowski Dawid, Michał Kajczuk i inni. 2022. "*Prototyping a LTE-A Base Station with a Universal Hardware Platform*". Komunikat zgłoszony na konf. MIKON 2022.

[5] ETSI TS 136 141 LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) conformance testing (3GPP TS 36.141 version 10.12.0 Release 10), Sect. 6.

WYBRANE METODY CHARAKTERYZACJI MIKROFALOWEJ MATERIAŁÓW MAGNETYCZNIE ŻYROTROPOWYCH

CHOSEN MICROWAVE CHARACTERIZATION METHODS OF MAGNETICALLY GYROTROPIC MATERIALS

Adam Pacewicz

Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska,
Nowowiejska 15/19, 00-665 Warszawa,
adam.pacewicz@pw.edu.pl

DOI: 10.15199/59.2022.4.81

Streszczenie: **Rezonansowe metody charakteryzacji mikrofalowej materiałów magnetycznie żyrotropowych mają potencjał, aby stać się standardem w przemyśle. Szczególnie znaczenie posiadają materiały ferrytowe, znajdujące zastosowanie w konstrukcji cyrkulatorów, izolatorów czy strojonych filtrów. W pracy zawarto podsumowanie prac nad rozwojem metod rezonansowych do charakteryzacji materiałów ferrimagnetycznych o kształcie sferycznym i planarnym, przede wszystkim szerokości linii rezonansu ferromagnetycznego i magnetyzacji nasycenia.**

Abstract: **Resonance methods for microwave characterization of magnetically gyrotropic materials have the potential to become an industry standard. Ferrite materials that are used in the construction of circulators, insulators and tuned filters are of particular importance. The paper summarizes the work on the development of resonance methods for the characterization of spherical and planar ferrimagnetic materials, especially the ferromagnetic linewidth and saturation magnetization.**

Słowa kluczowe: **rezonans ferromagnetyczny, kulka YIG, cienka warstwa YIG, metody rezonansowe, szerokość linii rezonansu ferromagnetycznego**

Keywords: **ferromagnetic resonance, YIG sphere, thin YIG film, resonance methods, ferromagnetic linewidth**

1. WSTĘP

Dynamiczna przenikalność magnetyczna ferromagnetyka namagnesowanego polem stałym do nasycenia wzdłuż kierunku $+z$ kartezjańskiego układu współrzędnych wyraża się tensorem Poldera $\bar{\bar{\mu}}$:

$$\bar{\bar{\mu}} = \mu_0 \begin{bmatrix} \mu & j\kappa & 0 \\ -j\kappa & \mu & 0 \\ 0 & 0 & \mu_{||} \end{bmatrix}, \tag{1}$$

którego wyrazy można opisać modelem Landaua-Lifszyca-Gilberta (LLG) [1]:

$$\mu = 1 + \frac{H_{0r} + j\,\alpha\,\hat{w}}{H_{0r}^2 - \hat{w}^2 + 2\,j\,\alpha\,H_{0r}\,\hat{w}}, \tag{2}$$

$$\kappa = \frac{\hat{w}}{H_{0r}^2 - \hat{w}^2 + 2\,j\,\alpha\,H_{0r}\,\hat{w}}, \tag{3}$$

$$H_{0r} = \frac{H_0}{M_s}, \tag{4}$$

$$\hat{w} = \frac{\hat{f}}{\gamma M_s}, \tag{5}$$

$$\alpha = \frac{\Delta H}{2H_0}, \tag{6}$$

gdzie μ_0 – przenikalność magnetyczna próżni, $\mu_{||} \approx 1$, H_0 – wewnętrzne stałe pole magnetyczne, $\hat{f}$ – częstotliwość zespolona, ΔH – szerokość linii rezonansu ferromagnetycznego, γ – współczynnik żyromagnetyczny, M_s – magnetyzacja nasycenia, j – jednostka urojona.

Bezpośrednie wyznaczenie wyrazów tensora Poldera (1) μ i κ metodami rezonansowymi, których dokładność jest co do zasady znacznie wyższa niż metod transmisyjno - odbiciowych [2], jest możliwe tylko dla pewnych geometrii badanej próbki, np. cylindrycznej całkowicie wypełniającej wysokość wnęki [3]. Jeżeli natomiast przyjąć, że μ i κ są dane przez model LLG, metody rezonansowe pozwalają na bardzo dokładne wyznaczenie wartości szerokości linii rezonansu ferromagnetycznego ΔH, zależnej od częstotliwości. Mogą one także zostać wykorzystane do wyznaczenia wartości magnetyzacji nasycenia M_s. W dalszej części pracy zostaną omówione metody poświęcone materiałom ferromagnetycznym o kształcie sferycznym i planarnym namagnesowanym stycznie.

W próbkach o ww. geometrii obserwuje się rezonanse Kittla na częstotliwościach zależnych od zewnętrznego pola magnetycznego H_{ext}. W przybliżeniu magnetostatycznym dla rodzajów podstawowych wyrażają się one zależnościami:

$$f = \gamma H_{ext} \tag{7}$$

dla kulek oraz

$$f = \gamma\sqrt{H_{ext}(H_{ext} + M_s)} \tag{8}$$

dla warstw namagnesowanych stycznie.

Metody rezonansowe można podzielić ze względu na stosowany układ sprzęgający próbkę z torem pomiarowym na szerokopasmowe i wąskopasmowe, wykorzystujące dyskretne częstotliwości. Układy szerokopasmowe są stosowane gdy rezonans Kittla charakteryzuje się wystarczająco dużą wartością dobroci (z praktycznego punktu widzenia ok. 100 lub więcej). Ich zaletą jest moż-

liwość prowadzenia pomiarów na dowolnych częstotliwościach mieszczących się w paśmie pracy układu. Błędy pomiarowe dla wysokostratnych ferromagnetyków można znacznie ograniczyć wykorzystując metodę polegającą na sprzężeniu rezonansu Kittla z rodzajem układu rezonansowego o wysokiej dobroci.

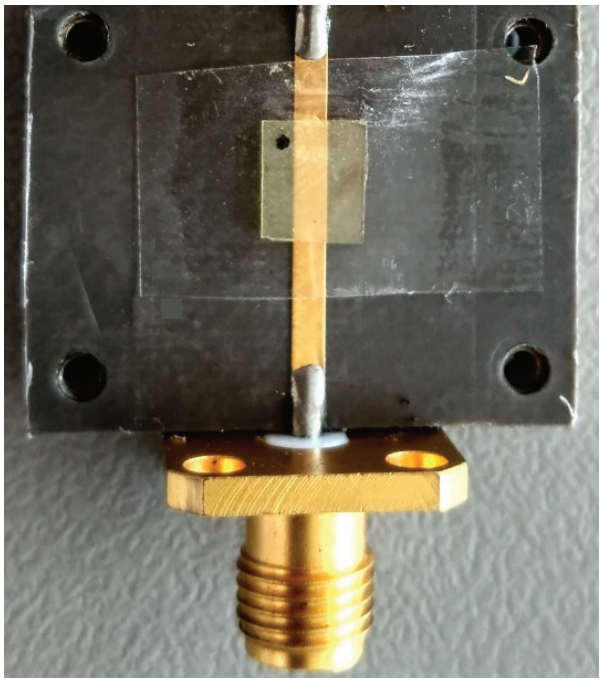
2. UKŁADY SZEROKOPASMOWE

2.1. Charakteryzacja ΔH warstw ferromagnetycznych przy użyciu linii transmisyjnych w technologii planarnej

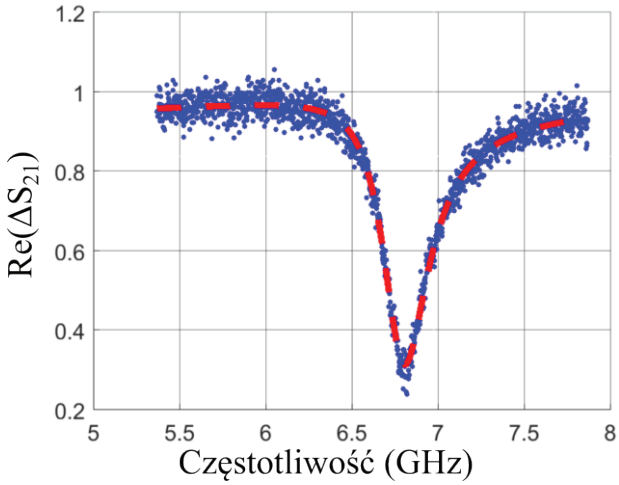
Wyznaczenie szerokości linii rezonansu ferromagnetycznego ΔH stycznie namagnesowanej cienkiej warstwy ferromagnetycznej (Rys. 1) za pomocą planarnej linii transmisyjnej polega na pomiarze widma wyrazu macierzy rozproszenia S_{21} w polu rezonansowym próbki i w polu referencyjnym, przy którym rezonans Kittla nie występuje. Następnie ze stosunku zespolonych widm (ΔS_{21}) można wyekstrahować ΔH za pomocą dopasowania krzywej rezonansowej [4] (Rys. 2 i 3). Wyznaczona w ten sposób szerokość linii jest szerokością zewnętrzną, tj. będącą szerokością połówkową części urojonej przenikalności efektywnej μ_{eff} danej zależnością

$$\mu_{eff} = \frac{\mu^2 - \kappa^2}{\mu}. \tag{9}$$

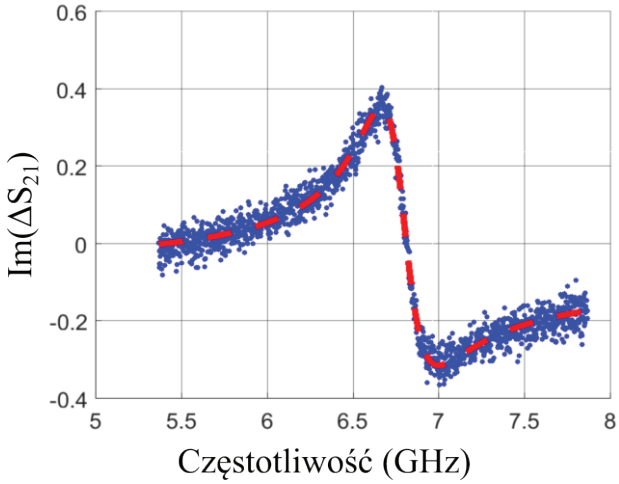
Wartość ΔH na ogół rośnie z częstotliwością i jej przykładowe widmo przedstawiono na Rys. 4.



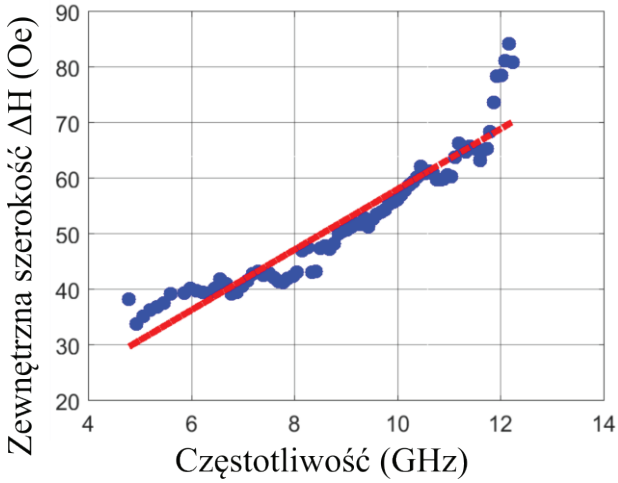
Rys. 1. Cienka warstwa (t = 50 nm) granatu itrowo-żelazowego YIG umieszczona na linii mikropaskowej.



Rys. 2. Część rzeczywista znormalizowanego widma S_{21} wraz z dopasowaniem [4].



Rys. 3. Część urojona znormalizowanego widma S_{21} wraz z dopasowaniem [4].



Rys. 4. Przykładowe widmo częstotliwościowe ΔH cienkiej ferromagnetycznej warstwy wraz z linią trendu.

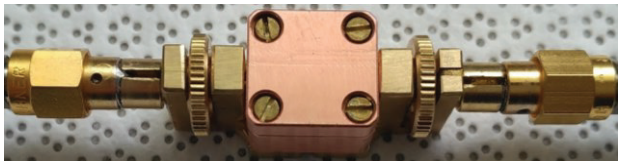
2.2. Charakteryzacja ΔH kulek ferrytowych we wnęce podkrytycznej

Charakteryzacja materiałów ferromagnetycznych opisaną wyżej metodą rezonansową za pomocą linii transmisyjnych będącymi strukturami otwartymi ma te

wady, że nie można łatwo regulować poziomem sprzężenia próbki z torem pomiarowym ani dokładnie określić strat promieniowania. Do charakteryzacji widma ΔH nisko- i średniostratnych kulek ferromagnetycznych zaproponowano [5] metodę wykorzystującą zamkniętą wnękę podkrytyczną sprzężoną z torem pomiarowym za pomocą regulowanych pętli. Dla niewielkich kulek ΔH typowo wyraża się zależnością

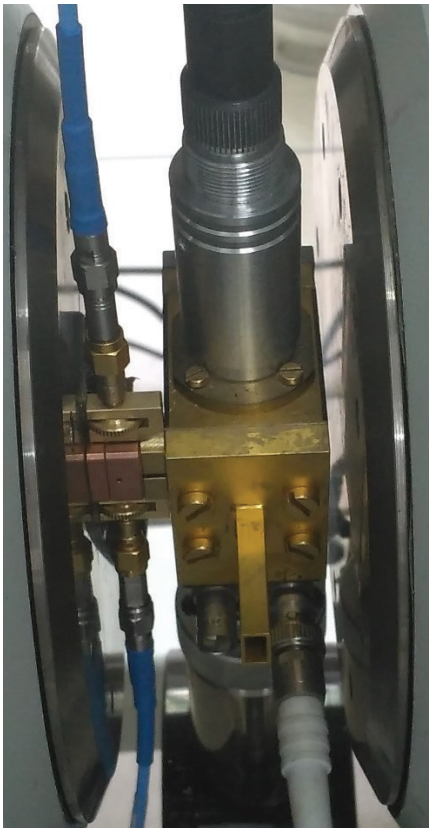
$$\Delta H = \frac{H_0}{Q}, \tag{10}$$

gdzie Q jest dobrocią nieobciążoną układu. Podobne układy mogą służyć do wyznaczenia magnetyzacji nasycenia próbki [6].



(a)

(b)

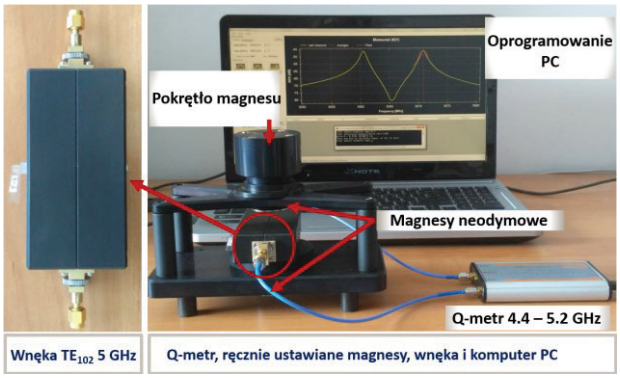


(c)

Rys. 5. Wnęka podkrytyczna wykorzystana do szerokopasmowej charakteryzacji rezonansowej kulek ferromagnetycznych. Widok z góry (a), wnętrze (b), wnęka umieszczona pomiędzy nabiegunnikami elektromagnesu (c).

3. UKŁADY WĄSKOPASMOWE

Alternatywnymi metodami wyznaczenia ΔH ferromagnetyków są metody oparte o teorię perturbacji [7] wykorzystujące np. prostokątne wnęki rezonansowe. Wymagają one przynajmniej względnej znajomości przykładanego stałego pola magnetycznego. Ponadto, warunkiem ich stosowalności jest słabe zaburzenie rodzaju wnęki przez rodzaj próbki obciążającej wnękę. Warunek ten jest zazwyczaj spełniony w przypadku charakteryzacji cienkich (o grubości np. rzędu dziesiątek nanometrów) warstw ze względu na bardzo niską ilość materiału, lecz często bywa niespełniony np. w przypadku kulek ferrytowych o rozmiarach makroskopowych, zwłaszcza o niskich wartościach ΔH. W takich przypadkach powstający systematyczny błąd pomiarowy można wyznaczyć poprzez modelowanie elektromagnetyczne [7]. Przy silnym sprzężeniu rodzaju próbki i wnęki dochodzi do rozszczepienia krzywej rezonansowej rodzaju wnęki na dwie, na podstawie dobroci których można wyznaczyć ΔH [8]. W takim przypadku na ogół nie ma potrzeby znajomości wartości przykładanego pola magnetycznego, co pozwala na zminiaturyzowanie układu do charakteryzacji ΔH. Realizację takiego układu wykonaną na Politechnice Warszawskiej przedstawiono na Rys. 6.



Rys. 6 Zminiaturyzowany układ do charakteryzacji kulek ferrytowych składający się z wnęki prostokątnej z umieszczoną wewnątrz próbką, Q-metru, ręcznie regulowanego magnesu stałego oraz komputera z oprogramowaniem.

LITERATURA

[1] D. M. Pozar, Microwave Engineering, 4th ed. John Wiley & Sons, Inc., 2011.

[2] P. Queffelec, S. Mallegol, and M. LeFloc'h, "Automatic measurement of complex tensorial permeability of magnetized materials in a wide microwave frequency range," IEEE Trans. Microw. Theory Tech., vol. 50, no. 9, pp. 2128–2134, Sep. 2002.

[3] Jerzy Krupka, "Measurements of All Complex Permeability Tensor Components and the Effective Line Widths of Microwave Ferrites Using Dielectric Ring Resonators," IEEE Trans. Microw. Theory Tech., vol. 39, no. 7, pp. 1148–1157, 1991.

[4] Sangita S. Kalarickal, Pavol Krivosik, Mingzhong Wu, Carl E. Patton, Michael L. Schneider, Pavel

Kabos, T. J. Silva, and John P. Nibarger, "Ferromagnetic resonance linewidth in metallic thin films: Comparison of measurement methods," J. Appl. Phys., vol. 99, no. 9, pp. 1–7, 2006.

[5] Adam Pacewicz, Jerzy Krupka, Bartlomiej Salski, Pavlo Aleshkevych, and Pawel Kopyt, "Rigorous broadband study of the intrinsic ferromagnetic linewidth of monocrystalline garnet spheres," Sci. Rep., vol. 9, no. 1, p. 9434, Dec. 2019.

[6] Jerzy Krupka, Adam Pacewicz, Bartlomiej Salski, Paweł Kopyt, Jeremy Bourhill, Maxim Goryachev, and Michael Tobar, "Electrodynamic improvements to the theory of magnetostatic modes in ferrimagnetic spheres and their applications to saturation magnetization measurements," J. Magn. Magn. Mater., vol. 487, p. 165331, Oct. 2019.

[7] A. Pacewicz, J. Krupka, P. Aleshkevych, B. Salski, P. Kopyt, and R. Frender, "Broad-ferromagnetic-linewidth non-metallic gyromagnetic spheres: A comparison of linewidth characterization methods," J. Appl. Phys., vol. 127, no. 16, p. 163903, Apr. 2020.

[8] Jerzy Krupka, Pavlo Aleshkevych, Bartlomiej Salski, and Pawel Kopyt, "Ferromagnetic linewidth measurements employing electrodynamic model of the magnetic plasmon resonance," Meas. Sci. Technol., vol. 29, no. 2, p. 025501, Feb. 2018.

ZDALNA REALIZACJA ZAJĘĆ W LABORATORIACH RADIOKOMUNIKACYJNYCH
REMOTE LEARNING IN LABORATORIES OF RADIOCOMMUNICATION

Jarosław Sadowski[1]; Jacek Stefański[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jarsad@eti.pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jstef@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.82

Streszczenie: **Referat zawiera opis założeń i budowy dwóch zestawów narzędzi do zdalnej realizacji zajęć laboratoryjnych, wykonanych w laboratoriach Politechniki Gdańskiej: zestaw symulatorów w postaci interaktywnych stron www oraz system zdalnego dostępu do sprzętowych stanowisk laboratoryjnych.**
Abstract: **This paper presents assumptions and realization of two tools for remote learning in laboratory classes: online simulators in form of interactive webpages and system for remote control of laboratory stands.**

Słowa kluczowe: **zdalne nauczanie, zdalny dostęp, symulatory**

Keywords: **remote learning, remote access, simulators**

1. WSTĘP

Pandemia COVID-19 w znaczący sposób utrudniła realizację nauczania w ostatnich dwóch latach na wszystkich stopniach edukacji, w tym na uczelniach wyższych. Niektóre formy zajęć, takie jak wykłady czy ćwiczenia, mogły być realizowane z użyciem gotowych, uniwersalnych systemów i platform do telekonferencji oraz pracy grupowej. Te same narzędzia były także używane z powodzeniem do realizacji spotkań online czy organizacji konferencji naukowych, a zdobyte w ten sposób doświadczenia oraz rozpowszechnienie narzędzi do wideokonferencji niewątpliwie spowodowały większe zainteresowanie takimi formami spotkań online także po przywróceniu tradycyjnych zajęć w salach (kontaktowych). Ale zajęcia praktyczne, takie jak laboratoria czy niektóre typy projektów studenckich nie dają się tak łatwo zrealizować w formie zdalnej i wymagają opracowania dedykowanych narzędzi, dostosowanych do specyfiki danych zajęć, ich tematyki, ale też celu zajęć i kompetencji, jakie powinien pozyskać student po zrealizowaniu tych zajęć. Niniejszy referat przedstawia przykład narzędzi do zdalnego nauczania przedmiotów laboratoryjnych w obszarze radiokomunikacji, zrealizowanych w Katedrze Systemów i Sieci Radiokomunikacyjnych Politechniki Gdańskiej.

2. ZAŁOŻENIA WSTĘPNE I WYMAGANIA

Podstawowym założeniem podczas opracowywania narzędzi do zdalnej realizacji zajęć laboratoryjnych było możliwie wierne odtworzenie funkcjonowania rzeczywistych stanowisk laboratoryjnych, używanych do realizacji zajęć stacjonarnych. Było to zgodne z wytycznymi władz uczelni odnośnie umożliwienia płynnego przejścia pomiędzy nauczaniem kontaktowym i zdalnym bez modyfikacji planów zajęć, dostarczając studentom w miarę możliwości taką samą wiedzę, doświadczenia i wrażenia niezależnie od formy realizacji laboratoriów.

W literaturze można znaleźć przykłady różnych podejść do nauczania zdalnego przedmiotów praktycznych. Część z nich bazuje na gotowym oprogramowaniu do automatyzacji pomiarów jak np. LabView [4, 5], które jednak nie zawsze sprawdza się, gdy celem realizacji zajęć jest nie tylko uzyskanie wyników pomiarów ale też nauczenie kolejności wykonywania czynności w celu uzyskania określonych rezultatów. Inne rozwiązania, np. opisane w [1, 3], zakładają użycie kamer internetowych do prezentacji wyglądu stanowiska i stanu przyrządów. Zrealizowane eksperymenty wykazały jednak, że błyszczące ekrany urządzeń z wyświetlaczami LCD lub CRT bardzo utrudniają uzyskanie dostatecznej jakości obrazu przy zmiennych warunkach oświetlania i powodują efekt mory, a dodatkowo w wielu przyrządach opisy i etykiety są wykonane na tyle małą czcionką, że nie są czytelne na obrazie z kamer. Z kolei w [2] zaproponowano zdalny dostęp do stanowiska laboratoryjnego przez udostępnienie komputera z oprogramowaniem sterującym stanowiskiem w laboratorium przez funkcję „zdalny pulpit". Metoda ta może jednak powodować potencjalne problemy z naruszeniem licencji na oprogramowanie jeśli licencja pozwala tylko na użytkowanie danego oprogramowania w obrębie uczelni, dlatego użycie komercyjnego oprogramowania zostało ostatecznie odrzucone podczas projektowania narzędzi do zdalnej realizacji zajęć laboratoryjnych. Ponadto przyjęto następujące założenia:

- brak konieczności zakładania kont użytkownika w kolejnym systemie, a jeśli uwierzytelnianie użytkowników jest niezbędne to powinno być realizowane przez centralny serwer e-nauczania (np. Moodle) dla minimalizacji gromadzenia danych osobowych;
- pożądana jest możliwość współdziałania studentów w grupach;
- należy unikać konieczności instalacji jakiegokolwiek oprogramowania na urządzeniach (komputerach) użytkowników;

– dostęp do infrastruktury uczelni przez funkcję „zdalnego pulpitu” nie może być używany z powodu potencjalnych problemów z licencjonowaniem oprogramowania oraz zbyt dużymi uprawnieniami użytkowników zdalnych;
– opracowane narzędzia powinny być w miarę możliwości niezależne od platformy stosowanej przez użytkowników, zdalny dostęp powinien być gwarantowany z większości komputerów z systemem Windows, MacOS i Linux a w miarę możliwości także z platform mobilnych (Android, iOS);
– pożądane stosowanie rozwiązań i narzędzi (np. kompilatorów) typu Open Source, które nie narzucają ograniczeń na metodę i cel użycia;

Biorąc pod uwagę powyższe założenia wykonano narzędzia do zdalnej realizacji zajęć laboratoryjnych, które można podzielić na dwie grupy: symulatory oraz interfejsy do zdalnego sterowania stanowiskami sprzętowymi przez interfejs www.

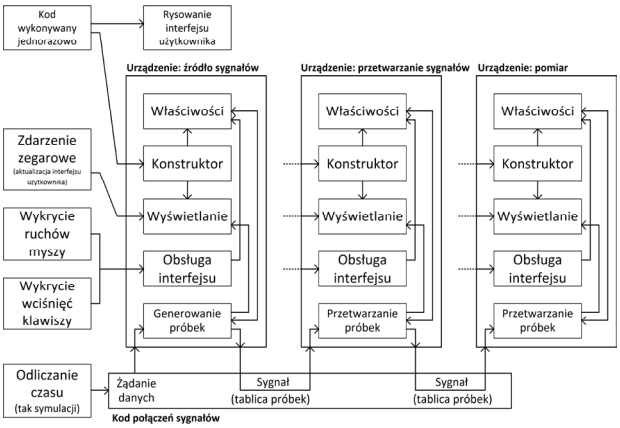
3. SYMULATORY STANOWISK LABORATORYCJNYCH

Wykonane symulatory stanowisk laboratoryjnych stanowią zasadniczo interaktywne strony www, które umożliwiają odtworzenie funkcjonowania rzeczywistych stanowisk sprzętowych, zarówno pod względem zachowania badanych urządzeń, funkcjonowania przyrządów pomiarowych jak i sposobu obsługi stanowiska przez użytkowników. Wszystkie symulatory zostały napisane w formie programów w języku javascript i są uruchamiane bezpośrednio w przeglądarce internetowej.

Stanowisko laboratoryjne do odtworzenia w symulatorze musi być podzielone na zestaw wirtualnych urządzeń, tj. części kodu odpowiedzialnej za zachowanie i wygląd wydzielonego fragmentu stanowiska, takiego jak źródło sygnału, urządzenie pomiarowe lub urządzenie wykonujące operacje na sygnałach. Każde wirtualne urządzenie jest zdefiniowane jako obiekt z zestawem obowiązkowych metod odpowiedzialnych za następujące zadania: konstruktor definiuje początkowe wartości zmiennych i wykonuje fragment kodu rysujący cały wygląd urządzenia na ekranie. Czynności wykonywane przez użytkownika, takie jak klikanie myszą, przewijanie „rolki” myszy czy naciskanie klawiszy są obsługiwane przez procedury obsługi zdarzeń, które zamieniają np. współrzędne kursora myszy na zmiany właściwości obiektu zapisywane w postaci zmiennych lokalnych. Dodatkowo efektem czynności wykonywanych przez studentów może być aktualizacja wyglądu urządzenia na ekranie, np. zmiana wyświetlanych wartości na wirtualnych wyświetlaczach.

Chociaż w większości przypadków kod odpowiedzialny za obsługę interfejsu użytkownika stanowi największą część całego kodu symulatora, najważniejszą funkcję symulatora realizuje jedna metoda odpowiedzialna za przetwarzanie próbek sygnałów. W symulatorach wszystkie sygnały są reprezentowane w postaci tablic próbek. Urządzenia będące źródłami sygnałów, jak np. generatory czy nadajniki, wytwarzają periodycznie tablice próbek w procedurze wyzwalanej licznikiem czasu. Te tablice próbek są następnie podawane jako dane wejściowe do procedur przetwarzania sygnałów w kolejnych wirtualnych urządzeniach, takich jak np. wzmacniacze, filtry, mieszacze itp. Urządzenia z tej grupy mogą mieć więcej niż jedno wejście i wyjście. Z kolei pojedyncze wejście mają urządzenia, które można ogólnie zakwalifikować jako przyrządy pomiarowe: woltomierze, oscyloskopy, analizatory widma itp. Przetwarzanie sygnałów (próbek) przez urządzenia z tej ostatniej grupy najczęściej skutkuje aktualizacją wyświetlanych danych, jak wyniki pomiarów w formie liczbowej lub graficznej (wykresy).

Metody realizujące generowanie i przetwarzanie tablic próbek są wywoływane dziesięć razy na sekundę. Kolejność ich wykonywania wynika ze struktury połączeń pomiędzy wirtualnymi urządzeniami i jest kontrolowana przez fragment kodu opisany na rys. 1 jako „kod połączeń sygnałów”.



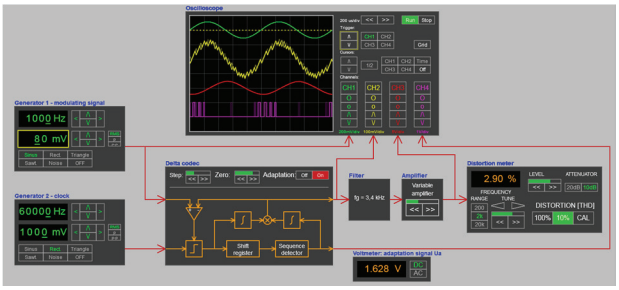
Rys. 1. Schemat blokowy struktury oprogramowania symulacyjnego

Podział programu symulatora na obiekty pozwala na uruchomienie w jednym symulatorze wielu instancji danego wirtualnego urządzenia, np. kilku takich samych mierników mierzących parametry sygnałów w różnych miejscach układu. Poza tym pozwala to na łatwe przenoszenie kodu przy tworzeniu kolejnych symulatorów.

Przygotowane symulatory urządzeń można podzielić na trzy kategorie. Pierwsza to urządzenia ogólnego przeznaczenia, jak woltomierze czy oscyloskopy, których sposób obsługi przez użytkowników nie jest krytyczny a wygląd w symulatorze nie musi odpowiadać żadnemu rzeczywistemu urządzeniu. Przykładem urządzeń z tej grupy może być oscyloskop przedstawiony na wyglądzie symulatora na rys. 2. Grupa druga to urządzenia specjalizowane, jak np. tester radiokomunikacyjny lub miernik zniekształceń, w których poza samym uzyskiwaniem wyników istotny jest też wygląd oraz sposób obsługi przez użytkowników (np. miernik zniekształceń na rys. 2). Natomiast w trzeciej grupie znajdują się symulatory modeli laboratoryjnych podlegających badaniom, np. modulatorów czy kodeków. W przypadku rzeczywistych stanowisk urządzenia te są budowane jako modele dydaktyczne o uproszczonej strukturze odpowiedniej do prezentowania zasad ich działania. Z tego powodu wirtualne modele tych urządzeń zostały zaprezentowane w symulatorach w formie schematów blokowych, a przykładem urządzenia z tej grupy może być koder delta przedstawiony na rys. 2.

Symulacja działania wszystkich wirtualnych urządzeń jest realizowana w dziedzinie czasu. Ogólna zasada symulacji bazuje na metodzie różnic skończonych w dziedzinie czasu (ang. Finite Differences in Time Domain), która do stabilnej pracy wymaga by częstotliwość próbkowania sygnałów była znacznie wyższa niż to wynika z kryterium Nyquista. Dobierając częstotliwość próbkowania należało mieć na uwadze, że niektóre symulatory wymagają zachowania rzeczywistej skali czasu, np. w przypadku gdy wyświetlane wartości są zależne od czasu realizacji ćwiczenia lub gdy częścią ćwiczenia jest odsłuch sygnałów z symulatora.

Wygląd całego symulatora dla przykładowego ćwiczenia dotyczącego badania właściwości kodowania delta z adaptacją pokazano na rys. 2. Przepływ sygnałów w symulowanym układzie jest reprezentowany strzałkami pomiędzy wyświetlanymi modelami urządzeń.



Rys. 2. Przykładowy wygląd symulatora stanowiska do badania kodowania delta

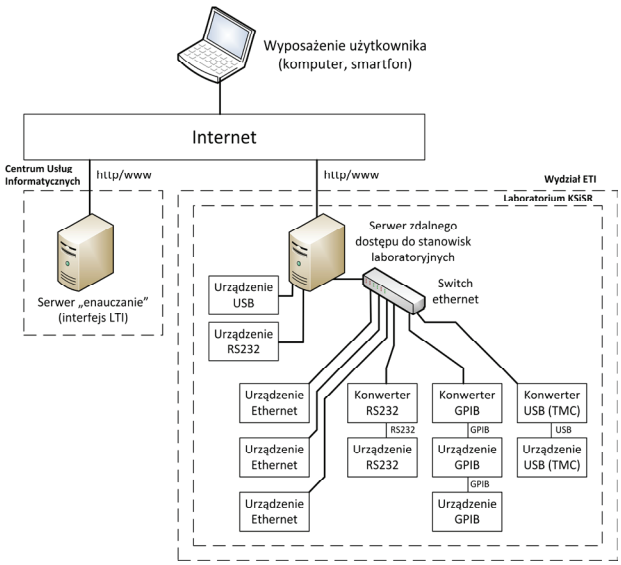
W ramach dotychczasowych prac wykonano po pięć symulatorów do przedmiotu Miernictwo Radiokomunikacyjne, Sygnały Telekomunikacyjne i Technika Bezprzewodowa, ponadto pojedyncze symulatory do laboratoriów z Podstaw Telekomunikacji, Techniki Satelitarnej i Urządzeń Radiokomunikacyjnych. Niewątpliwie do zalet symulatorów należy zaliczyć możliwość ich łatwej integracji z różnymi platformami e-learningowymi, możliwość jednoczesnego uruchamiania dowolnej liczby instancji, a także swoboda w realizacji zadań przez poszczególnych studentów w dogodnych terminach. Nie bez znaczenia jest także możliwość wykonania symulatorów do stanowisk laboratoryjnych, których nie ma w ofercie nauczania kontaktowego np. z uwagi na brak urządzeń pomiarowych. Natomiast najczęściej zgłaszanym przez studentów problemem przy używaniu symulatorów był brak możliwości pracy w grupach i utrudnione konsultowanie postępów prac pomiędzy studentami.

4. ZDALNY DOSTĘP DO STANOWISK LABORATORYJNYCH

Dokładność pomiarów realizowanych przy użyciu symulatorów stanowisk laboratoryjnych zależy w znacznym stopniu od dokładności odwzorowania właściwości mierzonego układu w modelu symulacyjnym i ograniczeń samej implementacji symulatora. Wad tych nie mają rozwiązania, w których interfejs www służy tylko do zdalnego dostępu do przyrządów pomiarowych w rzeczywistym stanowisku laboratoryjnym, dlatego jako uzupełnienie koncepcji symulatorów wykonany został prototyp systemu zdalnego dostępu do laboratoriów, ujednolicony pod względem wyglądu i obsługi z wykonanymi wcześniej symulatorami. Przy realizacji elementów systemu zdalnego dostępu przyjęto następujące założenia:

- stanowiska laboratoryjne będą dostępne dla studentów w zdefiniowanych terminach;
- studenci muszą zarezerwować termin realizacji danego ćwiczenia;
- ćwiczenia mogą być realizowane indywidualnie lub w grupach, w tym drugim przypadku wszyscy studenci w grupie mają jednocześnie dostęp do tych samych urządzeń pomiarowych;
- należy unikać konieczności instalowania jakiegokolwiek oprogramowania na urządzeniach użytkowników;
- nie należy używać funkcji „zdalny pulpit" ani oprogramowania z licencjami ograniczającymi użycie tylko do sieci uczelnianej;
- nawet urządzenia, które mają fabrycznie wbudowane interfejsy zdalnego sterowania, nie mogą być bezpośrednio udostępnione w sieci Internet;
- wszystkie polecenia zdalnego sterowania, wysyłane z interfejsu www, są weryfikowane pod kątem poprawności w celu wyeliminowania ryzyka wprowadzania urządzeń w laboratorium w nieprawidłowy stan lub przekroczenia dopuszczalnych zakresów nastaw.

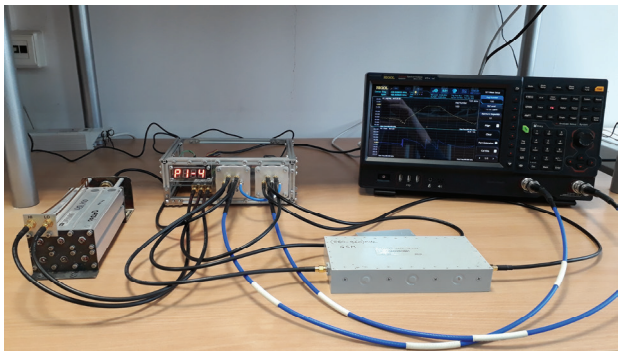
Od strony sprzętowej większość funkcji systemu zdalnego dostępu spełnia serwer zbudowany w oparciu o komputer przemysłowy, pracujący pod kontrolą systemu operacyjnego Linux. Do definiowania i rezerwowania terminów realizacji zajęć służy baza danych SQL z prostym interfejsem www, który jest zintegrowany z platformą e-nauczania (np. Moodle) poprzez interfejs Learning Tools Interoperability (LTI), pozwalający na obsługę uwierzytelniania użytkowników i przekazywania ich danych (identyfikatory, nazwiska) bez konieczności gromadzenia tych danych w serwerze zdalnego dostępu.



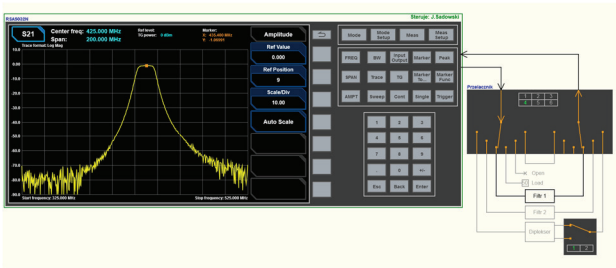
Rys. 3. Schemat blokowy części sprzętowej systemu zdalnego dostępu do stanowisk laboratoryjnych

Po uwierzytelnieniu użytkowników na ich komputerach uruchomiony zostaje kod interfejsu zdalnego dostępu w postaci programu w języku javascript. Program ten wysyła polecenia do serwera zdalnego dostępu w postaci pliku JSON. Polecenia te są najpierw przetwarzane przez program obsługi grup użytkowników, zintegrowany z serwerem www typu Apache przez interfejs FastCGI. Zadaniem tego programu jest łączenie poleceń od kilku użytkowników do sterowania jednym stanowiskiem i rozdzielanie wyników pomiarów z jednego stanowiska dla wszystkich studentów w grupie, realizującej dane ćwiczenie, a ponadto informowanie pozostałych studentów w grupie o czynnościach (kliknięcia, zmiany nastaw przyrządów) realizowanych przez daną osobę. Same polecenia sterujące przyrządami są przesyłane do kolejnego programu, który odpowiada za przebieg danego ćwiczenia, tj. zamienia komendy na ciąg instrukcji sterujących poszczególnymi urządzeniami pomiarowymi, wchodzącymi w skład danego stanowiska laboratoryjnego, niezależnie od sposobu podłączenia urządzeń do serwera. Ten program weryfikuje też zakresy zmian nastaw i wartości w ramach ćwiczenia oraz nie pozwala na wykonanie przez kontrolowane przyrządy poleceń spoza zakresu dopuszczalnych w danym ćwiczeniu.

Przykładem ćwiczenia, zrealizowanego w formie zdalnego dostępu do stanowiska laboratoryjnego, może być pomiar parametrów filtrów i diplekserów antenowych. Stanowisko do zajęć tradycyjnych składało się z analizatora obwodów (VNA), zestawu filtrów i dipleksera, oraz przewodów połączeniowych, przejściówek i kalibratorów. Na potrzeby zdalnego dostępu stanowisko zostało uzupełnione o zestaw przekaźników RF sterowanych przez mikrokomputer Raspberry PI. Wygląd stanowiska umożliwiającego zdalne sterowanie zamieszczono na rys. 4, natomiast rys. 5 zawiera wygląd interfejsu www do sterowania tym stanowiskiem.



Rys. 4. Przykładowe stanowisko laboratoryjne umożliwiające zdalne sterowanie i prezentację wyników



Rys. 5. Interfejs www do zdalnego sterowania stanowiskiem laboratoryjnym z rys. 4.

W przypadku zdalnego dostępu do przyrządów pomiarowych czynności realizowane przez studentów mogą prowadzić do nieprawidłowych wskazań stanowiska czy wręcz uniemożliwić pomiary. Dlatego poza bieżącą kontrolą poprawności poleceń sterujących, program obsługujący przebieg poszczególnych ćwiczeń powinien przywracać wszystkie przyrządy pomiarowe do ustalonego stanu przed udostępnieniem sterowania kolejnej grupie studenckiej.

Warto wspomnieć, że część urządzeń pomiarowych nie była projektowana do pracy ciągłej bez nadzoru, wiec zdalne laboratoria wiążą się zwykle z koniecznością zapewnienia nadzoru technicznego nad urządzeniami w trakcie realizacji ćwiczeń. W ramach dotychczasowych prac wykonano pełne oprogramowanie serwera zdalnego dostępu i kilka stanowisk do przedmiotu Miernictwo Radiokomunikacyjne, a dalszy rozwój systemu będzie zależny od decyzji co do możliwości wprowadzenia zdalnego nauczania do stałej oferty dydaktycznej.

5. PODSUMOWANIE

Mimo, iż najlepsze efekty dydaktyczne można osiągnąć podczas realizacji zajęć praktycznych z bezpośrednim dostępem studentów do stanowisk laboratoryjnych, narzędzia do e-learningu opracowane w okresie nauki zdalnej mogą nadal znajdować zastosowanie jako uzupełnienie oferty dydaktycznej w modelu nauczania hybrydowego, do realizacji zajęć fakultatywnych czy poszerzenia zakresu realizowanych ćwiczeń.

LITERATURA

[1] Andrieu G., Fredon T., Farah S., Aknin N., "Remote Laboratory Implementation for the Study of Transmission Lines", *2021 International e-Engineering Education Services Conference (e-Engineering)*, 2021: 107-110

[2] Bekasiewicz A., Pankiewicz B., Wojcikowski M., Klosowski M., Koziel S., "Application of Open-Hardware-Based Solutions for Rapid Transition From Stationary to the Remote Teaching Model During Pandemic", *IEEE Transactions on Education*, 2021, vol. 64, no. 3: 299-307

[3] Martin S., Fernandez-Pacheco A., Ruipérez-Valiente J. A., Carro G.,. Castro M, "Remote Experimentation Through Arduino-Based Remote Laboratories", *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 2021, vol. 16, no. 2: 180-186

[4] Mohammed A. K., El Zoghby H. M., Elmesalawy M. M., "Remote Controlled Laboratory Experiments for Engineering Education in the Post-COVID-19 Era: Concept and Example", *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 2020: 629-634

[5] Pavel I., Brânzilă M., Sărmășanu C., Donose C., "LabVIEW Based Control and Monitoring of a Remote Test-Bench Experiment for Teaching Laboratories", *2021 International Conference on Electromechanical and Energy Systems (SIELMEN)*, 2021: 398-402

SYSTEMY 5G W ZASTOSOWANIACH MILITARNYCH
5G SYSTEMS IN MILITARY APPLICATIONS

Marek Suchański[1]; Joanna Śliwa[2]; Krzysztof Kosmowski[3]

[1] Wojskowy Instytut Łączności – Państwowy Instytut Badawczy, Zegrze Południowe, m.suchanski@wil.waw.pl
[2] Wojskowy Instytut Łączności – Państwowy Instytut Badawczy, Zegrze Południowe, j.sliwa@wil.waw.pl
[3] Wojskowy Instytut Łączności – Państwowy Instytut Badawczy, Zegrze Południowe, k.kosmowski@wil.waw.pl

DOI: 10.15199/59.2022.4.83

Streszczenie: **Współczesne operacje wojskowe wymagają sieci łączności o zwiększonej pojemności i niezawodności, dlatego w opracowaniu omówiono sieci bezprzewodowe 5G i wyartykułowano ich główne funkcjonalności stanowiące o przydatności tych systemów do zastosowań w domenie militarnej. Jednocześnie omówiono zasadnicze zagrożenia dla ich bezpieczeństwa wynikające z rozwiązań programowych wykorzystanych do ich budowy.**
Abstract: **Modern military operations require communication networks with increased capacity and reliability, therefore the study discusses 5G wireless networks and articulates their main functionalities, which make these systems useful for military applications. At the same time, the article discusses the main threats to 5G systems' security resulting from their software – based architecture.**

Słowa kluczowe: **architektura 5G, bezpieczeństwo systemów 5G, NFV, SDN.**

Keywords: **5G architecture, 5G systems' security, NFV, SDN.**

1. WSTĘP

Gwałtowny wzrost liczby i różnorodności łączących się urządzeń i znaczący wzrost objętości i rodzaju przesyłanych w sieciach informacji, przy ograniczeniach możliwości technologii sieci komórkowych czwartej generacji (4G) stał się dla przemysłu i operatorów inspiracją do podjęcia działań mających na celu zdefiniowanie, rozwój i implementację systemów piątej generacji (5G). Systemy 5G spełniają oczekiwania użytkowników w zakresie zwiększenia pojemności sieci, wyższej przepływności, większej efektywności wykorzystania widma, zwiększonej szerokości pasm, mniejszych opóźnień, zmniejszonego zużycia energii, zwiększonej niezawodności, większej gęstości połączeń i wyższej mobilności. Taką efektywność sieci uzyskano dzięki zastosowaniu wirtualizacji sieci zdefiniowanej programowo, SDN. Ta architektura obejmuje modularność funkcji sieciowych, które mogą być wdrożone, konfigurowane i skalowane na życzenie tak by dostosować je do różnorodnych scenariuszy użycia w inteligentny i efektywny kosztowo sposób. Powyższe walory systemów 5G stały się przyczyną podjęcia analiz możliwości ich wykorzystania w domenie militarnej. Badania takie podjęto w wielu ośrodkach w tym m.in. w *European Defence Agency* (EDA) [1] i agendach NATO takich jak np. NATO *Communication and Information Agency* (NCIA) [2]. W prezentowanym opracowaniu przedstawiono analizę funkcjonalności systemów 5G stanowiących nową jakość dla zastosowań militarnych. Jednocześnie oczywistym jest, że od wszelkich systemów wprowadzanych do zastosowań militarnych wymaga się zapewnienia poufności i odporności na przewidywalne ataki przeciwnika. Ponieważ w systemach 5G występuje ogromna liczba rozwiązań programowych z otwartym kodem źródłowym tworzących obszar szczególnego ryzyka, dlatego w opracowaniu omówiono również problemy związane z zabezpieczeniem systemów 5G przed cyberatakami.

2. CHARAKTERYSTYKA SYSTEMÓW 5G

Prace nad 5. generacją systemów rozpoczęto w 2012 r. i w efekcie w roku 2015 Międzynarodowa Unia Telekomunikacyjna (ITU – *International Telecommunication Union*) opublikowała wizję kierunków rozwoju systemów radiokomunikacji mobilnej „*IMT for 2020 and beyond*" (IMT – *International Mobile Telecommunication*) [3].

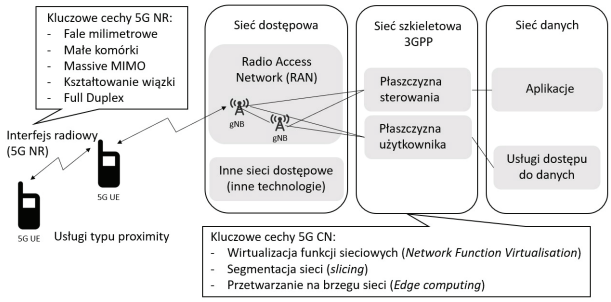
Wizja ta zawiera podstawowe scenariusze działania systemów 5G oraz zestaw 14-tu zdolności takich jak np. szczytowa prędkość transmisji, efektywność widmowa, opóźnienie itp., których osiągnięcie zapewni realizowalność tych scenariuszy. Opublikowane przez ITU-R (R oznacza sektor radiokomunikacyjny ITU) scenariusze stanowią syntezę szerokiego spektrum przypadków użycia sugerowanych przez producentów, organizacje regionalne i operatorów [4]. Scenariusze te dotyczą 3 podstawowych klas użytkowania systemu 5G, a mianowicie: rozszerzony mobilny szerokopasmowy dostęp do Internetu (*Enhanced Mobile Broadband*, eMBB); ultra niezawodna komunikacja o niskich opóźnieniach (*Ultra Reliable Low Latency Communications,* URLLC) oraz masowa komunikacja pomiędzy maszynami (*Massive Machine Type Communications,* mMTC).

Zapewnienie powyższych funkcjonalności wymaga nowatorskiego podejścia do filozofii działania sieci NGN (*New Generation Network*) i otwiera nową erę w radiokomunikacji, która zmienia całkowicie zasady jej funkcjonowania, prowadząc, do metamorfozy rozwiązań i zasad działania, zamieniając stopniowo rozwiązania

sprzętowe na programowo-informatyczne. Sieć 5G staje się więc siecią radioinformatyczną, a nie tylko radiokomunikacyjną [5].

Podczas prac nad ogólną architekturą systemów 5G przyjęto nową (względem systemów 4G) filozofię rozdziału funkcjonalności pomiędzy siecią dostępu radiowego RAN i siecią szkieletową CN. Uproszczony schemat funkcjonalny systemu (Rys. 1) przedstawiony przez 3GPP (*3rd Generation Partnership Project*) w 15 edycji standardu IMT-2020 ilustruje główne komponenty sieci komórkowej 5G [6]:

- **CN** (*Core Network*) – odpowiada za funkcjonalności nie związane z dostępem radiowym a niezbędne do stworzenia pełnej sieci takie jak np. uwierzytelnianie, taryfikacja, zestawianie połączeń „end to end" itd. Zainstalowanie tych funkcji w 5G CN zapewnia obsługę różnych technologii RAN przez tę samą sieć szkieletową.
- **RAN** (*Radio Access Network*) – odpowiada za funkcjonalności radiowe takie jak np. kolejkowanie, obsługa zasobów radiowych, protokoły retransmisji, kodowanie itd. RAN tworzony jest przez sieć stacji bazowych nowej generacji (gNB)
- **UE** (*User Equipment*) – wyposażenie użytkownika takie jak np. smartfony, terminale, pojazdy, inteligentne budynki itp.
- **Data Network** – usługi zapewniane przez operatora, aplikacje użytkownika. Może to być Internet lub inne sieci.



Rys. 1 Schemat funkcjonalny sieci 5G

Sieć szkieletowa 5G (5GCN) jest implementowana w postaci oprogramowania na wydajnych serwerach, które służą do realizacji chmury obliczeniowej i wówczas oprogramowanie realizuje określone usługi związane
z obsługą połączeń i innymi funkcjami sieciowymi. Z tego powodu architekturę sieci szkieletowej 5G określa się często jako architekturę SBA (*Service Based Architecture*), czyli architekturę bazującą na realizacji obsługi zadań (realizacji usług) lub architekturę zorientowaną usługowo. Koncepcja architektury SBA wynika z zastosowania takich technik jak wirtualizacja funkcji sieciowych (NFV – *Network Function Virtualization*) i sieci definiowanych programowo.

Wirtualizacja sieci oznacza zwykle [5] równoległą realizację złożonych funkcji sieciowych w postaci oprogramowania, a więc zastępowanie funkcji sieciowych przez oprogramowanie, z wykorzystaniem mechanizmów naśladowania urządzeń i funkcji sieciowych przez to oprogramowanie.

Wirtualizacja sieci stwarza operatorom możliwość udostępniania klientom zoptymalizowanych jej zasobów w postaci wirtualnych podsieci logicznych (zwanych warstwami, bądź plastrami – z ang. *slices*) dopasowanych do ich potrzeb. Każdy klient może dzięki temu zostać operatorem wirtualnym. Taka funkcjonalność nazywana jest określeniem *Network Slicing*.

Przeprowadzona w NCIA [2] analiza charakterystyk systemów 5G doprowadziła do wytypowania 5 niżej wymienionych funkcjonalności czyniących te systemy perspektywicznymi dla zastosowań militarnych:

- zakres wykorzystywanych częstotliwości,
- Nowe Radio 5G (*5G New Radio*),
- sieć szkieletowa 5G,
- usługi bezpośrednie/bliskie (*Proximity Services*),
- *Non-Terrestrial Networks (NTN)*.

Omówienie powyższych właściwości systemów 5G przedstawiono w dalszej części prezentowanego opracowania.

3. FUNKCJONALNOŚCI 5G ISTOTNE DLA ZASTOSOWAŃ MILITARNYCH

3.1. Zakresy wykorzystywanych częstotliwości

Istotnym aspektem każdej technologii radiowej jest jej przydatność do pracy w określonych zakresach częstotliwości, przewidzianych regulacjami międzynarodowymi. W tym aspekcie zakres częstotliwości w jakim planowane jest wdrożenie 5G NR stanowi o jego atrakcyjności również z punktu widzenia zastosowań militarnych. Fakt, że technologia ta będzie wdrażana na całym świecie powoduje, że kluczowe elementy np. układy scalone, czy też anteny będą produkowane komercyjnie również na zakresy częstotliwości będące w kręgu zainteresowania odbiorców militarnych. Zmniejsza to istotnie barierę ekonomiczną wprowadzenia technologii. Jako przykład można podać zakres częstotliwości przewidziany do wdrożenia 5G w Japonii, czyli 4,5 ÷ 4,9 GHz, który z kolei mieści się w paśmie NATO 4,4÷5,0 GHz (*Tactical Relay*).

Z kolei zakres fal milimetrowych (*mmWaves*) jest atrakcyjny ze względu na jego właściwości propagacyjne. Zjawiska fizyczne (silne tłumienie sygnału) w połączeniu z rozwiązaniami w zakresie techniki antenowej jak *massive MIMO (Multiple Input-Multiple Output)* i kształtowanie wiązki stanowią istotną korzyść z punktu widzenia możliwości ukrycia emisji elektromagnetycznej. Umożliwia to budowę trudnych do wykrycia systemów łączności o wysokiej przepustowości.

Kształtowanie wiązki oprócz aspektów związanych ze skrytością transmisji jest również pożądaną cechą ze względów użytkowych, eliminuje bowiem potrzebę pozycjonowania anten na masztach polowych.

3.2. Usługi bezpośrednie oraz funkcjonalność IAB

Z punktu widzenia scenariuszy operacyjnych typu „*infrastructure-less*" interesujące są tak zwane usługi bezpośrednie oraz zintegrowana sieć dostępowa (*5G Proximity Services (ProSe) and Integrated Access – Backhaul (IAB)*). Mogą one być wykorzystane do tworzenia sieci wielopoziomowych (*multi-tired*) i/lub kratowych (*mesh*).

Usługi typu *proximity* pozwalają na bezpośrednią wymianę danych pomiędzy terminalami użytkownika bez angażowania stacji bazowych. Do tego celu przewidziano interfejs nazywany *5G Sidelinks*. Z kolei funkcjonalność typu IAB zapewnia bezprzewodowe łączenie sieci dostępowych ze szkieletowymi i w ten sposób eliminuje się połączenia kablowe co jest szczególnie istotne w sieciach o dużej gęstości – w szczególności na polowych stanowiskach dowodzenia. Do tego celu przewidziano interfejs IAB. To ostatnie rozwiązanie stwarza możliwości rozwinięcia autonomicznego klastra 5G (*private network*) na polu walki.

Wirtualizacja sieci szkieletowej 5G i elementów sieci dostępowej oraz koncepcja segmentacji oferują nowe, atrakcyjne dla wojska możliwości. Segmentacja umożliwia na przykład wydzielenie na danym obszarze fragmentu publicznej infrastruktury 5G jako infrastruktury dedykowanej oraz wykorzystanie jej wraz z gotowymi komponentami militarnych (tzw. *system-in-a-box*) do uruchomienia prywatnej wojskowej sieci 5G.

3.3. Non-Terrestrial Networks

Kolejnym rozwiązaniem, które otwiera nowe możliwości w obszarze łączności specjalnej są sieci typu *Non-Terrestrial*. Wdrożenie tych rozwiązań umożliwi rozszerzenie zasięgu łączności taktycznej i dostarczenie usług aplikacyjnych dla obszaru operacji połączonych (sojuszniczych). Na przykład powietrzne systemy 5G mogłyby oświetlać obszar operacji w sposób podobny do systemów satelitarnych zapewniając łączność szerokopasmową mobilnym oddziałom na ziemi.

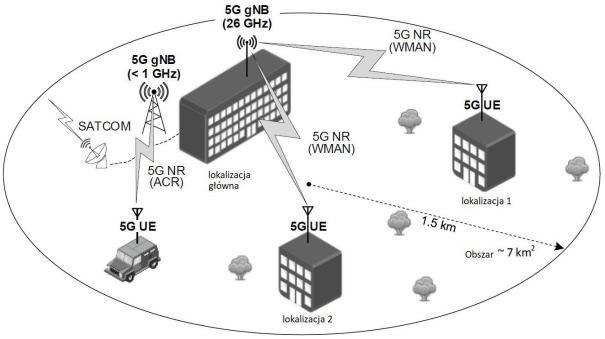
3.4. Scenariusze wykorzystania

W analizach dotyczących potencjału wykorzystania technologii 5G na potrzeby zastosowań militarnych w NCIA [2] zidentyfikowano 10 scenariuszy w czterech różnych obszarach zastosowań:

- mobilna infrastruktura telekomunikacyjna (*communication infrastructure scenario,* CIS) w operacjach ekspedycyjnych;
- operacje taktyczne na lądzie;
- komunikacja na morzu oraz;
- komunikacja statyczna.

W scenariuszach operacji ekspedycyjnych rozpatrywane jest utworzenie dużych i/lub małych stanowisk dowodzenia z wykorzystaniem własnych, rozwijanych elementów 5G w celu zapewnienia łączności w ramach stanowiska dowodzenia.

Poniżej, na Rys. 2 przedstawiono przykładowy scenariusz dla dużego stanowiska dowodzenia. Rozpatrywane jest wykorzystanie prywatnej sieci 5G dla zbudowania bezprzewodowej sieci metropolitalnej (WMAN, *Wireless Metropolitan Area Network*) wraz z usługą bezprzewodowej transmisji danych na tym obszarze (ACR, *Area Coverage Radio*). Prywatna komórka 5G pracująca w paśmie mmWave jest wykorzystana jako element sieci stacjonarnej w topologii punkt-wielopunkt w paśmie 26 GHz (zharmonizowany zakres NATO), natomiast dla zapewnienia usługi ACR przewiduje się implementację w zakresie poniżej 1 GHz, najlepiej w paśmie NATO 225-400 MHz.



Rys. 2 Duże stanowisko dowodzenia w ramach operacji ekspedycyjnej [2]

Scenariusz adaptujący inne rozpatrywane rozwiązania to na przykład scenariusz taktyczny z wykorzystaniem sieci NTN. W tej koncepcji systemy 5G są rozpatrywane jako rozszerzenie możliwości typowych sieci pola walki (CNR, *Combat Net Radio*). Łączność dalekiego zasięgu zapewnia prywatna stacja bazowa 5G umieszczona na jednostce latającej – co minimalizuje jednocześnie problem niekorzystnego ukształtowania terenu, czy też zwiększonych odległości pomiędzy stanowiskami dowodzenia różnych szczebli. Dla zapewnienia łączności na poziomie kompani przewiduje się tutaj komórki taktyczne w górnym zakresie pasma NATO 400 MHz, co zapewni łączność między terminalami taktycznymi umożliwiającymi również pracę w trybie *proximity*. Usługi tego typu wzmacniałyby możliwości taktycznych sieci łączności w przypadkach niesprzyjających warunków z punktu widzenia walki radioelektronicznej (WRE).

4. ZAGROŻENIA I WYZWANIA DLA BEZPIECZEŃSTWA SIECI 5G

Infrastruktura sieci 5G jest budowana w wykorzystaniem sieci SDN, które zakładają możliwość wykorzystania generycznych urządzeń posiadających zdolności przetwarzania (*computing*), przechowywania danych (*storage*) oraz przesyłania danych (*networking*) do budowy zróżnicowanych funkcji sieciowych, zarządzanych przez elastyczne i konfigurowane dynamicznie kontrolery i orkiestratory. To uniezależnienie od infrastruktury jest wyzwaniem dla bezpieczeństwa sieci.

Istnieje konieczność zapewnienia bezpieczeństwa wymiany danych między kontrolerami a funkcjami sieciowymi, która odbywa się z wykorzystaniem interfejsów oprogramowania (*Application Programming Interface* – API). API jest interfejsem między komponentami programowymi oraz zestawem procedur, protokołów i rozwiązań informatycznych do budowy aplikacji komputerowych. Niezbędne jest odpowiednie zabezpieczenie interfejsów poprzez wykorzystanie mechanizmów zapewnienia poufności (np. szyfrowanie) oraz uwierzytelnienia i autoryzacji dostępu do funkcji sieciowych.

Infrastruktura sieci 5G jest wrażliwa na standardowe ataki znane ze świata IT (*information technology*), związane z podatnościami oprogramowania i protokołów komunikacyjnych. Jednakże wyzwanie związane z komunikacją między głównymi komponentami sieci szkieletowej jest tym większe, gdy sieć ma być wykorzystywana w infrastrukturze krytycznej bądź w działaniach militarnych. W tym wypadku mechanizmy automatycznej aranżacji usług mogą być bardzo przydatne jako te, które w znaczący sposób mogą wesprzeć realizację usług o wysokich wymaganiach jakościowych (np. skrócić czas przetwarzania połączenia poprzez przeniesienie funkcji autoryzacji użytkowników na brzeg sieci, tzw. *Edge*), czy też zapewnić dodatkowy mechanizm analizy ruchu poprzez uruchomienie specjalnej metody filtrowania dla konkretnego przepływu. Jednak realizacja tego procesu wymaga wieloskokowej komunikacji między komponentami programowymi, która musi być zabezpieczona. W tym wypadku rozwiązaniem może być zastosowanie architektury typu *Zero Trust*, która zakłada, że wewnątrz sieci może znajdować się atakujący, dlatego też wymaga się uwierzytelniania i autoryzacji za każdym razem, kiedy realizowana jest interakcja między elementami sieci i kiedy zasoby mają zostać udostępnione.

Segment RAN jest narażony na ataki związane z powszechną dostępnością medium transmisyjnego. Mogą to być ataki ukierunkowane na zakłócanie sygnału radiowego, nieuprawniony dostęp do sieci radiowej lub „podsłuchiwanie" (zdekodowanie i detekcja sygnału innego użytkownika) [8]. Wprowadzona w sieci 5G otwarta specyfikacja interfejsu radiowego (ORAN) stanowi wyzwanie dla ich bezpieczeństwa. Nieodpowiednio zdefiniowane i źle zabezpieczone aplikacje ORAN, np. w warstwie fizycznej czy MAC (*medium access control*) mogą być podatne na wspomniane ataki.

Sieci 5G są również wrażliwe na ataki na dostępność. Szczególne wymagania jakości dla segmentu URLLC są możliwe do spełnienia m.in. dzięki zastosowaniu przetwarzania brzegowego (*Multi-Access Edge Computing* – MEC). Jednak niedostępność lub duże opóźnienia w dostępie do kluczowych usług sieciowych mogą spowodować znaczące utrudnienia w realizacji podstawowych wymagań jakościowych. Dlatego też ataki typu DoS, czy też DDoS (*Distributed / Denial of Service*) są bardzo znaczącym zagrożeniem dla sieci 5G [9]. Szczególnym atakiem typu DoS jest tzw. *signaling storm*, który ma na celu sparaliżowanie sieci dostępowej poprzez zajęcie zasobów przeznaczonych na przesyłanie informacji sygnalizacyjnej (*Control Plane* – CP) [10]. Wysyłane są duże ilości informacji sygnalizacyjnej w tym samym czasie przez wiele urządzeń. W rezultacie pozostali użytkownicy sieci nie będą mogli być obsługiwani. Ataki typu *signaling storm* mogą być rezultatem celowego skoordynowanego wysyłania informacji sygnalizacyjnej przez urządzenia atakującego, ale też błędów w aplikacjach mobilnych powodujących częste zrywanie połączenia.

W środowisku działań militarnych warto wspomnieć również o atakach typu *Man in the Middle* (MITM), np. polegających na stworzeniu tzw. fałszywej stacji bazowej (*Fake Base Station*), tzn. nadajnika emitującego informacje sygnalizacyjne typowe dla stacji bazowej 5G [7]. W sieciach 5G stosowane jest dwustronne uwierzytelnianie, tak więc w przypadku, kiedy fałszywa stacja bazowa nie jest zintegrowana z siecią szkieletową, użytkownik nie podłączy się do niej i jego dane nie zostaną przechwycone. Inną istotną kwestią jest także zapewnienie bezpieczeństwa podczas procedury przełączania użytkownika pomiędzy stacjami bazowymi (*Handover*). Na tym etapie atakujący może starać się wywołać błąd uwierzytelniania użytkownika w sieci 5G, poprzez modyfikacje przesyłanych kluczy. Ten rodzaj ataku można również sklasyfikować jako szczególny przypadek ataku MITM.

5. PODSUMOWANIE

W artykule przedstawiono charakterystykę systemów 5G w kontekście wsparcia współczesnych operacji wojskowych. Zidentyfikowano i scharakteryzowano istotne funkcjonalności stanowiące o korzyściach wynikających z zastosowania systemów 5G w militarnych systemach łączności. Jedną z takich funkcjonalności jest np. wykorzystanie milimetrowego zakresu częstotliwości stwarzającego możliwość stosowania systemów antenowych typu MIMO co w połączeniu z kształtowaniem wiązki antenowej zapewnia budowę na stanowiskach dowodzenia trudnych do wykrycia systemów łączności o wysokiej przepustowości. W artykule omówiono również zagrożenia dla bezpieczeństwa militarnych sieci 5G wynikające z wykorzystywanych do ich budowy rozwiązań informatycznych. Podatność na zagrożenia bezpieczeństwa takich sieci wynika z ich architektury opartej o otwarte, dobrze zidentyfikowane interfejsy oraz mechanizmy wirtualizacji.

LITERATURA

[1] EDA, 2021 *5G technologies for defence*. White paper, Version 1.0.

[2] Capella G., Kaprulu A., Bastos L., 2020, Potential of 5G technologies for NATO applications, *NCIA Technical Raport*, NCIA/TR/2020/NCBO14792/01.

[3] IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond, Recommendation ITU-R M.2083.

[4] Dahlman E., Parkvall S., Skold J., 2018, "5G NR The next generation wireless access technology". *Elsevier Ltd.*

[5] Gajewski S., Gajewska M., 2018 ”Filozofia sieci 5G-radioinformatyka i wirtualizacja. Część 1. Radioinformatyczny system 5G". *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne,* 11/2018.
[6] 3GPP TS 23.501 V15.4.0 System Architecture for the 5G System. Stage 2 (Release 15).
[7] Holtrup G., Lacube W., Percia D. David, Mermoud A., Bovet G., and Lenders V.,2021, “*5G System Security Analysis*”. arXiv preprint arXiv:2108.08700.
[8] Arjoune Y. and Faruque S., 2020, "Smart Jamming Attacks in 5G New Radio: A Review". *10th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 1010-1015.
[9] Javed M. A. and Niazi S. K., 2019, "5G Security Artifacts (DoS /DDoS and Authentication)". *International Conference on Communication Technologies (ComTech)*, pp. 127-133.
[10] Francois F., Abdelrahman O. H. and Gelenbe E., 2015, "Impact of Signaling Storms on Energy Consumption and Latency of LTE User Equipment". *IEEE 17th International Conference on High Performance Computing and Communications*, pp. 1248-1255.

WĄSKOPASMOWA KOMPENSACJA DIPP W MOBILNYM ŁĄCZU FRONTHAUL DSB-RFOF

NARROWBAND DIPP COMPENSATION IN THE MOBILE FRONTHAUL DSB-RFOF LINK

Zbigniew Zakrzewski

Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich, Bydgoszcz, zbizak@pbs.edu.pl

DOI: 10.15199/59.2022.4.84

Streszczenie: **W opracowaniu przedstawiono zjawisko dispersion induced power penalty (DIPP), które jest istotne w przypadku transportowania sygnału radiowego poprzez ścieżkę optyczną. Format DSB-RFoF oraz proste techniki modulacji i demodulacji optycznej IM/DD powodują, że podatność sygnału na DIPP wzrasta. Zaproponowana technika przestrajania lasera do obszaru okna z minimalnym DIPP sprawdza się, ale w sygnale CP-OFDM wymagana jest kompensacja poziomów podnośnych. Taką metodę, zsynchronizowaną ze zjawiskiem DIPP, zaproponowano w opracowaniu.**

Abstract: **The paper presents the dispersion induced power penalty (DIPP) phenomenon, which is important in the case of transporting a radio signal over an optical path. The DSB-RFoF format and the simple IM/DD optical modulation and demodulation techniques make the signal more susceptible to DIPP. The proposed technique to tune the laser to a window region with a minimum DIPP works fine, but compensation for subcarrier levels is required in the CP-OFDM signal. Such a method, synchronized with the DIPP phenomenon, is proposed in the paper.**

Słowa kluczowe: **DIPP, DSB-RFoF, dyspersja chromatyczna, radio poprzez światłowód, RFoF.**

Keywords: **chromatic dispersion, DIPP, DSB-RFoF, radio-over-fiber, RFoF.**

1. WSTĘP

Pierwotna architektura sieci w domenie radiowej systemów mobilnych typu Distributed Radio Access Network (D-RAN), także znana jako All-in-One Macro Base Station, zapewnia prostotę, gdyż polega na konstruowaniu stacji bazowych jako punktowo zespolonych systemów do przetwarzania i emisji sygnałów radiowych. Stosowanie takiego rozwiązania jest zasadne, gdy wykorzystanie stacji bazowej jest równomierne w czasie, zaś obciążenie szczytowe jest znacznie poniżej sprzętowych możliwości wyposażenia. Gwałtownie rosnące zapotrzebowanie na szerokopasmowy dostęp do Internetu zmieniło paradygmat w podejściu do projektowania nowych systemów komunikacji bezprzewodowej, szczególnie tych, które mogą służyć do świadczenia bardziej dynamicznych i wyrafinowanych usług sieciowych. W celu eliminacji wad rozwiązania D-RAN wprowadzono architekturę Centralized/Cloud Radio Access Network (C-RAN).

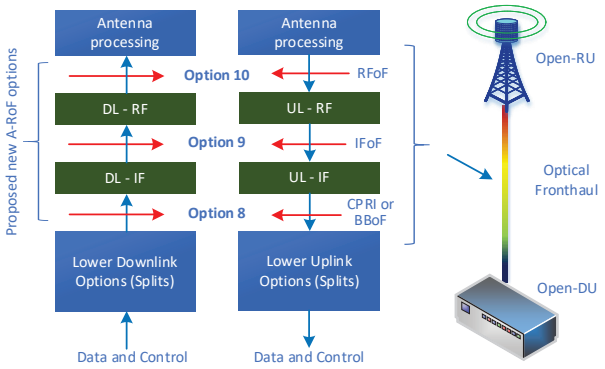
Koncepcja C-RAN znalazła zastosowanie w sieciach mobilnych od drugiej generacji (2G), zaś architektura z nią związana została upowszechniona dopiero w pierwszym dziesięcioleciu XXI wieku. Pojawienie się na rynku światowym systemów 4G wyłącznie z domeną komutacji pakietów, zrewolucjonizowało podejście do sposobów unowocześniania architektury C-RAN. Na pierwszym planie pojawił się światłowód oraz sieci wykorzystujące jego potencjał teletransmisyjny. W przypadku systemów mobilnych 5G, zarówno w domenie rdzeniowej jak też radiowej, trudno jest sobie wyobrazić funkcjonowanie tak wymagających sieci bez rozwiązań opartych na telekomunikacji światłowodowej. Zaproponowana w ramach projektu 5GPPP koncepcja sieci Xhaul [4,5] wyjątkowo dobrze pokazuje stopień integracji domen światłowodowej oraz radiowej, co znalazło szerokie zastosowanie w rozwiązaniach 5G [3].

W poniższym opracowaniu szczególną uwagę zwrócimy na wykorzystanie sieci światłowodowych w domenie radiowej kolejnych generacji systemów mobilnych. Sieci następnej generacji dostępu radiowego NG-RAN bardzo mocno są oparte na sieciach optycznych, co zapewnia konstruowanie efektywnych i szerokopasmowych interfejsów dla łączy backhaul (BH), midhaul (MH) oraz fronthaul (FH). Sieci bachhaul stanowią bardzo ważny szerokopasmowy łącznik między siecią rdzeniową (5GC) i siecią radiową NG-RAN. Nieco inne wykorzystanie mają sieci midhaul oraz fronthaul, które pełnią funkcje sieci dystrybuujących cyfrowe sygnały sterujące oraz sygnały radiowe o różnym stopniu przetworzenia. W tym miejscu ma zastosowanie technika radio-over-fiber (RoF) [6], która polega na przenoszeniu sygnałów radiowych z pasm podstawowego (BBoF), pośredniego (IFoF) oraz radiowego (RFoF) poprzez tor światłowodowy lub ścieżkę optyczną [20]. Tego rodzaju metodę zwiększania zasięgu dosyłania sygnału radiowego zdefiniowano po raz pierwszy w [15], gdzie określono jednostkę do wprowadzania sygnału radiowego na nośną optyczną oraz jednostkę służącą do odwracania tego procesu. Sygnały radiowe mogą być przesyłane poprzez medium światłowodowe w wersji scyfryzowanej, w technice digitized-radio-over-fiber (D-RoF) lub w wersji analogowej, w technice analog-radio-over-fiber (A-RoF) [19].

Technika D-RoF jest powszechnie wykorzystywana w systemach dostępu radiowego oraz w mikrofalowych systemach teletransmisyjnych. W tym zakresie najbardziej rozpoznawalny jest interfejs CPRI [7], który zmonopolizował rynek w tej dziedzinie i pozwala na dostarczanie scyfryzowanych sygnałów radiowych z pasma podstawowego. W celu rozszerzenia funkcjonalności interfejsu D-RoF, wprowadzono nowe rozwiązania zwiększające dynamikę pracy tego łącza. Do najbardziej rozpoznawalnych dynamicznych interfejsów D-RoF zaliczamy eCPRI

[8] oraz NGFI [13]. Wstępny zarys tych rozwiązań został przedstawiony przez 3GPP [2] oraz 5GPPP [3] do zastosowań w systemie mobilnym 5G. Upowszechnienie wdrażania interfejsów z rodziny NGFI w systemach 5G zapoczątkowało trend poszukiwania rozwiązań, które obniżą koszty transportowania, przygotowanych w chmurze brzegowej, scyfryzowanych sygnałów radiowych do zespołów radiowych (RU). Spowodowało to pojawienie się nowej koncepcji open radio access network (O-RAN) [23]. Do głównych graczy w tym zakresie, zajmujących się standaryzacją, można zaliczyć O-RAN Alliance, 3GPP oraz IEEE.

Technika A-RoF pozostaje ciągle w sferze badań, gdyż jest oparta na transportowaniu sygnałów poprzez światłowód w wersji pierwotnej. Tego rodzaju rozwiązanie jest szczególnie użyteczne, gdy istnieje potrzeba przesłania sygnału radiowego wysoko przetworzonego na poziomie nośnej pośredniej (IFoF) lub emisyjnej nośnej radiowej (RFoF). Technika analogowego transportowania sygnałów radiowych może być zrealizowana w architekturze fiber-to-the-antenna (FTTA) lub z wykorzystaniem całkowicie optycznej sieci transportowej. Propozycje rozwiązań umożliwiających jednoczesną pracę interfejsów D-RoF oraz A-RoF przedstawiono w pracach [26,27]. Systemy optyczne wykorzystujące zwielokrotnienie falowe (C/M/DWDM) z węzłami całkowicie optycznymi oraz elastyczne sieci optyczne (EON) stanowią bardzo dobrą strukturę sieciową nadającą się do transportowania sygnałów A-RoF. Na tej podstawie można stwierdzić, że sieci całkowicie optyczne dobrze wpisują się w koncepcję O-RAN [22], gdyż ich pojawienie się w domenie lokalnej oraz regionalnej zapewni transport sygnałów optycznych zmodulowanych w dowolnych nowoczesnych i klasycznych formatach.



Rys. 1. Nowe sprzęgi A-RoF umożliwiające transportowanie sygnałów radiowych w całkowicie optycznej światłowodowej sieci fronthaul [26,27]

Transmisja sygnałów optycznych w technice A-RoF jest bardzo efektywna spektralnie oraz pozwala na budowanie sygnałów na obrzeżu sieci lub w chmurze do postaci końcowej. Należy jednak brać pod uwagę fakt, że sygnał pochodzący z interfejsu A-RoF posiada strukturę sygnału analogowego i podlega silnemu oddziaływaniu zjawisk zachodzących w torze światłowodowym. W pracy [27] autor zaproponował wprowadzenie 2 nowych sprzęgów, w odniesieniu do modelu 3GPP [2], do obsługi transportu w formacie IFoF (Option 9) oraz RFoF (Option 10) – rys. 1. W systemie mobilnym B5G/6G planowane jest zastosowanie wysokich częstotliwości milimetrowych dochodzących do 100GHz. Wysoka częstotliwość radiowa zapewnia masowy i szerokopasmowy dostęp radiowy, ale na niezbyt dużym obszarze. Transportowanie sygnałów radiowych w formacie RFoF przy tak wysokiej radiowej częstotliwości nośnej wymaga stosowania specjalnych zabiegów, gdyż tor światłowodowy jest dyspersyjny. Mamy tutaj na myśli odcinki torów lub ścieżek optycznych, które nie powinny przekraczać 20km, jednak odziaływanie zjawisk, szczególnie związanych ze względnym opóźnieniem jest zauważalne. Masowy dostęp oraz lokalne lub regionalne zastosowanie interfejsu RFoF wymusza stosowanie prostych i tanich technik modulacji oraz demodulacji, dlatego rozwiązanie polegające na modulacji intensywności światła oraz bezpośredniej detekcji (IM-DD) wydaje się być optymalne. Podczas modulacji IM pojawia się sygnał w formacie DSB. Wstęgi boczne przenoszą informację, ale w punkcie prostego odbiornika DD podlegają detekcji kwadratowej mocno obniżającej czułość odbiornika. Zjawisko to jest ściśle zależne od częstotliwości radiowej RF, długości toru światłowodowego oraz współczynnika dyspersji chromatycznej występującego w kanale optycznym i jest określane jako DIPP.

2. WPŁYW DIPP NA TRANSPORT I ODBIÓR SYGNAŁU DSB-RFOF

Po raz pierwszy wpływ zjawiska DIPP na odbiór sygnału optycznego w formacie DSB został przedstawiony w opracowaniu [21]. W wielu kolejnych eksperymentach, przeprowadzonych w późniejszych latach, zostało to potwierdzone [10,12,14,24,25]. W międzyczasie pojawiły się nowe formaty modulacji sygnałów radiowych oraz wzrosło zapotrzebowanie na poszukiwanie rozwiązań RoF możliwych do wykorzystania w mobilnych systemach komunikacji bezprzewodowej. W wielu publikacjach przedstawiono metody i techniki osłabiania lub kompensowania wpływu DIPP na efektywny odbiór zmodulowanych sygnałów optycznych przenoszących mikrofalowe sygnały radiowe.

Sygnał zmodulowany w formacie RFoF (Option 10 – rys. 1), pochodzący z nadajnika optycznego, można określić podstawową zależnością:

$$E_{IM}(t)=\sqrt{1+m_{IM}\cos(2\pi f_{RF}t)}\exp\left(j2\pi\frac{c}{\lambda_0}t\right) \quad (1)$$

gdzie m_{IM} jest to wskaźnik głębokości modulacji, f_{RF} - nośna radiowa, zaś λ_0 - długość optycznej fali nośnej. Gdy wskaźnik m_{IM} jest mały, wówczas zależność (1) mona zapisać w postaci:

$$E_{IM}(t)=\left(1+m_{IM}\cos(2\pi f_{RF}t)\right)\exp\left(j2\pi\frac{c}{\lambda_0}t\right) \quad (2)$$

Po trygonometrycznym przekształceniu zależności (2) i uzależnieniu dwuwstęgowego sygnału DSB od zjawisk dyspersyjnych, charakterystycznych dla toru światłowodowego, uzyskamy:

$$E_{IM}(t,\bullet)=\exp\left(j2\pi\frac{c}{\lambda_0}t\right)+ \\ +\frac{m_{IM}}{2}\left\{\exp\left[j2\pi\left(\frac{c}{\lambda_0}+f_{RF}\right)t-\Phi(\bullet)\right]+\right. \\ \left.+\exp\left[j2\pi\left(\frac{c}{\lambda_0}-f_{RF}\right)t-\Phi(\bullet)\right]\right\} \quad (3)$$

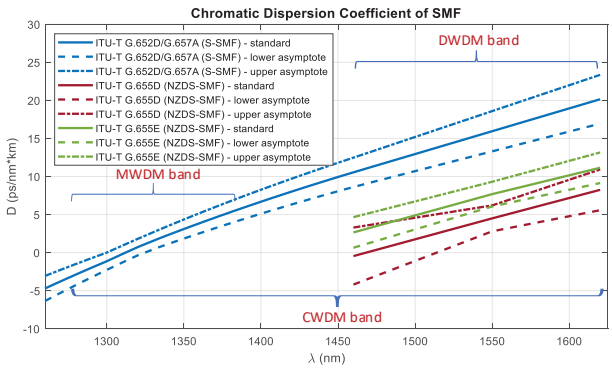
Z punktu widzenia fotodetektora pracującego w formacie bezpośredniej detekcji (DD), ścieżka optyczna jest wąskopasmowym filtrem z płaską amplitudowo charakterystyką transmitancyjną oraz liniowym współczynnikiem grupowego opóźnienia. Na tej podstawie możemy określić, że funkcja przejścia dla toru światłowodowego może być reprezentowana przez [11]:

$$H(f_{RF},\lambda_0)=\exp\left(-j\Phi(f_{RF},\lambda_0)\right), \quad (4)$$

zaś

$$\Phi(f_{RF},\lambda_0)=\pi D(\lambda_0)L\frac{\lambda_0^2}{c}f_{RF}^2 \quad (5)$$

gdzie L stanowi długość ścieżki optycznej, zaś $D(\lambda_0)$ jest to współczynnik wypadkowej efektywnej dyspersji chromatycznej, który zależy od optycznej długości fali nośnej.



Rys. 2 Charakterystyki współczynnika dyspersji chromatycznej światłowodów telekomunikacyjnych [16-18] stosowanych w optycznych sieciach ze zwielokrotnieniem falowym xWDM.

Przykładowe charakterystyki dyspersji chromatycznej, wykorzystywane do obliczeń, zostały przedstawione na rysunku 2.

Sygnał elektryczny za fotodetektorem jest proporcjonalny do natężenia światła na niego padającego, zatem można tutaj zastosować regułę detekcji kwadratowej [12]:

$$i_{IM}(t)\propto\left|E_{IM}(t,f_{RF},\lambda_0)\right|^2 \quad (6)$$

Po podstawieniu zależności (3) do (6) oraz pominięciu składników nieprzenoszących informacji w sygnale zmodulowanym, otrzymujemy

$$i_{IM}(t,f_{RF},\lambda_0)\approx \\ \approx 2m_{IM}\cos\left(\pi D(\lambda_0)L\frac{\lambda_0^2}{c}f_{RF}^2\right)\cos(2\pi f_{RF}t) \quad (7)$$

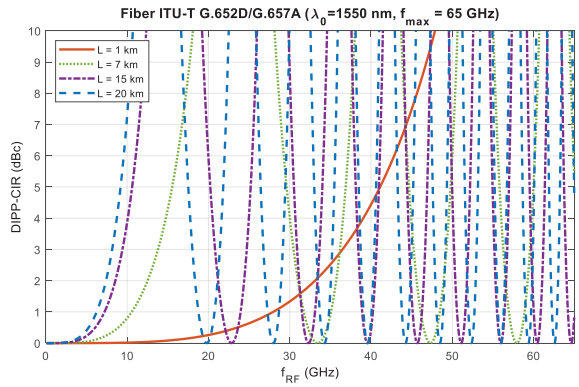
Finalnie moc sygnału obserwowanego po stronie fotodetektora można wyrazić przy pomocy przybliżonej zależności:

$$P_{IM}(f_{RF},\lambda_0)\approx\cos^2\left(\pi D(\lambda_0)L\frac{\lambda_0^2}{c}f_{RF}^2\right). \quad (8)$$

Praktyczne jest, aby wyniki obliczeń przedstawiać w odniesieniu do stałego komponentu, jakim jest nieprzenosząca informacji nośna optyczna oraz w skali logarytmicznej, aby zwiększyć zakres odczytów na wykresach. W tym celu wprowadzimy określenie DIPP – carrier to interference ratio (DIPP-CIR) oraz zależność stosowaną do obliczeń w dalszej części opracowania:

$$\text{DIPP-CIR}_{IM}= \\ =10\log\frac{P_C}{P_{SBs}}=10\log\cos^2\left(\pi D(\lambda_0)L\frac{\lambda_0^2}{c}f_{RF}^2\right)= \\ =20\log\cos\left(\pi D(\lambda_0)L\frac{\lambda_0^2}{c}f_{RF}^2\right) \quad (9)$$

W praktyce sygnał ten będzie podlegał dodatkowemu osłabieniu w wyniku tłumienia w ścieżce optycznej. W optycznym interfejsie fronthaul łączne tłumienie nie powinno przekraczać 20 dB (według O-RAN Alliance [22]), jednak w ścieżce optycznej bazującej na sieci pasywnej PON tłumienie to może być osiągnięte bardzo szybko ze względu na obecność sprzęgaczy optycznych.



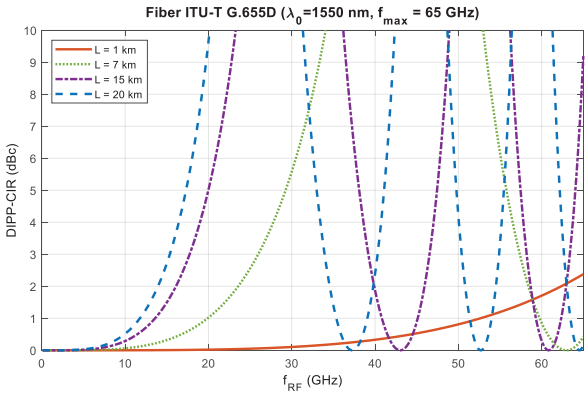
Rys. 3 Wyniki obliczeń DIPP-CIR przy propagacji sygnału DSB w ścieżce optycznej G.652D/G.657A o zmiennej długości (przypadek nośnej RF o zmiennej częstotliwości w zakresie do 65GHz oraz stałej nośnej optycznej o długości fali 1550nm)

Wyniki obliczeń i symulacji dotyczących zjawiska DIPP występującego w łączu fronthaul podczas transportowania sygnału DSB-RFoF zostały zaprezentowane przez autora w pracy [28]. Do obliczeń oraz symulacji wykorzystano charakterystyki dyspersyjne światłowodów przedstawione na rysunku 2. Założono, że długość ścieżki optycznej fronthaul nie przekroczy 20 km, co jest zgodne z wymaganiami 3GPP oraz O-RAN Alliance. W celu tzw.

ominięcia zjawiska DIPP zaproponowano metodę przestrajania lasera, jako źródła nośnej optycznej w rejon, w którym występuje konstruktywna interferencja składowych tworzących wstęgi boczne (SBs). Zakres przestrajania jest ściśle zależny od tzw. okien dobrej transmisji, które można wyznaczyć metodą odcięcia w ustalonym progu QoOCH [28]:

$$\text{QoOCh} = \begin{cases} \text{strong} \rightarrow \text{DIPP-CIR}_{IM} < 3\text{dB} \\ \text{weak} \rightarrow 3\text{dB} < \text{DIPP-CIR}_{IM} < 10\text{dB} \\ \text{unusable} \rightarrow \text{DIPP-CIR}_{IM} > 10\text{dB} \end{cases} \quad (10)$$

Podczas ustalania progów określono, że dopuszczalne tłumienie wynikające ze zjawiska DIPP może wynosić 10dB (10), gdyż trzeba mieć jeszcze zapas na tłumienie absorpcyjne i rozproszeniowe występujące w światłowodzie jednomodowym.



Rys. 4 Wyniki obliczeń DIPP-CIR przy propagacji sygnału DSB w ścieżce optycznej G.655D o zmiennej długości (przypadek nośnej RF o zmiennej częstotliwości w zakresie do 65GHz oraz stałej nośnej optycznej o długości fali 1550nm)

Przykładowe wyniki obliczeń DIPP-CIR zostały przedstawione na rysunkach 3 i 4. Parametry dobrano w taki sposób, aby można było porównać efekty wygaszania wstęg w tych samych warunkach, to znaczy w tym samym zakresie częstotliwości radiowych (RF) sygnałów modulujących oraz tej samej optycznej długości fali nośnej. Długości ścieżek optycznych są także takie same. Obliczenia zostały wykonane dla dwóch różnych światłowodów telekomunikacyjnych, tj. G.652D (lub G.657A) oraz G.655D. Współczynniki dyspersji chromatycznej dla tych dwóch włókien przy długości fali 1550nm wynoszą odpowiednio 15.95 oraz 4.5 ps/nm*km (wielkości możliwe do odczytania z wykresów średnich wartości umieszczonych na rys. 2). Porównanie tych charakterystyk pokazuje, że przy spadającym współczynniku dyspersji chromatycznej okna dobrej transmisji są coraz szersze, co powoduje, że zbocza charakterystyk są mniej strome.

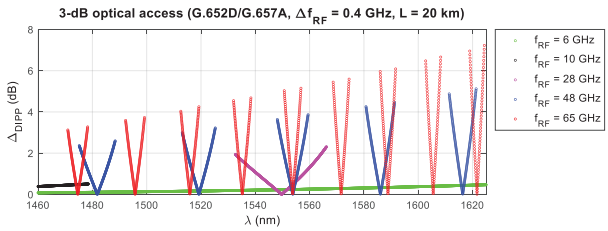
3. WĄSKOPASMOWA KOMPENSACJI DIPP W ZAKRESIE KANAŁU RADIOWEGO

Powyższe obliczenia zostały wykonane dla niezmodulowanej nośnej radiowej. Rzeczywisty sygnał radiowy składa się z większej liczby składowych częstotliwościowych, które łącznie muszą zmieścić się w wyznaczonym kanale częstotliwościowym.

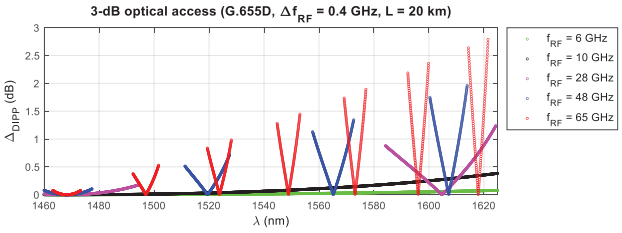
W interfejsie 5G-NR w kierunku „downlink" ma zastosowanie modulacja CP-OFDM, która pozwala na dynamiczne przydzielanie zasobów radiowych różnym terminalom bezprzewodowym (UE) w ramach tego samego kanału częstotliwościowego. Najmniejsza jednostka składa się z 12 podnośnych OFDM, zaś częstotliwościowa szerokość tego zasobu jest zależna od parametru numerologicznego μ, który między innymi definiuje odstęp między podnośnymi OFDM. W wyniku połączenia wszystkich bloków zasobów radiowych, tworzy się sygnał, który cechuje się szerokością dochodzącą do 400 MHz. Oczywiście cały kanał częstotliwościowy nie jest pokryty tym sygnałem, gdyż wymagane jest stosowanie częstotliwościowych odstępów ochronnych [1], które ograniczają rzeczywistą częstotliwościową szerokość zmodulowanego sygnału OFDM.

W celu określenia wpływu częstotliwościowej szerokości sygnału na nierównomierność przenoszenia podnośnych OFDM w łączu RFoF, wprowadzimy zależność na stopień zróżnicowania poziomów skrajnych podnośnych należących do radiowego kanału RF. Do tego celu wykorzystamy zależność (9), w której częstotliwościowa odległość skrajnej podnośnej OFDM od środkowej nośnej RF jest określona jako połowa szerokości kanału częstotliwościowego:

$$\begin{aligned} &\Delta\text{DIPP-CIR}_{(\text{dB})}\left(f_0, \Delta f_{RF}\right) = \\ &= \left|\text{DIPP-CIR}_{(\text{dBc})}\left(f_0 - \frac{\Delta f_{RF}}{2}\right) - \right. \\ &\left. -\text{DIPP-CIR}_{(\text{dBc})}\left(f_0 + \frac{\Delta f_{RF}}{2}\right)\right| \end{aligned} \quad (11)$$



Rys. 5 Różnicowy DIPP-CIR w funkcji optycznej długości fali nośnej dla wybranych radiowych częstotliwości nośnych oraz 3-dB progu (włókno G.652D/G.657A)



Rys. 6 Różnicowy DIPP-CIR w funkcji optycznej długości fali nośnej dla wybranych radiowych częstotliwości nośnych oraz 3-dB progu (włókno G.655D)

Wartość ΔDIPP-CIR, określoną przez zależność (11), można odnieść do pierwszego lub drugiego progu (10), który określa dopuszczalną wartość DIPP-CIR dla środkowej częstotliwości nośnej f_0 kanału optycznego.

Na rysunkach 5 i 6 przedstawiono wyniki obliczeń, które pokazują jak mocno różną się poziomem mocy

skrajne składowe sygnału CP-OFDM, które są przenoszone na nośnej radiowej RF oraz nośnej optycznej z zakresu pasma systemu DWDM (optyczne pasma S+C+L). Obliczenia wykonano dla kanału o szerokości 400MHz oraz dwóch generacji jednomodowych światłowodów telekomunikacyjnych.

Oddziaływanie DIPP można znacząco obniżyć w przypadku kompensacji dyspersji chromatycznej. Jest to jednak proces, który zwykle realizuje się po stronie optycznej. Ścieżka optyczna fronthaul bazuje na światłowodowym łączu lokalnym lub łączu systemu xWDM, gdzie dyspersją zarządza się w odniesieniu do strumieni cyfrowych. Innym rozwiązaniem jest wprowadzenie częściowej wstępnej kompensacji nierównomierności w sygnale RF, która jest obserwowana za fotodetektorem. Rozwiązanie to nie zredukuje obniżenia czułości odbiornika DSB-RFoF, ale pozwoli na poprawne odebranie większej porcji bloków PRB zawierających podnośne CP-OFDM w przypadku kierunku DL (5G-NR).

Zbiór wartości DIPP-CIR poszczególnych podnośnych CP-OFDM za fotodetektorem można wyrazić przy pomocy wektora:

$$\mathbf{A}_{D_N} = \left[\text{DIPP-CIR}\left(f_1\right),\ldots,\text{DIPP-CIR}\left(f_N\right)\right]. \quad (12)$$

W następnym kroku należy określić środek korygowanego kanału częstotliwościowego $\Delta f_{RF}/2$ i wyznaczyć wielkość DIPP-CIR dla tej częstotliwości podnośnej:

$$A_{D_{\Delta f}} = \text{DIPP-CIR}\left(\frac{\Delta f_{RF}}{2}\right). \quad (13)$$

Zanegowana charakterystyka elementów wektora $\mathbf{A}_{D_N}$ umożliwi wstępną korekcję przyszłego efektu DIPP, co po związaniu (12) i (13) można określić poprzez:

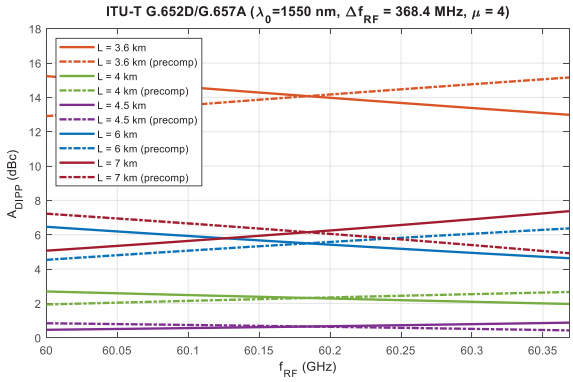
$$\mathbf{A}_C = -\left(\mathbf{A}_{D_N} - A_{D_{\Delta f}}\right) = A_{D_{\Delta f}} - \mathbf{A}_{D_N}. \quad (14)$$

Jeśli chcemy porównać charakterystyki (12) i (14) w odniesieniu do ważonego punktu $\Delta f_{RF}/2$, wówczas należy ponownie dodać wartość DIPP-CIR dla tej częstotliwości podnośnej:

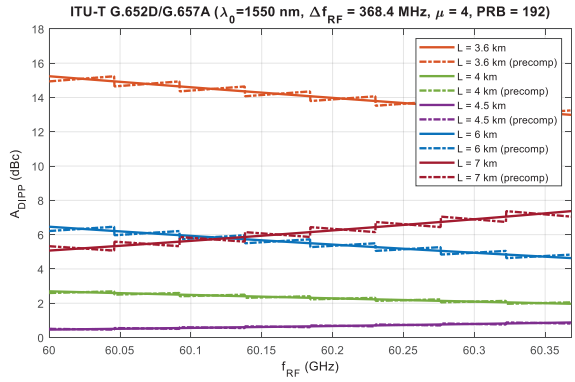
$$\mathbf{A}_{C_{DIPP}} = A_{D_{\Delta f}} - \mathbf{A}_{D_N} + A_{D_{\Delta f}} = 2A_{D_{\Delta f}} - \mathbf{A}_{D_N}. \quad (15)$$

Charakterystyka wektora punktów kompensujących może być także wyznaczona w węższym zakresie, tj. w paśmie bloku PRB lub grupy bloków PRB, które są przydzielane konkretnemu terminalowi mobilnemu. Mechanizm wąskopasmowej kompensacji DIPP może być przeprowadzony na przykład w ramach procesu cyfrowego kształtowania wiązki (DBF) w lokalizacji O-DU (kierunek DL) – rys. 1.

Wyniki przykładowych obliczeń wskazujących na możliwość kompensacji DIPP w zakresie kanału częstotliwościowego i zasobu PRB odpowiednio zostały przedstawione na rysunkach 7 i 8. Posłużono się zależnością (15), gdzie wektor wartości DIPP-CIR został ustalony dla częstotliwości podnośnych w paśmie 60GHz. Do obliczeń został wzięty kanał o szerokości 400MHz (5G-NR), który mieści 1536 podnośnych ustawionym względem siebie w odległości definiowanej przez parametr numerologiczny $\mu = 4$. Kanał optyczny został zlokalizowany wokół nośnej o długości fali 1550nm.



Rys. 7 Przykłady charakterystyk kompensacji DIPP, po przejściu sygnału poprzez światłowód G.652D/G657A, w zakresie całego kanału częstotliwościowego 60GHz.



Rys. 8 Przykłady charakterystyk wąskopasmowej kompensacji DIPP, po przejściu sygnału poprzez światłowód G.652D/G.657A, w zakresie zasobu PRB

Obliczenia wykonano dla jednego z najgorszych przypadków, tj. dla światłowodu G.652D (lub G.657A), który w środku optycznego pasma C wykazuje wysoką wartość współczynnika dyspersji chromatycznej. Tego rodzaju obliczenia mogą być wykonane dla dowolnej optycznej nośnej, która mieści się w dopuszczalnym oknie DIPP-CIR. Po stronie O-DU tego rodzaju obliczenia mogą być wykonane tylko w przypadku znajomości wartości dwóch parametrów ścieżki optycznej, tj. jej długości oraz współczynnika dyspersji chromatycznej dla ustalonej optycznej nośnej. Wielkości te mogą być pozyskane przy pomocy prostego testu OTDR lub obliczeniowego [9] przeprowadzonego pomiędzy O-DU oraz O-RU.

4. PODSUMOWANIE I KIERUNEK DALSZYCH BADAŃ

Potrzeba wąskopasmowej kompensacji zjawiska DIPP pojawiła się w trakcie prowadzenia symulacyjnych badań nad przenoszeniem sygnału radiowego poprzez ścieżkę optyczną w formacie DSB-RFoF [28]. Wprawdzie procesor jest w stanie przeprowadzić normalizację

oraz korekcję poziomów podnośnych OFDM, ale tylko w wąskim zakresie. Jest to proces podobny do tego, który został wprowadzony w technice modulacji DMT.

Przewidywanie zjawiska DIPP i reagowanie na nie na pewno podniesie efektywność pracy demodulatora CP-OFDM. Na rysunkach 7 i 8 można dostrzec, że nie we wszystkich przypadkach warto polegać na jednej metodzie wstępnej kompensacji. W przypadku lokalizacji zasobu radiowego PRB położonego w obszarze o niższym DIPP-CIR, wskazana jest korekcja wąskopasmowa w zakresie samego bloku. Dla przypadków przeciwnych warto jest polegać na korekcji w całym kanale częstotliwościowym. Należy przy tym zaznaczyć, że jest to tylko korekcja poziomów poszczególnych podnośnych. Dodatkowym efektem, powodowanym przez dyspersję chromatyczną ścieżki optycznej, występującym w synale zmodulowanym N-QAM jest szum fazowy, który można kompensować także poprzez określenie odpowiedniej podnośnej OFDM w kanale lub bloku, która stanowi wyznacznik synchronizacyjny.

Kolejnym etapem prowadzenia prac badawczych będzie przeprowadzenie symulacji komputerowej, która potwierdzi wyniki badań obliczeniowych. Planowane jest wykonanie tych badań w aplikacji VPIphotonics Design Suite.

LITERATURA

[1] 3GPP TR 38.104 v17.5.0. 2022. NR. Base Station (BS) radio transmission and reception (Release 17).

[2] 3GPP TR 38.801 v14.0.0. 2017. Study on new radio access technology: Radio access architecture and interfaces.

[3] 5GPPP. 2020. Architecture Working Group. "View on 5G Architecture".

[4] 5G-XHaul, D2.3. 2017. "Architecture of Optical/Wireless Backhaul and Fronthaul and Evaluation".

[5] Camps-Mur, D. et al. 2019. "5G-XHaul: A Novel Wireless-Optical SDN Transport Network to Support Joint 5G Backhaul and Fronthaul Services". *IEEE Communications Magazine*, 57: 99-105.

[6] Cooper, A.J. 1990. "Fiber/Radio for the provision of cordless/mobile telephony services in the access network". *Electron. Lett*., 26: 2054-2056.

[7] CPRI Industry Forum (Ericsson, Huawei, NEC, and Nokia). 2015. CPRI Specification 7.0.

[8] CPRI Industry Forum (Ericsson, Huawei, NEC, and Nokia). 2019. eCPRI Specification 2.0.

[9] Devaux, F., Sorel, Y., Kerdiles, J.F. 1993. "Simple measurement of fiber dispersion and of chirp parameter of intensity modulated light emitter". *Journal of Lightwave Technology*. 11(12): 1937-1940.

[10] Elrefaie, A.F. and Lin, C. 1995. "Clipping distortion and chromatic dispersion limitations for 1550 nm video trunking systems". *Proc. IEEE Symp. Computers and Communications*, Alexandria, 328–337.

[11] Elrefaie, A.F., Wagner, R.E., Atlas, D.A., Daut, D.G. 1988. "Chromatic dispersion limitations in coherent lightwave transmission systems". *Journal of Lightwave Technology*. 6(5): 704-709.

[12] Gliese, U., Nielsen, S.N. and Nielsen, T.N. 1996. "Limitations in distance and frequency due to chromatic dispersion in fiber-optic microwave and millimeter-wave links". *IEEE MTT-S Int. Microwave Symp. Dig*., 1547–1550.

[13] IEEE, 1914. 2019. (1914.1 Standard for Packet-Based Fronthaul Transport Networks, 1914.3 Standard for Radio Over Ethernet Encapsulations and Mappings). Next Generation Fronthaul Interface.

[14] Ih, C.S. and Gu, W. 1990. "Fiber Induced Distortions in a Subcarrier Multiplexed Lightwave System". *IEEE J. Select. Areas Commun*., 8: 1296–1303.

[15] ITU-R, F.1332-l. Radio-Frequency Signal Transport through Optical Fibres. 1999.

[16] ITU-T, G.652. 2016. Characteristics of a single-mode optical fibre and cable.

[17] ITU-T, G.655. 2009. Characteristics of a non-zero dispersion-shifted single-mode optical fibre and cable.

[18] ITU-T, G.657. 2016. Characteristics of a bending-loss insensitive single-mode optical fibre and cable.

[19] ITU-T, Series G, Supplement 55. Study Group 15. 2021. "Radio-over-Fibre (RoF) Technologies and Their Applications".

[20] Lee, C.H. 2013. Microwave Photonics, 2'd ed., *CRC Press*, *Taylor & Francis Group*.

[21] Meslener, G.J. 1984. "Chromatic dispersion induced distortion of modulated monochromatic light employing direct detection". *IEEE J. of Quantum Electron*. 20: 1208-1216.

[22] O-RAN Alliance. 2022. O-RAN Open X-haul Transport WG9. WDM-based Fronthaul Transport 2.0. Technical Specification ORAN-WG9.

[23] O-RAN Alliance. 2022. Use Cases and Overall Architecture Workgroup. O-RAN Architecture Description 6.0. Technical Specification ORAN-WG1.

[24] Park, J., Elrefaie, A.F. and Lau, K.Y. 1996. "Fiber chromatic dispersion effects on multichannel digital millimeter-wave transmission". *IEEE Photon. Technol. Lett.*, 8: 1716–1718.

[25] Schmuck, H. 1995. "Comparison of optical millimeter-wave system concepts with regard to chromatic dispersion". *Electron. Lett*., 31, 1848–1849.

[26] Zakrzewski, Z. 2017. "Optical RRH working in an all-optical fronthaul network". *Proceedings of the 8th International Conference on Photonics, Devices and Systems, SPIE*, 10603: 1–12.

[27] Zakrzewski, Z. 2020. "D-RoF and A-RoF Interfaces in an All-Optical Fronthaul of 5G Mobile Systems". *Appl. Sci*. 10 (1212).

[28] Zakrzewski, Z. 2021. "Optical channel selection avoiding DIPP in DSB-RFoF fronthaul interface". *Entropy*. 23 (1554).

WPŁYW NIELINIOWOŚCI NA JAKOŚĆ SYGNAŁU ODBIERANEGO W SYSTEMIE M-MIMO OFDM

Marcin Wachowiak[1], Paweł Kryszkiewicz[2]

[1]Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, marcin.r.wachowiak@student.put.poznan.pl
[2]Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, pawel.kryszkiewicz@put.poznan.pl

DOI: 10.15199/59.2022.4.85

Streszczenie: **Połączenie techniki wieloantennowej M-MIMO z techniką multipleksacji OFDM może powodować, że nadawany sygnał będzie podlegał silnemu zniekształceniu przez nieliniowości stopni wyjściowych. W pracy zbadano wpływ zniekształcenia nieliniowego na efektywną charakterystykę promieniowania i jakość sygnału odbieranego w systemie. Otrzymane wyniki pozwalają zaobserwować podobną efektywną charakterystykę kierunkową zarówno dla sygnału pożądanego jak i zniekształceń w przypadkach bezpośredniej widzialności jak i kanału dwuścieżkowego.**[1]

Abstract: **Combining M-MIMO technology with the OFDM technique may result in significant distortion of the transmitted signals by the nonlinear front-end amplifiers. In this work, the impact of the frontend nonlinearity on the radiation pattern and the received signal quality is studied. Obtained results show that the radiation pattern is similar both for the desired and distortion signal components in cases of direct visibility and two-path propagation.**

Słowa kluczowe: **Modulacja wielotonowa OFDM, nieliniowość stopnia wyjściowego, system wieloantenowy M-MIMO**
Keywords: **Orthogonal Frequency Division Multiplexing, front-end nonlinearity, Massive Multiple Input Multiple Output**

1. WSTĘP

Rozwój technologiczny oraz cyfryzacja wielu stref życia przekłada się na wzrost wymaganej przepływności systemów komunikacji bezprzewodowej. Liczba użytkowników sieci komórkowych oraz wymagania jakości obsługi (ang. Quality of Service - QoS) systematycznie rosną powodując wzmożone obciążenie infrastruktury. Co roku ruch generowany przez terminale mobilne zwiększa się w przybliżeniu o 50% [3]. W celu pokrycia rosnącego zapotrzebowania wdrażane są nowe generacje sieci komórkowych, zawierające nowe techniki zwiększające osiąganą efektywność widmową oraz przepływność. Kluczowym rozwiązaniem umożliwiającym dalsze zwiększenie tych parametrów jest technika wieloantenowa M-MIMO (ang. Massive Multiple-Input Multiple-Output).

Prace dotyczące systemów M-MIMO wykazały, że ich pojemność rośnie wraz ze zwiększaniem liczby anten oraz nie jest odgórnie ograniczona [2]. Oznacza to, że dzięki zastosowaniu odpowiedniego prekodowania, gdy liczba anten dąży do nieskończoności, również pojemność systemu będzie dążyła do nieskończenie dużej wartości. Należy mieć na uwadze, że te rozważania nie uwzględniają wielu niekorzystnych efektów, które są obecne w rzeczywistych systemach transmisji bezprzewodowej. Jedną z istotnych niedoskonałości łańcucha nadawczego jest nieliniowość wzmacniaczy mocy, która przekłada się na zniekształcenie sygnału nadawanego. Liniowość wzmacniacza ma szczególne znaczenie w przypadku modulacji OFDM, gdzie występują wysokie wahania amplitudy obwiedni zespolonej [4]. W przypadku systemu M-MIMO OFDM każda z anten jest podłączona do osobnego modulatora oraz wzmacniacza o potencjalnie nieliniowej charakterystyce. Wyznaczenie sumarycznej charakterystyki zniekształceń nie jest jednak prostym zagadnieniem.

Pierwotne analizy [1] zakładały, że wynikowe zniekształcenie sygnału przez tor nadawczy może być modelowane jako addytywny szum biały, nieskorelowany między antenami. Założenie o braku korelacji skutkuje dookólną charakterystyką promieniowania zniekształcenia. W pracy [6] pokazano przypadek dwóch podnośnych nadawanych w systemie wieloantenowym przez nieliniowość opisaną wielomianem trzeciego stopnia, który podważa słuszność wcześniejszych założeń. W tym przypadku sygnał zniekształceń nieliniowych jest skorelowany między antenami, a charakterystyka promieniowania produktów intermodulacji pokrywa się z charakterystyką promieniowania sygnału pożądanego.

W kolejnych pracach wpływ nieliniowości wzmacniacza został dokładniej przeanalizowany w kontekście wpływu na emisje pozapasmowe i ich charakterystykę promieniowania. W [8] autorzy dokonali dokładnej analizy charakterystyki promieniowania mocy zniekształcenia dla transmisji na jednej podnośnej oraz modelu nieliniowości jako wielomianu z pamięcią. Następnie analiza została rozszerzona dla przypadku OFDM dla kanału LOS (ang. Line of Sight) oraz pojedynczego i kilku użytkowników. Dodatkowo wykazano, że dla dużej liczby użytkowników lub dla dużej liczby ścieżek propagacji oraz takiej samej mocy nadawczej dla każdego odbiorcy, charakterystyka promieniowania zniekształcenia jest zbliżona do dookólnej.

W niniejszej pracy przeprowadzono analizę wpływu nieliniowości na system Massive MIMO z oceną charakterystyki kierunkowej sygnału użytecznego i zniekształceń, a także stosunku mocy sygnału do mocy zniekształcenia w odbiorniku. Rozważono w pełni cyfrową implementację nadajnika tzn. o liczbie łańcuchów cyfrowego przetwarzania sygnału równiej liczbie anten nadawczych. Jako model nieliniowego wzmacniacza zastosowano ogranicznik

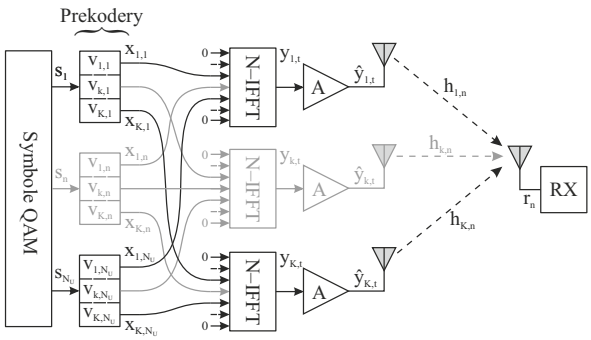
[1]Praca została sfinansowana przez Narodowe Centrum Nauki w ramach projektu OPUS nr 2021/41/B/ST7/00136.

mocy, który został pominięty w powyżej cytowanych pracach. Jest on jednak bardzo istotny, ponieważ może charakteryzować połączenie układu cyfrowego zniekształcenia wstępnego (ang. Digital Predistortion) i nieliniowości wzmacniacza. Pokazano, że jest to charakterystyka optymalna dla modulacji OFDM [9]. Wpływ zniekształcenia został zbadany dla różnych modeli kanałów: bezpośredniej widzialności LOS, modelu dwuścieżkowego oraz nieskorelowanego Rayleigha. Dotychczas kanał dwuścieżkowy nie był analizowany w kontekście zniekształceń w systemie M-MIMO, a może być często spotykanym w przypadku bezpośredniej widoczności nadajnika i odbiornika. Zastosowano prekodowanie MRT (ang. Maximum Ratio Transmission) dla jednego użytkownika.

Struktura artykułu przedstawia się następująco: w sekcji 2 przedstawiono szczegółowy opis rozważanego systemu, w sekcji 3 zdefiniowano parametry modelu symulacyjnego oraz omówiono otrzymane wyniki, natomiast sekcja 4 podsumowuje całą pracę.

2. MODEL SYSTEMU

Rozważmy system przedstawiony na Rys. 1. Nadawane symbole QAM podlegają prekodowaniu, a następnie nadaniu przez K równoległych łańcuchów nadawczych składających się z modulatora OFDM o maksymalnie N podnośnych, nieliniowego wzmacniacza i elementu antenowego. Sygnały dodają się w jednoantenowym odbiorniku. W następnych podsekcjach opisane zostanie przetwarzanie sygnału w każdym z wymienionych elementów modelu.



Rys. 1. Model systemu.

2.1. Kanał radiowy

Z uwagi na użycie modulatora OFDM można założyć, że kanał radiowy jest płaski w pasmie pojedynczej podnośnej. Dla n-tej podnośnej i k-tej anteny jest oznaczony jako pojedyncza liczba zespolona $h_{k,n}$. Rozważymy 3 rodzaje kanałów:

- LOS (ang. Line of Sight): modelowany przez tłumienie wolnej przestrzeni i obrót fazy sygnału wynikający z odległości od elementu antenowego do odbiornika;
- dwuścieżkowy: prócz ścieżki bezpośredniej widzialności uwzględnia również dodatkową ścieżkę odbitą od podłoża ze współczynnikiem odbicia -1, wyznaczoną na podstawie położenia anten;
- Rayleigha: modelowany jako niezależne zmienne zespolone gaussowskie dla każdej anteny, podnośnej i azymutu.

2.2. Prekodowanie

Kierowanie wiązki odbywa się za pomocą prekodera MRT, który maksymalizuje moc w punkcie odbioru. Jednocześnie normalizowana jest średnia moc emitowana ze wszystkich elementów antenowych na danej podnośnej do jedynki [7]. Współczynniki prekodowania dla k-tej anteny na n-tej podnośnej obliczamy jako:

$$v_{k,n} = \frac{h^*_{k,n}}{\sqrt{\sum_{k=1}^{K} |h_{k,n}|^2}}, \tag{1}$$

gdzie $*$ oznacza sprzężenie zespolone. Prekodowanie odbywa się poprzez przemnożenie nadawanego na n-tej podnośnej symbolu QAM tzn. s_n przez współczynnik prekodowania dla n-tej podnośnej i k-tej anteny otrzymując symbol $x_{k,n}$:

$$x_{k,n} = s_n v_{k,n}. \tag{2}$$

2.3. Modulacja OFDM

Prekodowane symbole są następnie poddawane modulacji OFDM, poprzez zastosowanie odwrotnej transformacji Fouriera IFFT (ang. Inverse Fast Fourier Transform) rozmiaru N. Niezerowe są jedynie N_u symbole o indeksach $\mathcal{N} = \{-N_\mathrm{u}/2, ..., -1, 1, ..., N_\mathrm{u}/2\}$. Pozostałe $N - N_\mathrm{u}$ podnośne, tzn. składowa stała oraz podnośne blisko krawędzi pasma, są modulowane zerami jak w rzeczywistych systemach OFDM. Na wyjściu IFFT t-ta próbka symbolu OFDM dla k-tej anteny jest obliczana jako:

$$y_{k,t} = \frac{1}{\sqrt{N}} \sum_{n \in \mathcal{N}} x_{k,n} e^{j2\pi \frac{n}{N} t} \tag{3}$$

gdzie $t \in \{-N_\mathrm{CP}, ..., N-1\}$, a N_CP to liczba próbek przedrostka cyklicznego.

2.4. Nieliniowy stopień wyjściowy

Zmodulowany sygnał jest poddawany przetworzeniu przez model nieliniowości identyczny dla każdego łańcucha nadawczego:

$$\hat{y}_{k,t} = \mathcal{A}(y_{k,t}), \tag{4}$$

który w przypadku zastosowanego ogranicznika amplitudy (ang. soft-limiter) można opisać przez

$$\hat{y}_{k,t} = \begin{cases} y_{k,t} & \text{dla } |y_{k,t}|^2 \leqslant P_\mathrm{max} \\ \sqrt{P_\mathrm{max}} e^{j \arg y_{k,t}} & \text{dla } |y_{k,t}|^2 > P_\mathrm{max} \end{cases}, \tag{5}$$

gdzie P_max oznacza maksymalną moc nadawczą powyżej której sygnał jest ograniczany zachowując fazę sygnału wejściowego.

Często do opisu systemu, zamiast podawania wartości $P_{\max}$, stosowany jest parametr IBO (ang. Input Backoff) definiowany jako stosunek mocy maksymalnej do mocy średniej na wejściu wzmacniacza, wyrażony w decybelach:

$$IBO\ [dB] = 10log_{10}\left(\frac{P_{max}}{\mathbb{E}[|y_{k,t}|^2]}\right). \tag{6}$$

Zakładamy tu, że moc średnia liczona jest zarówno po poszczególnych próbkach symbolu OFDM jak i po poszczególnych antenach tzn.

$$\mathbb{E}[|y_{k,t}|^2] = \frac{\bar{E}_s}{NK}\sum_{n=1}^{N_u}\sum_{k=1}^{K}|v_{k,n}|^2, \tag{7}$$

gdzie $\bar{E}_s$ jest mocą średnią symbolu s_n. Przy tej definicji wszystkie K wzmacniacze pracują z identycznym poziomem odcięcia.

Można dokonać dekompozycji sygnału na wyjściu wzmacniacza zgodnie z zasadami regresji liniowej jako

$$\hat{y}_{k,t} = \alpha_k y_{k,t} + d_{k,t} \tag{8}$$

gdzie α_k oznacza współczynnik wzmocnienia, a $d_{k,t}$ sygnał zniekształceń nieskorelowany z symbolem danych $y_{k,t}$. Parametr α_k jest zdefiniowany na podstawie wzoru:

$$\alpha_k = \frac{\mathbb{E}\left[\hat{y}_{k,t}y^*_{k,t}\right]}{\mathbb{E}\left[y_{k,t}y^*_{k,t}\right]}, \tag{9}$$

gdzie $\mathbb{E}$ oznacza wartość oczekiwaną dla danej anteny k. Wartość oczekiwana jest estymowana poprzez uśrednienie po po chwilach czasu t. W przypadku gaussowskiego charakteru próbek $y_{k,t}$ o tej samej mocy dla każdej anteny wartość ta jest stała dla wszystkich anten i wynosi [10]:

$$\alpha_k = 1 - e^{-\gamma^2} + \frac{\sqrt{\pi\gamma}}{2}\,\mathrm{erfc}\,(\gamma). \tag{10}$$

gdzie $\gamma = 10^{\frac{IBO}{20}}$ może być interpretowana jako stosunek maksymalnej wartości obwiedni zespolonej sygnału do wartości średniej w skali liniowej.

2.5. Odbiór sygnału

Transmitowane sygnały z poszczególnych anten w dziedzinie czasu są poddawane splotowi z odpowiadającymi im kanałami szerokopasmowymi oraz sumowane w odbiorniku. W odbiorniku OFDM stosowana jest szybka transformacja Fouriera (ang. FFT), która pozwala na przedstawienie sygnału odbieranego na n-tej podnośnej jako:

$$r_n = \sum_{k=1}^{K}\mathcal{F}_{[n,t=0,\ldots,N-1]}\{\hat{y}_{k,t}\}h_{k,n} + w_n, \tag{11}$$

gdzie w_n to próbka szumu białego na n-tej podnośnej w odbiorniku, a $\mathcal{F}_{[n,t=0,\ldots,N-1]}\{\cdot\}$ oznacza transformację Fouriera dla $t = 0, ..., N-1$ na n-tej podnośnej.

Korzystając z (8) oraz (3) sygnał odebrany można rozwinąć do:

$$r_n = \sum_{k=1}^{K}\alpha_k h_{k,n}x_{k,n} + \sum_{k=1}^{K}h_{k,n}\tilde{d}_{k,n} + w_n, \tag{12}$$

gdzie

$$\tilde{d}_{k,n} = \mathcal{F}_{[n,t=0,\ldots,N-1]}\{d_{k,t}\}. \tag{13}$$

Uwzględniając współczynniki prekodowania symboli z (1) otrzymujemy:

$$r_n = \sum_{k=1}^{K}\alpha_k\frac{h_{k,n}h^*_{k,n}}{\sqrt{\sum_{\tilde{k}=1}^{K}\left|h_{\tilde{k},n}\right|^2}}s_n + \sum_{k=1}^{K}h_{k,n}\tilde{d}_{k,n} + w_n. \tag{14}$$

Widoczne jest, że niezależnie od charakterystyki nieliniowości obecny będzie zysk prekodowania tzn. $h_{k,n}h^*_{k,n}$. Będzie on jednak pomniejszony proporcjonalnie do wartości α_k. Dodatkowo, choć w wyniku dekompozycji (8) sygnał nieliniowości $d_{k,t}$ jest nieskorelowany z sygnałem pożądanym $y_{k,t}$ nie można prosto wyciągnąć wniosku co do charakterystyki kierunkowej składnika $\sum_{k=1}^{K}h_{k,n}\tilde{d}_{k,n}$ tzn. czy poszczególne składniki sumy dodają się w fazie dla pewnych kierunków. W tym celu zostaną przeprowadzone symulacje komputerowe, których wyniki zostaną przenalizowane w kolejnym rozdziale.

3. WYNIKI SYMULACJI

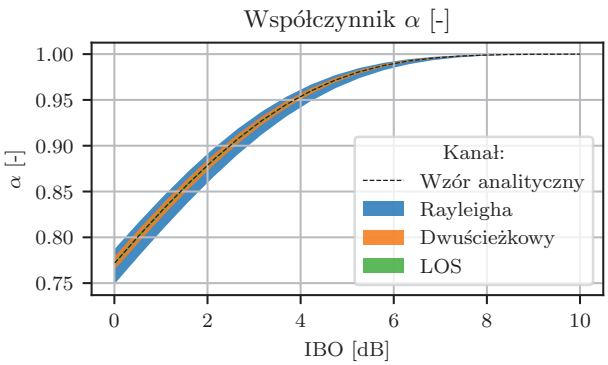
W celu analizy propagacji sygnału pożądanego i zniekształceń przeprowadzona została symulacja systemu M-MIMO z liniową macierzą antenową składającą się z K elementów odległych od sąsiednich o połowę długości fali (przy częstotliwości nośnej $f_c = 3.5GHz$). Przyjęto, że nadajnik znajduje się na wysokości 15 m, natomiast odbiornik jest umieszczony w odległości 300 m na wysokości 1.5 m. Transmisja była prekodowana dla kąta azymutu $45°$, dla którego macierz kanału była w pełni znana. W każdym przypadku przesyłano 10 symboli OFDM, co zapewniało wystarczającą zbieżność wyników. Parametry symulacji zostały przedstawione w Tab. 1.

Tab. 1. Parametry symulacji

Parametr	*Symbol*	*Wartość*	*Jednostka*
Szerokość pasma podnośnej	Δf	15	[kHz]
Częstotliwość nośna	f_c	$3,5$	[GHz]
Liczba podnośnych	N	1024	$[-]$
Liczba symboli konstelacji QAM	m	64	$[-]$
Rozdzielczość wyznaczania charakterystyki promieniowania	$\angle_{res}$	1	$[°]$
Liczba anten	K	$16-128$	$[-]$

Na pierwszym etapie zostało zbadane czy współczynnik α_k jest niezależny od anteny nadawczej tzn. czy $\forall_{k,k'}\alpha_k = \alpha_{k'}$. Porównano wartości wyznaczone analitycznie z wynikami estymacji dla różnych wartości IBO

oraz wybranych kanałów propagacyjnych. Rys. 2 przedstawia zakres wartości α_k obliczonych niezależnie dla każdej anteny. Widoczne jest, że dla kanałów bezpośredniej widzialności zakresy wartości parametru α pokrywają się z wartością wyznaczoną na podstawie wzoru, a w przypadku kanału Rayleigha rozrzut, choć niezerowy, jest niewielki i bliski wartości analitycznej. Pozwala to z dobrym przybliżeniem wykorzystać stałą, obliczoną analitycznie wartość α. Sugeruje też, że sygnał po prekodowaniu ma zbliżony do gaussowskiego rozkład.



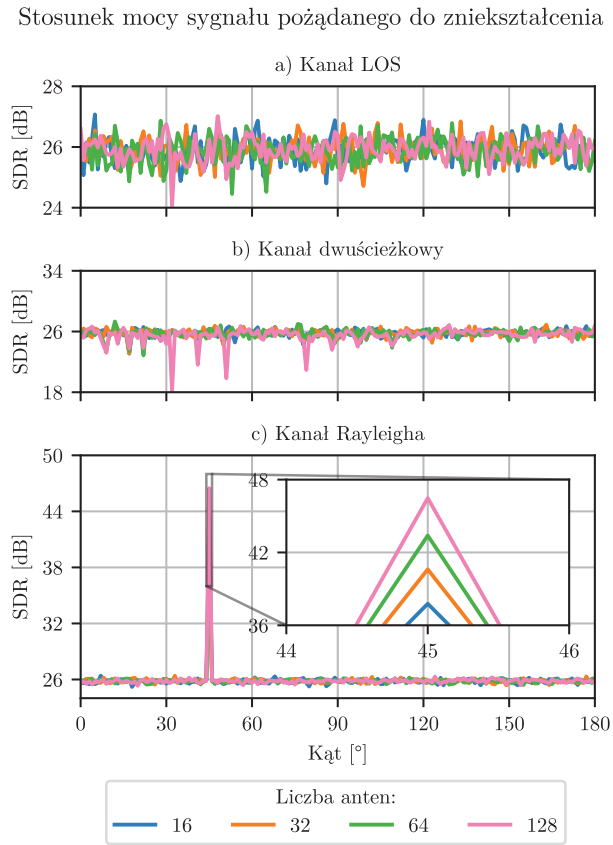
Rys. 2. Zakresy wartości parametru α w zależności od IBO dla 64 anten i wybranych kanałów propagacyjnych.

Kolejnym ważnym problemem jest wpływ prekodowania w systemie wieloantenowym na stosunek mocy sygnału do mocy zniekształceń SDR (ang. Signal to Distortion Ratio) w odbiorniku. Został on wyznaczony jako iloraz sumarycznej mocy sygnału pożądanego do sygnału zniekształconego przypadającego na pasmo używanych podnośnych $\mathcal{N}$. Wyniki przedstawiono na Rys. 3 dla wszystkich trzech rozważanych kanałów radiowych.

Wyniki dla kanału LOS pokrywają się z wcześniejszymi analizami, dyskutowanymi we wstępie, według których sygnał pożądany oraz zniekształcony są kierowane w taki sam sposób, co skutkuje stałym stosunkiem sygnału pożądanego do zniekształcenia niezależnie od liczby anten i azymutu używanego w prekodowaniu. Nieznaczne wahania wartości SDR w zależności od azymutu czy liczby anten wynikają z ograniczonej długości trwania i dokładności symulacji.

Dla kanału dwuścieżkowego kąty kierowania wiązki oraz stosunek sygnału pożądanego do zniekształcenia przedstawiają się bardzo podobnie jak w przypadku kanału LOS. Mimo dodania dodatkowej ścieżki reprezentującej odbicie, stosunek sygnału pożądanego do zniekształcenia jest stały niezależnie od kąta oraz od liczby anten.

W przypadku kanału Rayleigha widać znaczący wzrost wartości SDR wynikający z zastosowania prekodowania dla azymutu 45°. Stosunek sygnału pożądanego do zniekształcenia rośnie proporcjonalnie do logarytmu liczby anten tzn. $10\log_{10} K$. Dla innych kątów, dla których kanał jest nieskorelowany, nie widać takiego zysku. W przypadku kanału Rayleigha, szeroko dyskutowanego w wielu publikacjach, np. [1], można rzeczywiście domniemywać, że efekty nieliniowe będą mogły być pomijalne przy zastosowaniu wystarczająco dużej liczby anten nadawczych. Trzeba brać jednak pod uwagę, że rzeczywiste kanały radiowe zazwyczaj odbiegają od tego wyidealizowanego modelu [5].

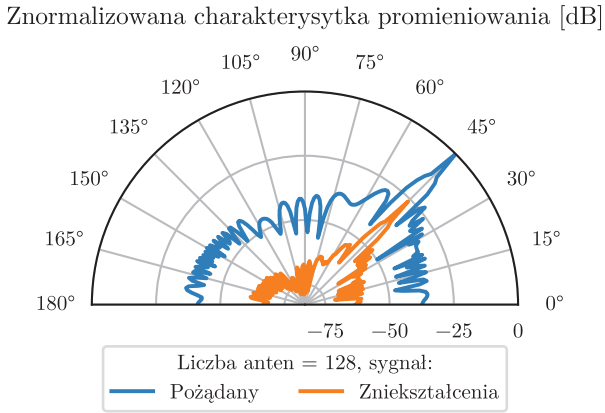


Rys. 3. Stosunek mocy sygnału pożądanego do zniekształcenia w zależności od kąta i liczby anten dla wybranych kanałów propagacyjnych i IBO=5 dB.

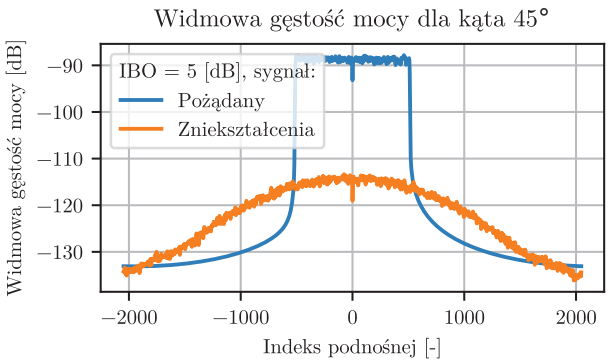
Praktycznym, a wcześniej nierozważanym kanałem radiowym w kontekście wpływu nieliniowości na sygnał M-MIMO jest kanał dwuścieżkowy. Stąd kolejne wyniki skupią się na tym przypadku. Na Rys. 4 pokazano znormalizowaną charakterystykę kierunkową dla sygnału pożądanego i sygnału zniekształceń. Charakterystyka jest silnie kierunkowa z uwagi na zastosowanie aż 128 anten. Podobną kierunkowość przejawia jednak również sygnał zniekształcenia. Porównując wartości charakterystyki dla sygnału pożądanego oraz zniekształcenia można odczytać wartość SDR, która zgodnie z wcześniejszymi analizami wynosi około 26 dB dla IBO=5 dB. Warto podkreślić, że dla rozważanej liniowej macierzy antenowej występuje symetria charakterystyki antenowej, pozwalająca nie pokazywać charakterystyki dla azymutów 180-360°.

Na Rys. 5 pokazano widmowe gęstości mocy sygnału pożądanego i zniekształcenia dla K=128 anten i IBO równego 5 dB. Oba sygnały zachowują charakterystykę podobną do tej znanej dla nadawania sygnału OFDM w systemie jednoantenowym. Widoczne jest, że zniekształ-

cenie nieliniowe ma szersze pasmo niż sygnał pożądany powodując zarówno zakłócenia w pasmie podnośnych użytecznych i poza ich pasmem.



Rys. 4. Charakterystyka promieniowania sygnału pożądanego oraz zniekształcenia w zależności od kąta dla 128 anten, IBO=5 dB oraz kanału dwuścieżkowego.



Rys. 5. Widmowa gęstość mocy sygnału odbieranego dla kąta prekodowania dla 128 anten i kanału dwuścieżkowego.

4. WNIOSKI

W artykule został przeanalizowany wpływ zniekształcenia nieliniowego w postaci ogranicznika amplitudy na sygnały w systemie M-MIMO. Na podstawie przeprowadzonych analiz można zauważyć, że zarówno w przypadku bezpośredniej widoczności jak i propagacji dwuścieżkowej zysk prekodowania nie wpływa na poprawę wartości SDR w odbiorniku. Zyski prekodowania wieloantenowego uwidaczniają się dla kanału Rayleigha, gdzie można zaobserwować wzrost SDR wraz ze zwiększającą się liczbą anten. Można stwierdzić, że systemy M-MIMO pracujące w paśmie fal milimetrowych oraz środowisku propagacji z małą liczbą ścieżek mogą być ograniczone przez nieliniowości stopni końcowych. Konieczne może być zastosowanie technik ograniczających wpływ nieliniowości na jakość odbioru w systemach wieloantenowych.

LITERATURA

[1] Björnson, Emil, Hoydis, Jakob, Kountouris, Marios, Debbah, Mérouane, 2014, “Massive mimo systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits”. *IEEE Transactions on Information Theory*, 60 (11): 7112–7139.

[2] Björnson, Emil, Hoydis, Jakob, Sanguinetti, Luca, 2018, “Massive MIMO has unlimited capacity”. *IEEE Transactions on Wireless Communications*, 17 (1): 574–590.

[3] Ericsson, 2021, “Ericsson mobility report”. Technical report, Ericsson, dostępny pod adresem: https://www.ericsson.com/4ad7e9/assets/local/reports-papers/mobility-report/documents/2021/ericsson-mobility-report-november-2021.pdf.

[4] Gharaibeh, Khaled M., 2011, *Nonlinear Distortion in Wireless Systems: Modeling and Simulation with MATLAB*. John Wiley & Sons, New York, ISBN 9781119961734.

[5] Gunnarsson, Sara, Flordelis, Jose, Van der Perre, Liesbet, Tufvesson, Fredrik, 2018, “Channel hardening in massive mimo-a measurement based analysis”. *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 1–5.

[6] Larsson, Erik G., Van Der Perre, Liesbet, 2018, “Out-of-band radiation from antenna arrays clarified”. *IEEE Wireless Communications Letters*, 7 (4): 610–613.

[7] Marzetta, Thomas L., Larsson, Erik G., Yang, Hong, Ngo, Hien Quoc, 2016, *Fundamentals of Massive MIMO*. Cambridge University Press.

[8] Mollén, Christopher, Gustavsson, Ulf, Eriksson, Thomas, Larsson, Erik G., 2018, “Spatial characteristics of distortion radiated from antenna arrays with transceiver nonlinearities”. *IEEE Transactions on Wireless Communications*, 17 (10): 6663–6679.

[9] Raich, R., Qian, Hua, Zhou, G.T., 2005, “Optimization of sndr for amplitude-limited nonlinearities”. *IEEE Transactions on Communications*, 53 (11): 1964–1972.

[10] Rowe, H. E., 1982, “Memoryless nonlinearities with gaussian inputs: Elementary results”. *The Bell System Technical Journal*, 61 (7): 1519–1525.

INTELIGENTNA PREDYKCJA STANU ZAJĘTOŚCI WIDMA Z UWZGLĘDNIENIEM WPŁYWU ZANIKÓW

INTELLIGENT PREDICTION OF SPECTRUM OCCUPANCY THAT INCLUDES FADING EFFECT INFLUENCE

Małgorzata Wasilewska[1]; Hanna Bogucka[2]

[1] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, malgorzata.wasilewska@put.poznan.pl
[2] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, hanna.bogucka@put.poznan.pl

DOI: 10.15199/59.2022.4.86

Streszczenie: **W pracy przedstawiono algorytm predykcji wolnych zasobów w sieciach radiowych 5G. Sygnał 5G nadawany przez użytkownika pierwotnego (PU) poddawany jest zanikom występującym w kanale, co uniemożliwia poprawną detekcję i tym samym właściwą ochronę transmisji PU. Zaproponowany algorytm wykorzystuje możliwości głębokiego uczenia maszynowego w celu rozpoznania zależności czasowo-częstotliwościowych występujących w odebranym sygnale, a także rozpoznania stopnia zaniku. Znając te informacje, algorytm dokonuje lepszej detekcji wolnych zasobów, przy jednoczesnej ochronie transmisji PU.**

Abstract: **In this paper, we present a 5G spectrum resources prediction algorithm. 5G signal, transmitted by the primary user (PU) is transmitted through fading channel, which makes negatively affects prediction performance and proper protection of PU's transmission. The proposed algorithm applies deep learning for estimating fading level and recognizing time-frequency patterns in a received signal. Having this information, the algorithm can perform better signal prediction and PU's transmission protection.**

Słowa kluczowe: **CNN, ML, detekcja sygnałów, 5G**

Keywords: **CNN, ML, spectrum sensing, 5G**

1. WSTĘP

Wraz z szybko zwiększającą się liczbą urządzeń bezprzewodowych, konieczne staje się opracowanie niezawodnych metod umożliwiających efektywne wykorzystanie zasobów radiowych. Sposobem na poprawę stopnia wykorzystania widma jest zastosowanie inteligentnych metod pozyskiwania świadomości o warunkach transmisyjnych w środowisku radiowym. Jedną z takich metod jest detekcja zajętych zasobów radiowych (ang. *spectrum sensing* – SS) [6]. Wolne zasoby mogą być wykorzystane do transmisji, pod warunkiem, że użytkownicy licencjonowani (ang. *Primary Users* – PU) mają zapewnione pierwszeństwo w wykorzystaniu tych zasobów i ich transmisja będzie chroniona, tzn. niezakłócana.

Uczenie maszynowe (ang. machine learning – ML) okazało się bardzo pożytecznym narzędziem w SS [4]. Główną przewagą ML nad tradycyjnymi metodami detekcji sygnałów jest umiejętność rozpoznawania złożonych zależności występujących w odebranych sygnałach [1]. Sygnały transmitowane bezprzewodowo poddawane są czasowym zmianom intensywności ruchu telekomunikacyjnego. Ruch telekomunikacyjny zależy także w dużej mierze od lokalizacji nadajnika. Dodatkowo, odbierane sygnały poddawane są efektom propagacji radiowej, takim jak zaniki, czy shadowing. W efekcie, detekcja sygnału może okazać się w znacznym stopniu utrudniona. Należy zatem uwzględnić występowanie tych efektów w procesie SS. Zaniki częstotliwościowe charakteryzują się pewną powtarzalnością, zarówno w czasie, jak i w częstotliwości. Algorytm ML powinien być w stanie rozpoznać te zależności i zastosować zdobytą wiedzę w celu poprawy jakości SS.

Znając zależności i wzorce występujące w sygnale możliwa jest nie tylko detekcja, ale także predykcja sygnału, a tym samym predykcja stanu zajętości widma w przyszłości. W poprzedniej pracy [5], skupiliśmy się na predykcji zajętości zasobów (ang. *spectrum prediction* – SP), przy założeniu, że kanał jest stały w czasie. Tym razem użytkownik dokonujący detekcji może się poruszać względem PU, co wiąże się ze zmieniającym się wpływem kanału w czasie i częstotliwości.

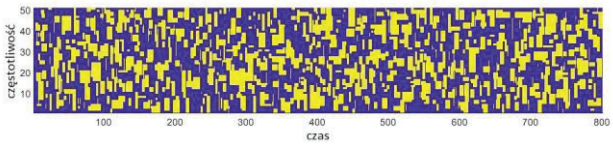
W zaproponowanym algorytmie, wykorzystywana jest splotowa sieć neuronowa (ang. *convolutional neural netork* – CNN). Typowym zastosowaniem CNN jest przetwarzanie i analiza obrazów. W naszym zastosowaniu CNN użyta jest do analizy spektrogramów utworzonych z wartości energii odebranych sygnałów. Podobne podejście zaprezentowano w [2], gdzie CNN również wykorzystano do predykcji typu oraz liczby transmitujących użytkowników, jednak dane użyte na wejściu sieci składały się z jednowymiarowych sekwencji danych w czasie. Długoterminowa predykcja przeprowadzona na danych czasowo-częstotliwościowo-przestrzennych zaprezentowana została w [3]. W tym celu użyto hybrydowego algorytmu łączącego CNN z siecią rekurencyjną.

W naszej pracy CNN używane jest nie tylko do predykcji zajętości widma, ale także do estymacji stopnia zaniku. Silne zaniki uniemożliwiają poprawną detekcję, przez co użytkownik przeprowadzający SS może zdecydować się na transmisję, co w konsekwencji prowadzi do zwiększenia poziomu interferencji i zakłóceń sygnału nadawanego przez PU. W ramach naszego eksperymentu założono, że ochrona transmisji PU jest priorytetowa, a równocześnie maksymalizowana jest efektywność prawidłowej detekcji i predykcji wolnych zasobów.

Reszta artykułu podzielona jest na następujące części. W części drugiej, opisano model systemu i zdefiniowano rozpatrywany problem badawczy. W części trzeciej, opisano zastosowany algorytm wykorzystujący sieci CNN. W części czwartej, przedyskutowano założenia symulacji. W części piątej przedstawiono otrzymane wyniki. Na koniec, w części szóstej, zawarto wnioski i podsumowanie.

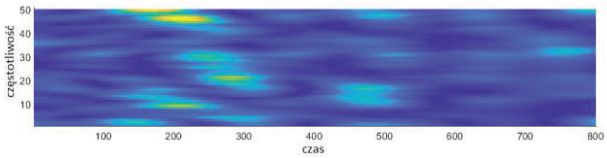
2. MODEL SYSTEMU

W ramach eksperymentu, rozpatrywany jest pewien obszar przestrzenny, w którym użytkownik wtórny (ang. *secondary user* – SU) chce dokonać detekcji i predykcji w czasie i częstotliwości sygnału 5G nadawanego w transmisji w dół przez stację bazową (PU). W momencie, gdy znalezione zostaną wolne zasoby SU może rozpocząć swoją transmisję. Jeżeli dane zasoby czasowo częstotliwościowe zostają zajęte, SU zobowiązany jest do wstrzymania transmisji. Przykład zajętych przez PU i wolnych zasobów prezentuje Rys. 1, gdzie obszary zaznaczone na żółto oznaczają zajęte zasoby, a obszary niebieskie zasoby wolne.

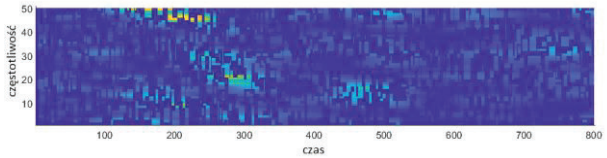


Rys. 1. Przykład przydziału zasobów RB w czasie i częstotliwości (kolor żółty oznacza zajęte RB, kolor niebieski wolne RB)

Detekcja wolnych RB jest trudniejsza, jeżeli transmitowany sygnał przesyłany w kanale wielodrogowym, poddany jest zjawisku zaników i efektowi Dopplera. Wpływ kanału na każdy RB przedstawiono na Rys. 2. Ciemniejsze obszary oznaczają zasobu, na które przypadło większe tłumienie. Sygnał o takiej samej zajętości zasobów, jak przedstawiono na Rys. 1. poddany działaniu kanały z Rys. 2 przedstawiono na Rys. 3.



Rys. 2. Wizualizacja wpływu kanału z zanikami na sygnał, przy założeniu, że wszystkie zasoby RB są zajęte

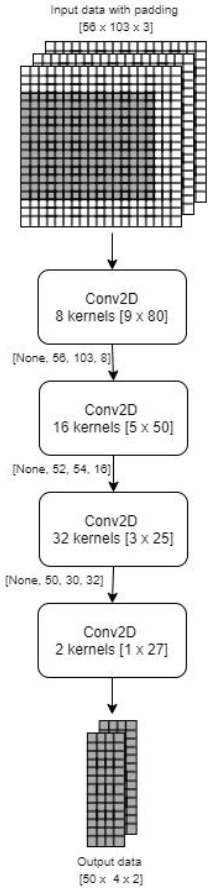


Rys. 3. Sygnał poddany działaniu kanału z zanikami (wartości energii obliczone na każdy RB)

W rozpatrywanym scenariuszu, SU porusza się w obszarze, w którym średni stosunek sygnału do szumu (ang. *signal-to-noise ratio* – SNR) jest stały. Chwilowa wartość SNR dla danego RB zależy od aktualnego wpływu kanału. Dodatkowo wartość SNR zależy od efektu shadowingu.

3. CNN DLA DETEKJI, PREDYKCJI I ESTYMACJI KANAŁU

Do detekcji, predykcji i estymacji kanału wykorzystano algorytm uczenia maszynowego, a konkretnie algorytm CNN. Detekcja dotyczy estymacji zajętości zasobów RB w aktualnej chwili na wszystkich częstotliwościach z rozpatrywanego pasma. Predykcja przyszłego stanu pasma, polega na przewidywaniu zajętości RB w kolejnych chwilach czasowych. Przewidywanie jest korzystne dla SU przede wszystkim dlatego, że umożliwia mu przygotowanie się do transmisji i zaniechanie dalszych obliczeń wykonywanych w ramach detekcji.



Rys. 4. Model sieci CNN. W zaproponowanym algorytmie używane są dwie takie sieci.

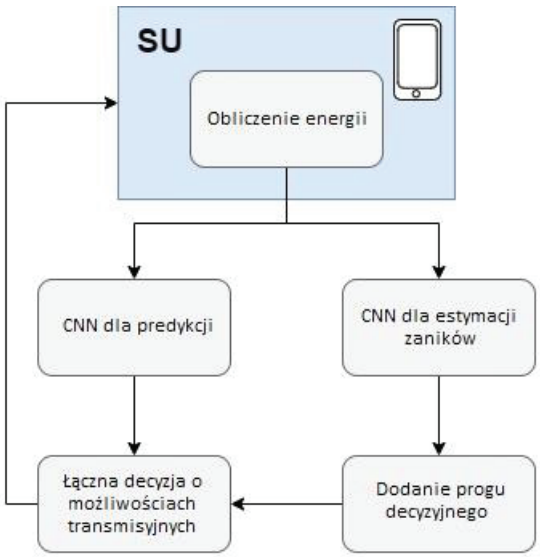
CNN użyto również do estymacji poziomu zaników przypadających na poszczególne RB. Informacja o stopniu zaników staje się sposobem oceny wiarygodności działania algorytmu detekcji/predykcji. Znając wpływ zaników na odbierany sygnał, możliwe jest zapewnienie lepszej ochrony transmisji PU. W ramach ochrony PU, SU musi unikać wykorzystywania zasobów, w których prawidłowa detekcja jest niemożliwa do przeprowadzenia.

W zaproponowanym algorytmie używane są dwa modele CNN. Jeden oblicza najprawdopodobniejszy, teraźniejszy i przyszły stan zajętości zasobów, drugi estymuje wpływ kanału dla tych samych zasobów. Oba modele mają taką samą architekturę oraz akceptują na

wejściu ten sam typ danych. Danymi wejściowymi są spektrogramy, na które składają się wartości energii przypadające na poszczególne RB. Model CNN przedstawiono na Rys. 4. Wynikiem działania CNN przystosowanego do detekcji i predykcji (CNN_1) jest zestaw decyzji w całym paśmie o zajętości aktualnych RB oraz trzech przyszłych RB z trzech kolejnych chwil czasowych. Drugi model (CNN_2) ocenia stopień zaniku dla tych samych zasobów RB.

Kiedy poziom zaników jest już oszacowany, ważne jest podjęcie decyzji o wartości progowej poziomu zaników, która oznaczałaby konieczność ochrony potencjalnej transmisji PU. Wybierana jest taka wartość progowa, która zapewnia największe prawdopodobieństwo ochrony transmisji PU i równocześnie najmniejsze prawdopodobieństwo utraty okazji do transmisji przez SU. Wartość progowa wybierana jest na etapie uczenia sieci CNN i podczas fazy testowej pozostaje stała.

Poprzez połączenie wyników CNN_1 i CNN_2 otrzymywana jest ostateczna decyzja o zajętości zasobów, która umożliwia SU podjęcie decyzji o transmisji. Schemat całego algorytmu przedstawiono na Rys.5.



Rys. 5. Zaproponowany algorytm łączący decyzje dwóch sieci CNN w decyzję końcową o zajętości zasobów.

4. SYMULACJA

W ramach eksperymentu, SU stara się zdobyć informację o dostępności zasobów w obecności sygnału 5G nadawanego w transmisji w dół przez stację bazową (PU). Przyjęte pasmo to 10 MHz, które składa się z 50 RB. Każdy RB składa się z 7 symboli OFDM, transmitowanych w czasie 0.5 ms na 12 podnośnych. Wartości energii używane jako dane wejściowe przez sieci CNN, obliczane są przez SU na podstawie zebranych próbek sygnału tylko z pierwszego symbolu OFDM przypadającego na dany RB. Daje to czas na podjęcie decyzji o stanie zajętości rozpatrywanego RB i na potencjalną transmisję. PU transmituje sygnał losowo w czasie i częstotliwości, jednak wybrane do tego celu RB sąsiadują ze sobą. Sygnał odbierany przez SU jest zniekształcony poprzez shadowing oraz zaniki. Do zasymulowania kanału użyto modelu „The Extended Vehicular A model". Założona częstotliwość Dopplera wynosi 50 Hz.

Model CNN_1 używa nieliniowej funkcji aktywacji o nazwie ReLu (ang. Rectified Linear unit) we wszystkich warstwach oprócz ostatniej warstwy, dla której wybrano funkcję nazywaną powszechnie funkcją „Softmax". CNN_1 ma za zadanie dokonać klasyfikacji danych wejściowych, dlatego jako funkcję strat przyjęto funkcję kategoryzującą „Sparse Categorical function". Aby wyznaczyć współczynniki sieci, użyto optymalizatora „Adam", dla którego przyjęto współczynnik uczenia równy 0.0001. Wynikiem działania CNN_1 są wartości prawdopodobieństwa zajętości wszystkich rozpatrywanych RB, które są reprezentowane przez liczby 0, lub 1. Liczba 0 oznacza, że dany RB został uznany za niewykorzystywany przez PU, natomiast liczba 1 oznacza, że RB jest zajęty.

Model CNN_2 używa funkcji sigmoidalnej jako funkcji aktywacji we wszystkich warstwach. Podobnie jak w CNN_1 użyto optymalizatora „Adam" ze współczynnikiem uczenia równym 0.001. Błąd średniokwadratowy posłużył jako funkcja strat.

Aby uzyskać reprezentatywne rezultaty, wybrano 11 lokalizacji SU wokół nadajnika PU. Każda lokalizacja charakteryzuje się innym średnim SNR. Wartości SNR wybrano z zakresu od -24 dB do 16 dB.

Założono, że SU jest zdolny wytrenować swoje własne sieci dla swojej lokalizacji, lub może pobrać gotową sieć z zewnętrznego serwera.

Poziom zaników stanowi ocenę jakości kanału i przyjmuje wartości z zakresu od 0 do 1, gdzie 0 oznacza najgorsze zaobserwowane warunki detekcji, natomiast 1 oznacza najlepszy kanał. Aby znaleźć najlepszą wartość progową zaników, przetestowano parę możliwych wartości: [0.0001, 0.0005, 0.001, 0.002, 0.003, 0.005, 0.01, 0.02, 0.03, 0.05, 0.1, 0.15, 0.2, 0.3].

5. WYNIKI

5.1. Sposób uzyskania wyników i metoda ich ewaluacji

Jako metryk oceny uzyskanych wyników predykcji, użyto wartości prawdopodobieństwa poprawnej detekcji P_d oraz prawdopodobieństwa fałszywej decyzji potwierdzającej obecność sygnału, tzw. prawdopodobieństwa fałszywego alarmu P_{fa}. Celem jest maksymalizacja P_d, przy równoczesnej minimalizacji P_{fa}. Wartość P_d stanowi miarę jakości ochrony PU, natomiast im niższa wartość P_{fa}, tym więcej jest możliwości transmisyjnych dla SU.

Najlepsza wartość progowa zaników powinna spełniać:

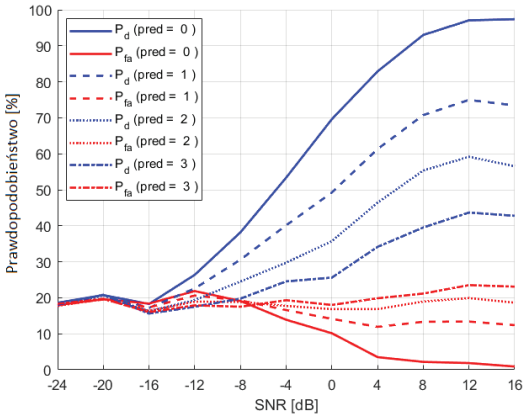
$$arg\max_{thr} P_{\mathrm{d}}^{thr} - P_{\mathrm{d}}^{\mathrm{CNN}_1} + P_{\mathrm{fa}}^{\mathrm{CNN}_1} - P_{\mathrm{fa}}^{thr}, \qquad (1)$$

gdzie $P_{\mathrm{d}}^{\mathrm{CNN}_1}$ i $P_{\mathrm{fa}}^{\mathrm{CNN}_1}$ to wartości P_d i P_{fa}, które zostałyby uzyskane, gdyby cały algorytm korzystał tylko z sieci CNN_1, natomiast P_{d}^{thr} i P_{fa}^{thr} to P_d i P_{fa} uzyskane w wyniku działania algorytmu korzystającego zarówno z

CNN_1 jak i CNN_2 wg przedstawionego na Rys. 5. schematu, przy założeniu, że wartość *thr* to wartość progowa zaników.

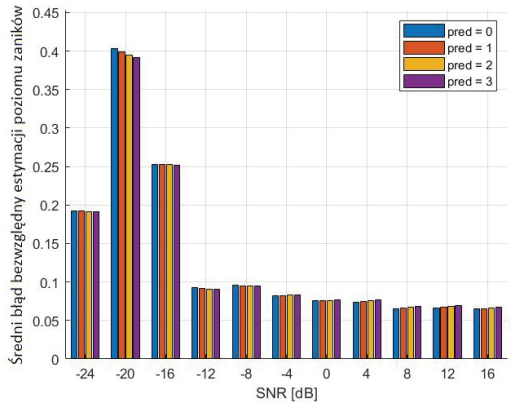
5.1. Analiza wyników

Na Rys. 6. przedstawiono wyniki działania CNN_1 dla detekcji (oznaczenie *pred* = 0) oraz dla predykcji trzech nadchodzących RB (*pred* = 1, 2, 3). Można zaobserwować, że zarówno P_d jak i P_{fa} przyjmują wartości ok. 20% dla niskich wartości SNR. Wraz ze zwiększającą się średnią wartością SNR wartości P_d i P_{fa} coraz bardziej się od siebie różnią, osiągając stałe wartości dla SNR równego ok. 16 dB. Wyraźna jest rozbieżność między wynikami dla detekcji i predykcji. Maksymalna wartość P_d dla *pred*=1 osiąga wartość ok. 75%, czyli predykcja jednego RB wprzód jest o ponad 20% gorsza niż detekcja zajętości aktualnie trwającego RB. Wraz ze zwiększającym się parametrem pred, maksymalne wartości P_d są coraz niższe. Niższe wyniki P_d dla predykcji nie gwarantują niższych wartości P_{fa}. Wartość P_{fa} dla *pred*=1 osiąga poziom ok 10%, podczas gdy dla detekcji P_{fa} było bliskie zeru. Z obserwacji rosnących P_{fa} i malejących P_d wraz ze wzrastającym *pred* można wywnioskować, że przy pewnej wartości *pred* oba prawdopodobieństwa nie będą się różnić od siebie dla żadnej wartości SNR.



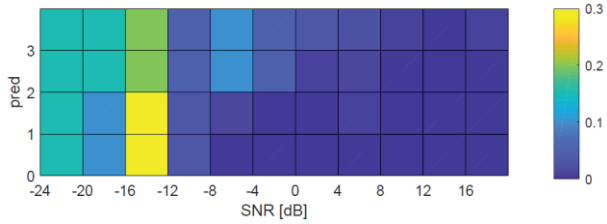
Rys. 6. Prawdopodobieństwo detekcji i predykcji oraz prawdopodobieństwa fałszywego alarmu dla CNN_1

Na Rys. 7. Przedstawiono wyniki działania CNN_2. Są to wartości bezwzględnego błędu estymacji poziomu zaników. Maksymalny błąd uzyskano dla wartości SNR=-20 dB. Wraz ze zwiększającym się SNR, błąd estymacji poziomu zaników maleje. Co ciekawe, nie ma dużej różnicy w wynikach dla różnych wartości parametru pred. Jest to spowodowane tym, że kanał zmienia się wystarczająco powoli, aby poziomy zaników w kolejnych sąsiadujących blokach RB był dosyć podobne do siebie.



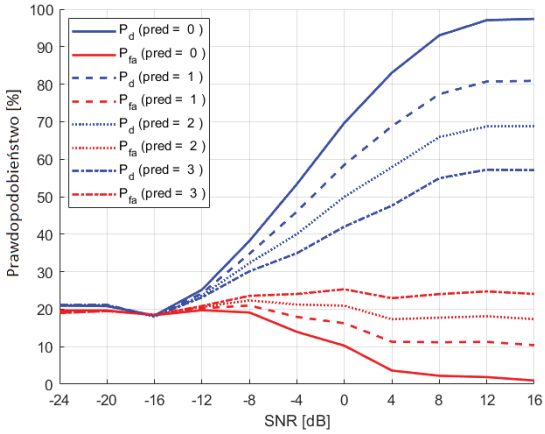
Rys. 7. Błąd bezwzględny estymacji poziomu zaników dla detekcji i predykcji

Po dokonaniu wstępnej predykcji za pomocą CNN_1 i estymacji poziomu zaników za pomocą CNN_2 następuje wybór najlepszej wartości progowej zaników, w celu podjęcia decyzji o zwiększeniu ochrony PU w przypadku silnych zaników. Okazuje się, że w zależności od SNR i *pred* można wyznaczyć różne, optymalne *thr*, które przedstawiono na Rys. 8. Dla niższych SNR, preferowane są wyższe wartości *thr*, natomiast im wyższy SNR, tym niższa optymalna wartość *thr*. Podobnie jak dla wyników działania CNN_2, na wyniki dla niskich SNR wpływa stosunkowo wysoki poziom szumu i zaniki. Większość z uzyskanych optymalnych wartości *thr* jest mniejsza od 0.05, a dla SNR ≥ 8dB *thr* nie jest większy od 0.005.



Rys. 8. Najlepsze wartości progowe zaników w zależności od parametru pred i SNR

Ostateczne wyniki działania całego algorytmu przedstawiono na Rys. 9. Aby je uzyskać, założono wspólną dla wszystkich *pred* i SNR wartość progową zaników *thr* = 0.003. Można zauważyć, że wprowadzenie wartości progowej *thr* do podejmowania ostatecznej decyzji o zajętości zasobów wpłynęło pozytywnie na wyniki uzyskane przez CNN_1. Krzywe P_d rosną szybciej wraz ze zwiększającą się wartością SNR. Dla najwyższych SNR i *pred*=1, *pred*=2 i *pred*=3, P_d wzrosło odpowiednio o 7.44%, 12.25% oraz 14%. Można zaobserwować lekkie pogorszenie w wynikach P_{fa} i różnią się od wyników CNN_1 w granicach 1% do 2%. Dla detekcji, czyli dla *pred*=0 zarówno P_d jak i P_{fa} są zbliżone do analogicznych wyników uzyskanych przez CNN_1.



Rys. 8. Ostateczne wyniki działania algorytmu predykcji z uwzględnieniem oceny zaników

6. WNIOSKI

W niniejszym artykule zaproponowano i zbadano nowy algorytm bazujący na ML użyty dla poprawy predykcji wolnych zasobów w sieci 5G. Poprawa predykcji jest znacząca. Główny cel, tj. ochrona transmisji PU, został osiągnięty. Pomimo zwiększenia wartości prawdopodobieństwa detekcji i predykcji w celu ochrony PU, okazało się, że możliwe jest utrzymanie minimalnych strat okazji do transmisji dla SU. Poprawa jakości predykcji umożliwia SU lepsze przygotowanie się do przyszłej transmisji lub kontynuowania już rozpoczętej transmisji.

LITERATURA

[1] Bhamare Devyani, Poonam Suryawanshi. 2018. „Review on reliable pattern recognition with machine learning techniques". *Fuzzy Information and Engineering*, 10 (3): 362–377.

[2] Omotere Oluwaseyi, Fuller John, Qian Lijun Han Zhu. „Spectrum occupancy prediction in coexisting wireless systems using deep learning". 2018. IEEE 88th Vehicular Technology Conference: 1–7.

[3] Shawel Bethelhem, Woldegebreal Dereje, Pollin Sofie. 2019 „Convolutional lst m-based long-term spectrum prediction for dynamic spectrum access". 27th European Signal Processing Conference (EUSIPCO): 1–5

[4] Tian Jinnfeng, Cheng Peng, Chen Zhuo et al. 2019. „A machine learning-enabled spectrum sensing method for OFDM systems". *IEEE Transactions on Vehicular Technology*, 68 (11): 11374–11378.

[5] Wasilewska Małgorzata, Bogucka Hanna, Kliks Adrian. 2020 „Spectrum Sensing and Prediction for 5G Radio". *Big Data Technologies and Applications,* 176–194.

[6] Wasilewska Małgorzata, Kliks Adrian, Bogucka Hanna, et al. 2021. „Artificial Intelligence for Radio Communication Context-Awareness". *IEEE Access*

ESTYMACJA I ANALIZA STANU KANAŁU RADIOWEGO INTERFEJSU LTE NA POTRZEBY REALIZACJI HETEROGENICZNEGO ALGORYTMU MULTILINK Z UŻYCIEM GŁĘBOKIEGO UCZENIA

ESTIMATION AND ANALYSIS OF THE RADIO CHANNEL STATE OF THE LTE INTERFACE FOR THE IMPLEMENTATION OF THE HETEROGENEOUS MULTILINK ALGORITHM WITH THE USE OF DEEP LEARNING

Krzysztof K. Cwalina, Olga Błaszkiewicz, Alicja Olejniczak, Piotr Rajchowski, Jarosław Sadowski

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk
kkcwalina@eti.pg.edu.pl, olga.blaszkiewicz@pg.edu.pl, alicja.olejniczak@pg.edu.pl,
piorajch@eti.pg.edu.pl, jaroslaw.sadowski@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.87

Streszczenie: **W niniejszym artykule przedstawiono estymację i analizę stanu kanału radiowego dla interfejsu LTE na potrzeby realizacji heterogenicznego algorytmu trasowania przesyłanych pakietów pomiędzy dostępnymi interfejsami radiowymi. Udowodniono przydatność i zwiększenie efektywności predykcji metryki BLER opracowanego modelu głębokiego uczenia względem modelu liniowego. Przedstawiono także metodykę oraz przeanalizowano istotność parametrów wejściowych uzyskanego modelu głębokiego uczenia.**

Abstract: **In the article the performed estimation and analysis of the radio channel state for the LTE air interface for the implementation of a heterogeneous algorithm for routing the transmitted packets between available radio interfaces is presented. The usefulness and an efficiency increase of the BLER metric prediction of the developed deep learning model in relation to the linear model were proven. The methodology and the significance of the input parameters of the obtained deep learning model was also analyzed and presented.**

Słowa kluczowe: **głębokie uczenie, heterogeniczne sieci, kanał radiowy, LTE**

Keywords: **deep learning, heterogeneous networks, radio channel, LTE**

1. WSTĘP

Ze względu na swoją przydatność i uniwersalność zastosowań w różnorodnych, trudnych pod względem propagacji fal warunkach propagacyjnych nowoczesne urządzenia telemetryczne i sterujące często realizowane są jako urządzenia heterogeniczne, tj. posiadające więcej niż jeden interfejs do komunikacji radiowej. Taki sposób komunikacji tychże urządzeń determinuje zarówno możliwość realizacji transmisji z użyciem wszystkich interfejsów radiowych równolegle np. zwiększając tym samym potencjalnie chwilową przepływność lub z użyciem wybranego pojedynczego interfejsu radiowego.

Innym sposobem korzystania z dostępnych kilku interfejsów radiowych jest wybór łącza pozwalającego w danej chwili na uzyskanie najwyższych parametrów jakościowych, a więc poprawa niezawodności komunikacji. Algorytm, który umożliwi trasowanie przesyłanych pakietów pomiędzy dostępnymi interfejsami radiowymi w zależności od ich parametrów funkcjonalnych oraz parametrów łącza radiowego określono jako algorytm multilink (A-ML). Niewątpliwie, sposób wyboru takiego interfejsu oraz jego parametrów pracy, by móc przesłać aktualne dane pomiarowe lub sterujące jest zagadnieniem niezmiernie złożonym, szczególnie w zróżnicowanych środowiskach propagacyjnych oraz ze zmiennym obciążeniem sieci.

Celem prowadzonych prac jest realizacja A-ML do efektywnego wyboru łącza komunikacyjnego z dostępnych interfejsów komunikacyjnych w heterogenicznym urządzeniu telemetrycznym. Jak opisano w [3], wynikowa wartość efektywności algorytmu multilink zależy od efektywności predykcji wybranych parametrów jakościowych danego interfejsu radiowego oraz oczywiście od właściwego algorytmu decyzyjnego, który determinuje sposób trasowania pakietów w urządzeniu na podstawie obliczonych estymat. Ze względu na złożoność realizacji całego A-ML, zdecydowano się na jego dekompozycję i na tym etapie prac w niniejszym artykule przeprowadzono badania interfejsu radiowego LTE (Long Term Evolution), jako jednego z potencjalnie występujących w docelowych urządzeniach.

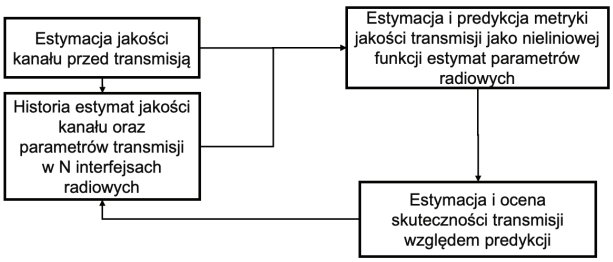
Warto zaznaczyć, iż niniejszy artykuł stanowi rozwinięcie prac przedstawionych przez autorów w [3] m.in. o interpretację i analizę istotności parametrów modelu sieci neuronowej dla interfejsu radiowego LTE. Ma to na celu przedstawienie proponowanej metodyki oraz wyniki pośredniego wyznaczenia istotności parametrów wejściowych i zależności wartości historycznych na potrzeby projektowania rozwiązań do estymacji i analizy stanu kanału radiowego z użyciem głębokiego uczenia.

2. WYDZIELONY INTERFEJS RADIOWY NA POTRZEBY REALIZACJI A-ML

Niewątpliwie, efektywność A-ML determinowana jest przez efektywność estymacji i predykcji parametrów jakościowych każdego interfejsu radiowego osobno [3]. Mając na względzie niniejsze, można dokonać dekompozycji algorytmu wyboru łącza komunikacyjnego dla każdego interfejsu radiowego i określić efektywność jego

pracy (przekładającą się na końcową efektywność A-ML), co przedstawiono na rys. 1.

Decyzję o wyborze interfejsu komunikacyjnego dla kolejnej transmisji stanowi prawdopodobieństwo warunkowe realizacji transmisji z określonymi parametrami (np. błędu bitowej stopy błędów, błędu blokowej stopy błędów, opóźnienia itd.) w określonych warunkach propagacyjnych. Warunki propagacyjne determinują tzw. jakość kanału radiowego, która jest estymowana przez terminale ruchome i stacje bazowe danej sieci i wyrażana jest parametrycznie jako np. moc sygnału odbieranego, stosunek sygnału do szumu, stosunek sygnału do szumu i interferencji itd.



Rys. 1. Schemat blokowy zdekomponowanego algorytmu multilink dla pojedynczego interfejsu radiowego

Na potrzeby realizacji A-ML trasującego pakiety z danymi użytkownika konieczna jest estymacja możliwych do zmierzenia parametrów jakościowych kanału radiowego przed realizacją kolejnej transmisji. Należy jednak pamiętać, że zbiór parametrów możliwych do estymacji jest ograniczony i w ogólności definiowany przez użyty moduł radiowy pełniący rolę interfejsu radiowego w heterogenicznym urządzeniu telemetrycznym. Wyznaczone estymaty jakości kanału radiowego (parametry kanału radiowego) stanowią informację użyteczną nie tylko dla aktualnie planowanej transmisji, ale także dla kolejnych, przy czym długość historii analizowanych próbek określono na drodze badań pomiarowych i prac analitycznych [3].

Estymacja i predykcja metryk jakości przed transmisją realizowana może być na wiele sposobów, przy czym w ogólności możemy podzielić je na metody liniowe oraz metody nieliniowe. Wydaje się, iż odwzorowanie analizowanych estymat jakości kanału i predykcja wybranych metryk jakości dla poszczególnych interfejsów radiowych powinna być funkcją nieliniową, przy czym funkcja liniowa umożliwi określenie dolnej granicy akceptowanych rozwiązań. Górną, teoretyczną granicą jest możliwość predykcji metryk jakości interfejsów radiowych z prawdopodobieństwem równym 1.

Metryki jakości stanowić mogą zarówno parametry stricte definiujące łącze radiowe w łączu w górę, czy w łączu w dół (np. bitowa stopa błędów, blokowa stopa błędów), jak i parametry transmisji zależne od charakterystyki pracy sieci operatora komórkowego (np. opóźnienie pomiędzy dwoma urządzeniami na poziomie warstwy IP).

Po realizacji procesu decyzyjnego na podstawie przewidywanych metryk jakości kanału (transmisji) estymowana jest skuteczność transmisji definiowana jako wyznaczenie błędów wartości oczekiwanych z rzeczywistymi wartościami, które uzyskano w trakcie transmisji danych użytecznych. Parametry te przekazywane są do bloku pamięci przechowującego historię danych ze zdefiniowaną długością pamięci, który będzie aktualizowany w trakcie pracy urządzenia i wykorzystywany do zwiększenia efektywności realizacji kolejnych transmisji.

3. STANOWISKO I SCENARIUSZE POMIAROWE

Na potrzeby badań zrealizowano stanowisko pomiarowe opisane szczegółowo w [1, 2], które zawiera w szczególności: tester radiokomunikacyjny CMW500 firmy Rohde&Schwarz emulujący stację bazową LTE eNodeB (Evolved Node B), moduł radiowy RM500Q-GL firmy Quectel, komputer jednopłytowy do zarządzania testerem i modułem radiowym w celu realizacji zautomatyzowanych pomiarów.

Z użyciem opisanego stanowiska pomiarowego zrealizowano kampanię pomiarową. W tab. 1 przedstawiono parametry pracy emulowanej stacji bazowej eNodeB.

Tab. 1. Parametry pracy emulowanej stacji bazowej

Parametr	*Wartość*	
Pasmo pracy	OB1	UL: 1920-1980 MHz DL: 2110-2170 MHz
Szerokość pasma pracy	10 MHz	
Profil środowiskowy	EP5 Medium	
Odchyłka dopplerowska	1 Hz	
Zakres zmian SNR	[-3 dB; 1 dB; 30 dB]	
Widmowa gęstość mocy sygnału nadawanego	-75 dBm/15 kHz	
Liczba przesłanych podramek w ramach pojedynczej transmisji	100	
Odstęp czasu pomiędzy pomiarami	maks. 300 ms	

Określono, iż analiza w kontekście realizacji algorytmu multilink skupiona będzie w szczególności na parametrach dostępnych i możliwych do pozyskania z modułów radiowych przed potencjalną transmisją z użyciem interfejsu radiowego LTE w docelowych warunkach pracy A-ML. Pozyskiwano parametry jakościowe transmisji oraz estymaty jakości kanału radiowego zgodne ze standardem, tj.: BLER [%], RSRP (Reference Signal Received Power) [dBm], RSRQ (Reference Signal Received Quality) [dB], RSSI (Reference Signal Strength Indicator) [dBm], SINR (Signal to Interference plus Noise Ratio) [dB], MCS (Modulation and Coding Scheme).

W trakcie pomiarów zmienną wartość SNR uzyskiwano poprzez zmianę mocy sygnału wyjściowego, a nie zmianę poziomu interferencji (filtrowanego szumu białego) względem sygnału użytecznego.

4. ANALIZA UZYSKANYCH WYNIKÓW

Podczas pomiarów uzyskano ponad 100 000 zestawów danych pomiarowych. Metrykę jakości transmisji stanowi blokowa stopa błędów BLER, zdefiniowana jako liczba błędnie odebranych bloków zasobów w kanale PDSCH (Physical Downlink Shared Channel) do liczby wszystkich przesłanych bloków zasobów. Przy czym, operacja obliczania BLER realizowana jest na podstawie

obliczonych sum kontrolnych CRC (Cyclic Redundancy Code) po dekodowaniu kanałowym w odbiorniku, czyli jeszcze w warstwie fizycznej interfejsu radiowego.

Z oczywistych względów wybrany parametr jakościowy BLER powinien być wyliczany (predykcja) przed realizacją transmisji. Uwzględniając opisaną w rozdziale 2. koncepcję A-ML oraz badania pomiarowe przedstawione w [3] określono, iż BLER stanowi nieliniową funkcję względem estymat jakościowych stanu kanału radiowego i schematu modulacyjno-kodowego. W takim przypadku estymata $\hat{y}$ stanu interfejsu radiowego wyrażona może być wzorem

$$\hat{y} = BLER = f\begin{pmatrix} \boldsymbol{RSRP_b}, \boldsymbol{RSRQ_b}, \\ \boldsymbol{RSSI_b}, \boldsymbol{SINR_b}, \boldsymbol{MCS_b} \end{pmatrix}, \quad (1)$$

gdzie argumentami wejściowymi są wektory wartości pozyskanych przed realizacją transmisji (oznaczone indeksem b) także dla danych historycznych, których długości odpowiadają oknom czasowym, które poddawane są analizie w procesie predykcji jakości kolejnej transmisji.

4.1. Model liniowy

Wyznaczając liniową zależność funkcji (1) z użyciem wielowymiarowej regresji wielorakiej zdefiniowano dolną granicę akceptowanych rozwiązań [3]

$$\hat{y}_{lin} = BLER = a + b \cdot RSRP_b + c \cdot RSRQ_b + d \cdot RSSI_b + e \cdot SINR_b + f \cdot MCS_b, \quad (2)$$

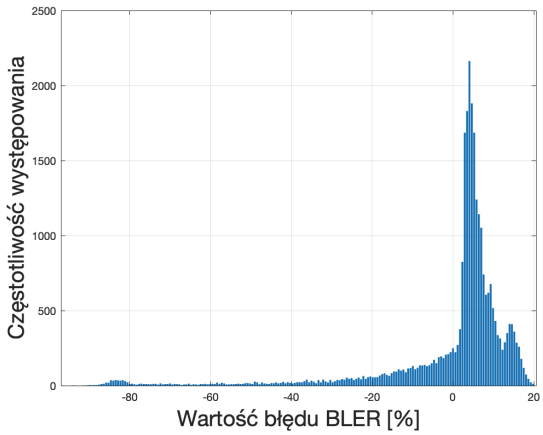
gdzie $a = 0{,}207$, $b = -0{,}718$, $c = -1{,}673$, $d = 0{,}937$, $e = -0{,}732$ oraz $f = -0{,}102$.

Zgodność modelu liniowego względem danych empirycznych zbadano wyznaczając współczynnik determinacji R^2, który zawiera się w zakresie [0; 1], gdzie wartość 1 oznacza idealne dopasowanie modelu do danych empirycznych. Wartość R^2 obliczano dla wszystkich danych i wartości modelu, które zgodne są z matematycznymi obliczeniami zgodnie z modelem, nawet jeśli nie są zgodne z ich fizyczną interpretacją (np. wartości BLER mniejsze od 0% lub większe od 100%). Ograniczenie wartości z modelu realizowano już po faktycznym wykonaniu regresji. Zgodnie z oczekiwaniami, wyznaczona wartość R^2 równa 0,07 wskazuje, iż model liniowy w tym przypadku nie opisuje w sposób dokładny pomierzonych parametrów radiowych z uzyskiwaną metryką jakościową BLER.

W celu ujednolicenia sposobu oceny efektywności wybranych metod (zarówno liniowej jak i nieliniowej) przyjęto, iż metrykę oceny stanowić będzie wartość pierwiastka błędu średniokwadratowego RMSE (Root Mean Square Error).

W trakcie badań [3] określono, iż największą efektywność predykcji metryki BLER interfejsu LTE dla modelu liniowego uzyskuje się używając filtrowanych dolnoprzepustowo wartości parametrów jakościowych z uwzględnieniem okna $W = 15$ wartości historycznych. Wartość średnia błędów wynosiła 0%, natomiast wartość pierwiastka błędu średniokwadratowego RMSE wyniosła 16,99%. Na rys. 2 przedstawiono histogram estymowanych błędów predykcji metryki BLER dla predykcji z użyciem filtru FIR (Finite Impulse Response) o długości $W = 15$.

Ze względu na ograniczoną efektywność, niezbędna jest realizacja nieliniowej funkcji do predykcji metryki jakości transmisji, przy czym postać tej funkcji jest nieznana i w ogólności 5-cio wymiarowa. Mając na względzie niniejsze zdecydowano, iż znalezienie wielowymiarowej funkcji minimalizującej błąd metryki jakości transmisji może zostać efektywnie zrealizowane z użyciem metod głębokiego uczenia.



Rys. 2. Histogram estymowanych błędów predykcji metryki BLER dla predykcji z użyciem filtru FIR o długości W = 15

4.2. Model głębokiego uczenia

Niewątpliwie, problem predykcji jakości kolejnej transmisji jest problemem regresyjnym, a pozyskane dane pomiarowe umożliwiają realizację głębokiego uczenia i fazy uczenia tzw. metodą uczenia nadzorowanego (supervised learning). Jako metrykę efektywności pracy głębokiego uczenia wybrano błąd średniokwadratowy MSE (Mean Square Error). Umożliwiło to zarówno minimalizację błędu średniego, ale także skupienie wyników wokół wartości średniej przy minimalizacji prawdopodobieństwa występowania dużej różnicy względem rzeczywistej wartości blokowej stopy błędów.

Węzły rozproszonych sieci telemetrycznych nie dysponują dużymi mocami obliczeniowymi i należy w ogólności założyć, iż minimalizacja energochłonności takiego urządzenia jest jednym z głównych czynników, które należy uwzględnić. Przyjęto, iż sieć powinna być jak najmniej złożona oraz w jak największym stopniu wspierać możliwość zrównoleglania obliczeń, dlatego też wybrano architekturę sieci w pełni połączonych bez sprzężenia zwrotnego FFNN (FeedForward Neural Network). Sieci rekurencyjne charakteryzują się sekwencyjnym sposobem przetwarzania danych, natomiast sieci splotowe w porównaniu z sieciami FFNN o ograniczonej liczbie warstw i węzłów mogłyby zwiększyć liczbę obliczeń koniecznych do realizacji predykcji, stąd nie były brane pod uwagę w niniejszych badaniach.

Optymalna architektura sieci jest w ogólności nieznana i pomimo prób zdefiniowania reguł przedstawio-

nych w literaturze możliwe są one do stosowania wyłącznie w ograniczonych przypadkach. Dlatego też wyznaczono hiperparametry analizowanych architektur FFNN w zakresie: liczby warstw ukrytych, liczby węzłów w warstwach ukrytych, postaci funkcji aktywacji oraz liczby węzłów wejściowych (tj. liczby uwzględnianych danych historycznych). Biorąc pod uwagę stosunkowo ograniczoną liczbę parametrów do sprawdzenia, zdecydowano się użyć dwuetapowej metody siatkowej (Grid Search), tj. zgrubnego określenia zależności efektywności predykcji BLER jako funkcji zmiennych parametrów architektury sieci a następnie rozszerzonej (dokładniejszej) analizy.

Wszystkie dane wejściowe w sieci, ze względu na inną charakterystykę zmian, podlegały normalizacji do zakresu wartości [0, 1]. Na podstawie uzyskanych wartości funkcji straty do optymalizacji procesu uczenia zastosowano zmodyfikowany algorytm stochastycznego spadku gradientu z adaptacyjnym współczynnikiem uczenia ADAM (*Adaptive Moment Estimation*), z parametrami $\alpha = 0{,}001$, $\beta_1 = 0{,}999$ i $\varepsilon = 10^{-8}$.

Szczegółowe wyniki oraz opis analityczny dla poszczególnych etapów wyboru optymalnej architektury FFNN autorzy przedstawili w [3]. W tab. 2 przedstawiono wybrane zakresy parametrów w etapie drugim procesu poszukiwania optymalnych hiperparametrów głębokiej sieci oraz parametry procesu uczenia.

Tab. 2. Zakresy zmienności hiperparametrów i parametry uczenia w drugim etapie metody siatkowej.

Parametr	*Wartość*
Liczba warstw ukrytych	[1, 10]
Liczba węzłów w warstwach ukrytych	[20; 20; 100]
Funkcja aktywacji	tansig
Liczba węzłów wejściowych	5 parametrów i ich wielokrotność w zależności od długości analizowanej historii, tj. [0; 30]
Współczynnik uczenia	0,001
Liczba iteracji	maks. 5000 lub maks. 20 błędnych walidacji
Sposób podziału zbiorów danych	60% dane uczące, 20% dane walidacyjne i 20% dane testowe

Poprzez wskaźnik błędnych walidacji rozumiana jest realizacja licznika, którego wartość inkrementowana jest w przypadku braku zmniejszenia błędu średniokwadratowego dla danych walidacyjnych. W przypadku osiągnięcia wartości progowej, proces uczenia jest kończony ze względu na brak wzrostu efektywności uczenia w kolejnych 20. iteracjach.

Na podstawie zrealizowanych badań architektur sieci możliwe jest jednoznaczne określenie znaczenia akwizycji historycznych (poprzednich) estymat parametrów radiowych względem realizacji modelu głębokiego uczenia opartego wyłącznie na chwilowych wartościach parametrów radiowych, niezależnie od stopnia skomplikowania modelu. Efektywność modeli głębokiego uczenia, wyrażona jako macierz błędów RMSE dla wartości chwilowych, jest znacząco mniejsza w przypadku analizowania wyłącznie wartości chwilowych i znajdowała się w zakresie od 16,6% do 17,9%.

Dla modeli uwzględniających historyczne estymaty wartości macierzy błędów zawierają się w przedziale od 10,7% do 17,5%. Należy w tym przypadku zwrócić uwagę na występujące przykłady architektur uzyskujących wartości błędu większe od modelu liniowego, które nie osiągały zbieżności ze względu na osiąganie granicznej wartości progowej błędów walidacji. Oznacza to, iż modele te silnie zależały od rozkładu danych uczących oraz inicjalizowanych wartości wag. W tab. 3 przedstawiono obliczone najmniejsze wartości RMSE dla metryki BLER dla wybranych liczebności analizowanych historycznych parametrów radiowych.

Tab. 3. Obliczone najmniejsze wartości RMSE dla wybranych liczebności analizowanych danych wejściowych.

Liczba wartości historycznych	5	10	15	20	25	30
RMSE [%]	11,0	10,7	11,4	11,1	11,0	11,2
Liczba warstw ukrytych	7	4	2	3	3	3
Liczba węzłów	20	20	40	20	20	20

Na podstawie uzyskanych wyników można określić, iż zwiększanie liczby węzłów oraz liczebności warstw w ogólności nie umożliwiało zwiększenia dokładności predykcji metryki BLER podczas kolejnej transmisji, a najlepsze wyniki uzyskiwano w większości dla liczby węzłów w warstwach ukrytych równej 20 i różnej liczby warstw ukrytych. Jednakże, można zauważyć odwrotną proporcjonalność pomiędzy liczbą warstw ukrytych a liczbą uwzględnionych wartości historycznych w kontekście uzyskiwanego RMSE. Zwiększenie pamięci danych historycznych umożliwia zmniejszenie w modelu liczby warstw przy uzyskiwaniu podobnej wartości RMSE.

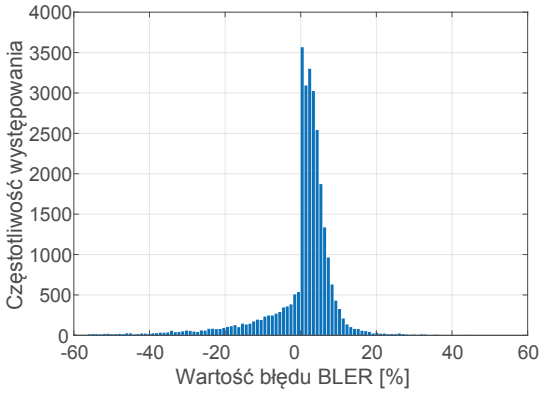
Różnice pomiędzy poszczególnymi modelami nie są większe niż 0,7% błędu RMSE względem najefektywniejszego modelu. Dla analizowanych macierzy błędów, 2% uzyskiwanych modeli posiada wynikowe wartości błędu MSE blokowej stopy błędów większe od modelu liniowego, przy czym 34% uzyskiwanych modeli charakteryzuje się ponad dwukrotnym zyskiem względem modelu liniowego. Wskazuje to na znaczący wzrost efektywności predykcji metryki BLER z użyciem modelu głębokiego uczenia stanowiącego nieliniową funkcję argumentów parametrów radiowych dla interfejsu LTE oraz dużą zbieżność opracowywanych modeli dla zdefiniowanego zagadnienia.

Na podstawie uzyskanych wyników określono architekturę dla najmniejszego RMSE równego 10,7%, której parametry przedstawiono w tab. 4.

Tab. 4. Wyznaczone parametry architektury FFNN.

Parametr	*Wartość*
Liczba warstw ukrytych	4
Liczba węzłów w warstwach ukrytych	20
Funkcja aktywacji	tansig
Liczba węzłów wejściowych	55

Na rys. 3 przedstawiono histogram estymowanych błędów predykcji metryki BLER dla predykcji z użyciem głębokiego uczenia.



Rys. 3. Histogram estymowanych błędów predykcji metryki BLER dla predykcji z użyciem głębokiego uczenia

Uzyskane wyniki potwierdzają, iż możliwe jest z zadanym prawdopodobieństwem określenie jakości transmisji na podstawie pomierzonych przed transmisją parametrów kanału i sieci z zastosowaniem metody głębokiego uczenia. Nieliniowa postać funkcji (1) estymowana z użyciem wyznaczonego modelu głębokiej sieci FFNN charakteryzuje się zerową wartością średnią błędu predykcji, a rozkład błędów wykazuje największą przystawalność do rozkładu gaussowskiego z RMSE równym 10,7% co jest niemal 40% zwiększeniem efektywności względem modelu liniowego, którego najmniejsza wartość RMSE wynosiła 17,01%.

4.3. Analiza istotności parametrów wejściowych

Ze względu na swoją charakterystykę wypracowane modele głębokich sieci neuronowych, a w szczególności sieci FFNN, stanowią duże wyzwanie pod względem analizy ich interpretowalności. Związane jest to także z analizą istotności poszczególnych parametrów wejściowych i ich wpływu na kształt estymowanej przez model sieci nieliniowej funkcji przenoszenia.

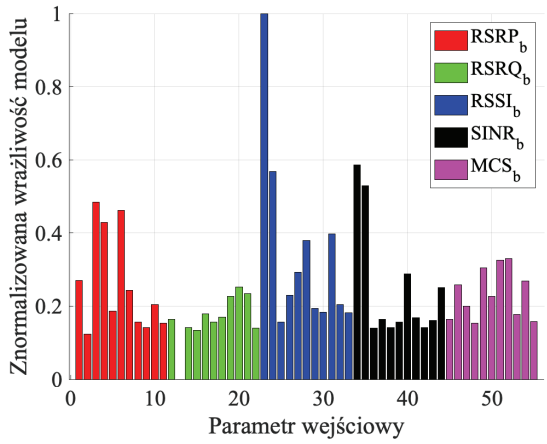
Przy zastosowaniu klasycznych metod regresyjnych, forma opracowanego modelu wraz ze współczynnikami regresji jest wystarczająca do określenia istotności poszczególnych argumentów wejściowych funkcji. Takie podejście, ze względu na charakterystykę głębokich sieci neuronowych nie może zostać zastosowane. Jedną z proponowanych metod analizy modelu jest zastosowanie diagramu Hintona, który wizualizuje względną istotność każdego węzła wejściowego sieci poprzez wizualizację wartości wag poszczególnych neuronów połączonych z analizowanymi węzłami wejściowymi [5]. W takim przypadku, zakłada się, iż analizując diagram Hintona dla pierwszej warstwy ukrytej sieci możliwe jest określenie istotności parametrów wejściowych na końcową postać modelu głębokiej sieci neuronowej.

Jednakże, założenie, iż analizowane wagi dla pojedynczej sieci mają odzwierciedlać niejako współczynniki regresji całego modelu, szczególnie nieliniowego, może być obarczone dużym błędem, m.in. ze względu na interpretację braku istotności parametrów zależnych liniowo, co wykazano w [4]. Udowodniono także, iż wrażliwość modelu na parametry wejściowe jest proporcjonalna do wariancji wyników uzyskiwanych na wyjściu modelu, co oznacza, iż analizując macierz wariancji wartości wyjściowych możemy określić istotność parametrów wejściowych.

Do analizy istotności argumentów funkcji (1) modelowanej proponowaną głęboką siecią FFNN zastosowano metodę modelowania fragmentarycznego proponowaną w [4]. Oparta jest ona na wspomnianej analizie wariancji danych wyjściowych, przy założeniu, iż sprawdzane są fragmentarycznie wydzielone dane wejściowe sieci dla pojedynczego węzła wejściowego przy zerowych wartościach wejściowych pozostałych węzłów (zero padding). W takim przypadku można zdefiniować wrażliwość modelu, tj. wektor wariancji $\boldsymbol{s} = \{s_1, s_2, \dots, s_{N_{we}}\}$, względem liczby parametrów wejściowych N_{we} jako

$$s_i = \sigma^2_{\hat{y}_{dl}(x_i)},\ i = 1, 2, \dots, N_{we}\ , \tag{3}$$

gdzie $\sigma^2_{\hat{y}_{dl}(x_i)}$ to wariancja wartości estymowanych przez model głębokiego uczenia względem pojedynczego parametru wejściowego x_i. Uzyskany wektor wariancji $\boldsymbol{s}$ jest następnie normalizowany względem swojej wartości maksymalnej, uzyskując $\tilde{\boldsymbol{s}}$. Na rys. 4 przedstawiono obliczoną znormalizowaną wrażliwość modelu proponowanej metody głębokiego uczenia.



Rys. 4. Znormalizowana wrażliwość uzyskanego modelu głębokiego uczenia

Biorąc pod uwagę przedstawione wyniki oraz wartości średnie wrażliwości modelu dla grup parametrów wejściowych, które wynoszą kolejno $\mu_{\tilde{s}_{RSRP}}$ =0,26, $\mu_{\tilde{s}_{RSRQ}}$ =0,16, $\mu_{\tilde{s}_{RSSI}}$ =0,34, $\mu_{\tilde{s}_{SINR}}$ =0,24 i $\mu_{\tilde{s}_{MCS}}$ =0,23 można określić, iż wszystkie grupy parametrów radiowych wpływają na końcowe estymaty modelu i nie ma podstaw do redukcji lub całkowitego usunięcia żadnego z nich. Widoczny jest brak liniowej zależności wrażliwości modelu względem wartości historycznych, czy też tożsamych istotności wartości historycznych dla wszystkich grup parametrów (co zwykle jest jednym z założeń przy klasycznym modelowaniu tego typu zależności [2]) oraz

w ogólności najmniejsza istotność $\boldsymbol{RSRQ_b}$ w uzyskanym modelu głębokiego uczenia. Dodatkowo zgodnie z oczekiwaniami, najbardziej aktualne estymaty parametrów $\boldsymbol{RSSI_b}, \boldsymbol{SINR_b}, \boldsymbol{RSRP_b}$ istotnie wpływają na wrażliwość modelu względem wartości historycznych, co jest spowodowane dynamiką kanału radiowego. Przy czym, widoczne jest, iż pomimo normalizacji wartości wejściowych, wartości estymat $\boldsymbol{RSSI_b}$ mają dominujący wpływ na predykcję metryki BLER. Natomiast pojedyncza wartość $RSRQ_{b,2}$, jest nieistotna i mogłaby zostać zredukowana, przy czym nie wpłynie to znacząco na zmniejszenie złożoności obliczeniowej modelu.

5. PODSUMOWANIE

W niniejszym artykule przedstawiono zrealizowaną estymację i analizę stanu kanału radiowego dla interfejsu radiowego LTE na potrzeby realizacji heterogenicznego algorytmu trasowania przesyłanych pakietów pomiędzy dostępnymi interfejsami radiowymi w zależności od ich parametrów funkcjonalnych oraz parametrów łącza radiowego – algorytmu multilink. W tym celu zrealizowano badania pomiarowe z użyciem opracowanego stanowiska pomiarowego, wyznaczono referencyjne wartości dla modelu liniowego, a następnie udowodniono przydatność i niemal 40% zwiększenie efektywności predykcji metryki BLER opracowanego modelu głębokiego uczenia względem modelu liniowego. Zwrócono uwagę na odwrotną proporcjonalność pomiędzy liczbą warstw ukrytych a liczbą uwzględnionych wartości historycznych w kontekście uzyskiwanego RMSE. Zwiększenie pamięci danych historycznych umożliwia zmniejszenie w modelu liczby warstw przy uzyskiwaniu podobnej wartości RMSE. Warto zaznaczyć, iż zaprezentowane wyniki oraz wnioski mogą znaleźć szersze zastosowanie w praktyce niż wyłącznie w A-ML, np. w kontekście, projektowania podobnych rozwiązań dla interfejsu LTE bazujących na estymacji i predykcji jakości transmisji.

Dodatkowo przedstawiono metodykę oraz przeanalizowano istotność parametrów wejściowych uzyskanego modelu głębokiego uczenia poprzez obliczenie znormalizowanych wartości wrażliwości modelu i wskazano uzyskane zależności, które będą przydatne przy realizacji rozszerzonej formy opracowanego modelu zarówno dla innych interfejsów radiowych, jak i dla rozszerzenia modelu dla interfejsu LTE i umożliwienie jego adaptacji dla różnych warunków bez ponownego wyznaczania całego modelu (transfer learning).

Wskazano brak liniowej zależności wrażliwości modelu względem wartości historycznych, czy też tożsamych istotności wartości historycznych dla wszystkich grup parametrów oraz w ogólności najmniejszą istotność parametru $\boldsymbol{RSRQ_b}$ w uzyskanym modelu głębokiego uczenia. Ponadto, najbardziej aktualne estymaty parametrów $\boldsymbol{RSSI_b}, \boldsymbol{SINR_b}, \boldsymbol{RSRP_b}$ istotnie wpływają na wrażliwość modelu względem wartości historycznych, co jest spowodowane dynamiką kanału radiowego.

***

Praca została zrealizowana w ramach projektu pt. „Koncentrator danych elektroenergetycznych z innowacyjną funkcjonalnością decyzyjną, funkcjonalnością bramy i funkcjonalnością multilink, pracujący w środowisku i na brzegu sieci AMI, SCADA, HAN, IoT", współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego, umowa nr POIR.01.01.01-00-0651/19-00 z dnia 09.03.2020 r.

LITERATURA

[1] Błaszkiewicz, O., Burczyk, R., Czapiewska, A., Gajewska, M., Gajewski, S., Marczak, A., Miszewski, M., Sadowski, J., Stefański, J. 2021. Stanowisko badawcze do automatycznej analizy jakości transmisji koncentratora KODEŚ z interfejsami LTE, NB-IoT oraz ISM. *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne*.

[2] Błaszkiewicz, O., Burczyk, R., Czapiewska, A., Gajewska, M., Gajewski, S., Sadowski, J. 2021. Estymacja jakości transmisji w systemie LTE na podstawie wskaźników mierzonych przez moduł radiowy. *XX Krajowa Konferencja Elektroniki*.

[3] Cwalina, K. K., Rajchowski, P., Olejniczak, A., Błaszkiewicz, O., Burczyk, R. 2021. Channel State Estimation in LTE-Based Heterogenous Networks Using Deep Learning. *Sensors*, wol. 21, wyd. 22.

[4] Despagne, F., Massart, D.-L. 1998. Variable selection for neural networks in multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, wol. 40, wyd. 2.

[5] Skapura, David M. 1995. Building Neural Networks. Addison-Wesley Professional. Edycja 1.

MODELOWANIE INTERFERENCJI W SYSTEMACH RADIOKOMUNIKACYJNYCH Z TRANSMISJĄ OFDM ZA POMOCĄ PARAMETRÓW STATYSTYCZNYCH

INTERFERENCES MODELING IN RADIOCOMMUNICATION SYSTEMS WITH OFDM TRANSMISSION USING STATISTICAL PARAMETERS

Monika Zamłyńska[1]

[1] Akademia Wojsk Lądowym im. Generała Tadeusza Kościuszki, Wrocław, monika.zamlynska@gmail.com

DOI: 10.15199/59.2022.4.88

Streszczenie: **Modelowanie systemów radiokomunikacyjnych wymaga opisu kanału z zakłóceniami w sposób, który odzwierciedla rzeczywistość. Ze względu na stochastyczny charakter zakłóceń, do takiego modelowania można wykorzystać narzędzia statystyczne, a interferencje modelować za pomocą zmiennych losowych z dobranymi parametrami w taki sposób, aby ich statystyki opisowe były tożsame ze statystykami rzeczywistych sygnałów. W przypadku systemów z transmisją OFDM (Orthogonal Frequency Division Multiplexing) świetnie sprawdza się zmienna gaussowska.**

Abstract: **Modeling of radiocommunication systems requires the description of the noisy channel in a way that reflects reality. Due to the stochastic nature of the distortions, such modeling can use statistical tools and the interferences can be modeled using random variables with selected parameters in such a way that their descriptive statistics are identical to the statistics of real signals. In the case of systems with OFDM transmission (Orthogonal Frequency Division Multi-plexing), the Gaussian variable works well.**

Słowa kluczowe: **interferencje, OFDM (Orthogonal Frequency-Division Multiplexing), testy normalności, zespolona zmienna losowa.**

Keywords: **complex random variable, interferences, normality tests, OFDM (Orthogonal Frequency-Division Multiplexing).**

1. WSTĘP

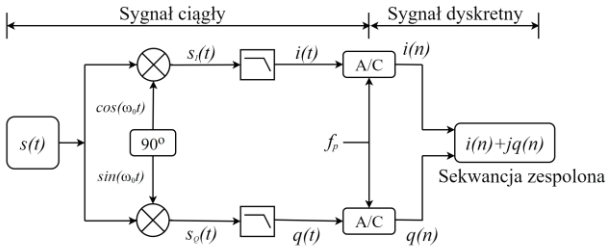
Szumy i interferencje są czynnikami, które zakłócają przekazywanie informacji. Z punktu widzenia modelowania matematycznego szum to sygnał, który ma charakter stochastyczny [1]. Dzięki tej własności możliwe jest wyliczenie statystyk opisowych, m.in.: średniej arytmetycznej, wartości oczekiwanej, wariancji, odchylenia standardowego, a także określić parametry, które pomogą w symulacjach matematycznych.

Szczególnym przypadkiem szumów są te posiadające normalny rozkład prawdopodobieństwa. Rozkłady normalne są charakterystyczne dla wielu sygnałów, choćby takich generowanych przez rezystory, czy też elementy układów elektronicznych [7]. Z kolei podejście do modelowania statystycznego sygnału OFDM (Orthogonal Frequency Division Multiplexing) za pomocą zmiennej losowej o rozkładzie normalnym zaprezentowano w [3, 6]. Jednakże postęp, jaki wymusza rozwój systemów radiokomunikacyjnych, wymaga również właściwego modelowania występujących w nich zakłóceń interferencyjnych.

2. MODELOWANIE INTERFERENCJI

2.1. Model matematyczny

Z punktu widzenia stochastycznego zarówno szum termiczny, jak i każdy sygnał interferencyjny o częstotliwości radiowej, są sygnałami ciągłymi z czasem ciągłym. Jednakże współczesne urządzenia odbiorcze oraz pomiarowe przetwarzają wszelkie sygnały w domenie cyfrowej, tj. po ich dyskretyzacji i kwantyzacji. W praktyce reprezentowane są one w dziedzinie liczb zespolonych, za pomocą składowej *I* (Synfazowej) oraz *Q* (Kwadraturowej) o czasie dyskretnym. Model takiego przekształcenia przedstawiono na rys. 1.



Rys. 1. Przekształcenie sygnału rzeczywistego w dziedzinę liczb zespolonych.

Z punktu widzenia systemu telekomunikacyjnego, szum opisuje się za pomocą mocy średniej. Ze względu na stochastyczny charakter szumu, ma ona związek ze statystyką opisową, jaką jest wariancja σ^2:

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}[x(n) - \hat{x}]^2 \tag{1}$$

Szczególnym przypadkiem jest tzw. szum o rozkładzie normalnym. Dla takiego sygnału wartość średnia $\hat{x}$ wynosi 0, ponieważ nie ma on składowej stałej. Zatem równanie (1) przyjmuje postać:

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}[x(n)]^2 \tag{2}$$

Ze względu na reprezentację sygnału interferencyjnego za pomocą próbek I/Q, moc średnią sygnału opisać jako superpozycję mocy dla składowych rzeczywistej i zespolonej [7]:

$$P_x = P_{Ix} + P_{Qx} \tag{3}$$

gdzie konkretne składowe dane są następującymi zależnościami:

$$P_{Ix} = \frac{1}{N}\sum_{n=1}^{N}[I_n]^2 \quad (4)$$

$$P_{Qx} = \frac{1}{N}\sum_{n=1}^{N}[Q_n]^2 \quad (5)$$

Są to moce średnie dla składowej rzeczywistej *I* oraz składowej urojonej *Q*. W przypadku szumu normalnego składowe *I* oraz *Q* można traktować jako niezależne zmienne losowe, dla których statystyki opisowe liczymy osobno.

Do modelowania zakłóceń świetnie nadaje się szum biały, czyli szum o charakterze gaussowskim bez składowej stałej (wartość oczekiwana takiego sygnału musi być zerowa). Aby zadaną interferencję można było opisać szumem gaussowskim, to funkcja gęstości prawdopodobieństwa dla zespolonych zmiennych losowych tego zakłócenia musi mieć normalny rozkład gęstości prawdopodobieństwa zarówno dla składowej rzeczywistej *I*, jak i dla składowej urojonej *Q*.

2.2. Testy zgodności z rozkładem normalnym

Testy przeprowadzono w narzędziach Statistica i PQStat, gdzie postawiono następujące hipotezy:

- H_0: badany rozkład jest normalny ($p > \alpha$ i rozkład nie jest skośny oraz jego kurtoza odpowiada kurtozie rozkładu normalnego),
- H_1: badany rozkład nie jest normalny ($p < \alpha$ lub rozkład jest skośny lub jego kurtoza nie odpowiada kurtozie rozkładu normalnego).

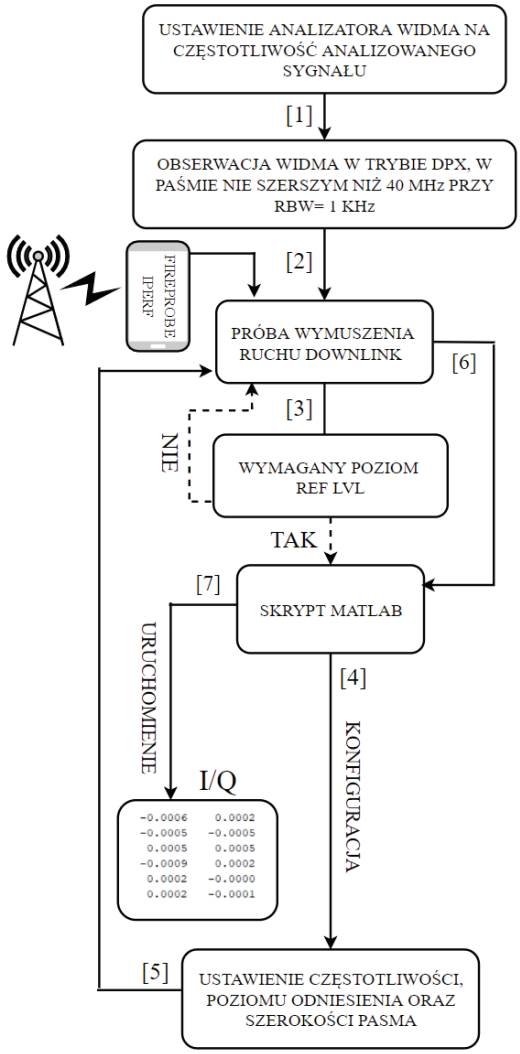
Parametr *p* to prawdopodobieństwo wystąpienia wszystkich możliwych wartości badanej zmiennej losowej [8], a parametr *α* to poziom istotności. Program wylicza wartość statystyki testowej, oraz wartość *p* dla tej statystyki (czyli część pola pod krzywą, która odpowiada wartości statystyki testowej). Wartość *p* pozwala wybrać spośród hipotezy zerowej H_o i alternatywnej H_1 tę bardziej prawdopodobną. Przy czym zawsze zakłada się prawdziwość hipotezy zerowej H_0, a zebrane w danych dowody mają dostarczyć wystarczającej ilości argumentów przeciwko tej hipotezie [8].

Aby mieć większą pewność, że mamy się do czynienia z rozkładem normalnym, hipoteza H_0 musiała zostać przyjęta we wszystkich testach, zarówno dla składowej rzeczywistej *I*, jak i zespolonej *Q*. Testy poszczególnych sygnałów odbyły się dla 5000 próbek, a założony poziom istotności wynosił $\alpha = 0{,}05$, zgadzając się, że w *5%* sytuacji odrzucimy hipotezę zerową H_0, gdy jest ona prawdziwa. Wartość statystyki *p* musi spełniać warunek $p > \alpha$, aby przyjąć hipotezę H_0.

2.3. Układ pomiarowy

W symulacji przeanalizowano trzy rodzaje zakłóceń: szum termiczny, zakłócenie od Wi-Fi 6 (Wireless Fidelity 6) i zakłócenie od sieci komórkowych 5G NR (New Radio) TDD (Time Division Duplex) w paśmie 2600 MHz. Wymienione sygnały zarejestrowanego za pomocą analizatora widma czasu rzeczywistego Tektronix RSA306B. Próbki *I*/*Q* zarejestrowano dla szerokości pasma 10 MHz, co w tym modelu urządzenia odpowiada częstotliwości próbkowania wynoszącej 14 MHz. Poziom odniesienia (reference level) podczas wszystkich pomiarów wynosił -30 dBm, co pozwoliło na uniknięcie zniekształceń wynikających z przestrojenia toru analogowego urządzenia pomiarowego oraz nasycenia przetwornika A/C.

Próbki *I*/*Q* dla sygnału Wi-Fi 6 i 5G TDD w paśmie 2600 MHz zostały zarejestrowane w warunkach wymuszonego obciążenia punktu dostępowego oraz stacji bazowej za pomocą aplikacji mobilnej Fireprobe firmy V-SPEED [1]. Schemat procedury pomiarowej przedstawiono na rys. 2.



Rys. 2. Schemat procedury pomiarowej do rejestracji próbek I/Q.

Z kolei próbki I/Q szumu termicznego zostały zarejestrowane przy podłączeniu do wejścia analizatora widma obciążenia bezodbiciowego, co odpowiada rejestracji szumów własnych urządzenia.

2.4. Wyniki testów zgodności statystycznej

Próbki poddano testom zgodności z rozkładem normalnych, realizując osobno testy zgodności dla składowej *I* oraz osobno dla składowej *Q*. Wyniki przedstawiono w Tabeli 1-7. Na rys. 3-5 przedstawiono rozkłady gęstości prawdopodobieństwa dla szumu termicznego, Wi-Fi 6 oraz 5G TDD o częstotliwości 2600 MHz z otrzymanych wyników, oraz dopasowany do nich rozkład normalny.

Tab. 1. Statystyki opisowe dla szumu termicznego, Wi-Fi 6 i 5G TDD (2600 MHz).

Analizowana zmienna	*Szum termiczny*		*Wi-Fi 6*		*5G TDD (2600 MHz)*	
	I	Q	I	Q	I	Q
Średnia	$4{,}77 \cdot 10^{-7}$	$-3{,}08 \cdot 10^{-8}$	$1{,}00 \cdot -10^{-5}$	$3{,}4 \cdot 10^{-5}$	$1{,}80 \cdot 10^{-5}$	$5{,}00 \cdot 10^{-6}$
Wariancja:	$7{,}28 \cdot 10^{-11}$	$7{,}36 \cdot 10^{-11}$	$3{,}00 \cdot 10^{-6}$	$3{,}00 \cdot 10^{-6}$	$2{,}41 \cdot 10^{-7}$	$2{,}42 \cdot 10^{-7}$
Odchylenie standardowe:	$9{,}00 \cdot 10^{-6}$	$9{,}00 \cdot 10^{-6}$	$1{,}75 \cdot 10^{-3}$	$1{,}66 \cdot 10^{-3}$	$4{,}91 \cdot 10^{-4}$	$4{,}92 \cdot 10^{-4}$
Minimum:	$-3{,}40 \cdot 10^{-5}$	$-3{,}30 \cdot 10^{-5}$	$-6{,}46 \cdot 10^{-3}$	$-5{,}86 \cdot 10^{-3}$	$-1{,}72 \cdot 10^{-3}$	$-1{,}91 \cdot 10^{-3}$
Maksimum:	$2{,}80 \cdot 10^{-5}$	$3{,}70 \cdot 10^{-5}$	$6{,}15 \cdot 10^{-3}$	$7{,}58 \cdot 10^{-3}$	$1{,}78 \cdot 10^{-3}$	$1{,}72 \cdot 10^{-3}$

Tab. 2. Wyniki testu Kołomogrova-Smirnova dla szumu termicznego, Wi-Fi 6 i 5G TDD (2600 MHz).

Analizowana zmienna	*Szum termiczny*		*Wi-Fi 6*		*5G TDD (2600 MHz)*	
	I	Q	I	Q	I	Q
Statystyka D:	0,008	0,002	0,009	0,001	0,011	0,011
Wartość p:	**0,913**	**0,912**	**0,758**	**0,432**	**0,550**	**0,550**

Tab. 3. Wyniki testu Lillieforce'a dla szumu termicznego, Wi-Fi 6 i 5G TDD (2600 MHz).

Analizowana zmienna	*Szum termiczny*		*Wi-Fi 6*		*5G TDD (2600 MHz)*	
	I	Q	I	Q	I	Q
Statystyka D:	0,008	0,008	0,009	0,012	0,011	0,011
Wartość p:	**0,635**	**0,635**	**0,341**	**0,081**	**0,132**	**0,132**

Tab. 4. Wyniki testu Shapiro-Wilka dla szumu termicznego, Wi-Fi 6 i 5G TDD (2600 MHz).

Analizowana zmienna	*Szum termiczny*		*Wi-Fi 6*		*5G TDD (2600 MHz)*	
	I	Q	I	Q	I	Q
Statystyka W:	1,000	1,000	1,000	0,999	1,000	1,000
Statystyka Z:	-0,150	-1,227	0,297	0,857	-0,722	-0,722
Wartość p:	**0,560**	**0,890**	**0,383**	**0,196**	**0,765**	**0,765**

Tab. 5. Wyniki testu D'Agostino-Pearsona dla szumu termicznego, Wi-Fi 6 i 5G (2600 MHz).

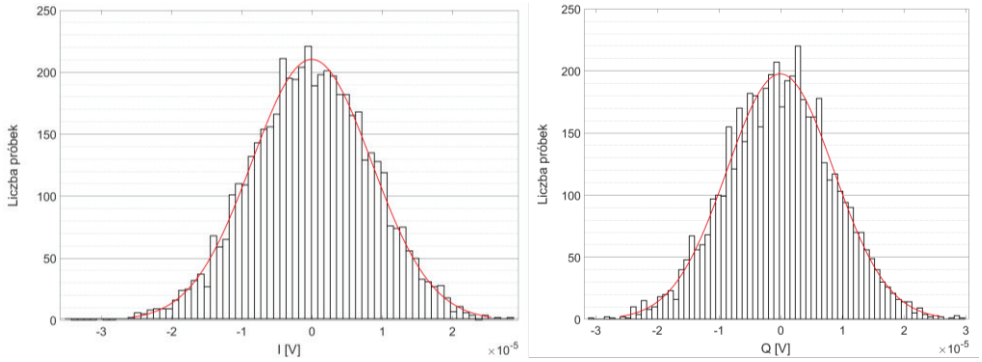
Analizowana zmienna	*Szum termiczny*		*Wi-Fi 6*		*5G TDD (2600 MHz)*	
	I	Q	I	Q	I	Q
Statystyka K^2:	2,603	1,059	2,747	1,784	0,846	0,846
Wartość p:	**0,272**	**0,589**	**0,253**	**0,410**	**0,765**	**0,655**

Tab. 6. Wyniki testu skośności dla szumu termicznego, Wi-Fi 6 i 5G (2600 MHz).

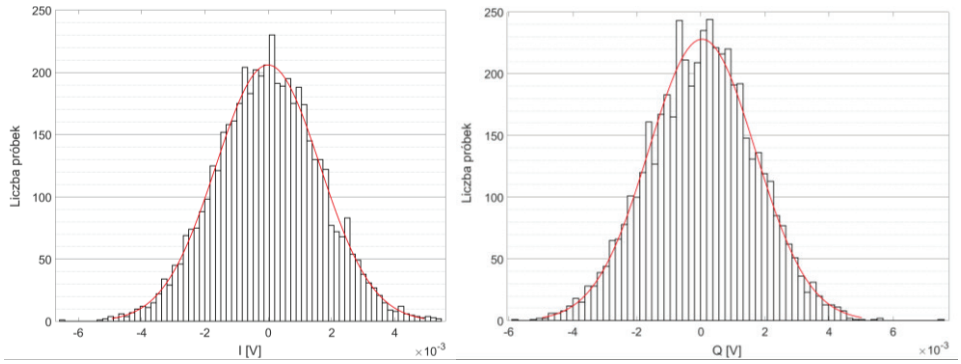
Analizowana zmienna	*Szum termiczny*		*Wi-Fi 6*		*5G TDD (2600 MHz)*	
	I	Q	I	Q	I	Q
Skośność:	-0,045	-0,003	-0,051	-0,0324	-0,010	-0,010
Statystyka Z:	-1,300	0,080	-1,467	-0,938	-0,290	-0,290
Wartość p:	**0,194**	**0,937**	**0,142**	**0,348**	**0,772**	**0,655**

Tab. 7. Wyniki testu kurtozy dla szumu termicznego, Wi-Fi 6 i 5G (2600 MHz).

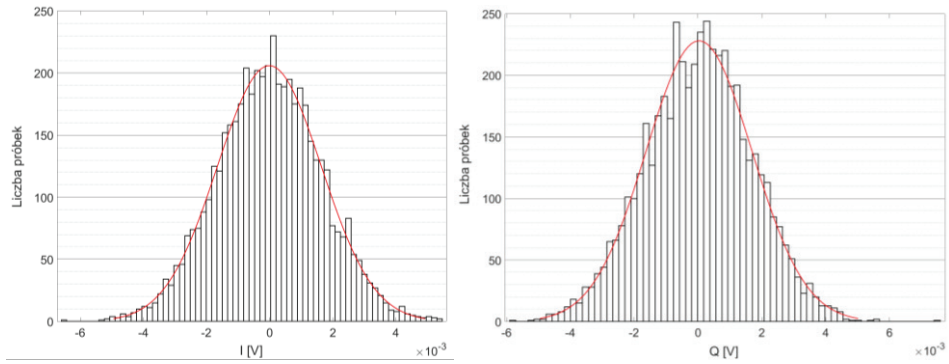
Analizowana zmienna	*Szum termiczny*		*Wi-Fi 6*		*5G TDD (2600 MHz)*	
	I	Q	I	Q	I	Q
Kurtoza:	-0,066	-0,070	0,052	-0,066	-0,061	-0,061
Statystyka Z:	-0,956	-1,026	0,771	-0,951	-0,873	-0,873
Wartość p:	**0,339**	**0,305**	**0,441**	**0,341**	**0,383**	**0,383**



Rys. 3. Histogramy dla składowych I i Q szumu termicznego.



Rys. 4. Histogramy dla składowych I i Q dla WIFI6.



Rys. 5. Histogramy dla składowych I i Q zakłócenia od 5G TDD (2600 MHz).

Z Tabel 1-7 wynika, że wartość statystyki testowej p jest zawsze większa od założonego poziomu istotności α. Oznacza to, że w każdym z testów zgodności przyjęto hipotezę H_0. Implikuje to na fakt, iż zakłócenia od szumu termicznego, Wi-Fi 6 sieci 5G TDD w paśmie 2600 MHz mają rozkład normalny, zatem mogą być modelowane za pomocą szumu gaussowskiego, w którym wartość średnia, minimum, maksimum oraz wariancja będę takie same, jak w zakłóceniu rzeczywistym. Wartość średnia według Tabeli 1-7 jest zbliżona do 0, zatem wybranym zakłóceniem będzie szum biały. W modelu pominięto niezerową wartość średnią oraz ekstrema globalne zarejestrowanego sygnału interferencyjnego.

3. OCENA DOKŁADNOŚCI MODELU INTERFERENCJI

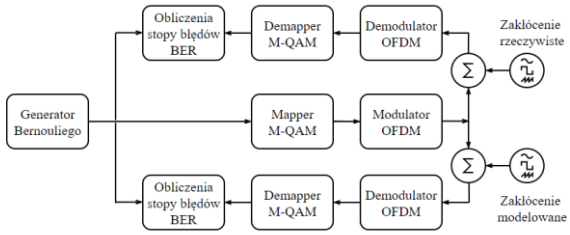
3.1. Model toru radiokomunikacyjnego

Na potrzeby oceny dokładności modelowania interferencji opracowano symulację w środowisku Matlab/Simulink (rys. 6). Sygnał użyteczny pochodzi z generatora Bernouliego, który jest poddawany modulacji wielowartościowej M-QAM (Quadrature Amplitude Modulation) i technice OFDM.

Następnie, sygnał ten rozdzielany jest na dwa kanały:

- pierwszy (z dodawanym zakłóceniem, które zostało wcześniej zarejestrowane przez analizator widma czasu rzeczywistego),
- drugi (z dodawanym modelem zakłócenia, tj. generatorem zespolonej zmiennej losowej o rozkładzie normalnym).

Następnie, oba zakłócone sygnały poddawane są demodulacji i estymacji poziomu bitowej stopy błędów BER (Bit Error Rate). Symulacje przeprowadzono dla dwóch wartościowości modulacji: 64-QAM i 256-QAM, ze względu na to, że te wartościowości są najczęściej używane w systemach 5G [2]. Z uwagi na uproszczony model pominięto kodowanie korekcyjne FEC (Forward Error Correction).

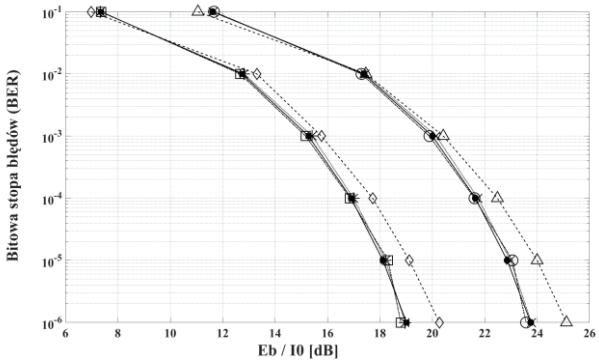


Rys. 6. Schemat toru telekomunikacyjnego użytego w symulacji z wprowadzonymi zakłóceniami do kanału komunikacyjnego.

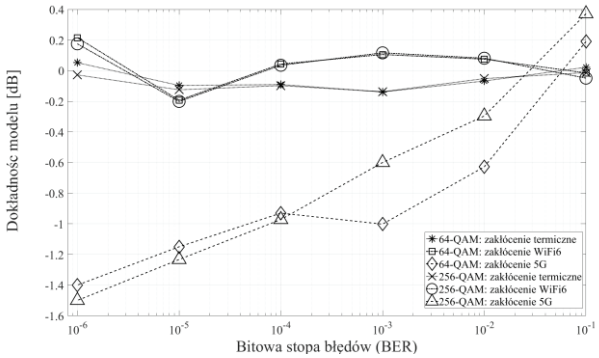
3.2. Wyniki

Na rys. 7 przedstawiono zależność bitowej stopy błędów od jakości sygnałów w modelowanym torze radiokomunikacyjnym wyrażonej poprzez E_b/I_0. Energia przypadająca na bit informacji (E_b) została wyestymowana na podstawie mocy sygnału użytecznego (tj. na wyjściu modulatora OFDM) i prędkości bitowej. Widmowa gęstość mocy zakłócenia I_0 odpowiada mocy sygnału zakłócającego podzielonej przez szerokość pasma, w której moc ta jest mierzona.

Z kolei dokładność modelu interferencji dla zadanych poziomów BER przedstawiono na rys. 8. Zgodnie z wcześniejszymi oczekiwaniami, przedstawiony model interferencji sprawdza się niemalże idealnie, tj. błąd modelu nie przekracza 0,15 dB, dla zarejestrowanego szumu termicznego. Zbliżony poziom dokładności, wynoszący 0,22 dB przy poziomie BER równym 10^{-6}, uzyskano również dla sygnału interferencyjnego pochodzącego od punktu dostępowego sieci Wi-Fi 6. Z kolei dla sygnału zakłócającego pochodzącego od stacji bazowej 5G TDD pracującej w paśmie 2600 MHz, dokładność modelu nie przekracza 1,6 dB przy poziomie BER wynoszącym 10^{-6}, co również należy uznać za dobry wynik.



Rys. 7. Bitowa stopa błędów BER dla zadanych poziomów E_b/I_0.



Rys. 8. Dokładność modelu interferencji dla zadanych poziomów BER.

4. WNIOSKI

W pracy pokazano, że zmienna losowa o rozkładzie normalnym nadaje się nie tylko do modelowania zakłóceń termicznych, ale również interferencji pochodzących od współczesnych systemów radiokomunikacyjnych, które korzystają z transmisji OFDM, czyli m.in. Wi-Fi 6, 5G, a w przyszłości najprawdopodobniej również 6G. Takie podejście znajduje uzasadnienie nie tylko w ujęciu statystycznym, ale również zostało ono poparte symulacją komputerową uproszczonego toru radiokomunikacyjnego. Dalsze prace będę sprowadzać się do rozszerzenia toru o kodowanie korekcyjne FEC oraz inne elementy występujące w nadajniku i odbiorniku urządzeń 5G NR (New Radio), a także większej gamy sygnałów radiokomunikacyjnych zarejestrowanych przy różnych częstotliwościach próbkowania. Pozwoli to na bardziej precyzyjne określenie zakresu stosowalności opisanego sposobu modelowania interferencji.

LITERATURA

[1] FireProbe. https://www.fireprobe.net/pl (dostęp: 13.04.2022).

[2] Holma, Harri, Antii Toskala, Takehiro Nakamura. 2020. *5G technology: 3GPP new radio*. John Wiley & Sons.

[3] Kotzer, Igal, Smadar Har-Nevo, Sasha Sodin, Simon Litsyn, 2012. „A model for OFDM signals with applications." Proceedings of 18th European Wireless Conference 1 : 1–7.

[4] Lyons, Richard G. 2010. *Understanding digital signal processing*. Pearson.

[5] Razali, Nornadiah Mohd, Yap Bee Wah. 2011. „Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests." *Journal of Statistical Modeling and Analytics* 2 (1) : 21–33.

[6] Rudziński, Adam. 2014. „Normalized Gaussian approach to statistical modeling of OFDM signals. *Journal of Telecommunications and Information Technology* 1 : 54–61.

[7] Szabatin, J. 2007. *Podstawy teorii sygnałów*. Wydawnictwa Komunikacji i Łączności.

http://manuals.pqstat.pl/statpqpl [dostep: 2022-06-22]

PLATFORMY DO WDRAŻANIA KOMUNIKACJI W STANDARDZIE LORAWAN

LORAWAN COMMUNICATION IMPLEMENTATION PLATFORMS

Joanna Szewczyk[1]; Mariusz Nowak[2]; Piotr Remlein[1]; Aleksandra Głowacka[2]

[1] Politechnika Poznańska, Instytut Radiokomunikacji, Poznań, Joanna.Szewczyk@cs.put.poznan.pl, Piotr.Remlein@put.poznan.pl

[2] Politechnika Poznańska, Instytut Informatyki, Poznań, Mariusz.Nowak@cs.put.poznan.pl, Aleksandra.Glowacka@student.put.poznan.pl

DOI: 10.15199/59.2022.4.89

Streszczenie: **Wraz z dynamicznym rozwojem Internetu Rzeczy (IoT), wzrasta potrzeba stosowania coraz to nowszych energooszczędnych rozwiązań komunikacyjnych. Sieci LPWAN (ang. Low-Power Wide-Area Network) umożliwiają niskoenergetyczną komunikację dalekiego zasięgu. Istotnym jest by nowe rozwiązania cechowały się łatwością konfiguracji umożliwiającą szybkie wdrożenie. W artykule dokonano przeglądu platform do komunikacji urządzeń IoT w standardzie LoRaWAN (ang. LoRa Wide Area Network) pod kątem stopnia złożoności konfiguracji. Opisano sposób konfiguracji budowanego na Wydziale Informatyki i Telekomunikacji, Politechniki Poznańskiej eksperymentalnego systemu LPWAN do badań związanych z Inteligentnymi budynkami.**

Abstract: **With the rapid growth of the Internet of Things (IoT), there is a growing need for newer energy-efficient communication solutions. Low-Power Wide-Area Networks (LPWAN) enable low-energy long-range communication. It is important for new solutions to be easy to configure and quick to implement. The paper reviews platforms for IoT device communication in LoRaWAN (LoRa Wide Area Network) standard in terms of configuration complexity. The paper describes the configuration of an experimental LPWAN system being built at the Faculty of Computing and Telecommunications, Poznan University of Technology for research related to Intelligent buildings.**

Słowa kluczowe: **Inteligentne Budynki, Internet Rzeczy, LoRaWAN, TTS.**

Keywords: **Intelligent Buildings, Internet of Things, LoRaWAN, TTS.**

1. WSTĘP

Internet Rzeczy (IoT – ang. Internet of Things) jest dziedziną, która dynamicznie się rozwija. Urządzenia IoT otaczają nas niemalże wszędzie, poprawiając jakość życia poprzez automatyzację wielu procesów. Pozwalają na minimalizowanie zużycia energii, co wpływa na redukcję kosztów, a także ograniczanie emisji gazów cieplarnianych do atmosfery.

Do realizacji komunikacji w rozwiązaniach IoT najczęściej wykorzystuje się transmisję bezprzewodową. Wiele istniejących rozwiązań opartych jest na komunikacji poprzez Wi-Fi lub Bluetooth Low Energy (BLE) [1]. Rozwiązania te mają swoje wady i zalety. Jednak coraz popularniejsze staję się wykorzystanie sieci LPWAN. LPWAN charakteryzują się działaniem w nielicencjonowanym pasmie ISM, dalekim zasięgiem oraz małym zużyciem energii.

W artykule dokonano analizy dostępnych serwerów sieciowych do komunikacji urządzeń IoT w standardzie LoRaWAN oraz przedstawiono sposób implementacji serwera sieciowego LoRa (ang. Long Range) na platformie The Things Stack (TTS). Wybór platformy TTS wynika z faktu migracji w 2021 roku wcześniejszego rozwiązania – TTN (The Things Network) do nowej implementacji – TTS. Mimo łatwości wdrażania rozwiązań opartych na sieciach dalekiego zasięgu, zdarzają się problemy, stosunkowo trywialne, a potrafiące utrudnić i opóźnić prace.

2. ANALIZA SERWERÓW SIECIOWYCH

W tym rozdziale, zostały scharakteryzowane wybrane, najpopularniejsze rozwiązania chmurowe, pozwalające na wdrażanie autorskich rozwiązań, wykorzystujących komunikację w sieciach LPWAN, a także platform pozwalających na przetwarzanie i analizę danych gromadzonych na serwerach sieciowych.

LoRaWAN jest jednym z protokołów sieci LPWAN – niskoenergetycznych sieci dalekiego zasięgu. Standard LoRaWAN pozwala na komunikację dwukierunkową w nielicencjonowanym paśmie ISM, o częstotliwości 868 MHz (dla Europy). Dzięki temu, wykorzystanie technologii bezprzewodowej LoRaWAN staje się coraz bardziej popularne w rozwiązaniach dla Internetu przedmiotów i sieci sensorowych [2]-[4].

2.1. TTN (TTS)

Serwis The Things Network migrował do The Things Stack Community Edition w 2021 roku. The Things Stack jest serwerem sieci LoRaWAN, który umożliwia projektowanie różnych scenariuszy wdrożeniowych, obsługuje wszystkie istniejące wersje LoRaWAN (w tym najnowszą 1.1), wszystkie tryby pracy A, B i C oraz wszystkie parametry regionalne wydane przez LoRa Alliance. TTS umożliwia użytkownikom uruchomienie własnego serwera i wykorzystanie dedykowanych rozwiązań lub aplikacji internetowych. Dzięki tym możliwościom TTS jest bardzo dobrym rozwiązaniem dla osób początkujących jak i doświadczonych.

Pierwotnie wdrażanie autorskich rozwiązań na platformie TTN pozwalało na transmisję „w górę” średnio co

30 sekund, w ciągu 24 godzin przez jedno urządzenie i maksymalnie dziesięciu komunikatów w transmisji „w dół", w ciągu 24 godzin. Proces migracji serwisu The Things Networks do The Thing Stack, sprawił, że zwiększyła się możliwa liczba transmitowanych informacji w łączu w dół. W serwisie TTN można było wysłać 10 komunikatów w transmisji „w dół" na dzień, natomiast serwis TTS pozwala na 10 tysięcy komunikatów „w dół" w ciągu 24 godzin [5], [6].

2.2. Chirpstack

Chirpstack jest platformą typu open-source, która zapewnia funkcjonalności dla sieci LoRaWAN. Wymaga uruchomienia serwera na tzw. bramce, gdzie następnie można korzystać z lokalnej wersji interfejsu do zarządzania urządzeniami. Uruchomiony serwer pozwala również na integrację z platformami do analizy i wizualizacji danych. Rozwiązanie Chirpstack pozwala na komercyjne wykorzystanie swoich funkcjonalności oraz posiada wsparcie dla urządzeń wszystkich klas w standardzie LoRaWAN oraz wspiera wszystkie wersje LoRa [7].

2.3. Openstack

Openstack jest serwerem chmurowym pozwalającym na wdrażanie autorskich rozwiązań z obszaru IoT z wykorzystaniem komunikacji w standardzie LoRaWAN [8].

2.4. Thethings.io

Kolejną platformą wykorzystywaną w LoRa Thethings.io. Posiada prosty interfejs API (ang. Application Programming Interface) pozwalający użytkownikom na realizację autorskich rozwiązań. Platforma Thethings.io pozwala również na integrację z TTS za pomocą protokołów HTTPS i MQTT [9].

Na podstawie analizy parametrów zestawionych w tabeli 1, zasadnym jest wybór platformy TTN (TTS).

Przedstawione platformy oprócz komunikacji w systemie LoRaWAN, pozwalają na łatwą integrację z popularnymi platformami do przetwarzania, analizowania i wizualizacji danych, takich jak: Thingspeack, Arduino IoT Cloud, AWS IoT, Azure IoT Hub, Google Cloud i Node-Red.

Tab. 1. Charakterystyka platform wykorzystywanych z LoRa. Źródło: [5][11].

	TTN (TTS)	Chirpstack	Open-Stack	Thethings.io
Łatwość wdrożenia własnych rozwiązań	Łatwa konfiguracja. Ogromna społeczność korzystająca z tej platformy, dzięki czemu łatwo znaleźć pomoc. Otwarty kod źródłowy.	Średnio zaawansowana konfiguracja. Wymaga umiejętności pracy ze skryptami.	Konieczność specjalistycznych umiejętności i wiedzy	Średnio zaawansowana konfiguracja.
Koszt	Bezpłatne dla własnego użytku. Dla użytku komercyjnego: €190 / $230 miesięcznie	Bezpłatne dla użytku prywatnego i komercyjnego.	Koszt obliczany na podstawie połączenie poziomu usług i infrastruktury	Opłata zależna od liczby urządzeń i utrzymania platformy. Bezpłatne testowanie na 15 dni.
Obsługiwane protokoły	TCP/IP, UDP, DHCP, MQTT, HTTP, Websockets, AMQP, gRPC	MQTT, HTTP, DHCP, Protobuf Websockets, AMQP, gRPC	TCP/IP, UDP, ICMP, DHCP, ARP	MQTT, HTTP, Websockets, CoAP

3. OPIS SYSTEMU

Na Wydziale Informatyki i Telekomunikacji, Politechniki Poznańskiej, został zaimplementowany system eksperymentalny, służący do przeprowadzenia badań nad stopniem złożoności wybranego serwera sieciowego, dedykowanego dla rozwiązań korzystających z komunikacji w standardzie LoRaWAN.

3.1. Wykorzystane urządzenia i technologie

Do implementacji badanego systemu wykorzystano bramę LoRa – Lorank 8, której specyfikację przedstawiono w Tab. 2, platformę Arduino Uno Rev 3 oraz nadajnik RFM95W (specyfikacja nadajnika w Tab. 3). Platforma Arduino wraz z nadajnikiem jest odpowiedzialna za wysyłanie wiadomości testowych do serwisu TTS.

Nadajnika RFM95W jest wyposażony w moduł LoRa, który umożliwia komunikację o dużym zasięgu, z dużą odpornością na zakłócenia i małym zużyciu energii.

Do realizacji połączenia urządzenia końcowego z TTS poprzez bramę LoRa, wykorzystano bibliotekę LMiC (loRaMAC in C), którą można znaleźć na platformie github [15].

Na rysunku 2 przedstawiono lokalizację ekperymentalnego systemu, który znajduję się w budynku CWiBT (Centrum Wykładowe i Biblioteka Techniczna) Politechniki Poznańskiej, w Poznaniu przy ulicy Piotrowo 2. W projekcie wykorzystano jedną bramkę i jedno urządzenie końcowe. Docelowo zestaw eksperymantalny będzie składał się z około 10 urządzeń końcowych z czujnikami do kontrolowania mikroklimatu w pomieszczeniach wspomnianego budynku.

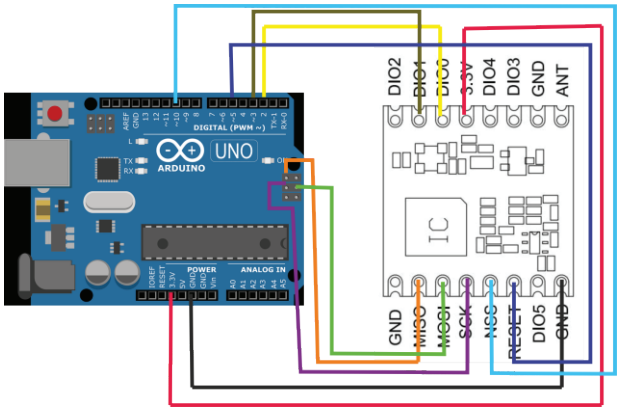
Tab. 2. Parametry wykorzystanej bramki [12].

Pasmo częstotliwości	868 MHz
Czułość	-138 dBm
Moc maksymalna	+27 dBm (500mW)
Demodulatory LoRa	49
Kanały równoległe	8
Maks. połączonych węzłów	~60 tysięcy (przy założeniu, że węzły wysyłają dane raz na godzinę. W innych przypadkach, liczba węzłów, które może obsłużyć około 10-20 tysięcy)
Procesor	1GHz, ARM Cortex A8
OS	Debian / Angstrom Linux
Wi-Fi	Opcjonalnie (przez USB)
Natężenie prądu	1A
Maks. natężenie prądu USB	500mA
Zasilacz sieciowy	5VDC, 2A

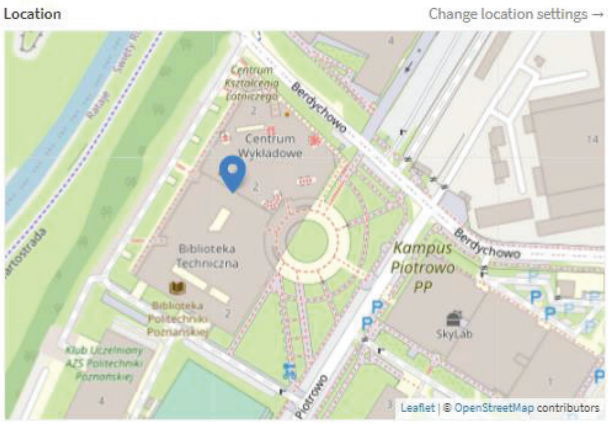
Tab. 3. Najważniejsze parametry nadajnika RFN95W [13].

Maksymalny budżet łącza	168 dB
Modulacja	FSK, GFSK, MSK, GMSK, LoRa™ i OOK
Programowalna przepływność	do 300 kbps
Odporność na zakłócenia	IIP3 = -12,5 dBm
Wysoka czułość	-148 dBm
Wbudowane czujniki	Czujnik temperatury i wskaźnik niskiego poziomu baterii
Stała moc wyjściowa RF (ang. radio frequency - częstotliwość radiowa) w zależności od napięcia zasilania	+20 dBm - 100 mW
Sprawność wzmacniacza mocy	+14 dBm
Niski prąd RX (rezystancja mierzona)	10,3 mA

Na Rys. 1 przedstawiono schemat platformy Arduino Uno R3 z nadajnikiem RFM95W.

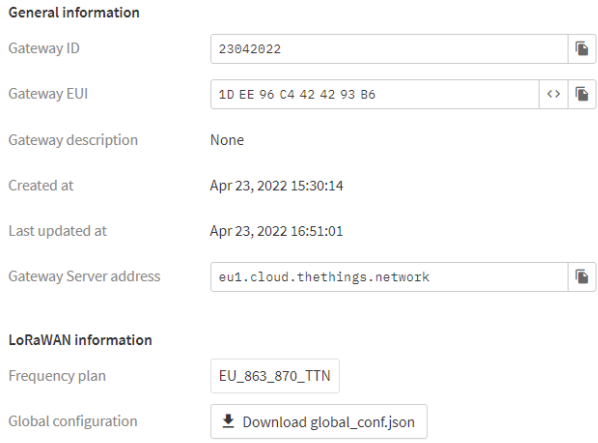


Rys. 1. Schemat połączenia Arduino z RFM95W. Źródło: opracowanie własne na podstawie: [14].



Rys. 2. Lokalizacja zestawu eksperymentalnego. Źródło: Zrzut ekranu pochodzi z panelu administracyjnego na platformie TTS.

Konfigurację na serwerze webowym rozpoczynamy od dodania bramy. W celu dodania bramy do sieci TTS należy pobrać z bramy informację o jej identyfikatorze - Gateway ID. Kluczowym elementem konfiguracji połączenia bramy z serwerem webowym, jest pobranie pliku global_conf.json (Rys. 3).

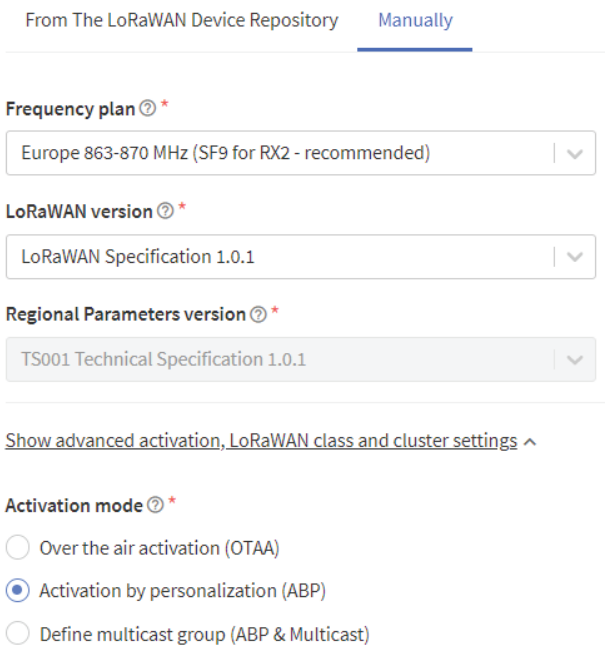


Rys. 3. Dane konfiguracyjne bramy na platformie TTS. Źródło: Zrzut ekranu pochodzi z panelu administracyjnego na platformie TTS.

W przypadku wykorzystanej w projekcie bramy, pojawił się problem z rozróżnianiem wielkości znaków dla

Gateway ID. Konieczna była zmiana parametru Gateway ID w pliku local_conf.json. Drugim problemem był brak synchronizacji zegara systemowego bramki z zegarem serwera. Rozwiązaniem tego problemu było zainstalowanie pakietu NTP (Network Time Protocol – protokół synchronizacji czasu).

Po dodaniu w serwisie bramy, należy zdefiniować urządzenie końcowe. W tym celu najpierw należy dodać w serwisie aplikację i w utworzonej aplikacji dodać konfigurowane urządzenia. Można je dodać z repozytorium LoRaWAN lub wprowadzić dane konfiguracyjne ręcznie (Rys. 4).



Rys. 4. Dane konfiguracyjne podczas definiowania urządzenia końcowego na platformie TTS. Źródło: Zrzut ekranu pochodzi z panelu administracyjnego platformy TTS.

Podczas dodawania urządzeń należy wybrać sposób aktywacji. Domyślnie zaznaczony jest sposób OTAA, który jest rekomendowany i lepszy od ABP, ponieważ klucze sesji są przydzielane dynamicznie i zmieniają się przy każdej nowej sesji. Jednak sposób aktywacji ABP pozwala na lepszą kontrolę nad urządzeniem, dlatego na etapie prototypowania wybrano ten sposób [16].

4. PODSUMOWANIE

W niniejszym artykule przedstawiono analizę serwerów sieciowych do komunikacji w sieci LoRaWAN oraz sposób konfiguracji budowanego systemu eksperymentalnego na wybranym serwerze sieciowym, wykorzystując łatwo dostępne i niedrogie urządzenia oraz oprogramowanie typu open source. Eksperymentalny system będzie w przyszłości rozbudowywany i wykorzystywany do badań w dziedzinie Inteligentnych budynków oraz w dziedzinie bezpieczeństwa komunikacji w sieciach LPWAN.

Praca finansowana ze środków przyznanych przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach zadania nr 0312/SBAD/8162

LITERATURA

[1] Hossein Motlagh Naser, Mohammadrezaei Mahsa, Hunt Julian, Zakeri Behnam. 2020. “Internet of Things (IoT) and the Energy Sector”. *Energies 13*, 494.
[2] LoRa Alliance website. Available online: https://lora-alliance.org/about-lorawan/ (dostęp 20.04.2022).
[3] Adefemi Alimi Kuburat Oyeranti, Ouahada Khmaies, Abu-Mahfouz Adnam M., Rimer Suvendi. 2020. “A Survey on the Security of Low Power Wide Area Networks: Threats, Challenges, and Potential Solutions”. *Sensors 20*, 5800.
[4] Nowak Mariusz. Derbis Piotr, Kurowski Krzysztof, Różycki Rafał, Waligóra Grzegorz. 2021. “LPWAN Networks for Energy Meters Reading and Monitoring Power Supply Network in Intelligent Buildings”. *Energies 14, 7924.*
[5] The ThingNetwork website. Available online: https://www.thethingsnetwork.org/docs/quick-start (dostęp 20.04.2022).
[6] The ThingStack website. Available online: https://www.thethingsindustries.com/docs/ (dostęp 20.04.2022).
[7] Chirpstack website. Available online: https://www.chirpstack.io (dostęp 20.04.2022).
[8] OpenStack website. Available online: https://www.openstack.org/community/ (dostęp 20.04.2022).
[9] TheThings.io website. Available online: https://thethings.io (dostęp 20.04.2022).
[10] Gassara Mouna & Elleuchi Manel, Abid Mohamed. 2021. “Cloud-based platforms for LoRa internet of things: a survey”. *International Journal of Informatics and Communication Technology (IJ-ICT). 10. 54. 10.11591/ijict.v10i1.pp54-64.*
[11] So Jaeyoung, Kim Daehwan, Kim Hongseok, Lee Hyunseok, Park Suwon. 2016. "LoRaCloud: LoRa platform on OpenStack". *IEEE NetSoft Conference and Workshops (NetSoft), pp. 431-434.*
[12] Podręcznik użytkownika bramki Lorank 8. https://github.com/Ideetron/Larank/blob/master/lorank8v1/manual.pdf (dostęp 20.04.2022).
[13] Dokumentacja modułu radiowego RFM95W. https://www.hoperf.com/modules/lora/RFM95.html (dostęp 20.04.2022).
[14] Poradnik łączenia urządzenia końcowego opartego na Arduino i RFM95W z TTN. https://www.mobilefish.com/developer/lorawan/lorawan_quickguide_build_lora_node_rfm95_arduino_uno.html (dostęp 20.04.2022).
[15] Bilblioteka LMiC. https://github.com/mcci-catena/arduino-lmic (dostęp 20.04.2022).
[16] Aktywacja OTAA i ABP – różnice. https://www.thethingsindustries.com/docs/devices/abp-vs-otaa/ (dostęp 20.04.2022).

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

Analiza wpływu bezprzewodowej synchronizacji węzłów oraz liczby lokalizowanych etykiet na skuteczność odbioru pakietów w ultraszerokopasmowym systemie lokalizacyjnym

Analysis of the influence of the wireless anchor node synchronization and number of localized tags on the packet-reception success ratio in the ultrawideband localization system

Vitomir Djaja-Jośko

Politechnika Warszawska, Warszawa, vitomir.djaja-josko@pw.edu.pl

DOI: 10.15199/59.2022.4.90

Streszczenie: **Ostatnimi czasy coraz większą popularnością cieszą się usługi wykorzystujące informacje o położeniu osób oraz obiektów wewnątrz pomieszczeń. Konsekwencją tego jest dynamiczny rozwój systemów dostarczających takich informacji. Ze względu na swoje właściwości, na szczególną uwagę zasługują systemy ultraszerokopasmowe. Jednakże, aby mogły one poprawnie funkcjonować, konieczna jest synchronizacja węzłów tworzących ich infrastrukturę. W referacie przeanalizowano wpływ bezprzewodowej synchronizacji węzłów oraz liczby lokalizowanych etykiet na skuteczność odbioru pakietów. W tekście pokrótce opisano dwie metody bezprzewodowej synchronizacji oraz zamieszczono wyniki badań symulacyjnych.**

Abstract: **Nowadays, a growing interest in the indoor Location Based Services (LBS) can be observed. As a result, there is a lot of research put into the development of the systems capable of delivering indoor localization data. Owing to their capabilities in that matter, ultrawideband systems deserve a significant attention. However, to work properly, anchor nodes comprising the fixed infrastructure need to be synchronized. In the paper the influence of the wireless synchronization and the number of localized tags on the packet reception success rate is analysed. A brief description of two wireless synchronization methods along with simulation results is presented.**

Słowa kluczowe: **bezprzewodowa synchronizacja, IOT, systemy lokalizacyjne, UWB.**

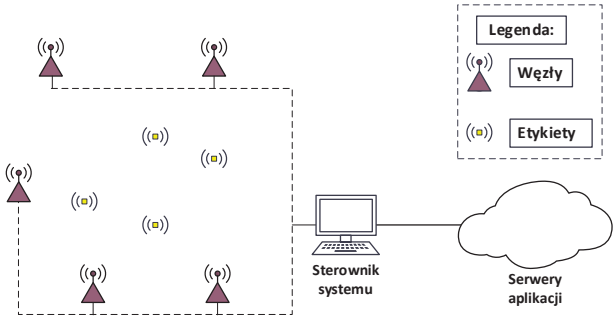
Keywords: **IOT, localization systems, ultrawideband, wireless synchronization.**

1. WSTĘP

Ostatnimi czasy można zaobserwować rosnące zainteresowanie usługami wykorzystującymi do działania informacje o położeniu osób oraz obiektów wewnątrz pomieszczeń. Do licznych przykładów takich rozwiązań można zaliczyć systemy do lokalizacji wózków w magazynach [12], systemy do asystowania pasażerom na lotniskach czy systemy wspomagające osoby starsze w życiu codziennym [14].

Wśród wielu rozwiązań umożliwiających uzyskanie informacji o lokalizacji osób lub obiektów wewnątrz pomieszczeń, takich jak systemy wizyjne, ultradźwiękowe, inercyjne czy laserowe, warto wyróżnić systemy radiowe. Ich cechy takie jak możliwość propagacji sygnałów przez ściany, skalowalność, powszechność czy elastyczność implementacji predestynują je do użycia w systemach lokalizacyjnych. Do wyznaczania położenia wykorzystują one między innymi pomiary poziomu odbieranego sygnału, kąta lub kierunku przybycia sygnału, czasu lub różnicy czasów przybycia sygnałów. Ostatnie z tych metod są wykorzystywane w systemach ultraszerokopasmowych.

Ogólna architektura ultraszerokopasmowego systemu lokalizacyjnego została przedstawiona na rys. 1.



Rys. 1. Ogólna architektura systemu lokalizacyjnego

Składa się ona z lokalizowanych etykiet (pełniących najczęściej rolę prostych nadajników), węzłów tworzących infrastrukturę, sterownika systemu oraz ewentualnie serwerów aplikacji.

Ze względu na to, że w systemach ultraszerokopasmowych do wyznaczania lokalizacji wykorzystywane są wyniki pomiarów czasu (bądź różnicy czasów) przybycia sygnałów, konieczne jest zapewnienie synchronizacji węzłów pomiarowych rozumianej jako:

- ujednolicanie skali czasu dla wszystkich węzłów,
- redukcję wpływu opóźnień wprowadzanych przez węzły oraz niedokładności wzorców częstotliwości oraz nieidealności urządzeń.

2. STAN WIEDZY

W literaturze można znaleźć różne prace dotyczące zagadnienia synchronizacji węzłów w ultraszerokopasmowych systemach lokalizacyjnych. Wśród rozwiązań komercyjnych dominująca wydaje się być synchronizacja przewodowa. Dwa najpopularniejsze systemy dostępne

na rynku, tj. Dart firmy Zebra Technologies [15] oraz Dimension4 firmy Ubisense [2] wykorzystują architekturę, w której sygnały synchronizacyjne są rozprowadzane drogą przewodową od sterownika systemu do węzłów, bezpośrednio, bądź przez połączenie szeregowe (tzw. daisy-chain).

Wśród prac naukowych można wyróżnić rozwiązanie zaproponowane przez M. Kuhn'a, C. Zhang'a, B. Merkl'a i innych, w którym synchronizacja odbywa się z wykorzystaniem impulsów transmitowanych drogą przewodową [5] oraz rozwiązanie opisane przez K. Yu, I. Sharp'a i Y.J. Guo, wykorzystujące węzeł referencyjny połączony przewodowo z węzłami pomiarowymi [4].

Innym podejściem do problemu jest synchronizacja bezprzewodowa. W literaturze można znaleźć wiele rozwiązań tego typu, przy czym najliczniejsze zaliczają się do jednej z dwóch grup: synchronizacji z wyznaczaniem czasu propagacji oraz synchronizacji z węzłem referencyjnym. Dobrymi przykładami należącymi do pierwszej grupy są rozwiązania S. Y. Park'a, H.S. Ahn'a i W. Yu [11], D. Kreiser'a i O. Sonom'a [8] i W. Yan'a, Z. Jing'a i Z. Nailong'a [13]. Wszystkie polegają na wyznaczaniu czasu propagacji między węzłami a etykietami, co wiąże się z koniecznością wymiany znaczącej liczby pakietów.

Do przykładowych prac, w których opisano bezprzewodową synchronizację wykorzystującą węzeł referencyjny, transmitujący pakiety referencyjne można zaliczyć referaty autorstwa C. McElroy'a, D. Neirynck'a i M McLaughlin'a [1], K. Hamahuchi'ego i R. Kohno'a [3], czy J. Tiemann'a, F. Eckermann'a i C. Wietfeld'a [7].

Istotnym problemem systemów wykorzystujących bezprzewodową synchronizację węzłów może być ich ograniczony obszar działania, wynikający z konieczności odbioru pakietów referencyjnych przez wszystkie węzły pomiarowe. Odpowiedzą na to, może być zastosowanie większej liczby węzłów referencyjnych retransmitujących pakiety referencyjne, co opisali autorzy w [9].

Bezprzewodowa synchronizacja może znacząco ułatwić instalację systemu (ze względu na brak konieczności przewodowego łączenia wszystkich urządzeń), ale może również prowadzić do zwiększenia zajętości łącza radiowego, wynikającego z konieczności transmisji pakietów referencyjnych. To z kolei przekłada się na skuteczność odbioru przez węzły pomiarowe pakietów nadawanych przez lokalizowane etykiety. Warto wspomnieć, że ta istotna kwestia często jest pomijana w literaturze.

W referacie przeprowadzono analizę wpływu dwóch metod bezprzewodowej synchronizacji węzłów w ultraszerokopasmowym systemie lokalizacyjnym oraz liczby lokalizowanych etykiet, na skuteczność odbioru pakietów, rozumianą jako stosunek liczby odebranych pakietów, do liczby pakietów nadanych.

W rozdziale trzecim opisano analizowane metody synchronizacji. Rozdział czwarty dotyczy badań symulacyjnych skuteczności odbioru pakietów. Całość została podsumowana w rozdziale piątym.

3. METODY BEZPRZEWODOWEJ SYNCHRONIZACJI WĘZŁÓW

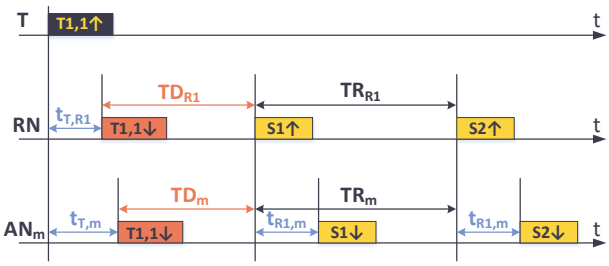
Analizowane metody bezprzewodowej synchronizacji wykorzystują co najmniej jeden węzeł referencyjny umieszczony w znanym położeniu, nadający pakiety referencyjne, odbierane przez węzły pomiarowe. Metody różnią się sposobem transmisji pakietów referencyjnych:

- w metodzie 1 są one wyzwalane odbiorem pakietu z etykiety,
- w metodzie 2 są one nadawane okresowo.

Obie metody zostały opisane w literaturze, między innymi w [6] oraz [9].

3.1. Opis metody 1

Metoda 1 (z wyzwalaną transmisją pakietów referencyjnych) została z założenia opracowana do działania z jedną etykietą (T), która transmituje pakiety. Poglądowy schemat transmisji przedstawiono na rys. 2.



Rys. 2. Schemat transmisji w metodzie 1

Po odbiorze pakietu z etykiety, węzeł referencyjny (RN) retransmituje dwa pakiety. Pierwszy (S1), nadawany z opóźnieniem TD_{R1} jest wykorzystywany do ujednolicenia skali czasu w węzłach pomiarowych, natomiast rolą drugiego (S2) jest umożliwienie kompensacji niedokładności ich wzorców częstotliwości. Okres TR_{R1} pomiędzy pakietami S1 i S2 jest stały, jego wartość jest znana w węzłach pomiarowych (AN_m). TD_m i TR_m to zmierzone odcinki czasu, natomiast $t_{T,R1}$, $t_{T,m}$ i $t_{R1,m}$ to czasy propagacji między poszczególnymi urządzeniami.

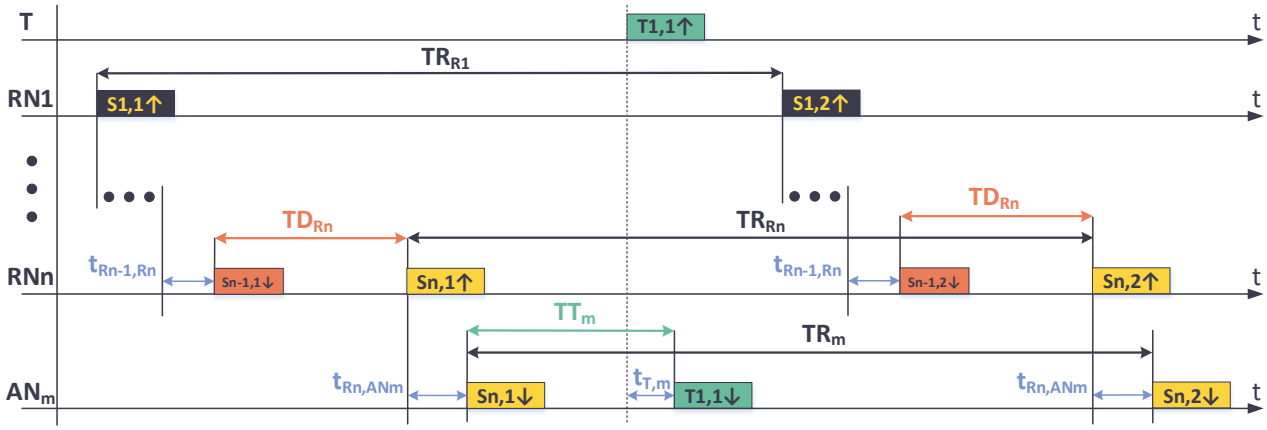
3.2. Opis metody 2

W metodzie 2 transmisja pakietów referencyjnych odbywa się niezależnie od pakietów nadawanych przez etykiety (T). Schemat transmisji został przedstawiony na rys. 3. Pakiety referencyjne (S1,1; S1,2) są wysyłane okresowo (TR_{R1}) przez nadrzędny węzeł referencyjny (RN1). Jeżeli wykorzystywane są dodatkowe podrzędne węzły referencyjne (RNn), kolejno, z predefiniowanymi opóźnieniami (TD_{Rn}) retransmitują one te pakiety. Na podstawie czasów odbioru pakietów referencyjnych oraz zmierzonych odcinków czasu (TT_m, TR_m) węzły pomiarowe (AN_m) są wstanie ujednolicić swoją skalę czasu względem węzła referencyjnego oraz skorygować wpływ opóźnień oraz niedokładności ich wzorców częstotliwości. Czasy propagacji między urządzeniami oznaczono jako $t_{RN-1,RN}$, $t_{Rn,ANm}$ i $t_{T,m}$.

3.3. Podsumowanie metod

Analizując schematy transmisji przedstawione na rys.2 oraz na rys. 3 można dojść do następujących wniosków:

- ze względu na specyfikę metody 1, nie można w niej zbadać skuteczności odbioru pakietów,



Rys. 3. Schemat transmisji w metodzie 2

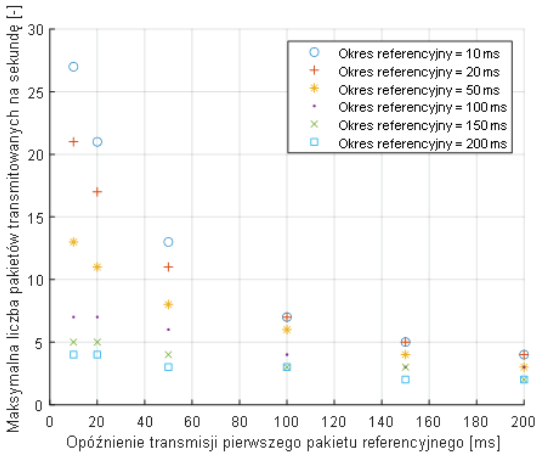
a jedynie maksymalną liczbę pakietów możliwych do nadania w trakcie sekundy, zależną od czasów TD_{R1} i TR_{R1},

- w metodzie 2 wpływ na skuteczność odbioru pakietów wpływ mają czynniki takie jak liczba etykiet, częstość transmisji pakietów, liczba węzłów referencyjnych oraz okres ich transmisji.

4. BADANIA SYMULACYJNE

W badaniach przyjęto, że oprócz metody synchronizacji na skuteczność odbioru pakietów wpływ mają również następujące czynniki:

- czas trwania transmitowanych pakietów – na potrzeby symulacji przyjęto typowe wartości, mogące pojawić się w rzeczywistym systemie lokalizacyjnym, tj. 2 ms dla pakietów referencyjnych i 4,5 ms dla pakietów z etykiet,
- stosunek mocy sygnału pożądanego do mocy sygnałów zakłócających – do jego oceny przyjęto model potęgowy (n = 2) uwzględniający wolne zaniki (σ = 4 dB),
- czas niezbędny na przetworzenie pakietów w węzłach – przyjęto 2,5 ms.



Rys. 4. Maksymalna liczba pakietów transmitowanych na sekundę w metodzie 1

4.1. Badania metody 1

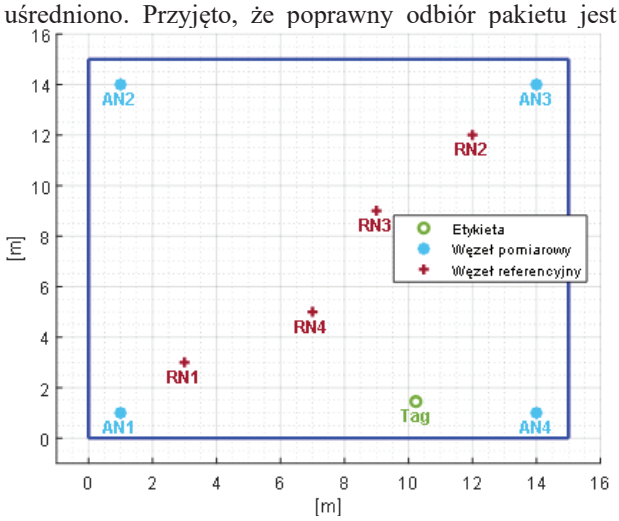
Określenie maksymalnej liczby pakietów nadawanych przez etykietę, przekładające się bezpośrednio na częstość wyznaczania jej położenia, zostało przeprowadzone w funkcji długości opóźnienia transmisji pierwszego pakietu referencyjnego (TD_{R1}) oraz długości okresu referencyjnego (TR_{R1}). Wyniki przedstawiono na rys. 4.

4.2. Badania metody 2

Celem badań metody 2 było określenie skuteczności odbioru pakietów dla różnych warunków pracy systemu takich jak liczba etykiet, ich okres nadawania, liczba węzłów referencyjnych i długość okresu referencyjnego.

Zasymulowano pomieszczenie o rozmiarze 15 x 15 metrów (rys. 5), w którym umieszczono cztery węzły pomiarowe, od jednego do czterech węzłów referencyjnych nadających z okresem od 0,5 s do 2 s oraz od 1 do 10 etykiet nadających z okresami od 0,1 s do 3 s (dodatkowo randomizowanymi poprzez dodanie losowego składnika o rozkładzie jednostajnym z zakresu od 0 do połowy ustawionego okresu).

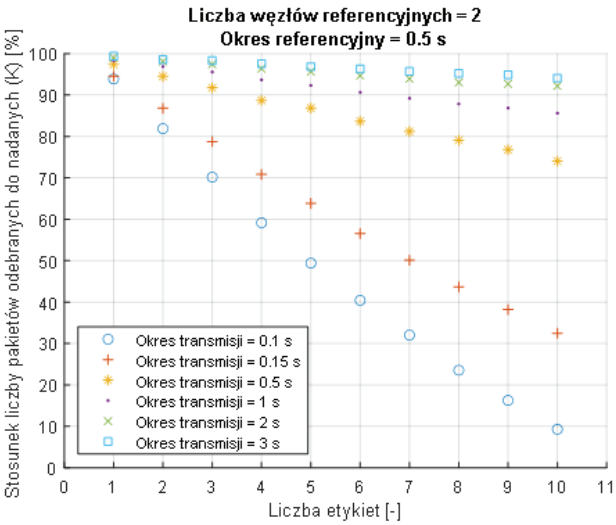
Położenia etykiet były losowane na początku każdej symulacji i założono, że poruszają się one z prędkością 1 m/s. Dla każdej kombinacji parametrów przeprowadzono 100 symulacji, symulowano 30 sekund pracy systemu. Uzyskane wyniki skuteczności odbioru pakietów uśredniono. Przyjęto, że poprawny odbiór pakietu jest



Rys. 5. Rozkład węzłów w symulacji metody 2

możliwy, kiedy dwa pakiety (uwzględniając czas potrzebny na ich przetworzenie) nie nakładają się na siebie w odbiorniku, bądź różnica poziomu odbieranych sygnałów jest większa od 20 dB. Analizowane były wszystkie kombinacje etykiet i węzłów pomiarowych.

Na rys. 6 przedstawiono przykładowe wyniki symulacji.



Rys. 6. Przykładowe wyniki symulacji dla metody 2

Otrzymane wyniki symulacji są zgodne z oczekiwaniami. Kluczowy wpływ na skuteczność odbioru pakietów ma okres transmisji w etykietach. Wraz z jego wzrostem, znacząco maleje skuteczność. Przyczynia się do tego również wzrost liczby etykiet, ale przede wszystkim przy bardzo krótkich okresach (poniżej 1 s). Co istotne – przy długich okresach, rzędu 2-3 s (co jest zupełnie wystarczające dla wielu scenariuszy wykorzystania systemów lokalizacyjnych), liczba etykiet nie ma decydującego wpływu.

Badania wykazały również, że wzrost liczby węzłów referencyjnych oraz ich okres transmisji nie mają decydującego wpływu na skuteczność odbioru pakietów, co przemawia na korzyść bezprzewodowej synchronizacji węzłów. Ze względu na ograniczone miejsce, wyników nie można było zamieścić w pracy, ale można je znaleźć w tekście rozprawy doktorskiej [10].

5. PODSUMOWANIE

W referacie przedstawiono problem wpływu metody bezprzewodowej synchronizacji węzłów na skuteczność odbioru pakietów. Opisano pokrótce metody oraz omówiono badania symulacyjne. Wykazały one, że dominujący wpływ ma liczba etykiet oraz ich okres transmisji, a nie liczba węzłów referencyjnych.

Wyniki trudno odnieść do aktualnego stanu wiedzy, ze względu na to, że większość autorów nie prowadzi badań w tym obszarze.

LITERATURA

[1] McElroy Ciaran, Neirynck Dries, McLaughlin Michael. 2014. "Comparison of wireless clock synchronization algorithms for indoor location systems". *Proc. of the 2014 IEEE International Conference on Communications Workshops (ICC)*: 157-162.

[2] „Dimension4 – Brochure", Ubisense Limited, 2019. https://ubisense.com/wp-content/uploads/2019/12/Dimension-4-Large-Brochure-17Sep2019.pdf.

[3] Hamaguchi Kiyoshi, Kohno Ryuiji. 2008. "Development of experimental TDOA system test-bed for indoor applications". *Proc. of the 2008 IEEE International Conference on Ultra-Wideband*: 201-204.

[4] Yu Kegen, Sharp Ian, Guo Y Jay. 2009. "Ground-Based Wireless Positioning", *Wiley*, Hoboken, NJ, USA.

[5] Kuhn Michael, Zhang Cemin, Merkl Brandin, [i in.]. 2008. "High accuracy UWB localization in dense indoor environments". *Proc. of the 2008 IEEE International Conference on Ultra-Wideband*: 129-132.

[6] Djaja-Jośko Vitomir, Kołakowski Jerzy. 2016. „Metoda bezprzewodowej synchronizacji węzłów i korekcji wyników pomiarów TDOA w ultraszerokopasmowym systemie lokalizacyjnym". *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne, SIGMA NOT.* LXXXVIII (6): 213-216.

[7] Tiemann Janis, Eckermann Fabian, Wietfeld Christian. 2016. "Multi-user interference and wireless clock synchronization in TDOA-based UWB localization". *Proc. of the 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*: 1-6.

[8] Kreiser Dan, Sonom Olonbayar. 2014. "Multiple sensor node synchronization for parallel UWB communication and ranging". *Proc. of the 2014 IEEE International Conference on Ultra-WideBand (ICUWB)*: 398-402.

[9] Djaja-Jośko Vitomir. 2020. „Novel Method for the Wireless Synchronization of the Anchors in the UWB Localization System Utilizing Two Reference Nodes". *Proc. of the 23rd International Microwave and Radar Conference (MIKON 2020)*: 69–73.

[10] Djaja-Jośko Vitomir. 2021, "Nowe metody bezprzewodowej synchronizacji w ultraszerokopasmowych systemach lokalizacyjnych". Rozprawa doktorska. PW.

[11] Park SangYoung, Ahn Hyo-Sung, Yu Wonpil. 2007. "Round-trip time-based wireless positioning without time synchronization". *Proc. of the 2007 International Conference on Control, Automation and Systems*: 2323-2326.

[12] Reinke Christof, Beinschob Patrick. 2013. "Strategies for contour-based self-localization in large-scale modern warehouses". *Proc. of the 2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)*: 223-227.

[13] Yan Wang, Jing Zhao, Nailong Zhang. 2015. "The Designing of Indoor Localization System Based on Self-Organized WSN Using Pulson UWB Sensors". *Proc of the 2015 2nd International Conference on Information Science and Control Engineering*: 965-969.

[14] Vlasenko Iuliia, Nikolaidis Ioanis, Stroulia Eleni. 2015. "The Smart-Condo: Optimizing Sensor Placement for Indoor Localization". *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45 (3): 436-453.

[15] „Zebra UWB Technology – Product Datasheet" ZiH Corp, 2018. https://www.zebra.com/content/dam/zebra_new_ia/en-us/solutions-verticals/product/location_solutions/zebra-ultra-wideband/spec-sheet/uwb-portfolio-spec-sheet-en-us.pdf.

ANALIZA ZANIKÓW SZYBKOZMIENNYCH W RADIOWYCH SIECIACH BAN PRACUJĄCYCH W REWERBERACYJNYM ŚRODOWISKU PROPAGACYJNYM

FAST FADING ANALYSIS FOR BODY AREA NETWORKS
OPERATING IN REVERBERATION PROPAGATION ENVIRONMENT

Sławomir J. Ambroziak[1], Filipe D. Cardoso[2], Paweł Kosz[1], Manuel M. Ferreira[2], Luís M. Correia[3]

[1]Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, ul. Narutowicza 11/12, 80-233 Gdańsk
{slawomir.ambroziak, pawkosz}@pg.edu.pl
[2]EST Setúbal, Polytechnic Institute of Setúbal / INESC-ID, Estefanilha, 2910-761 Setúbal, Portugal
{filipe.cardoso, manuel.ferreira}@estsetubal.ips.pt
[3]Instituto Superior Técnico, University of Lisbon / INESC-ID, Rua Alves Redol, 9, 1000-029 Lisbon, Portugal
luis.m.correia@tecnico.ulisboa.pt

DOI: 10.15199/59.2022.4.91

Streszczenie: **W referacie przedstawiono badania i analizę zaników szybkozmiennych występujących w sieciach BAN, pracujących na częstotliwości 2,45 GHz w środowisku zamkniętym o kształcie cylindrycznym z metalowymi ścianami. Badania przeprowadzono przy użyciu symulatora zweryfikowanego pomiarowo. Uzyskane wyniki wskazują, że w analizowanym środowisku propagacyjnym zaniki szybkozmienne można modelować przy pomocy rozkładu Rayleigha, nawet w przypadku występowania warunków bezpośredniej widoczności anten łącza radiowego.**

Abstract: **In the paper the research and analysis of fast fading in Body Area Networks operating at 2.45 GHz in indoor environment with cylindrical shape and metallic walls has been presented. The investigations have been performed with the use of simulator that has been empirically validated. The results shows that fast fading may be modelled by a Rayleigh distribution, even when line-of-sight conditions occurs.**

Słowa kluczowe: **sieci BAN, charakterystyka zaników szybkozmiennych, modelowanie propagacji fal radiowych.**

Keywords: **Body Area Networks, fast fading characterization, propagation modelling.**

1. WSTĘP

Coraz większe zainteresowanie aplikacjami monitorującymi zdrowie człowieka powoduje, że uwaga naukowców i projektantów systemów bezprzewodowych coraz częściej kierowana jest na radiowe sieci BAN (Body Area Networks). Sieci te zbudowane są ze stosunkowo małych, energooszczędnych urządzeń, które mogą być umieszczone wewnątrz ciała ludzkiego, na jego powierzchni, bądź w bezpośrednim sąsiedztwie, i które zwykle mają na celu monitorowanie parametrów życiowych użytkownika oraz ich przesyłanie między sobą, a także – za pośrednictwem węzła koordynującego – do urządzenia zewnętrznego [1].

Zastosowanie sieci BAN na statkach i promach pasażerskich daje możliwość zwiększenia komfortu oraz bezpieczeństwa pasażerów i pracowników podczas rejsów, co ze strony armatorów powoduje zainteresowanie takimi ich aplikacjami, które umożliwiałyby monitorowanie stanu zdrowia, dostarczałyby informacji o położeniu osób na statku/promie, a nawet świadczyłyby usługi rozrywkowe. Środowisko propagacyjne, w którym takie sieci miałyby pracować – przez wzgląd na metalowe ściany, sufity i podłogi – należy uznać za nietypowe, wyjątkowo trudne i mające charakter rewerberacyjnego środowiska propagacyjnego. W takim stanie rzeczy istnieje potrzeba prowadzenia badań mających na celu zrozumienie i opisanie za pomocą statystycznych modeli właściwości kanału radiowego dla tego typu środowisk.

W literaturze przedmiotu znaleźć można liczne prace dotyczące modelowania kanału radiowego w sieciach BAN pracujących w różnych środowiskach [2-4], szczególnie dla zastosowań medycznych [5,6], jednakże większość z nich dotyczy łatwo dostępnych typowych środowisk wewnątrzbudynkowych, takich jak sala szpitalna, pomieszczenie biurowe, czy też komora bezodbiciowa. Dostępne są również nieliczne prace dotyczące badań propagacyjnych na pokładach statków [7-11], jednakże nie uwzględniają one specyfiki sieci BAN.

Pierwsze w skali światowej badania dotyczące sieci BAN w trudnym środowisku promu pasażerskiego zostały opisane w [12,13], gdzie zaprezentowano badania pomiarowe, których wyniki posłużyły do opracowania empirycznych modeli tłumienia propagacyjnego oraz zaników dla sieci BAN pracujących na częstotliwościach 868 MHz, 2,45 GHz i 6,489 GHz [14,15]. Modele te dotyczą jednak korytarzy i kabin pasażerskich, podczas gdy w niniejszym artykule nacisk położono na pomieszczenie o kształcie cylindrycznym z hemisferycznym sufitem. Nietypowy kształt tego środowiska oraz fakt wykonania jego ścian, sufitu i podłogi z materiałów przewodzących powoduje, że istnieje potrzeba dokładnego zbadania warunków propagacyjnych w nim występujących.

W rozdziale 2 scharakteryzowano propagację fal radiowych wewnątrz metalowych struktur kolistych. Rozdział 3 zawiera opis wyników badań pomiarowych, natomiast w rozdziałach 4 i 5 przedstawiono analizę wpływu wielokrotnych odbić, wymiarów środowiska i położenia nadajnika na zaniki szybkozmienne.

2. PROPAGACJA FAL RADIOWYCH WEWNĄTRZ METALOWYCH STRUKTUR KOLISTYCH

Na wstępie przeanalizowano propagację fal radiowych wewnątrz struktur kolistych z idealnie przewodzącymi ścianami, przy czym nacisk położono na sytuacje, w których zaobserwować można różną liczbę odbić składowych odbitych sygnału radiowego w zależności od wzajemnego położenia nadajnika i odbiornika wewnątrz tej struktury. Propagacja wewnątrz struktur metalowych jest zwykle przedstawiana w kontekście falowodów, czy tez wnęk rezonansowych [16], których fizyczne rozmiary są rzędu długości fali, podczas gdy dla analizowanej w niniejszym referacie częstotliwości (2,45 GHz) długość fali, wynosząca 12,2 cm, jest zdecydowanie mniejsza niż wymiary analizowanego pomieszczenia o promieniu 8 m.

Wobec powyższego w analizach rozważana jest propagacja w wolnej przestrzeni z uwzględnieniem odbić fali radiowej od przewodzących ścian, dla których współczynnik odbicia jest równy (bądź bardzo zbliżony) do -1. W takich warunkach powierzchniowa gęstość mocy składowych odbitych docierających do anteny odbiorczej przyjmuje wartości porównywalne z powierzchniową gęstością mocy składowej bezpośredniej – przy założeniu porównywalnej długości drogi propagacyjnej dla obu składowych. Dużego znaczenia, z punktu widzenia natężenia pola elektrycznego sygnału odbieranego, nabiera natomiast liczba (krotność) odbić poszczególnych składowych odbitych. Dla ustalonego wzajemnego położenia nadajnika i odbiornika natężenie pola sygnału odbieranego jest sumą wektorową składowej bezpośredniej i składowych odbitych.

Natężenie pola elektrycznego, E_r, w punkcie odbioru, przy założeniu zastosowania anten izotropowych, uwzględnieniu R_m składowych odbitych i maksymalnej krotności odbić równej n_r, można wyrazić poprzez [21]

$$E_{r\left[\frac{\mathrm{V}}{\mathrm{m}}\right]} = \sum_{i=1}^{n_r}\sum_{j=1}^{R_m} \frac{\sqrt{30\cdot P_{t[\mathrm{W}]}}}{d_{i,j[\mathrm{m}]}} \Gamma_i e^{-j\frac{2\pi d_{i,j[\mathrm{m}]}}{\lambda_{[\mathrm{m}]}}}, \tag{1}$$

gdzie $d_{i,j}$ to długość drogi propagacyjnej dla j-tej składowej i dla i-krotnego odbicia, P_t to moc nadawana, λ to długość fali, a Γ_i to współczynnik odbicia dla i-krotnego odbicia. Natomiast moc sygnału odbieranego można przy takich założeniach wyrazić w sposób następujący [17]

$$P_{r[\mathrm{dBW}]} = 132{,}8 + 20\log\left|E_{r[\mathrm{V/m}]}\right| - 20\log f_{[\mathrm{Hz}]}. \tag{2}$$

Analiza odbić wewnątrz struktur kołowych przedstawiona w [18,19] pokazuje, że dla dwóch punktów wewnątrz koła obliczenia przy uwzględnieniu pojedynczych odbić są stosunkowo łatwe. Jednakże uwzględnienie więcej niż jednokrotnego odbicia każdej składowej wymaga dużo większej złożoności obliczeniowej.

Z tego względu w niniejszej pracy zastosowano podejście iteracyjne, na potrzeby którego przyjęto założenie, że odbiornik umieszczony jest w punkcie $R(x_r,y_r)$, a nadajnik w punkcie $T(x_t,y_t)$ wewnątrz koła o promieniu r. Kiedy rozważane są składowe odbite tylko jeden raz, odcięte, x_i, punktów odbicia na brzegu koła mogą być wyznaczone jako pierwiastki następującego wielomianu 4-tego stopnia [18]

$$y = \sum_{p=0}^{4} c_p x^p, \tag{3}$$

przy czym:

$$c_0 = vx_r^2 - vr^2\left(1 + 2wx_r + w^2x_r^2\right), \tag{4a}$$

$$c_1 = 4vx_r + 2wvx_r^2 + 2w, \tag{4b}$$

$$c_2 = v\left(1 + 2wx_r\right) + w^2 - 2u, \tag{4c}$$

$$c_3 = -2u\left(vr + w\right), \tag{4d}$$

$$c_4 = u^2, \tag{4e}$$

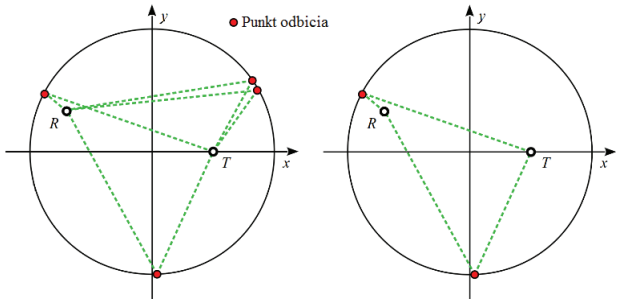
gdzie:

$$u = \frac{2}{r^2}, \qquad v = \frac{1}{x_r^2 + y_r^2}, \qquad w = \frac{1}{x_t}. \tag{5}$$

Rzędne, y_i, punktów odbicia na brzegu koła można następnie wyznaczyć z następującej zależności

$$y_i = y_r \frac{2x_t x_i^2 - r^2\left(x_i + x_t\right)}{2x_t x_r x_i - r^2\left(x_r + x_t\right)}. \tag{6}$$

Na rys. 1. przedstawiono w sposób graficzny wpływ wzajemnego położenia nadajnika i odbiornika wewnątrz struktury kolistej na liczbę punktów odbicia, a – co za tym idzie – na liczbę składowych odbitych. Jak można zauważyć mogą wystąpić dwie sytuacje, w których istnieją cztery lub dwa punkty odbicia.

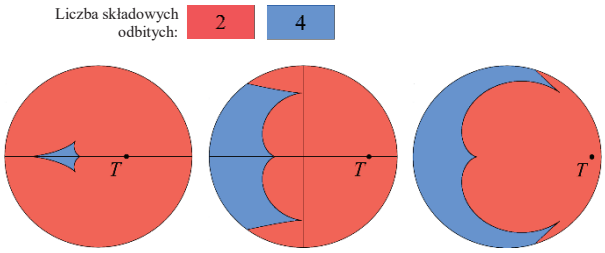


Rys. 1. Ilustracja wpływu wzajemnego położenia nadajnika i odbiornika na liczbę składowych odbitych [21]

Taki stan rzeczy wynika z faktu, że rozwiązaniem równania (3) mogą być albo cztery liczby rzeczywiste i wtedy występują cztery składowe odbite albo para liczb rzeczywisty i para liczb zespolonych, co skutkuje tylko dwoma składowymi odbitymi. Obszary wewnątrz struktury kolistej, w których występują różne liczby składowych odbitych, nazywane są obszarami katakaustycznymi [18].

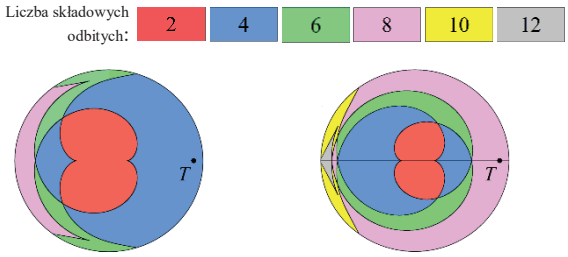
Na rys. 2. zaprezentowano obszary katakaustyczne dla różnego położenia nadajnika i przy uwzględnieniu tylko jednokrotnych odbić. Fakt istnienia tych obszarów implikuje różne warunki propagacyjne w nich występujące, co może mieć istotne znaczenie podczas charakteryzowania m.in. zaników szybkozmiennych.

Z kolei na rys. 3. zaobserwować można obszary katakaustyczne powstałe przy uwzględnieniu również odbić dwukrotnych i trzykrotnych oraz dla nadajnika w punkcie $T(0{,}9r;\ 0)$. Dla przypadku odbić dwukrotnych można wyróżnić obszary, w których występują 2, 4, 6 i 8 składowych odbitych, natomiast dla przypadku odbić trzykrotnych można dodatkowo wyróżnić obszary, w których występuje 10 i 12 składowych odbitych.



(a) $T(0{,}3r;\ 0)$ (b) $T(0{,}7r;\ 0)$ (c) $T(0{,}9r;\ 0)$

Rys. 2. Obszary katakaustyczne przy uwzględnieniu odbić jednokrotnych [21]



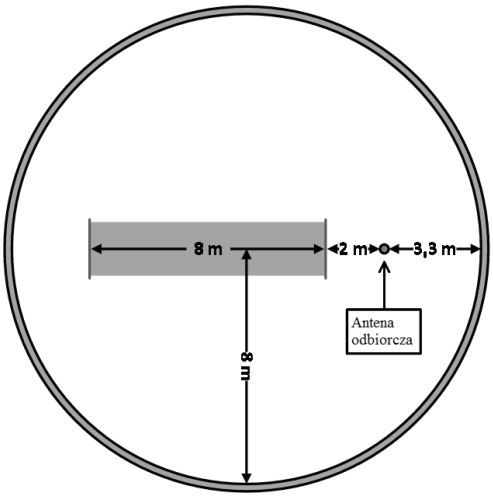
(a) dla odbić dwukrotnych (b) dla odbić trzykrotnych

Rys. 3. Obszary katakaustyczne przy uwzględnieniu odbić dwukrotnych i trzykrotnych [21]

3. WYNIKI BADAŃ POMIAROWYCH

Badania pomiarowe zrealizowano na pokładzie ósmym w pomieszczeniu dyskoteki promu pasażerskiego M/F Wawel. Należy przypomnieć, że elementy konstrukcyjne pomieszczenia wykonane są ze stali. Badania skupiały się na pomiarach tłumienia fal radiowych o częstotliwości 2,45 GHz w łączu pomiędzy węzłem znajdującym się na ciele człowieka (on-body) a węzłem zewnętrznym (off-body). Pomiary zrealizowano przy zastosowaniu stanowiska pomiarowego i metodyki opisanej w [13,14,20] dla przypadku montażu anteny nadawczej na ciele mężczyzny (1,72 m, 60 kg) oraz dla trzech typowych miejsc jej montażu: prawa strona głowy (HE_R), przód tułowia (TO_F) i lewy nadgarstek (AB_L). Jako antenę odbiorczą zastosowano dwupolaryzacyjną antenę tubową. Na rys. 4. przedstawiono rzut badanego środowiska wraz z miejscem instalacji anteny odbiorczej i zacienionym obszarem poruszania się użytkownika.

Podczas pomiarów zrealizowano trzy scenariusze dynamiczne: zbliżania (Z), oddalania (O) oraz pętli, przy czym ostatni scenariusz nie jest rozpatrywany w tym artykule. Scenariusz Z polegał na rozpoczęciu marszu w odległości 10 m od anteny odbiorczej i zatrzymaniu się w odległości 2 m. Z kolei scenariusz O stanowił sytuację odwrotną. Podczas realizacji powyższych scenariuszy jedyną przeszkodą zaburzającą linię bezpośredniej widoczności obu anten łącza radiowego było ciało ludzkie. Warunki propagacyjne były zatem uzależnione od orientacji użytkownika w środowisku, przy czym można wyróżnić trzy następujące przypadki: LoS (Line-of-Sight) – kiedy główny kierunek promieniowania anteny nadawczej był zwrócony w stronę anteny odbiorczej; NLoS (Non-Line-of-Sight) – kiedy główny kierunek promieniowania anteny nadawczej był zwrócony w stronę przeciwną; oraz QLoS (Quasi-Line-of-Sight) – kiedy główny kierunek promieniowania anteny nadawczej był ortogonalny względem głównego kierunku promieniowania anteny odbiorczej.



Rys. 4. Rzut badanego środowiska propagacyjnego

W tab. 1. przedstawiono estymaty parametrów rozkładów prawdopodobieństwa dla głębokości zaników szybkozmiennych, przy czym wzięto pod uwagę dwa rozkłady: Rice'a (z parametrem s_{Ri} i parametrem skali σ_{Ri}) oraz Rayleigha (z parametrem skali σ_{Ra}), a podczas wyznaczania estymat parametrów rozkładów zastosowano metodę największej wiarygodności.

Tab. 1. Parametry rozkładów gęstości prawdopodobieństwa dla głębokości zaników szybkozmiennych (na podstawie pomiarów)

			Rice			Rayleigh
			s_{Ri}	σ_{Ri}	K [dB]	σ_{Ra}
LoS	Z	TO_F	0,98	0,53	2,3	0,87
QLoS	Z	HE_R	0,02	1,00	-37,0	1,00
		AB_L	0,65	0,86	-5,4	0,97
	O	HE_R	0,07	1,04	-26,4	1,04
		AB_L	0,03	1,00	-33,5	1,00
NLoS	O	TO_F	0,08	0,98	-24,8	0,98

Jak można zaobserwować współczynnik Rice'a, K, jest dodatni tylko dla warunków LoS, dla których wynosi 2,3 dB. Dla warunków QLoS parametr ten zmienia się w zakresie od -37 do -5,4 dB, przy czym największą wartość przyjmuje dla scenariusza Z/AB_L, w którym ruch anteny zamontowanej na nadgarstku powoduje cykliczną zmianę warunków propagacyjnych. Z kolei dla warunków NLoS współczynnik Rice'a wynosi -24,8 dB. Przyczyną tak małych wartości tego współczynnika jest rewerberacyjny charakter badanego środowiska, którego elementy

konstrukcyjne wykonane są z materiałów przewodzących. W takim środowisku moc składowej bezpośredniej oraz składowych odbitych (nawet wielokrotnie) są na porównywalnym poziomie.

4. WPŁYW KROTNOŚCI ODBIĆ NA ZANIKI SZYBKOZMIENNE

W celu zbadania wpływu krotności odbić sygnału radiowego na parametry rozkładów prawdopodobieństwa głębokości zaników szybkozmiennych opracowano symulator, który umożliwia określenie mocy sygnału odbieranego dla dowolnego położenia odbiornika wewnątrz dwuwymiarowej struktury kołowej ograniczonej przewodzącymi ścianami, zgodnie z opisem w rozdziale 2. Symulator zaimplementowano przy użyciu oprogramowania Matlab®. Należy przy tym mieć na uwadze fakt, że w rzeczywistości ściany pomieszczenia nie są idealnymi przewodnikami, co – jak wiadomo – ma wpływ na wartość współczynnika odbicia. Ponadto dynamika rzeczywistych scenariuszy, obecność elementów wyposażenia (np. stoły, lampy, itp.) i innych osób w badanym środowisku, hemisferyczny sufit, a także charakterystyki zastosowanych anten nie zostały uwzględnione w symulatorze. Wynika to z faktu, że na obecnym etapie prowadzonych prac, celem było opracowanie możliwie prostego narzędzia symulacyjnego, umożliwiającego przeprowadzenie analiz, przy jednoczesnym odwzorowaniu warunków rzeczywistych na akceptowalnym poziomie.

Badania symulacyjne przeprowadzono z uwzględnieniem maksymalnej krotności odbić od 1 do 5 oraz dla struktury kołowej o promieniu 8 m i dla częstotliwości 2,45 GHz. Położenie zarówno nadajnika jak i odbiornika wewnątrz koła wybierane było losowo, a cały proces powtarzany był 6000 razy. Parametry rozkładów prawdopodobieństwa głębokości zaników szybkozmiennych również i w tym przypadku wyznaczono przy pomocy metody największej wiarygodności.

W tab. 2. przedstawione zostały wyniki symulacji. Jak należało się spodziewać dla warunków LoS współczynnik Rice'a rośnie od -12,1 do 2,9 dB wraz ze zmniejszaniem maksymalnej krotności odbić. Wynika to z faktu, że im mniej składowych odbitych jest uwzględnianych podczas obliczania całkowitej mocy odbieranej, tym mniejsza jest wartość parametru skali σ_{Ri} i tym większy współczynnik K. W warunkach NLoS, czyli bez uwzględnienia składowej bezpośredniej, zmiany współczynnika Rice'a mają podobny charakter, przy czym przyjmują zdecydowanie niższe wartości, tj. od -32,8 do 0,4 dB.

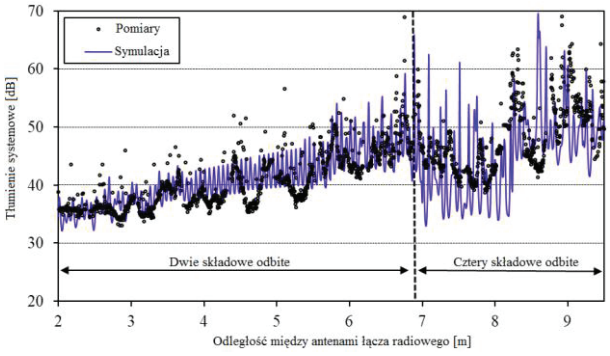
Porównując wyniki symulacji z wynikami pomiarów można zauważyć, że dla warunków LoS empiryczna wartość współczynnika Rice'a znajduje się pomiędzy wartościami symulacyjnymi uzyskanymi dla n_r=1 i n_r=2. Z kolei dla NLoS wartości empiryczne zbliżone są do symulacyjnych dla n_r=2, a różnica wynosi 5,2 dB, przy czym dla wartości na poziomie -25 i -30 dB różnica ta jest nieistotna. Przyczyną większych wartości współczynnika K uzyskanych na podstawie danych empirycznych jest fakt, że podczas pomiarów zastosowano anteny kierunkowe, co miało wpływ na ograniczenie liczby składowych odbitych docierających do odbiornika.

Tab. 2. Parametry rozkładów gęstości prawdopodobieństwa dla głębokości zaników szybkozmiennych (na podstawie symulacji)

			Maksymalna krotność odbić, n_r				
			1	2	3	4	5
LoS	Rayleigh	σ_{Ra}	0,84	0,86	0,88	0,89	0,89
	Rice	s_{Ri}	0,97	0,92	0,85	0,53	0,29
		σ_{Ri}	0,49	0,57	0,64	0,76	0,83
		K [dB]	2,9	1,1	-0,5	-6,1	-12,1
NLoS	Rayleigh	σ_{Ra}	0,85	0,89	0,91	0,93	0,93
	Rice	s_{Ri}	0,87	0,04	0,03	0,03	0,03
		σ_{Ri}	0,59	0,89	0,91	0,93	0,93
		K [dB]	0,4	-30,0	-32,6	-32,8	-32,8

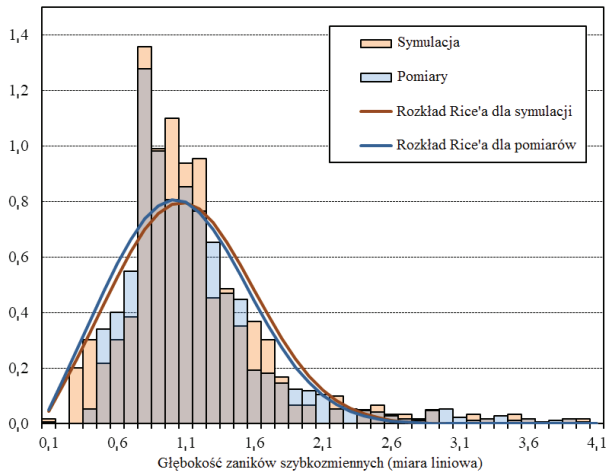
W ogólności można stwierdzić, że w analizowanym środowisku odbicia dwukrotne (obok jednokrotnych) pełnią bardzo istotną rolę, co wydaje się być oczywiste, biorąc pod uwagę fakt, że ściany wykonane są z materiałów przewodzących. Pomimo faktu, że środowisko symulacyjne jest uproszczone do dwóch wymiarów, a co za tym idzie nie uwzględnia odbić od podłogi i hemisferycznego sufitu, można wysnuć wniosek, że zaniki szybkozmienne w metalowym cylindrycznym środowisku propagacyjnym można modelować z zadowalającą dokładnością przy uwzględnieniu maksymalnie dwukrotnych odbić.

Na rys. 5. przedstawiono przykładowe przebiegi tłumienia systemowego uzyskane w wyniku pomiarów i symulacji dla scenariusza Z/TO_F, czyli dla warunków LoS [21]. Jak można zauważyć istnieje duża korelacja między oboma zestawami danych. Na szczególne podkreślenie zasługuje fakt, że w obu przypadkach można wyraźnie zidentyfikować obszary katakaustyczne, w których występują dwie i cztery składowe odbite.



Rys. 5. Przykładowe przebiegi tłumienia systemowego uzyskane w wyniku pomiarów i symulacji (Z/TO_F, LoS)

Rozkłady gęstości prawdopodobieństwa głębokości zaników szybkozmiennych wyznaczonych na podstawie wyżej przedstawionych danych symulacyjnych i pomiarowych zaprezentowano na rys. 6. Uzyskane wartości współczynnika Rice'a wynoszą 0,7 i 1,2 dB, odpowiednio dla pomiarów i symulacji. Są to wartości bardzo zbliżone do tych uzyskanych dla całego zestawu danych pomiarowych (patrz tab. 1., LoS). Również i to porównanie pokazuje, że wystarczające jest uwzględnienie w symulatorze maksymalnej krotności odbić wynoszącej 2.



Rys. 6. Porównanie rozkładów gęstości prawdopodobieństwa dla pomiarów i symulacji (Z/TO_F, LoS) [21]

5. WPŁYW PROMIENIA STRUKTURY KOŁOWEJ I OBSZARU KATAKAUSTYCZNEGO NA ZANIKI SZYBKOZMIENNE

W kolejnym etapie prowadzonych prac przeanalizowano wpływ unormowanego względem długości fali promienia struktury kołowej, r/λ, oraz obszaru katakaustycznego, w którym znajduje się odbiornik, na parametry zaników szybkozmiennych. Symulacje przeprowadzono dla unormowanego promienia wynoszącego 33, 49, 66, 82 i 98, co dla rozpatrywanej częstotliwości odpowiada fizycznemu promieniowi o wartościach odpowiednio 4 m, 6 m, 8 m, 10 m i 12 m. Dla każdego położenia nadajnika wyznaczone zostały (z uwzględnieniem maksymalnej krotności odbić równej 2) obszary katakaustyczne, w których występują dwie lub cztery składowe odbite. W tab. 3. zestawione zostały parametry rozkładów gęstości prawdopodobieństwa głębokości zaników szybkozmiennych dla różnych wartości unormowanego promienia, warunków propagacyjnych i obszarów katakaustycznych.

Tab. 3. Parametry rozkładów gęstości prawdopodobieństwa dla różnych wartości unormowanego promienia, warunków propagacyjnych i obszarów katakaustycznych

			Rice			Rayleigh
		r/λ	s_{Ri}	σ_{Ri}	K [dB]	σ_{Ra}
LoS	Dwie składowe odbite	33	0,96	0,52	2,3	0,85
		49	0,96	0,50	2,7	0,85
		66	0,96	0,50	2,7	0,84
		82	0,96	0,49	2,8	0,84
		98	0,97	0,48	3,1	0,83
	Cztery składowe odbite	33	0,86	0,63	-0,3	0,88
		49	0,89	0,60	0,4	0,87
		66	0,90	0,59	0,7	0,87
		82	0,92	0,57	1,1	0,86
		98	0,93	0,55	1,6	0,85
NLoS	Dwie składowe odbite	33	0,08	0,94	-24,4	0,91
		49	0,07	0,91	-25,3	0,88
		66	0,04	0,91	-30,1	0,88
		82	0,05	0,89	-28,0	0,86
		98	0,04	0,89	-30,0	0,86
	Cztery składowe odbite	33	0,05	0,95	-28,6	0,95
		49	0,05	0,91	-28,2	0,91
		66	0,04	0,90	-30,1	0,90
		82	0,04	0,88	-29,9	0,88
		98	0,04	0,87	-29,8	0,87

Dla warunków LoS i obszaru z dwiema składowymi odbitymi współczynnik Rice'a zmienia się w zakresie od 2,3 do 3,1 dB i nieznacznie rośnie wraz ze wzrostem promienia unormowanego – jednakże wzrost ten wynosi zaledwie 0,8 dB, co jest wartością nieistotną. Z kolei dla obszaru z czterema składowymi odbitymi współczynnik ten również nieznacznie rośnie od -0,3 do 1,6 dB i jest od 2,6 do 1,5 dB mniejszy niż dla obszaru z dwiema składowymi odbitymi, przy czym różnica ta maleje wraz ze wzrostem unormowanego promienia.

Analizując wyniki dla warunków NLoS można zaobserwować, że dla obszaru, w którym występują dwie składowe odbite współczynnik Rice'a zmienia się od -30,1 do -24,4 dB. Dla obszaru z czterema składowymi odbitymi wartości tego współczynnika są nieznacznie mniejsze niż dla obszaru z dwiema składowymi i zmieniają się w zakresie od -30,1 do -28,2 dB. Jak wiadomo dla współczynników Rice'a poniżej -3 dB rozkłady gęstości prawdopodobieństwa Rice'a i Rayleigha praktycznie się pokrywają, dlatego zaniki w tym przypadku mogą być modelowane drugim z wymienionych rozkładów, którego zastosowanie wymaga mniejszej złożoności obliczeniowej przez wzgląd na mniejszą liczbę parametrów.

Biorąc pod uwagę rozmiar obszaru katakaustycznego z czterema składowymi odbitymi, w którym zaniki szybkozmienne są bardziej dotkliwe, można zauważyć, że zmniejsza się on wraz ze zmniejszaniem się odległości nadajnika od środka struktury kołowej (patrz rys. 2.). Im mniejszy jest ten obszar, tym mniejsze jest prawdopodobieństwo, że odbiornik znajdzie się w miejscu, w którym występują cztery składowe odbite. Mając powyższe na względzie można stwierdzić, że umieszczenie nadajnika w środku pomieszczenia o cylindrycznym kształcie spowoduje występowanie tylko jednego obszaru katakaustycznego z dwiema składowymi odbitymi, co pozwoli uzyskać największe wartości współczynnika Rice'a.

6. PODSUMOWANIE

W referacie przedstawiono zagadnienia związane z propagacją fal radiowych wewnątrz metalowych struktur kołowych, ze szczególnym uwzględnieniem zaników szybkozmiennych. Zaproponowano uproszczony model symulacyjny, którego dokładność została zweryfikowana empirycznie na podstawie wyników pomiarów przeprowadzonych w pomieszczeniu dyskoteki o kształcie cylindrycznym, znajdującej się na promie pasażerskim.

Zaobserwowano, że w środowisku tego typu, noszącym cechy środowiska rewerberacyjnego (przez wzgląd na przewodzące ściany, sufit i podłogę), bardzo istotne są składowe dwukrotnie odbite. W ujęciu globalnym, dla wszystkich scenariuszy współczynnik Rice'a rozkładu gęstości prawdopodobieństwa głębokości zaników szybkozmiennych przyjmuje stosunkowo małe wartości i zmienia się w zakresie od -30,1 dB (dla NLoS) do 3,1 dB (dla LoS). Uwzględniając fakt, że w sieciach BAN sytuacja bezpośredniej widoczności anten występuje niezwykle rzadko, co ma bezpośredni związek z ruchem użytkownika, można oczekiwać, że w rzeczywistych warunkach uśredniona wartość współczynnika Rice'a będzie

osiągać wartości znacznie poniżej 3,1 dB. Powyższe pozawala stwierdzić, że zaniki szybkozmienne w takich warunkach mogą być z wystarczającą dokładnością modelowane za pomocą rozkładu Rayleigha, do opisu którego wystarczający jest jeden zamiast dwóch parametrów.

Zauważono również, że najlepszym z punktu widzenia planowania sieci miejscem instalacji nadajnika w analizowanym środowisku jest jego środek. Rozpatrując bilans energetyczny łącza radiowego i kwestię pokrycia radiowego wydaje się to być rzeczą oczywistą. Jednakże, jak dowiedziono w referacie, jest to również istotne z punktu widzenia wpływu zaników szybkozmiennych na moc sygnału odbieranego.

Przyszłe prace dotyczyć będą przeprowadzenia badań pomiarowych mających na celu ocenę wpływu różnych scenariuszy propagacyjnych, różnych charakterystyk promieniowania anten i różnego charakteru ruchu użytkowników na zaniki szybkozmienne oraz opracowanie symulatora uwzględniającego powyższe parametry.

7. PODZIĘKOWANIA

Zaprezentowane w referacie badania prowadzone były w ramach COST Action CA20120, "*Intelligence-Enabling Radio Communications for Seamless Inclusive Interactions*" (INTERACT), a także w ramach porozumienia o współpracy pomiędzy Wydziałem Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej, a Instituto Superior Técnico, University of Lisbon.

LITERATURA

[1] A. Reichman, J. Takada, "Body communications", in R. Verdone and A. Zanella (eds.), *Pervasive Mobile and Ambient Wireless Communications*, Springer, Londyn, UK, 2012.

[2] IEEE, "Channel Model for Body Area Network", IEEE P802.15 Working Group for Wireless Personal Area Networks, Nowy Jork, USA, 2009.

[3] D.B. Smith i inni, "Propagation Models for Body Area Networks: A Survey and New Outlook", *IEEE Antennas and Propagation Magazine*, vol. 55, nr 5, 2013.

[4] S.J. Ambroziak, L.M. Correia, K. Turbic, "Radio channel measurements in body-to-body communications in different scenarios," *URSI Asia-Pacific Radio Science Conference (URSI AP-RASC)*, Seul, Korea Południowa, 2016.

[5] P.A. Catherwood, W. Scanlon, "Link Characteristics for an Off-Body UWB Transmitter in a Hospital Environment", *Loughborough Antennas & Propagation Conference*, Loughborough, UK, 2009.

[6] P. Cui i inni, "Measurement and Modeling of Wireless Off-Body Propagation Characteristics Under Hospital Environment at 6–8.5 GHz", *IEEE Access*, vol. 5, str. 10915-10923, 2017.

[7] H. Kdouh i inni, "Measurements and path loss models for shipboard environments at 2.4 GHz", *41st European Microwave Conference*, Manchester, UK, 2011.

[8] A. Mariscotti, M. Sassi, A. Qualizza, M. Lenardon, "On the propagation of wireless signals on board ships", *IEEE Instrumentation & Measurement Technology Conference*, Austin, USA, 2010.

[9] X.H. Mao, Y. H. Lee, B.C. Ng, "Wideband channel characterization along a lift shaft on board a ship", *IEEE Antennas and Propagation Society International Symposium*, Toronto, Kanada, 2010.

[10] E. Balboni, J. Ford, R. Tingley, K. Toomey, J. Vytal, "An empirical study of radio propagation aboard naval vessels", *IEEE-APS Conference on Antennas and Propagation for Wireless Communications*, Waltham, USA, 2000.

[11] G.B. Tait, "Electromagnetic Environment Characterization of Below-Deck Spaces in Ships", *IEEE International Symposium on Electromagnetic Compatibility*, Detroit, USA, 2008.

[12] K.K. Cwalina, S.J. Ambroziak, P. Rajchowski i L.M. Correia, "Radio channel measurements in 868 MHz off-body communications in a ferry environment", *XXXII General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS),* Montreal, Kanada, 2017.

[13] P.T. Kosz, S.J. Ambroziak, L.M. Correia, "Radio channel measurements in off-body communications in a ferry passenger cabin", *XXXII General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS),* Montreal, Kanada, 2017.

[14] P.T. Kosz, S.J. Ambroziak, J. Stefanski, K.K. Cwalina, L.M. Correia, K. Turbic, "An empirical system loss model for body area networks in a passenger ferry environment", *Baltic URSI Symposium (URSI)*, Poznan, Polska, 2018.

[15] K.K. Cwalina, S. Ambroziak, P. Rajchowski, "An Off-Body Narrowband and Ultra-Wide Band Channel Model for Body Area Networks in a Ferryboat Environment," *Applied Sciences*, vol. 8, nr 6, str. 1-16, 2018.

[16] A. Weisshaar, G.C. Alexander, P.C. Magnusson, V.K. Tripathi, "Transmission Lines and Wave Propagation", CRC press, 2000.

[17] J.D. Parsons, *The Mobile Radio Propagation Channel*, John Wiley & Sons, Chichester, UK, 2000.

[18] G. Glaeser, "Reflections on Spheres and Cylinders of Revolution", *Journal for Geometry and Graphics*, vol. 3, nr 2, str. 121–139, 1999.

[19] E.H. Lockwood, *A Book of Curves*, Cambridge University Press, Cambridge, UK, 1961.

[20] S.J. Ambroziak, "Measurement Stand and Methodology for Research of the Off-Body and Body-to-Body Radio Channels in WBANs with Different Diversity Schemes", *International Journal of Antennas and Propagation*, Vol. 2019, Article ID 3837190.

[21] F.D. Cardoso, P.T. Kosz, M.M. Ferreira, S.J. Ambroziak, L.M. Correia, „Fast Fading Characterization for Body Area Networks in Circular Metallic Indoor Environments", *IEEE Access*, vol. 8, str. 43817-43825, 2020.

TŁUMIENIE SYGNAŁÓW RADIOWYCH PRZEZ ROŚLINNOŚĆ DLA CZĘSTOTLIWOŚCI 28 GHZ

SIGNAL ATTENUATION BY FOLIAGE FOR 28 GHZ FREQUENCY

Maciej Nikiforuk[1]; Krzysztof Cichoń[2];

DOI: 10.15199/59.2022.4.92

[1]mackos.nikos@outlook.com
[2] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań, Krzysztof.cichon@put.poznan.pl

Streszczenie: **Rozwój systemów radiokomunikacyjnych powoduje poszukiwanie nowych, coraz wyższych, pasm częstotliwości dla zapewnienia rosnących potrzeb użytkowników. Wzmożone zainteresowanie tzw. pasmem fal milimetrowych wzmaga pytania o adekwatność dotychczas stosowanych modeli propagacyjnych dla wyższych zakresów częstotliwości. Jednym z interesujących aspektów jest tłumienie roślinności, które zostało poddane szczegółowej analizie w niniejszej pracy. Poza analizą komponentów tłumienia listowia w niniejszej pracy skupiono się na wykonaniu kampanii pomiarowej w otoczeniu budynków Politechniki Poznańskiej. Uzyskane wyniki porównano z zewnętrznymi pracami pomiarowymi i dokonano kalibracji współczynników tłumienia listowia odpowiednich dla rodzimych gatunków roślinności.**

Abstract: **The development of radio communication systems makes it necessary to seek for higher frequency bands to meet the growing needs of users. An increased interest in the so-called millimeter waves band raises questions about the adequacy of the use of propagation models for higher frequency ranges. One interesting aspect is radio signal attenuation in foliage, analyzed in detail in the paper. In addition to the analysis of the components of the foliage attenuation, the paper focuses on the implementation of the measurement campaign in the vicinity of the buildings of Poznań University of Technology. The obtained results were compared to external measurement works and there was performed the calibration of the foliage attenuation coefficients appropriate for native species of vegetation.**

Słowa kluczowe: **Tłumienie listowia, pasmo 28 GHz, propagacja radiowa.**

Keywords: **Foliage attenuation, 28GHz bandwidth, wireless propagation.**

1. WSTĘP

Gwałtowny rozwój systemów radiokomunikacyjnych wymaga poszukiwania i analizy nowych pasm częstotliwości. Zainteresowaniem cieszą się coraz to wyższe pasma częstotliwości, czego dowodem było ujęcie w standardzie systemu piątej generacji grupy częstotliwości określonej jako FR2, od 24,25 GHz do 52,6 GHz. Pasma z tego zakresu są czasem określane jako „fale milimetrowe" (ang. *mmWaves*). Niemniej jednak, w pracach [10][3] wskazano, że wraz ze wzrostem częstotliwości nośnej należy wziąć pod uwagę pomijane dotychczas aspekty a w [8] wśród pięciu kluczowych aspektów różniących pasma częstotliwości poniżej 6 GHz oraz pasma „fal milimetrowych" znalazło się powiększone tłumienie sygnałów radiowych przez roślinność. W zaleceniu Międzynarodowej Unii Telekomunikacyjnej [13] wskazano, że tłumienie sygnału radiowego w terenie leśnym dla częstotliwości 2 GHz wynosi 0,4 dB/m, a dla częstotliwości 30 GHz piętnastokrotnie więcej, czyli 6 dB/m. Szczególnie interesującym przypadkiem jest zatem sytuacja, gdy system komunikacji mobilnej jest umieszczony w pobliżu parku centralnego lub obszaru mieszkalnego. Scenariusz ze stratami spowodowanymi przez roślinność wymaga modelowania łącza jako dodatkowej, nadmiarowej straty.

Analiza wpływu listowia na tłumienie sygnału w pasmach milimetrowych spotkało się z zainteresowaniem badaczy. W artykułach [7-8] przygotowano i zbadano dwa różne scenariusze pomiarowe dla częstotliwości 28 GHz. Na podstawie pomiarów stwierdzono, że odbierane sygnały składają się z sygnału o dobrze określonej składowej kierunkowej oraz o składowej dodatkowej dyfuzyjnej. Na podstawie punktów pomiarowych dostrojono i wyprowadzono odpowiednie współczynniki. Wiodący charakter eksperymentów przeprowadzonych w Korei Południowej podkreśla fakt, że odzwierciedlają duży udział teorii radiacyjnego transferu energii. W pracy [7] badano sezonową zmienność tłumienia propagacyjnego. Pomiary prowadzono w grudniu, gdy liście ledwo zakrywały drzewa, oraz w październiku, gdy liście były bujne.

Choć istnieje kilka modeli tłumienia roślinności dla różnych pasm częstotliwości, to analiza prac oraz pomiarów wykazała, że jednym z najczęściej stosowanych opisów jest model wykładniczy zależny od częstotliwości, o którym mowa w [6-7]. Współczynniki dla tego modelu uzyskano na podstawie serii pomiarów. W [10] zaproponowano model tłumienia sygnałów radiowych od 1 GHz do 30 GHz. Bazując na tłumieniu dla pojedynczego drzewa, model pozwala obliczyć stratę tłumienia propagacyjnego spowodowaną przez drzewo, współczynnik Rice'a oraz skumulowany rozkład mocy. Model wymaga dodatkowych parametrów wejściowych, takich jak promień gałęzi, grubość gałęzi, gęstość liści i gałęzi itp. Model daje dobry wgląd w efekty i czynniki wpływające na fale milimetrowe [10].

W niniejszej pracy zdecydowano się na następujący porządek logiczny. W rozdziale drugim zaprezentowano ogólny zarys problematyki tłumienia listowia dla fal milimetrowych, w rozdziale trzecim odniesiono się do przygotowania kampanii pomiarowej. W rozdziale czwartym zaprezentowano wyniki uzyskane w trakcie pomiarów oraz porównano je z przyjętymi modelami. Rozdział piąty zajmuje się omówieniem wniosków.

2. MODELE TŁUMIENIA LISTOWIA

W modelowaniu tłumienia listowia zakłada się, że nadajnik znajduje się w oddali od ośrodka składającego się z listowia, podczas gdy odbiornik znajduje się w środku obszaru zalesionego. Wtedy dv to odległość w listowiu, dla której fala propaguje się przez listowie. Spośród dostępnych modeli analitycznych tłumienia listowia dwa są powszechnie stosowane.

2.1. Wprowadzenie do równania przepływu radiacyjnego

Równanie przepływu radiacyjnego (ang. *Radiative Transfer Equation* - RTE) [4][11][12] próbuje modelować roślinność jako jednorodny ośrodek losowo wypełniony swoistymi rozpraszaczami, które charakteryzują się następującym zestawem parametrów: współczynnik emisyjności j_v i współczynnik absorpcji α. Natomiast I stanowi intensywność sygnału. RTE jest zwykle wyrażane w postaci następującego równania różniczkowego.

$$\frac{dI}{dd_v} = \underbrace{j_v}_{emisja} - \underbrace{\alpha I}_{\substack{tłumienie \\ rozpraszanie}} \tag{1}$$

Lewa strona równania dI/dd_v opisuje zmianę natężenia sygnału przy niewielkiej zmianie odległości. Pierwszy wyraz po prawej stronie reprezentuje wzrost natężenia wynikający z rozpraszania otaczających elementów, natomiast drugi wyraz określa zmniejszenie intensywności podczas absorpcji i rozpraszania. Dzieląc obie strony przez współczynnik alfa i podstawiając iloczyn $dd_v \cdot \alpha$ jako $d\tau$, można zapisać równanie w bardziej zamkniętej formie.

$$\frac{dI}{d\tau} = S_v - I \tag{2}$$

W równaniu (2), S_v reprezentuje funkcję źródła [4]. Zakładając, że funkcja ta jest stała, można szybko wyprowadzić równanie opisujące tłumienie ośrodka w funkcji odległości.

$$\mathrm{I}(\tau) = I_0 \exp(-\tau) + S_v \cdot (1 - \exp(-\tau)) \tag{3}$$

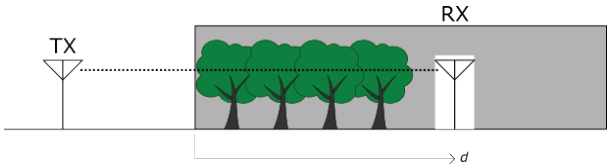
Równanie (3) przedstawia natężenie sygnału w funkcji tau. Dla $\tau >> 1$ można zignorować pierwszy składnik prawej strony równania, ponieważ drugi składnik staje się dominujący.

2.2. Model (MA) z maksymalnym tłumieniem

Pierwszy model zaproponowany w rekomendacji ITU P.833-9 [10] rozwiązuje problem tłumienia obszaru zalesionego w postaci funkcji wykładniczej. Model znalazł zastosowanie w modelowaniu tłumienia roślinności parków miejskich w centrach miast. Na Rys. 1. przedstawiono przykładowy scenariusz bezprzewodowego łącza radiowego. Anteny nadawcza i odbiorcza mają tę samą wysokość, dlatego też ścieżka sygnału jest równoległa do ziemi. W modelu odbiornik znajduje się w obszarze zalesionym, gdzie odległość d jest mierzona od granicy lasu. Pod uwagę brane jest tylko tłumienie listowia, natomiast odległość od nadajnika do granicy lasu nie jest uwzględniana w modelu. Tak więc nadmierne tłumienie listowia A_{ev} definiuje się jako:

$$A_{ev} = A_m \left[1 - \exp\left(-\frac{d \cdot \gamma}{A_m}\right)\right] \tag{4}$$

gdzie A_{ev} jest nadmiernym tłumieniem listowia wyrażonym w [dB], A_m to maksymalne tłumienie pojedynczej stacji bazowej na danej głębokości listowia, d to odległość przez zalesienie w metrach, a γ to współczynnik tłumienia dla małej odległości listowia wyrażony w [dB/m].



Rys. 1. Tłumienie listowia dla modelu MA

Parametry modelu podlegają kalibracji na podstawie pomiarów. A_{ev} może być postrzegana jako docelowa wartość strat leśnych lub jako nadmiarowe tłumienie listowia. γ to współczynnik specyficzny dla roślinności i gatunku, który znajdujemy na podstawie pomiarów i zmian częstotliwości. Model z Aalborga [5] jest przykładem takiego wyznaczania współczynników. Pomiary wykonano dla częstotliwości 24 GHz (scenariusz S1) w Brazylii dla obszaru mocno zalesionego składającego się z takich drzew jak kauczukowiec brazylijski, abiu, Mangueira oraz paulownia. W modelu z Aalborga (S1) wyznaczono γ jako 4,25 dB/m, a A_m jako 142,86 dB.

2.3. Model z maksymalnym wykładniczym zanikiem (MED)

Innym sposobem symulacji tłumienia ulistnienia jest uwzględnienie zależności częstotliwości i odległości ważonych trzema współczynnikami, jak w (5).

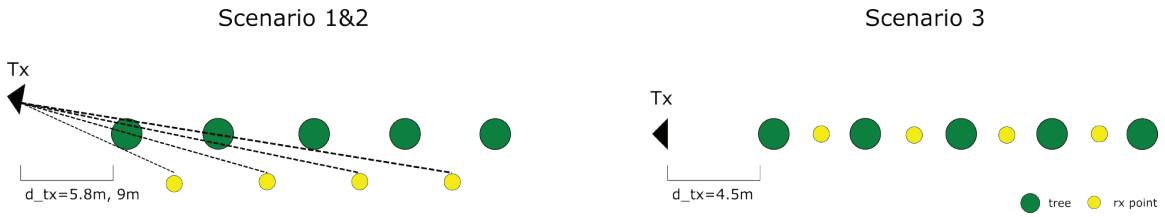
$$L_v = a \cdot f^b \cdot d_v^c \tag{5}$$

gdzie L_v to tłumienie listowia wyrażone w dB, f to częstotliwość podana w MHz, d_v to odległość, na jaką fala przechodzi przez ulistnienie w metrach, *a, b, c* to współczynniki specyficzne dla rozważanego scenariusza pomiarowego na podstawie przeprowadzonych pomiarów. Na przykład w modelu FITU-R [9] współczynniki są następujące a = 0,390, b = 0,390, c = 0,250, natomiast dla innych modeli odpowiednio: COST-235 [6][7] a = 13,77, b = −0,009, c = 0,26 i Weissberger [6][7] a = 0,187, b = 0,284, c = 0,588.

3. PRZYGOTOWANIE POMIARÓW

3.1. Scenariusz pomiarowy

Zasadniczą motywacją podczas przygotowania scenariusza pomiarowego było dostosowanie go do warunków możliwej transmisji na terenie miejskim. Zatem przygotowane scenariusze pomiarowe są dwojakie: z jednej strony wysokość zawieszenia anteny nadawczej jest większa niż odbiorczej, podobnie jak w przypadku stacji bazowych, których anteny znajdują się na większej wysokości nad poziomem terenu niż anteny w terminalu użytkownika [7]. W takim przypadku transmitowany sygnał przechodzi przez drzewo pod zadanym kątem. Z drugiej strony wysokości zawieszenia anteny nadaw-



Rys. 2. Tłumienie listowia dla modelu MA

czej i odbiorczej mogą być podobne, a fala może propagować się przez drzewo poziomo względem terenu. Jest to szczególnie ważne w parkach miejskich lub przy znacznych odległościach transmisji. Opisuje to rozważany scenariusz III. Na Rys. 2. przedstawiono schemat dla rozważanych scenariuszy pomiarowych, natomiast na Rys. 3. ujęto widok z lotu ptaka obszarów pomiarowych na placu Centrum Wykładowego Politechniki Poznańskiej z zaznaczonymi punktami pomiarowymi.

Tab. 1. Parametry scenariuszy pomiarowych

Scenariusz	*1*	*2*	*3*
Tx-Rx odległość[m]	4.5 - 29.5	4.5 -27	3 -33
Tx/Rx wysokość [m]	5 / 4	5.2 / 3.6	2.7 /2.6
Średnia wysokość drzew [m]	8	7.39	4.64
Średnia szerokość korony [m]	5	4.3	1.4
Średni dystans pomiędzy pniami [m]	5.74	4.25	6

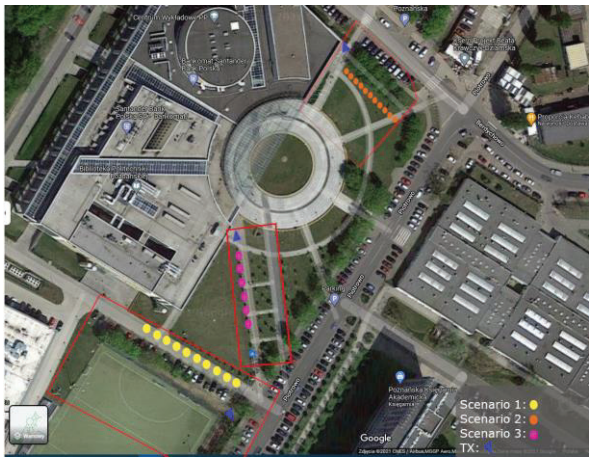
W Tab. 1 przedstawiono wybrane parametry scenariuszy pomiarowych. Chociaż scenariusze 1 i 2 były podobne pod względem założeń pomiarów, dotyczyły różnych gatunków drzew, odpowiednio platana klonolistnego dla scenariusza pierwszego, klonu jawora dla scenariusza drugiego oraz klonu pospolitego dla scenariusza trzeciego.

Pomiary przeprowadzono we wrześniu w suchy dzień o średniej temperaturze 17°C. W pomiarach uznano, że odległość między anteną nadawczą a koroną drzewa jest wystarczająca do wystąpienia pola dalekiego. Dlatego obliczona odległość Fraunhofera [9], która wyniosła 0,672 m była minimalną odległością dTX. Jak wynika z Rys. 2, rzeczywiste odległości były znacznie większe niż odległość Fraunhofera.

3.2. Konfiguracja sprzętowa

Pomiary wykonano dla następującej konfiguracji sprzętowej. Po stronie nadajnika zastosowano generator sygnału [1] z falą ciągłą (CW) o częstotliwości nośnej 28 GHz nadawanej z mocą 0 dBm. Zarówno po stronie nadajnika, jak i odbiornika zastosowano anteny tubowe o szerokości kąta połowy mocy[2] wynoszącym 16 stopni i wzmocnieniu 18 dBi. Pomiary wykonano dla polaryzacji pionowej.

Podczas pomiarów zastosowano procedurę polegającą na wykonaniu trzech serii pomiarowych na scenariusz, a następnie pobraniu wartości średniej. We wszystkich scenariuszach nadajnik jest zamocowany w miejscu, podczas gdy odbiornik porusza się po ścieżkach, a pomiary wykonywane są w punktach zaznaczonych punktami na Rys. 3. Odległość między poprzedzającymi punktami wynosi 2,5 m w scenariuszu I i II lub 6 m w scenariuszu III.



Rys. 3. Aleje pomiarowe przy Centrum Wykładowym Politechniki Poznańskiej

4. WYNIKI POMIARÓW

Pomiary przeprowadzono dla trzech scenariuszy, jak opisano w rozdziale 3. Na Rys. 4 i 5 przedstawiono zależność tłumienia ulistnienia od głębokości listowia dla pierwszego i trzeciego scenariusza pomiarowego. Wyniki pomiarów przedstawiono w postaci symboli – okręgów. Zestawiono je z modelami referencyjnymi: „Aalborg", „FITU-R", „COST", „Weissberger". Dla scenariuszy I i II model COST-235 najdokładniej odzwierciedlał pomierzone wartości tłumienia. Z drugiej strony, Scenariusz III, przedstawiony na Rys. 5., charakteryzuje się najszybszym wzrostem tłumienia, co jest widoczne w modelu jako wysoka wartość γ. W tym przypadku model FITU-R jest modelem najbliższym pod względem najmniejszego błędu średniokwadratowego. Jest to również model o największej wartości γ.

Porównanie pomiarów z modelami tłumienia listowia stało się podstawą do przeprowadzenia kalibracji. W tym celu wyznaczono parametry dla modelu MA i MED dla każdego scenariusza oraz średniokwadratowy błąd. Konieczność uzupełnienia tej grupy o typowe rodzime parkowe gatunki pozwoliła uzyskać współczynniki przedstawione w Tab. 3. Wartości RMSE dla wyznaczonych współczynników nie przekraczały 10 dB.

Następnie porównano modele MED i MA z porównywanymi punktami pomiarowymi przedstawionymi w pracy [9]. Różnice wynikają z różnych gatunków drzew i liczby przeprowadzonych punktów pomiarowych. Ostateczny wniosek jest taki, że w scenariuszach wzdłuż alei drzew uzyskane modele MED i MA są nie-

Tab. 3 Parametry modelu tłumienia propagacyjnego w roślinności po kalibracji

Nazwa modelu i typ	Częstotliwość	Parametry modelu					RMSE [dB]		
		γ	A_m	a	b	c	Scen.1	Scen.2	Scen.3
Model 1 (MA)	28 GHz	1,99	50,32	-	-	-	3,45	-	-
		2,18	61,33	-	-	-	-	3,98	-
		15,58	54,26	-	-	-	-	-	4,32
Model 2 (MED)	28 GHz	-	-	33,13	-0,25	0,78	3,71	-	-
		-	-	20,32	-0,2	0,84	-	3,87	-
		-	-	0,25	0,39	0,46	-	-	9,75
Weissberger (MED)	$f \leq 95$GHz	-	-	0,19	0,28	0,59	7,54	9,97	32,33
COST-235 (MED)	9,6 GHz $\leq f \leq$ 57,6 GHz	-	-	13,77	-0,009	0,26	6,86	5,91	24,94
FITU-R (MED)	10 GHz $\leq f \leq$ 40 GHz	-	-	0,39	0,39	0,25	19,39	16,24	9,76
Aalborg (MA)	24 GHz	4,25	142,86	-	-	-	33,52	28,49	20,91

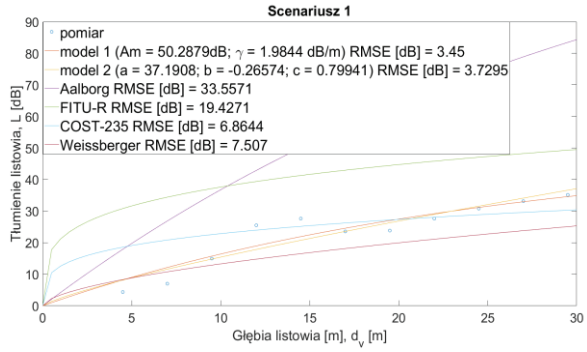
mal identyczne. Kluczową różnicą jest poziom tłumienia, który można powiązać z dokładną formą drzewa (liczba i gęstość liści konkretnego przykładu drzewa). W scenariuszu III anteny nadawczo-odbiorcze zostały umieszczone na tej samej wysokości i skierowane do siebie. Dlatego w tych pomiarach kluczowa była odległość między drzewami i gęstość korony. Uzyskana charakterystyka była jednak dość precyzyjna w porównaniu z teoretycznym opisem MA.

5. WNIOSKI

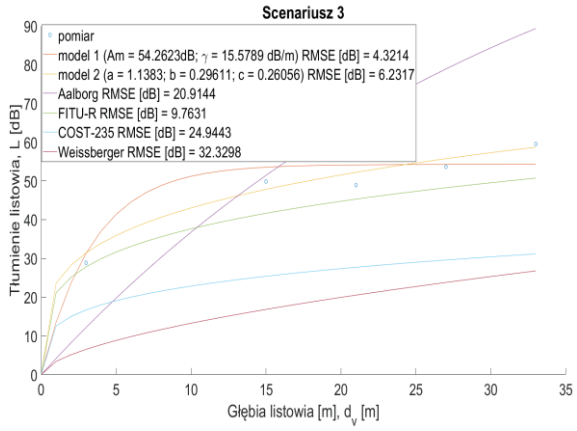
W referacie przedstawiono pomiary tłumienia propagacyjnego w listowiu w paśmie 28 GHz, które przeprowadzono za pomocą kierunkowych anten tubowych dla typowych obszarów parkowych porośniętych roślinnością w Poznaniu. Głównym celem kampanii było uzyskanie dopasowanych parametrów dla modeli tłumienia liści. Pomiary i przygotowanie modeli zostało przeprowadzone według głównych założeń dla trzech scenariuszy. Zauważono, że scenariusz trzeci charakteryzował się największym tłumieniem. Tak duże tłumienie przeszkody było spowodowane nie tylko specyficzną odległością między drzewami, ale przede wszystkim dużą gęstością listowia. W pozostałych scenariuszach tłumienie było nieporównywalnie mniejsze. Podczas gdy w pierwszym i drugim scenariuszu drzewa miały swego rodzaju prześwity, w trzecim scenariuszu żadne drzewo nie miało widocznych luk.

6. PODZIĘKOWANIA

Praca została sfinansowana dzięku wpsarciu Politechniki Poznańskiej w ramach projektu: 0312/SBAD/8161.



Rys. 4. Tłumienie listowia a głębia listowia dla scen. I



Rys. 5. Tłumienie listowia a głębia listowia dla scen. III

Bibliografia

[1] Anritsu. *Rf/microwave signal generator.* 2021. https:// dl.cdn-anritsu.com/en-us/test-measurement/files/Brochures-Datasheets-Catalogs/datasheet/11410-00515V.pdf.

[2] Cisco. " Antenna Patterns and Their Meaning." *Cisco Systems*, 2007.

[3] Erik Dahlman, Stefan Parkvall, Johan Skold. *5G NR The next generation wireless access tehcnology.* Elsevier, 2018.

[4] George B. Rybicki, Alan P. Lightman. *Radiative Processes in Astrophysics.* 2004 WILEY-VCH Verlag GmbH Co. KGaA, 1985.

[5] Ignacio Rodriguez, Renato Abreu, Erika P. L. Almeida, Mads Lauridsen, Alexandre Loureiro and Preben Mogensen. "24 GHz cmWave Radio Propagation Through Vegetation: Suburban Tree Clutter Attenuation." *2016 10th European Conference.* IEEE, 2016.

[6] Junghoon Ko, Sooyoung Hur, Yun-Seok Noh, Kuyeon Whang, Jeongho Park, Dong-Jo Park and Dong-Ho Cho. " Measurements and Analysis of Radio Propagation at 28 GHz in Vegetated Areas of Typical Residential Environments." *IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, VOL. 68, NO. 5*, May 2020.

[7] Junghoon Ko, Yun-Seok Noh, Yong-Chan Kim, Sooyoung Hur, Sung-rok Yoon, DH. Park, K. Whang, Dong-Jo Park, Dong-Ho Cho. "28 GHz Millimeter-Wave Measurements and Models for Signal Attenuation in Vegetated Areas." *11th European Conference on Antennas and Propagation (EUCAP).* 2017.

[8] Krzysztof Cichoń, "5 Key Differences Between mmWave and sub-6 GHz", https://rimedolabs.com/blog/5-key-differences-between mmwave-and-sub-6ghz/

[9] Ravaioli, Fawwaz T. Ulaby — Umberto. *Fundamentals of Applied Electromagnetics, Seventh Edition.* Pearson, 2015, 2010.

[10] Ryszard, J. Katulski. *Propagacja fal radiowych w sieciach 5G/IoT.* Wydawnictwa Komunikacji i Łączności, 2020.

[11] Schwering, R. A. Johnson and F. *A transport theory of millimiter wave propagation in woods and forests. Center for communications systems.* 1985.

[12] Telmo R. Fernandes, Rafael F. S. Caldeirinha, M.O. Al-Nuaimi and Jurgen Richter. *Radiative Energy Transfer Based Model for Radiowave Propagation in Inhomogeneus Forests,* 2006.

[13] Union, International Telecommunication. *Radiocommunication Sector of ITU. Recommendation ITU-R P.833-9, Attenuation in vegetation, P Series Radio-wave propagation.* 2016.

TWORZENIE MAP TŁUMIENIA PROPAGACYJNEGO Z WYKORZYSTANIEM METODY IDW

CREATING PROPAGATION ATTENUATION MAPS USING THE IDW METHOD

Bartosz Majewski[1]; Krzysztof Malon[2]; Michał Kryk[3]; Jan M. Kelner[4]

[1] Wojskowa Akademia Techniczna, Warszawa, bartosz.majewski@student.wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, krzysztof.malon@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Warszawa, michal.kryk@wat.edu.pl
[4] Wojskowa Akademia Techniczna, Warszawa, jan.kelner@wat.edu.pl

DOI: 10.15199/59.2022.4.93

Streszczenie: **W nowoczesnych systemach komunikacji bezprzewodowej dąży się do rozwiązań technologicznych mających na celu zwiększenie wydajności sieci radiowych, w tym także bardziej efektywnego wykorzystania zasobów widmowych. Niezmiernie istotnym elementem współczesnych systemów jest budowanie świadomości sytuacyjnej oraz wspieranie funkcji kognitywnych. Możliwe jest to między innymi dzięki zastosowaniu map środowiska radiowego REM reprezentujących złożone i wielodziedzinowe informacje dla zastosowań radia kognitywnego. Jednym z kluczowych obszarów REM są mapy tłumienia propagacyjnego pozwalające na określenie m.in. zasięgów łączności sieci bezprzewodowych oraz poziomów sygnałów radiowych w określonych lokalizacjach. W artykule przedstawiono algorytm budowania map tłumienia propagacyjnego z wykorzystaniem metody interpolacji ważonej odwrotną odległością IDW.**

Abstract: **In modern wireless communication systems, technological solutions aim to increase the efficiency of radio networks, including the more effective use of spectrum resources. An essential element of current systems is building situational awareness and supporting cognitive functions. It is possible by using, e.g., Radio Environment Maps (REM), representing complex and multi-domain information for cognitive radio applications. One of the critical areas of REM is propagation attenuation maps allowing for the determination of communication ranges of wireless networks or levels of radio signals in specific locations. The article presents an algorithm for building propagation attenuation maps using the Inverse Distance Weighted (IDW) interpolation method.**

Słowa kluczowe: **interpolacja, komunikacja bezprzewodowa, mapy środowiska radiowego, mapy tłumienia propagacyjnego.**

Keywords: **interpolation, wireless communications, radio environment maps, propagation attenuation maps.**

1. WSTĘP

Obecnie w szeroko pojętych dziedzinach wojskowych jak i cywilnych świadomość warunków środowiska radiowego w czasie pracy staje się coraz bardziej pożądaną i ważną informacją uzupełniającą wykorzystanie zasobów w różnych typach systemów dynamicznych. Traktując widmo radiowe jako ograniczone zasoby w systemie bezprzewodowym, ważnym aspektem jest rozważne i efektywne jego wykorzystanie. Często jednak zdarza się, iż zasoby są wykorzystywane w sposób nieefektywny, głównie przez licencjonowanych tradycyjnych użytkowników łączności radiowej, np. nadawcy radiowi, telewizyjni, operatorzy telefonii komórkowej. Liczne testy oraz kampanie pomiarowo-badawcze udowodniły, że niektóre z licencjonowanych zakresów częstotliwości są niewykorzystywane w przestrzeni i w czasie, a nielicencjonowane i licencjonowane pasma częstotliwości dla operatorów sieci komórkowych są przepełnione. Coraz powszechniej stosowanym podejściem do wykorzystania zasobów radiowych i poznania środowiska radiowego jest sensing (monitorowanie widma) [12][9]. Kolejną możliwością pozyskania wiedzy o środowisku radiowym jest wykorzystanie zasobów bazy danych wraz z ich geolokalizacją. Taka baza danych umożliwia wprowadzenie nielicencjonowanych wtórnych urządzeń w częściach pasm częstotliwości. Dzięki informacjom zawartym w bazach można uzyskać w czasie rzeczywistym obraz zajętości widma w poszczególnych pasmach w interesującej użytkownika lokalizacji. Radia kognitywne CR (ang. Cognitive Radio) korzystają z takich baz danych geolokalizacyjnych w celu pobrania informacji o dostępnych kanałach, łącznie z ograniczeniami czasowymi i maksymalnymi wartościami mocy nadawania. Sama geolokalizacyjna baza danych wykazuje pewne ograniczenia związane z wysoce dynamicznym środowiskiem radiowym. Pożądanym rozwiązaniem jest zatem zastosowanie koncepcji połączenia monitorowania widma z geolokalizacją i bazą danych tj. map środowiska radiowego REM (ang. Radio Environment Map).

2. MAPY ŚRODOWISKA RADIOWEGO

Mapy środowiska radiowego reprezentują złożone i wielodziedzinowe informacje dla zastosowań radia kognitywnego. Rozpowszechnianie udostępnionych informacji zawartych w REM pozwala na korzystanie z funkcji kognitywnych tj.: świadomość sytuacyjna (ang. situational awarness), dedukcja (ang. deducing), uczenie (ang. learning), planowanie (planning) oraz wspomaganie podejmowania decyzji (ang. decision-making support). REM zawiera informacje przede wszystkim o obszarze działań, dane pochodzące z czujników (węzłów sieci) czy też dane propagacyjne [3].

REM jest wykorzystywane na szeroką skalę zarówno w dziedzinach cywilnych jak i wojskowych. Przede wszystkim jest to element dynamicznego zarządzania widmem. Mówiąc o systemach i sieciach bezprzewodowych obserwowany jest ciągły ich rozwój,

a w szczególności rozwój telefonii komórkowej i związanych z nią technologii np. LTE (ang. Long Term Evolution), 4G+, 5G, 6G. Szybkie unowocześnianie kolejnych technologii i rozwiązań nie trwa w takim tempie w przypadku systemów wojskowych, które muszą spełniać bardziej krytyczne wymagania pod względem mobilności, niezawodności, bezpieczeństwa oraz ukrycia połączeń [4].

Siły militarne każdego z państw działają w różnych uwarunkowaniach terenowych zaczynając od nizinnych, a kończąc na górzystych, czy też pustynnych. Każde środowisko będzie miało inny wpływ na zapewnienie chociażby łączności między jednostkami. Ze względu na m.in. dużą mobilność systemy łączności taktycznej opierające się na mobilnych sieciach ad-hoc (MANET) idealnie pasują do różnych typów terenu. Tworzenie sieci taktycznych wymaga planowania z uwzględnieniem różnych warunków propagacyjnych i właśnie w tym celu wykorzystywane mogą być mapy środowiska radiowego [7][8]. REM jest źródłem informacji o określonych zasięgach nadajników, realnych możliwościach realizacji misji itp. co przekłada się na skuteczne prowadzenie operacji wojskowych, ciągłość łączności i znajomość zmieniającego się środowiska radiowego w rejonie misji [14].

Zainteresowanie wypracowanymi standardami i narzędziami dotyczącymi tworzenia wspólnego obrazu operacyjnego widma wyraziło też NATO (ang. North Atlantic Treaty Organization). Zakłada się, iż tworzone mapy środowiska radiowego mają współpracować za pomocą standaryzowanych interfejsów z systemami sensorowymi w celu uaktualniania mapy widma na podstawie informacji pozyskiwanych w czasie rzeczywistym. Systemy planowania częstotliwości miałyby dostęp do rzeczywistej informacji o zajętości danego przedziału widma, a także posiadałyby dane statystyczne na temat środowiska radiowego [13].

W ramach wielu badań nad dynamicznym dostępem do widma stwierdzono, iż REM jest przede wszystkim wsparciem w podejmowaniu decyzji w kognitywnych sieciach radiowych. Jej wykorzystanie przyczynia się do minimalizacji zakłóceń, lepszego zarządzania zasobami radiowymi oraz optymalizacji sieci [2].

Jednym z wielu zastosowań jest także wykorzystanie REM do działania DSM (ang. Dynamic Spectrum Managment) oraz do poprawienia wyników uzyskiwanych na drodze mechanizmu DSM/CDSM (ang Centralized Dynamic Spectrum Managment) w celu pełnej świadomości sytuacji elektromagnetycznej [11].

Podsumowując, należy zatem uznać REM jako istotny element w procesie dynamicznego zarządzania widmem. Na podstawie informacji, które dostarcza REM można oszacować, a także modelować dalsze szczegóły konkretnego obszaru środowiska radiowego dotyczące m.in. poziomów zakłóceń, poziomów tłumienia propagacyjnego, stosowania odpowiednich rodzajów transmisji [1].

Podstawowa funkcjonalność REM to budowanie dynamicznej mapy zakłóceń radiowych. W artykule przedstawiono tworzenie map tłumienia propagacyjnego z wykorzystaniem wybranej bezpośredniej metody konstrukcji RF-REM, tj. metody interpolacji ważonej odwrotną odległością.

3. MAPY TŁUMIENIA PROPAGACYJNEGO

Mapy tłumienia propagacyjnego są jedną z reprezentacji map środowiska radiowego odpowiedzialną za zobrazowanie poziomów tłumień sygnału radiowego na określonych kierunkach. Jednym z algorytmów stworzenia mapy tłumienia propagacyjnego PAM (ang. Propagation Attenuation Map) jest ten zaproponowany w [3]:

1) Wyznaczenie profili terenowych na podstawie lokalizacji nadajnika, kierunków azymutalnych, a także danych cyfrowych o ukształtowaniu terenu.

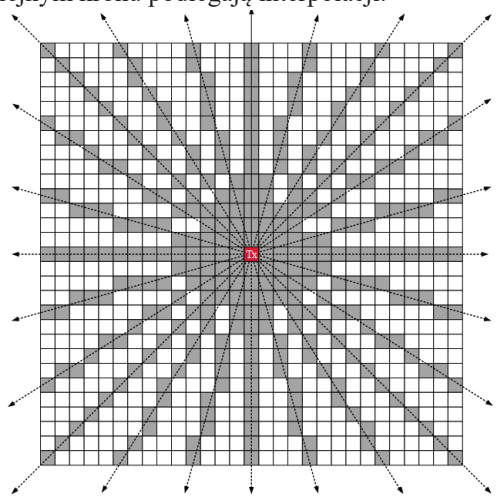
2) Wykorzystanie metody PEM (ang. Parabolic Equation Method) oraz określenie tłumienia propagacyjnego pomiędzy nadajnikiem, a odbiornikami na danej wysokości.

3) Skonstruowanie PAM dla konkretnych profili terenowych.

4) Wykorzystanie algorytmu interpolacji liniowej celu wyznaczenia poziomów tłumień w pozostałych punktach (pomiędzy wartościami obliczonymi).

5) Wizualizacja wyników w postaci gotowych map tłumienia propagacyjnego.

Na Rys. 1 przedstawiono macierz mapy tłumienia propagacyjnego przed interpolacją. Punkty oznaczone kolorem szarym reprezentują wartości obliczone z wykorzystaniem metody PEM, natomiast białe – punkty, które w kolejnym kroku podlegają interpolacji.



Rys. 1. Macierz mapy tłumienia propagacyjnego przed interpolacją

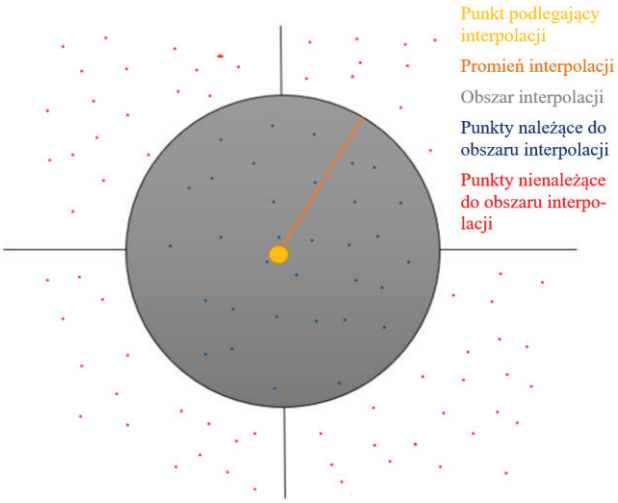
4. METODA IDW

W ramach niniejszego artykułu liniowa metoda interpolacji zaproponowana w algorytmie tworzenia PAM w [3] została zastąpiona metodą IDW (ang. Inverse Distance Weighted).

4.1. Opis metody

IDW czyli metoda interpolacji ważonej odwrotną odległością jest jedną z najpopularniejszych metod uży-

wanych w procesie tworzenia map środowiska radiowego, czy też map powierzchni wykorzystywanych w geostatystyce. Idea tej metody opiera się na informacji dotyczącej znanych wartości i dokonaniu operacji przypisywania wag w zależności od tego, jak daleko znajdują się te wartości od punktu, którego poziom będzie interpolowany [5]. IDW jest matematyczną metodą pozwalającą oszacowywać nieznane wartości zmiennych na podstawie już znanych znajdujących się wewnątrz okręgu, który jest definiowany przez wartość tzw. promienia poszukiwań [10]. Długość tego promienia pozwala na modyfikację obszaru poszukiwań i kontrolę obejmowanego obszaru na bazie którego oszacowywana będzie poszukiwana wartość. Ideę metody IDW oraz charakterystyczne punkty przedstawiono na Rys. 2.



Rys. 2. Idea metody IDW

Równanie opisujące metodę IDW ma postać średniej ważonej:

$$Z_j = \frac{\sum_{i=1}^{n} z_i w_{ij}}{\sum_{i=1}^{n} w_{ij}} \tag{1}$$

gdzie:

Z_j – interpolowana wartość (nieznana) w j-tej lokalizacji,
z_i – znana wartość w i-tej lokalizacji,
w_{ij} – wartość wagi,
n – liczba punktów wykorzystywanych do interpolacji.

Wartość wagi jest to odwrotność odległości między punktem w i-tej (znanej), a j-tej (nieznanej – interpolowanej) lokalizacji dzielona na sumę odwrotności odległości wszystkich punktów branych pod uwagę do wyliczania wartości nieznanej co przedstawione jest wzorem [6]:

$$w_{ij} = \frac{\frac{1}{d_{ij}^{p}}}{\sum_{i=1}^{n} \frac{1}{d_{ij}^{p}}} \tag{2}$$

gdzie:

d_{ij}^{p} – określony dystans (odległość) od wartości znanej w i-tej lokalizacji do wartości interpolowanej w j-tej lokalizacji,
p – wykładnik potęgowy.

Dystans stanowi odległość, na bazie której wyliczane są wartości wag. Parametr ten można wyznaczyć na podstawie położenia punktu interpolowanego oraz punktów o znanych wartościach. Im wyższa wartość dystansu pomiędzy punktami, tym niższe wartości wag co świadczy o większym wpływie punktów leżących bliżej na wartość wynikową.

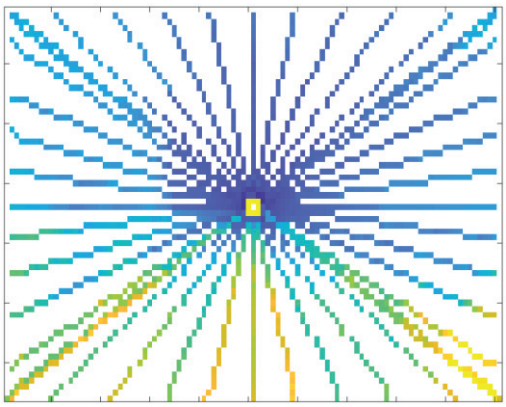
Za pomocą wartości wykładnika potęgowego (dodatnia liczba rzeczywista) można natomiast kontrolować wpływ znanych próbek punktowych bazując na ich odległości od punktu interpolowanego. Przy zwiększaniu wartości wykładnika potęgowego można zauważyć, że interpolowane wartości zaczynają się zbliżać do wartości najbliżej położonego punktu o znanym poziomie. W celu wygładzenia powierzchni należy zmniejszyć wartość tego parametru, co skutkuje większym wpływem pozostałych punktów otaczających wartość poszukiwaną.

Kolejnym istotnym parametrem związanym z metodą IDW jest liczba punktów, na podstawie których dokonywana jest interpolacja. Ograniczenie liczby rozważanych punktów wejściowych może poprawić szybkość przetwarzania. Należy również wziąć pod uwagę, że punkty wejściowe położone daleko od lokalizacji, w której dokonywana jest interpolacja, mogą mieć słabą korelację przestrzenną lub jej brak i z tego względu należy je wyeliminować z obliczeń. Liczbę punktów można zdefiniować wprost lub określić promień w obrębie którego punkty będą uwzględnianie w procesie interpolacji.

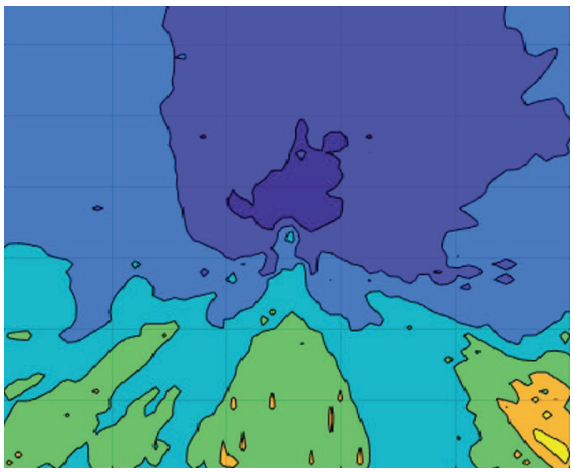
Do najistotniejszych zalet metody IDW można zaliczyć stosunkowo łatwą implementację oraz szybkość i wydajność obliczeniową.

4.2. Przykładowe wyniki

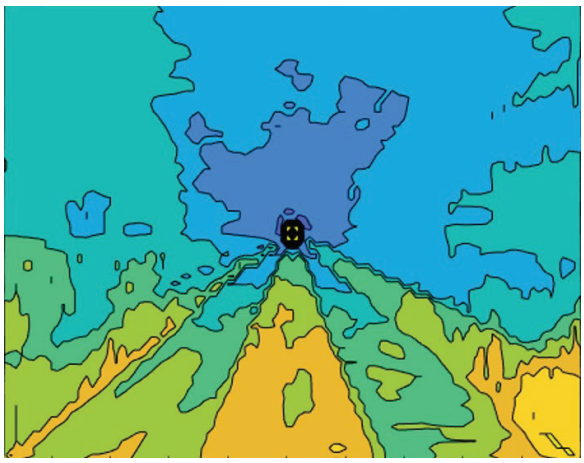
Na Rys. 3 zamieszczono przykładową mapę tłumienia propagacyjnego przed interpolacją zawierającą wartości obliczone na podstawie PEM wzdłuż zadanych profili. W celu wizualnego porównania rezultatów działania zaimplementowanej metody, na kolejnych zobrazowaniach przedstawiono kompletne mapy PAM dla dwóch przypadków. Rys. 4 przedstawia mapę tłumienia propagacyjnego wyznaczoną z wykorzystaniem metody interpolacji IDW, natomiast Rys. 5 - po interpolacji liniowej zdefiniowanej w [3].



Rys. 3. Przykładowa mapa tłumienia propagacyjnego przed interpolacją



Rys. 4. Mapa tłumienia propagacyjnego po interpolacji z wykorzystaniem metody IDW



Rys. 5. Mapa tłumienia propagacyjnego po interpolacji liniowej

5. PODSUMOWANIE

W artykule przedstawiono zastosowanie metody interpolacji ważonej odwrotną odległością w algorytmie tworzenia map tłumienia propagacyjnego. Zaprezentowane rozwiązanie bazuje na obliczonych, z wykorzystaniem metody równania parabolicznego, punktach rozmieszczonych wzdłuż zadanych profili mających swój początek w miejscu lokalizacji nadajnika. Podstawą obliczeń są m.in. lokalizacja nadajnika, dane dotyczące ukształtowania terenu oraz wysokość anteny odbiorczej. Zadaniem interpolacji jest wyznaczenie wartości tłumienia w pozostałych punktach mapy. Tworzone mapy tłumienia propagacyjnego mogą znaleźć zastosowanie m.in. w mapach środowiska radiowego czy też w systemach planowania sieci radiowych.

W ramach kierunków dalszych prac autorzy planują rozszerzenie badań związanych z wykorzystaniem metody interpolacji IDW w porównaniu ze stosowaną wcześniej metodą interpolacji linowej w zakresie dokładności oraz złożoności obliczeniowej (czasu obliczeń).

LITERATURA

[1] Denkovski D., Atanasovski V., Gavrilovska L., Riihijärvi J., Mähönen P. 2012. „Reliability of Radio Environment Map: Case of Spatial Interpolation Techniques". *7th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*: 248–253.

[2] Gajewski Piotr. 2018. „Propagation Models in Radio Environment Map Design". *Baltic URSI Symposium*: 234–237.

[3] Kryk Michał, Malon Krzysztof, Kelner Jan Marcin. 2021. „Algorithm of Creating Propagation Attenuation Maps Based on Parabolic Equation Method". *Signal Processing Symposium (SPSympo)*: 138–142.

[4] Kryk Michał. 2021. „Creating Radio Environment Maps (REMs) based on Various Path Loss Models". *Proceedings of the 2021 37th International Business Information Management Conference (IBIMA)*: 9254–9264.

[5] Li Zhengquan, Wang Kuo, Ma Hao, Wu Yaoxiang. 2018. „An Adjusted Inverse Distance Weighted Spatial Interpolation Method". *3rd International Conference on Communications, Information Management and Network Security (CIMNS 2018)*, 65: 128–132.

[6] Majdisova Zuzana, Skala Vaclav. 2017. „Radial Basis function approximations: comparison and applications". *Applied Mathematical Modelling*, 51: 728–743.

[7] Malon Krzysztof, Skokowski Paweł, Łopatka Jerzy. 2018. „Optimization of the MANET Topology in Urban Area Using Redundant Relay Points". *Proceedings of the 2018 19th International Conference on Military Communications and Information Systems (ICMCIS)*.

[8] Malon Krzysztof, Skokowski Paweł, Łopatka Jerzy. 2018. „Optimization of Wireless Sensor Network Deployment for Electromagnetic Situation Monitoring". *International Journal of Microwave and Wireless Technologies*.

[9] Malon, Krzysztof. 2021. „Dynamiczny dostęp do widma w radiowych sieciach doraźnych z węzłami kognitywnymi". *Wojskowa Akademia Techniczna*.

[10] Malvić Tomislav, Ivšinović Josip, Velić Josipa, Sremac Jasenka, Barudžija Uroš. 2020. „Application of the Modified Shepard's Method (MSM): A Case Study with the Interpolation of Neogene Reservoir Variables in northern Croatia". *Stats*, 3: 68–83.

[11] Romanik Janusz, Golan Edward, Zubel Krzysztof, Kaniewski Paweł. 2019. „Electromagnetic Situational awareness of Cognitive Radios supported by Radio Environment Maps". S*ignal Processing Symposium (SPSympo)*: 1–6.

[12] Skokowski, Paweł. 2021. „Budowanie świadomości sytuacji elektromagnetycznej w sieciach doraźnych z węzłami kognitywnymi". *Wojskowa Akademia Techniczna*.

[13] Suchański Marek, Kaniewski Paweł, Romanik Janusz, Golan Edward. 2018. „Radio Environment Map to Support Frequency Allocation in Military Communications Systems". *Baltic URSI Symposium*: 230–233.

[14] www.telekom.com/en/company/details/intelligent-5g-network-how-does-dynamic-spectrum-sharing-work-611104

OCENA EFEKTYWNOŚCI GENERACJI MAP TŁUMIENIA PROPAGACYJNEGO

EFFECTIVENESS EVALUATION FOR GENERATING PROPAGATION ATTENUATION MAP

Michał Kryk[1]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, michal.kryk@wat.edu.pl

DOI: 10.15199/59.2022.4.94

Streszczenie: **Nowoczesne systemy komunikacji bezprzewodowej wykorzystują różne rozwiązania technologiczne w celu zwiększenia wydajności tworzonych sieci radiowych. Obecnie jednym z kierunków pozwalających usprawnić zarządzanie zasobami radiowymi jest wykorzystanie mapy środowiska radiowego (REM). REM są coraz częściej wykorzystywane w mobilnych sieciach doraźnych (MANET), w szczególności w taktycznych sieciach wojskowych. Mapy tłumienia propagacyjnego (PAM) są kluczowymi elementami REM, które pozwalają na określenie zasięgów węzłów sieci radiowej. W niniejszym artykule zostanie przedstawiona ocena efektywności autorskiego algorytmu generowania PAM. Zaproponowany algorytm PAM bazuje na metodzie równań parabolicznych (PEM), interpolacji liniowej oraz profilach terenu wyznaczanych na podstawie map wysokościowych DTED.**

Abstract: **Modern wireless communication systems use various technological solutions to increase the efficiency of the created radio networks. Currently, the use of the radio environment map (REM) is one of the directions allowing improvement of radio resource management. REM is being used more often in emerging mobile ad-hoc networks (MANETs), in particular in tactical military networks. Propagation attenuation maps (PAMs) are the key elements of REM in determining the ranges of the radio network nodes. In this paper, an efficiency assessment of PAM generation is presented. The proposed PAM algorithm is based on the parabolic equation method (PEM), linear interpolation, and terrain profiles determined based on the DTED elevation maps.**

Słowa kluczowe: **interpolacja, komunikacja bezprzewodowa, mapa tłumienia propagacyjnego, mapa środowiska radiowego, metoda równania parabolicznego, propagacja fal radiowych.**

Keywords: **interpolation, parabolic equation method (PEM), propagation attenuation map (PAM), radio environment map (REM), radio wave propagation, wireless communications.**

1. WSTĘP

W mobilnych sieciach ad-hoc (MANET) [1–3] nie zawsze można zastosować rozwiązania stosowane w sieciach komórkowych. W tym przypadku, nie występują stacje bazowe o stałych lokalizacjach, a każdy węzeł sieci można traktować jako stację ruchomą. Pełna mobilność wszystkich węzłów MANET oraz zmienność terenu, na którym działa sieć, przyczyniła się do coraz bardziej powszechnego wykorzystania map środowiska radiowego (REM) w tego typu sieciach [4–6]. Mapy tłumienia propagacyjnego (PAM) są natomiast jednym z podstawowych elementów REM. Do tworzenia PAM można wykorzystać statystyczne i deterministyczne modele propagacyjne [7]. PAM pozwala na wyznaczenie wartości tłumień sygnału od dowolnego węzła sieci traktowanego jako nadajnik. Na tej podstawie możliwe jest określenie zasięgów poszczególnych węzłów MANET i jej rekonfiguracja.

W niniejszym artykule skupiono się na ocenie efektywności generowania PAM z wykorzystaniem nowatorskiego algorytmu [8,9]. W pierwszym etapie algorytmu, bazując na mapach wysokościowych DTED, wyznaczenie są profile terenu wychodzące promieniście z pozycji wybranego węzła. Dla każdego profilu, metodą równania parabolicznego (PEM) wyliczane są tłumienia wzdłuż na wysokości anteny hipotetycznego węzła odbiorczego. W ten sposób otrzymywana jest tzw. macierz rzadka tłumień. Macierz ta jest następnie uzupełniana do pełnej macierzy tłumień. W tym celu zastosowany jest autorskiego algorytmu interpolacji liniowej.

Pozostała część artykułu ma następujący układ. W sekcji 2 przedstawiono opis autorskiego algorytmu. Przegląd metryk wykorzystanych do oceny efektywności generowania PAM oraz wyniki z przeprowadzonych badań zawarto odpowiednio w sekcjach 3 i 4. W części końcowej przedstawiono wnioski.

2. ALGORYTM PAM

Ogólna procedura generowania PAM jest przedstawiona na Rys. 1. Główna idea opiera się na wyznaczeniu macierzy rzadkiej dla danych profili terenu. Proponowane podejście umożliwia określenie wartości tłumienia propagacyjnego we wszystkich dyskretnych punktach wokół stałej pozycji TX (Rys. 2). W kolejnym kroku wyznaczana jest pełna macierz PAM na podstawie metody interpolacji liniowej. Analizowany obszar zawiera się w kwadracie o długości boku $2R_0$, gdzie R_0 oznacza promień analizowanego obszaru terenu.

Indeksy poszczególnych pikseli są skorelowane ze współrzędnymi geograficznymi, a odległość między nimi zależy od dostępnych danych topograficznych, tj. rozdzielczości DTED. Obliczenia według metody PEM wykonywane są na zdefiniowanych kierunkach wokół TX w zależności od rozdzielczości kątowej.

Przykładową procedurę wyznaczania tłumienia dla azymutu ($\alpha_n = n\Delta\alpha$) zaznaczono na zielono (Rys. 2).

Pierwszym krokiem jest odczyt profilu terenu wzdłuż analizowanego azymutu α_n na podstawie map wysokościowych DTED.

Wejście: *lokalizacja nadajnika, kierunek azymutu (rozdzielczość kątowa), mapy w formacie DTED, wysokość anteny odbiorczej (h_R)*

Wyjście: *mapa tłumienia propagacyjnego (PAM) – tzw. macierz pełna PAM*

1 ***Przygotowanie wstępne:*** *wczytać mapę DTED*

Powtarzać dla każdego azymutu (α)

2 *Obliczyć profil terenu na podstawie lokalizacji nadajnika dla danego kierunku azymutu*

3 *Numeryczne rozwiązanie PEM dla danego profilu terenu*

$$\frac{\partial^2}{\partial z^2}E(x,z)-2\mathrm{j}\,k_0\frac{\partial}{\partial x}E(x,z)+k_0^2\left(n^2-1\right)E(x,z)=0$$

4 *Odczytać tłumienie propagacyjne na podstawie PEM dla wysokości anteny odbiorczej*

$$L(D)=L(x=D,z=h_R)$$

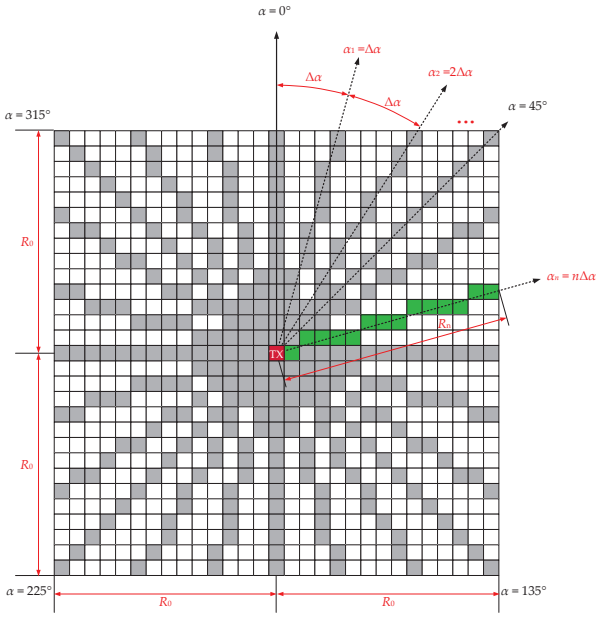
5 *Wyznaczyć elementy tzw. macierzy rzadkiej PAM dla zadanych kierunków azymutów i profili terenu*

$$PAM(\alpha)=L(D)$$

Koniec

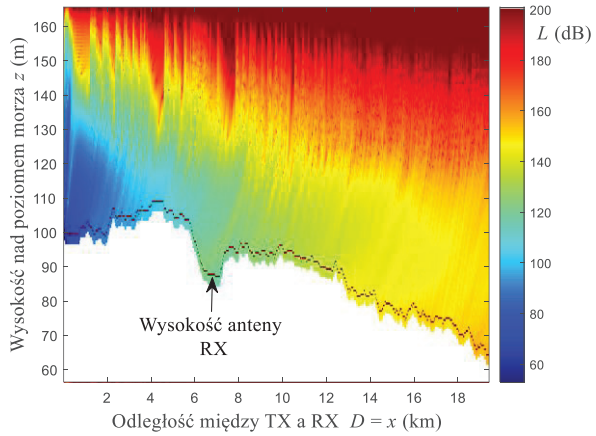
6 *Wyznaczyć pozostałe punkty macierzy PAM za pomocą metody interpolacyjnej – pełna macierz PAM*

Rys. 1. Procedura generowania macierzy PAM

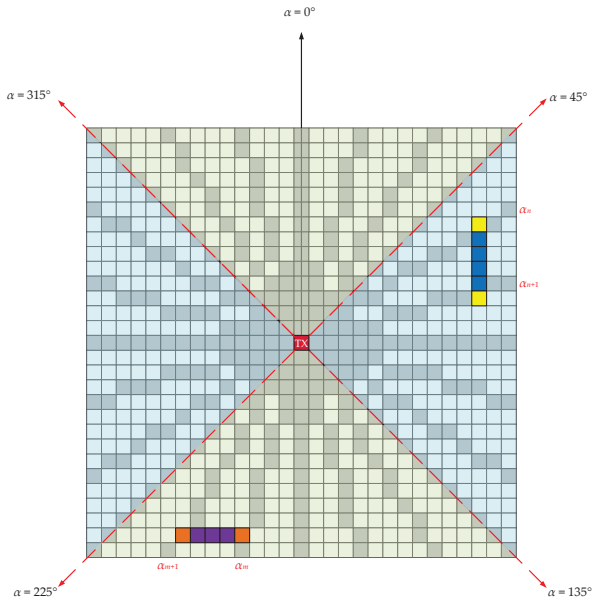


Rys. 2. Wyznaczanie profili terenu wzdłuż zadanych azymutów oraz tworzenie macierzy rzadkiej PAM

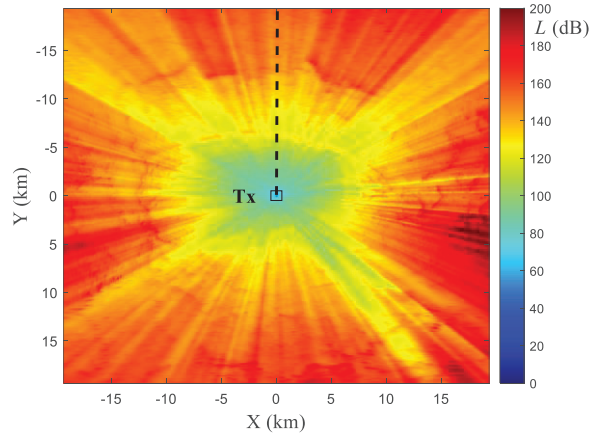
W kolejnej fazie wykonywane są obliczenia numeryczne na bazie PEM mające na celu określenie tłumienia sygnału nad analizowanym profilem terenu (Rys. 3). Następnie z otrzymanego zobrazowania PEM, w zależności od wysokości anteny odbiorczej wyznaczany jest wektor tłumienia dla danego azymutu (Rys. 3). Te dane umieszczane są w odpowiednich komórkach macierzy PAM (Rys. 2 i 4).



Rys. 3. Wyznaczenie tłumienia na bazie PEM nad profilem terenu dla wybranego azymutu



Rys. 4. Metoda interpolacji liniowej umożliwiająca tworzenie pełnej macierzy PAM



Rys. 5. PAM uzyskana na podstawie interpolacji liniowej

Na Rys. 4 zostały przedstawiony przykłady interpolacji liniowej dla kolumny i wiersza macierzy PAM. W przypadku interpolacji wierszy, punkty fioletowe są wartościami zerowymi, które są interpolowane za pomocą punktów pomarańczowych wyznaczonych dla dwóch sąsiednich profili na bazie PEM. Interpolacja w wierszach jest wykonywana dla azymutów $0° \leq \alpha \leq 45°$, $135° \leq \alpha \leq 225°$, $315° \leq \alpha \leq 360°$. Analogiczna sytuacja występuje dla procesu interpolacji w kolumnach. Niebieskie punkty są obliczane za pomocą dwóch żółtych punktów dla azymutu $45° \leq \alpha \leq 135°$ oraz $225° \leq \alpha \leq 315°$.

Ostatecznym wynikiem tej procedury jest pełna macierz PAM, którą zilustrowano na Rys. 5. Porównanie wyznaczania PAM metodą interpolacji liniowej z metodą IDW przedstawiono w [10].

3. METRYKI EFEKTYWNOŚCI

W tej sekcji opisano metryki wykorzystane do oceny efektywności generacji macierzy PAM [9]. Zgodnie z przedstawionym opisem algorytmu w poprzedniej sekcji, w rzadkiej macierzy PAM istnieją punkty, dla których nie wyznaczono tłumienia. Wartości tłumienia w tych lokalizacjach muszą być interpolowane. Z drugiej strony są punkty w PAM, w których obliczenia PEM były wykonywane wielokrotnie dla kilku sąsiednich azymutów. Wynika to z dużej gęstości profili w pobliżu TX. W tym celu został wprowadzony parametr redundancji *r*, jako liczba obliczeń wartości tłumienia w danym punkcie. W konsekwencji możliwe jest zdefiniowanie współczynnika obliczonych punktów dla analizowanego obszaru:

$$CPR[\%] = 100 \cdot CP/TP \tag{1}$$

gdzie *CP* i *TP* są odpowiednio liczbą obliczonych punktów i całkowitą liczbą punktów na analizowanym obszarze (macierzy PAM).

Aby określić liczbę wszystkich wykonanych obliczeń należy zsumować wartości redundancji *r* dla wszystkich punktów w analizowanym obszarze:

$$CN = \sum_{s=1}^{TP} r_s \tag{2}$$

gdzie *s* jest indeksem punktu, r_s jest wartością redundancji dla określonego punktu *s*.

Wprowadzona wcześniej miara *CP* uwzględnia liczbę punktów, w których obliczenia zostały wykonane przynajmniej raz ($r \geq 1$).

Korelując ten parametr z redundancją można zdefiniować nową miarę CP_r oznaczającą liczbę punktów, w których obliczenia zostały wykonane dokładnie *r* razy. W związku z tym wcześniej wprowadzona metryka może być również wyrażona jako

$$CN = \sum_{r=1}^{r_{max}} r \cdot CP_r \tag{3}$$

gdzie r_{max} jest maksymalną wartością dla analizowanego obszaru.

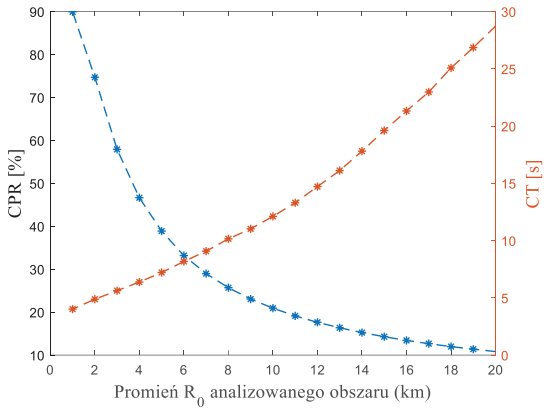
Znormalizowaną liczbę punktów, w których obliczenia zostały wykonane dokładnie *r*-razy, można zdefiniować w następujący sposób:

$$NCP_r[\%] = 100 \cdot CP_r/CP \tag{4}$$

Dodatkowo do oceny wydajności tworzenia PAM wykorzystano czas obliczeń *CT* wyrażony w sekundach. Miara uwzględnia wykonanie wszystkich kroków algorytmu z wyłączeniem obliczania profili terenu. Wartości bezwzględne tej miary zależą od mocy obliczeniowej używanego komputera. Dlatego parametr ten służy do porównywania różnych wariantów, np. różnej wielkości analizowanego obszaru.

4. OCENA EFEKTYWNOŚCI GENEROWANIA PAM

Przedstawione w tej sekcji wyniki dotyczą badań przeprowadzonych dla różnych wielkości analizowanego obszaru (kwadrat o długości boku $2R_0$). Rozdzielczość kątowa $\Delta\alpha$ generowania profili terenów dla wszystkich testów jest równa 1°. Rys. 6 pokazuje wpływ zwiększenia promienia R_0 na obliczoną wartość współczynnika punktów *CPR* (kolor niebieski) i czas obliczeń *CT* potrzebny do uzyskania wyników (kolor pomarańczowy).

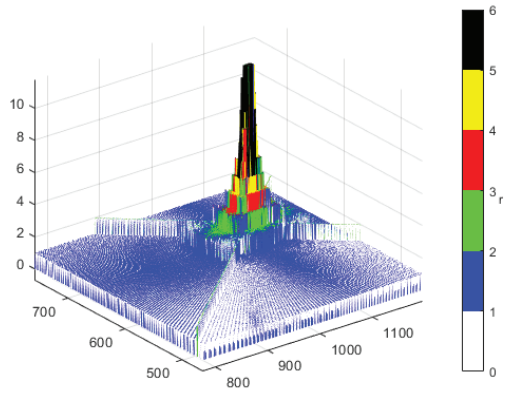


Rys. 6. Metryki CPR i CT w funkcji promienia R_0.

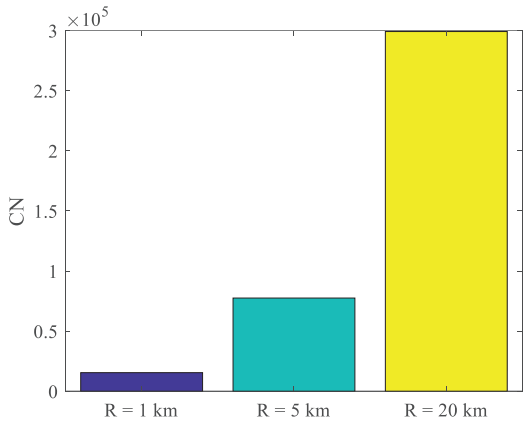
Jak wspomniano w poprzedniej sekcji, metryka *CT* zależy od mocy obliczeniowej i służy do celów porównawczych. Im większa wartość promienia, tym dłuższy czas potrzebny na wykonanie obliczeń. Np. w przypadku dziesięciokrotnego wzrostu R_0 (od 2 do 20 km), *CT* wzrosło około sześciokrotnie. W przypadku analizowania małych obszarów zdecydowana większość punktów macierzy PAM jest obliczana tylko na podstawie PEM. Niektóre punkty są obliczane wielokrotnie, jak pokazano na Rys. 7. Im większy obszar analizy, tym *CPR* znacznie maleje, a więcej punktów należy wyznaczyć metodą interpolacji. W przypadku promienia R_0 równego 1 km około 90% punktów macierzy PAM jest wyznaczone na podstawie PEM. Zwiększenie R_0 do 10 km skutkuje spadkiem do poziomu zaledwie około 21%. W takim przypadku 79% punktów musi być interpolowanych.

Rys. 7 przedstawia efekt podwójnego obliczania punktów w generowaniu końcowego PAM. W szczególności największa redundancja obliczeń dotyczy obszaru obok TX. Dlatego wzrost liczby obliczeń *CN* nie przekłada się proporcjonalnie na wzrost *CPR*.

Inną miarą, niezależną od mocy obliczeniowej używanego komputera, jest całkowita liczba obliczeń *CN*. Wartość bezwzględną tej miary dla trzech odległości przedstawiono na Rys. 8.



Rys. 7. Redundancja r dla punktów macierzy PAM zilustrowana dla obszaru w pobliży TX



Rys. 8. CN dla różnych promieni R_0

5. WNIOSKI

W artykule skupiono się na nowatorskiej metodzie tworzenia PAM i ocenie efektywności generacji takich map. Proponowane podejście opiera się na profilach terenu DTED, PEM oraz interpolacji liniowej. Tym samym opracowane rozwiązanie uwzględnia wpływ topografii na obliczanie tłumienia propagacyjnego. Z drugiej strony zaproponowane dwuetapowe podejście do PAM można uznać za uniwersalne. Oznacza to, że DTED, PEM i interpolację liniową można zastąpić innymi, które zapewniają obliczenie pełnej macierzy tłumienia. W artykule przedstawiono opis algorytmu generacji PAM na podstawie PEM dla zadanej topografii terenu. Zdefiniowano również metryki oceny skuteczności algorytmu w zależności od wielkości PAM, które stanowią podstawę do oceny proponowanego rozwiązania dla różnych PAM.

Analizując przedstawione w pracy wyniki można stwierdzić, że w zakresie do około 3 km ponad 50% punktów w PAM wyznaczana jest na podstawie PEM. W przypadku większych obszarów należy pamiętać, że większość punktów jest interpolowana, co wpływa na dokładność wyznaczenia wartości tłumienia. Zatem dokładność zaproponowanej metody wyznaczania tłumienia wzrasta z promieniem R_0. Szczegółowy opis algorytmu i analizę efektywności zaproponowanej metody generacji PAM przedstawiono w [9].

Przyszłe prace nad rozwojem PAM będą koncentrować się nad ocenie wpływu gęstości profili na dokładność szacowania tłumienia propagacji, wykorzystaniu PAM do oceny zasięgu radiowego węzłów, porównaniu metody interpolacji liniowej z innymi metodami (np. IDW [10] lub Kriging [5,6]) oraz modyfikacji algorytmu PAM dla anten kierunkowych.

LITERATURA

[1] Malon Krzysztof, Skokowski Paweł, Łopatka Jerzy. 2018. "Optimization of the MANET topology in urban area using redundant relay points". *2018 19th International Conference on Military Communications and Information Systems (ICMCIS)*: 1–4.

[2] Skokowski Paweł. 2021. *Budowanie świadomości sytuacji elektromagnetycznej w sieciach doraźnych z węzłami kognitywnymi*. Wojakowa Akademia Techniczna, Warszawa.

[3] Skokowski Paweł, Malon Krzysztof, Łopatka Jerzy 2022. "Building the electromagnetic situation awareness in MANET cognitive radio networks for urban areas". *Sensors*, 22(3): 716.

[4] Pesko Marko, Javornik Tomaž, Košir Andrej, Štular Mitja, Mohorčič Mihael. 2014. "Radio environment maps: The survey of construction methods". *KSII Transactions on Internet and Information Systems*, 8(11): 3789-3809.

[5] Kaniewski Paweł; Romanik Janusz, Golan Edward, Zubel Krzysztof. 2021. "Spectrum awareness for cognitive radios supported by radio environment maps: Zonal approach". *Applied Sciences*, 11(7): 2910.

[6] Suchański Marek, Kaniewski Paweł, Romanik Janusz, Golan Edward, Zubel Krzysztof. 2020. "Radio environment maps for military cognitive networks: Density of small-scale sensor network vs. map quality". *EURASIP Journal on Wireless Communications & Networking*, 2020: 189.

[7] Kryk Michał. 2021. "Creating radio environment maps (REMs) based on various path loss models". *2021 37th International Business Information Management Conference (IBIMA)*: 9254–9264.

[8] Kryk Michał, Malon Krzysztof, Kelner Jan M. 2021. "Algorithm of creating propagation attenuation maps based on parabolic equation method". *2021 Signal Processing Symposium (SPSympo)*: 138–142.

[9] Kryk Michał, Malon Krzysztof, Kelner Jan M. 2022. "Propagation attenuation maps based on parabolic equation method". *Sensors*, 22(11): 4063.

[10] Majewski Bartosz, Malon Krzysztof, Kryk Michał, Kelner Jan M. 2022. „Tworzenie map tłumienia propagacyjnego z wykorzystaniem metody IDW". *Przegląd Telekomunikacyjny*, 95(9).

MULTIPLEKSACJA PROMIENIOWANIA THZ Z WYKORZYSTANIEM NIE-POOSIOWYCH OPTYCZNYCH STRUKTUR DYFRAKCYJNYCH

MULTIPLEXING OF THE THZ RADIATION USING OFF-AXIS DIFFRACTIVE OPTICAL ELEMENTS

Paweł Komorowski[1]

[1] Instytut Optoelektroniki, Wojskowa Akademia Techniczna, Warszawa, pawel.komorowski@wat.edu.pl

DOI: 10.15199/59.2022.4.95

Streszczenie: **Powszechnie uważa się, że zakres THz będzie podstawą przyszłych sieci 6G i nie podlega wątpliwości, że konieczne jest opracowywanie rozwiązań zwiększających możliwości przesyłowe tego typu systemów. Przedstawione badania mają na celu eksplorację możliwości multipleksacji terahercowych kanałów transmisji danych z wykorzystaniem struktur dyfrakcyjnych, przekierowujących promieniowanie w wiele punktów ogniskowych. W celu zaprojektowania tego typu elementów, opracowano i zaimplementowano algorytmy optymalizacyjne (w tym unikatowy algorytm, bazujący na splotowej dyfrakcyjnej sieci neuronowej). Opracowane struktury mogą z powodzeniem być stosowane dla dużego zakresu częstotliwości terahercowych i sub-terahercowych, umożliwiając tym samym multipleksację kanałów transmisji danych w przyszłych sieciach telekomunikacyjnych, niezależnie od wykorzystywanego przez nie pasma.**

Abstract: **It is commonly believed, that THz waves will be the base of the future 6G telecommunication systems and all development of methods increasing the bandwidth of those systems is necessary. The novel solutions for the designing and optimization, based on the diffractive optical elements for the THz spectral range, have been proposed in this work. Described structures allow redirecting THz radiation into multiple focal points, also with frequency separation. The design and optimization algorithms for the diffractive elements have been developed and implemented (including a unique method, based on the convolutional diffractive neural network). Designed structures can work in a wide range of terahertz and sub-terahertz radiation. Therefore, they can be considered an appealing solution, regardless of the band, used by future telecommunication systems.**

Słowa kluczowe: **multipleksacja, optyczne struktury dyfrakcyjne, promieniowanie terahercowe, struktury wytwarzane technologią druku 3D**

Keywords: **3D printed structures, diffractive optical elements, multiplexing, terahertz radiation**

1. WSTĘP

Jeszcze do niedawna promieniowanie terahercowe pozostawało jednym z ostatnich niezbadanych regionów spektrum promieniowania elektromagnetycznego. Wiązało się to z brakiem odpowiednich źródeł i detektorów, mogących pracować w tym zakresie częstotliwości. W ostatnich trzech dekadach obserwuje się jednak ciągły, szybko przyspieszający rozwój tej dziedziny (Sizov 2019) i nie ma powodów, żeby przypuszczać, że w najbliższej przyszłości miałoby to ulec zmianie. Obecnie, technologie terahercowe są jednym z ważniejszych tematów badań wśród naukowców na całym świecie, znajdując coraz więcej potencjalnych i realnych zastosowań. Jednym z najbardziej perspektywicznych i najintensywniej badanych obszarów zastosowań promieniowania THz jest transmisja danych (Marczewski 2020). Przewiduje się, że bezprzewodowe łącza terahercowe będą w stanie osiągać przepływności przekraczające terabit na sekundę (Tbps), do tej pory osiągane jedynie w zaawansowanych sieciach światłowodowej komunikacji optycznej. W ostatnich latach obserwuje się również szybki rozwój piątej generacji telekomunikacji bezprzewodowej (5G), napędzany poprzez ciągle rosnące zapotrzebowanie na coraz wyższe szybkości przesyłania danych Nie jest jeszcze pewne które pasma i jakie technologie będzie wykorzystywać technologia 6G, ale częstotliwości terahercowe wydają się być najbardziej prawdopodobnym rozwiązaniem (O'Hara et al. 2019).

Wraz z rozwojem technologii terahercowej (w tym nadajników, detektorów i modulatorów), pozwalającej na przekraczanie kolejnych barier szybkości, dystansu i efektywności przesyłania informacji, rośnie również zainteresowanie adaptowaniem znanych z innych systemów rozwiązań zwiększających pojemność informacyjną łączy danych. Należy tu wymienić implementację zaawansowanych schematów modulacji amplitudowo-fazowej, zwiększanie liczby nadajników i odbiorników (MIMO – ang. multiple input, multiple output) czy zwielokrotnianie liczby kanałów przesyłu danych wewnątrz pojedynczego medium transmisyjnego (multipleksacja). To ostatnie rozwiązanie stanowiło główną inspirację do podjęcia badań przedstawionych w niniejszej pracy, której celem jest opracowanie struktur dyfrakcyjnych działających w zakresie promieniowania terahercowego, umożliwiających multipleksację i demultipleksację sygnałów terahercowych, propagujących się w pojedynczym kanale transmisyjnym. Tego typu struktury muszą uwzględniać nie-poosiowe przekierowanie wiązki oraz często asymetryczny rozkład natężenia promieniowania. W takim przypadku do uzyskania pożądanego działania konieczne staje się stosowanie iteracyjnych metod projektowania oraz dodatkowych metod optymalizacyjnych.

W niniejszej pracy, oprócz typowych metod iteracyjnych, zaproponowane zostało nowatorskie podejście, wykorzystujące dyfrakcyjną sieć neuronową, która optymalizuje pojedynczy element dyfrakcyjny w celu osiągnięcia zadanego rozkładu natężenia w płaszczyźnie wyjściowej układu. Splotowe sieci neuronowe (CNN – ang. convolutional neural network) są powszechnie wykorzystywane w widzeniu maszynowym i rozpoznawaniu obiektów. Należy jednak zauważyć, że w przypadku tego

typu sieci to jądra splotów podlegają zmianom i optymalizacji w procesie uczenia. Zupełnie inna sytuacja występuje w przypadku algorytmów propagacji, gdzie jądro zdefiniowane jest przez całkę dyfrakcyjną. Tym niemniej, sama operacja matematyczna splotu nie ulega zmianom i z tej perspektywy proponowana sieć jest obliczeniowo podobna do sieci splotowych. Jest to więc sieć hybrydowa, korzystająca z operacji splotu oraz realizująca dyfrakcję światła.

Metody zwiększania pojemności informacyjnej terahercowych kanałów transmisji danych proponowane w niniejszej pracy są realizowane poprzez przekierowywanie i skupianie promieniowania THz w więcej niż jedną plamkę ogniskową. W ten sposób padające na strukturę promieniowanie, pochodzące z pojedynczego kanału transmisji może zostać podzielone na pewną liczbę kanałów odseparowanych przestrzennie. Kanały te mogą różnić się między sobą częstotliwością promieniowania lub szczeliną czasową, w której się propagują. Na tej podstawie zostały zaproponowane terahercowe elementy dyfrakcyjne, realizujące (de)multipleksację z podziałem częstotliwości lub czasu. Warto zauważyć, że po odwróceniu biegu promieniowania, analogiczne struktury mogą łączyć sygnały pochodzące z pewnej ilości nadajników w jeden kanał transmisji.

2. PROJEKTOWANIE I WYTWARZANIE STRUKTUR

Opis działania struktury dyfrakcyjnej, a także wyznaczenie rozkładu natężenia powstającego za nią, wymagają poprawnej metody opisu propagacji promieniowania. Bazuje ona najczęściej na tzw. całkach dyfrakcyjnych, transformujących zespolone amplitudy pola uwzględniając odległość propagacji. Symulacje przedstawione w niniejszej pracy zostały przeprowadzone z wykorzystaniem splotowej metody propagacji promieniowania (Sypek 1995). Do projektowania struktur wykorzystano dwa typy algorytmów optymalizacyjnych – iteracyjny (Siemion et al. 2020) oraz bazujący na sieci neuronowej (Surma et al. 2021).

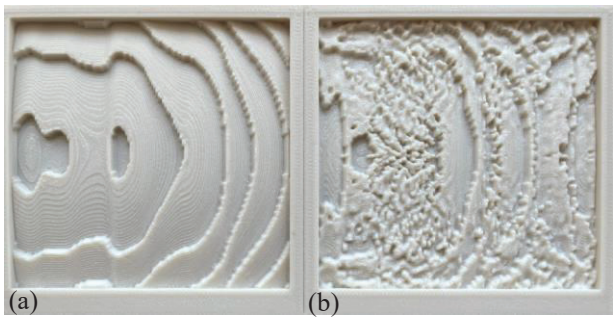
Rozmiary detali struktur dyfrakcyjnych zależą od projektowej długości fali, która w przypadku promieniowania terahercowego rzędu pojedynczych milimetrów lub ułamków milimetra. Jest to już rozdzielczość osiągana przez wiele powszechnie dostępnych i wydajnych kosztowo metod druku przestrzennego. W dodatku istnieje cała gama materiałów dostępnych do druku 3D, równocześnie posiadających dobre, bądź bardzo dobre parametry optyczne w zakresie THz (Squires and Lewis 2018). Mowa tu o między innymi o poliamidzie 12 (PA 12), polistyrenie wysokoudarowym (HIPS) czy cyklicznym kopolimerze olefinowym (COC, skomercjalizowany jako TOPAS).

3. WYNIKI EKSPERYMENTALNE

3.1. Multipleksacja z podziałem czasu

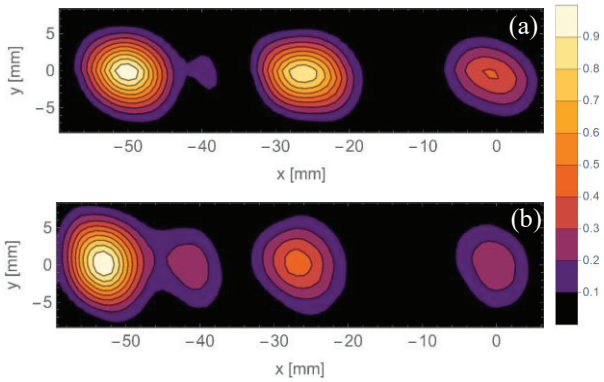
Koncepcja multipleksacji sygnałów z podziałem czasu z wykorzystaniem struktur dyfrakcyjnych polega na jednoczesnym skupianiu wiązki w różnych odległościach od środka struktury. Dzięki temu promieniowanie trafia na poszczególne detektory w różnych chwilach, co umożliwia ich demultipleksację. Struktury wieloogniskowe umożliwiałyby zwiększenie szybkości skanowania poprzez zwielokrotnienie ilości informacji docierającej do detektorów w różnych chwilach czasowych. Kluczowym aspektem, umożliwiającym rozróżnienie sygnałów pochodzących z poszczególnych ognisk, jest czas przelotu promieniowania od emitera do detektora. Widać więc, że rozumowanie to można przenieść również na grunt telekomunikacji terahercowej oraz multipleksacji kanałów transmisji.

W ramach pracy zaprojektowano struktury przekierowujące promieniowanie terahercowe w trzy plamki ogniskowe. Przeanalizowano symetryczne oraz asymetryczne rozkłady plamek względem głównej osi optycznej elementu. Na przypadku soczewki trój-ogniskowej został również przetestowany unikalny algorytm optymalizacyjny oparty o sieć neuronową. Zdjęcia dwóch z wytworzonych struktur przedstawia *Rys. 1*.



Rys. 1 Fotografie wytworzonych soczewek trój-ogniskowych zaprojektowanych z wykorzystaniem: (a) algorytmu iteracyjnego, (b) sieci neuronowej.

Struktury zaprojektowane z wykorzystaniem obydwu algorytmów optymalizacyjnych zostały zweryfikowane eksperymentalnie. Element dyfrakcyjny oświetlany był falą płaską, uformowaną przez zwierciadło paraboliczne. Promieniowanie padające na zwierciadło pochodziło z powielacza częstotliwości, bazującego na diodach Schottky'ego. Pomiar dokonywany był z wykorzystaniem diody Schottky'ego, umieszczonej na zmotoryzowanym uchwycie, poruszającym się w trzech osiach. Umożliwia to skanowanie natężenia w dowolnej płaszczyźnie lub osi. Poprzeczne przekroje natężenia w płaszczyźnie prostopadłej do osi optycznej przedstawia *Rys. 2*.

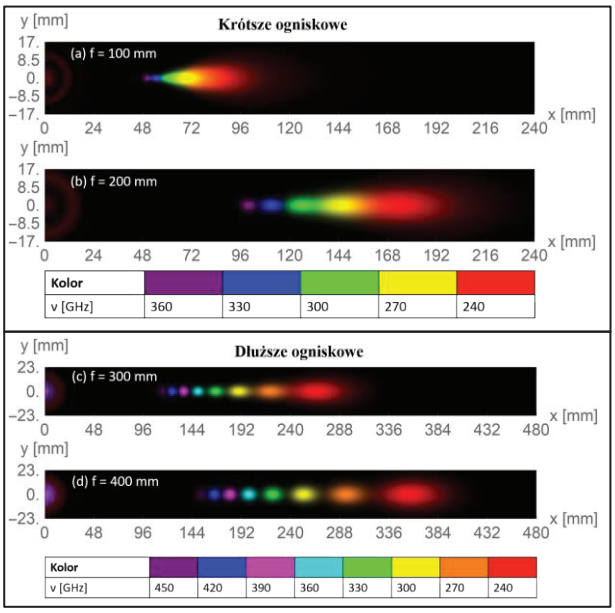


Rys. 2 Poprzeczne przekroje natężeń obserwowane za soczewkami trójogniskowymi optymalizowanymi (a) algorytmem iteracyjnym oraz (b) siecią neuronową.

Dla obydwu struktur można dostrzec trzy wyraźne ogniska – w przypadku algorytmu iteracyjnego (a) ich pozycje pokrywają się niemal dokładnie z projektem, natomiast w przypadku sieci neuronowej (b) lewe i środkowe ogniska są nieznacznie bardziej oddalone od osi optycznej. Można jednakże zaobserwować, że dla struktury projektowanej algorytmem iteracyjnym ogniska przyjmują eliptyczny kształt i są rozciągnięte w poziomie (zostało to przewidziane na etapie symulacji), który to efekt został zniwelowany przy optymalizacji siecią neuronową.

3.2. Multipleksacja z podziałem częstotliwości

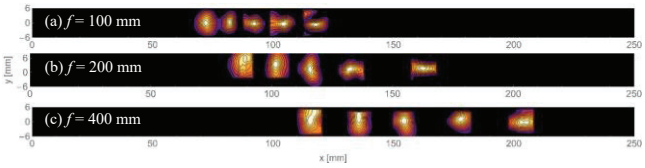
Koncepcja multipleksacji w dziedzinie częstotliwości zakłada separację przestrzenną kanałów różniących się częstotliwością (długością fali) przesyłanego sygnału. W podejściu optyki dyfrakcyjnej może to zostać zrealizowane poprzez strukturę skupiającą promieniowanie w różnych punktach jednej płaszczyzny ogniskowej. Położenie punktów ogniskowych zależeć natomiast będzie od częstotliwości skupianego promieniowania terahercowego. Jedną z najprostszych struktur optycznych realizujących takie zadanie jest siatka dyfrakcyjna. Separacja poszczególnych rzędów ugięcia siatek dyfrakcyjnych następuje w strefie dalekiej (Fraunhofera), co sprawia, że projektowany układ optyczny musiałby być bardzo dużych wymiarów. Z tego względu konieczne jest zoptymalizowanie działania struktury dla bliższych odległości. Zostało to zrealizowane poprzez nałożenie na siebie struktury siatki dyfrakcyjnej i soczewki skupiającej promieniowanie w konkretnej odległości ogniskowej. W tym przypadku separacja rzędów ugięcia nastąpi w płaszczyźnie ogniskowej struktury. Zaprojektowano cztery struktury będące połączeniem binarnej siatki dyfrakcyjnej i soczewek skupiających różniące się ogniskowymi użytych soczewek, zmieniającymi się od $f_1 = 100$ mm poprzez $f_2 = 200$ mm i $f_3 = 300$ mm, aż po $f_4 = 400$ mm. Bazowa



Rys. 3 Obliczone rozkłady natężenia w płaszczyźnie ogniskowej dla struktur o ogniskowych (a) 100 mm, (b) 200 mm, (c) 300 mm, (d) 400 mm.

częstotliwość, dla której projektowana była struktura to $\nu = 300$ GHz, co przekłada się na długość fali $\lambda = 1$ mm. Teoretyczne położenia ognisk otrzymane w wyniku modelowania numerycznego przedstawia *Rys. 3*. Poszczególne panele na *Rys. 3*. przedstawiają fragmenty płaszczyzn ogniskowych soczewek wycięte w taki sposób, aby oś optyczna znajdowała się na środku ich lewej krawędzi. We wszystkich przypadkach obserwowana jest wyraźna separacja ognisk powstających dla poszczególnych częstotliwości promieniowania. Wokół osi optycznej można dostrzec nieugięte promieniowanie zerowego rzędu w postaci koncentrycznych kół i pierścieni.

Zaprojektowane elementy zostały wytworzone metodami druku przestrzennego oraz zweryfikowane eksperymentalnie. Badano położenie ognisk w zależności od częstotliwości padającego promieniowania, co przedstawia *Rys. 4*.



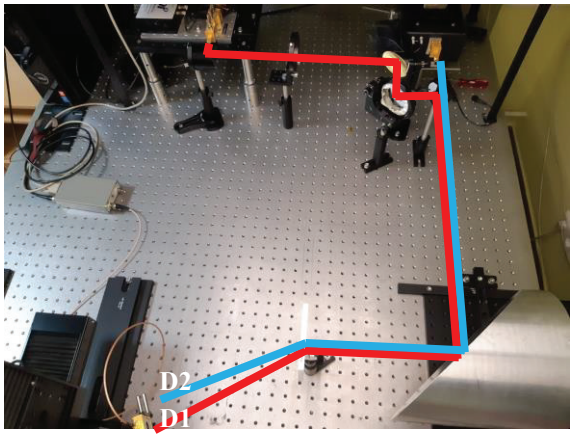
Rys. 4 Zmierzone rozkłady natężenia odpowiadające kolejnym plamkom ogniskowym dla struktur o różnych ogniskowych.

Skany wykonano niezależnie dla każdego ogniska i każdej częstotliwości, a następnie naniesiono na pojedynczy wykres dla każdej struktury (o różnych ogniskowych). Puste przestrzenie pomiędzy ogniskami zostały wypełnione zerami. W pierwszej kolejności można ocenić, że położenia ognisk i ich separacja wynikająca z pomiarów eksperymentalnych odpowiada symulacjom numerycznym. W przypadku ogniskowej $f_1 = 100$ mm plamki ogniskowe dla kolejnych kanałów częstotliwościach różniących się o $d\nu = 30$ GHz są ledwie rozróżnialne i częściowo nachodzą na siebie. Odległości między kolejnymi maksimami nie przekraczają kilkunastu milimetrów, co mając na uwadze przeciętny rozmiar detektorów rzędu co najmniej kilku milimetrów praktycznie uniemożliwia ich jednoczesną, niezależną detekcję. Zgodnie z przewidywaniami teoretycznymi, separacja ognisk poprawia się wraz ze zwiększaniem długości ogniskowej. W przypadku dłuższych ogniskowych wszystkie plamki są już klarownie rozdzielone. Dla ogniskowej $f_4 = 400$ mm odległości pomiędzy kolejnymi ogniskami przekraczają kilkanaście milimetrów i dla żadnej pary częstotliwości nie nachodzą na siebie.

3.3. Równoległa transmisja danych

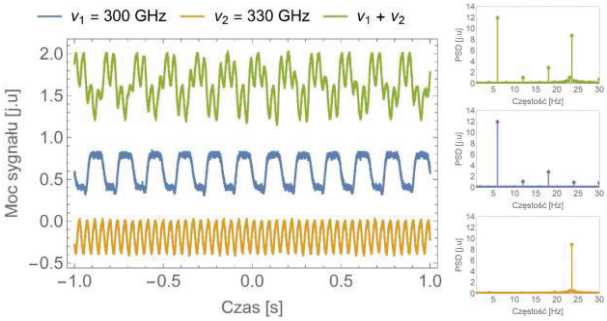
Pomiary przedstawione w powyższym podrozdziale 3.2 zostały również uzupełnione o eksperyment przedstawiający równoległą transmisję dwóch kanałów danych pojedynczym kanałem optycznym. Wyniki w nim przedstawione dobitnie pokazują, że zaprojektowana struktura przekierowuje promieniowanie terahercowe o różnych częstotliwościach w różne punkty przestrzeni. Działanie to może więc zostać wykorzystane do separacji prze-

strzennej kanałów transmisji danych o różnych częstotliwościach, czyli multipleksacji z podziałem częstotliwości. W celu unaocznienia takiej funkcjonalności zaprojektowano eksperyment, w którym promieniowanie terahercowe o dwóch różnych częstotliwościach propagowało się wspólnym torem optycznym, a następnie oświetlało zaprojektowaną strukturę, gdzie częstotliwości były separowane przestrzennie. Pomiar wykonywany był w dwóch punktach D1 oraz D2 odpowiadających miejscom występowania plamek ogniskowych dla obydwu częstotliwości (odpowiednio ν_1 = 300 GHz oraz ν_2 = 330 GHz). Zdjęcie układu eksperymentalnego do dwukanałowej terahercowej transmisji danych przedstawia *Rys. 5*.



Rys. 5 Fotografia układu eksperymentalnego wykorzystanego do przeprowadzenia dwukanałowej transmisji danych.

Poszczególne częstotliwości promieniowania były modulowane niezależnie od siebie (dla rozróżnienia odpowiednio częstościami nazwanymi ω_1 oraz ω_2), co symulowało przesyłanie niezależnych ciągów danych. Przebiegi czasowe zarejestrowane w poszczególnych ogniskach oraz referencyjny przebieg zmierzony w torze optycznym wraz z odpowiadającymi im transformatami Fouriera przedstawia *Rys. 6*.



Rys. 6 Przebiegi czasowe sygnałów o częstotliwościach ν_1 = 300 GHz (niebieska linia), ν_2 = 330 GHz (pomarańczowa linia) oraz referencyjny przy włączonych obydwu nadajnikach (linia zielona).

Z łatwością można zaobserwować, że w poszczególnych ogniskach rejestrowany jest jedynie przebieg odpowiadający zadanej częstotliwości promieniowania przekierowanej w daną plamkę ogniskową. Co więcej, obecność promieniowania o innej częstotliwości w optycznym kanale transmisji danych nie wpływa w żaden sposób na obserwowany przebieg. Jest to wyraźnie widoczne w widmach otrzymanych po transformacji Fouriera (PSD), w których widoczne są jedynie częstości modulacji promieniowania o częstotliwości 300 GHz w pozycji D1 oraz 330 GHz w pozycji D2.

4. PODSUMOWANIE

Głównym celem omawianych badań było zaprojektowanie, wytworzenie i zbadanie właściwości optycznych struktur dyfrakcyjnych, pozwalających na zwielokrotnianie pojemności informacyjnej terahercowych kanałów transmisji danych poprzez multipleksację w dziedzinie częstotliwości oraz czasu. W niniejszej pracy zostały przedstawione i zweryfikowane praktycznie iteracyjne metody projektowania i optymalizacji takich elementów. Zaproponowany został również unikatowy algorytm optymalizacyjny, wykorzystujący sieci neuronowe, co pozwoliło na zaprojektowanie struktur o niespotykanych wcześniej rozkładach fazowych. Wszystkie zaprojektowane struktury dyfrakcyjne zostały wytworzone technikami druku przestrzennego oraz poddane weryfikacji eksperymentalnej, która potwierdziła ich poprawne, zgodne z założeniami działanie. Ponadto, praktyczna możliwość wykorzystania opracowywanych elementów została wykazana w eksperymencie realizującym równoległą transmisję dwóch kanałów danych, różniących się częstotliwością wykorzystywanego promieniowania, w pojedynczym torze optycznym. Kanały te zostały skutecznie rozdzielone za pomocą struktury multipleksującej promieniowanie THz z podziałem częstotliwości.

LITERATURA

Marczewski, Jacek 2020. “Digital Communication – Optical vs. THz Links.” *Opto-Electronics Review* 28 (3): 176–81.

O’Hara, John F., Sabit Ekin, Wooyeol Choi, and Ickhyun Song. 2019. “A Perspective on Terahertz Next-Generation Wireless Communications.” *Technologies* 7 (2): 43.

Siemion, Agnieszka, Paweł Komorowski, Mateusz Surma, Izabela Ducin, Piotr Sobotka, Michał Walczakowski, and Elzbieta Czerwińska. 2020. “Terahertz Diffractive Structures for Compact In-Reflection Inspection Setup.” *Optics Express* 28 (1).

Sizov, F F. 2019. “Optoelectronics and Optoelectronic Devices Brief History of THz and IR Technologies” 6582: 67–79.

Squires, A. D., and R. A. Lewis. 2018. “Feasibility and Characterization of Common and Exotic Filaments for Use in 3D Printed Terahertz Devices.” *Journal of Infrared, Millimeter, and Terahertz Waves* 39 (7): 614–35.

Surma, Mateusz, Mateusz Kaluza, Patrycja Czerwińska, Paweł Komorowski, and Agnieszka Siemion. 2021. “Neural-Network Based Approach to Optimize THz Computer Generated Holograms.” *Photonics Letters of Poland* 13 (4): 88–90.

Sypek, Maciej. 1995. “Light Propagation in the Fresnel Region. New Numerical Approach.” *Optics Communications*.

MODELOWANIE WIELOZAKRESOWEJ SIECI TAKTYCZNEJ Z DYNAMICZNYM ZARZĄDZANIEM WIDMEM

MODELLING OF MULTIBAND TACTICAL NETWORK WITH DYNAMIC SPECTRUM MANAGEMENT

Piotr Gajewski[1]; Jerzy Łopatka[2]; Piotr Łubkowski[3];

[1] Wojskowa Akademia Techniczna, Warszawa, piotr.gajewski@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, jerzy.lopatka@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Warszawa, piotr.lubkowski@wat.edu.pl

DOI: 10.15199/59.2022.4.96

Streszczenie: **W artykule przedstawiono rozwiązanie wysokiej jakości symulatora taktycznej sieci mobilnej MAENA. Omówiono zasady budowy modelu sieci z zaawansowanymi technikami efektywnego wykorzystania zasobów widmowych i unikania zakłóceń. Proponowane rozwiązania zostały sprawdzone w heterogenicznych scenariuszach sieciowych. W pracy przedstawiono przykładowe wyniki symulacji w złożonym środowisku propagacyjnym.**

Abstract: **The article presents a solution of a high-quality MAENA tactical mobile network simulator. Principles of the network model construction with advanced techniques of efficient use of spectral resources and disturbances avoiding are discussed. The proposed solutions have been tested in heterogeneous network scenarios. The paper presents example simulation results in complex propagation environment.**

Słowa kluczowe: **mobilne sieci ad hoc, dynamiczne zarządzanie widmem, radio kognitywne, komunikacja wielopasmowa.**

Keywords: **Mobile Ad hoc Networks, Dynamic Spectrum Management, Cognitive Radio, Multi-band Communication.**

1. WSTĘP

Łączność bezprzewodowa jest jedynym środkiem komunikacyjnym w systemach funkcjonujących w warunkach dynamicznie zmiennego środowiska. W szczególności dotyczy to taktycznych systemów wojskowych, gdzie zmieniają się sposoby funkcjonowania systemu użytkownika wynikające ze zmian: konfiguracji systemu połączeń, środowiska propagacyjnego oraz wysokiej mobilności środków komunikacyjnych, przy dużym zróżnicowaniu usług i ograniczeń widmowych wynikających zarówno z dostępnych zasobów, jak i aktywnego przeciwdziałania oponenta.

Spełnienie wymagań współczesnego pola walki wywołuje konieczność stosowania nowych technologii komunikacyjnych, zapewniających adaptacyjne dostosowanie systemu łączności do środowiska. Wdrażanie nowych technologii wymaga posiadania narzędzi komputerowych zarówno do badania systemów oraz ich planowania i zarządzania w czasie realizacji misji. Najbardziej obiecującą technologią, wspierającą realizację złożonych wymagań w heterogenicznym systemie zintegrowanym, zapewniającym dowodzenie, sterowanie, komunikację, wywiad, i rozpoznanie C4ISR (Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance)), są mobilne sieci ad-hoc (MANET – Mobile Ad hoc NETwork) z dynamicznym zarządzaniem widmem (DSM – Dynamic Spectrum Management), wykorzystujące algorytmy sztucznej inteligencji (AI – Artificial Inteligence) w urządzeniach radia kognitywnego (CR- Cognitive Radio) [1].

Aby zastosować MANET w obszarze taktycznym, należy wziąć pod uwagę rodzaj sił i ich organizację, gdyż różne sieci zapewniają komunikację między węzłami na danym szczeblu, a także między różnymi szczeblami dowodzenia, uwzględniając różnorodność źródeł danych. Dywersyfikacja rodzajów przesyłanych informacji, wykorzystywanych pasm częstotliwości, różnych kształtów fal (waveform), priorytety węzłów i usług tworzą heterogeniczne środowisko o dużej złożoności ruchu telekomunikacyjnego. Aby uniknąć przerw w komunikacji spowodowanych przeszkodami terenowymi lub dużymi odległościami, wykorzystuje się węzły wspierające umieszczone na dronach. Coraz szerzej stosuje się też bezprzewodowe sieci sensorowe, od systemów monitorowania środowiska, po osobiste wyposażenie żołnierzy.

W [2] autorzy przedstawili ogólną charakterystykę kognitywnych sieci MANET oraz opis symulatora CORASMA (COgnitive RAdio for dynamic Spectrum MAnagement) dedykowanego badaniu rozwiązań kognitywnych dla współdzielenia widma w pasmie UHF z wykorzystaniem DSM. W tym rozwiązaniu celem DSM jest optymalizacja przydziału dostępnego widma i jego wykorzystania w czasie, biorąc pod uwagę mobilność i potencjalnie nieoczekiwane zakłócenia, zapewniając w ten sposób, że globalny zasób widma jest zawsze używany we właściwym miejscu i czasie.

W pracy opisano problemy związane z modelowaniem takich sieci na przykładzie symulatora opracowanego w ramach międzynarodowego projektu Europejskiej Agencji Obrony - MAENA (Multi-bAnd Efficient Networks for ad hoc communicAtions).

2. SYMULATOR MAENA

2.1. Modelowanie sieci MANET

Ogólnie model jest uproszczoną reprezentacją systemu rzeczywistego, odzwierciedlającą jego główne ce-

chy. Złożoność modelu zależy od wierności odwzorowania szczegółów systemu w środowisku obliczeniowym zgodnie z badanymi problemami.

Do budowania modeli symulacyjnych często wykorzystuje się doświadczenia z istniejących rozwiązań w narzędziach komercyjnych, takich jak OMNET++, NS2, OPNET itd., biorąc pod uwagę specyficzne właściwości systemów opartych na MANET. Każde z tych narzędzi umożliwia modelowanie określonych aspektów MANET, a ich ograniczenia omówiono w [4]. Aby umożliwić modelowanie pełnego stosu sieci MANET, z nowymi aplikacjami i z uwzględnieniem szczegółowych modeli środowiska radiowego, konieczna jest budowa symulatora dedykowanego. Wspomniany symulator CORASMA został wykorzystany do zbadania różnych aspektów dotyczących DSA w CR MANET na poziomie taktycznym. Symulator ten został zbudowany z wykorzystaniem środowiska OMNET++ i wybranych procedur biblioteki INET/MANET jako tła do opracowania sieci składającej się z węzłów wyposażonych w funkcjonalność CR.

2.2. Założenia projektu MAENA

Wdrażanie systemów łączności taktycznej w bieżących operacjach wojskowych napotyka na niemal krytyczny niedobór widma. Przyczynami tego niedoboru są: rosnące wymagania dotyczące szerokości pasma nowowprowadzonych wojskowych systemów komunikacyjnych o zwiększonych wymaganiach dotyczących szybkości transmisji danych oraz rosnące potrzeby cywilne dotyczące wykorzystania widma.

Aby sprostać tym wyzwaniom, w projekcie MAENA przyjęto trojakie podejście:

- dynamiczne zarządzanie widmem (DSM) w celu elastycznego wykorzystania widma w czasie wykonywania misji i unikanie zakłóceń;

- komunikacja wielopasmowa w celu poprawy dostępności widma i odporności łączności na zmiany środowiska propagacyjnego;

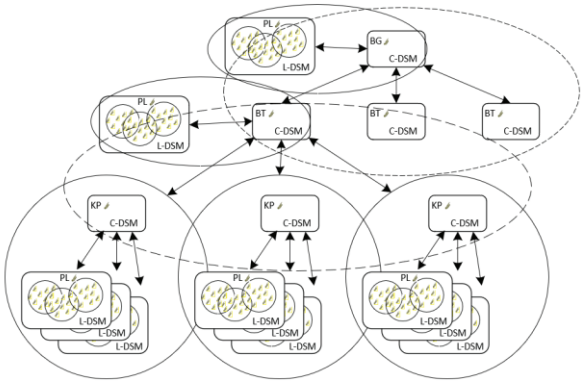
- ścisła integracja danych i płaszczyzn sterowania do poprawy efektywności rozwiązań komunikacyjnych.

2.3. Charakterystyka symulatora MAENA

Celem symulatora MAENA jest modelowanie mobilnych taktycznych sieci VHF i UHF ad hoc (Rys.1) do oceny wydajności rozwiązań opracowanych i zaimplementowanych w ramach projektu. Symulator MAENA wykonuje realistyczne modelowanie komunikacji za pomocą pełnego stosu protokołów realizowanych w węzłach radiowych, w połączeniu z bardzo wiernym modelowaniem propagacji radiowej i przetwarzaniem sygnału radiowego w warstwie fizycznej na poziomie próbek IQ. Symulator umożliwia zdefiniowanie węzła wyposażonego w maksymalnie cztery interfejsy radiowe obsługujące komunikację w kanałach UHF i/lub VHF.

Główną cechą symulatora MAENA jest możliwość modelowania i testowania różnych rozwiązań wdrażanych w ramach koncepcji radia nowej generacji, w tym waveformów, stosów protokołów, aplikacji i zmiennych polityk dotyczących ich zachowania i sposobu użytkowania. Obejmują one również rozwiązania dedykowane do zarządzania siecią, w tym zarówno lokalny DSM (L-DSM), jak i koordynowany, centralny DSM (C-DSM) [5].

Koncepcja X-Band DSM została zaproponowana w projekcie MAENA, aby obsłużyć wiele waveformów występujących w węźle radiowym. Międzypasmowe DSM koordynuje sensing i wykorzystanie widma w różnych sieciach radiowych, albo za pośrednictwem jednostek L-DSM, albo bezpośrednio współdziałając z indywidualnymi sensorami widma.



Rys. 1. Uogólniony schemat taktycznego systemu łączności

Każdy węzeł jest wyposażony w narzędzia autokonfiguracji sieci (grupowania), rozpoznawania środowiska radiowego, zarządzania zasobami (w tym widmem), adaptacyjnej konfiguracji struktury sygnału.

Dla realizacji funkcji adaptacji i optymalizacji oraz do oceny charakterystyk systemu opracowany został zbiór metryk umożliwiających wyznaczanie chwilowych wartości parametrów transmisji i ich statystycznych właściwości, ale również zdefiniowane zostały metryki globalnej oceny realizacji funkcji komunikacji (metryki efektywności systemu) [6].

3. BUDOWANIE SCENARIUSZA

Symulator wykorzystuje interfejs człowiek-maszyna (HMI) do definiowania scenariusza, struktury jednostek wojskowych, sieci komunikacyjnych i ich parametrów, a także do wczytywania tła mapy i układu zdefiniowanych węzłów na załadowaną mapę.

Pozycje i trasy węzłów tworzone są przy użyciu cyfrowej mapy terenu (DTM – Digital Terrain Model) zawierającej również cyfrowy model terenu (DTEM - Digital Terrain Elevation Model) oraz przeszkody terenowe, takie jak np. budynki, lasy itd. Szczegółowe modele propagacyjne umożliwiają wyznaczanie strat propagacji z założoną dokładnością. Zdefiniowane są również przepływy ruchu użytkowników, które mają być przekazywane przez sieć. Dodatkową funkcją dostępną w symulatorze jest możliwość tworzenia niewspółpracujących węzłów, które generują zakłócenia degradujące komunikację.

Dla tak przygotowanego scenariusza możliwe jest rozpoczęcie symulacji z dedykowanym ziarnem symulacji. Aby ułatwić analizę, otrzymane wyniki przedstawiane są w formie graficznej. Badania przeprowadzono dla dwóch rodzajów scenariuszy:

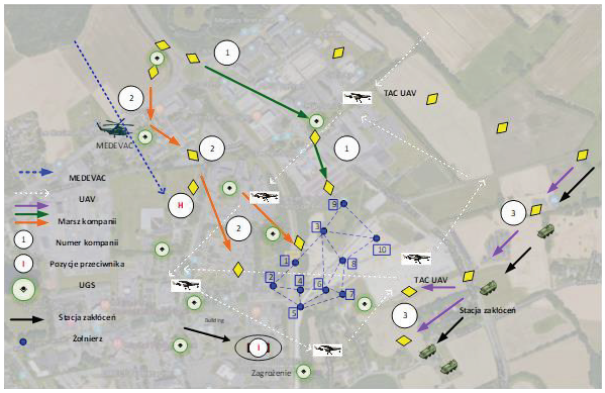
- scenariusze techniczne, składające się zwykle z kilku węzłów radiowych, przeznaczone do analizy określonej cechy funkcjonowania sieci radiowej;
- scenariusze operacyjne, przedstawiające analizy efektywności sieci dla złożonych operacji sił zbrojnych.

4. WYNIKI BADANIA PRZYKŁADOWYCH SCENARIUSZY OPERACYJNYCH

4.1. Scenariusz operacji miejskiej

W dalszej części przedstawiono wyniki symulatora MAENA, uzyskane dla wojskowego scenariusza operacji pokojowej, prowadzonej wewnątrz miasta w sytuacji aktywnych działań przeciwnika, takich jak atak ogniowy, zakłócanie i ewakuacja medyczna.

Scenariusz obejmuje system komunikacyjny składający ze 183 węzłów. Zdefiniowano 4 sieci UHF i 28 sieci VHF. Łączna liczba interfejsów radiowych przewidzianych do wykorzystania podczas symulacji wynosi 419 (183 dla pasma UHF, 236 dla pasma VHF). Uproszczony schemat scenariusza operacyjnego przedstawiono na rysunku 2.

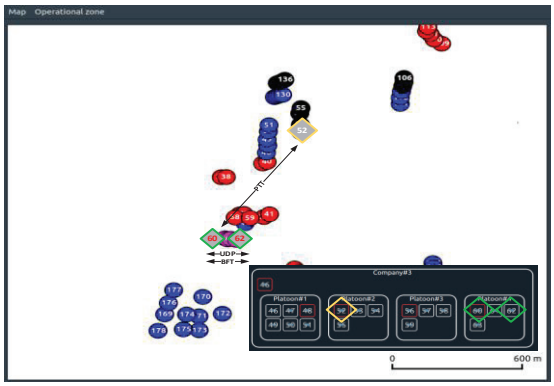


Rys. 2. Scenariusz operacji miejskiej

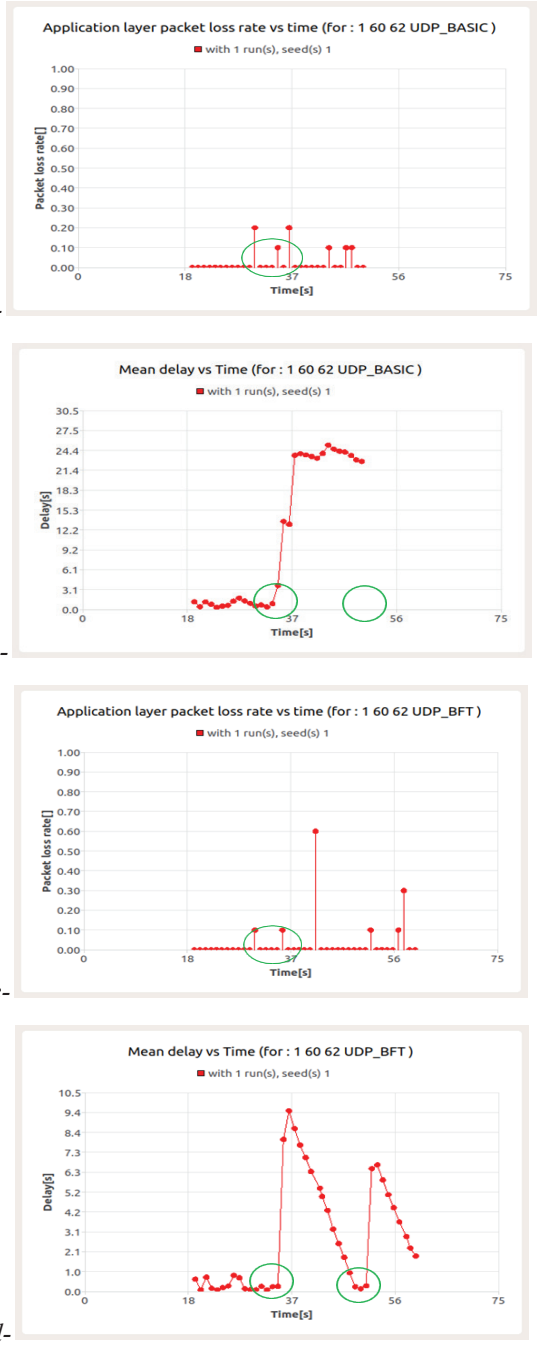
W operacji można wyróżnić 3 fazy: przemieszczenia jednostek do obszaru działań (kolorowe strzałki określają trasy poszczególnych jednostek), działanie spieszonej jednostki w obszarze rozwinięcia (granatowe kropki), ewakuacja, z użyciem śmigłowca, żołnierzy rannych podczas wymiany ognia z przeciwnikiem. Operacja jest wspierana przez bezzałogowe sensory naziemne UGS (zielone kółka) oraz dwa UAV realizujące transfer danych z UGS do węzłów naziemnych. Poniżej zamieszczono wybrane wyniki symulacji.

4.2. Transfer pomiędzy węzłami w obszarze rozwinięcia

W pierwszym przypadku przedstawiono wyniki symulacji dla transferu pomiędzy wybranymi węzłami (Rys. 3). Zaprezentowani następujące sesje: PTT (Push To Talk) realizowaną między 25 a 30 s symulacji, UDP trwającą od 20 do 50 s oraz BFT (Blue Force Tracking) trwającą między 20 a 60 s. Wyniki przedstawiono na rysunku 4. Zaobserwowane wzrosty straty i wzrost opóźnień wynikają z chwilowej utraty łączności w wyniku przesłonięcia terenu na drodze propagacji.



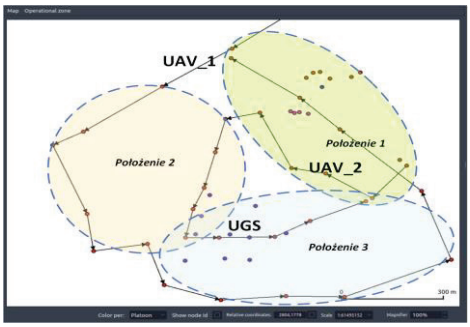
Rys. 3. Transfer informacji między węzłami 60 – 62 i 60-52



Rys. 4. Chwilowe straty pakietów i opóźnienia dla transmisji UDP (a,b) i BFT (c, d)

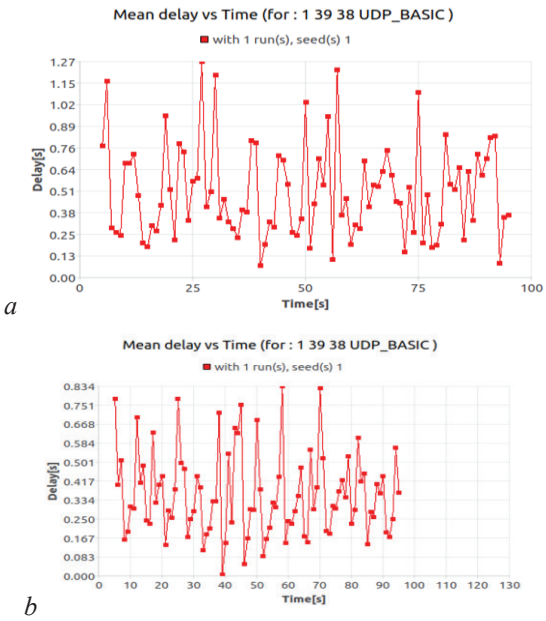
4.3. Transfer pomiędzy UAV

W tym przypadku scenariusz badawczy zakłada przesyłanie meldunków w postaci komunikatów UDP pomiędzy dwoma węzłami UAV_1 i UAV_2, które przemieszczają się po założonych trasach (Rys. 5). Przyjęto, że węzeł UAV_1 przesyła pakiety UDP o rozmiarze 500B z częstotliwością co 1 sek. Dla tak przyjętych uwarunkowań przeprowadzone zostały symulacje, których celem było oszacowanie strat pakietów oraz opóźnienia transferu danych w relacji UAV_1 – UAV_2.



Rys. 5. Scenariusz badawczy z transferem UAV

W tym scenariuszu nie zaobserwowano strat pakietów z uwagi na stosunkowo niewielkie odległości pomiędzy węzłami i brak przeszkód terenowych. Chwilowe zmiany opóźnień pakietów przedstawione są na Rys. 6.



Rys. 6. Chwilowa zmienność opóźnienia dla a- większej i b- mniejszej odległości węzłów

Zmiany opóźnienia są związane ze zmianą trajektorii lotu. Wartości średnie są większe w przypadku a, gdy węzły znajdują się w oddalonych obszarach 1 i 3 (Rys. 5).

5. WNIOSKI

Wysokiej jakości symulacja sieci MANET jest krytycznym etapem projektowania rzeczywistego systemu, umożliwiającym weryfikację przyjętych założeń, modelowanie rzeczywistego środowiska radiowego, używanych urządzeń, stosów protokołów i proponowanych mechanizmów zarządzania widmem. Ponieważ wykorzystanie widma jest ściśle związane ze strukturą sieci, liczbą węzłów i ich rozmieszczeniem, dostępne zasoby widmowe są głównym czynnikiem ograniczającym wykorzystanie poszczególnych usług, wpływając na wydajność sieci przy zmiennym ruchu.

Ponieważ prezentowane rozwiązanie jest unikatowe, wyniki naszych badań nie mogą być porównywane z innymi opublikowanymi rozwiązaniami poświęconymi celom cywilnym. Proponowane narzędzie symulacyjne zostało zweryfikowane i zatwierdzone poprzez serię testów i eksperymentów.

6. PODZIĘKOWANIA

Za znaczący wkład i wspólny wysiłek w realizację projektu MAENA Autorzy dziękują wszystkim partnerom: Królewska Akademia Wojskowa, RMA, Belgia, Uniwersytet w Oulu, UoO, Finlandia, Thales SIX GTS Francja, TF, Francja, Thales Deutschland, TD, Niemcy, Fraunhofer FKIE, FKIE, Niemcy, Rohde & Schwarz, R&S, Niemcy, Stiftelsen SINTEF, Norwegia, Instytut Badań Obronnych, FFI, Norwegia, Wojskowa Akademia Techniczna WAT, Polska, Wojskowy Instytut Łączności WIŁ, Polska, Radmor, Polska, Instytut Badań Obronnych, FOI, Szwecja, Saab AB, Szwecja.

Niniejsza praca badawcza została wykonana w ramach Projektu MAENA EDA nr B-1476-IAP4-GP i współfinansowana przez Wojskową Akademię Techniczną w ramach Projektu Badawczego nr. UGB/22-744/2022/WAT „Zastosowania wybranych technik informatyki, komunikacji i rozpoznania na obszarach cywilnych i wojskowych”.

LITERATURA

[1] Brubank J.L.; Chomento P.F.; Haberman B.K.; Kasch W.T.: Key Challenges of Military Tactical Networking and the Elusive Promise of MANET Technology. IEEE Comm. Mag. 2006, 44, 39–45.

[2] Gajewski P.; Łopatka J.; Malon K.: HiFi Symulator Mobilnych Sieci Doraźnych Manet z Węzłami Kognitywnymi, Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne, 8-9/2014, str. 1202-1209.

[3] Gajewski P.; Łopatka J.; Łubkowski P.: Performance Analysis of Public Safety Cognitive Radio MANET for Diversified Traffic, Sensors 2022, 22(5), 1927.

[4] Jie L.; Wang L.: Simulation tools for MANETs: A systematic survey. Int. J. Comput. Sci. Eng. Surv. vol. 2 2019, pp 1045 -1053.

[5] Łopatka J, Dołowski J., Bryś R., Grochowina B., “Efficiency Assessment of Coordinated Hierarchical Dynamic Spectrum Access for Mobile Ad Hoc Networks”, 38th IBIMA Conference, November 2021, Seville, Spain.

[6] Łopatka, J.; Gajewski, P.; Malon, K.; Krygier, J. Performance Monitoring of Cognitive Radio Mobile ad hoc Network with Dynamic Spectrum Access. In Proceedings of the AFRICON-2017, Cape Town, South Africa, 18–20 September 2017.

ASPEKTY REGULACYJNE ORAZ ZARZĄDZANIE CZĘSTOTLIWOŚCIAMI W PRZYPADKU MAŁYCH ORAZ BARDZO MAŁYCH SATELITÓW

REGULATORY ASPECTS AND FREQUENCY MANAGEMENT FOR SMALL AND VERY SMALL SATELLITES

Radosław Mądral; Radosław Tyniów

[1] Urząd Komunikacji Elektronicznej, Wrocław, Radoslaw.Madral@uke.gov.pl
[2] Urząd Komunikacji Elektronicznej, Wrocław, Radoslaw.Tyniow@uke.gov.pl

DOI: 10.15199/59.2022.4.97

Streszczenie: **W referacie zaprezentowano zagadnienia związane z krajowymi oraz międzynarodowymi regulacjami dotyczącymi zgłaszania oraz rejestracji częstotliwości radiowych wykorzystywanych przez urządzenia radiowe umieszczone na pokładzie małych satelitów wynoszonych na niskie orbity okołoziemskie.**
Abstract: **The paper presents issues related to national and international regulations concerning the notification and registration of radio frequencies used by radio devices located on board small satellites launched into low orbits around the Earth.**

Słowa kluczowe: **Gospodarka Zasobami Częstotliwości, Radiokomunikacja Satelitarna, Satelity Nie-geostacjonarne**

1. WSTĘP

Możliwość korzystania z małych satelitów ułatwiła wielu krajom oraz organizacjom pozarządowym, w tym uniwersytetom, instytutom edukacyjnym i badawczym oraz prywatnemu przemysłowi, przyłączenie się do eksploracji i wykorzystania przestrzeni kosmicznej, oraz obniżyła finansowy próg wejścia w opracowywanie technologii kosmicznych.

Satelity mogą być pogrupowane w różne kategorie np. według ich masy. W tym przypadku klasyfikacja może być następująca: mini to satelity o wadze mniejszej niż 1000 kg, mikro to satelity o wadze mniejszej niż 100 kg, nano to satelity o wadze mniejszej niż 10 kg, pico to satelity o wadze mniejszej niż 1 kg i najmniejsze femto to satelity o wadze mniejszej niż 0,1 kg. Jednak do tej pory nie uzgodniono powszechnie akceptowanego standardu w zakresie definicji małego lub bardzo małego satelity. Mały satelita niekoniecznie jest fizycznie mały, ponieważ może mieć rozkładaną lub zmieniającą kształt budowę, niekoniecznie jest lekki i nie musi być też mniej skomplikowany lub mniej wydajny w porównaniu do satelity, który nie jest uważany za mały. Do typowych cech małego satelity należą: a) stosunkowo krótki okres budowy; b) stosunkowo nieliczne zespoły budujące takiego satelitę; c) niewielkie wymagania dotyczące budowy i testowania; oraz d) niewielkie koszty rozwoju, budowy i eksploatacji.

W odniesieniu do wynoszenia na orbitę i eksploatacji satelitów muszą być spełnione pewne istotne wymagania wynikające z prawa międzynarodowego:

1. Notyfikacja i rejestracja, w Międzynarodowym Związku Telekomunikacyjnym (ITU), częstotliwości radiowych wykorzystywanych przez urządzenia radiowe umieszczone na pokładzie satelity;

2. Uwzględnienie w projektowaniu i eksploatacji satelity wymogów mających na celu ograniczanie ilości śmieci kosmicznych;

3. Rejestracja obiektu kosmicznego poprzez Sekretarza Generalnego Organizacji Narodów Zjednoczonych.

W niniejszym artykule opisano głównie pierwszy z powyższych punktów, czyli procedurę rejestracji w Głównym Międzynarodowym Rejestrze Częstotliwości (MIFR) ITU urządzenia radiowego umieszczonego na pokładzie satelity. Pierwszym etapem tej procedury jest złożenie w Urzędzie Komunikacji Elektronicznej (UKE) wniosku o wydanie pozwolenia na używanie urządzenia radiowego, które będzie umieszczone na pokładzie satelity.

2. Regulamin Radiokomunikacyjny ITU (RR)

W przypadku służby radiokomunikacji satelitarnej, podobnie jak w przypadku innych służb radiokomunikacyjnych, wykorzystanie przydziału częstotliwości może powodować szkodliwe zakłócenia, dlatego każda stacja satelitarna musi zostać zgłoszona do międzynarodowej koordynacji.

Regulamin Radiokomunikacyjny ITU jest wiążącym międzynarodowym traktatem regulującym wykorzystanie przez administracje krajów członkowskich zasobów widma i zasobów orbitalnych. Określa on prawa i obowiązki państw członkowskich w zakresie wykorzystania tych zasobów. Głównym celem i obowiązkiem jest rejestracja przydziału częstotliwości w Głównym Międzynarodowym Rejestrze Częstotliwości (MIFR) i związana z tym międzynarodowa koordynacja takiego przydziału.

W Regulaminie Radiokomunikacyjnym ITU nie ma przepisów regulacyjnych dotyczących wyłącznie małych satelitów. Rozróżniane są tylko satelity lub systemy geostacjonarne (GSO) i inne niż GSO.

Postanowienia Regulaminu dotyczące wykorzystania małych satelitów znajdują się w następujących częściach Regulaminu Radiokomunikacyjnego:

- Artykuł 1 – zawiera definicje terminologii używanej w innych częściach RR,
- Artykuł 5 – zawiera tablicę przeznaczeń częstotliwości informującą o wykorzystaniu poszczególnych zakresów częstotliwości przez różne służby radiokomunikacyjne,
- Artykuły 9 i 11 – zawierają procedury publikacji wstępnych informacji (API - Advance Publication Information) koordynacji (CR/C) oraz notyfikacji przydziału do Rejestru MIFR,
- Artykuły 21/22 – zawierają informacje o poziomach dopuszczalnych mocy (PFD w przypadku służb ziemskich, EIRP w przypadku ochrony służb satelitarnych, EPFD w przypadku ochrony satelitów GSO przed satelitami NGSO),
- Artykuł 25 – zawiera postanowienia dotyczące służby amatorskiej i amatorskiej satelitarnej,
- Artykuł 29A – zawiera postanowienia dotyczące służb radiowych związanych z obserwacją Ziemi,
- Załącznik 1 – zawiera postanowienia dotyczące klasyfikacji rodzajów emisji i niezbędnych szerokości pasm,
- Załącznik 4 – zawiera wykazy i tabele parametrów stacji satelitarnej wymaganych do jej poprawnego zgłoszenia.

Pozostałe kryteria i procedury techniczne zawarte są w odpowiednich zaleceniach i raportach ITU-R.

3. Podstawowe zasady korzystania z Tablicy Przeznaczeń Częstotliowści

Struktura przeznaczeń częstotliwości (art. 5 RR) i związane z nią zasady stanowią podstawę planowania oraz wdrożenia służb radiokomunikacyjnych. Tablica Przeznaczeń Częstotliwości obejmuje zakres od 8,3 kHz do 3 000 GHz i jest podzielona na mniejsze podzakresy, które mogą być przeznaczone do wykorzystania przez jedną lub kilka spośród ponad czterdziestu zdefiniowanych służb radiokomunikacyjnych (art. 1 RR). Służby radiowe są identyfikowane jako pierwszej lub drugiej ważności (te ostatnie nie powinny powodować szkodliwych zakłóceń oraz żądać ochrony przed zakłóceniami od tych pierwszych). Ponadto Tablica zawiera przypisy (uwagi), które doprecyzowują sposób wykorzystania poszczególnych podzakresów.

Poniżej podano typowe służby radiokomunikacyjne wykorzystywane przez małe satelity oraz w tabeli 1 przykładowe zakresy częstawości przeznaczone dla tych służb:

- Amateur-Satellite Service/ Służba Amatorska Satelitarna,
- Earth Exploration-Satellite Service (EESS)/ Służba Badań Ziemi Satelitarna,
- Meteorological-Satellite Service/ Służba Meteorologiczna Satelitarna,
- Space Research-Satellite Service/ Służba Badań Ziemi Satelitarna,
- Space Operation Service (SOS)/ Służba Operacji Kosmicznych

Tab. 1. Przykładowe zakresy częstotliwości dla służby EESS (Earth Exploration-Satellite Service)

Zakres częstotliwości	*Przeznaczenie*	*Kategoria*
401 – 403 MHz	EESS (Ziemia-kosmos)	Pierwszej ważności
2025 – 2110 MHz	EESS (Ziemia-kosmos), (kosmos-kosmos) SOS (kosmos-Ziemia), (kosmos-kosmos)	Pierwszej ważności
2200 – 2290 MHz	EESS (kosmos-Ziemia), (kosmos-kosmos) SOS (kosmos-Ziemia), kosmos-kosmos)	Pierwszej ważności
7190 – 7250 MHz	EESS (Ziemia-kosmos)	Pierwszej ważności
8025 – 8400	EESS (kosmos-Ziemia)	Pierwszej ważności
13.75 –14 GHz	EESS	Drugiej ważności

4. Złożenie wniosku o wydanie pozwolenia radiowego

Operator przymierzający się do wyniesienia małego satelity, na pokładzie którego będzie umieszczone urządzenie radiowe powinien uzyskać od organu regulującego zarządzanie częstotliwościami odpowiednie pozwolenie. W przypadku Polski jest to Prezes Urzędu Komunikacji Elektronicznej (UKE). Rozpoczęciem procesu jest złożenie wniosku o wydanie pozwolenia na używanie urządzenia radiowego umiejscowionego na pokładzie satelity, wraz z następującymi załącznikami:

- dokumentem potwierdzającym uiszczenie opłaty skarbowej,
- dokumentem potwierdzającym spełnianie wymagań zasadniczych przez urządzenie radiowe (najczęściej będzie to deklaracja zgodności).

Wraz z wnioskiem należy również dostarczyć plik w odpowiednim formacie, zawierający informacje o wykorzystywanych częstotliwościach radiowych oraz niektórych parametrach technicznych satelity. Do przygotowania odpowiedniego pliku należy skorzystać z programów opracowanych przez ITU:

- SpaceCap – oprogramowanie do tworzenia zgłoszeń koordynacyjnych i notyfikacyjnych,
- SpaceVal - oprogramowanie do walidacji przygotowanych zgłoszeń koordynacyjnych i notyfikacyjnych,
- GIMS - oprogramowanie umożliwiające dołączanie dodatkowych plików graficznych (np. charakterystyki anten).

Wszczęcie procedury zgłaszania przydziału częstotliwości do ITU, zgodnie z artykułem 9.1 Regulaminu Radiokomunikacyjnego, należy rozpocząć nie wcześniej niż siedem lat i nie później niż dwa lata przed planowanym

terminem wprowadzenia do użytku (BiU – Bring into Use) sieci lub systemu satelitarnego.

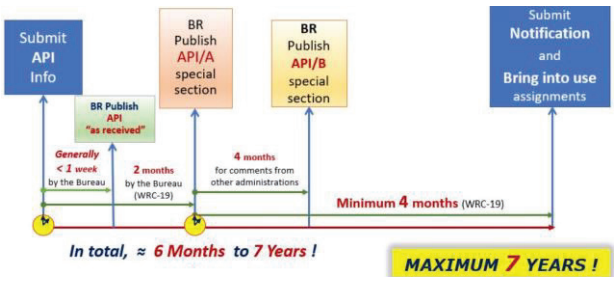
5. Zasady regulacyjne i procedury koordynacyjne mające zastosowanie do małych satelitów

Po otrzymaniu wniosku z odpowiednimi załącznikami oraz pliku z parametrami technicznymi UKE wszczyna procedurę, po zakończeniu której będzie można wydać pozwolenie radiowe. Pierwszym krokiem jest rozpoczęcie 'koordynacji' międzynarodowej wykorzystywanych przez satelitę częstotliwości. Uzgodnienie wykorzystania częstotliwości stanowi podstawowy składnik międzynarodowych regulacji dotyczących radiokomunikacji, ponieważ umożliwia wdrażanie nowych przydziałów częstotliwości, unikając szkodliwych zakłóceń w stosunku do innych istniejących i planowanych użytkowników.

Małe satelity mogą wykorzystywać zakresy częstotliwości podlegające koordynacjom jak i takie, które takiemu procesowi nie podlegają. O tym czy pasmo częstotliwości podlega koordynacji, stanowią przypisy (uwagi) w tablicy przeznaczeń częstotliwości, jednak w przeważającej części przypadków małe satelity wykorzystują pasma częstotliwości, które nie podlegają koordynacji, dlatego tylko taki przypadek zostanie opisany w dalszej części artykułu.

W przypadku częstotliwości niepodlegających koordynacji stosuje się postanowienia art. 9 podsekcja IA, które obejmują procedurę komentowania i rozwiązywania problemów związanych ze zgłaszanymi przez inne administracje zastrzeżeniami, określoną w punkcie 9.3 artykułu 9 RR.

Odpowiednie procesy zostały przedstawione na rysunku 1, zaczynając od wysłania pliku z parametrami technicznymi, na wpisaniu do Głównego Międzynarodowego Rejestru Częstotliwości (MIFR) kończąc.



Rys 1 Proces obsługi zgłoszeń dla sieci satelitarnych w zakresach częstotliwości nie podlegających koordynacji (źródło ITU)

5.1. Publikacja wstępnych informacji (API)

Pierwszym etapem procedury ITU jest dostarczenie pliku z parametrami technicznymi zgłaszanego satelity i umieszczonego na nim urządzenia radiowego (tzw. API), celem wszczęcia przez ITU międzynarodowego procesu komentowania, poprzez jego publikację w tzw. części API/A w Międzynarodowym Informacyjnym Cyrkularzu Częstotliwości (BRIFIC - International Frequency Information Circular). Celem procedury API określonej w sekcji I art. 9 RR jest informowanie wszystkich administracji o każdym planowanym systemie satelitarnym wykorzystującym satelitę NGSO wraz z jego ogólnym opisem. Ta obowiązkowa procedura zapewnia mechanizm formalny, dzięki któremu każda administracja może dokonać wstępnej oceny prawdopodobnego wpływu planowanej sieci satelitarnej na stacje planowanych lub istniejących systemów satelitarnych i ich stacji naziemnych w odpowiednich pasmach częstotliwości i w razie potrzeby mieć możliwość zgłoszenia komentarzy i uwag.

Zakres informacji wymaganych do zgłoszenia jest określony w Załączniku 4 Regulaminu Radiokomunikacyjnego, w tym:

- nazwa satelity, odpowiedzialna administracja i operator satelity, poziomy mocy, oznaczenie emisji, charakterystyki anten, pasma częstotliwości,
- parametry orbit,
- obszary obsługi,
- stacje naziemne.

Wszystkie zgłoszenia powinny być przygotowane w formacie zgodnym z oprogramowaniem ITU SpaceCap.

5.2. Procedury regulacyjne dotyczące komentowania i rozwiązywania problemów

Od momentu otrzymania zgłoszenia, ITU ma dwa miesiące na jego opublikowanie w części A BRIFIC – sekcja specjalna API/A. Następnie po opublikowaniu, administracje innych krajów mają cztery miesiące na zgłaszanie komentarzy/zastrzeżeń. Zastrzeżenia nie mają określonej formy i mogą być przedstawiane w sposób ogólny, natomiast niektóre administracje już na tym etapie wskazują konkretne przydziały, które według nich mogą być zakłócane. Po tym terminie ITU publikuje w części B BRIFIC – sekcja specjalna API/B listę administracji, które przesłały komentarze.

W następującym po tym okresie zarówno administracja zgłaszająca przydział jak i administracje zgłaszające komentarze powinny współpracować, wymieniać wszelkie dostępne dodatkowe istotne informacje, oraz dołożyć wszelkich starań, aby wspólnie szukać sposobów na rozwiązanie istniejących trudności.

Do wszystkich zgłoszonych zastrzeżeń należy się odnieść. Oficjalne odpowiedzi wysyła odpowiedzialna administracja, ale wkład merytoryczny do tych odpowiedzi powinien dostarczyć operator danego satelity i od sprawności w reagowaniu i przekazywaniu niezbędnych informacji w dużej mierze zależy czas procedowania wniosku.

Podczas wymiany korespondencji ze składającymi komentarze administracjami obowiązują poniższe zasady:

- zainteresowane strony powinny dołożyć wszelkich możliwych starań, aby rozwiązać trudności za pomocą wzajemnie akceptowalnych modyfikacji parametrów swoich sieci satelitarnych,
- w pierwszej kolejności strona odpowiedzialna za planowaną sieć satelitarną powinna dołożyć wszelkich

możliwych starań aby znaleźć rozwiązanie dla zgłaszanych zastrzeżeń, ale bez rozważania możliwości dostosowania się do sieci innych administracji,

- jeśli nie można znaleźć takich rozwiązań, można poprosić inne administracje o rozważenie innych możliwości uzyskania kompatybilności pomiędzy przydziałami,
- każda ze stron może zwrócić się o pomoc do Biura Radiokomunikacji ITU.

6. Procedury notyfikacji i rejestracji

Po osiągnięciu porozumienia z komentującymi administracjami, można przystąpić do ostatniej fazy procesu, czyli notyfikacji ostatecznych parametrów i rejestracji przydziału w Głównym Międzynarodowym Rejestrze Częstotliwości (MIFR).

MIFR stanowi jeden z filarów międzynarodowych ram regulacyjnych dotyczących łączności radiowej, ponieważ zawiera wszystkie wykorzystania częstotliwości zgłoszone do ITU. Każdy nowy użytkownik widma, przed wyborem konkretnych częstotliwości powinien wziąć pod uwagę wpisane do niego przydziały. Z tych powodów zgłoszenie przydziału i rejestracja w MIFR, stanowi ważne zobowiązanie dla administracji, zwłaszcza w odniesieniu do tych przydziałów częstotliwości, które mają międzynarodowe implikacje.

W MIFR powinny się znaleźć:

- przydziały częstotliwości służące do nadawania i odbioru na stacjach naziemnych i satelitarnych,
- przydziały częstotliwości mogące powodować szkodliwe zakłócenia,
- przydziały częstotliwości rejestrowane wyłącznie w celach informacyjnych.

Proces notyfikacji i rejestracji przydziału w MIFR jest opisany artykule 11 RR. Po otrzymaniu kompletnego zgłoszenia przydziału częstotliwości, ITU najpierw je publikuje w części I-S BRIFIC jako potwierdzenie otrzymania danych, następnie po szczegółowych badaniach w zależności od otrzymanego wyniku jest opublikowany w części II-S, jeżeli jest on korzystny, lub w części III-S, jeżeli wynik jest niekorzystny.

Po tym jak ITU opublikuje przydział częstotliwości w części II-S, UKE może wydać pozwolenie radiowe. A sam przydział częstotliwości będzie brany pod uwagę we wszelkich przyszłych działaniach prowadzonych na poziomie krajowym, regionalnym lub międzynarodowym.

7. Opłaty

7.1. Opłata skarbowa

Zgodnie z art. 6 ust 1, pkt 3 Ustawy o opłacie skarbowej z dnia 16 listopada 2006 r. (t.j. Dz. U. z 2020 r. poz. 1546 z późn. zm.), z chwilą złożenia wniosku o wydanie pozwolenia radiowego powstaje obowiązek zapłaty opłaty skarbowej określonej w załączniku do ustawy, w wysokości:

- 1939 zł – od pozwoleń radiowych wydawanych w związku z wykonywaniem działalności polegającej na świadczeniu usług telekomunikacyjnych, dostarczaniu sieci telekomunikacyjnych lub udogodnień towarzyszących (załącznik do ustawy – część III, poz. 32),
- 82 zł – od pozostałych pozwoleń radiowych (załącznik do ustawy – część III, poz. 44),

Organem właściwym w sprawach pobierania opłaty skarbowej za czynności wykonane przez Urząd Komunikacji Elektronicznej jest Urząd Miasta Stołecznego Warszawy.

7.2. Opłaty związane z procedowaniem przydziału w ITU

Niektóre czynności w ITU podlegają opłatom, które musi ponieść wnioskodawca. Czynności oraz wysokości opłat są określone w Decyzji 482. Dla sieci satelitarnych wykorzystujących zakresy częstotliwości niepodlegające koordynacji opłaty wynoszą:

- zgłoszenie API – 570 CHF
- notyfikacja – 7030 CHF

W przypadku nieuregulowania płatności w terminie sześciu miesięcy, licząc od daty zgłoszenia powyższych czynności do ITU, procedowanie przydziału satelitarnego zostaje anulowane, jednak opłaty za każdą z powyższych czynności są w dalszym ciągu wymagane.

7.3. Opłaty za prawo do wykorzystywania częstotliwości

Wnioskodawca, który uzyskał prawo do wykorzystywania częstotliwości określonych w pozwoleniu radiowym uiszcza roczne opłaty zgodnie z rozporządzeniem Rady Ministrów z dnia 7 grudnia 2017 r. w sprawie rocznych opłat za prawo do dysponowania częstotliwością (Dz.U. z 2017 r., poz. 2407). Obowiązek uiszczania opłat powstaje w momencie doręczenia decyzji (pozwolenia radiowego).

8. Podsumowanie

W ostatnich latach można zaobserwować gwałtowny wzrost liczby wynoszonych satelitów i związane z tym zwiększenie nakładu pracy wymaganej do zapewnienia ich bezkonfliktowego współistnienia i funkcjonowania. Oprócz problemów czysto technicznych istnieją również wyzwania regulacyjne, które dodatkowo komplikują proces lokowania satelity na orbicie okołoziemskiej. Zaprezentowane w referacie zagadnienia mogą być pomocne przy określaniu ram czasowych działań zmierzających do wyniesienia w przestrzeń kosmiczną satelity nie geostacjonarnego.

LITERATURA

[1] Regulamin Radiokomunikacyjny ITU, edycja 2020.
[2] Strona internetowa ITU - Small Satellites Support: https://www.itu.int/en/ITU-R/space/Pages/supportSmallSat.aspx, dostęp: 05.05.2022

SENSOR MONITORINGU WIDMA Z WYKORZYSTANIEM RADIA PROGRAMOWALNEGO

Maciej Mazuro[1]; Paweł Skokowski[2]

[1]Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, maciej.mazuro@student.wat.edu.pl
[2]Wojskowa Akademia Techniczna, Instytut Systemów Łączności, Warszawa, pawel.skokowski@student.wat.edu.pl

DOI: 10.15199/59.2022.4.98

Streszczenie: **Era cyfrowa i masowa digitalizacja przynoszą coraz większą liczbę urządzeń emitujących fale radiowe, a co za tym idzie następuje zagęszczenie spektrum elektromagnetycznego. Konieczne zatem jest zarządzanie widmem w taki sposób, aby użytkownicy nie zakłócali się wzajemnie i nie wywoływali interferencji. Narzędziem, które zwiększa świadomość elektromagnetyczną i jest pierwszym krokiem do efektywnego przydziału pasm częstotliwości jest sensor monitoringu widma oparty o np. metodę detekcji energii odbieranych sygnału.**

Abstract: **The digital age and mass digitization bring an increasing number of devices emitting radio waves, and thus the electromagnetic spectrum becomes denser. Therefore, it is necessary to manage the spectrum in such a way that users do not interfere with each other and do not cause interference. A tool that increases electromagnetic awareness and is the first step in the effective allocation of frequency bands is a spectrum monitoring sensor based on a method of detecting the energy of the received signal.**

Słowa kluczowe: **radio programowalne, detektor energii, widmo elektromagnetyczne, fale radiowe**

Keywords: **software defined radio, energy detector, electromagnetic spectrum, radio waves**

1. KONCEPCJA

Wraz ze znaczącym rozwojem technologicznym nastąpił wzrost zapotrzebowania na dostęp do spektrum elektromagnetycznego. Transmisje radiowe nie są już zarezerwowane do transmisji mowy lub przesłania krótkiej wiadomości tekstowej SMS (ang. Short Message Service). Obecnie mamy do czynienia z potężną liczbą urządzeń cyfrowych, które komunikują się za pomocą fal radiowych. Prawie wszystkie urządzenia mają dostęp do Internetu i pozwalają na długie godziny surfowania po sieci, oferta platform streamingowych umożliwia oglądanie ulubionego serialu w każdym zakątku świata. Nawet nasze domy stały się inteligentne i przesyłają nam na bieżąco obraz z domowej kamery, co pozwala na podglądanie naszego pupila, który został sam w mieszkaniu. Żyjemy w czasach, kiedy nasza lodówka może wysłać nam informację, co potrzebujemy kupić, aby przyrządzić nasze ulubione danie.

Ogrom informacji, które nas otaczają wymusza skuteczne zarządzanie widmem elektromagnetycznym, aby każdy użytkownik został obsłużony w wybranym przez niego czasie z jednoczesnym zapewnieniem najwyższej jakości świadczonych usług. Narzędziem niezwykle skutecznym i przydatnym w kontroli środowiska elektromagnetycznego jest sensor monitoringu widma [1][2]. Poniższy artykuł prezentuje koncepcje sensora monitoringu widma, zaprojektowanego w oparciu o radio programowalne SDR (ang. Software Defined Radio). System działa według techniki detektora energii z estymacją poziomu szumu. Odbiornik firmy Ettus Research pozwala na odbiór próbek rzeczywistych sygnałów radiowych, które przetwarzane są w środowisku programistycznym Matlab. Następnie w ramach działania opracowanego programu obliczana jest energia próbek sygnału, które zostają porównane z wartością progową. Pozwala to na określenie zajętości kanału z zadanym prawdopodobieństwem.

2. OPIS PROBLEMU

2.1. Podstawy detekcji

Podstawą do przeprowadzenia detekcji sygnału i określenia zajętości kanału jest oszacowanie wartości energii odbieranego sygnału. Jest to niezbędne do dalszych operacji, ponieważ na podstawie zmierzonego poziomu energii można oceniać, czy kanał jest zajęty. Otrzymana wartość jest zestawiana z wartością progową, a w przypadku przekroczenia wartości odniesienia można z zadanym prawdopodobieństwem stwierdzić zajętość kanału.

Teoria decyzji stanowi bazę do wykrywania transmisji radiowych. W nowoczesnych systemach cyfrowych przesyłane są ciągi znaków – 0 i 1, które następnie trafiają do odbiorników. Te interpretują otrzymane wartości logiczne i podejmują decyzję czy otrzymany bit ma wartość 0 lub 1. Źródło sygnału może być w dwóch stanach – stan logiczny niski, w którym transmisja nie jest prowadzona oraz logiczny stan wysoki odpowiadający transmisji danych. Stany te określane są na podstawie rachunku prawdopodobieństwa jako hipoteza zerowa H_0 i hipoteza alternatywna H_1.

Wyróżniane są cztery możliwe sytuacje:

- wybrano hipotezę H_0 kiedy prawdziwa jest hipoteza H_0;
- wybrano hipotezę H_0 kiedy prawdziwa jest hipoteza H_1;
- wybrano hipotezę H_1 kiedy prawdziwa jest hipoteza H_0;
- wybrano hipotezę H_1 kiedy prawdziwa jest hipoteza H_1;

Pierwszy i czwarty przypadek odpowiada prawidłowemu działaniu odbiornika, a drugi (błędne wykrycie) i trzeci

(fałszywy alarm) jest błędny. W technikach radarowych stosuje się testy statystyczne, które mają na celu minimalizację prawdopodobieństwa fałszywego alarmu P_{FA} i maksymalizację prawdopodobieństwa detekcji P_D [3]. Do często wykorzystywanych narzędzi statystycznych należą m.in. test Neymana-Pearsona [4] lub algorytm największego prawdopodobieństwa [5].

2.2. Techniki detekcji sygnałów radiowych

Podczas detekcji widma odbiornik działa w oparciu o dwie hipotezy [12]:

$$y(k) = \begin{cases} w(k) \rightarrow hipoteza\ H_0 \\ h \cdot s(k) + w(k) \rightarrow hipoteza\ H_1 \end{cases}, \quad (1)$$

gdzie *y(k)* jest odebranym sygnałem przez radio programowalne, *w(k)* jest użytecznym sygnałem, *s(k)* stanowi zakłócenia występujące w kanale, a współczynnik *h* opisuje amplitudę sygnału oraz przesunięcie fazy.

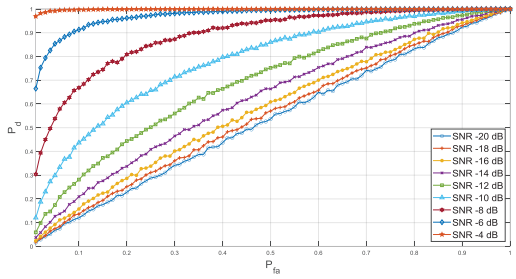
W ramach powyższego równania odebrany sygnał *y(k)* może w całości stanowić szum – przyjęcie hipotezy H_0 lub sygnał w odbiorniku jest sumą szumu i sygnału użytkownika *s(k)* wymnożonego przez współczynnik *h*, reprezentujący wartość amplitudy i przesunięcia fazy.

W technikach detekcji sygnałów radiowych wyróżniamy trzy bazowe metody oraz dwa rozwiązania określane mianem dodatkowych. Do najbardziej popularnych i najczęściej stosowanych technik należy detekcja oparta na energii odbieranego sygnału [6][7][8], detekcja prowadzona w oparciu o wzorce sygnałowe [9] oraz detektory wykrywające właściwości cyklostacjonarne [10]. Nowe podejście w zakresie analizy widmowej przyniosło rozwiązania takie jak metoda kowariancji [11] oraz wektorów własnych [12], która pomimo osiągania lepszych rezultatów cechuje się zdecydowanie większą złożonością.

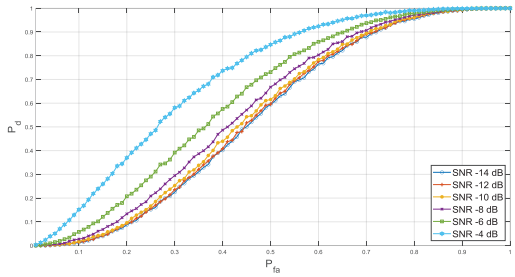
Wszystkie omówione podejścia dokonują wyboru spośród dwóch hipotez, ale na podstawie innych warunków. W technice opartej na energii sygnałów czynnikiem decydującym jest poziom energii odebranego sygnału, cyklostacjonarne podejście oparte jest na wykrywaniu okresowości w badanych przebiegach [13][14], a ostatnia z podstawowych metod stosuje znajdujące się w bazie wzorcowe sygnały do porównania z przetwarzanym impulsem.

3. PROPONOWANE ROZWIĄZANIE

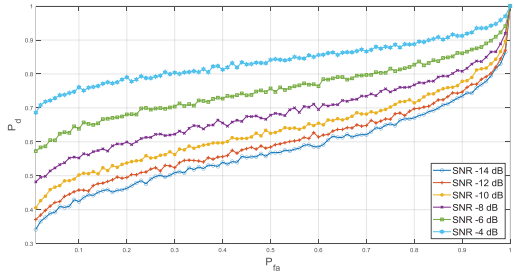
W ramach praktycznego rozwiązania zaimplentowano algorytmy trzech wybranych technik detekcji energii w środowisku programistycznym Matlab Simulink. Do badań została wybrana technika detekcji energii ze względu na niewymaganie wiedzy a priori na temat badanego sygnałów [15] oraz metody kowariancji i wektorów własnych ze względu na łatwość implementacji w środowisku tablicowym. Efekty uzyskanej pracy zostały zestawione w ramach charakterystyki działania odbiorników ROC (*ang. Receiver Operating Characteristic*), które pozwalają zestawić prawdopodobieństwo detekcji sygnału z prawdopodobieństwem fałszywego alarmu w funkcji stosunku poziomu sygnału użytecznego (informacja) do poziomu szumu tła SNR (ang. Signal-to-Noise Ratio).



Rys. 1 Charakterystyki ROC detektora energii



Rys. 2 Charakterystyki ROC detektora kowariancji



Rys. 3 Charakterystyki ROC detektora wektorów własnych

Na podstawie uzyskanych charakterystyk (Rys. 1), (Rys. 2), (Rys. 3) oraz założeń przyjętych w implementacji praktycznej – brak dostępu do bazy sygnałów wzorcowych oraz braku znajomości struktury odbieranych sygnałów, do implementacji sprzętowej został wybrany detektor energii.

4. IMPLEMENTACJA

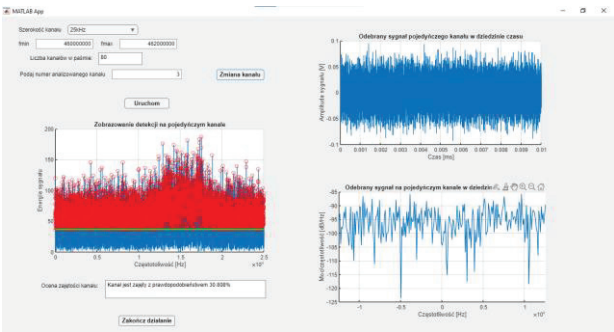
4.1. Rozwiązanie sprzętowe

Do praktycznej realizacji wykorzystane zostało radio programowalne USRP B200mini. Jest to jedno z najpopularniejszych urządzeń peryferyjnych firmy Ettus Research, które umożliwia odbiór sygnałów w paśmie od 70 MHz, aż do 6 GHz z definiowaną przez użytkownika szerokością pojedynczego kanału (do wartości 56 MHz). W ramach zaproponowanego rozwiązania realizowany jest odbiór sygnałów w zadanym paśmie, przetworzenie sygnału z dziedziny czasu do dziedziny częstotliwościowej przy wykorzystaniu szybkiej transformaty Fouriera (FFT). Moduł uzyskanych wartości jest równoznaczny z wartością energii poszczególnych próbek. Sama energia jest niewystarczająca do estymacji zajętości kanału częstotliwościowego. Należy dodatkowo obliczyć wartości progowe (progi detekcji), z którymi następnie zostaną po-

równane wartości energii próbek [16]. Jeśli energia sygnału jest większa niż wartość próbki odbiornik przyjmuje hipotezę alternatywną za prawdziwą – w odebranych danych znajdują się informację użyteczne, wyemitowane przez źródło sygnału. W przeciwnym przypadku hipoteza zerowa zostaje uznana za prawdziwą. Do ustalenia wartości odniesienia wykorzystane zostały energia szumu w paśmie obliczana na podstawie odchylenia standardowego. Stosunek energii sygnału w paśmie do energii zakłóceń pozwala na ustalenie wartości SNR, na podstawie której ustalona zostaje liczbowa wartość progowa. Aby analizować pojedyńczy kanał i określić jego zajętość użytkownik musi określić numer interesującego go kanału w interfejsie użytkownika.

4.2. Dedykowana aplikacja

Sprzętowy detektor energii odbieranych sygnałów posiada dedykowaną aplikacje, która pozwala na zdefiniowanie interesującego nas pasma do monitoringu, szerokości pojedynczego kanału i rozpoczęciu procesu odbioru sygnałów radiowych. Następnie użytkownik ma możliwość wyboru interesującego go kanału i zobrazowanie wyników analizy, które zawierają przebiegi czasowe, częstotliwościowe oraz wykres widmowy wraz z zaznaczonymi wynikami detekcji transmisji radiowych. Dodatkowo pozyskiwana jest informacja o zajętości badanego kanału oraz prawdopodobieństwo z jakim stwierdzono detekcję (Rys. 4).



Rys. 4 Detekcja energii w paśmie 460-462MHz

Graficzny interfejs użytkownika jest niezwykle przydatnym narzędziem, który pozwala w prosty i przejrzysty sposób zapoznać się z możliwościami technicznymi algorytmu działającego na platformie radia programowalnego. Aplikacja została opracowana na podstawie narzędzia App Designer, które pozwala na proste i intuicyjne tworzenie interfejsów użytkowników z gotowych bloków funkcyjnych oraz ich interakcję z istniejącymi już modułami programu.

5. WYNIKI TESTÓW

W ramach przeprowadzonych badań sprawdzeniu poddane zostały pasma częstotliwości, na których można zaobserwować dużą zajętość. Wybrane pasma są wykorzystywane przez różnych użytkowników, instytucje i reprezentują odmienne technologie radiokomunikacyjne. Pasmo 460-462MHz to pasmo trankingowe łączności Ministerstwa Spraw Wewnętrznych i Administracji (MSWiA) oraz Ministerstwa Obrony Narodowej (MON), 890-892MHz to fragment pasma GSM dla łącza w górę. Analizując charakterystyki widma z zaznaczonymi wynikami detekcji i na podstawie informacji o prawdopodobieństwie zajętości danego kanału można stwierdzić, że na danym paśmie transmisja radiowa jest prowadzona. W ramach aplikacji możliwy jest podgląd przebiegu odbieranego sygnału w dziedzinie czasu, częstotliwości oraz wykres widmowy, na którym zaznaczone są momenty detekcji sygnału na podstawie energii odebranego sygnału i zestawienia go z poziomem szumu, przy założeniu, że w odbieranym sygnale mamy do czynienia z addytywnym białym szumem gaussowskim. Jest to swoiste uproszczenie metodyki ze względu na losowy i zmienny charakter zakłóceń, które mogą być estymowane przy pomocą wariancji.

W zależności od zadeklarowanej szerokości analizowanego pasma różne wartości prawdopodobieństwa zajętości kanału będą wskazywały na transmisję radiową. Dla szerszych pasm kanałów np. 100kHz lub 200kHz jedna podnośna wykorzystywana w systemach LTE zajmuje około 10% lub 5% szerokości całego pasma. W takim przypadku, pomimo mniejszej liczby próbek przekraczających próg detekcji mamy do czynienia z przesyłaniem informacji. Parametr prawdopodobieństwa detekcji jest tylko wartością pomocniczą, która wskazuje z jak dużą pewnością możemy określić status zajętości pasma.

Tab. 1. Zestawienie wyników detekcji

Pasmo [MHz]	*Kanał*	*Zajęty*	*Prawodpodobieństwo*
100-110	48	tak	31,4%
460-462	33	tak	33,3%
460-462	76	tak	33,4%
790-820	15	tak	32,1%
890-915	11	tak	33,5%
890-915	38	tak	33,3%
890-915	65	tak	33,4%

W ramach przeprowadzonych symulacji zbadane zostały wybrane zakresy pasma częstotliwości m.in. pasmo 100-110 MHz, które jest zakresem wykorzystywanym przez radiofonię, zakres od 460-462MHz stanowiący pasmo trankingowe służb, 790-820 MHz, które odpowiada pasmu stacji bazowych oraz 890-915 MHz, w zakresie którym występuje transmisja w systemach GSM. W każdym z badanych pasm przeanalizowane było kilka kanałów i w większości z nich można zaobserwować prawdopodobieństwo detekcji (zajętość kanału) przekraczające 20 i 30%, co świadczy o wysokich poziomach energii odbieranego sygnału. W przypadku wolnych częstotliwości wartości są minimalne i potrafią nie przekroczyć setnych wartości procenta. Ciekawą zależnością są zbliżone wartości prawdopodobieństwa zajętości wybranych kanałów na poszczególnych częstotliwościach. Może to świadczyć o równomiernej zajętości analizowanego pasma.

6. WNIOSKI

Analizując otrzymane wyniki i porównując je z dostępną literaturą [1] można stwierdzić prawidłowe działanie opracowanego sensora widma. Opracowany program

prawidłowo przetwarza odebrane sygnały, a zaproponowany detektor porównuje wartości energii poszczególnych próbek z wartością progową. Skuteczność działania potwierdziły próby przeprowadzone w mniej eksploatowanych zakresach częstotliwości, gdzie tylko pojedyńcze próbki przekraczały wartości progowe. Wybrane pasma częstotliwości i prowadzone na nich transmisję są wykrywane z dużym prawdopodobieństwem, co pozwala na łatwe i szybkie określenie zajętości kanału oraz występowanie użytkowników pierwotnych. Jest to niezwykle istotny aspekt w kontekście dalszego rozwoju sieci telekomunikacyjnych i zwiększającej się liczby abonentów, którzy chcą mieć stały i nieprzerwany dostęp do łącza radiowego w wybranej przez siebie formie. Zaimplementowany detektor opiera się na estymacji poziomu szumu z wykorzystaniem właściwości statystycznych addytywnego szumu gaussowskiego, co zdecydowanie upraszcza algorytm działania, a równocześnie daje dużą skuteczność w sytuacji, kiedy nie jest znany charakter odbieranego sygnału. Zastosowanie detektora opierającego się na detekcji cech charakterystycznych, wzorców sygnałowych lub bazującego na kowariancji sygnałów mogłoby przynieść lepsze rezultaty, jednak duża złożoność obliczeniowa wymagana dla tych metod dyskwalifikuje je jako docelowe implementacje na komputerach klasy mikro typu Raspberry Pi.

Kolejne etapy rozwoju sensora monitoringu widma zakładają stworzenie bezzałogowego systemu powietrznego, w ramach którego radio programowalne SDR zostanie umieszczone wraz z mikrokomputerem Raspberry Pi na UAV, co znacząco zwiększy mobilność i zasięg pracy detektora oraz umożliwi prowadzenie operacji z dala od centrum dowodzenia, co jest szczególnie ważne w przypadku prowadzenia operacji specjalnych przez służby mundurowe. Docelowe możliwości systemu zakładają prowadzenie monitoringu widma elektromagnetycznego w czasie rzeczywistym, przekazanie do operatora w centrum dowodzenia narzędzi umożliwiających zobrazowanie wyników detekcji oraz zintegrowanie rozwiązania z systemami świadomości sytuacyjnej, będącymi na wyposażeniu służb.

LITERATURA

[1] Skokowski P., 2021. „Budowanie świadomości sytuacji elektromagnetycznej w sieciach doraźnych z węzłami kognitywnymi". Warszawa: Wojskowa Akademia Techniczna.

[2] Skokowski P., Malon K., Łopatka J., "Building the Electromagnetic Situation Awareness in MANET Cognitive Radio Networks for Urban Areas". Sensors (2022).

[3] Hillenbrand J., Weiss T., Jondral F., "Calculation of detection and false alarm probabilities in spectrum pooling systems", IEEE Commun. Lett., vol. 9, no. 4, ss. 349–351, 2005.

[4] Neyman J.; Pearson E. S., „On the Problem of the Most Efficient Tests of Statistical Hypotheses", Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 231, ss. 289-337, 1933.

[5] Jakubiak A., „Probabilistyczne metody detekcji sygnałów na tle zakłóceń", Oficyna Wydawnicza Politechniki Warszawskiej, 2013.

[6] Arslan H, Yucek T. 2009. „A survey of spectrum sensing algorithms for cognitive radio applications". *IEEE Communications Surveys & Tutorials*, 11 (1) : 116–130.

[7] Kandeepan, S., Giorgetti A., Chiani M., "Periodic Spectrum Sensing Performance and Requirements for Detecting Legacy Users with Temporal and Noise Statistics in Cognitive Radios," P-roceedings of IEEE Globecom/Workshop, Hawaii, 2009.

[8] Mariani A., Giorgetti A., Chiani A., "Effects of noise power estimation on energy detection for cognitive radio applications", IEEE Trans. Commun., vol. 59, no. 12, ss. 3410-3420, 2011.

[9] Sun H., Nallanathan A., Wang C., Chen Y., "Wideband spectrum sensing for cognitive radio networks: A survey", IEEE Wireless Communications, vol. 20, ss. 74–81, 2013.

[10] Kyouwoong, K., Akbar I. A., Bae K. K., Urn J-S., Spooner C. M, "Cyclostationary Approaches to Signal Detection and Classification in Cognitive Radio" IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN), Dublin, Ireland, ss. 212–215, 2007.

[11] Zeng, Y., and Y. C. Liang, "Spectrum-Sensing Algorithms for Cognitive Radio Based on Statistical Covariances," IEEE Transactions On Vehicular Technology, Vol. 58, ss. 1804–1815, No. 4, 2009.

[12] W. Su, J. Kosinski, "A survey of digital modulation recognition methods," in Proc. Int. Signal Process. Conf., Dallas, 2003.

[13] Skokowski P., Łopatka J., *Metody wykrywania i analizy sygnałów DSSS*, VII Konferencja Naukowo Techniczna Systemy Rozpoznania i Walki Elektronicznej KNTWE 2008, ISBN: 978-83-61486-04-6.

[14] Skokowski P., Łopatka, "Efficient Method of 3G Signal Detection", PIERS Online, Vol. 7, No. 1, ss. 51-54, 2011.

[15] Malon K., Skokowski P., Kelner J. M., Łopatka J., Koncepcja nadzorcy w symulatorze radia kognitywnego CORASMA, Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, ISSN: 1230-3496, 2014.

[16] Skokowski P., Malon K., Łopatka J., "Properties of centralized cooperative sensing in cognitive radio networks," Proc. SPIE 10418, XI Conference on Reconnaissance and Electronic Warfare Systems, 1041807 (20 April 2017); https://doi.org/10.1117/12.2269996

ANALIZA SYMULACYJNA RZECZYWISTEJ EKSPOZYCJI NA PROMIENIOWANIE ELEKTROMAGNETYCZNE OD SYSTEMÓW WIELOANTENOWCH 5G

SIMULATION ANALYSIS OF ACTUAL ELECTROMAGNETIC FIELD EXPOSURE FROM 5G MULTIANTENNA SYSTEMS

Marcin Rybakowski[1,2]; Kamil Bechta[3]

[1] Nokia Solutions and Networks Sp. z o.o., Wrocław, marcin.rybakowski@nokia.com
[2] Politechnika Wrocławska, Wydział Elektroniki, Fotoniki i Mikrosystemów, Wrocław
[3] Nokia Solutions and Networks Sp. z o.o., Wroclaw, kamil.bechta@nokia.com

DOI: 10.15199/59.2022.4.99

Streszczenie: **Systemy wieloantenowe są kluczową technologią 5-tej generacji standardu systemów komórkowych gwarantującą znaczący wzrost pojemności tych systemów. Zastosowanie złożonych algorytmów formowania i sterowania wiązkami antenowymi komplikuje jednak sposób oceny ekspozycji na pole elektromagnetyczne od takich systemów. Podejście dotychczas stosowane wymaga stosowania bardzo dużych stref ochronnych ponieważ zakłada nierealistyczne działanie systemów wieloantenowych. Zaproponowane w międzynarodowym standardzie IEC 62232 podejście statystyczne daje bardziej realistyczną ocenę, ale wymaga stosowania złożonego modelowania systemu i jest mocno zależne od scenariusza i parametrów systemu. W artykule tym przedstawiamy analizę symulacyjną wybranych systemów wieloantenowych pod kątem ekspozycji na pole elektromagnetyczne i badamy wpływ wybranych parametrów scenariusza na współczynnik redukcji mocy stosowany w nowej metodzie.**

Abstract: **Multi-antenna systems are the key technology of the 5th generation of the cellular systems that guarantees a significant increase in capacity of these systems. However, the use of complex beamforming algorithms complicates the methods of assessing electromagnetic field exposure from such systems. The approach used so far turns out to be very conservative. The statistical approach proposed in international standard IEC 62232 gives a more realistic assessment but requires the use of complex system modelling and is heavily dependent on the scenario and system parameters. In this article, we present a simulation analysis of selected multi-antenna systems in terms of exposure to electromagnetic field and examine the impact of selected scenario parameters on the power reduction factor used in the new method.**

Słowa kluczowe: **5G, ekspozycja na pole elektromagnetyczne, Massive MIMO, sterowanie wiązką antenową, systemy wieloantenowe**

Keywords: **5G, beamforming, EMF exposure, Massive MIMO, multiantenna systems**

1. WSTĘP

Systemy wieloantenowe typu *Massive MIMO (*ang. Multiple Input Multiple Output*)* są najważniejszą technologią wprowadzoną do standardu 5G w celu sprostania stale rosnącemu zapotrzebowaniu na przepustowość danych w systemach komórkowych. Wyposażenie stacji bazowych (BS) w duże układu antenowe pozwala bardzo znacząco zwiększyć wydajność widmową komórki radiowej, głównie dzięki dwóm uzupełniającym się technikom: kształtowaniu wiązki i multipleksowaniu przestrzennemu. Stosując kształtowanie wiązki, BS skupia energię transmisji w kierunku konkretnego terminala użytkownika (UE), co znacznie zwiększa moc sygnału odbieranego. W przypadku multipleksowania przestrzennego wiele strumieni jest wysyłanych przez BS do kilku aktywnych UE, które są rozdzielone przy użyciu algorytmu formowania wiązek antenowych. Ze względu na znaczące zyski obiecane przez technologie wieloantenowe stacje bazowe wyposażone w dziesiątki elementów antenowych zostają wdrażane obecnie w sieciach 5G. Techniki te mają być rozwijane w następnych generacjach systemów komórkowych nawet w kierunku ekstremalnie dużych systemów wieloantenowych gdzie układy antenowe będą budowane z setek czy nawet tysięcy elementów antenowych.

Systemy wieloantenowe oraz techniki kształtowania i sterowania wiązkami antenowymi wprowadzają wyzwania do oceny ekspozycji na pole elektromagnetyczne (PEM) w otoczeniu stacji bazowej. Tradycyjne metody stosowane w poprzednich standardach systemów komórkowych stosowane przez operatorów w celu estymacji stref ochronnych gdzie natężenie PEM przekracza normy obowiązujące w danym kraju bazowały na uproszczonym podejściu [1], [2]. W modelu tym zakłada się maksymalną wartość mocy promieniowanej (EIRP*)* w każdym kierunku. W przypadku systemów wieloantenowych, gdzie duże układy antenowe dysponują dużo większym zyskiem, ale jednocześnie bardzo wąską wiązką promieniowania anteny, obecna metoda powoduje estymacje bardzo dużych stref ochronnych. Ocena ekspozycji zakłada średnią wartość natężenia PEM mierzoną przez wiele minut (6 lub 30 minut) [1], a okres taki jest bardzo duży w odniesieniu do czasów przełączania czy kształtowania wiązek antenowych oraz adaptacji zasobów radiowych dla poszczególnych UE, które kształtują się na poziomie milisekund lub krótszym. Powoduje to, że średnia rzeczywista ekspozycja na PEM w danym kierunku jest dużo mniejsza niż przy obecnie stosowanej metodzie. Dla systemów wieloantenowych zostało rozwinięte w ostatnim czasie tak zwane podejście rzeczywiste do oceny ekspozycji na PEM (ang. actual EMF exposure approach*)*, które w ocenie tej zakłada dynamikę zmian PEM spowodowaną bardzo szybką adaptacją kształtowania wiązek antenowych [2], [6]. Podejście to umożliwia bardziej reali-

styczną ocenę ekspozycji na PEM od stacji bazowych wyposażonych w systemy wieloantenowe. Międzynarodowe organy standaryzacyjne, takie jak Międzynarodowa Komisja Elektromagnetyczna (IEC) i Międzynarodowa Unia Telekomunikacyjna (ITU) wprowadziły tę metodę i opisują jak należy ją stosować dla stacji bazowych z systemami wieloantenowymi. Ostatnia wersja nowej edycji międzynarodowego standardu dotyczącego oceny ekspozycji na PEM tj. IEC 62232 Ed3 CDV, opisuje tę nową metodę w szczegółach [2].

W artykule tym przedstawimy koncepcję tej metody i skupimy się na modelowaniu statystycznym wybranego systemu komórkowego ze stacjami bazowymi wieloantenowymi. Zostaną pokazane wyniki symulacji oparte na realistycznym modelowaniu systemu typu *Massive MIMO* pracującego na częstotliwości 3.5 GHz ze sterowanymi wiązkami antenowymi, realistycznym modelowaniu propagacji fal radiowych oraz obsługi terminali. Przedstawiona zostanie analiza ekspozycji na PEM od stacji bazowych używających różnej wielkości układów antenowych, różnej liczby jednoczesnych wiązek promieniowania oraz obsługujących rożną liczbę terminali.

2. RZECZYWISTA EKSPOZYCJA NA PEM OD SYSTEMÓW WIELOANTENOWYCH

Znając moc nadawczą P_{tx} w watach, charakterystykę zysku anteny nadawczej G_{tx} i dozwoloną gęstość mocy S_{max} w W/m^2 możemy obliczyć odległość bezpieczną od anteny D_{min} w metrach używając wzoru na tłumienie fali elektromagnetycznej w wolnej przestrzeni

$$D_{min}(\theta,\phi) = \sqrt{\frac{P_{tx}G_{tx}(\theta,\phi)}{4\pi S_{max}}}, \qquad (1)$$

gdzie Θ jest kątem elewacji, ϕ jest kątem azymutu i $P_{tx}G_{tx}$ = EIRP jest zastępczą mocą promieniowaną izotropowo wyrażoną w watach. Regulacje zwykle wymagają użycia największej możliwej mocy nadawczej i zysku antenowego danej stacji bazowej do szacowania ekspozycji na pole elektromagnetyczne żeby zagwarantować duże marginesy bezpieczeństwa dla wyników obliczeń natężenia PEM i wynikające z tego strefy bezpieczne. Limity natężenia PEM specyfikowane są jako średnia wartość w czasie (najczęściej 6 lub 30 minut [1]) gdzie deterministyczny współczynnik F_{TDC} wynikający z techniki dupleksu TDD (ang. Time Division Duplex), będący frakcją czasu przeznaczonego do transmisji w porównaniu do czasu przeznaczonego do odbioru, może być zastosowany do obliczenia średniej mocy transmitowanej w (1) przy pomocy wzoru

$$P_{tx} = P_{txmax}F_{TDC}, \qquad (2)$$

gdzie P_{txmax} jest maksymalną konfigurowaną mocą promieniowaną i F_{TDC} jest współczynnikiem dupleksu. W przypadku dupleksu FDD (*Frequency Division Duplex)* współczynnik ten wynosi 1, w przypadku dupleksu TDD jego wartość zależy od konfiguracji ramki i wynosi np. 0,75 co oznacza, że stacja bazowa w danym okresie transmituje przez 75% czasu i odbiera przez 25% czasu [2]. Ocena ekspozycji na PEM przy pomocy wzorów (1) i (2) jest stosunkowo prosta w przypadku użycia klasycznych anten ze statyczną charakterystyką promieniowania.

Stacje bazowe stosujące systemy wieloantenowe typu *Massive MIMO* mogą bardzo dynamicznie generować wiązki antenowe o zmiennej charakterystyce promieniowania, która adaptuje się do aktualnej charakterystyki kanału radiowego każdego z obsługiwanych użytkowników. Średnia wartość natężenia PEM zależy więc nie tylko od deterministycznych współczynników takich jak współczynnik dupleksu F_{TDC}, ale również od współczynników stochastycznych określonych na podstawie analiz statystycznych, zazwyczaj z dystrybuanty (*CDF – Cummulative Distribution Function)* średniej mocy promieniowanej w danym kierunku. W wyniku tego bardziej realistyczna odległość bezpieczna jest określona poprzez użycie rzeczywistej efektywnej mocy transmitowanej P_{tx_actual} obliczanej ze wzoru poniżej:

$$P_{tx\_actual} = P_{txmax}F_{TDC}F_{PR}, \qquad (3)$$

gdzie F_{PR} jest współczynnikiem redukcji mocy [2]. F_{PR} jest odczytywany z 95 lub 99 centyla dystrybuanty średniej wartości EIRP. Współczynnik ten jest kluczowy do prawidłowej estymacji realnej ekspozycji na PEM systemów wieloantenowych. Model statystyczny użyty do jego obliczenia powinien uwzględniać następujące najważniejsze parametry:

- Współczynniki układu antenowego i wynikające z tego charakterystyki układu antenowego
- Ilość aktywnych strumieni MIMO (ilość wiązek antenowych)
- Dystrybucję przestrzenną użytkowników terminali
- Liczbę aktywnych użytkowników oraz czas trwania transmisji do poszczególnego użytkownika

Warto zaznaczyć, że nie są to wszystkie zmienne, które są ważne przy tworzeniu modelu statystycznego.

Różne modele statystyczne powstały do określenia wartości współczynnika F_{PR} [4], [5], [6]. Symulacje systemowe przedstawione w tym artykule stosują jeden z najdokładniejszych modeli statystycznych opisany w następnej sekcji. Model ten został zastosowany w publikacji [3] dla analizy rzeczywistej ekspozycji od systemu antenowego 8x8 z algorytmem kształtowania wiązki typu *eigen-bemforming*. W niniejszej publikacji model ten został zastosowany do analizy również większych układów antenowych wraz z innym algorytmem sterowania wiązką typu *Grid of Beams (*GoB).

3. MODEL STATYSTYCZNY ORAZ ZAŁÓŻENIA SYMULACYJNE

W modelu symulatora systemowego wykorzystano statystyczny model przestrzenny 3D propagacji fal radiowych zgodny z 3GPP 38.901 UMa (Urban Macro) [7]. Symulacje wykonano dla sieci komórkowej zbudowanej z 7 komórek, gdzie każda z nich posiada 3 sektory, w których znajdują się stacje bazowe wieloantenowe typu *Massive MIMO*. W scenariuszu stacje bazowe są umieszczone na wysokości 25 m, a odległość między nimi wynosi 500 m. System pracuje na częstotliwości 3.5GHz i używa pasma

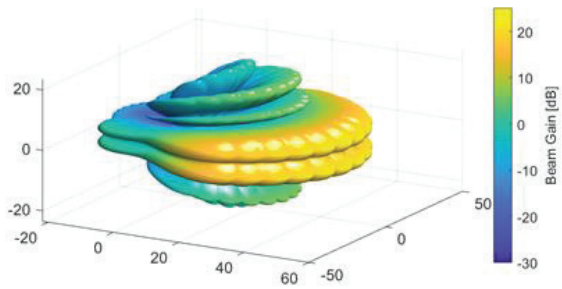
o szerokości 20MHz. Symulacje wykonano dla 3 różnych systemów wieloantenowych o parametrach przedstawionych w Tabeli 1. Wszystkie anteny są dwupolaryzacyjne (+/- 45 stopni), a zysk pojedynczego elementu antenowego wynosi 5,2dBi. Odległość pomiędzy elementami antenowymi wynosi połowę długości fali. Układ antenowy jest pogrupowany w mniejsze podukłady antenowe tak jak na Rysunku 1 i każdy z nich jest podłączony do osobnego układu nadawczo-odbiorczego TRX.

Tab. 1. Parametry systemów wieloantenowych

Parametr	32TRX_8x8	64TRX_12x8	64TRX_12x16
Liczba układów nadawczo-odbiorczych TRX	32	64	64
Liczba elementów antenowych w układzie (VxH, V- w pionie, H -w poziomie)	8x8	12x8	12x16
Maksymalny zysk anteny [dBi]	23,3	25	28
Moc nadawcza całkowita [W]	128	200	200
Liczba wiązek antenowych na polaryzację	24	GoB_1: 24 GoB_2: 74	50
Liczba wiązek w pionie i poziomie	2x12	GoB_1: 2x12 GoB_2: 6x12	2x25

Rys. 1. Konfiguracje układów antenowych

Przykładowa charakterystyka promieniowania 3D pokazująca wszystkie możliwe wiązki promieniowania jest przedstawiona na Rysunku 2 dla systemu wieloantenowego 64TRX_12x8.



Rys. 2. Zbiorcza charakterystyka promieniowania układu 64TRX_12x8

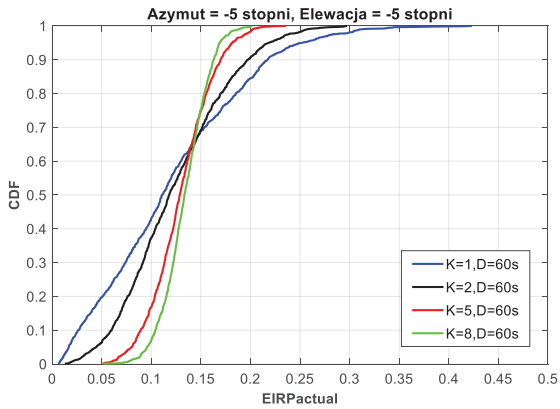
Na Rysunku 2 widzimy dwa rzędy wiązek antenowych, po 12 wiązek w każdym z nich. Charakterystyki te są identyczne dla obu polaryzacji. Rozkład wiązek w pionie oraz poziomie dla innych anten jest przedstawiony w Tabeli 1. Terminale użytkowników używają pojedynczej anteny dookólnej i są losowo rozmieszczone w komórce, gdzie 20% terminali jest na zewnątrz budynków, a 80% terminali jest wewnątrz budynków, których wysokość jest zmienna od 4 do 8 pięter (jeden scenariusz zakłada wyższe budynki od 15 do 20 pięter, co jest opisane w sekcji 4). Liczba jednocześnie obsługiwanych terminali K zmienia się od 1 do 8, a czas trwania pojedynczego połącznia D zmienia się od 10 do 360 sekund. System wieloantenowy może obsługiwać jednocześnie wiele użytkowników dzięki multipleksacji przestrzennej poprzez użycie osobnych wiązek antenowych o najmniejszej korelacji wzajemnej (interferencji między-wiązkowej). Każdy terminal wykorzystuje pełną moc i pasmo w czasie transmisji (*full buffer*). Terminale są obsługiwane w ciągu 6 minut co oznacza, że jeżeli parametr D jest krótszy od 6 minut, nowe lokalizacje terminali są losowo wybierane i transmisja odbywa się ponownie przez okres D. Podejście to umożliwia realistyczne zamodelowanie systemu, w którym stacja bazowa obsługuje wielu użytkowników rozmieszczonych w obrębach komórki. Wymaga to od systemu wieloantenowego ciągłego skanowania i zmiany wiązek antenowych co zostało zamodelowane w stosowanym symulatorze systemowym. W celu otrzymania dokładnej reprezentacji statystycznej symulacje zostały powtórzone od 50 do 100 razy.

W sekcji 4 przestawimy wyniki symulacji dla zaprezentowanych różnych systemów antenowych i zbadamy wpływ liczby obsługiwanych terminali K i czasu trwania transmisji D na współczynnik redukcji mocy F_{PR}.

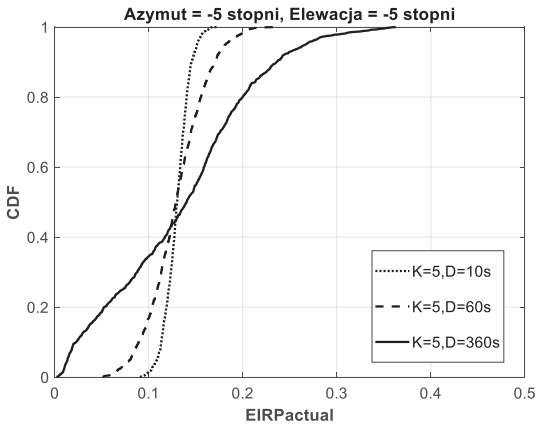
4. ANALIZA WYNIKÓW SYMULACJI

Przykładowe wyniki symulacji przedstawione na Rysunkach 3 i 4 dla stacji bazowej 64TRX_12x8 ilustrują dystrybuanty unormowanej mocy promieniowanej, która jest stosunkiem średniej mocy promieniowanej uśrednionej w okresie 6 minut do maksymalnej mocy promieniowanej danej stacji bazowej. Wyniki przedstawione są dla kierunku największego promieniowania, który charakteryzuje się największą ekspozycją na PEM. W przypadku zastosowanych scenariuszy jest to kierunek o kącie azymutu -5° i kącie elewacji -5°. Kąt azymutu wynika z tego, że w tym kierunku znajduje się wiązka antenowa o największym zysku ponieważ jest to kierunek najbliższy kierunkowi symetrii anteny (ang. Broadside), gdzie układ antenowy może posiadać największy zysk. W naszym przypadku jest on przesunięty o -5° ponieważ tam znajduje się najbliższa wiązka ze zbioru wiązek antenowych. Natomiast wartość kąta elewacji wynika z tego, że antena zawieszona jest powyżej budynków, gdzie znajdują się terminale użytkowników. Z tego powodu wiązka z drugiego rzędu wiązek antenowych o nachyleniu -5° jest najczęściej wybierana. Na Rysunku 3 możemy zaobserwować wpływ liczby jednocześnie obsługiwanych użytkowników przez wiele wiązek

antenowych na unormowaną moc promieniowaną stacji bazowej 64TRX_12x8 . Widać, że średnia moc promieniowana maleje wraz ze wzrostem liczby teminali. Natomiast na Rysunku 4 przedstawiony jest wpływ czasu transmisji do jednego terminala na unormowaną moc promieniowaną, która maleje wraz ze skróceniem czasu transmisji. Krótszy czas transmisji oznacza, że wiązki antenowe są częściej przełączane w ciągu czasu obserwacji 6 minut. Zauważyć można, że dla wszystkich wartości K i D unormowana moc promieniowana jest duża mniejsza od 1, co oznacza, że podejście dotychczasowe nie jest dobrą metodą do analizy ekspozycji na PEM od systemów wieloantenowych. Oba wykresy pokazują, że im większa zmienność przełączania wiązek antenowych, tzn. K rośnie oraz D maleje, tym ekspozycja na pole elektromagnetyczne od systemów wieloantenowych jest mniejsza.



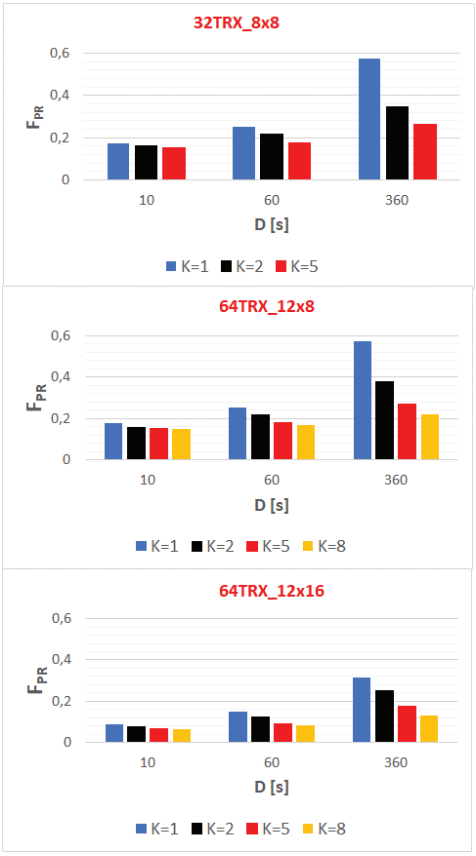
Rys. 3. Dystrybuanta unormowanej mocy promieniowane stacji bazowej 64TRX_12x8 dla K=1, 2, 5, 8 i D=60s



Rys. 4. Dystrybuanta unormowanej mocy promieniowane stacji bazowej 64TRX_12x8 dla K=5 i D=10s, 60s, 360s

Zgodnie w wytycznymi IEC 62232 [2] współczynnik redukcji mocy F_{PR} odczytujemy z 95 centyla dystrybuanty unormowanej mocy promieniowanej. Wartości te dla wszystkich typów systemów wieloantenowych zastosowanych w symulacjach możemy porównać na Rysunku 5. Zauważyć możemy, że im większa antena, tym współczynnik F_{PR} maleje. Współczynnik ten kształtuje się różnie w zależności od scenariusza i konfiguracji i dla bardziej realnych scenariuszy ruchu telekomunikacyjnego, czyli K ≥ 2 i D ≤60s i dla systemów wieloantenowych 32TRX_8x8 oraz 64TRX_12x8 utrzymuje się na zbliżonym poziomie 0,15 – 0,22. W przypadku systemów 64TRX_12x16 współczynnik ten przyjmuje wartości w zakresie 0,07 – 0,13. Oznacza to, że znacząco mniejsza średnia moc promieniowana jest emitowana w porównaniu do maksymalnej teoretycznej stosowanej w obecnej metodzie. Dzięki temu strefa bezpieczna może być zmniejszona o około 40-50% dla anten 32TRX_8x8 i 64TRX_12x8 (dla K ≥ 2 i D ≤60s) . Wyniki są podobne mimo, że układy antenowe się różnią, ale powodem jest większa moc nadawcza zastosowana w 64TRX_12x8. Dla anteny 64TRX_12x16 strefy bezpieczne mogą być zmniejszone aż o 25-35%. Dla przykładu, odległość bezpieczna dla stacji bazowej 64TRX_12x16 z parametrami z Tabeli 1 wynosi 27,7 metrów, a w podejściu realistycznym z oszacowanym współczynnikiem redukcji mocy wartość ta spada aż do 7,1-9,7 metrów w zależności od parametrów D i K.

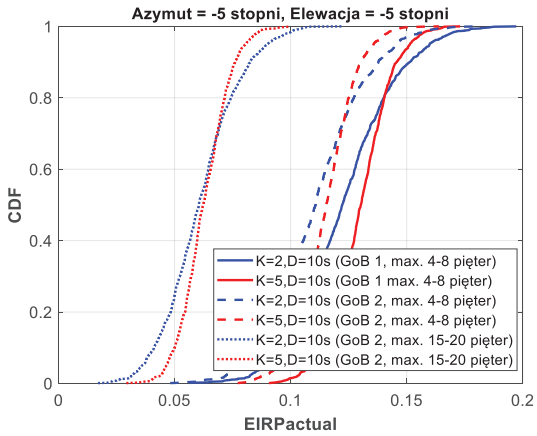
Można zaobserwować, że dla anteny 32TRX_8x8 i 64TRX_12x8 wartości współczynnika F_{PR} są bardzo zbliżone do siebie. Wynika to ze scenariusza gdzie mamy tylko dwa rzędy wiązek w elewacji i dość niskie budynki, w których znajdują się terminale. Wobec tego największą dynamikę przełączania wiązek mamy w płaszczyźnie poziomej, a ponieważ obie anteny mają tę samą liczbę elementów antenowych w tej płaszczyźnie, to wyniki są zbliżone.



Rys. 5. Wartość współczynnika redukcji mocy dla różnych rodzajów systemów wieloantenowych

Scenariusz został zmodyfikowany poprzez znaczne zwiększenie maksymalnej liczby pięter do wartości od 15 do 20 oraz zastosowania 6 zamiast 2 rzędów wiązek w

zbiorze wiązek antenowych (GoB_1 i GoB_2 w Tabeli 1). Wyniki są przedstawione na Rysunku 6, gdzie można zauważyć, że mimo zwiększenia liczby wiązek w elewacji kiedy mamy niższe budynki, nie wpływają one bardzo znacząco na zmianę unormowanej mocy promieniowanej, ponieważ prawdopodobieństwo wyboru wiązek z dodatkowych rzędów jest mała. Dopiero kiedy użytkownicy są rozlokowani na wyższych piętrach wiązki te będą wybierane dużo częściej, a większa zmienność przełączania wiązek powoduje spadek ekspozycji na PEM. Oznacza to, że w całościowej analizie bardzo ważny jest scenariusz dla którego występuje ocena ekspozycji oraz liczba wiązek antenowych w dwóch płaszczyznach.

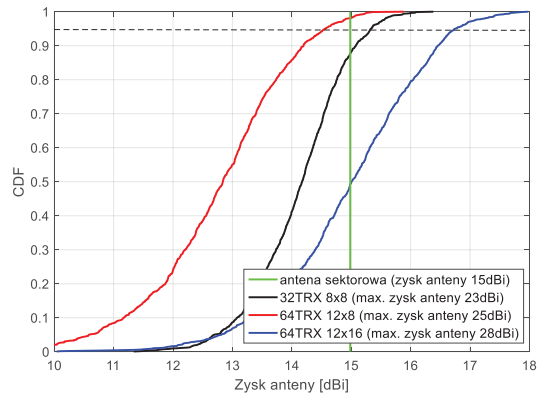


Rys. 6. Dystrybuanta unormowanej mocy promieniowane stacji bazowej 64TRX_12x8 dla scenariuszy z budynkami o różnej wysokości

Interesujące jest przedstawienie wyników symulacji w postaci dystrybuanty średniego zysku antenowego jak na Rysunku 7 dla jednego z przypadków przy K=2 i D=10s. Dla porównania został tam dodany wykres zysku typowej anteny sektorowej o wartości 15dBi. Jest to stała wartość ponieważ tego typu antena sektorowa nie stosuje dynamicznego kształtowania wiązki. Porównując zysk anteny sektorowej ze średnim zyskiem systemów wieloantenowych ich wartość może być niewiele większa lub nawet mniejsza. Skupiając się na 95 centylu tego średniego zysku antenowego widać, że wartość ta jest dużo mniejsza niż maksymalny zysk tych układów antenowych np. 95 centyl średniego zysku anteny 12x16 jest równy tylko 16,8dBi przy maksymalnym zysku tej anteny wynoszącym aż 28dBi.

5. PODSUMOWANIE

W publikacji tej zaprezentowano analizę symulacyjną rzeczywistej ekspozycji na PEM od systemów wieloantenowych. Podejście statystyczne bazujące na realistycznym modelu systemu wieloantenowego pod względem propagacji fali radiowej, rozmieszczenia stacji bazowych, rozkładu terminali, rodzaju algorytmu kształtowania i formowania wiązki oraz ruchu telekomunikacyjnego pokazuje, że rzeczywista ekspozycja na PEM od tego typu systemów jest duża mniejsza i zależy znacząco od parametrów systemu i scenariusza. Wyznaczony współczynnik redukcji mocy pozwala na znaczące zmniejszenie stref ochronnych wokół anteny. Ponieważ mogą zdarzyć sie przypadki przekroczenia ustalonych sterf ochronnych ze względu na dużą dynamikę systemu stacje bazowe muszą być wyposażone w mechanizm pomiaru i kontroli mocy promieniowanej, tak jak to jest rekomendowane w standardzie IEC 62232 [2]. Jednakże, z wyników symulacji możemy wywnioskować, że kontrola EIRP będzie się odbywać bardzo rzadko i tylko na zasadzie prewencyjnej.



Rys. 7. Dystrybuanta średniego zysku antenowego różnych systemów wieloantenowych w kierunku największej ekspozycji na PEM (dla K=2, D=10s)

LITERATURA

[1] International Commission on Non-Ionizing Radiation Protection, "Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz)," Health Phys., vol. 118, no. 5, pp. 483–524, Marzec 2020.

[2] "Determination of RF field strength, power density and SAR in the vicinity of radiocommunication base stations for the purpose of evaluating human exposure", International Electrotechnical Commission (IEC) IEC 62232 Ed3, Lipiec 2021.

[3] P. Baracca, A. Weber, T. Wild, and C. Grangeat, "A statistical approach for RF exposure compliance boundary assessment in massive MIMO systems", WSA 2018; 22nd International ITG Workshop on Smart Antennas, Marzec 2018.

[4] B. Thors, A. Furuskar, D. Colombi, and C. Tornevik, "Time-averaged realistic maximum power levels for the assessment of radio frequency exposure for 5G radio base stations using massive MIMO," IEEE Access, vol. 5, pp. 19 711–19 719, Wrzesień 2017.

[5] D. Pinchera, M. Migliore, and F. Schettino, "Compliance boundaries of 5G massive MIMO radio base stations: A statistical approach," IEEE Access, vol. 8, pp. 182 787–182 800, Październik 2020.

[6] Bo Xu at al. "A Summary of Actual Maximum Approach Studies on EMF Compliance of 5G Radio Base Stations", EuCAP 2022, Marzec 2022.

[7] 3GPP Radio Access Network Working Group, "Study on Channel Model for Frequencies from 0.5 to 100 GHz (Release 15)", 3GPP TR 38.901 V15.1.0.

ZASTOSOWANIE METOD KALIBRACYJNYCH W PEŁNOFALOWEJ SYMULACJI ELEKTROMAGNETYCZNEJ OBWODÓW PLANARNYCH NA CZĘSTOTLIWOŚCIACH SUBTERAHERCOWYCH

Mateusz Wrzesiński; Paweł Bajurko

Politechnika Warszawska, Instytut Radioelektroniki i Technik Multimedialnych, Warszawa,
{mateusz.wrzesinski.stud, pawel.bajurko}@pw.edu.pl

DOI: 10.15199/59.2022.4.100

Streszczenie: **We wrotach obwodu planarnego w pełnofalowych symulacjach elektromagnetycznych występuje nieciągłość powodująca niedopasowanie wrót do połączonej z nimi linii transmisyjnej. Powoduje to błędy w wynikach symulacji. Problem ten można rozwiązać stosując metody kalibracyjne, podobnie jak w mikrofalowej technice pomiarowej. W referacie przedstawiono przykłady ilustrujące to zagadnienie. Omówiono także kwestię wielomodowego sprzężenia między wrotami symulowanego obwodu, które ogranicza skuteczność kalibracji modelu z elektrycznie grubym podłożem. Zagadnienie to jest szczególnie ważne w przypadku obwodów pracujących na częstotliwościach subterahercowych.**

Abstract: **A discontinuity exists in the ports of the planar circuit in full-wave electromagnetic simulations, causing a mismatch between the port and the connected transmission line. This causes uncertainty in the simulation results. This problem can be solved by applying calibration methods, analogously to microwave measurement technology. The paper presents examples to illustrate this issue. A multi-mode coupling between the ports of the simulated circuit, which limits the calibration performance for a model with an electrically thick substrate, is also discussed. This is a matter of concern for circuits operating at subterahertz frequencies.**

Słowa kluczowe: **Kalibracja, obwody planarne, linia mikropaskowa, symulacja pełnofalowa.**

Keywords: **Calibration, planar circuits, microstrip line, full-wave simulation.**

1. WSTĘP

Wraz z rozwojem telekomunikacji na potrzeby łączności bezprzewodowej przeznaczane są coraz wyższe pasma radiowe. Pociąga to za sobą konieczność projektowania i symulacji obwodów elektronicznych pracujących na wysokich częstotliwościach, w zakresie mikrofal i fal milimetrowych. Wykorzystywane w tym celu narzędzia analizy symulacyjnej bazują na metodach, które można podzielić na dwie główne grupy: metody obwodowe i metody pełnofalowe [11].

Metody obwodowe zakładają określony rodzaj fali w liniach transmisyjnych i innych podzespołach w obwodzie (zazwyczaj rodzaj podstawowy) i bazują na analitycznych rozwiązaniach dla tego rodzaju [14]. Dzięki temu zapotrzebowanie tych metod na zasoby obliczeniowe jest stosunkowo niewielkie, a czas analizy krótki. Często wynik analizy jest znany wręcz natychmiastowo, przez co korzystanie z metod obwodowych jest niezwykle użyteczne. Wadą metod obwodowych jest ograniczony zakres stosowalności zaimplementowanych rozwiązań analitycznych oraz pominięcie części zjawisk, które mogą występować w obwodzie, takich jak wzbudzanie wyższych rodzajów fal czy sprzężenia między odległymi fragmentami obwodu. Problemy te narastają w wyższych pasmach częstotliwości. Szczególnie w zakresie częstotliwości subterahercowych często korzystne jest zastosowanie metod pozbawionych wspomnianych ograniczeń.

Metody pełnofalowe nie ograniczają rozwiązania do z góry określonej klasy zjawisk, na przykład tylko wybranych rodzajów fal. W rozwiązaniu dopuszczone są wszystkie rodzaje, które możliwe są przy warunkach brzegowych określonych strukturą i parametrami analizowanego obwodu. Dzięki temu symulacja uwzględnia ogół zjawisk opisywanych równaniami Maxwella, w tym efekty pasożytnicze lub „drugoplanowe" z punktu widzenia funkcji obwodu, które mogą jednak istotnie wpływać na jego działanie. Zaletą metod pełnofalowych jest także brak ograniczeń na typ symulowanego obwodu. Na przykład w obwodzie mikropaskowym bez przeszkód może zostać umieszczony obiekt o strukturze nieplanarnej. Główną wadą metod pełnofalowych jest duża złożoność obliczeniowa analizy, duże wymagania na zasoby pamięci komputerowej i długi czas obliczeń [2].

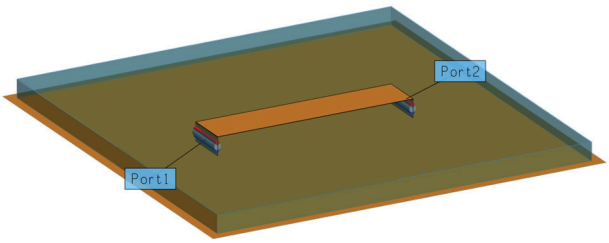
Należy odnotować, że w ostatnich latach pojawiły się hybrydowe metody symulacji, pozwalające wykorzystywać różne kombinacje metod obwodowych i pełnofalowych w analizie złożonych obwodów. Na przykład w modelu symulowanym metodą pełnofalową mogą być umieszczane elementy lub sekcje symulowane metodami obwodowymi [3, 5] albo – przeciwnie – w modelu symulowanym metodami obwodowymi mogą być uwzględniane niektóre efekty pasożytnicze jak sprzężenia sąsiadujących fragmentów obwodu [4]. Takie podejście pozwala osiągnąć kompromis między złożonością obliczeniową a dokładnością rozwiązania.

W pełnofalowej analizie symulacyjnej obwodów pracujących na wysokich częstotliwościach istotnym problemem jest nieciągłość występująca we wrotach badanego obwodu. Powoduje ona niedopasowanie między źródłem dołączanym do wrót a fragmentem obwodu bezpośrednio za wrotami, na skutek czego następują odbicia

fal zniekształcające wynik obliczeń. Ten sam problem występuje po stronie obciążenia. Wyodrębnienie parametrów obwodu wymaga wykorzystania metod kalibracyjnych [6, 8, 9, 10, 12, 13], podobnie jak to ma miejsce w mikrofalowej technice pomiarowej, gdzie kalibracja jest niezbędna do usunięcia wpływu kabli i przyrządów pomiarowych z danych pomiarowych. Niniejsza praca przedstawia ilustrację tego problemu na przykładzie prostego obwodu zawierającego odcinek linii mikropaskowej oraz osiąganą poprawę po zastosowaniu kalibracji na wrotach zasilających symulowany obwód.

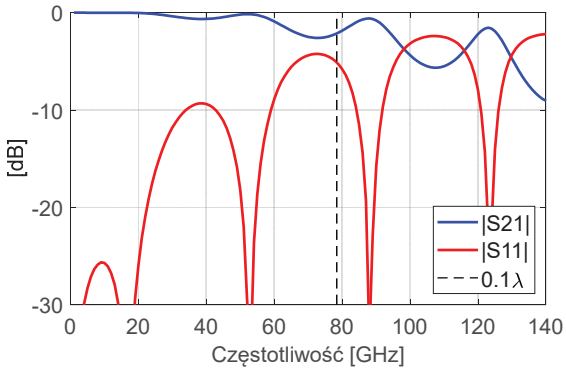
2. SYMULACJA PEŁNOFALOWA ODCINKA LINII MIKROPASKOWEJ

Na rys. 1 przedstawiono model odcinka 50-omowej linii mikropaskowej stworzony w symulatorze elektromagnetycznym Altair FEKO [1]. Omawiany przykład jest celowo skrajnie prosty, aby pokazać rozpatrywany problem w sposób niezaburzony innymi efektami. Szerokość linii wynosi 0,462 mm, długość odcinka 2,2 mm. Linia wykonana jest z idealnego przewodnika na bezstratnym podłożu o grubości 0,203 mm i względnej przenikalności elektrycznej 3,55. Na częstotliwości 78,4 GHz grubość podłoża odpowiada 1/10 długości fali w ośrodku o takiej przenikalności. Zwykle grubość 1/10 długości fali uważa się za granicę stosowalności podłoża w obwodach mikropaskowych, poniżej której udział niepożądanych rodzajów fal jest akceptowalnie mały. Symulację przeprowadzono w zakresie do częstotliwości 140 GHz, celowo przekraczając granicę 1/10 długości fali, aby móc zaobserwować efekty związane ze wzbudzaniem w strukturze rodzajów fal innych niż rodzaj podstawowy i ich wpływ na kalibrację. W modelu użyto wrót zasilających typu „edge port" o impedancji 50 Ω. Wrota wzbudzają falę wprowadzając różnicę potencjałów w miejscu ich umieszczenia, ale nie jest na nich wykonywana wewnętrzna kalibracja usuwająca związaną z nimi nieciągłość.



Rys. 1. Model odcinka linii mikropaskowej w symulatorze FEKO z wrotami typu „edge port"

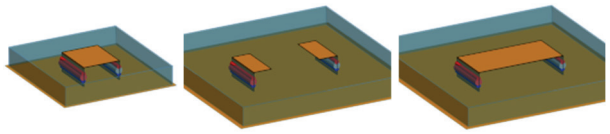
Symulacja obwodu została przeprowadzona metodą momentów. Obliczone parametry rozproszenia obwodu przedstawione są na rys. 2. Dzięki prostemu modelowi (odcinek bezstratnej linii transmisyjnej) można łatwo ocenić odchyłki wyznaczonych parametrów od idealnej charakterystyki (|S21| = 0 dB, brak odbić). Pod tym względem wynik symulacji można uznać za zadowalający tylko w zakresie do częstotliwości ok. 24 GHz. Na wyższych częstotliwościach rosną odbicia i straty, mimo że sama linia pozostaje bezstratna i dopasowana falowo do wrót zasilających. Efektów tych nie można wytłumaczyć niewielką dyspersją, charakterystyczną dla linii mikropaskowej [14]. Są one wynikiem problemu nakreślonego we wstępie. Zafalowania widoczne na charakterystyce powstają na skutek interferencji fal odbijających się od nieciągłości występujących we wrotach zasilających. Powyżej 60 GHz, a więc jeszcze przed granicą 0,1 długości fali, wpływ nieciągłości jest na tyle duży, że bez korekcji dane z symulacji stają się praktycznie nieużyteczne.



Rys. 2. Parametry rozproszenia linii z rys. 1, wynik symulacji bez korekty

3. KALIBRACJA

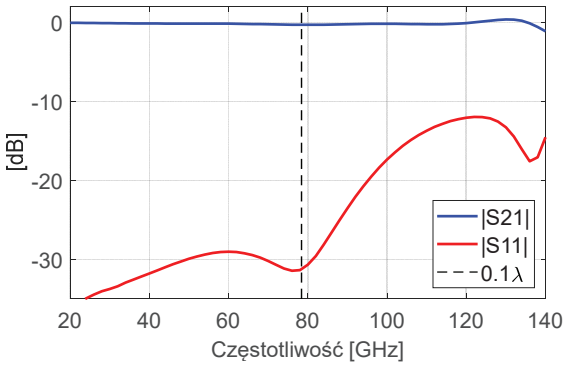
Jak wspomniano we wstępie, w celu odtworzenia prawidłowych wyników symulacji można posłużyć się metodami kalibracyjnymi. Do korekcji wyników z rys. 2 posłużyła metoda kalibracji TRL (Through-Reflect-Line) [7], oparta o trzy elementy wzorcowe: bezpośrednie połączenie wejścia i wyjścia (*through*), rozwarty obwód (*reflect*) i odcinek linii (*line*). Kalibracja TRL działa poprawnie dla długości elektrycznych odcinka linii z zakresu od 20° do 160°. Modele użytych wzorców kalibracyjnych przedstawione są na rys. 3. Wzorzec *through* ma długość równą dwukrotności grubości podłoża. Jako częstotliwość środkową dla wzorca „line" wybrano 80 GHz, co pozwala zastosować kalibrację TRL w zakresie częstotliwości 20–140 GHz.



Rys. 3. Wzorce kalibracyjne „through", „reflect" i „line"

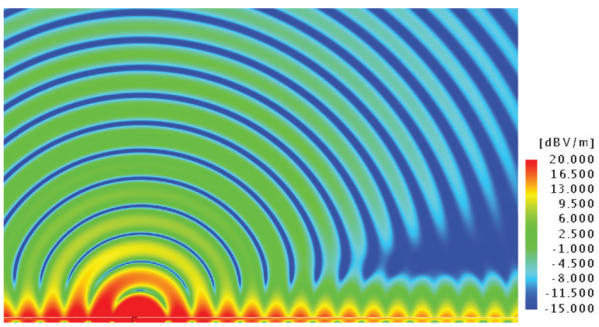
Wynik symulacji po kalibracji przedstawiono na rys. 4. Odchyłka współczynnika transmisji od poziomu 0 dB do częstotliwości 120 GHz mieści się w granicach 0,3 dB. Współczynnik odbicia do częstotliwości 95 GHz utrzymuje się na poziomie poniżej −20 dB. Poziom błędów spadł radykalnie w porównaniu do wyniku bez korekty (rys. 2), co świadczy o skutecznym usunięciu nieciągłości na wrotach zasilających w wyniku wykonanej kalibracji.

Na częstotliwościach powyżej granicy 0,1 długości fali na rys. 4 można zaobserwować, że skuteczność kalibracji zaczyna wyraźnie się pogarszać. Wzrasta poziom współczynnika odbicia, a w okolicy 140 GHz narastają też błędy współczynnika transmisji. Dzieje się tak na sku-



Rys. 4. Parametry rozproszenia linii z rys. 1 po korekcji metodą TRL

tek wielomodowej transmisji między wrotami zasilającymi. Kalibracja TRL zakłada, że transmisja między wrotami odbywa się tylko na jednym rodzaju fali. Symulacja pełnofalowa uwzględnia jednak wszystkie zjawiska opisane równaniami Maxwella, w tym występowanie innych rodzajów fal oprócz rodzaju podstawowego. Ich udział wzrasta z częstotliwością i staje się zauważalny po przekroczeniu granicy 0,1 długości fali. Aby zilustrować obecność wielu rodzajów w symulowanych strukturach na rys. 5 przedstawiono chwilowy rozkład indukcji pola elektrycznego na częstotliwości 140 GHz w otoczeniu pojedynczego wzorca *reflect*, widocznego jako niewielki czarny element w dolnej części rysunku. Na rozkładzie wyraźnie widoczne są dwie fale, których źródłem jest wzorzec: fala emitowana w wolną przestrzeń w postaci koncentrycznych półokręgów oraz fala związana z podłożem widoczna w dolnej części rysunku. Można zauważyć, że fale te różnią się długością (a więc także stałą propagacji). Jak wspomniano fale te występują równolegle do rodzaju podstawowego w linii mikropaskowej, mają swój udział w transmisji między wrotami i uniemożliwiają skuteczną kalibrację na wysokich częstotliwościach metodami nieuwzględniającymi ich obecności.



Rys. 5. Rozkład indukcji pola elektrycznego na częstotliwości 140 GHz w otoczeniu wzorca „reflect”

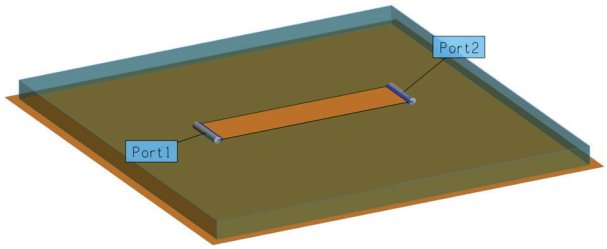
Uzyskane wyniki można ocenić pod względem zakresu częstotliwości, w którym spełniają określone wymagania. W tab. 1 zamieszczono zestawienie wymagań, odpowiadających im zakresów częstotliwości oraz grubości elektrycznej podłoża na maksymalnej częstotliwości z zakresu. Jak można zaobserwować, zastosowanie kalibracji wyraźnie rozszerza zakres częstotliwości, w którym rezultat symulacji spełnia wymagania.

Tab. 1. Zakresy częstotliwości i grubości podłoża, dla których wynik symulacji spełnia wymagania – wrota „edge port”

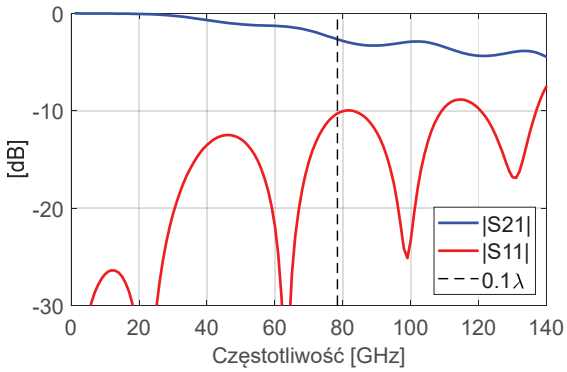
Wymaganie 1	$\Delta\lvert S21\rvert \leq 0{,}1$ dB, $\lvert S11\rvert \leq -20$ dB
Bez kalibracji	22 GHz (0,028λ)
Z kalibracją	33 GHz (0,042λ)
Wymaganie 2	$\Delta\lvert S21\rvert \leq 0{,}3$ dB, $\lvert S11\rvert \leq -15$ dB
Bez kalibracji	26 GHz (0,033 λ)
Z kalibracją	105 GHz (0,1342λ)

4. ZALEŻNOŚĆ SKUTECZNOŚCI KALIBRACJI OD SPOSOBU POBUDZENIA

Symulatory elektromagnetyczne zwykle udostępniają szereg sposobów na wzbudzenie rozkładu pola w analizowanej strukturze. Zależnie od wybranego sposobu nieciągłość oraz udział niepożądanych rodzajów generowanych we wrotach zasilających może się zmieniać. Na rys. 6 przedstawiono badaną linię mikropaskową zasilaną przez wrota typu „microstrip port” dostępne w symulatorze FEKO, zaś na rys. 7 jej parametry rozproszenia obliczone bez korekty. Różnice między charakterystykami uzyskanymi dla różnych typów wrót (rys. 2 i rys. 7) są konsekwencją odmiennego charakteru nieciągłości występującej na wrotach.



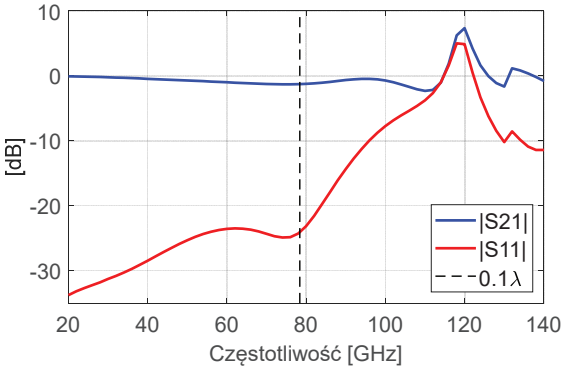
Rys. 6. Model odcinka linii mikropaskowej w symulatorze FEKO z wrotami typu „microstrip port”



Rys. 7. Parametry rozproszenia linii z rys. 6, wynik symulacji bez korekty

Po przygotowaniu zestawu kalibracyjnego z wrotami typu „microstrip port” i wykonaniu kalibracji, wynik symulacji przedstawia się jak na rys. 8. Odchyłka współczynnika transmisji w zakresie do granicy 1/10 długości fali sięga 1,3 dB, co jest wynikiem wyraźnie gorszym niż przy zastosowaniu wrót typu „edge port” (0,3 dB, rys. 4). Również powyżej granicy 1/10 długości fali błędy narastają z częstotliwością znacznie szybciej niż w przypadku

wrót „edge port”. Najprawdopodobniej świadczy to o silniejszym sprzężeniu wrót „microstrip port” z rodzajami innymi od rodzaju podstawowego w linii mikropaskowej niż ma to miejsce w przypadku wrót typu „edge port”.



Rys. 8. Parametry rozproszenia linii z rys. 6 po korekcji metodą TRL

W przypadku wrót „microstrip port” kalibracja praktycznie nie wpływa na zakresy częstotliwości, w których wynik symulacji spełnia wymagania wymienione w tab. 2. Dla tego typu wrót konwencjonalne metody kalibracji (nierozróżniające rodzajów fal) nie dają oczekiwanej poprawy dokładności wyników.

Tab. 2. Zakresy częstotliwości i grubości podłoża, dla których wynik symulacji spełnia wymagania – wrota „microstrip port”

Wymaganie 1	$\Delta\lvert S21\rvert \leq 0{,}1$ dB, $\lvert S11\rvert \leq -20$ dB
Bez kalibracji	24 GHz (0,030λ)
Z kalibracją	24 GHz (0,030λ)
Wymaganie 2	$\Delta\lvert S21\rvert \leq 0{,}3$ dB, $\lvert S11\rvert \leq -15$ dB
Bez kalibracji	34 GHz (0,043λ)
Z kalibracją	32 GHz (0,041λ)

5. PODSUMOWANIE

Występowanie nieciągłości we wrotach zasilających obwód mikropaskowy symulowany metodą pełnofalową powoduje błędy symulacji rosnące z częstotliwością i z grubością podłoża. Ma to szczególne znaczenie w przypadku analizy obwodów pracujących na częstotliwościach subterahercowych, gdzie podłoża mogą być stosunkowo grube elektrycznie. W zaprezentowanych przykładach zaobserwowano, że wynik symulacji bez dodatkowej korekty można uznać za prawidłowy (obciążony nieistotnym błędem), jeśli grubość podłoża nie przekracza ok. 0,03 długości fali. Zastosowanie kalibracji pozwala skorygować nieciągłość we wrotach i rozszerzyć zakres stosowalności symulacji na obwody o grubszych podłożach, lecz skuteczność kalibracji jest uzależniona od typu wrót zasilających obwód. Jeśli procedura kalibracji jest wykonywana samodzielnie przez użytkownika, kwestia skuteczności kalibracji wymaga indywidulanego zbadania. Niektóre symulatory elektromagnetyczne mają wbudowane procedury kalibracji wrót zasilających. Jeśli procedury automatycznej kalibracji wrót są zaimplementowane w kodzie programu, zazwyczaj jest to wyraźnie odnotowane w dokumentacji symulatora. Należy jednak mieć na uwadze, że w każdym pełnofalowym symulatorze wielomodowa propagacja między wrotami obwodu na zbyt grubym podłożu będzie negatywnie wpływać na dokładność wyniku.

LITERATURA

[1] Altair Feko, *Altair Engineering Inc.*, [online]. Dostęp: https://www.altair.com/feko/.

[2] Balanis Constantine A., „Advanced Engineering Electromagnetics”, 2nd Edition, John Wiley & Sons.

[3] Barchanski A., "Linking circuit simulation with full-wave solver for board-level EMC design," in *IEEE Electromagnetic Compatibility Magazine*, vol. 4, no. 3, pp. 52-58, 3rd Quarter 2015.

[4] Cadence AWR, [online]. Dostęp: https://www.awr.com/.

[5] Cetiner B., Coccioli R., Housmand B., Itoh T., "Combination of Circuit and Full Wave Analysis for Pre-Matched Multifinger FET," *2000 30th European Microwave Conference*, 2000, pp. 1-4.

[6] Duan J. and Zhu L., "Numerical Short-Open-Load (SOL) Calibration Technique for Accurate Extraction of Electrically-Small Planar/Non-Planar Microstrip-Line Circuits," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 70, no. 4, pp. 2067-2076, April 2022.

[7] Engen G. F., Hoer C. A., "Thru-Reflect-Line: An Improved Technique for Calibrating the Dual Six-Port Automatic Network Analyzer," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 27, no. 12, pp. 987-993, Dec. 1979.

[8] Farina M. and Rozzi T., "A short-open deembedding technique for method of moments based electromagnetic analyses", *IEEE Trans. Microw. Theory Tech.*, vol. 49, no. 4, pp. 624-628, Apr. 2001.

[9] Li L., Wu K. and Zhu L., "Numerical TRL calibration technique for parameter extraction of planar integrated discontinuities in a deterministic MoM algorithm", *IEEE Microw. Wireless Compon. Lett.*, vol. 12, no. 12, pp. 485-487, Dec. 2002.

[10] Massey J. W., Hajiaboli A. and Okhmatovski V. I., "On the Reciprocity Relation in General Multiport Microwave Circuits and Errata to Vector-Short-Open-Calibration Deembedding," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 2, pp. 1250-1254, Feb. 2021.

[11] Pozar D. M., "Microwave engineering education: From field theory to circuit theory," *2012 IEEE/MTT-S International Microwave Symposium Digest*, 2012, pp. 1-3.

[12] Rautio J. C. and Okhmatovski V. I., "Unification of double-delay and SOC electromagnetic deembedding," in *IEEE Transactions on Microwave Theory and Techn.*, vol. 53, no. 9, pp. 2892-2898, Sept. 2005.

[13] Rautio J. C., "A de-embedding algorithm for electromagnetics", *Int. J. Microwave Millimeter-Wave Computer-Aided Eng.*, vol. 1, no. 3, pp. 282-287, July 1991.

[14] Rosłoniec Stanisław, 1999, „Liniowe obwody mikrofalowe”, WKiŁ.

POPRAWA WYDAJNOŚCI PRACY SIECI GESTYCH STANDARDU IEEE 802.11AX

Marek Natkaniec [1]; Mateusz Kras [2]

[1] AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, natkanie@agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, matikrass@wp.pl

DOI: 10.15199/59.2022.4.101

Streszczenie**: Praca koncentruje się na problemie optymalizacji działania sieci gęstych standardu IEEE 802.11ax. Cel pracy został zrealizowany przy użyciu symulatora NS-3. Analizowano różne topologie pracy sieci z uwzględnieniem kolorowania BSS (Basic Service Set). Przedstawiono proces poprawy wydajności pracy sieci dla zmian wartości mocy nadajnika, progu CCA Threshold (Clear Channel Assessment Threshold) oraz zysku energetycznego anteny. Określono wpływ czułości odbiornika na wydajność pracy sieci.**
Abstract**: The paper focuses on the optimization of IEEE 802.11ax dense networks. The results were obtained with the use of the NS-3 simulator. Various network topologies were analyzed with BSS coloring mechanism enabled or disabled. The process of improving the network performance for changes in the transmitter power value, CCA Threshold, and antenna gain was presented. The influence of receiver sensitivity on network performance was determined.**

Słowa kluczowe**: sieci gęste, CCA Threshold, kolorowanie sieci, IEEE 802.11ax.**

Keywords**: dense networks, CCA Threshold, network coloring, IEEE 802.11ax.**

1. WSTĘP

Rodzina standardów IEEE 802.11 koncentruje się na specyfikacji warstwy fizycznej i warstwy MAC (*Medium Access Control*) dla bezprzewodowych sieci lokalnych. Od czasu publikacji pierwszej wersji standardu IEEE 802.11 upłynęło już ponad 20 lat. W tym czasie komitet standaryzacyjny opracował wiele rozszerzeń poprawiających ogólne funkcjonowanie sieci. Standard IEEE 802.11ax nazywany również siecią Wi-Fi szóstej generacji to odpowiedź na coraz większe oczekiwania dotyczące szybkości i małych opóźnień transmisji, rosnącą liczbę urządzeń oraz coraz bardziej zagęszczone sieci bezprzewodowe w takich środowiskach jak stadiony, centra handlowe czy budynki mieszkalne [1]. Nowe rozszerzenie oferuje szybkość transmisji do 9,6 Gbit/s, co daje ok. 37% wzrost w porównaniu do standardu IEEE 802.11ac, redukcję opóźnienia transmisji oraz większą niezawodność pracy sieci bezprzewodowej. Duży nacisk został w standardzie położony na bardziej efektywne wykorzystanie pasma oraz niwelowanie negatywnego wpływu interferencji. W stosunku do rozszerzenia IEEE 802.11ac rozszerzono warstwę fizyczną m.in. wprowadzając technikę OFDMA (*Orthogonal Frequency Division Multiple Access*) i modulację 1024-QAM oraz warstwę MAC dodając m.in. drugi wektor alokacji sieci NAV oraz ramkę trigger wspierającą działanie trybu UL MU-MIMO. W standardzie IEEE 802.11ax na poziomie warstwy MAC zaproponowano również nowe mechanizmy związane z zarzadzaniem energią oraz mechanizmy pozwalające na pracę w tzw. gęstych środowiskach [2].

Praca składa się z pięciu rozdziałów. Rozdział 2 zawiera przegląd literatury. W rozdziale 3 omówiono koncepcję sieci gęstych. Z kolei w rozdziale 4 zawarto wyniki badań symulacyjnych przeprowadzonych dla dwóch różnych scenariuszy pracy sieci WLAN. Ostatni rozdział podsumowuje niniejszą pracę.

2. PRZEGLĄD LITERATURY

W ostatnich latach ukazało się wiele publikacji na temat analizy wydajności pracy najnowszego rozszerzenia IEEE 802.11ax [3], [4]. Ze względu na ograniczenia w implementacji nowych mechanizmów standardu w sprzęcie sieciowym, znaczna cześć publikacji została oparta na przeprowadzonych symulacjach bądź też modelach analitycznych.

Mechanizmy zwiększające wykorzystanie przestrzenne kanału radiowego zostały przetestowane w [5] oraz [6]. W pracy [5] przeprowadzono testy kolorowania BSS dla wielu scenariuszy uwzględniając różny poziom zagęszczenia sieci. Zbadano 3 obszary, w których umieszczono 10 punktów dostępowych i 10 stacji. W centralnie umieszczonym BSS zastosowano kolorowanie, aby porównać osiągane przepustowości z resztą sieci. W najbardziej zagęszczonym środowisku uzyskano 450-procentowy wzrost przepustowości dzięki zastosowaniu kolorowania sieci. Uzyskano również znacznie mniejsze opóźnienia i zajętości kanału. W artykule [6] zbadano wpływ zmiany parametru CCA Threshold na uzyskiwane szybkości transmisji. Omówione testy przeprowadzono dla dwóch sieci, gdzie w każdej z nich pracowało maksymalnie 15 urządzeń. Zmiana wartości z -82 dBm na -50 dBm spowodowała wzrost przepustowości o około 79% przy 15 urządzaniach w sieci, natomiast dla 2 urządzeń w sieci wzrost ten wynosił 44%. Zwiększając CCA Threshold zaobserwowano spadek średnich opóźnień oraz jego zmienności o około 60-70%. Z analizy literatury wynika, że mechanizmy które zostały wdrożone w rozszerzeniu IEEE 802.11ax znacząco poprawiają wydajność sieci. Niestety prace te nie wskazują jaka będzie wydajność dla scenariuszy z bardzo dużą liczbą urządzeń WLAN zlokalizowanych w budynkach.

3. SIECI GĘSTE

Rozwój technologii, wzrost mobilności oraz popularności Internetu Rzeczy spowodował znaczny wzrost liczby urządzeń podłączonych do sieci Internet. Oprócz wzrostu liczby urządzeń nieustanie wzrasta również liczba sieci bezprzewodowych. Przewiduje się, że w 2022 roku sieci Wi-Fi będą obsługiwać ponad połowę całego ruchu internetowego [7].

Aby obsłużyć zwiększającą się liczbę użytkowników projektuje się sieci składające się z dużej liczby punktów dostępowych. Sieć gęstą można zdefiniować jako dowolne środowisko o dużej koncentracji użytkowników połączonych bezprzewodowo i korzystających z usług sieciowych. Ideą sieci gęstych jest duża liczba punktów dostępowych o małym zasięgu, aby oferować jak największą pojemność. Wysokie zagęszczenie można zaobserwować w centrach miast, galeriach handlowych, stadionach czy budynkach biurowych i mieszkalnych. Szczególnie problematyczne okazują się obszary gęsto zabudowane wielopoziomowymi budynkami mieszkalnymi. Obecnie prawie w każdym mieszkaniu znajduje się punkt dostępowy sieci Wi-Fi. Urządzenia te pracują zwykle z domyślnymi ustawieniami producenta, ponieważ znaczna większość użytkowników nie posiada wystarczającej wiedzy i świadomości na temat konfiguracji poszczególnych parametrów.

Ze względu na znaczną liczbę urządzeń współistniejących na danym obszarze zaobserwować można kilka negatywnych skutków mających znaczny wpływ na wydajność sieci. Mowa tutaj o przypadkach występowania stacji ukrytej i związanych z tym zjawisk takich jak efekt zakleszczenia (*Deadlock effect*), efekt eksponowanego węzła (*Exposed node effect*) czy efekt wzmocnienia sygnału na skutek interferencji. Niestety mechanizm wstępnej rezerwacji kanału radiowego RTS/CTS stosowany zazwyczaj w przypadku występowania stacji ukrytych nie zawsze rozwiązuje opisywane problemy [8]. W sieciach gęstych poprawiana konfiguracja urządzeń ma szczególne znaczenie dla osiąganej przepustowości z punktu widzenia całego badanego obszaru. Jeden źle skonfigurowany węzeł sieciowy może zakłócić transmisje we wszystkich sąsiadujących urządzeniach, ponieważ rywalizują one o ten sam kanał radiowy. Podczas planowania sieci gęstych niezwykle istotne jest znalezienie odpowiedniego balansu pomiędzy wartościami takich parametrów jak moc nadajnika i zysk anteny, CCA Threshold oraz stosowanie anten kierunkowych.

W kolejnym rozdziale podjęto próbę optymalizacji wybranych ustawień urządzeń sieciowych. Skonfigurowanie punktów dostępowych we właściwy sposób pozwoli zmniejszyć interferencje i kolizje ramek.

4. ANALIZA SYMULACYJNA SIECI GESTYCH STANDARDU IEEE 802.11ax

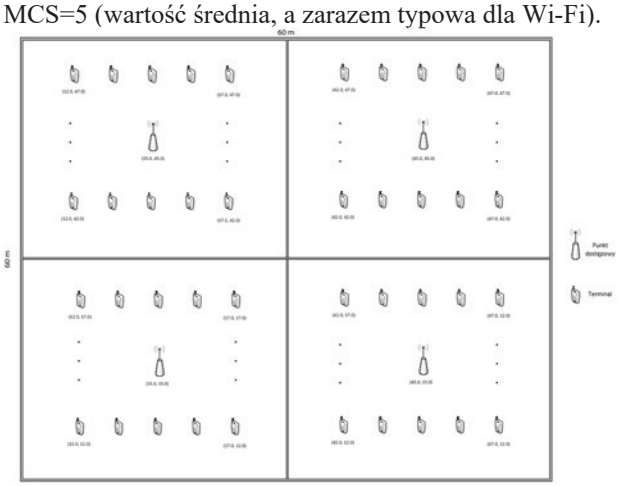
Na potrzeby przeprowadzenia testów wydajności sieci pracującej zgodnie ze standardem IEEE 802.11ax wybrano symulator NS-3 w wersji 3.31 [9]. NS-3 jest symulatorem zdarzeń dyskretnych co oznacza, że wszelkie zmiany stanu następują w dyskretnych chwilach czasu, a zdarzenia trwają nieskończenie krótko. Oprogramowanie NS-3 zostało wybrane ze względu na rozbudowaną implementację standardu IEEE 802.11ax, szczegółową dokumentację i bardzo duże możliwości konfiguracji scenariuszy badawczych. Najistotniejsze parametry symulacji wspólne dla wszystkich scenariuszy zebrano w Tabeli 1.

Tab. 1. Wartości parametrów symulacji

Parametr	*Wartość*
Pasmo częstotliwości	5 [GHz]
Szerokość kanału	20 [MHz]
Rx Noise Figure	7 [dB]
Okres ochronny	800 [ns]
Model propagacyjny	HybridBuildingsPropagationLossModel
MCS	11
RTS/CTS	Wył. / Wł.
Wielkość kolejki MAC	50 ramek
Protokół transportowy	UDP
Typ ruchu	CBR
Rozmiar ramki	1500 B

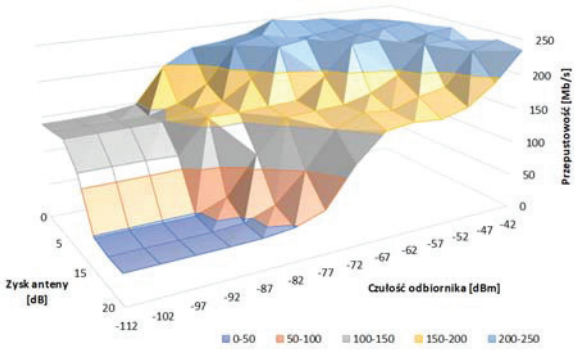
4.1. Scenariusz 1 – środowisko biurowe

W tym scenariuszu przetestowano działanie sieci w środowisku biurowym. Topologia sieci widoczna jest na Rys. 1 i składa się z 4 pomieszczeń. Na środku każdego pomieszczenia ustawiono punkt dostępowy na wysokości 3 m. Do każdego z urządzeń AP dołączono 25 stacji rozmieszczonych w siatce po 5 urządzeń w każdym rzędzie. Stacje generują ruch typu CBR i wysyłają ramki do AP z MCS=5 (wartość średnia, a zarazem typowa dla Wi-Fi).



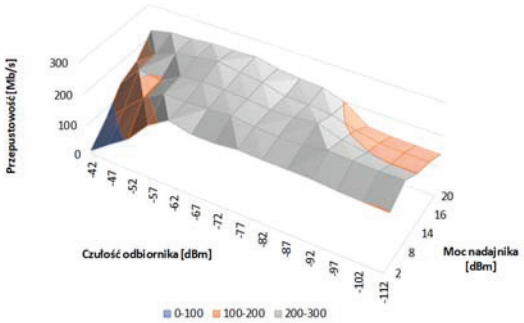
Rys. 1. Gęste środowisko biurowe

W trakcie symulacji przyjęto, że próg CCA Threshold = -92 dBm, urządzenia wyposażone są jedną antenę, a parametrami zmiennymi są: TxGain = 0-20 [dBi], RxGain = 0-20 [dBi], RxSensitivity = -112 - -42 [dBm] i TxPower = 2-20 [dBm]. Ruch oferowany ustawiono na 300 Mbit/s. Zmieniając wartości zysku energetycznego anteny oraz czułości odbiornika dla mocy nadajnika równej 16 dBm zebrano 98 punktów symulacyjnych, z których wykreślono powierzchnie dla przepustowości, średniego opóźnienia, jego zmienności oraz strat pakietów. Z uwagi na ograniczenia odnośnie maksymalnej liczby stron, w artykule przedstawiono jedynie wyniki uzyskiwanych wartości przepustowości (Rys. 2).



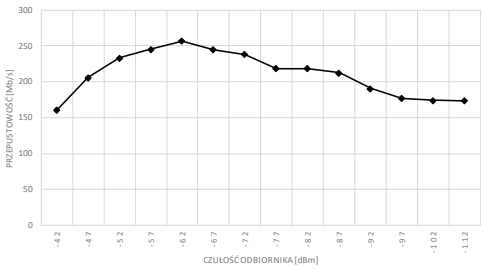
Rys. 2. Wykres przepustowości w funkcji zysku anteny oraz czułości odbiornika dla scenariusza 1

Stwierdzono, że zbyt duże wartości zysku energetycznego anten oraz wysoka czułość odbiornika powoduje wykrywanie transmisji nawet przez bardzo odległe stacje. Taka sytuacja prowadzi do drastycznego wzrostu liczby kolizji oraz długiego oczekiwania na dostęp do kanału radiowego, co w rezultacie przynosi spadek przepustowości. Kolejno, przeprowadzono analizę wpływu mocy nadajnika oraz czułości odbiornika na uzyskiwaną wydajność pracy sieci przy zysku anteny równym 0 dB (Rys. 3).



Rys. 3. Wykres przepustowości w funkcji mocy nadajnika oraz czułości odbiornika dla scenariusza 1

Maksymalną przepustowość zaobserwowano dla mocy nadajnika 16 dBm i względem takiej płaszczyzny dokonano przekroju powierzchni (Rys. 4). Z badań wynika, że sieć pracuje najefektywniej dla czułości odbiornika -62 dBm oraz mocy nadajnika 16 dBm. Przeprowadzone symulacje pokazują, że przy innych parametrach stacje są z reguły w stanie odebrać sygnały z sieci znajdujących się w sąsiadujących pomieszczeniach, co powoduje wzajemna blokadę możliwości nadawania (działanie algorytmu CSMA/CA) i spadek wydajności pracy sieci.

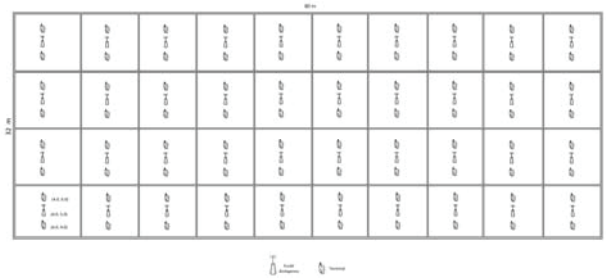


Rys. 4. Wykres przepustowości w funkcji czułości odbiornika dla scenariusza 1

Przeprowadzone badania pokazują, że zbyt duże wartości zysku anteny oraz czułości odbiornika nie są wskazane dla sieci gęstych. Należy również pamiętać, że planując zakupy urządzeń sieci WLAN jednym z podstawowych kryteriów jest zapewnienie wysokiej czułości odbiornika, zarówno po stronie AP jak i stacji klienckich. Gwarantuje to uzyskanie bezbłędnych transmisji nawet przy dużych odległościach między urządzeniami, ale niestety może oznaczać niską wydajność dla sieci gęstych. Ponieważ po zakupie urządzeń, na czułość odbiornika użytkownik nie ma już żadnego wpływu, w celu optymalizacji pracy sieci należy rozważyć możliwość zmian innych parametrów zdefiniowanych w standardzie IEEE 802.11ax. Jednym z takich parametrów jest próg wykrywania nośnej – CCA Threshold.

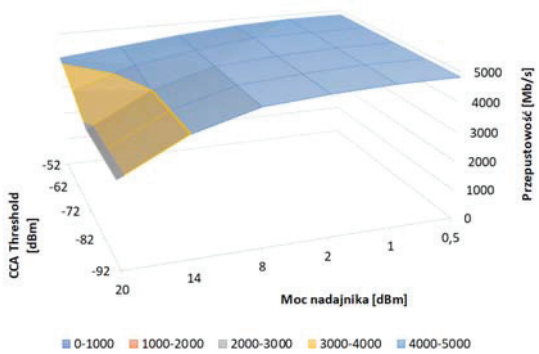
4.2. Scenariusz 2 – zabudowa mieszkaniowa

W drugim scenariuszu skupiono się na gęstej zabudowie mieszkaniowej, w której w ramach każdej z sieci pracowały dwa urządzenia klienckie, a sumaryczna liczba mieszkań (oraz AP) wyniosła 40 (Rys. 5). W każdym mieszkaniu na wysokości ok. 3 m znajduje się centralnie umieszczony punkt dostępowy.



Rys. 5. Gęsta zabudowa mieszkaniowa

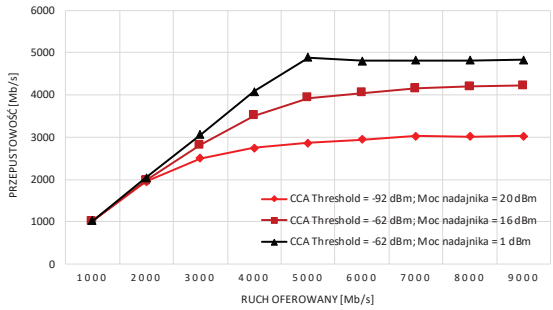
W przeprowadzonych badaniach przyjęto sumaryczny ruch oferowany równy 5,2 Gbit/s. Założono również, że Rx Sensitivity = -68 dBm, Tx Power = 0,5-20 [dBm] i próg CCA Threshold = -52 - -92 dBm. Dla zmiany wartości CCA Threshold oraz mocy nadajnika zebrano wyniki dla 30 punktów symulacyjnych i wykreślono powierzchnię, która obrazuje wielkość ruch realizowanego.



Rys. 6. Wykres przepustowości w funkcji progu CCA Threshold oraz mocy nadajnika dla scenariusza 2

Dla wyznaczonych parametrów optymalnych – moc nadajnika 1 dBm i CCA Threshold -62 dBm, najgorszych – moc nadajnika 20 dBm i CCA Threshold -92 dBm, a

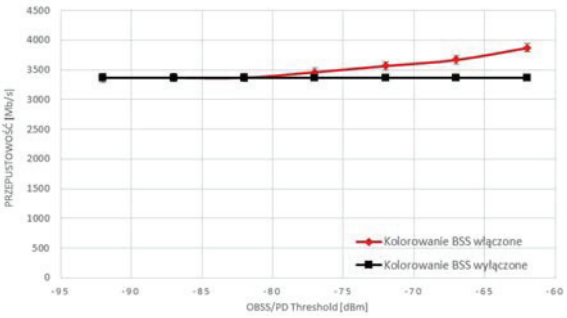
także domyślnych wartości z symulatora NS-3 – moc nadajnika 16 dBm i CCA Threshold -62 dBm przedstawiono wykres przepustowości w funkcji ruchu oferowanego.



Rys. 7. Wykres przepustowości w funkcji ruchu oferowanego dla scenariusza 2

Porównując najlepszą konfigurację z najgorszą otrzymano 60-procentowy wzrost przepustowości, 22-procentowy spadek opóźnienia, 57-procentowy spadek zmienności opóźnienia oraz o 30% mniej straconych ramek. Uzyskane wyniki potwierdzają słuszność przeprowadzania optymalizacji parametrów dla konkretnych warunków pracy. Z analizy otrzymanych metryk wynika, że zbyt niski próg CCA w połączeniu z dużą mocą nadajnika prowadzi do degradacji wydajności. Uzasadnienie tego faktu jest identyczne jak w poprzednim scenariuszu - urządzenia sieciowe wykrywają transmisję od oddalonych stacji i traktują kanał radiowy jako zajęty, co przekłada się na przepełnienie kolejek i zrzucanie ramek, a w konsekwencji spadek przepustowości.

Dla scenariusza drugiego przeanalizowano również działanie mechanizmu kolorowania BSS, który został zaproponowany dla sieci gęstych w standardzie IEEE 802.11ax. Na Rys. 8 porównano przepustowość otrzymaną z włączonym oraz wyłączonym mechanizmem kolorowania BSS - przy założeniu zmian wartości parametru OBSS/PD (*Overlapping BSS Preambule Detection Threshold*) tego mechanizmu.



Rys. 8. Wykres przepustowości w funkcji OBSS/PD przy włączonym oraz wyłączonym kolorowaniu sieci

Największy wzrost przepustowości, ok. 15-procentowy uzyskano dla progu OBSS/PD wynoszącego -62 dBm. Zbyt mała wartość tego parametru sprawia, że mechanizm kolorowania nie przynosi żadnego zysku, ponieważ zbyt niski poziom OBSS/PD sprawia, że AP jest w stanie odebrać zbyt dużo różnych transmisji. Mniejszy zysk z kolorowania niż przedstawiony w innych artykułach naukowych może być spowodowany tym, że ściany w rozważanych budynkach zapewniają częściową separacje poszczególnych sieci.

5. PODSUMOWANIE

Celem pracy było przetestowanie nowych rozwiązań wdrożonych w standardzie IEEE 802.11ax w zastosowaniu do sieci gęstych. W pracy przetestowano 2 topologie różniące się względem sobą liczbą pomieszczeń oraz rozmieszczonych w nich urządzeń sieciowych. Zaprezentowane wyniki pozwalają stwierdzić, że w sieciach gęstych kolizje ramek w znacznym stopniu ograniczają wydajność pracy sieci. Stwierdzono, że dla wydajnej pracy sieci gęstych niezwykle istotne są takie parametry jak próg wykrywania nośnej CCA Threshold, moc nadajnika, zysk anteny oraz czułość odbiornika. Ustawienie zbyt dużych wartości mocy nadawczej oraz zbyt małych wartości parametru CCA Threshold prowadzi do pogorszenia wydajności sieci. Parametry te jak pokazano w pracy mogą posłużyć do optymalizacji działania gęstych sieci WLAN.

PODZIĘKOWANIA

Praca została sfinansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

LITERATURA

[1] IEEE Std 802.11ax-2021 (Amendment to IEEE Std 802.11-2020), vol., no., pp.1-767, 19 May 2021.
[2] Khorov E., Kiryanov A., Lyakhov A. and Bianchi G., "A Tutorial on IEEE 802.11ax High Efficiency WLANs", *in IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 197-216, 2019.
[3] Szymakowski M. Natkaniec M., Prasnal Ł., Analiza symulacyjna sieci standardu IEEE 802.11ax. Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, pp. 415–419, 2019.
[4] Rochim A. F., Harijadi B., Purbanugraha Y. P., Fuad S., and Nugroho K. A., Performance comparison of wireless protocol IEEE 802.11ax vs 802.11ac. *In 2020 International Conference on Smart Technology and Applications (ICoSTA)*, pages 1–5, 2020
[5] Wilhelmi F., Barrachina-Muñoz S. and Bellalta B., "On the Performance of the Spatial Reuse Operation in IEEE 802.11ax WLANs", *2019 IEEE CSCN*, GRANADA, Spain, 2019, pp. 1-6.
[6] Malhotra A., Maity M., and Dutta A., "How much can we reuse? An empirical analysis of the performance benefits achieved by spatial-reuse of IEEE 802.11ax" *In 2019 11th Int. Conf. on Comm. Systems Networks*, 2019. Bengaluru, India.
[7] Usha Andra Taru Khurana Thomas Barnett, Jr. Shruti Jain. Cisco visual networking index, 2018.
[8] Zhong Z., Kulkarni P., Cao F., Fan Z., and Armour S., "Issues and challenges in dense WiFi networks", *In 2015 IWCMC*, pages 947–951, 2015. Dubrovnik, Croatia
[9] Symulator ns-3. [Online] http://www.nsnam.org/.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

ANALIZA DZIAŁANIA MECHANIZMU KOLOROWANIA W SIECIACH GESTYCH STANDARDU IEEE 802.11AX

AN ANALYSIS OF BSS COLORING MECHANISM IN IEEE 802.11AX DENSE NETWORKS

Marek Natkaniec[1]; Natalia Bieryt[2]

[1] AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, natkanie@agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, bieryt.natalia@gmail.com

DOI: 10.15199/59.2022.4.102

Streszczenie: **W artykule przedstawiono analizę działania mechanizmu kolorowania zdefiniowanego w standardzie IEEE 802.11ax. Przeanalizowano i porównano wydajność pracy sieci gęstych dla różnych scenariuszy pracy. Zaobserwowano pozytywny wpływ działania mechanizmu kolorowania na jakość świadczonych usług QoS. Dzięki przeprowadzonym badaniom pokazano, że właściwy dobór parametrów mechanizmu kolorowania w standardzie IEEE 802.11ax ma silny wpływ na QoS i wydajność pracy sieci gęstych.**

Abstract: **The paper presents an analysis of BSS color scheme defined in IEEE 802.11ax standard. The efficiency of dense networks for different scenarios was analyzed and compared. A positive impact of the coloring mechanism on the quality of the services was observed. It was shown that the proper selection of the coloring mechanism parameters in the IEEE 802.11ax standard has a strong impact on QoS and the performance of dense networks.**

Słowa kluczowe: **kolorowanie sieci, sieci gęste, IEEE 802.11ax, QoS.**

Keywords: **network coloring, dense networks, IEEE 802.11ax, QoS.**

1. WSTĘP

W ostatnich latach aplikacje oferujące usługi dostarczania strumieni wideo oraz dźwięku zyskały ogromną popularność, powodując tym samym coraz większe wymagania dotyczące jakości transmisji. Konieczność nieustannej poprawy jakości świadczonych usług QoS (*Quality of Service*) jest spowodowana również dynamicznym rozwojem tak zwanego Internetu rzeczy IoT (*Internet of Things*), który korzysta z najważniejszej funkcjonalności oferowanej przez sieci bezprzewodowe jaką jest mobilność. Jednak aplikacje często posiadają różne wymagania co do priorytetu ruchu, co powoduje potrzebę dużego zróżnicowania w zakresie jakości świadczonych usług. Sytuację pogarsza fakt istnienia, szczególnie w miejscach silnie zurbanizowanych, tzw. sieci gęstych (*dense networks*), czyli sieci w których występuje bardzo duża liczba stacji oraz punktów dostępowych rozmieszczonych na stosunkowo małym obszarze. Powoduje to poważne problemy, gdyż taka liczba urządzeń bezprzewodowych generujących różnorodne strumienie danych zwiększa częstotliwość wystąpień kolizji.

Standard IEEE 802.11ax poza dalszym zwiększaniem możliwości transmisji oferuje też szereg nowych funkcjonalności pozwalających na poprawę wydajności pracy w warunkach współistnienia dużej liczby urządzeń sieciowych [1]. Celem niniejszej pracy jest analiza symulacyjna mechanizmu kolorowania sieci zdefiniowanego w standardzie IEEE 802.11 oraz określenie wpływu jego parametrów na jakość świadczonych usług.

Praca składa się z sześciu rozdziałów. Rozdział 2 przedstawia przegląd literatury. W rozdziale 3 przybliżono krótko standard IEE 802.11ax, a w kolejnym rozdziale omówiono mechanizm kolorowania sieci. W rozdziale 5 zawarto wyniki badań symulacyjnych przeprowadzonych dla dwóch różnych scenariuszy pracy sieci bezprzewodowej. Artykuł kończy rozdział 6, który zawiera podsumowanie niniejszej pracy.

2. PRZEGLĄD LITERATURY

W publikacji [2] analizowano wpływ współistnienia stacji pracujących w standardzie IEEE 802.11ax ze starszymi rozszerzeniami standardu. Zaobserwowano, że wydajność sieci wzrosła wraz z większym procentowym udziałem stacji wspierających najnowsze rozszerzenie standardu sieci bezprzewodowych. Autorzy zwrócili również uwagę na zjawisko lepszego traktowania użytkowników znajdujących się bliżej zasocjowanego punktu dostępowego spowodowane większymi interferencjami stacji znajdujących się na skraju sieci z sąsiednimi sieciami WLAN (*Wireless Local Area Network*). Kolejno w [3] wyciągnięto wniosek, że stosując mechanizm kolorowania sieci oraz wyższe wartości modulacji i schematy kodowania, nie zawsze osiągamy większe przepustowości, co wynika z dużych wymagań SINR (*Signal-to-Interference-plus-Noise Ratio*) dla tych trybów modulacji. Autorzy zaproponowali również algorytm, ustawiający odpowiedni próg OBSS/PD, który ma zapewnić sprawiedliwe traktowanie użytkowników w sieci. W wyniku jego zastosowania zaobserwowano zysk w przepustowości wynoszący 57%. W artykule [4] zbadano wpływ zmiany wartości progu OBSS_PD na osiąganą przepustowość w sieci dla siedmiu punktów dostępowych, z których każdy jest skojarzony z dziesięcioma stacjami. Zysk dla podanej topologii wyniósł 13% dla wartości OBSS_PD threshold

równej -72dBm, którą to poprzez przeprowadzenie licznych symulacji, uznano za najbardziej korzystną dla podanego scenariusza.

3. STANDARD IEEE 802.11ax

Nowy standard IEEE 802.11ax wprowadza wiele zmian, mających na celu poprawę efektywności widmowej. Jednak mimo to wciąż zapewnia kompatybilność z innymi wersjami standardu IEEE 802.11. Pasma częstotliwości, w których może pracować, to zarówno 2,4 GHz, jak i 5 GHz. Podobnie jak we wcześniejszym standardzie IEEE 802.11ac dostępne są cztery szerokości kanałów: 20 MHz, 40 MHz, 80 MHz oraz 160 MHz. Znaczącą zmianą jest czterokrotne zmniejszenie odległości między podnośnymi w technice OFDM (*Orthogonal Frequency-Division Multiplexing*). Skutkiem czego dostępna jest czterokrotnie większa liczba kanałów radiowych, które zostały rozmieszczone z odstępem 78,125 kHz. Zabieg ten doprowadził również do czterokrotnego wydłużenia czasu trwania symbolu, który wyniósł 12,8 µs, co spowodowało zwiększoną odporność sieci na różne zakłócenia. Dłuższy symbol wymaga odpowiedniej wartości trwania okresu ochronnego GI (*Guard Interval*), który zgodnie z nowym standardem przyjął wartości 800 ns, 1600 ns oraz 3200 ns [5].

Warstwa fizyczna wprowadza również kolejną modulację jaką jest 1024-QAM, dzięki której jeden symbol jest w stanie przenieść 10 bitów informacyjnych. Do tej pory lista dostępnych trybów modulacji wynosiła 9. Dzięki zastosowaniu nowej modulacji rozszerzono ten zakres do 11. Nowe indeksy 10 i 11 to modulacje 1024-QAM, stosujące sprawność kodu korekcyjnego kolejno 3/4 oraz 5/6.

IEEE 802.11ax zdefiniował dwa tryby pracy, jakimi są Single User, czyli tryb, w którym jednocześnie może nadawać jedna stacja, oraz Multi User, który pozwala na równoległe transmisje wielu klientów. Transmisje Multi User zostały uznane za kluczowe w rozwiązaniu problemów sieci gęstych. Mechanizmami, które pozwalają na jednoczesną transmisję wielu użytkowników są między innymi Multi-User MIMO oraz nie używana wcześniej w standardach rodziny IEEE 802.11 technika OFDMA (*Orthogonal Frequency Division Multiple Access*). Poprzednie rozszerzenie IEEE 802.11ac również posiadało możliwość transmisji w trybie MU-MIMO, jednak transmisja ta mogła się odbywać jedynie w „łączu w dół" [6]. Najnowsze rozszerzenie zniwelowało to ograniczenie poprzez umożliwienie korzystania z techniki MU-MIMO zarówno przy transmisji w „łączu w dół" jak i „w górę", przeznaczając na to osiem równoległych strumieni. Nowo wprowadzona technika OFDMA umożliwiła podział kanału na jednostki zasobów, nazywane RU (*Resource Units*), które mogą składać się z kilku podnośnych, a te z kolei przydzielane są konkretnemu użytkownikowi.

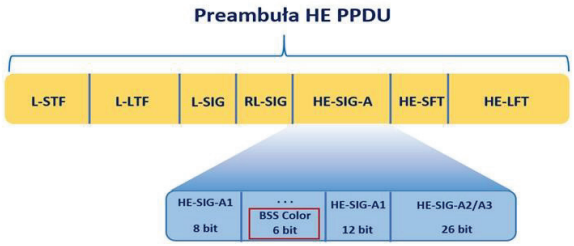
Rozszerzenie IEEE 802.11ax dodatkowo zdefiniowało cztery formaty ramek PPDU (*PHY Protocol Data Unit*). Pierwsza z nich przeznaczona została do transmisji typu SU (*Single User*), kolejna, tak zwana Extended Range SU PPDU do transmisji na dalsze odległości jedynie w paśmie o szerokości 20 MHz, trzecia natomiast MU (*Multi User*) PPDU, została przedstawiona w celu transmisji do wielu użytkowników oraz ostatnia Triggered Based PPDU, przeznaczona do transmisji w „łączu w górę" przez wielu użytkowników [7].

Standard IEEE 802.11ax wprowadza kilka zmian w zakresie agregacji [7]. Maksymalna wielkość zagregowanej ramki dla A-MPDU wynosi 8 388 607 bajtów, natomiast A-MSDU może zawierać 11 398 bajtów [1]. Kolejną zmianą jest rozszerzenie mechanizmu TWT (*Target Wake Time*), oferowanego już w rozszerzeniu IEEE 802.11ah, który pozwala na wprowadzenie urządzenia w tryb uśpienia, w przypadku jego braku udziału w transmisji. Spowodowało to znaczne zredukowanie poziomu poboru mocy, a co za tym idzie wydłużenie czasu życia baterii. Stacja jest wybudzana co odcinek SP (*Service Period*) na określony zarezerwowany czas, podczas którego dostaje możliwość wymiany informacji z punktem dostępowym. Urządzenie wówczas nie jest wybudzane nawet w celu odebrania ramek *Beacon*, co ma na celu znaczne zaoszczędzenie energii [7].

IEEE 802.11ax posiada również możliwość wirtualizacji sieci [7], która polega na stworzeniu wirtualnych punktów dostępowych, nazywanych Virtual Access Point (VAP). Dzięki temu, możliwym jest posiadanie nawet do kilkunastu identyfikatorów BSSID, co daje okazję do wydzielenia na przykład sieci przeznaczonej dla gości, bez konieczności posiadania dwóch fizycznych punktów dostępowych. Wymagane jest jednak, aby każda z nich posiadała ten sam kolor sieci.

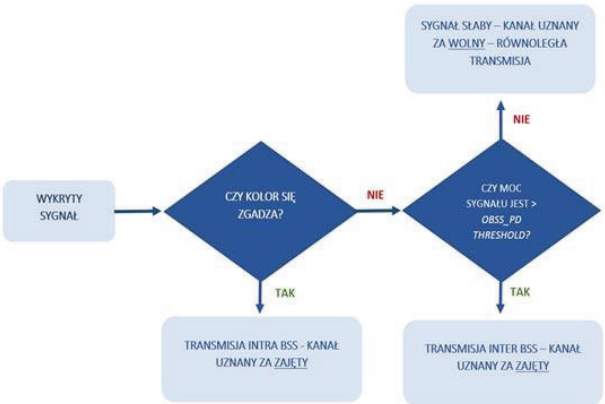
4. MECHANIZM KOLOROWANIA SIECI

Mechanizm kolorowania sieci jest metodą pozwalającą na efektywne wykorzystanie dostępnego pasma. Umożliwia on rozróżnienie transmisji pochodzących z różnych obszarów BSS (*Basic Service Set*), odbywających się w tym samym kanale, za pomocą koloru. Przez kolor rozumiemy sześciobitową wartość zawartą w preambule warstwy fizycznej, której miejsce zostało przedstawione na Rys. 1. Maksymalna liczba możliwych kolorów wynosi zatem 63, gdzie kolor 0 oznacza wyłączony mechanizm. Obok parametru CCA threshold zdefiniowany został kolejny próg, jakim jest OBSS_PD threshold, czyli tzw. *Overlapping BSS Preambule Detection Threshold*. Jego celem jest dopuszczenie możliwości równoległych transmisji a zarazem kontrola ich ilości, gdyż zbyt duża ich liczba powoduje zwiększenie poziomu interferencji.



Rys. 1. Preambuła warstwy fizycznej w IEEE 802.11ax

Pierwszym krokiem dokonywanym przez odbiorcę jest porównanie koloru zawartego w preambule przychodzącej ramki z kolorem swojej sieci. W przypadku niezgodności, odbiornik nie dekoduje ramki, oszczędzając przy tym energię. Następnie sygnał porównywany jest z nowo zdefiniowanym progiem OBSS_PD. Jeśli wykryty poziom mocy jest wyższy od zadanego progu, kanał uznany jest za zajęty. W przeciwnym wypadku sygnał jest na tyle słaby, że jest dopuszczona możliwość dokonania równoległej transmisji w kanale [3], [8]. Na Rys. 2 został przedstawiony algorytm działania mechanizmu kolorowania sieci.



Rys. 2. Algorytm działania mechanizmu kolorowania sieci w standardzie IEEE 802.11ax

Standard IEEE 802.11ax określa warunek, dotyczący tego, jakie wartości może przyjąć próg OBSS_PD:

$$OBSS_{PD} \leq \max(OBSS_{PD_{min}}, \min(OBSS_{PD_{max}}, OBSS_{PD_{min}} + (TX_{PWR_{ref}} - TX_{PWR}))) \quad (1)$$

gdzie

$OBSS_{PD_{min}} = -82\ dBm,\ OBSS_{PD_{max}} = -62\ dBm,$ $TX_{PWR_{ref}} = 21\ dBm$, a TX_{PWR} to moc nadawcza stacji wyrażona w dBm [1].

W wyniku wsparcia kolorowania sieci, w IEEE 802.11ax zostały wprowadzone również dwa typy wektorów alokacji sieci NAV (*Net Allocation Vector*), jeden przeznaczony dla transmisji z aktualnej sieci, a kolejny dla transmisji pochodzących z obcych obszarów BSS. Kanał jest uznany za wolny, gdy licznik obu wektorów NAV wyniesie zero [7].

Dodatkowo, AP (*Access Point*) posiada możliwość wykrycia konfliktu kolorów, czyli sytuacji powtórzenia się tego samego koloru w zachodzących na siebie wzajemnie sieciach. W przypadku wystąpienia takiego zdarzenia, stacja również posiada możliwość poinformowania o nim punktu dostępowego, z którym jest skojarzona. Punkt dostępowy dokonuje zmiany koloru poprzez wysłanie dedykowanej do tego ramki *BSS Color Change Announcement*, bądź też poinformowanie o tym przy udziale ramek *Beacon*, *Probe Response*, czy też *Reassociation Response*.

5. BADANIA SYMULACYJNE SIECI STANDARDU IEEE 802.11ax

W niniejszym rozdziale przedstawiono wyniki badań symulacyjnych wykonane przy pomocy symulatora zdarzeniowego NS (*Network* Simulator) w wersji 3.31 [9]. Narzędzie to pozwala na przeprowadzenie wysoko realistycznych symulacji sieciowych, w tym również bezprzewodowych sieci WLAN. W celu zapewnienia QoS, zastosowano funkcję EDCA (*Enhanced Distributed Channel Access*), która wciąż jest jedną z najważniejszych funkcji dostępu do kanału radiowego w rozszerzeniu IEEE 802.11ax. Najważniejsze parametry symulacji wspólne dla wszystkich scenariuszy zebrano w Tabeli 1. Na wszystkich rysunkach błąd każdego punktu symulacji dla 95% przedziału ufności nie przekraczał ±2%.

Tab. 1. Wartości parametrów symulacji

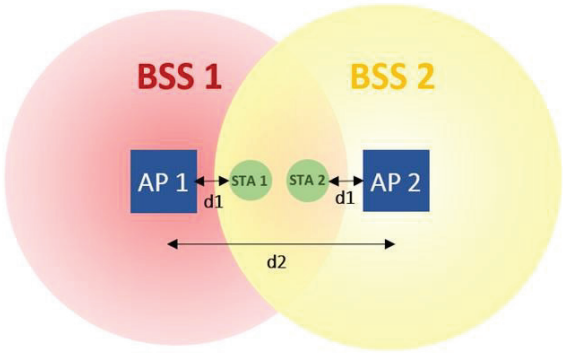
Parametr	*Wartość*
Pasmo częstotliwości	5 [GHz]
Szerokość kanału	20 [MHz]
Poziom mocy nadawczej stacji	15 [dBm]
Poziom mocy nadawczej AP	20 [dBm]
Czułość odbiornika	-92 [dBm]
Liczba anten	1
MCS	2 (QPSK ¾)
CCA ED threshold	-62 [dBm]
CCA SD threshold	-82 [dBm]
Próg OBSS_PD	(-82 dBm, -62 dBm)
Okres ochronny	800 [ns]
RTS/CTS	Wył. / Wł.
Fragmentacja	Wył.
Klasy ruchu	BK, BE, VI, VO
TXOP limit	BK = 0, BE = 0, VI=3, VO=1,5 [ms]
Protokół transportowy	UDP
Typ ruchu	CBR
Ruch oferowany	20 Mbit/s
Rozmiar ramki	1500 B
d1	20m
d2	(50m, 400m)

5.1. Scenariusz 1 – 2 AP

Topologia (Rys. 3) składa się z dwóch punktów dostępowych AP oraz dwóch stacji STA. Stacja numer jeden (STA 1) należy do sieci o identyfikatorze BSS 1, natomiast stacja numer dwa do sieci o identyfikatorze BSS 2. Stacja STA 1 nadaje do AP 1 oraz konsekwentnie STA 2 nadaje do AP 2. Parametr d2 to odległość pomiędzy dwoma punktami dostępowymi (AP1 oraz AP2), natomiast d1 to odległość stacji od punktu dostępowego, do którego nadaje.

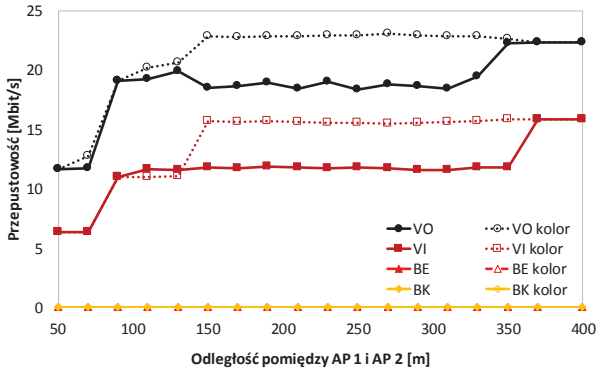
W scenariuszu przeanalizowano wpływ odległości d2 między AP1 oraz AP2 jak i wpływ zmiany parametru OBSS_PD na sumaryczną przepustowość obu sieci. Obie te sieci pracują w warunkach nasycenia, aby jak najlepiej zaobserwować działanie mechanizmu kolorowania sieci, jak również korzystają z tego samego kanału transmisyjnego. Zasięg sieci został ustalony poprzez stopniowe odsuwanie stacji od punktu dostępowego i obserwowanie zmieniającej się przepustowości. Zarówno dla BSS 1 jak

i BSS 2 wyniósł on około 180 metrów przy użyciu modulacji QPSK. Zatem aby całkowicie odizolować obie sieci, odległość między AP1 oraz AP2 musi wynosić minimum 360 metrów.



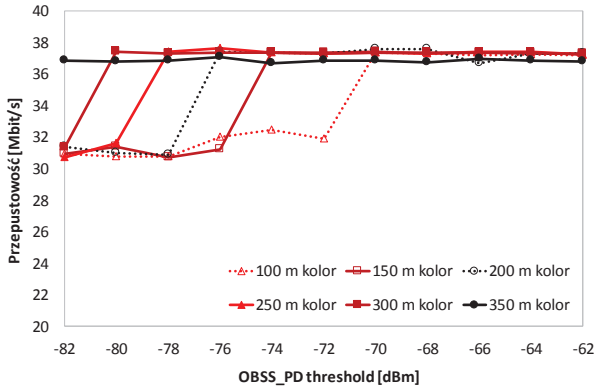
Rys. 3. Topologia z dwoma punktami AP

Dla podanej symulacji parametr OBSS_PD threshold został ustawiony na -72dBm, co oznacza, że gdy zostanie odebrany sygnał pochodzący z innej sieci, musi być on słabszy od zadanego progu, aby została dokonana równoległa transmisja. Rys. 4 przedstawia sumaryczną przepustowość dla obu sieci oraz czterech klas ruchu w zależności od odległości między punktami dostępowymi.



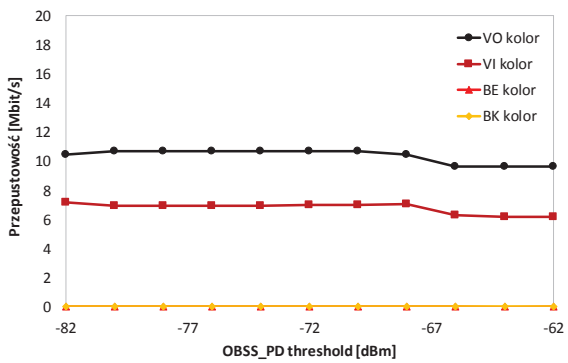
Rys. 4. Sumaryczna przepustowość dla obu sieci oraz czterech klas ruchu, w zależności od odległości d2

Z przeprowadzonych symulacji (Rys. 4) wynika, że gdy odległości między sieciami BSS 1 oraz BSS 2 są małe, ich zasięgi znacznie się pokrywają, a co za tym idzie dochodzi do licznych interferencji oraz kolizji. Zaobserwowano, że po rozsunięciu punktów dostępowych na odległość ok. 90-ciu metrów, poziom interferencji jest na tyle mały, że osiągane przepustowości są znacznie wyższe. Ten sam poziom przepływności utrzymuje się do momentu rozsunięcia stacji na odległość 300-stu metrów, a następnie stopniowo rośnie aż do osiągnięcia maksymalnej przepustowości wynoszącej około 38 Mbit/s (dla wszystkich klas ruchu), sumarycznie dla obu sieci. Zauważono również, że w przypadku włączonego kolorowania, przepustowość wyraźnie rośnie w okolicach 150-tego metra odległości. Wtedy to poziomy mocy sygnałów dochodzących z sąsiedniej sieci są słabsze od zadanego progu OBSS_PD, w wyniku czego dopuszczona zostaje równoległa transmisja. Procentowy zysk z zastosowania mechanizmu kolorowania sieci, dla podanego scenariusza badawczego, wyniósł około 20%. Wszystkie klasy ruchu zachowują się podobnie. Dla podanych parametrów, zysk wynikający z zastosowania mechanizmu kolorowania, zarówno dla klasy Video jak i Voice był bardzo podobny, czyli od 150-tego metra zaobserwowano poziom przepustowości taki, jaki sieci osiągałyby w warunkach odizolowania. Świadczy to o prawidłowym traktowaniu ruchu, który posiada różne priorytety przez mechanizm kolorowania. Niestety, w warunkach nasycenia osiągane przez klasy BE i BK wartości przepustowości oscylują w okolicach zera, co oznacza, że te klasy ruchu nie są prawidłowo obsługiwane w przypadku występowania sieci gęstej.



Rys. 5. Sumaryczna przepustowość dla obu sieci w zależności od wartości progu OBSS_PD w zależności od odległości d2

Kolejno przebadano wpływ zmiany wartości progu OBSS_PD na osiągane przepływności dla różnych odległości między AP1 oraz AP2 (Rys. 5). Zaobserwowano, że im mniejsza odległość między sieciami, tym wyższy próg jest wymagany, aby zaobserwować zysk wynikający z zastosowania mechanizmu kolorowania. Wynika to z braku możliwości dokonania równoległych transmisji przy niskiej wartości progu OBSS_PD. Przykładowo dla odległości 100 metrów między sieciami, moce sygnałów dochodzących z obcej sieci są wyższe niż -82 dBm, a zarazem niższe niż -70dBm, co dla tej odległości wymaga ustawienia progu OBSS_PD na wartość równą co najmniej -70 dBm, w celu osiągnięcia wyższej przepustowości. Dla odległości 350-ciu metrów możemy zastosować dowolną wartość progu z zakresu (-82 dBm, -62 dBm), gdyż zasięgi sieci pokrywały się nieznacznie i zysk płynący z mechanizmu kolorowania został zauważony przy zastosowaniu dowolnej wartości progu OBSS_PD.

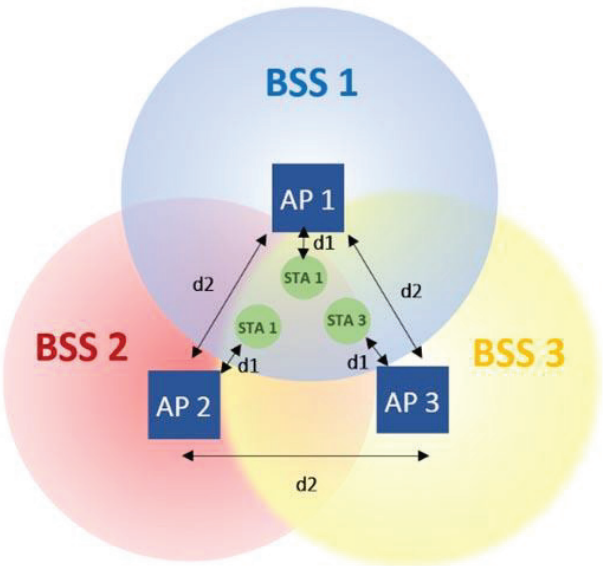


Rys. 6. Sumaryczna przepustowość dla czterech klas ruchu i obu sieci w zależności od wartości progu OBSS_PD dla odległości d2=50m

Zaobserwowano również, że przy niewielkiej odległości między AP1 oraz AP2 występuje spadek przepustowości dla wartości progu OBSS_PD większych niż -66 dBm (Rys. 6). Zjawisko to wynika ze zwiększonego poziomu interferencji, która z kolei wynika z dużej liczby równoległych transmisji w kanale, a co za tym idzie zwiększonej liczby strat pakietów. Ważną obserwacją jest również odpowiednie traktowanie klas ruchu, które jest stosowne do przyznanego im priorytetu dla dowolnej wartości parametru OBSS_PD threshold.

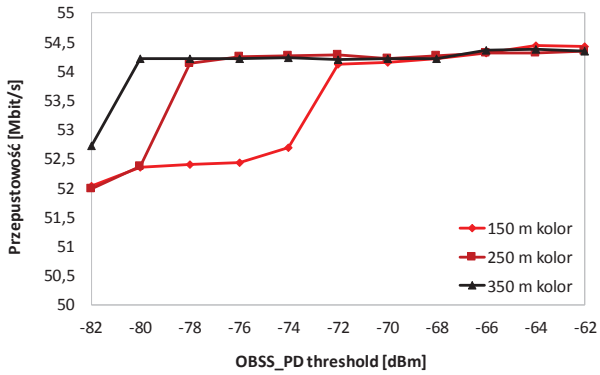
5.2. Scenariusz 2 – 3 AP

Drugi scenariusz rozszerzył topologię przedstawioną w punkcie 5.1 o kolejną sieć o identyfikatorze BSS 3. Wszystkie trzy sieci ustawione zostały w konfiguracji o współrzędnych tworzących trójkąt równoboczny. Stacje zostały umieszczone w obrębie stworzonego przez punkty dostępowe trójkąta, tak jak zostało to przedstawione na schemacie sieci (Rys. 7). Transmisja wszystkich trzech sieci odbyła się we wspólnym kanale. Zasięg każdej z nich wyniósł około 180 metrów. W podanym scenariuszu zbadano wpływ zmiany odległości d2 między sieciami oraz wartości progu OBSS_PD na osiągane przepustowości sieci oraz zysk wynikający z zastosowania mechanizmu kolorowania.



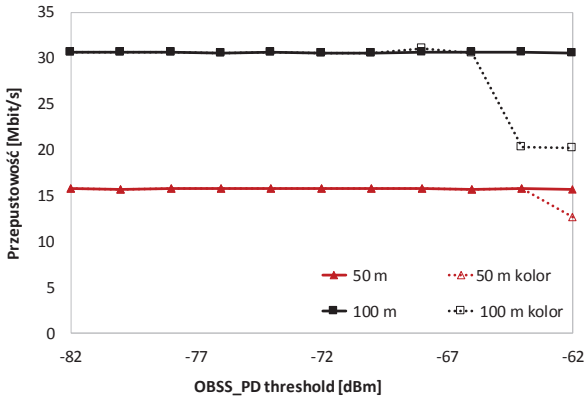
Rys. 7. Topologia z trzema punktami AP

Jako pierwszy przebadano wpływ zmiany wartości progu OBSS_PD na sumaryczną przepustowość trzech sieci (Rys. 8). Podobnie jak w poprzednim scenariuszu, poprzez zmniejszenie wartość parametru OBSS_PD threshold zysk wynikający z zastosowania kolorowania jest widoczny dla większych odległości między sieciami, który dla podanej topologii wyniósł około 4%. Zaobserwowano również, że klasy ruchu, przy odległości między punktami dostępowymi większymi niż 150 metrów, są traktowane zgodnie z przydzielonymi im priorytetami.



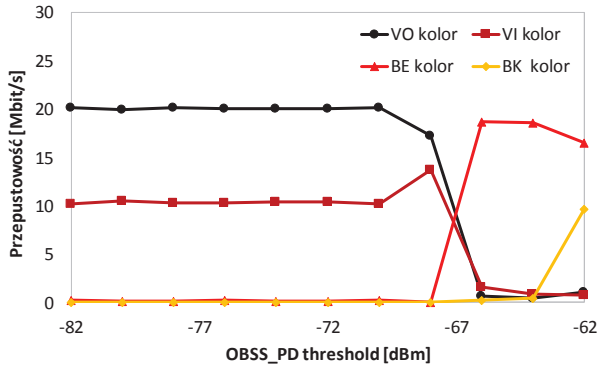
Rys. 8. Sumaryczna przepustowość trzech sieci w zależności od wartości OBSS_PD dla różnych odległości d2

W omawianym scenariuszu widoczny jest również skutek ustawienia zbyt wysokiego progu OBSS_PD, dopuszczającego więcej równoległych transmisji w tym samym kanale. Przy odległości 100 metrów między każdym punktem dostępowym, zastosowanie parametru OBSS_PD threshold większego lub równego -64 dBm, skutkuje spadkiem przepustowości nawet o 10 Mbit/s (Rys. 9).



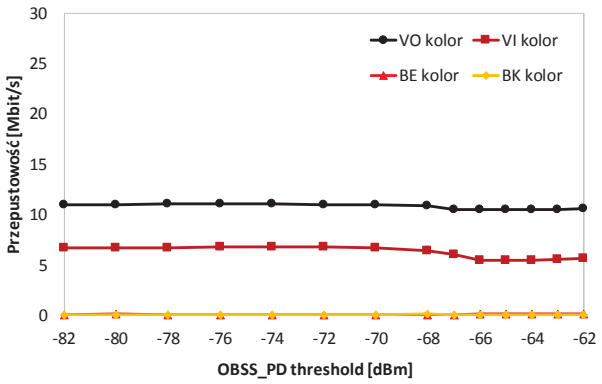
Rys. 9. Sumaryczna przepustowość dla trzech sieci, w zależności od progu OBSS_PD dla odległości d2 między sieciami równej 50 oraz 100 metrów

Ponadto zaobserwowano, że skutkiem ustawienia tak wysokiego progu jest nieodpowiednie traktowanie priorytetów klas ruchu, powodując ich całkowite odwrócenie, tak jak zostało to pokazane na Rys. 10.



Rys. 10. Sumaryczna przepustowość dla trzech sieci, z rozróżnieniem na poszczególne klasy ruchu, w zależności od progu OBSS_PD dla odległości d2 = 100 metrów

Podobne zjawisko zostało zaobserwowane w sieciach bezprzewodowych ad-hoc, stosujących funkcję EDCA w celu nadania różnych priorytetów, w których występuje tak zwany problem stacji ukrytych. W publikacji [10] zaobserwowano tendencję „mieszania się" priorytetów klas ruchu w obecności stacji ukrytych. Autorzy wskazują, że najbardziej znanym mechanizmem radzenia sobie z problemem, jakim są stacje ukryte, jest zastosowanie tak zwanego mechanizmu four-way-handshake, który opiera się na wymianie ramek RTS (*Request to Send*) i CTS (*Clear to Send*) przed wysłaniem ramki DATA. Autorzy wspominają również o wadach takiego rozwiązania, które obniża przepustowość sieci, co wynika ze znacznego zajęcia dostępnego pasma przez ramki kontrolne.



Rys. 11. Sumaryczna przepustowość dla trzech sieci, z rozróżnieniem na poszczególne klasy ruchu, w zależności od progu OBSS_PD dla odległości d2 = 100 metrów oraz przy zastosowaniu mechanizmu RTS/CTS

Sugerując się wnioskami pochodzącymi z wyżej opisanych publikacji włączono mechanizm RTS/CTS w celu próby eliminacji nierównego traktowania klas ruchu. Na Rys. 11 przedstawiono ruch pochodzący ze wszystkich trzech sieci, przynależący do czterech klas ruchu przy zastosowaniu ramek kontrolnych RTS oraz CTS. Zaobserwowano, że mechanizm poradził sobie ze zjawiskiem mieszania się klas o zadanych priorytetach, jednak ruch ten obniżył wydajność pracy sieci dla wszystkich klas ruchu.

6. PODSUMOWANIE

Kolorowanie sieci to mechanizm oferowany przez standard IEEE 802.11ax, który w odpowiednich okolicznościach oraz przy odpowiednim ustawieniu parametrów, przynosi wymierne korzyści w zakresie osiąganych przepustowości. Na zysk ten spory wpływ ma sama topologia sieci, czyli ustawienie stacji oraz punktów dostępowych, a zwłaszcza odległości między nimi. Badania pokazują, że największy zysk został osiągnięty w przypadku jednoczesnej pracy dwóch sieci w jednym kanale i wyniósł on około 20%. Manipulacja progiem OBSS_PD również daje dużą możliwość uzyskania zysku płynącego z zastosowania tego mechanizmu, gdyż jego wysoka wartość pozwala na większą liczbę równoległych transmisji. Należy jednak pamiętać, że może to prowadzić do znacznych interferencji, które mogą spowodować spadek osiąganej przepustowości. Mechanizm kolorowania sieci przy niskich wartościach progu OBSS_PD prawidłowo radzi sobie z obsługiwanymi priorytetami ruchu, a więc odpowiednio traktuje klasy definiowane przez funkcję EDCA. Są jednak przypadki, kiedy dochodzi do całkowitego zaburzenia pracy klas ruchu. Koniecznym więc może okazać się włączenie mechanizmu RTS/CTS, tak aby usługi były świadczone zgodnie z przydzielonym mu priorytetem. Należy również odpowiednio dobrać wartości progu OBSS_PD, który musi być na tyle wysoki, aby dopuścić równoległe transmisje w kanale, dzięki którym obserwujemy zysk w zakresie osiąganych przepustowości, ale równocześnie na tyle niski, aby nie spowodować dużych interferencji w kanale transmisyjnym.

PODZIĘKOWANIA

Praca została sfinansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

LITERATURA

[1] IEEE Std 802.11ax-2021 (Amendment to IEEE Std 802.11-2020), vol., no., pp.1-767, 19 May 2021.
[2] Selinis I., Filo M., Vahid S., Rodriguez J. and Tafazolli R., "Evaluation of the DSC algorithm and the BSS color scheme in dense cellular-like IEEE 802.11ax deployments," *2016 IEEE 27th International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, 2016.
[3] Selinis I., Katsaros K., Vahid S. and Tafazolli R., "Control OBSS/PD Sensitivity Threshold for IEEE 802.11ax BSS Color," *2018 IEEE 29th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Bologna, 2018, pp. 1-7.
[4] Šepić N., Kočan E. and Pejanović-Djurišić M., "Evaluating spatial reuse in 802.11ax networks with interference threshold adjustment," *2020 24th International Conference on Information Technology (IT)*, Zabljak, Montenegro, 2020, pp. 1-4.
[5] "Introduction to 802.11ax High-Efficiency Wireless", *National Instruments White paper*, Aug 2020
[6] Lee K., "Using OFDMA for MU-MIMO User Selection in 802.11ax-Based Wi-Fi Networks," *in IEEE Access*, vol. 7, pp. 186041-186055, 2019.
[7] E. Khorov, A. Kiryanov, A. Lyakhov and G. Bianchi, "A Tutorial on IEEE 802.11ax High Efficiency WLANs," *in IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 197-216, 2019.
[8] F. Wilhelmi, S. Barrachina-Muñoz and B. Bellalta, "On the Performance of the Spatial Reuse Operation in IEEE 802.11ax WLANs," *2019 IEEE Conference on Standards for Communications and Networking (CSCN)*, GRANADA, Spain, 2019, pp. 1-6.
[9] Symulator ns-3. [Online] http://www.nsnam.org/.
[10] K. Kosek, M. Natkaniec, L. Vollero „Problems with Correct Traffic Differentiation in Line Topology IEEE 802.11 EDCA Networks in the Presence of Hidden and Exposed Nodes" *Computational Science and Its Applications – ICCSA* 2009, Volume 5593.

BTS (BACKGROUND TRAFFIC SIMULATOR) – NARZĘDZIE DO GENEROWANIA RZECZYWISTEGO RUCHU DLA SIECI WLAN W ŚRODOWISKU NS-3

BTS (BACKGROUND TRAFFIC SIMULATOR) – THE TOOL FOR WLAN BACKGROUND TRAFFIC GENERATION IN NS-3 ENVIRONMENT

Szymon Stryczek[1]; Mikołaj Gwiazdowicz[2]; Marek Natkaniec[3]

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, szst@agh.edu.pl
[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, mikolaj@agh.edu.pl
[3] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, natkanie@agh.edu.pl

DOI: 10.15199/59.2022.4.103

Streszczenie: **Narzędzie BTS służy do zapisu ruchu sieciowego, a następnie odzwierciedlenia go w symulacji za pomocą modułu rozszerzającego symulator ns-3. Ta nowa metoda generowania ruchu tła, pozwala na wielokrotne testowanie scenariuszy symulacyjnych w warunkach rzeczywistych. Autorzy opisali różnice między dotychczas dostępnymi metodami, a proponowanym rozwiązaniem. Opisano działanie obu elementów opracowanego rozwiązania. Przedstawiono wpływ ruchu tła na realizację scenariusza symulacyjnego oraz wynik działania opisywanego narzędzia.**

Abstract: **The BTS tool is used to capture network traffic and then reproduce it in a simulation using the ns-3 simulator extension module. This new method of generating background traffic allows for repetitive testing of simulation scenarios under real conditions. The authors described the differences between the previously available methods and the proposed solution. The operation of both elements of the solution has been described. The influence of the generated background traffic on the simulation scenario was presented as well as the effects of the developed tool.**

Słowa kluczowe: **generowanie ruchu, ns-3, ruch tła, symulator, WLAN.**

Keywords: **background traffic, ns-3, simulator, traffic generation, WLAN.**

1. WSTĘP

Zaprezentowane w pracy narzędzie ma za zadanie odwzorowanie warunków rzeczywistego środowiska radiowego w symulatorze ns-3 [7]. Celem takiego odwzorowania mogą być badania wpływu warunków zewnętrznych na testowaną technologię. Narzędzie składa się z dwóch modułów, pierwszy służący do zarejestrowania ruchu w otoczeniu w określonym przedziale czasu, a drugi do jego reprezentacji w symulacji. Dodatkową funkcjonalnością pierwszego modułu jest możliwość przetworzenia już istniejących zapisów w formacie PCAP do formy akceptowalnej przez element symulacyjny. W referacie przedstawiono oferowane przez ns-3 metody generowania ruchu oraz przykłady ich wykorzystania. Opisywane narzędzie wprowadza do symulatora nową możliwość generowania ruchu, szczególnie istotną w symulacjach mających na celu odwzorowanie warunków rzeczywistych. W rozdziale trzecim przedstawiono zasady działania obu modułów. Istotnym aspektem opisywanego narzędzia jest możliwość wielokrotnego powtarzania eksperymentu w badanym otoczeniu. Rozdział czwarty poświęcono prezentacji wyników działania opracowanego rozwiązania w scenariuszu symulacyjnym.

2. PRZEGLĄD LITERATURY

Generowanie ruchu w symulatorze jest istotnym aspektem z punktu widzenia realizacji badań związanych ze współpracą wielu sieci na ograniczonym obszarze. Voicu et al. [10] zamodelowali środowisko współpracy sieci Wi-Fi i komórkowych w symulatorze ns-3. Wyniki zaprezentowano dla obu technologii pracujących w paśmie 5GHz, wykorzystując metody generowania ruchu w symulatorze. Autorzy w [6] do badań związanych ze współpracą różnych technologii w tym samym paśmie częstotliwości, również wykorzystują generowaniu ruchu. Narzędziem użytym do nasycenia kanałów transmisji Wi-Fi i LTE, była aplikacja iperf. Inną metodą generowania ruchu w ns-3 jest wykorzystanie interfejsu TAP. Pozwala to na udział rzeczywistej maszyny w symulacji. W ten sposób może zostać przekazany ruch z symulacji do fizycznej maszyny, jak również w przeciwnym kierunku. W artykule [9] wykorzystano interfejs TAP do połączenia rzeczywistych maszyn z symulowanym środowiskiem. Rozwiązanie to pozwala wprowadzić ruch do symulowanego urządzenia na poziomie warstwy drugiej. Następnie, ruch można nadawać z użyciem sieci bezprzewodowych, a zatem zgodnie z parametrami zadanymi w symulacji.

3. NARZĘDZIE BTS

Zaproponowane w artykule narzędzie BTS (*Background Traffic Simulator*) składa się z dwóch głównych komponentów. Pierwszym z nich jest sniffer, mający za zadanie zapewnić rzeczywiste dane do symulacji ruchu otoczenia. Drugim jest moduł rozszerzający symulator ns-3 o funkcjonalność symulacji rzeczywistego ruchu na podstawie danych zebranych przez pierwszy komponent. Dane na potrzeby symulacji można uzyskać bezpośrednio, nasłuchując otoczenia lub na podstawie zebranych wcześniej zapisów ruchu.

3.1. Sniffer

Moduł zapisu ruchu w sieciach WLAN, czyli element BTS typu sniffer, służy do rejestrowania ruchu w przestrzeni w celu późniejszego odwzorowania go w symulacji. Pozwala on na przeniesienie rzeczywistych warunków otoczenia do symulacji, w celu prowadzenia badań w warunkach odzwierciedlających realne warunki radiowe w przestrzeni. Rejestrator nasłuchuje i zapisuje ruch na wybranym przez operatora kanale. Wspiera on standardy IEEE 802.11 zgodne z wykorzystywanym interfejsem fizycznym [1]. Oznacza to, że zapis ruchu może być realizowany w technice odpowiedniej dla wykorzystywanej karty sieciowej. Wymaga się, aby interfejs sieciowy oferował tryb monitor. Moduł samoczynnie przełącza kartę sieciową do trybu monitor i po przeprowadzeniu operacji zapisu wraca do trybu zarządzanego. Program realizuje zadania z wykorzystaniem powłoki systemu Linux. Zapisy ruchu prowadzono wykorzystując Ubuntu 18.04 LTS.

Moduł może realizować zadania w dwóch tryb nazwanych odpowiednio:

- live – realizujący zadania na żywo,
- offline – przetwarzający dane z zarejestrowanych plików PCAP.

Każdy tryb przekształca dane wejściowe na plik CSV (ang. comma-separated values) w format odpowiednim dla modułu rozszerzającego symulator NS-3. Układ pliku wejściowego stanowi dziewięć kolumn:

1. id – numer porządkowy ramki,
2. timestamp – w formacie EPOCH podanym w sekundach z dokładnością do mikrosekund,
3. type – typ ramki,
4. subtype – podtyp ramki,
5. dbm – moc odebranego sygnału,
6. size – rozmiar ramki wyrażony w bajtach,
7. l4proto – protokół warstwy czwartej,
8. frequency – częstotliwość na jakiej odebrano ramkę wyrażona w megahercach,
9. rate – szybkość transmisji [Mbit/s].

Zapis na żywo w trybie „live" wymaga podania nazwy trybu jako wartość parametru „mode". Niezbędnym również jest podanie systemowej nazwy wykorzystywanego interfejsu (wartość parametru „dev"), kanału na którym będzie nasłuchiwać karta sieciowa (wartość parametru „ch") oraz podanie czasu rejestrowania (wartość parametru „dur") i nazwy pliku wyjściowego (wartość parametru „ofile"). Jeżeli istnieje w systemie plik o podanej nazwie, jego zawartość zostanie nadpisana. Praca w trybie „live" wymaga uprawnień użytkownika „root".

Przekształcenie danych z formatu PCAP do akceptowanego formatu w pliku CSV, wymaga określenia trybu jako „offline" (wartość parametru „mode"). Pozostałe obowiązkowe parametry to „ifile", przyjmujący ścieżkę do pliku wejściowego oraz „ofile" przyjmujący ścieżkę do pliku stanowiącego wynik przekształcenia. Pozostałe parametry w tym trybie to:

- „dur" przekształca wszystkie ramki od początku pliku do ramki o znaczniku czasowym równym lub największym mniejszym od sumy znacznika czasowego pierwszej zarejestrowanej ramki i wartości parametru dur,
- „start" definiuje wartość znacznika czasowego ramki, od której rozpocznie się przekształcanie zapisu ruchu. W przypadku, gdy wartość ta nie będzie zgodna ze znacznikiem żadnej ramki w pliku wejściowym, przekształcanie rozpocznie się od najmniejszego znacznika czasowego, większego od wartości przekazanej do programu,
- „stop" przyjmuje wartość znacznika czasowego odpowiadającego ostatniej przetwarzanej ramce z pliku PCAP. Jeżeli w pliku nie znajduje się ramka o wskazanym znaczniku czasowy, ostatnim przetworzonym wpisem będzie ramka o znaczniku czasowym o wartości największej, mniejszej od zadanej.

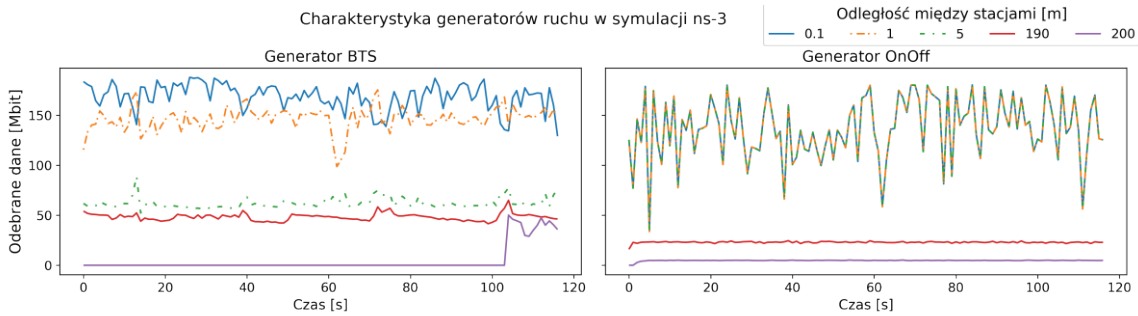
W celu uruchomia programu rejestrującego ruch wymagane jest, aby komputer prowadzący zapis posiadał zainstalowany interpreter języka Python 3.8.10 z zainstalowaną biblioteka Scapy 2.4.5. Jedynym wymaganiem sprzętowym jest posiadanie interfejsu sieciowego wspierającego tryb monitor.

Moduł ten pozwala wyodrębnić z całości rejestrowanego ruchu kluczowe dane dla odwzorowania rzeczywistego obciążenia kanału radiowego w środowisku symulacyjnym. Istotnym jest ograniczenie cech ruchu do minimum, aby optymalnie wykorzystać możliwości symulatora, zachowując zbliżony do realnego poziom nasycenia sieci. Parametry wyodrębnianie i zapisywane w pliku wyjściowym zostały tak dobrane, aby symulator mógł w sposób zbliżony do rzeczywistego reprezentować otocznie. Możliwość wykorzystania powszechnie stosowanych plików PCAP, pozwala odwzorować w symulacji ruch zapisany w dowolnym miejscu i czasie. Tryb rejestrujący ruch w czasie rzeczywistym można wykorzystać na przykład do zapisu ruchu z wykorzystaniem stacji mobilnej.

3.2. Rozszerzenie BTS symulatora ns-3

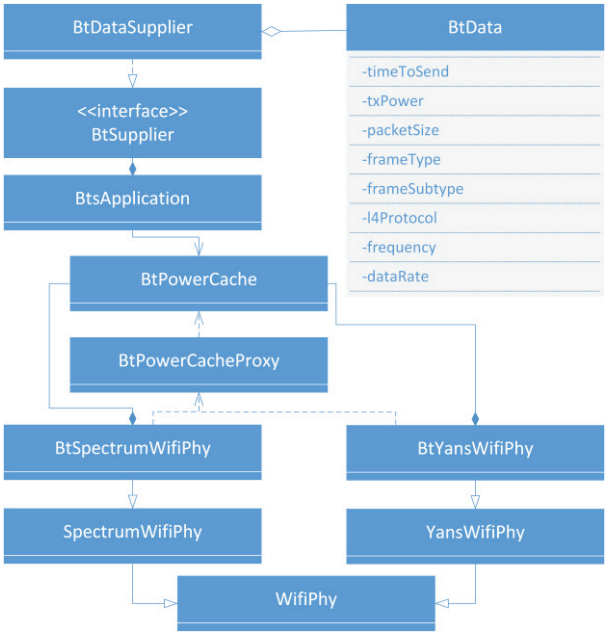
Narzędzie BTS umożliwia przekazanie zdarzeń zaobserwowanych w rzeczywistym ruchu bezprzewodowym do środowiska symulacyjnego ns-3. Dzięki niemu prostsze staje się badanie wpływu nowych lub niedostępnych jeszcze technologii na istniejący system i vice versa. Oprogramowanie jest rozszerzeniem symulatora ns-3 w wersji 3.35. Wprowadzone modyfikacje zostały zaimplementowane w taki sposób, aby uniknąć niezgodności z wcześniej opracowanymi scenariuszami ns-3 oraz kolejnymi wersjami symulatora. Aby umożliwić przekaz zbieranych informacji o mocy nadawanych danych, wymagana była modyfikacja warstwy fizycznej. W rozszerzeniu uwzględnione zostały obie dostępne implementacje warstwy pierwszej modelu OSI/ISO sieci WLAN: Yans i Spectrum.

Podczas prac nad rozszerzenie BTS podjęto decyzję o realizacji rozwiązania w formie odczytu z wcześniej przygotowanego pliku, zamiast przekazywania danych odbieranych na interfejsie sieci bezprzewodowej od razu do symulatora, poprzez interfejs TAP. Decyzja była podyktowana m.in. ograniczeniami wydajnościowymi.



Rys. 1. Charakterystyka generatorów ruchu w symulacji ns-3

Związane jest to z faktem, że ns-3 jest symulatorem zdarzeniowym. Ruch rzeczywisty możliwy jest do przekazywania na bieżąco tylko w przypadku, kiedy scenariusz symulowany jest w czasie równym lub szybszym niż czas rzeczywisty. Metoda ta charakteryzuje się jednak brakiem możliwości odtworzenia realnych warunków radiowych. Co więcej, bez zapisu ruchu lub wykorzystania znanego źródła ruchu jest ona jednorazowa. Wielokrotne odtworzenie w symulacji zdarzeń w przestrzeni, korzystając z wcześniej skonfigurowanego źródła ruchu powoduje, że nieidealne warunki rzeczywiste są kontrolowane, a zatem nie odzwierciedlają otoczenia.



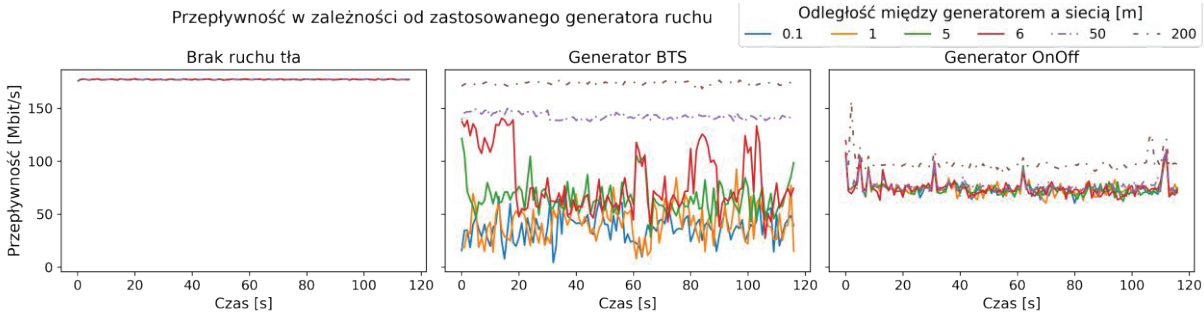
Rys. 2. Diagram UML głównych elementów rozszerzenia BTS symulatora ns-3.

Z punktu widzenia implementacyjnego można wyróżnić trzy istotne zbiory elementów wchodzących w skład narzędzia BTS. Są to źródło informacji o ruchu, zmodyfikowana warstwa fizyczna umożliwiająca manipulację danymi związanymi z ruchem w kanale radiowym oraz element łączący między warstwą aplikacji, a warstwą fizyczną. Szczegółową reprezentację modułu z uwzględnieniem istotnych klas przedstawia Rys. 2. Interfejs BtSupplier definiuje dostęp do danych zbieranych przez oprogramowanie typu sniffer dla klasy BtsApplication. Na podstawie opisu zebranych zdarzeń rzeczywistych reprezentowanych przez klasę BtData możliwe jest dalsze generowanie zdarzeń w symulatorze. Klasa BtPowerCache odpowiedzialna jest za przekazanie informacji o parametrach zdarzenia na podstawie jego identyfikatora do klas reprezentujących warstwę fizyczną. W celu zapewnienia poprawnej obsługi zdarzenia rozszerzone klasy warstw fizycznych wykorzystują klasę BtPowerCacheProxy. Modułowa struktura oraz zastosowanie interfejsów między warstwami pozwala na sprawną implementację rozwiązania BTS na potrzeby symulacji dedykowanych również innym standardom bezprzewodowym.

4. SYMULACJA RUCHU RZECZYWISTEGO W NS-3

Przewagą narzędzia BTS poza przeniesieniem charakterystyki intensywności ruchu rzeczywistego do symulacji jest też równoczesne zapewnienie dodatkowych metadanych opisujących zarejestrowane zdarzenia. Efektem tego jest bardziej realna charakterystyka intensywności w zależności od odległości od symulowanego źródła generowanego ruchu. Zjawisko to przedstawione zostało na Rys. 1. Na potrzeby generatora BTS ruch został zarejestrowany w pomieszczeniu, w którym do routera domowego przewodowo podłączono trzy komputery z włączoną aplikacją iperf 3 w trybie serwera [3].Następnie z każdą z tych maszyn połączono klienta w tej samej sieci, ale korzystającego z łączności bezprzewodowej. Wykorzystano sieć Wi-Fi w standardzie IEEE 802.11ax pracującej z użyciem pasma częstotliwości 2,4 GHz [2], [4], [8]. Punkt dostępowy znajdował się na środku pomieszczenia, a komputer z włączonym modułem zapisu ruchu sieciowego znajdował się po między drugą, a trzecią stacją pracującą w trybie klienta. Każda karta sieciowa w wykorzystywanych komputerach była wyposażona w trzy anteny, a użyty punkt dostępowy posiadał osiem anten. W celu uzyskania powyższych wyników przygotowano scenariusz testowy z wykorzystaniem modelu YANS [5]. Urządzenia sieciowe zamodelowano na podstawie użytych wcześniej urządzeń fizycznych. Do porównania wybrano generator ruchu typu OnOff w konfiguracji 200 Mbit/s zadanej przepływności oraz parametrami czasu aktywności i braku aktywności jako zmiennymi z rozkładu jednorodnego z zakresu odpowiednio [0;1] i [0;0,5].

W celu zaprezentowania wpływu zastosowania sposobu generowania ruchu tła na symulowaną sieć bezprzewodową rozszerzono poprzedni scenariusz o dodatkową parę węzłów, punkt dostępowy i stację, tworzącą sieć testową. Para węzłów skonfigurowana została podobnie jak wcześniej omawiane urządzenia fizyczne i umieszczona



Rys. 3. Przepływność sieci w zależności od zastosowanego generatora ruchu tła.

w odległości 5m od siebie względem osi y. W sieci zamodelowano ruch o stałej intensywności. Kolejne odległości testowej sieci względem generatora ruchu tła na osi x oraz wpływ tej odległości na przepustowość ruchu w sieci testowej przedstawia Rys. 3. Porównanie doskonale oddaje różnice w charakterystyce intensywności w zależności od odległości od symulowanego źródła generowanego ruchu spowodowane m.in. zróżnicowanymi poziomami mocy rejestrowanych ramek, w przeciwieństwie do standardowego profilu ruchu generatora OnOff. W przypadku scenariusza z generatorem BTS chwilowa wydajność sieci wykazuje dużo mniejszą stabilność dla niewielkich odległości, a dla znacznie większych umożliwia uzyskanie limitu przepustowości ze scenariusza kontrolnego. Zamodelowanie tak zróżnicowanego ruchu wykorzystując istniejące narzędzia symulatora ns-3 jest możliwe, jednak w większości wypadków będzie wiązać się ze sporym nakładem pracy względem uzyskiwanych rezultatów.

5. PODSUMOWANIE

Opracowane narzędzia pozwalają przenieść warunki otoczenia do symulacji i badać działanie znanych i nowych metod w zdefiniowanym otoczeniu. Wprowadza ono nową metodę generowania ruchu rozszerzając możliwości symulatora ns-3 o funkcjonalność generowanie ruchu zgodnego z występującym w badanym otoczeniu. Pozwala to testować nowe rozwiązania współużytkowania pasma przez różne technologie, jak również przewidywać problemy wdrażania kolejnych metod łączności bezprzewodowej w miejscu, gdzie pracują już inne systemy. Dotychczas wymagało to zaawansowanego modelowania ruchu w scenariuszach symulacyjnych lub wykorzystania kontrolowanych metod obciążenia infrastruktury sieciowej, takich jak aplikacje służące do transmisji fikcyjnych danych. Narzędzie BTS upraszcza ten proces i pozwala wielokrotnie powtarzać scenariusze testowe w rzeczywistych, wcześniej zarejestrowanych warunkach. Co więcej nie wyklucza użycia posiadanych już zapisów ruchu.

PODZIĘKOWANIA

Praca została sfinansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

LITERATURA

[1] IEEE. 2021. „IEEE Standard for Information Technology-- Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks--Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.” *IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016)*: 1-4379.

[2] IEEE. 2021. „IEEE Standard for Information Technology--Telecommunications and Information Exchange between Systems Local and Metropolitan Area Networks--Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 1: Enhancements for High-Efficiency WLAN.” IEEE Std 802.11ax-2021 (Amendment to IEEE Std 802.11-2020): 1-767.

[3] „iPerf - The ultimate speed test tool for TCP, UDP and SCTP.” https://iperf.fr/

[4] Khorov Evgeny, Anton Kiryanov, Andrey Lyakhov, Giuseppe Bianchi. 2019. „A Tutorial on IEEE 802.11ax High Efficiency WLANs.” *IEEE Communications Surveys & Tutorials* 21 (1): 197-216.

[5] Lacage Mathieu, Thomas R. Henderson. 2006. „Yet Another Network Simulator”. *Proceeding from the 2006 Workshop on Ns-2: The IP Network Simulator* (WNS2 '06): 12. doi: 10.1145/1190455.1190467.

[6] Makris Nikos, Agorastos Dimitrios Samaras, Virgilios Passas, Thanasis Korakis and Leandros Tassiulas. 2017. „Measuring LTE and WiFi coexistence in Unlicensed spectrum.” *European Conference on Networks and Communications (EuCNC)*: 1-6. doi: 10.1109/EuCNC.2017.7980769.

[7] Nsnam „ns-3 Network Simulator.” https://www.nsnam.org/.

[8] Natkaniec Marek, Łukasz Prasnal, Michał Szymakowski. 2019. „Analiza symulacyjna sieci standardu IEEE 802.11ax. ” *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne* 6: 415–419.

[9] Sahu Abhijeet, Ana Goulart, Karen Butler-Purry. 2016. „Modeling AMI network for real-time simulation in NS-3.” *2016 Principles, Systems and Applications of IP Telecommunications (IPTComm)*: 1-8, doi: 10.1109/IPTComm39427.2016.7780248.

[10] Voicu Andra M., Ljiljana Simić, Marina Petrova. 2021. „Modelling Broadband Wireless Technology Coexistence in the Unlicensed Bands.” *2021 IEEE 22nd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*: 129-138.doi:10.1109/WoWMoM51794.2021.00026.

PARADOKS PODŁĄCZONY CHOCIAŻ ODŁĄCZONY: SKRYTY KANAŁ STANDARDU IEEE 802.11 ZAPOBIEGAJĄCY ATAKOM NA SIECI WI-FI

CONNECTED WHILE DISCONNECTED PARADOX: AN IEEE 802.11 COVERT CHANNEL TO PREVENT WI-FI ATTACKS

Geovani Teca[1]; Marek Natkaniec[2];

[1] AGH University of Science and Technology, Cracow, teca@agh.edu.pl
[2] AGH University of Science and Technology, Cracow, natkanie@agh.edu.pl

DOI: 10.15199/59.2022.4.104

Streszczenie: **Rosnące zapotrzebowanie mobilnych urządzeń na nieprzerwaną łączność z Internetem w dowolnym miejscu i czasie sprawiają, że sieci bezprzewodowe standardu IEEE 802.11 stają się jednym z głównych sposobów zaspokojenia tego zapotrzebowania. Sieć Wi-Fi zapewnia łączność z internetem, ale także otwiera drzwi dla zagrożeń i ataków. W tym artykule przedstawiamy koncepcję skrytego kanału, który ma zapobiegać atakom na sieci Wi-Fi, zapewniając jednocześnie niezawodną i bezpieczną komunikację między stacją a punktem dostępowym.**

Abstract: **The growing demand from mobile devices for uninterrupted internet connectivity from everywhere and anytime transform the IEEE 802.11 wireless networks into one of the primary solutions to meet the demand. A Wi-Fi network is a means to provide internet connectivity to the devices, at the same time opening the door to threats and attacks. This paper presents a covert channel to prevent Wi-Fi attacks while assuring reliable and secure communication between the Wi-Fi station and the access point.**

Słowa kluczowe: **IEEE 802.11 sieci bezprzewodowe, Skryty kanał, Steganografia, Wi-Fi ataki.**

Keywords: **Covert channel, IEEE 802.11 wireless networks, Steganography, Wi-Fi attacks.**

1. INTRODUCTION

The overall development and deployment of electronic devices (e.g.: smartphones, laptops, smart-watches, IoT devices) transformed IEEE 802.11 wireless networks [1] into one of the primary means to assure internet connectivity to those devices. IEEE 802.11 is the standard for Wireless Local Area Network (WLAN), well-known as a Wi-Fi network. Wi-Fi network offers many commodities as it is simple to set up, grants users freedom of mobility, and has the possibility to extend its range without the concern about the physical space or length of cables. Those are the main features that led to the fast adoption and popularity of the IEEE 802.11 wireless networks.

The IEEE 802.11 wireless networks in one way provides Internet connectivity to wireless devices, and in the other way, it exposes the devices to threats and attacks, rising security concerns. During the authentication phase, the device exchanges keys, and during the data transmission, the device might send sensitive data, such as credit card or personal identification number. Both authentication keys or sensitive data can be under attack due to the nature of data transmission in IEEE 802.11 networks. The data in the IEEE 802.11 wireless networks is sent through electromagnetic waves propagating over the air, which means that no data transmission occurs in secret, exposing the data to threats and making it vulnerable to attacks [2], [3]. There is a high risk of sensitive data being intercepted, decrypted, and falling into malicious hands.

Among the most popular attacks in IEEE 802.11 network are: Rogue Access Point (RAP), an Access Point (AP) added to the network without consent, acting as an impostor, and impersonates the functions of the original AP that connects a Client Station (STA) to the internet. Packet sniffing, and capture attacks occur when a network device scans the wireless channel and collects the traffic for further analysis in order to find a security breach in the network. Replay attacks occur when a third-party listen to the ongoing communication in the wireless channel, intercepts it, modifies the data content, and replies as if it were the original sender, luring the receiver to engage in a data exchange with this third-party. Additionally, we mention the deauthentication attack, when the attacker clones the Media Access Control Address (MAC address) of the STA or AP and sends deauthentication request frame, forcing the STA to reauthenticate to the network, and during the reauthentication process, the STA might fall into many traps, such as associating with a RAP.

This paper presents the concept and implementation of a covert channel to prevent Wi-Fi attacks, by sending secret data without being connected to the AP. The secret data is inserted in the Most Significant Bit (MSB) of each value that represents the supported rate field of the probe request frame. The secret data is acknowledged by a probe response frame from the AP.

This research paper is structured as follows: Section 2 presents existing measures against IEEE 802.11 wireless networks attacks and existing IEEE 802.11 covert channels that could be deployed to prevent the attacks. Section 3 provides background knowledge regarding IEEE 802.11 wireless networks states and operations, the frame, and the field on which the covert channel is based. Section 4 describes the covert channel concept and operation. The achieved results are presented in section 5. As the last, section 6 is a brief conclusion of this work.

2. RELATED RESEARCHES

2.1. IEEE 802.11 wireless networks attacks

RAP detection by analyzing network traffic characteristics is presented in [4]. The technique consists of analyzing the AP behavior when the end-user generates traffic that demands a reaction from AP. A review of a variety of traffic sniffing attacks and countermeasures in [5] describes the different attacks that the STA could face, after network data traffic is collected, and several strategies against replay attacks are presented in [6]. To prevent deauthentication attack, a mechanism based on hashing the association ID value is described in [7].

2.2. IEEE 802.11 wireless networks covert channels

Covert channels can be considered as one of the mechanisms to prevent IEEE 802.11 wireless networks attacks. They allow sending secret data without previous association with the AP, making the data transfer go undetected through the wireless channel being targeted or monitored. The paper [8] presents a covert channel that is based on the field SSID of the probe request frame. Two methods to send secret data are presented in [9], based on the sequence number and the initialization vector field using regular IEEE 802.11 frames. The paper [10] presents a covert channel based on the 2-bit field protocol version of the IEEE 802.11 frame header.

3. BACKGROUND

In contrast to wired networks, where in most cases, plugging the ethernet cable into the device grants full access to the network, in IEEE 802.11 wireless networks, a device has to exchange information with the AP prior to having full network access. As described in figure 1, in the very first state, the STA is unauthenticated and unassociated (State 1). In this state, STA starts the network discovery process through passive or active scanning. In passive scanning, STA listens to beacon frames from AP that advertise the wireless network's existence. In active scanning, STA sends a probe request frame, searching for the available wireless network and providing its own information. When the probe request is received by the AP, it determines, if the information in the probe request matches the required parameters to associate with the network, and if the information matches, the AP issues a probe response, providing information regarding the wireless network. After discovering the existence of the wireless network, the STA issues an authentication request, providing its credentials. If the credentials match, the AP replies with an authentication response with a code indicating success. The STA passes to the authenticated and unassociated state (State 2). In order to have full network access to the network, STA issues an association request, the AP grants the association assigning an association ID to the connection, then STA passes to the last state: authenticated and associated (State 3).

In IEEE 802.11 wireless networks, the STA is exposed to attacks from the moment it announces its presence in the network. In-state 1, when using active scanning the STA can specify in the probe request the Service Set Identifier (SSID) which is the network name. From that moment the STA draws attention and could be a victim of a RAP attack. If state 1 draws attention to the STA, states 2 and 3 open the door for the intruders to attack, because during the authentication phase, the encryption key is exchanged, and after the association the data exchange takes place. A wireless channel compromised from the beginning could lead to the encryption key being captured, which would allow access to unauthorized devices to the network and the exchanged data between STA and AP can be intercepted and decrypted.

These days, IEEE 802.11 wireless networks have become popular, anyone with a wireless network device, when turning on the Wi-Fi icon, can discover a considerable number of Wi-Fi networks to connect to. Few networks are open and do not require a password, this implies that no data encryption is guaranteed and even for the networks that require a password, there is no total certainty regarding the legitimacy of the AP with which the STA establishes the connection. Indeed, many users connect to a Wi-Fi network without even seeing where the AP is or knowing who brought the Wi-Fi network into existence, putting themselves in dangerous situations. Data security and privacy are the main concerns to be addressed when establishing a Wi-Fi connection in an untrusted environment. Turns out that network steganography addresses the data security and privacy issues in the Wi-Fi network. Steganography is a mechanism for concealing a secret message within another explicit message. Network steganography is achieved using covert channels. In order to prevent data from being exposed to attacks in an untrusted environment, STA abstains from sending explicit data frames, instead, it sends the data as a secret message inside a regular non-data frame, without raising the awareness of the secret data exchange between STA and AP is taking place.

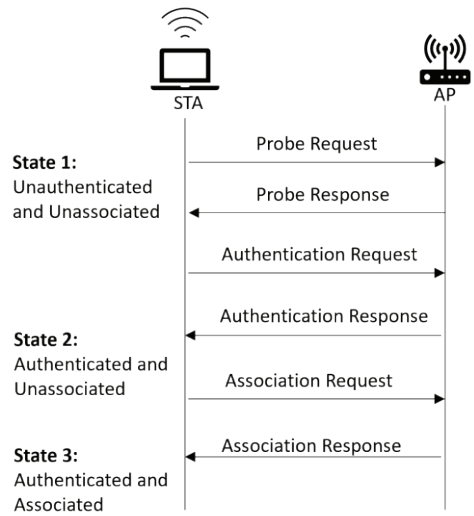


Fig. 1. IEEE 802.11 wireless networks state machine

As a regular non-data frame, the probe request is the frame broadcasted by STA to discover available wireless networks in the vicinities. The IEEE 802.11 probe request

frame structure is presented in figure 2, the frame is a typical IEEE 802.11 header, and the body contains the following fields such as the SSID to indicate the network name STA is searching (Empty SSID means, all available networks), field supported rate to indicate data rate which STA supports in Mbps.

In the supported rate field the STA can insert up to 8 values, each value represents the data rate and is 8 bits longer. The bits 0 through 6 represent the data rate value in decimal, the bit 7 (MSB) is to indicate whether the supported data rate is mandatory (set to 1) or optional (set to 0). When AP receives the probe request, among the parameters, that it takes into consideration are the supported data rates. In order to receive a probe response at least one mandatory data rate value in the probe request shall be the same mandatory data rate in the AP.

```
> IEEE 802.11 Probe Request, Flags: ........C
v IEEE 802.11 Wireless Management
  v Tagged parameters (78 bytes)
    > Tag: SSID parameter set: Krakow-Wifi
    > Tag: Supported Rates 6, 9, 12, 18, 24, 36, 48, 54, [Mbit/sec]
    > Tag: Extended Supported Rates Unknown Rate, [Mbit/sec]
    > Tag: Extended Capabilities (8 octets)
    > Tag: HT Capabilities (802.11n D1.10)
    > Tag: VHT Capabilities
```

Fig. 2. Probe request frame format with supported rate field highlighted

The possibility to specify whether a data rate value is mandatory or optional using the MSB, opens a window to create a covert channel that conceals the data into a secret message using the MSB of each data rate value. Having the covert channel created in an untrusted environment, the STA remains in state 1 and sends secret data to the AP with no concerns regarding data security and privacy, avoiding the threats and attacks that occur in states 2 and 3.

4. COVERT CHANNEL CONCEPT

We observe that the most popular Wi-Fi attacks take place when STA is in state 2 or 3. It is the attempt to connect to the network that exposes it to threats and makes it vulnerable to attacks. The concept presented in this paper is a covert channel that allows STA to send secret data to AP without performing authentication or association in untrusted environments. The secret message is inserted in the MSB of the values in the supported rate field carried in probe request frames while scanning the network. For each probe request, STA sets to 1 or 0 the MSB of the data rate value in the supported rate field. Each data rate value allows transmitting one bit of covert information. The scheme is presented in figure 3, demonstrating the scheme to encode a secret message as sequence **10101000**. STA sends a probe request setting the 6, 12, and 24 Mbps as mandatory data rate and the remainder data rate as optional.

The covert channel allows transmitting up to 8 bits in a single transmission, therefore before the opening of the covert channel, STA and AP must prove each other authenticity, which will allow eliminating the suspicion of the existence of an impostor from both sides. For that purpose, a 4-bit Cyclic Redundancy Check (CRC-4) is used, and STA and AP share three polynomial generators, the opening polynomial generator used to signal the opening, the data polynomial generator one to secret data, and the terminating polynomial generator to signal the covert channel termination. To signal the start of transmission, STA takes a 4-bit sequence and encodes it using the opening polynomial generator, producing an 8-bit sequence. Each bit from the sequence is then inserted in the MSB of each data rate value and sent in the probe request. The AP receives the probe request and decodes using the same starting polynomial and sends back a probe response, signaling that it is also ready to start the communication.

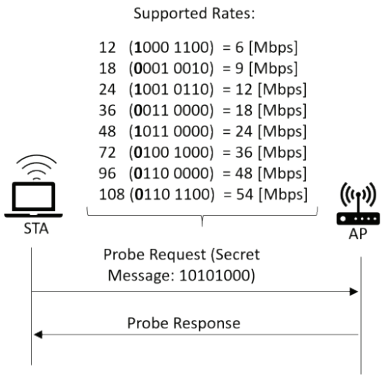


*Fig. 3. Pratical example of sending a covert message **10101000** using probe request*

After receiving the probe response that confirms that the covert channel is created, the STA sends the first covert message, as presented in figure 4, the STA after sending each covert message, schedules two periodic events. The first event is to schedule the next transmission and the second is to schedule the verification if a probe response has been received. If STA detects that the probe response has not been received, it retransmits right away the probe request.

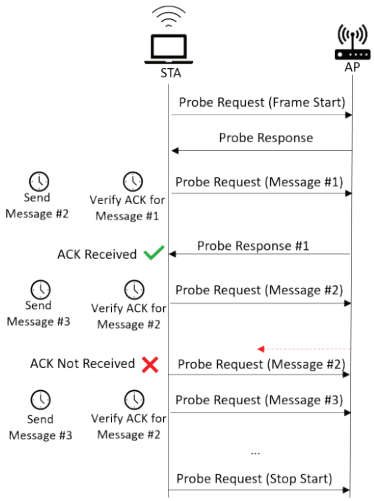


Fig. 4. Covert channel concept using probe request frame

In order to terminate the covert channel, the STA performs the same procedure as in opening the covert channel, but using the terminating polynomial, the AP decodes the secret message using the terminating polynomial. In general, the AP takes each MSB from the data rates in the received probe request, forms the 8-bit sequence, runs the CRC against the three polynomial generators to find a match, and determines whether the probe request signals opening, data, or terminating the covert channel.

The presented covert channel finds its practical use, for data exchanged between STA and AP in untrusted or hostile environments, where the users avoid being exposed to threats or attacks.

5. PERFORMANCE ANALYSIS

5.1. Simulation scenarios

The simulation environment was created using the ns-3 discrete-event network simulator [11]. The network was IEEE 802.11ac with a 20MHz channel with one AP and wireless STAs connected to the same AP. All the nodes were in constant position during the entire simulation. There was only one STA sending covert data (sending probe request all the time) and the remaining STAs generated UDP traffic in saturated conditions. To evaluate the performance of the covert channel, we have created different scenarios with different parameters. The parameters we intended to analyse the covert channel performance are:

- Throughput as the number bits per second sent and successfully acknowledged by the AP.
- Covert channel efficiency as the percentage of successful received frames over transmitted.
- Impact of retransmission mechanism, as a measure of how much the retransmission contributed to the covert channel efficiency.

In the first scenario, there is only one AP and one STA exchanging probe request and response. In the second scenario, we analyse the impact of external traffic on the covert channel by adding 5, 25 and 50 more STAs respectively, generating UDP traffic in saturation conditions. The variable parameter is the probe request interval, which indicates how often the STA generates the covert message in form of probe request. The simulation time was 300 seconds.

5.2. Simulation results

We observed that the covert channel achieves its maximum capacity which is 802 Mbps in isolated scenarios, only having the covert STA and the AP. However, adding four more regular Wi-Fi devices generating UDP traffic still has no big impact on the covert channel performance, as the results are slightly closed. The covert channels suffer degradation as observed when a network transits from small (having five stations) to a medium having 25 stations. The throughput degrades considerably down to more than 50% percent, and the registered throughput was 293 Mbps. Increasing the number of stations to 50 also had only a slight degradation in the covert channel throughput.

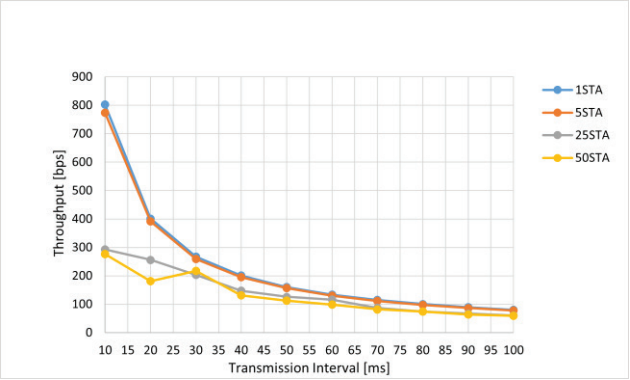


Fig. 5. Covert channel throughput analysis

Figure 6, presents the covert channel efficiency over the transmission interval. In isolated scenarios with only the covert STA and the AP, it was registered 100% efficiency as there was no external traffic to interfere or cause a collision with the covert STA. The same stable efficiency is observed with a network with four more additional stations with an efficiency of around 96%. We also observe that for the medium and larger network (25 and 50 stations respectively), shorter transmission interval leads to more collisions, as consequence the efficiency decreases considerably. The efficiency improves with a larger transmission interval, as the covert STA waits longer to transmit and has more chances to get a response.

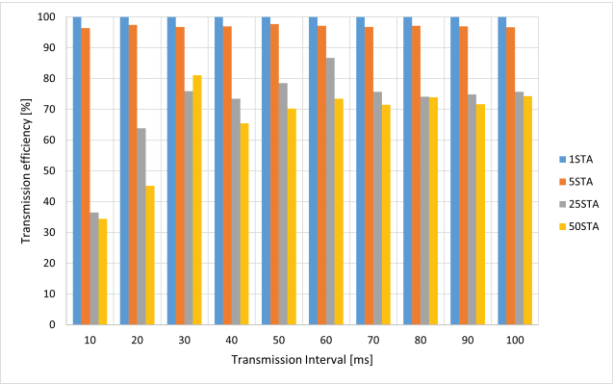


Fig. 6. Covert channel efficiency

The presented covert channel is reliable, providing retransmission in case of failure to receive the probe response and retransmission is a mechanism to improve the covert channel throughput and efficiency. In figure 7, which network size caused more frames to be retransmitted. For isolation, the scenario has previously presented with 100% of efficiency it registered no retransmission, for a small network with four more regular stations about 30% of the frames are retransmitted. The retransmission rate increases as the number of stations in the network also increase with shorter transmission interval.

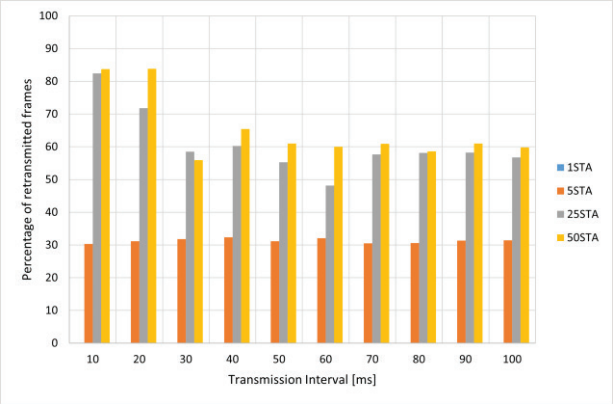


Fig. 7. Effect of retransmition on the covert channel perfomance

6. CONCLUSION

In this paper, we presented the current threats and attacks against IEEE 802.11 wireless networks. We described the popular attacks and threats and the developed countermeasure. We also proposed a covert channel to prevent the attacks avoiding the STA to going to unsaved zone (states 2 and 3) simultaneously allowing STA to send data to AP. The covert channel encodes the secret data in the MSB bit of supported rates values in the probe request with retransmission capabilities. The presented results proved that the covert channel performance stands out providing high values of throughput when compared with existing covert channels, even in dense network.

ACKNOWLEDGEMENTS

This work was supported by the Polish Ministry of Science and Higher Education with the subvention funds of the Faculty of Computer Science, Electronics and Telecommunications of AGH University.

REFERENCES

[1] IEEE Standard for Information Technology. 2021. "Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks--Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE Std 802.11-2020". *IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016)*: 1-4379.

[2] Umesh Kumar, Sapna Gambhir. 2014. "A Literature Review of Security Threats to Wireless Networks". *International Journal of Future Generation Communication and Networking*, 7 (4): 25-34.

[3] Noor Mardiana, Hassan Wan. 2013. "Wireless Networks: Developments, Threats and Countermeasures." *International Journal of Digital Information and Wireless Communications* 3: 125-140.

[4] Shetty Sachin, Song Min and Ma Liran. 2007. "Rogue Access Point Detection by Analyzing Network Traffic Characteristics." *MILCOM 2007 - IEEE Military Communications Conference*: 1-7.

[5] Prabadevi. Nagamalai Jeyanthi. 2018. "A Review on Various Sniffing Attacks and its Mitigation Techniques". *Indonesian Journal of Electrical Engineering and Computer Science,* 12 (3): 1117-1125.

[6] Aura Tuomas. 1997. "Strategies against Replay Attacks". *Proceedings 10th Computer Security Foundations Workshop:* 59-68.

[7] Arora Ananay. 2018. "Preventing wireless deauthentication attacks over 802.11 Networks".

[8] Sawicki Krzysztof. 2017. "Two-way complex steganographic system for authentication and authorization in IEEE 802.11 wireless networks". *ELEKTRONIKA - KONSTRUKCJE, TECHNOLOGIE, ZASTOSOWANIA*, 1: 24-28.

[9] Frikha Lilia, Trabelsi Trabelsi. El-Hajj Wassim. 2008. "Implementation of a Covert Channel in the 802.11 Header". *International Wireless Communications and Mobile Computing Conference*: 594-599.

[10] Goncalves Ricardo. Tummala Murali. Mceachen John. 2012. "Analysis of a MAC Layer Covert Channel in 802.11 Networks". *International Journal on Advances in Telecommunications,* 5 (3 & 4): 131-140.

[11] https://www.nsnam.org

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

ASYNCHRONICZNA METODA JEDNOCZESNEJ ESTYMACJI POŁOŻENIA I ORIENTACJI OBIEKTU ZA POMOCĄ DWÓCH NADAJNIKÓW

ASYNCHRONOUS METHOD Of SIMULTANEOUS OBJECT POSITION AND ORIENTATION ESTIMATION WITH TWO TRANSMITTERS

Jacek Stefański[1]; Jarosław Sadowski[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jstef@eti.pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jarsad@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.105

Streszczenie: **W artykule opisano nową metodę lokalizowania obiektów dla szerokiej gamy zastosowań, w tym Internetu Rzeczy. Zaproponowana metoda umożliwia estymację położenia i orientacji obiektu na płaszczyźnie lub w przestrzeni, również będącego w spoczynku, za pomocą sygnałów lokalizacyjnych wysyłanych jednocześnie z dwóch nadajników umieszczonych na obiekcie w znanej odległości od siebie. Przedstawiono matematyczną analizę metody oraz wyniki badań symulacyjnych efektywności pracy zaproponowanej metody.**

Abstract: **The paper proposes a new method of locating objects for all kinds of applications, including the Internet of Things. The proposed method enables the estimation of the position and orientation of an object on a plane or in space, also at rest, by means of location signals sent simultaneously from two transmitters placed on the object at a known distance from each other. The mathematical analysis of the method and the results of simulation tests of the effectiveness of the proposed method are presented.**

Słowa kluczowe: **bezprzewodowe sieci sensorowe, radiolokalizacja, radionawigacja.**

Keywords: **radiolocalization, radio navigation, wireless sensor networks.**

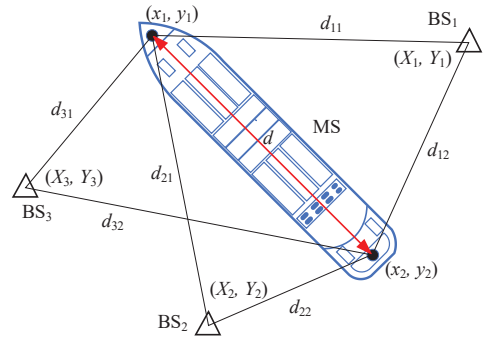
1. WSTĘP

Większość opisywanych w literaturze przedmiotu systemów przeznaczonych do estymacji położenia obiektów opiera się na synchronizacji czasowej pomiędzy elementami systemu, m.in. synchroniczna emisja sygnałów ze stacji bazowych lub synchroniczna detekcja sygnałów odbieranych z węzłów ruchomych. Niewiele ośrodków na świecie zajmuje lub zajmowało się asynchronicznymi metodami radiolokalizacyjnymi. Niniejszy artykuł stanowi propozycję wkładu do tej tematyki, w którym przedstawiono innowacyjną metodę radiolokalizacyjną polegającą na jednoczesnej transmisji sygnałów z dwóch nadajników umieszczonych na lokalizowanym obiekcie w znanej odległości od siebie (np. końce skrzydeł samolotu, dziób i rufa statku). Sygnały te są odbierane przez kilka odbiorników referencyjnych (stacji bazowych) pracujących asynchronicznie. W proponowanym rozwiązaniu współrzędne obu nadajników są estymowane jednocześnie, co prowadzi do jednoczesnego określenia położenia i orientacji lokalizowanego obiektu.

Artykuł jest zorganizowany w następujący sposób: w rozdziale 2 przedstawiono opis matematyczny zaproponowanej metody; w rozdziale 3 został opisany algorytm Newtona (NA) zastosowany do rozwiązywania układu nieliniowych równań pozycyjnych dla opisanej metody; z kolei w rozdziale 4 przedstawiono wyniki badań symulacyjnych efektywności pracy zaproponowanej metody radiolokalizacyjnej; wreszcie ostatnia sekcja stanowi podsumowanie zrealizowanych prac.

2. OPIS ZAPROPONOWANEJ METODY

Rozważmy przypadek dwuwymiarowy, w którym dwa zsynchronizowane nadajniki są umieszczone na lokalizowanym obiekcie (stacji ruchomej). Nadajniki znajdują się w znanej odległości d od siebie i jednocześnie emitują sygnały lokalizacyjne. Sygnały te są odbierane przez stacje bazowe (referencyjne) (BS). Stosując unikatowe identyfikatory w każdym nadajniku, poszczególne stacje bazowe mogą niezależnie odbierać te sygnały i określać różnicę odległości pomiędzy nadajnikami i antenami odbiorczymi [2]. Zaproponowany sposób lokalizowania obiektów opiera się na odbiorze sygnałów nadawanych przez lokalizowaną stację ruchomą (MS). Przykładowa struktura sieci radiowej, w której może być zaimplementowana ta metoda, została pokazana na rys. 1.



Rys. 1. Przykład struktury sieci radiowej, w której można zaimplementować zaproponowaną metodę

Na podstawie rys. 1 można zapisać układ równań nieliniowych, w którym współrzędne (x_1,y_1) oraz (x_2,y_2) są niewiadomymi:

$$\begin{cases} d_{n2} - d_{n1} = \Delta t_n \cdot v = \Delta d_n \quad n = 1,\ldots,N \\ d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 \end{cases}, \tag{1}$$

gdzie Δt_n stanowią różnicę czasów odbioru w poszczególnych stacjach bazowych sygnałów emitowanych z obu nadajników MS, v reprezentuje prędkość propagacji sygnałów radiowych w ośrodku propagacji (w przybliżeniu można założyć, że v jest równe prędkości światła w próżni $c \approx 3 \cdot 10^8$ m/s), a N to liczba stacji bazowych (w naszym przypadku $N = 3$, co jest minimalną liczbą odbiorników referencyjnych dla przypadku dwuwymiarowego). Ponadto d_{n1} i d_{n2} można opisać następującymi zależnościami:

$$\begin{cases} d_{n1} = \sqrt{\left(X_n - x_1\right)^2 + \left(Y_n - y_1\right)^2} \\ d_{n2} = \sqrt{\left(X_n - x_2\right)^2 + \left(Y_n - y_2\right)^2} \end{cases}, \tag{2}$$

gdzie (X_n,Y_n) są współrzędnymi stacji bazowych. Należy podkreślić, że współrzędne obu nadajników: (x_1,y_1) i (x_2,y_2) są estymowane jednocześnie, ale współrzędne te nie są niezależne, ponieważ odległość pomiędzy nadajnikami jest stała. Zatem dwa zbiory współrzędnych dwóch punktów jednego obiektu dają jednocześnie położenie tego obiektu, przez co można rozumieć położenie pierwszego nadajnika, drugiego nadajnika lub dowolną relację geometryczną pomiędzy współrzędnymi nadajników a śledzonym obiektem (np. położenie w środku pomiędzy dwoma nadajniki) oraz orientację obiektu w stosunku do zestawu odbiorników stacji referencyjnych. Zaletą zaproponowanej metody jest możliwość estymacji orientacji zarówno obiektów stacjonarnych, jak i ruchomych, natomiast metody oparte na pojedynczych urządzeniach umieszczanych na lokalizowanych obiektach mogą estymować jedynie kierunek ruchu, ale nie są w stanie określić orientacji obiektów nieruchomych.

Warto również zauważyć, że skoro stacje referencyjne mierzą tylko różnicę czasów dotarcia dwóch sygnałów, emitowanych w tym samym momencie, to ich działanie nie musi być synchronizowane z pracą pozostałych stacji bazowych i stąd asynchroniczna praca systemu.

Szczegółową analizę przykładowego algorytmu zastosowanego do rozwiązania układu równań (1), a także uzyskane wyniki zawarto w kolejnych rozdziałach.

3. OBLICZANIE POŁOŻENIA OBIEKTU

Układ równań (1) można wykorzystać do estymacji położenia stacji ruchomej, a dokładniej współrzędnych obu anten nadawczych (x_1,y_1) oraz (x_2,y_2), które określają jednocześnie położenie i orientację obiektu ruchomego, zakładając, że znane są:

- współrzędne stacji bazowych,
- różnice odległości Δd_n pomiędzy dwoma nadajnikami i poszczególnymi stacjami bazowymi,
- odległość d między nadajnikami.

W literaturze trudno znaleźć bezpośrednie algorytmy prowadzące do rozwiązania układów równań nieliniowych opisanych zależnością (1). Na potrzeby niniejszej pracy zastosowano uogólnioną wersję algorytmu Newtona (NA) [1]. W tym miejscu należy wyraźnie podkreślić, że do rozwiązania układu równań (1) zastosowano tylko jeden algorytm, ponieważ głównym celem tego artykułu jest przedstawienie nowej metody lokalizacji, a nie optymalizacja metod rozwiązywania tego układu równań.

Zakładając najprostszy przypadek dwuwymiarowy i $N = 3$, problem estymacji położenia MS w prezentowanej metodzie sprowadza się do rozwiązania układu równań nieliniowych postaci

$$\mathbf{f}(x_1, y_1, x_2, y_2) = \begin{cases} d_{12} - d_{11} = \Delta d_1 \\ d_{22} - d_{21} = \Delta d_2 \\ d_{32} - d_{31} = \Delta d_3 \\ d^2 = \left(x_2 - x_1\right)^2 + \left(y_2 - y_1\right)^2 \end{cases}, \tag{3}$$

gdzie niewiadome to współrzędne nadajników (x_1,y_1) i (x_2,y_2) w stacji ruchomej. Zlinearyzowaną wersję powyższego układu równań można zapisać jako

$$\mathbf{J}(x_1, y_1, x_2, y_2) \cdot \mathbf{h} = -\mathbf{f}(x_1, y_1, x_2, y_2), \tag{4}$$

gdzie $\mathbf{J}(x_1,y_1,x_2,y_2)$ jest macierzą Jakobiego, opisaną zależnością (5), a wektor $\mathbf{h} = [h_{x_1}, h_{y_1}, h_{x_2}, h_{y_2}]^T$ reprezentuje poprawkę współrzędnych nadajnika w kolejnych iteracjach,

$$\mathbf{J}(x_1, y_1, x_2, y_2) = \begin{bmatrix} \frac{x_1 - X_1}{d_{11}} & \frac{y_1 - Y_1}{d_{11}} & \frac{X_1 - x_2}{d_{12}} & \frac{Y_1 - y_2}{d_{12}} \\ \frac{x_1 - X_2}{d_{21}} & \frac{y_1 - Y_2}{d_{21}} & \frac{X_2 - x_2}{d_{22}} & \frac{Y_2 - y_2}{d_{22}} \\ \frac{x_1 - X_3}{d_{31}} & \frac{y_1 - Y_3}{d_{31}} & \frac{X_3 - x_2}{d_{32}} & \frac{Y_3 - y_2}{d_{32}} \\ \frac{x_2 - x_1}{d} & \frac{y_2 - y_1}{d} & \frac{x_1 - x_2}{d} & \frac{y_1 - y_2}{d} \end{bmatrix}. \tag{5}$$

Warto w tym miejscu podkreślić, że liczba równań w (1), a więc i liczba wierszy w (5) będzie większa dla przypadków, w których $N > 3$.

Algorytm służący do estymacji położenia MS został zaimplementowany w uniwersalnym środowisku obliczeń matematycznych MATLAB.

4. WYNIKI BADAŃ SYMULACYJNYCH

Jakość estymacji położenia obiektów z zastosowaniem opracowanej metody radiolokalizacyjnej oceniono na podstawie wyników badań symulacyjnych. Podczas badań symulacyjnych przyjęto następujący model: stacje bazowe zostały rozmieszczone równomiernie na okręgu o promieniu R; wewnątrz okręgu są generowane losowo współrzędne pierwszego nadajnika na obiekcie; następnie drugi nadajnik umieszczany jest na losowo wykreślonym azymucie w zadanej odległości d od położenia pierwszego nadajnika. Za pomocą algorytmu Newtona (NA) estymuje się współrzędne obu nadajników na obiekcie. W procesie rozwiązywania układu równań (1) dla każdego rozpatrywanego przypadku przyjęto wektor początkowy o współrzędnych początkowych i końcowych odpowiednio (0,0) i $(d,0)$. Odpowiada to sytuacji, w której nie jest znane przybliżone położenie lub orientacja śledzonego obiektu (np. z poprzednich pomiarów). Założono również, że odległość pomiędzy nadajnikami umieszczonymi na MS jest równa $d = 0{,}1R$.

Narzędzie symulacyjne opracowano w środowisku MATLAB. Podczas badań symulacyjnych uwzględniono błąd pomiaru różnicy odległości Δd_n pomiędzy dwoma

nadajnikami i stacjami bazowymi o rozkładzie normalnym (δ_d). Do modelowania tego błędu użyto funkcji *randn* w następujący sposób

$$\delta_d = \sigma \cdot randn\,, \tag{6}$$

gdzie σ jest odchyleniem standardowym pomiarów różnicy odległości. Podczas badania symulacyjnego każdy przypadek powtórzono 10 000 razy [3].

Na podstawie wyników uzyskanych z badań symulacyjnych wykreślono dystrybuantę CDF (*Cumulative Distribution Functions*) błędu bezwzględnego δ, opisanego zależnością (7)

$$\delta = \sqrt{\left(\hat{x} - x\right)^2 + \left(\hat{y} - y\right)^2}\,, \tag{7}$$

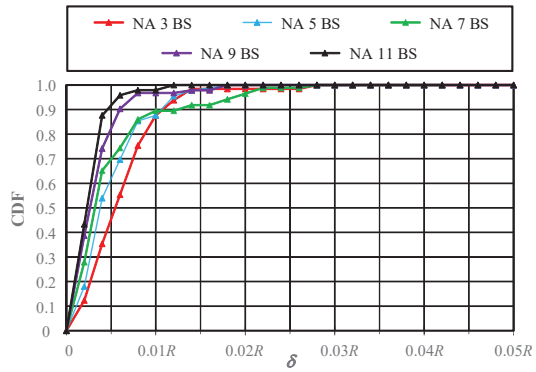
gdzie ($\hat{x}$, $\hat{y}$) reprezentują estymaty współrzędnych MS, natomiast (x,y) to rzeczywiste współrzędne MS, przy czym założono, że

$$\hat{x} = \frac{\hat{x}_1 + \hat{x}_2}{2}, \quad \hat{y} = \frac{\hat{y}_1 + \hat{y}_2}{2}\,, \tag{8}$$

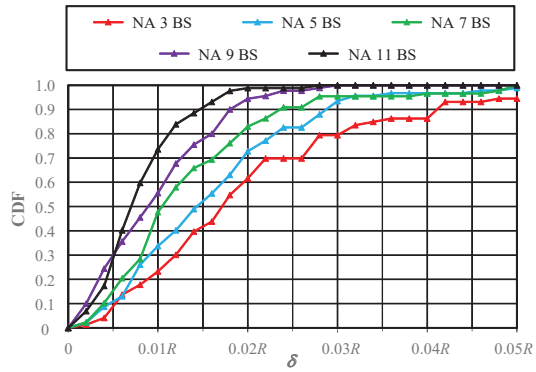
$$x = \frac{x_1 + x_2}{2}, \quad y = \frac{y_1 + y_2}{2}\,. \tag{9}$$

Jak wiadomo, NA jest algorytmem iteracyjnym, który jest wrażliwy na postać wektora startowego do pierwszej iteracji. Im bliższy poprawnemu rozwiązaniu jest wektor rozwiązań początkowych, tym szybciej algorytm znajduje to rozwiązanie. W naszych rozważaniach wektor początkowy był zawsze taki sam, co oznacza stałą pozycję w punkcie początkowym i stałą orientację. W badaniach symulacyjnych przyjęto kryterium zbieżności w postaci normy opisanej zależnością $\|(\mathbf{J}^T \cdot \mathbf{J})^{-1} \cdot \mathbf{J}^T \cdot \mathbf{f}\|$. Maksymalna liczba iteracji wynosiła 1 000. Jeżeli w trakcie działania algorytmu po 1 000 iteracji wartość normy nie spadła poniżej przyjętego progu, program symulacyjny przerywano i wybierano losową postać wektora startowego.

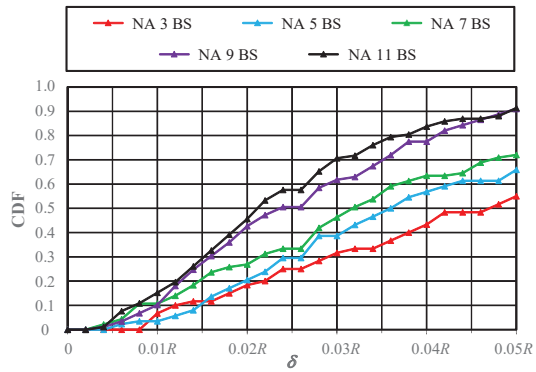
Dla przyjętej topologii sieci sensorowej, składającej się z różnej liczby stacji bazowych (od 3 do 11) analizowano dokładność estymacji położenia MS z zastosowaniem algorytmu Newtona i trzech wartości odchyleń standardowych σ błędów pomiaru różnic odległości. W celu uogólnienia wniosków otrzymane wyniki wyrażano w promieniu R badanego obszaru. Wyniki badań symulacyjnych przedstawiono na rysunkach od rys. 2 do rys. 4. Uzyskane wyniki są zgodne z oczekiwaniami. Na przykład, dla $\sigma = \sqrt{10^{-7}}R$ i dla 7 stacji bazowych w 90 % przypadków błąd bezwzględny δ nie przekraczał 0,01R, co dla promienia R = 1 000 m przekłada się na błąd estymacji położenia MS nie gorszy niż 10 m. Dla tej samej konfiguracji sieci stacji bazowych dla $\sigma = \sqrt{10^{-6}}R$ = 0,001R, błąd δ w 90 % przypadków nie jest gorszy niż 25 m. Jak można było oczekiwać, najgorsze wyniki uzyskano w przypadku, gdy odchylenie standardowe pomiarów różnicy odległości było największe z rozważanych (rys. 4). Przywołując powyższy przykład, w którym R = 1 000 m, tylko dla 11 stacji bazowych, bezwzględny błąd położenia obiektu w 90 % przypadków nie jest gorszy niż 50 m.



Rys. 2. Dystrybuanta (CDF) błędu bezwzględnego położenia obiektu δ z zastosowaniem algorytmu Newtona (NA). Założono, że odchylenie standardowe błędów pomiarów różnicy odległości $\sigma = \sqrt{10^{-7}}R$.



Rys. 3. Dystrybuanta (CDF) błędu bezwzględnego położenia obiektu δ z zastosowaniem algorytmu Newtona (NA). Założono, że odchylenie standardowe błędów pomiarów różnicy odległości $\sigma = \sqrt{10^{-6}}R$ = 0,001R.

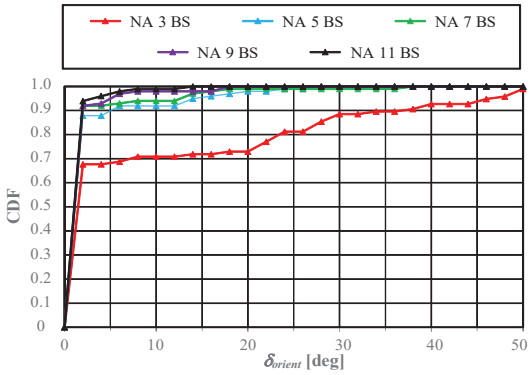


Rys. 4. Dystrybuanta (CDF) błędu bezwzględnego położenia obiektu δ z zastosowaniem algorytmu Newtona (NA). Założono, że odchylenie standardowe błędów pomiarów różnicy odległości $\sigma = \sqrt{10^{-5}}R$.

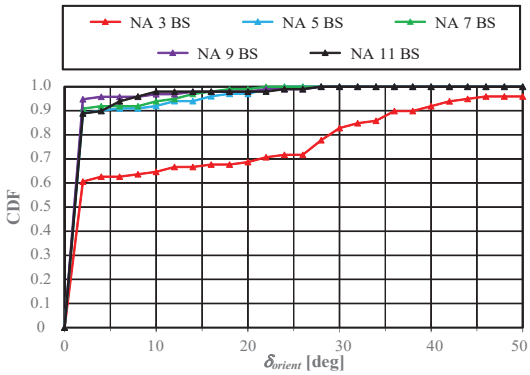
Niewątpliwą zaletą opracowanej metody jest możliwość jednoczesnego oszacowania położenia obiektu i jego orientacji. Stosując opisany powyżej model symulacyjny i przyjęte założenia przeprowadzono badania, w wyniku których uzyskano rodzinę krzywych CDF błędu orientacji δ_{orient} określoną zgodnie z zależnością

$$\delta_{orient} = \left| \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right) - \arctan\left(\frac{\hat{y}_2 - \hat{y}_1}{\hat{x}_2 - \hat{x}_1}\right) \right|. \quad (10)$$

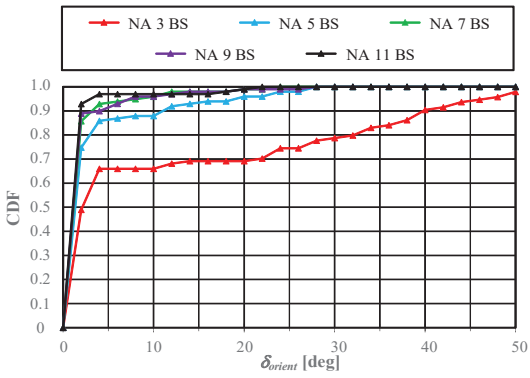
Wyniki badań symulacyjnych przedstawiono na rysunkach od rys. 5 do rys. 7. Z analizy uzyskanych wyników dotyczących błędu orientacji dla wszystkich rozpatrywanych przypadków należy zauważyć, że większość wyników ogranicza się do błędu poniżej 10 stopni.



Rys. 5. Dystrybuanta (CDF) błędu orientacji δ_{orient} z zastosowaniem algorytmu Newtona (NA). Założono, że odchylenie standardowe błędów pomiarów różnicy odległości $\sigma = \sqrt{10^{-7}}R$.



Rys. 6. Dystrybuanta (CDF) błędu orientacji δ_{orient} z zastosowaniem algorytmu Newtona (NA). Założono, że odchylenie standardowe błędów pomiarów różnicy odległości $\sigma = \sqrt{10^{-6}}R = 0{,}001R$.



Rys. 7. Dystrybuanta (CDF) błędu orientacji δ_{orient} z zastosowaniem algorytmu Newtona (NA). Założono, że odchylenie standardowe błędów pomiarów różnicy odległości $\sigma = \sqrt{10^{-5}}R$.

W krótkim podsumowaniu wyników badań symulacyjnych można stwierdzić, że opracowana nowa metoda lokalizacji obiektów na podstawie dwóch jednocześnie emitowanych sygnałów ma duży potencjał aplikacyjny. Przewidywaną dokładność estymacji położenia i orientacji można oszacować na podstawie wyników przeprowadzonych symulacji poprzez odpowiednie skalowanie błędów pomiarowych i obszaru działania systemu.

5. PODSUMOWANIE

W artykule przedstawiono nową asynchroniczną metodę radiolokalizacyjną. Nowością w tej metodzie jest jednoczesna transmisja sygnałów lokalizacyjnych przez dwa nadajniki umieszczone w znanej odległości od siebie na lokalizowanym obiekcie. Analiza matematyczna zaproponowanej metody została przeprowadzona z uwzględnieniem szczegółów implementacji algorytmu Newtona do rozwiązywania układu równań pozycyjnych.

Przedstawiono wyniki przeprowadzonych badań symulacyjnych zaproponowanej metody dla trzech przypadków (dla trzech różnych wartości wariancji błędów pomiarowych). Wykreślono trzy zestawy charakterystyk CDF w funkcji błędu bezwzględnego i błędu orientacji. Otrzymane wyniki wskazują, że zaproponowana metoda, opisana układem równań nieliniowych, prowadzi do uzyskania rzeczywistego położenia i orientacji lokalizowanego obiektu.

Zaproponowana metoda może mieć szerokie zastosowanie, zwłaszcza do lokalizowania obiektów o dużych gabarytach, takich jak statki i samoloty. Istotnym sektorem rynku mogą być również mobilne platformy autonomiczne, bezzałogowe lub wspomagane przez pilota, zarówno naziemne, morskie, jak i powietrzne. Przewagą zaproponowanego rozwiązania nad obecnie stosowanymi systemami nawigacyjnymi i lokalizacyjnymi, zarówno powietrznymi, jak i morskimi, jest pozyskiwanie nie tylko informacji lokalizacyjnych, ale również informacji o orientacji nadzorowanego obiektu. Takie rozwiązania nie są obecnie stosowane w praktyce. Na przykład orientacja statków w systemie automatycznej identyfikacji AIS (*Automatic Identification System*) jest zwykle pobierana z pokładowego żyrokompasu, a nie określana autonomicznie przez system lokalizacyjny. Jednak faktyczne zainteresowanie zaproponowaną techniką jednoczesnego określania położenia i orientacji obiektu będzie w dużej mierze zależeć od pracy i decyzji jednostek administracyjnych odpowiedzialnych za przepisy dotyczące bezpieczeństwa realizacji operacji (powietrznych lub morskich) lub określenie minimalnego wyposażenia pojazdów/platform niezbędnego do dopuszczenia do ruchu.

LITERATURA

[1] Foy Wade H. 1976. "Position-Location Solutions by Taylor-Series Estimation", *IEEE Transactions on Aerospace and Electronic Systems AES-12* (2): 187-194.

[2] Sadowski Jarosław, Stefański Jacek. 2021. „Sposób jednoczesnego wyznaczania położenia i orientacji obiektu, zwłaszcza będącego w ruchu". *Zgłoszenie patentowe w UPRP*, nr P.43919.

[3] Stefański Jacek, Sadowski Jarosław, 2020. "Sensor Position Estimation Method for IoT Using Mobile Reference Node". *IEEE Access* 8: 79287-79298.

DEMONSTRATOR RADARU PASYWNEGO Z ODWROTNĄ SYNTETYCZNĄ APERTURĄ

PASSIVE INVERSE SYNTHETIC APERTURE RADAR DEMONSTRATOR

Marcin Kamil Bączyk

Politechnika Warszawska, Instytut Systemów Elektronicznych, Warszawa, marcin.baczyk@pw.edu.pl

DOI: 10.15199/59.2022.4.106

Streszczenie: **W referacie przedstawiono demonstrator multistatycznego radaru pasywnego z odwrotną syntetyczną aperturą służącego do obrazowania poruszających się obiektów bez własnej emisji promieniowania. W pracy przedstawiono także wyniki przetwarzania rzeczywistych sygnałów zarejestrowanych za pomocą multistatycznego radaru pasywnego zbudowanego. Pokazano także otrzymane pasywne zobrazowania poruszającego się niewielkiego samolotu cywilnego.**

Abstract: **The paper presents a demonstrator of multistatic passive radar with an inverse synthetic aperture for imaging moving objects without its radiation emission. The article also presents the results of processing real signals recorded using a built multistatic passive radar. The obtained passive images of a moving small civil aircraft are also shown.**

Słowa kluczowe: **Radar pasywny, odwrotna syntetyczna apertura, demonstrator technologii.**

Keywords: **Passive radar, inverse synthetic aperture, technology demonstrator.**

1. WSTĘP

W radarze pasywnym do oświetlenia obiektu, zamiast własnej emisji, wykorzystuje się inne źródła promieniowania. Przeważnie są to, nie tworzone z myślą o wykrywaniu obiektów, nadajniki naziemnej telewizji lub radia [1]. Wynikający z tego brak nadajnika w radarze pasywnym i emisji własnej oznacza przede wszystkim brak możliwości wykrycia radaru na podstawie analizy widma sygnału, ale także brak konieczności asocjacji odrębnego pasma częstotliwości oraz niższy koszt budowy i utrzymania tej części radaru, która zwykle w radarach aktywnych jest znacznie bardziej złożona niż odbiornik.

Konstrukcja klasycznego radaru pasywnego wymaga rejestracji co najmniej dwóch sygnałów. Pierwszy z nich to sygnał referencyjny odbierany przez antenę skierowana bezpośrednio na nadajnik. Z kolei sygnał pomiarowy rejestrowany jest przez antenę skierowaną na wybrany obszar. Jeżeli w tym obszarze znajduje się obiekt, fala elektromagnetyczna emitowana przez niezależny nadajnik odbije się od tego obiektu i jego echo zostanie zarejestrowane przez kanał pomiarowy odbiornika. Na podstawie porównania obu sygnałów możliwe jest określenie opóźnienia sygnału echa odbitego od obiektu oraz przesunięcia częstotliwości spowodowanej efektem Dopplera [3].

Ponieważ przy pomocy pojedynczego odbiornika radaru pasywnego, rejestrującego sygnał referencyjny raz pomiarowy możliwe jest określenie jedynie odległości i prędkości bistatycznych do pełnej lokalizacji i śledzenie obiektów wykorzystuje się systemy multistatyczne, złożone z wielu odbiorników. Dopiero na podstawie pomiaru wykonanego przez wiele pojedynczych systemów możliwe jest określenie położenia obiektu [4]

Podobnie jak w radarach aktywnych w radarach pasywnych możliwe jest wykonanie zobrazowania poruszającego się obiektu. Techniki obrazowania radarowego wykorzystują wzajemną zmianę położenia lub orientacji w układzie nadajnik-obiekt-odbiornik. Możliwość wykonania zobrazowania w radarach wynika z faktu, że każdy punkt obiektu znajduje się na innej odległości oraz porusza się z inną prędkością względem radaru lub pary nadajnik-odbiornik. W ten sposób dysponując odpowiednio dużą rozróżnialnością odległościową oraz prędkościową możliwe jest wyznaczenie pozycji każdego z punktów obiektu i wykonanie jego zobrazowania.

W niniejszym referacie przedstawiono zbudowany na Politechnice Warszawskiej demonstrator radaru pasywnego z odwrotną syntetyczną aperturą składającego się z trzech niezależnych stacji odbiorczych. Opisano budowę pojedynczego systemu. W kolejnych punktach opisano wykonane pomiary oraz zaprezentowano wyniki przetworzenia sygnałów.

2. SPRZĘTOWY DEMONSTRATOR MULTISTATYCZNEGO RADARU PASYWNEGO

Nim przystąpiono do budowy rzeczywistego demonstratora radaru pasywnego określono wymagania jakie powinien on spełniać. Przede wszystkim budowany rejestrator miał umożliwiać ciągłą rejestrację sygnałów przez cały czas trwania manewrów obiektów kooperujących. Dodatkowo, ponieważ przeprowadzane później eksperymenty miały potwierdzić możliwość obrazowania multistatycznego, konieczne było zapewnienie odpowiedniej synchronizacji dla poszczególnych stacji pomiarowych. Istotne było również zapewnienie dostatecznie szerokiego pasma rejestracji sygnału, w celu jednoczesnej koherentnej rejestracji wielu kanałów częstotliwościowych naziemnej telewizji cyfrowej. Nie mniej ważnym

zagadnieniem było umożliwienie podglądu w czasie rzeczywistym, zarówno charakterystyki widmowej rejestrowanego sygnału, jak również wyniku przetwarzania sygnałów przez demonstrator radaru pasywnego w celu weryfikacji poprawności rejestrowanych danych. Ostatnie zagadnienie było istotne głównie ze względu na brak możliwości ponownego przeprowadzenia eksperymentu. Szybka reakcja na nieprawidłowe wyniki wstępnego przetwarzania sygnałów niejednokrotnie uchroniła przed utratą wartościowych danych.

Na potrzeby budowy demonstratora multistatycznego radaru pasywnego wykorzystano sprzęt dostępny w laboratorium Pracowni Technik Radiolokacyjnych Instytutu Systemów Elektronicznych. W skład pojedynczego systemu wchodzi:

- szafa NI-PXIe wraz z komputerem sterującym, układem odbiorczym sygnału GPS oraz układem synchronizującym,
- trzy urządzenia rejestrujące typu USRP X310 w paśmie od 10 MHz do 6 GHz oraz szerokości rejestrowanego sygnału do 160 MHz,
- macierz dyskowa zapewniająca transfer danych na poziomie 1,6GB/s umożliwiająca rejestrację do 400 milionów próbek na sekundę we wszystkich kanałach łącznie,
- jednostka obliczeniowa, pozwalająca na wstępne przetworzenie sygnałów i ich podgląd.

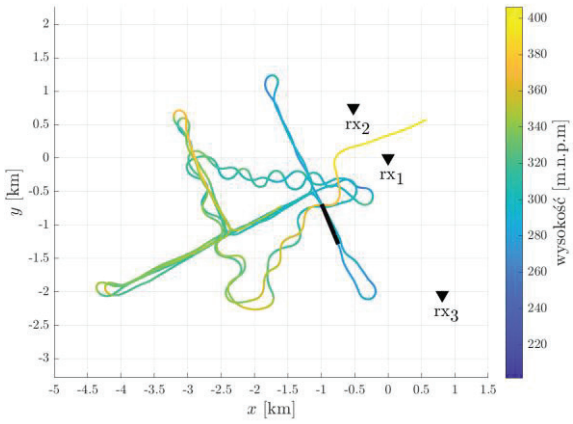
Wszystkie urządzenia połączone są bezpośrednio do szafy NI-PXIe przy pomocy interfejsu PCI Express. W trakcie pracy systemu dane rejestrowane przez urządzenia odbierane są przez komputer sterujący, a następnie zapisywane bezpośrednio na dysk oraz wysyłane do jednostki obliczeniowej. Wyniki wstępnie przetworzonego sygnału wyświetlane są na monitorze podłączonym do jednostki obliczeniowej. Dodatkowo jednostka sterująca przeprowadza szybką analizę fragmentów rejestrowanego sygnału, w celu weryfikacji podstawowych jego parametrów. Analiza wyświetlanych wyników pozwala na ocenę poprawności rejestracji sygnałów.

3. OPIS PRZEPROWADZONYCH POMIARÓW

Przedstawione w niniejszym referacie wyniki zostały uzyskane jako efekt przetworzenia sygnałów zarejestrowanych we Włoszech w prowincjach Piza oraz Livorno w czerwcu 2017 roku. Miejsce to zostało wybrane ze względu na dostępność dużej liczby nadajników naziemnej telewizji cyfrowej. Sygnały były emitowane przez wiele nadajników umieszczonych na różnych masztach, które ulokowane były na jednym wzniesieniu, w niedużej odległości od siebie. Dzięki temu rejestrację sygnału referencyjnego można było przeprowadzić za pomocą pojedynczej anteny.

W celu przetestowania różnych możliwych konfiguracji radaru multistatycznego w trakcie przeprowadzanych pomiarów wykorzystywanych było kilka różnych scenariuszy. Poniżej przedstawiono jeden z nich, dla którego zaprezentowano uzyskane wyniki w dalszej części rozdziału.

Na rysunku 1 przedstawiono poglądową mapę prezentującą pozycje odbiorników w omawianym scenariuszu oraz trajektorię obserwowanego obiektu, którym był niewielki samolot Cessna 182. Linią pogrubioną zaznaczono fragment jego trajektorii, po której poruszał się w trakcie integracji sygnału wykorzystanego do uzyskania zobrazowania. Trajektoria przedstawiona na rysunku 1 pozyskana została dzięki umieszczonemu na pokładzie samolotu rejestratorowi pozycji wyznaczanej przy użyciu sygnałów GPS. Ponieważ oprócz pozycji odbiorniki rejestrowały również czas, nie było konieczności dodatkowej kalibracji.



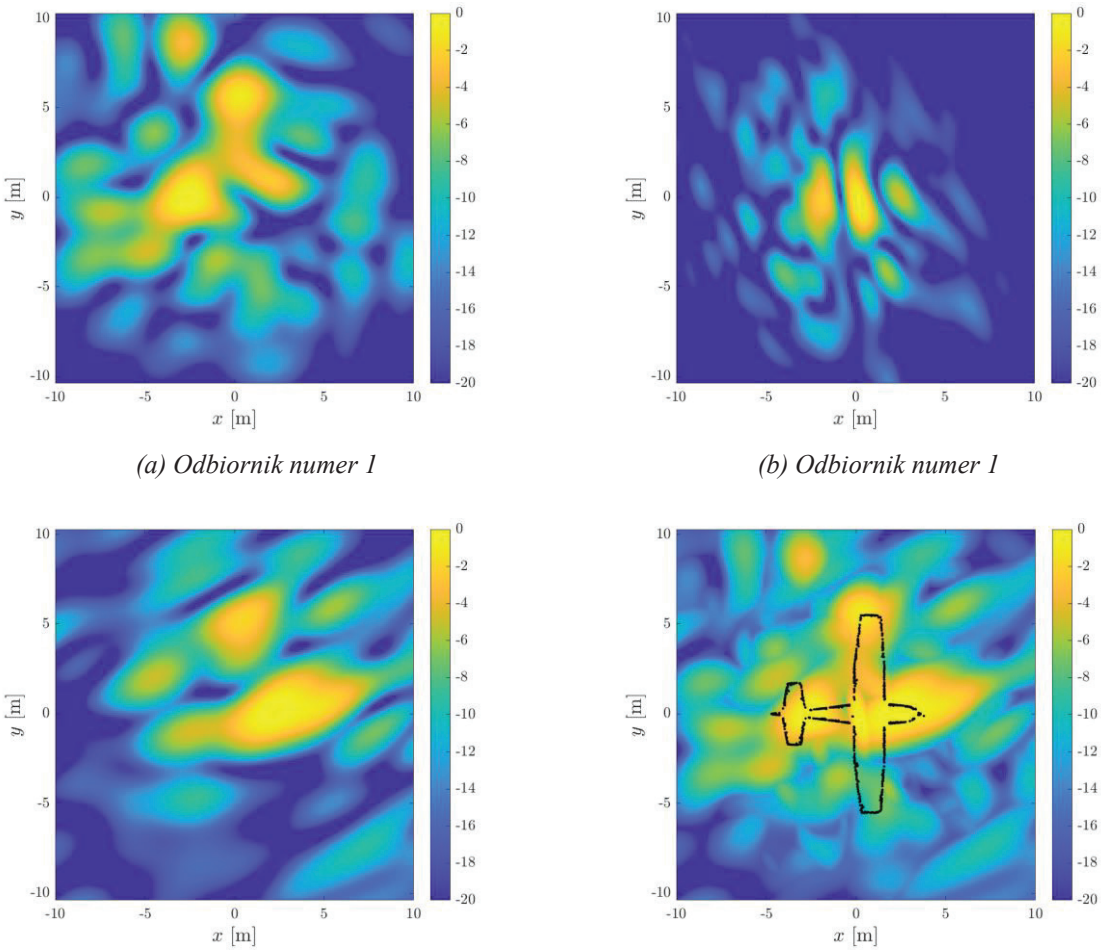
Rys. 1 Pozycje poszczególnych odbiorników, nadajników oraz trajektorii obrazowanego obiektu względem odbiornika numer 1.

4. PREZENTACJA WYNIKÓW

Na rysunku 2 przedstawiono wynik przetwarzania zarejestrowanych sygnałów w postaci zogniskowanych zobrazowań obserwowanego obiektu. Rysunki 2a-2c przedstawiają odpowiednio wyniki uzyskane dla sygnałów zarejestrowanych przez poszczególne odbiorniki. Niekoherentną sumę tych zobrazowań przedstawiono na ostatnim rysunku – 2d. Dodatkowo, dla ułatwienia wzrokowej identyfikacji poszczególnych elementów charakterystycznych echa obserwowanego obiektu na ostatnim rysunku naniesiono kontur modelu samolotu odpowiadającego obserwowanemu celowi.

Analizując przedstawione wyniki należy zwrócić uwagę, że obrazy obiektu wykonane poprzez przetworzenie sygnałów zarejestrowanych przez poszczególne odbiorniki są znacząco różne. Wynika to przede wszystkim z faktu obserwacji obiektu pod różnymi kątami, co skutkuje tym, że każdy z odbiorników rejestruje echa innych fragmentów obiektu.

Należy też zaznaczyć, że różnice w zobrazowaniach wynikają również z innego poziomu rejestrowanego sygnału. Każdy z torów odbiorczych charakteryzował się innym poziomem wzmocnienia. Informacja o nim została utracona w trakcie przeprowadzania samych pomiarów ze względu na niedoskonałości systemu rejestrującego. W związku z tym, wyniki przedstawione zostały w skali decybelowej wyrażonej względem maksymalnej intensywności danego zobrazowania.



(a) Odbiornik numer 1

(b) Odbiornik numer 1

(c) Odbiornik numer 1

(d) Suma poszczególnych zobrazowań

Rys. 2 wyniki zobrazowań obiektu dla poszczególnych odbiorników.

Odbiornik centralny w czasie rejestracji przetwarzanego sygnału znajdował się „najbardziej" z boku obserwowanego obiektu. Kierunek wyznaczany przez pozycję centralnego odbiornika oraz obiektu był zbliżony do kierunku wyznaczanego przez pozycję nadajników oraz obiektu. Z analizy wyrażenia:

$$\boldsymbol{k}(t,f) = 2\pi\frac{f}{c}\left(\frac{\boldsymbol{r_1}(t)}{r_1(t)} + \frac{\boldsymbol{r_2}(t)}{r_2(t)}\right),$$

gdzie f oznacza częstotliwość, t czas bieżący a $\boldsymbol{r_1}$ oraz $\boldsymbol{r_2}$ wektor położenia obiektu względem nadajnika i odbiornika, dla geometrii omawianej sceny radiolokacyjnej przedstawionej na rysunku 2 wynika, że rozróżnialność w obu kierunkach, odległościowymi oraz azymutalnym, jest największa dla sygnałów zarejestrowanych przez odbiornik centralny. Odczytana szerokość listka głównego punktu obrazu najsilniej odbijającego wynosi około 2,5 m, co odpowiada pasmu sygnału o szerokości około 56 MHz. Jest to szerokość pasma jaką zajmuje 7 sąsiadujących ze sobą sygnałów naziemnej telewizji cyfrowej o szerokości 8 MHz każdy. Z kolei rozróżnialność w kierunku azymutalnym, zależna od całkowitej zmiany kąta obserwacji z trakcie rejestracji sygnału, wynosi około 1 m. Z kolei, dla odbiorników bocznych, ze względu na znacznie większy kąt bistatyczny oraz mniejszą zmianę kąta pod jakim obserwowany jest obiekt w trakcie rejestracji sygnału, uzyskiwane rozróżnialności są nieznacznie mniejsze i wynoszą około 3 m oraz 2 m odpowiednio w kierunku odległościowym oraz azymutalnym.

5. DYSKUSJA WYNIKÓW

Występujące w przedstawionych w wyżej wynikach różnice w składowych zobrazowań nie sprowadzają się jedynie do osiąganych rozróżnialności. Zgodnie z rysunkiem 2 każdy z odbiorników obserwuje cel pod innym kątem. Odbiornik centralny, znajdujący się najbardziej na lewo od kierunku, w którym porusza się samolot, widzi jedynie jego boczny profil. Podobnie jak w przypadku obserwacji wizyjnej pozostałe odbiorniki boczne rejestrują odpowiednio lewy tył oraz lewy przód obiektu.

Interesujący jest fakt, że na obrazie odpowiadającemu centralnemu odbiornikowi nie jest widoczne echo końca lewego skrzydła, które jest wyraźnie widoczne na pozostałych dwóch obrazach. Prawdopodobnie echo skrzydła ukryte jest w listkach bocznych silnego echa kadłuba samolotu. Jego skuteczna powierzchnia odbijająca

jest znacznie większa od skutecznej powierzchni odbijającej końca skrzydła. W przypadku obserwacji obiektu niemal z boku stosunkowo duża i dostatecznie płaska powierzchnia samolotu może działać jak lustro dla fali elektromagnetycznej, co z kolei powoduje powstanie silnego echa pochodzącego od tej powierzchni. W opisywanym eksperymencie wszystkie nadajniki emitujące sygnały znajdowały się nad płaszczyzną wyznaczaną przez skrzydła samolotu, natomiast wszystkie odbiorniki poniżej tej płaszczyzny.

W związku z tym odbiorniki nie rejestrują sygnału odbitego płata, który jest oświetlany od góry. Jedynym fragmentem skrzydła, który jest oświetlany przez nadajniki oraz jest widoczny z poziomu odbiorników, jest jego boczny fragment. Co więcej, ponieważ nadajniki znajdowały się lekko z tyłu, echo skrzydła samolotu jest nieznacznie silniejsze w stosunku do reszty samolotu w odbiorniku numer 2 znajdującym się za obserwowanym samolotem.

Zróżnicowana charakterystyka odbijania fali elektromagnetycznej obserwowanego obiektu, odmienna geometria rozmieszczenia zarówno nadajników, jak i odbiorników, a także efekty przysłaniania różnych fragmentów poszycia sprawiają, że poszczególne obrazy różnią się. Dopiero złożenie wszystkich składowych w jeden obraz ukazuje lewy kontur obserwowanego obiektu, na podstawie którego możliwa jest wstępna klasyfikacja obserwowanego obiektu.

Podobne efekty można zauważyć w wynikach obrazowania uzyskanych dla symulowanych sygnałów. Pozycje poszczególnych nadajników i odbiorników oraz trajektoria obiektu wykorzystanych w symulacjach odpowiadały rzeczywistym wartościom przyjętym w przetwarzaniu sygnałów pomiarowych.

Najlepsza możliwa do uzyskania rozróżnialność odległościowa radaru pasywnego wykorzystującego pojedyncze pasmo DVB-T jest prawie dwukrotnie większa od rozmiarów obserwowanego rzeczywistego obiektu. Z tego względu stosunkowo trudno oszacować poprawność końcowego wyniku. Aby zweryfikować ostatecznie otrzymany wynik, należałoby przeprowadzić jednoczesny eksperyment, wykorzystując punktowe źródło sygnału o parametrach takich samych jak parametry sygnału DVB-T. Jest to z oczywistych powodów niemożliwe do wykonania.

Otrzymane wyniki wydają się nie być w pełni zogniskowane. Pomimo wyboru chwili, w której wpływ rozproszonego charakteru nadajnika jest minimalny, konieczne jest jego uwzględnienie i dodanie dodatkowego procesu ogniskowania, w którym dryfty fazy zostaną skompensowane [2]. Uzyskanie w pełni zogniskowanych zobrazowań jest niezbędne do ich koherentnej integracji. Niedoskonałości uzyskanych wyników nie pozwalają na identyfikację obserwowanego obiektu. Jednak ostateczny rezultat umożliwia określenie rozmiaru celu lub wręcz jego klasyfikację.

Radar aktywny, pracujący na przykład w paśmie X, o dużej szerokości wykorzystywanego pasma, będzie prawdopodobnie pod kątem jakości zobrazowań lepszym wyborem od radaru pasywnego. Dlatego tam gdzie jest to możliwe lepiej korzystać z radarów aktywnych. Natomiast w przypadku, gdy nie ma takiej możliwości lub promieniowanie własnego sygnału elektromagnetycznego jest niepożądane, radary pasywne są w stanie dostarczyć zobrazowania pozwalające na wstępne rozpoznanie celów znajdujących się w chronionej przestrzeni powietrznej.

6. PODSUMOWANIE

W referacie zaprezentowano wyniki eksperymentu przeprowadzonego w celu weryfikacji opracowanych w ramach rozprawy metod. Pokrótce opisano geometrię sceny radiolokacyjnej oraz systemy pomiarowe opracowane na potrzeby realizacji eksperymentu.

Otrzymane wyniki przetworzenia sygnałów rzeczywistych są zbieżne z niezaprezentowanymi w referacie wynikami przetworzenia sygnałów symulowanych.

Wyniki przeprowadzonego eksperymentu ostatecznie weryfikują pozytywnie możliwość zastosowania radaru pasywnego do obrazowania obiektów ruchomych. Koherentna synteza obrazów uzyskanych dla różnych pasm pozwala uzyskać rozróżnialność odległościową niezbędną do obrazowania niewielkich samolotów. Z kolei niekoherentna synteza obrazów uzyskanych dla różnych odbiorników obserwujących dany obiekt w danym momencie pod różnymi kątami pozwala na wydobycie większej liczby szczegółów obrazu obserwowanego obiektu.

LITERATURA

[1] Kulpa Krzysztof, Malanowski Mateusz. 2019. „From Klein Heidelberg to Modern Multistatic Passive Radar". *20th International Radar Symposium (IRS)* : 1-9

[2] Kantor Joshua M. 2020. „Minimum Entropy Autofocus Correction of Range-Varying Phase Errors in ISAR Imagery". *IEEE International Radar Conference (RADAR)* : 357–361.

[3] Zhao Yongsheng, Hu Dexiu, Zhao Yongjun, Liu Zhixin. 2020. „Moving target localization for multistatic passive radar using delay, Doppler and Doppler rate measurements". *Journal of Systems Engineering and Electronics* 939-949.

[4] Malanowski Mateusz , Kulpa Krzysztof. 2012. „Two Methods for Target Localization in Multistatic Passive Radar". *IEEE Transactions on Aerospace and Electronic Systems* 48(1) : 572-580.

SZYBKI ALGORYTM PROJEKCJI WSTECZNEJ OPARTY NA WSPÓŁRZĘDNYCH BARYCENTRYCZNYCH DO ZASTOSOWAŃ W AKTYWNYCH I PASYWNYCH ZOBRAZOWANIACH RADAROWYCH SAR

FAST BARYCENTRIC-BASED BACK PROJECTION ALGORITHM FOR PASSIVE AND ACTIVE SAR IMAGING

Krzysztof Radecki

Politechnika Warszawska, Instytut Systemów Elektronicznych, Warszawa, K.Radecki.1@elka.pw.edu.pl

DOI: 10.15199/59.2022.4.107

Streszczenie: **Algorytm projekcji wstecznej (ang. Back Projection Algorithm, BPA), stosowany w obrazujących z syntetyczną aperturą radarach, jest bardzo wymagającym rozwiązaniem pod względem złożoności obliczeniowej. Opisany w referacie algorytm projekcji wstecznej wykorzystujący współrzędne barycentryczne, dzięki innowacyjnej metodzie zawężania zakresu surowych danych radarowych potrzebnych do uzyskania zobrazowania, jest bardziej wydajny pod względem nakładu obliczeniowego w porównaniu do klasycznego algorytmu BPA i może być zastosowany w systemach radarowych pracujących w czasie rzeczywistym.**

Abstract: **Back projection (BP) algorithm is one of the most computationally demanding methods in Synthetic Aperture Radar (SAR) imaging technique. The huge computational complexity of this algorithm causes problems in real time radar systems. The article presents newly developed fast barycentic-based back projection algorithm which allows the classical well-known BP algorithm to be significantly accelerated. The described solution can be adapted to a real-time SAR systems.**

Słowa kluczowe: **SAR, BPA, projekcja wsteczna, współrzędne barycentryczne.**

Keywords: **SAR, BPA, back projection, barycentric coordinates.**

1. WSTĘP

Radar z syntetyczną aperturą (ang. *Synthetic Aperture Radar,* SAR) to technika, której początki sięgają lat 50. XX wieku. Służy ona do tworzenia wysokorozdzielczych zobrazowań powierzchni Ziemi. Radary SAR są powszechnie wykorzystywane w aplikacjach cywilnych oraz militarnych. Mogą pracować w każdych warunkach pogodowych zarówno w dzień jak i w nocy [3, 4].

System radaru obrazującego SAR jest realizowany poprzez umieszczenie radaru na platformie z anteną oświetlającą obszar Ziemi. Platforma będąca nośnikiem radaru to zazwyczaj samolot bądź satelita, poruszające się po linii prostej ze stałą prędkością [3, 4].

Minimalna odległość, która pozwala zaobserwować dwa różne obiekty na zobrazowaniu jest odwrotnie proporcjonalna do pasma sygnału [3, 4]. Dzisiejsze rozwiązania, wykorzystujące urządzenia pracujące w szerokim paśmie częstotliwości, pozwalają uzyskiwać obrazy SAR o rozróżnialności pojedynczych centymetrów. Dostęp do urządzeń takich jak radio programowalne (ang. *Software-Defined Radio,* SDR) oraz komercyjnych rozwiązań komponentów mikrofalowych z segmentu COTS (ang. *Commercial Off-The-Shelf*) dają możliwość tworzenia systemów SAR bez ponoszenia dużych kosztów [12].

Przez lata opracowano wiele algorytmów przetwarzania zobrazowań SAR. Powszechnie znanym rozwiązaniem jest algorytm projekcji wstecznej (ang. *Back Projection Algorithm,* BPA). Głównym założeniem algorytmu jest koherentna integracja odebranych sygnałów radarowych [2, 5]. Wadą BPA jest duży nakład obliczeniowy [6], który wręcz uniemożliwia zastosowanie tego algorytmu w systemach pracujących w czasie rzeczywistym.

Powstało wiele prac traktujących o przyspieszeniu algorytmu BPA [8, 11, 13]. Część zaproponowanych rozwiązań polega na podziale przetwarzanego obrazu na mniejsze obszary, a następnie na sumowaniu powstałych wyników. Jednak podczas łączenia wyników mogą powstawać różnice w porównaniu do rozwiązań uzyskiwanych za pomocą klasycznej techniki BPA.

W referacie przedstawiony został autorski algorytm projekcji wstecznej wykorzystujący współrzędne barycentryczne i pozwalający na szybsze uzyskiwanie wysokorozdzielczych zobrazowań SAR [9]. Opisane rozwiązanie polega na wyznaczaniu odwróconej wiązki syntetyczne apertury, co pozwala określić zakres danych radarowych potrzebnych do przetworzenia zobrazowania. Algorytm ten jest łatwy do implementacji na urządzeniach wspierających obliczenia równoległe i daje możliwość wykorzystania techniki BPA w systemach czasu rzeczywistego [9].

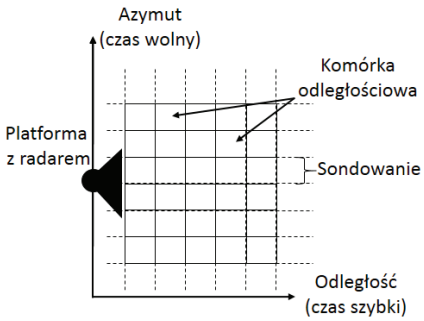
W rozdziale drugim referatu opisany został klasyczny algorytm BPA. Rozdział 3 przedstawia nowatorski algorytm projekcji wstecznej wykorzystujący współrzędne barycentryczne. Rozdział 5 i 6 to odpowiednio przedstawienie wyników oraz ich podsumowanie.

2. ALGORYTM PROJEKCJI WSTECZNEJ

Algorytm BPA w technice radarowej został niejako zapożyczony z medycyny, gdzie algorytm projekcji wstecznej był stosowany do rekonstrukcji obrazów tomograficznych [1]. BPA pozwala uzyskiwać zobrazowania radarowe o wysokiej rozdzielczości. Działanie algorytmu

wymaga znajomości relacji między położeniem rekonstruowanego punku, w którym tworzone jest obrazowanie, a położeniem radaru obserwującego scenę obrazowaną. Dane radarowe zarejestrowane w kolejnych momentach położenia platformy z radarem są integrowane.

Pierwszym krokiem w celu uzyskania zobrazowania SAR, jest wykonanie kompresji odległościowej. Surowe dane radarowe są formowane w postaci dwuwymiarowej macierzy danych radarowych (Rys. 1).



Rys. 1. Dwuwymiarowa macierz danych radarowych

Wiersze macierzy zawierają sondowania. Każde sondowanie składa się z komórek odległościowych. Kompresja odległościowa w przypadku radaru z liniowo zmodulowaną falą ciągłą FMCW (ang. *Frequency Modulated Continuous Wave*) polega na policzeniu szybkiej transformaty Fourdiera FFT (ang. *Fast Fourier Transform*) dla każdego wiersza macierzy z danymi:

$$R(m,n) = \sum_{k=0}^{N-1} r(m,k) e^{\frac{j2\pi kn}{N}},\ n = 0, \dots, N-1, \quad (1)$$

gdzie $r(m,k)$ to element (m,k) – tej komórki odległościowej bloku danych przed kompresją. $R(m,n)$ opisuje elementy komórek odległościowych po kompresji odległościowej z wiersza o indeksie m i kolumny o indeksie n. N to liczba komórek odległościowych w sondowaniu.

Następnie należy wyznaczyć odległość między obrazowanym punktem a położeniem platformy z radarem w momencie zarejestrowania sondowania:

$$R(m) = \sqrt{X(m) + Y(m) + Z(m)}, m = 0, \dots, M-1, \quad (2)$$

$$X(m) = (x_s(m) - x_p)^2, \quad (3)$$

$$Y(m) = (y_s(m) - y_p)^2, \quad (4)$$

$$Z(m) = (z_s(m) - z_p)^2, \quad (5)$$

gdzie $R(m)$ to odległość między obrazowanym punktem o współrzędnych x_p, y_p, i z_p a położeniem platformy z radarem w punkcie o współrzędnych $x_s(m)$, $y_s(m)$ i $z_s(m)$ w momencie zarejestrowania sondowania o indeksie m. M to liczba sondowań w bloku danych radarowych.

W kolejnym kroku można wyznaczyć indeks n komórki odległościowej dla rozpatrywanego sondowania m z bloku danych radarowych:

$$n = \frac{R(m) - 2r_{min}}{2\delta_r}, \quad (6)$$

gdzie r_{min} to minimalny zasięg oświetlania anteny nadawczej radaru, δ_r to rozmiar komórki odległościowej.

W celu wykonania kompresji w kierunku azymutalnym należy pomnożyć wartość komórki odległościowej po kompresji odległościowej przez wartość odpowiedzi impulsowej filtru dopasowanej o następującej postaci:

$$h(m) = \exp(j\frac{2\pi}{\lambda}R(m)), \quad (7)$$

gdzie λ to długość fali.

Moduł splotu odpowiedzi impulsowej z blokiem danych radarowych po kompresji odległościowej pozwala na uzyskanie zobrazowania SAR w rozpatrywanym punkcie P:

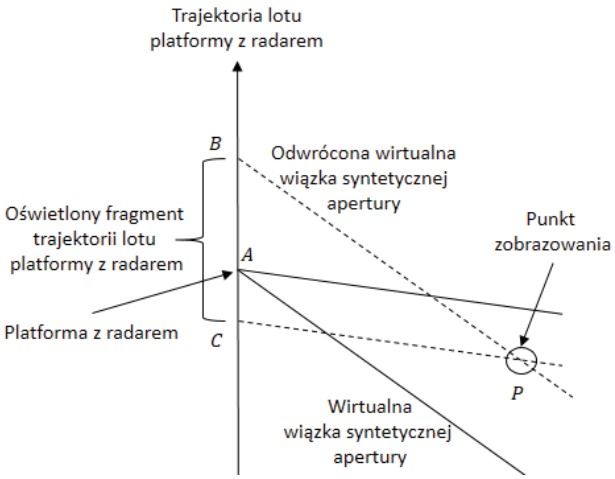
$$P = \left|\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} R(x_p - m, y_p - n) h(m)\right|. \quad (8)$$

Złożoność obliczeniowa BPA wynosi $O(n^3)$ [11]. Jeśli macierz danych radarowych zawiera 1000 komórek odległościowych, to liczba operacji potrzebna do uzyskania zobrazowania za pomocą algorytmu BPA wynosi 1000^3. Duży nakład obliczeniowy potrzebny do uzyskiwania zobrazowań praktycznie uniemożliwia wykorzystanie tego algorytmu w systemach pracujących w czasie rzeczywistym.

3. ALGORYTM PROJEKCJI WSTECZNEJ WYKORZYSTUJĄCY WSPÓŁRZĘDNE BARYCENTRYCZNE

Opracowany algorytm pozwala na zmniejszenie liczby wykonywanych operacji arytmetycznych podczas uzyskiwania zobrazowania przy jednoczesnym zachowaniu dużej dokładności względem klasycznego algorytmu BPA. Głównym elementem algorytmu wykorzystującego współrzędne barycentryczne jest tworzenie wirtualnej wiązki syntetycznej apertury i zawężenie zakresu surowych danych radarowych potrzebnych do rekonstrukcji punktu zobrazowania. Jeśli obrazowany punkt znajduje się w obszarze oświetlanym przez wirtualną antenę radaru, wówczas dane radarowe zarejestrowane w danym momencie są wykorzystywane przy rekonstrukcji obrazu w tym punkcie. Jeśli dany punkt znajduje się poza wirtualną wiązką anteny, wówczas dane radarowe zarejestrowane w tym położeniu radaru nie są wykorzystywane w rekonstrukcji obrazu. W praktycznej realizacji algorytmu lepiej jest odwrócić wirtualną wiązkę anteny tak, aby jej początek znajdował się w obrazowanym punkcie, a jej wiązka oświetlała trajektorię lotu platformy z radarem. Oświetlony fragment trajektorii lotu platformy z radarem określa zakres surowych danych radarowych potrzebnych do uzyskania zobrazowania w danym punkcie. Idea odwróconej wiązki syntetycznej apertury przedstawiona została na Rys. 2.

Mechanizm rozstrzygania czy dany rekonstruowany punkt znajduje się wewnątrz wyznaczonej wiązki anteny, czy znajduje się poza nią, został opracowany z wykorzystaniem współrzędnych barycentrycznych, które są wykorzystywane w grafice komputerowej [7].



Rys. 2. Idea wirtualnej wiązki syntetycznej apertury

Znając współrzędne P, B, C (Rys. 2) wierzchołków trójkąta opisującego wirtualną wiązkę syntetycznej apertury można określić, które punkty trajektorii platformy z radarem znajdują się wewnątrz wiązki wykorzystując współrzędne barycentryczne.

Pierwszym krokiem algorytmu wyznaczania współrzędnych barycentrycznych jest wyznaczenie wektorów $\overrightarrow{PB}$ i $\overrightarrow{PC}$:

$$\overrightarrow{PB} = (x_B - x_P, y_B - y_P), \tag{9}$$

$$\overrightarrow{PC} = (x_C - x_P, y_C - y_P), \tag{10}$$

gdzie x_P, y_P to współrzędne obrazowanego punktu, natomiast x_B, y_B oraz x_C, y_C to odpowiednio współrzędne punktów B i C leżących na trajektorii platformy z radarem (Rys. 2).

Położenie platformy z radarem w punkcie A opisane względem obrazowanego punktu P ma następującą postać:

$$A = P + \iota\overrightarrow{PC} + \nu\overrightarrow{PB}, \tag{11}$$

gdzie ι oraz ν to współrzędne barycentryczne. Zależność tę można uprościć wyznaczając wektor $\overrightarrow{PA}$:

$$\overrightarrow{PA} = (x_A - x_P, y_A - y_P), \tag{12}$$

$$\overrightarrow{PA} = \iota\overrightarrow{PC} + \nu\overrightarrow{PB}, \tag{13}$$

gdzie x_A, y_A to współrzędne punktu A położenia platformy z radarem.

W celu wyznaczenia współrzędnych barycentrycznych ι, ν należy wykonać kilka przekształceń. Aby uzyskać układ równań z dwiema niewiadomymi zmiennymi, należy wyznaczyć iloczyny skalarne wektora $\overrightarrow{PA}$ (13) z wektorami $\overrightarrow{PC}$ i $\overrightarrow{PB}$:

$$\overrightarrow{PA} \cdot \overrightarrow{PC} = \iota\left(\overrightarrow{PC} \cdot \overrightarrow{PC}\right) + \upsilon(\overrightarrow{PB} \cdot \overrightarrow{PC}), \tag{14}$$

$$\overrightarrow{PA} \cdot \overrightarrow{PB} = \iota\left(\overrightarrow{PC} \cdot \overrightarrow{PB}\right) + \upsilon(\overrightarrow{PB} \cdot \overrightarrow{PB}). \tag{15}$$

Po wyznaczeniu ι i ν układ równań przyjmuje następującą postać:

$$\iota = \frac{(\overrightarrow{PB}\cdot\overrightarrow{PB})(\overrightarrow{PC}\cdot\overrightarrow{PA}) - (\overrightarrow{PC}\cdot\overrightarrow{PB})(\overrightarrow{PB}\cdot\overrightarrow{PA})}{(\overrightarrow{PC}\cdot\overrightarrow{PC})(\overrightarrow{PB}\cdot\overrightarrow{PB}) - (\overrightarrow{PC}\cdot\overrightarrow{PB})(\overrightarrow{PC}\cdot\overrightarrow{PB})}, \tag{16}$$

$$\upsilon = \frac{(\overrightarrow{PC}\cdot\overrightarrow{PC})(\overrightarrow{PB}\cdot\overrightarrow{PA}) - (\overrightarrow{PC}\cdot\overrightarrow{PB})(\overrightarrow{PC}\cdot\overrightarrow{PA})}{(\overrightarrow{PC}\cdot\overrightarrow{PC})(\overrightarrow{PB}\cdot\overrightarrow{PB}) - (\overrightarrow{PC}\cdot\overrightarrow{PB})(\overrightarrow{PC}\cdot\overrightarrow{PB})}. \tag{17}$$

Warunki, jakie muszą spełniać współrzędne barycentryczne ι i ν aby dane radarowe ze stanu położenia platformy z radarem w punkcie A były integrowane podczas rekonstrukcji punktu P są następujące:

$$\iota \geq 0, \tag{18}$$

$$\upsilon \geq 0. \tag{19}$$

Nie jest sprawdzany warunek, czy suma współrzędnych barycentrycznych nie przekracza jeden, ponieważ w praktyce lot platformy z radarem nie jest idealnie prosty. Jedynym warunkiem jest położenie platformy z radarem między ramionami wyznaczonej wiązki.

4. WYNIKI

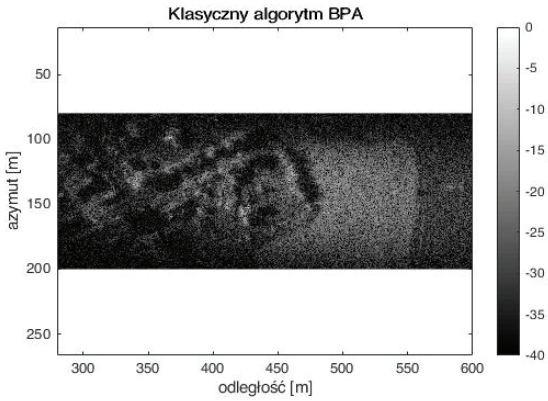
W eksperymencie zostały wykorzystane dane radarowe zarejestrowane z wykorzystaniem radaru aktywnego podczas kampanii pomiarowej w dniach 5-8 maja 2015 r. w Płockim Aeroklubie Ziemi Mazowieckiej (Rys. 3) [10]. Platforma na której umieszczony został radar to samolot Wilga PZL-104.

Rys. 3. Wilga PZL-104 jako nośnik radaru SAR

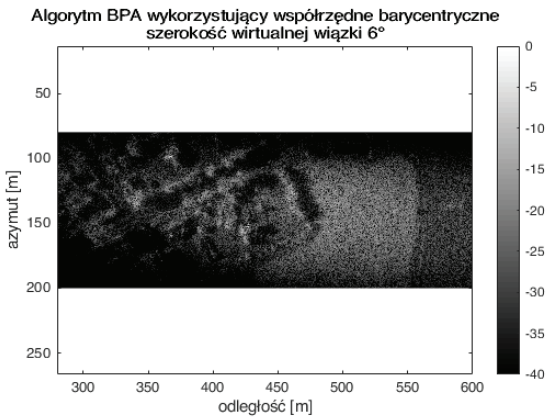
Parametry pracy radaru były następujące: częstotliwość nośna sygnału $35GHz$, pasmo emitowanego sygnału $1GHz$, częstotliwość próbkowania $2GHz$, częstotliwość powtarzania impulsów $2KHz$.

Rysunek 4 przedstawia wynik obrazowania uzyskany za pomocą klasycznego algorytmu BPA. Rysunek 5 przedstawia wynik uzyskany za pomocą algorytmu wykorzystującego współrzędne barycentryczne z szerokością wirtualnej wiązki syntetycznej apertury 6°.

Dane radarowe wykorzystane do uzyskania obrazowania zostały zarejestrowane w czasie $5s$ lotu platformy z radarem. Obrazowania składają się z 1000 punktów w kierunku azymutalnym i odległościowym. Obrazowania zostały uzyskane na urządzeniu wyposażonym w procesor Intel Core i7-4702MQ, 16 GB pamięci RAM i kartę graficzną GeForce GT 740M 2 GB.



Rys. 4. Obrazowanie uzyskane za pomocą klasycznego algorytmu BPA



Rys. 5. Obrazowanie uzyskane za pomocą algorytmu BPA wykorzystującego współrzędne barycentryczne

Klasyczny algorytm BPA, wykonywany sekwencyjnie, zwrócił wynik po 26 minutach. Nowo opracowany algorytm zwrócił wynik po 17 minutach. Po zrównoleglenniu obliczeń czasy uzyskania zobrazowań wyniosły odpowiednio $19s$ i $7s$.

5. PODSUMOWANIE

Opracowany algorytm pozwala na uzyskiwanie zobrazowań SAR o wysokiej rozdzielczości w znacznie krótszym czasie w porównaniu do klasycznego algorytmu BPA. Przedstawione rozwiązanie jest łatwe do implementacji na platformach obliczeniowych wyposażonych w układy graficzne i może znaleźć zastosowanie w systemach pracujących w czasie rzeczywistym. Zmieniając szerokości wirtualnej wiązki syntetycznej apertury można wpływać na jakość zobrazowania i czas potrzebny do jego uzyskania.

LITERATURA

[1] S Basu and Y Bresler. Fast filtered hierarchical backprojection algorithm for tomography. *Submitted to the IEEE Trans. on Image Processing*.

[2] Zhaoqiang Bi, Jian Li, and Zheng-She Liu. Super resolution sar imaging via parametric spectral estimation methods. *IEEE Transactions on Aerospace and Electronic Systems*, 35(1):267–281, 1999.

[3] Ian G Cumming and Frank H Wong. Digital processing of synthetic aperture radar data. *Artech house*, 1(3), 2005.

[4] Giorgio Franceschetti and Riccardo Lanari. Synthetic aperture radar processing crc press. *Electronic engineering systems series*, 1999.

[5] LeRoy A Gorham, Uttam K Majumder, Peter Buxa, Mark J Backues, and Andrew C Lindgren. Implementation and analysis of a fast backprojection algorithm. In *Algorithms for Synthetic Aperture Radar Imagery XIII*, volume 6237, page 62370G. International Society for Optics and Photonics, 2006.

[6] Damian Gromek, Krzysztof Kulpa, and Piotr Samczy´nski. Experimental results of passive sar imaging using dvb-t illuminators of opportunity. *IEEE Geoscience and Remote Sensing Letters*, 13(8):1124–1128, 2016.

[7] Kai Hormann and N Sukumar. *Generalized barycentric coordinates in computer graphics and computational mechanics*. CRC Press, 2017.

[8] HL Li, J Li, YX Hou, L Zhang, MD Xing, and Z Bao. Synthetic aperture radar processing using a novel implementation of fast factorized back-projection. 2013.

[9] Krzysztof Radecki, Piotr Samczyński, and Damian Gromek. Fast barycentric-based back projection algorithm for sar imaging. *IEEE Sensors Journal*, 19(22):10635–10643, 2019.

[10] P Samczynski, D Gromek, J Drozdowicz, M Wielgo, K Klincewicz, M Malanowski, K Kulpa, M Nowakowski, J Krzonkalla, M Mordzonek, et al. Experimental results of daniel-35 sar system tests. In *2015 16th International Radar Symposium (IRS)*, pages 571–576. IEEE, 2015.

[11] Lars MH Ulander, Hans Hellsten, and Gunnar Stenstrom. Synthetic-aperture radar processing using fast factorized back-projection. *IEEE Transactions on Aerospace and electronic systems*, 39(3):760–776, 2003.

[12] M Wielgo, P Samczy´nski, M Malanowski, K Ndini, K Kulpa, and P Baranowski. The sarenka sar system-experimental results of isar imaging. In *2014 15th International Radar Symposium (IRS)*, pages 1–4. IEEE, 2014.

[13] Song Zhou, Lei Yang, Lifan Zhao, Yuhao Wang, Huilin Zhou, Liangbing Chen, and Mengdao Xing. A new fast factorized back projection algorithm for bistatic forward-looking sar imaging based on orthogonal elliptical polar coordinate. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(5):1508–1520, 2019.

WPŁYW IMPLEMENTACJI ALGORYTMU CHANA NA JEGO EFEKTYWNOŚĆ
INFLUENCE OF CHAN'S ALGORITHM IMPLEMENTATION ON ITS EFFECTIVENESS

Agnieszka Czapiewska[1]; Robert Nicpoń[2]

[1] Politechnika Gdańska, Gdańsk, agnieszka.czapiewska@pg.edu.pl
[2] Politechnika Gdańska, Gdańsk, s165164@student.eti.pg.edu.pl

DOI: 10.15199/59.2022.4.108

Streszczenie: **W referacie przybliżono algorytm Chana służący do wyznaczania położenia w metodzie TDOA. W algorytmie tym estymaty współrzędnych położenia można uzyskać w kilku różnych etapach wykonywania procedur algorytmu. Ponadto wątpliwość budzi sposób wypełniania macierzy B. W referacie zbadano dokładność uzyskiwanych estymat położenia w zależności od sposobu implementacji algorytmu Chana.**

Abstract: **The paper presents the Chan algorithm used for determining the position in the TDOA method. In this algorithm, estimates of position coordinates can be obtained at several different stages of the execution of the algorithm's procedures. Moreover, the method of filling the matrix B raises doubts. The paper investigates the accuracy of the obtained position estimates depending on the implementation of the Chan algorithm.**

Słowa kluczowe: **algorytm Chana, radio lokalizacja, TDOA.**

Keywords: **Chan algorithm, radio localization, TDOA.**

1. WSTĘP

Możliwość lokalizowania osób i obiektów jest istotnym elementem rozwijającego się społeczeństwa przebywającego w tzw. obszarach smart city. Wiedza o położeniu osób i obiektów sprzyja optymalizacji proce-sów zachodzących nie tylko w przemyśle [6], [7], ale również w codziennej egzystencji ludzi [2], [3]. Ponadto możliwość lokalizowania osób stanowi jeden z elementów usług z zakresu bezpieczeństwa publicznego [1].

Sposobów wyznaczania położenia w literaturze jest opisanych bardzo dużo. Do wyznaczenia położenia można wykorzystać m. in. informację wizualną [11], zliczeniową [5] lub użyć fal elektromagnetycznych. W przypadku, gdy używane są fale elektromagnetyczne mówimy o radio lokalizowaniu. Systemy radio lokalizacyjne dzielone są ze względu na typ pomiarów, które wykorzystują do określania położenia. Do podstawowych zaliczamy: TOA (*Time of Arrival*) [8], w których wykonywane są pomiary czasu propagacji; TDOA (*Time Difference of Arrival*) [9], w których mierzona jest różnica czasu propagacji; AoA (*Angle of Arrival*) [10], gdzie wyznaczany jest kąt nadejścia sygnału; RSS (*Received Signal Strength*) [10], czyli pomiar mocy sygnału odbieranego. W każdej z tych metod używane są różne algorytmy wyznaczania położenia. A każdy z algorytmów pozwala uzyskać różną dokładność lokalizowania, zależnie od warunków, w których wykonywany był dany typ pomiarów. W niniejszym referacie poddano analizie znany algorytm Chana opracowany dla metody TDOA, w którym przy liczbie stacji referencyjnych większej niż 3 można uzyskać współrzędne położenia obiektu na różnych etapach realizacji algorytmu. Porównano dokładność wyznaczonych w ten sposób estymat współrzędnych. Ponadto w algorytmie tym można dwoma sposobami otrzymać macierz $\boldsymbol{B}$, co także zostało przeanalizowane pod kątem dokładności uzyskiwanych estymat współrzędnych położenia obiektu. W punkcie 2 omówiono algorytm Chana, wyróżniając etapy działań, w których można wyznaczyć współrzędne położenia oraz wyjaśniając dwa możliwe sposoby wyznaczenia macierzy $\boldsymbol{B}$. W punkcie 3 zaprezentowano wyniki symulacji dla kilku scenariuszy działania systemu lokalizacyjnego. Na końcu podsumowano uzyskane wyniki.

2. ALGORYTM CHANA

Algorytm Chana został zaproponowany w artykule [4] do wyznaczania położenia obiektów w systemach hiperbolicznych. W systemach tych mierzona jest różnica odległości. Algorytm Chana jest przybliżoną realizacją metody największej wiarygodności (ang. *maximum likelihood*). W referacie [4] autor prezentuje sposoby wyznaczania pozycji dla dwóch przypadków: gdy system składa się z trzech stacji referencyjnych oraz gdy stacji referencyjnych jest więcej niż trzy. Algorytm prezentowany dla większej liczby stacji jest zdecydowanie bardziej złożony oraz stwarza możliwości różnego sposobu implementacji niektórych operacji. Dlatego też w niniejszym referacie omówione zostaną jedynie procedury wyznaczania położenia algorytmem Chana dla większej liczby stacji referencyjnych niż 3. Liczbę stacji referencyjnych oznaczmy symbolem M. Wówczas należy wypełnić macierze $\boldsymbol{h}$ i $\boldsymbol{G_a}$:

$$\boldsymbol{h} = \frac{1}{2}\begin{bmatrix} r_{2,1}^2 - K_2 + K_1 \\ r_{3,1}^2 - K_3 + K_1 \\ \vdots \\ r_{M,1}^2 - K_M + K_1 \end{bmatrix}, \tag{1}$$

$$\boldsymbol{G_a} = -\begin{bmatrix} x_{2,1} & y_{2,1} & r_{2,1} \\ x_{3,1} & y_{3,1} & r_{3,1} \\ \vdots & \vdots & \vdots \\ x_{M,1} & y_{M,1} & r_{M,1} \end{bmatrix}, \tag{2}$$

Gdzie:

$$K_i = x_i^2 + y_i^2, \tag{3}$$

$$r_{i,1} = r_i - r_1, \tag{4}$$

$$x_{i,1} = x_i - x_1, \tag{5}$$

$$y_{i,1} = y_i - y_1. \tag{6}$$

Kolejnym krokiem jest aproksymacja wektora $\boldsymbol{z_a} = \begin{bmatrix} \boldsymbol{z_p} \\ r_1 \end{bmatrix}$, gdzie $\boldsymbol{z_p} = \begin{bmatrix} x \\ y \end{bmatrix}$, który pozwala na przybliżone określenie położenia obiektu oraz możliwość dalszych iteracji szacowania lokalizacji. Wstępną aproksymację wektora $\boldsymbol{z_a}$ otrzymujemy poprzez rozwiązanie równania:

$$\boldsymbol{z_a} \approx (\boldsymbol{G_a}^T \mathbf{Q}^{-1} \boldsymbol{G_a})^{-1} \boldsymbol{G_a}^T \mathbf{Q}^{-1} \boldsymbol{h}, \tag{7}$$

gdzie $\mathbf{Q}$ oznacza macierz kowariancji wektora szumu błędów pomiarów. Chan w swoim artykule pisze, że aby rozwiązać powyższe równanie nieliniowe należy przyjąć, że x, y a r_1 są wzajemnie niezależne. Wzór (7) może posłużyć do uzyskania pierwszego wyniku współrzędnych położenia obiektu i posłużyć do wyznaczenia macierzy $\boldsymbol{B}$, która teoretycznie powinna być wypełniona prawdziwymi wartościami:

$$\boldsymbol{B} = diag\{r_2^0, r_3^0, \dots, r_M^0\}. \tag{8}$$

Indeks górny „0" oznacza idealne wartości. Macierz $\boldsymbol{B}$ może być wyznaczona na podstawie pierwszej estymaty wektora $\boldsymbol{z_a}$. Przy czym można tę operację zrealizować dwoma sposobami:

$$r_i^0 = r_{i,1} - r_1 \tag{9}$$

lub

$$r_i^0 = r_{i,1} - \sqrt{(x_i - x_1)^2 + (y_i - y_1)^2}. \tag{10}$$

Znając macierz $\boldsymbol{B}$ można wyznaczyć macierz kowariancji wektora błędu:

$$\boldsymbol{\Psi} = c^2 \boldsymbol{BQB}. \tag{11}$$

Wówczas można wyznaczyć współrzędne poszukiwanej pozycji z większą dokładnością ze wzoru:

$$\boldsymbol{z_a} = \left(\boldsymbol{G_a}^T \boldsymbol{\Psi}^{-1} \boldsymbol{G_a}\right)^{-1} \boldsymbol{G_a}^T \boldsymbol{\Psi}^{-1} \boldsymbol{h}. \tag{12}$$

W kolejnej aproksymacji metodą MLE (ang. *Maximum Likelihood Estimation*) uzyskuje się wyrażenie:

$$\boldsymbol{z_a}' = \left(\boldsymbol{G_a}'^T \boldsymbol{\Psi}'^{-1} \boldsymbol{G_a}'\right)^{-1} \boldsymbol{G_a}'^T \boldsymbol{\Psi}'^{-1} \boldsymbol{h}', \tag{13}$$

gdzie:

$$\boldsymbol{h}' = \begin{bmatrix} (z_{a,1} - x_1)^2 \\ (z_{a,2} - y_1)^2 \\ z_{a,3}^2 \end{bmatrix}, \tag{14}$$

$$\boldsymbol{G_a}' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \tag{15}$$

$$\boldsymbol{\Psi}' = 4\boldsymbol{B}' \boldsymbol{cov}(\boldsymbol{z_a}) \boldsymbol{B}', \tag{16}$$

$$\boldsymbol{B}' = diag\{x^0 - X_1,\ \ y^0 - Y_1,\ \ R_1^0\}. \tag{17}$$

Ostatecznie współrzędne położenia obiektu wyznacza się następująco:

$$\boldsymbol{z_p} = \pm\sqrt{\boldsymbol{z_a}'} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}. \tag{18}$$

Właściwym rozwiązaniem jest to, które leży w obszarze zainteresowania.

Z powyższych rozważań wynika, że w algorytmie Chana zależności (7), (12) i (18) stanowią kolejne estymaty położenia lokalizowanego obiektu.

W publikacji Chana [4] wzór (7) jest oznaczony numerem 14b, a wzór (12) oznaczono numerem 14a.

3. BADANIA SYMULACYJNE

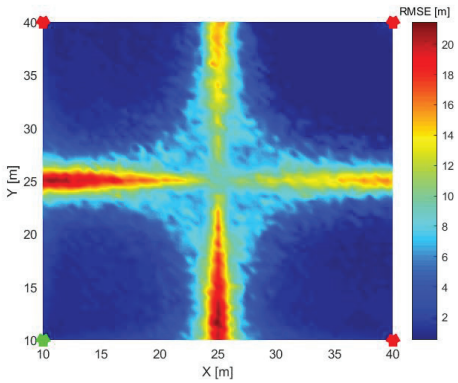
Zbadanie efektywności lokalizowania poszczególnych etapów algorytmu Chana oraz sposobu wypełniania macierzy $\boldsymbol{B}$, wymagało przeprowadzenia symulacji. Badania zrealizowano dla trzech scenariuszy rozstawienia czterech stacji referencyjnych: w kwadracie (współrzędne stacji referencyjnych: (10, 10), (10, 40), (40, 10), (40, 40)), czworokącie ((10, 10), (15, 30), (30, 15), (40, 40)) oraz trójkącie ((10, 10), (25, 10), (40, 10), (25, 40)). Rozstawienie stacji referencyjnych w kwadracie zaprezentowano graficznie na rys. 1, 2, w czworokącie na rys. 4, 5, w trójkącie na rys. 7, 8, gdzie stacje referencyjne oznaczono czerwonymi wielokątami, a wielokątem zielonym wyróżniono stację referencyjną, względem której mierzona była różnica czasu propagacji. Badania symulacyjne zawsze realizowano w obszarze ograniczonym przez kwadrat o przeciwległych wierzchołkach znajdujących się w punktach (10, 10) i (40, 40). W obszarze tym wyznaczono punkty pomiarowe z krokiem 0,5 m, w których symulowano 100 pomiarów różnicy odległości z błędem pomiaru o wariancji 0,5 m. Błąd RMSE (*Root Mean Square Error*) na wymienionych rysunkach przedstawiono jako mapę cieplną, gdzie kolor czerwony oznacza największe wartości błędów, natomiast niebieski najmniejsze. W celu wygodniejszego porównania efektywności poszczególnych sposobów wyznaczenia współrzędnych położenia lokalizowanego obiektu wyznaczono estymaty dystrybuanty średniokwadratowych błędów określania położenia, które zaprezentowano na rys. 3, 6 i 9 odpowiednio dla scenariuszy kwadratowego, czworokątnego i trójkątnego. Przyjęto następujące oznaczenia implementacji algorytmu:

- Chan v1 – użyto wzoru (9), wynik końcowy uzyskano z wzoru (18);
- Chan 14a v1 – użyto wzoru (9), wynik końcowy uzyskano z wzoru (12);
- Chan 14b v1 - użyto wzoru (9), wynik końcowy uzyskano z wzoru (7);
- Chan v2 – użyto wzoru (10), wynik końcowy uzyskano z wzoru (18).

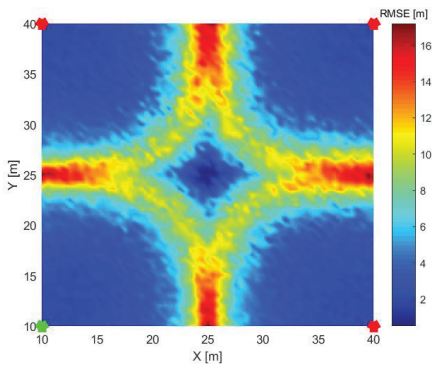
Na rys. 1, 4 i 7 zaprezentowano rozkład błędów RMSE na płaszczyźnie dla algorytmu Chan v1, natomiast na rys. 2, 5 i 8 dla Chan 14a v1. Nie prezentowano rozkładów błędów dla pozostałych wersji algorytmu Chana, ponieważ są one bardzo podobne do tych już zaprezentowanych, o czym świadczą wykresy dystrybuant (rys. 3, 6, 7).

Analizując uzyskane wyniki stwierdzono, że nieznacznie mniejsze błędy uzyskano, gdy macierz $\boldsymbol{B}$ wyznaczano z użyciem wzoru (10). Jednakże różnicę tę można zaobserwować tylko na rys. 3. Ponadto porównując rys.

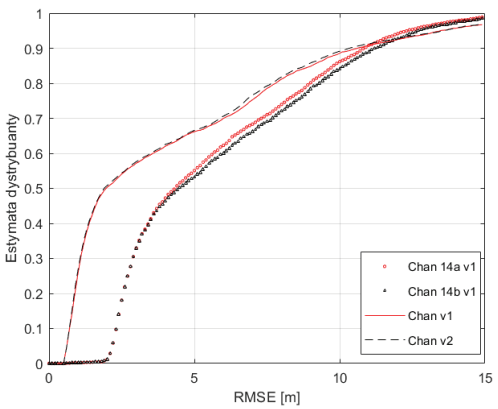
3, 6 i 9 stwierdzono, że uporządkowane rozstawienie stacji referencyjnych w przypadku algorytmu Chana jest niewskazane. Największe dokładności estymowanych współrzędnych uzyskano, gdy stacje referencyjne rozstawiono w czworokącie, a najmniejsze, gdy były rozstawione w kwadracie (tab. 1). Przeprowadzone badania dowodzą również, że zgodnie z przewidywaniami, estymaty współrzędnych położenia uzyskanych z wzoru (7) – w artykule [4] (14b), dają nieznacznie większe błędy niż te uzyskane z wzoru (12) – w artykule [4] (14a). Szczególnie



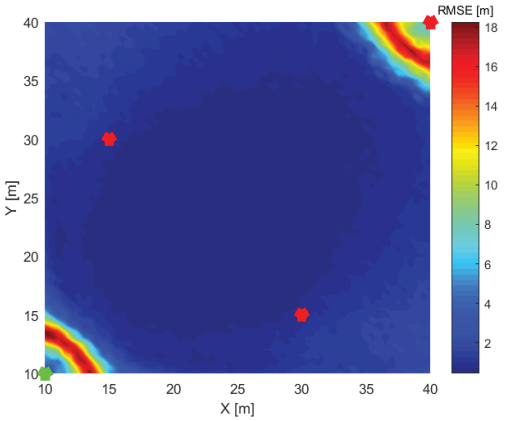
Rys. 1. Rozkład RMSE dla algorytmu Chana v1, stacje referencyjne rozstawione w wierzchołkach kwadratu



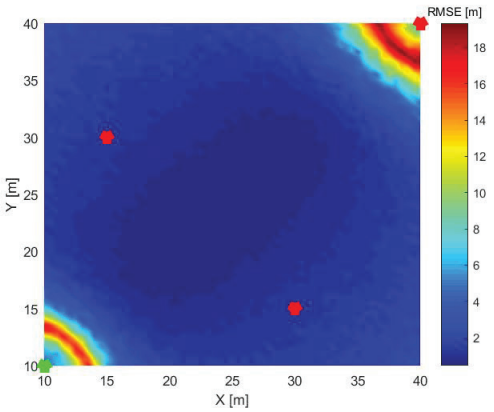
Rys. 2. Rozkład RMSE dla algorytmu Chana v1 14a, stacje referencyjne rozstawione w wierzchołkach kwadratu



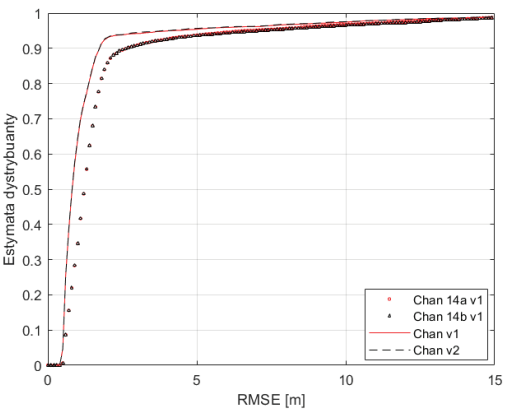
Rys. 3. Estymaty dystrybuanty RMSE, stacje referencyjne rozstawione w wierzchołkach kwadratu



Rys. 4. Rozkład RMSE dla algorytmu Chana v1, stacje referencyjne rozstawione w wierzchołkach czworokąta

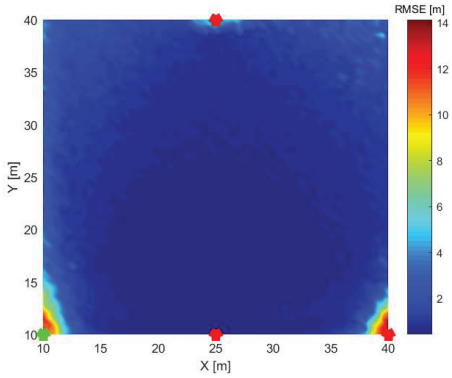


Rys. 5. Rozkład RMSE dla algorytmu Chana v1 14a, stacje referencyjne rozstawione w wierzchołkach czworokąta

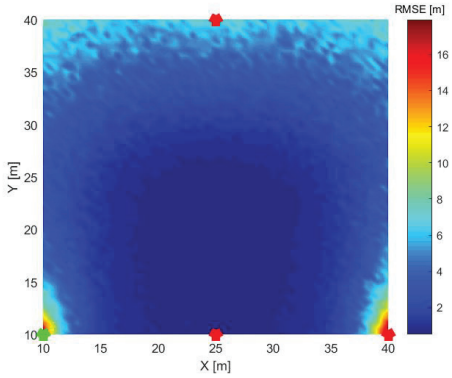


Rys. 6. Estymaty dystrybuanty RMSE, stacje referencyjne rozstawione w wierzchołkach czworokąta

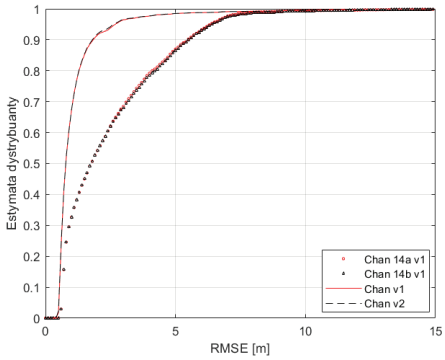
jest to widoczne przy niekorzystnym rozstawieniu stacji referencyjnych (rys. 3). Natomiast zawsze najmniejsze wartości błędów uzyskano z końcowego wzoru (18).



Rys. 7. Rozkład RMSE dla algorytmu Chana v1, stacje referencyjne rozstawione w trójkącie



Rys. 8. Rozkład RMSE dla algorytmu Chana v1 14a, stacje referencyjne rozstawione w trójkącie



Rys. 9. Estymaty dystrybuanty RMSE, stacje referencyjne rozstawione w trójkącie

Tab. 1. Błąd RMSE dla 90 % przypadków wyznaczonych współrzędnych ze wzoru (18)

Układ stacji referencyjnych	*RMSE [m]*
Kwadrat	10,4
Czworokąt	1,8
Trójkąt	1,9

Ponadto, należy zwrócić również uwagę, że w każdej z konfiguracji stacji referencyjnych istnieją obszary, w których uzyskane estymaty współrzędnych położenia są obarczone nawet dziesięciokrotnie większymi błędami niż w pozostałych obszarach. Dlatego też projektując system lokalizacyjnych oparty na metodzie TDOA i wykorzystujący algorytm Chana konieczne jest odpowiednie rozstawienie stacji referencyjnych, tak by obszar działania systemu pokrywał się z obszarami, gdzie błędy określania położenia są najmniejsze.

4. PODSUMOWANIE

W referacie zaprezentowano analizę algorytmu Chana dedykowanego do wyznaczania położenia w metodzie TDOA. Przeanalizowano sposób wyznaczania macierzy ***B*** i stwierdzono, że nie ma on dużego znaczenia dla efektywności całego algorytmu. Ponadto, potwierdzono, że wraz z kolejnymi aproksymacjami uzyskiwane współrzędne położenia obiektu cechują mniejsze błędy. Zauważono także, że dla algorytmu Chana, rozmieszczenie stacji referencyjnych ma bardzo duże znaczenie dla uzyskiwanych wartości błędów wyznaczania położenia.

LITERATURA

[1] Ansari A. R., Nasir S., Haq M., Cho S. 2018. "Accu-rate 3D Localization Method for Public Safety Ap-plications in Vehicular Ad-Hoc Networks". IEEE Access 6.

[2] Badawik A., Kołakowski J. 2015. „Zmodyfikowana metoda lokalizacji nadajnika z wykorzystaniem techniki RSS". Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne 4. ISSN: 1230-3496.

[3] Boreiko O., Teslyuk V., Beregovska C., Mykhailiuk A. 2015. „Model of Telecommunication Networks for Intelligent Building". 13th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM).

[4] Chan Y. T., Ho K. C. 1994. "A Simple and Efficient Estimator for Hiperbolic Location". IEEE Trans. on Signal Proc. 42 (8).

[5] Fang S., Li H., Yang M., Wang Z. 2022. "Inertial Navigation System Based Vehicle Temporal Relative Localization With Split Covariance Intersection Filter". IEEE Robotics and Automation Letters 7 (2).

[6] https://ubisense.com/other-industries/. dostęp: 5.05.2022.

[7] https://www.teldio.com/applications/rtls/. dostęp: 5.05.2022.

[8] Krishnaveni B. V., Reddy K. S., Reddy R. R. 2021. "An Introduction to the TOA measurement for UWB indoor localization Systems". 5th Conference on Information and Communication Technology (CICT).

[9] Wang K., Chen Z., Yan Q. 2021. "Research on Multi-platform Time Difference of Arrival and Frequency Difference of Arrival Joint Location Technology". IEEE 6th International Conference on Signal and Image Processing (ICSIP).

[10] Stefański J. 2021. „Badanie metod i projektowanie usług lokalizacyjnych w sieciach radiokomunikacyjnych". Wydawnictwo Politechniki Gdańskiej. Gdańsk. ISBN: 978-83-7348-417-7.

[11] Zhang T., Johnson-Roberson M. 2022. "Learning Cross-Scale Visual Representations for Real-Time Image Geo-Localization". IEEE Robotics and Automation Letters 7 (2).

MODELOWANIE ODDZIAŁYWANIA ZABURZEŃ HPM NA OBIEKTY IK

MODELLING THE IMPACT OF HPM DISTURBANCES ON CI OBJECTS

Piotr Słobodzian[1]

DOI: 10.15199/59.2022.4.109

[1] Politechnika Wrocławska, Wrocław, piotr.slobodzian@pwr.edu.pl

Streszczenie: **W artykule opisano ogólne zasady modelowania oddziaływania zaburzeń HPM na obiekty infrastruktury krytycznej (IK). Scharakteryzowano znane narzędzia wykorzystywane do modelowania skutków oddziaływania HPM. Następnie omówiono elementy modelowania występujące na pierwszym i drugim poziomie tzw. piramidy modelowania i symulacji skutków oddziaływania HPM.**

Abstract: **This paper describes the general principles of modeling the impact of HPM disturbances on objects of the critical infrastructure (IC objects). Current tools used to model the effects of HPM have been characterized. Then, the elements of modeling that belong to the first and second levels of the so-called modeling and simulation pyramids of the HPM impact have been discussed.**

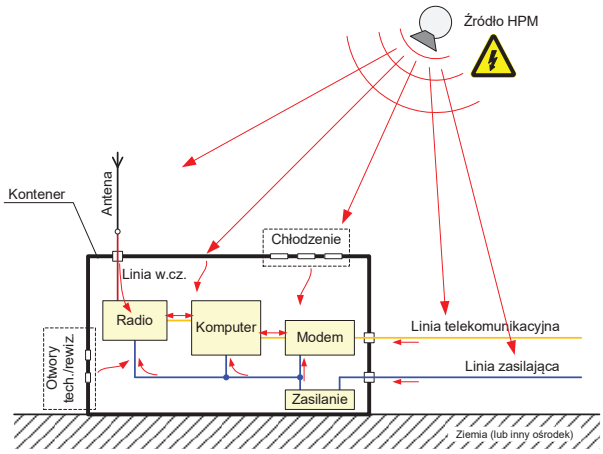
Słowa kluczowe: **impulsy HPM, modelowanie elektromagnetyczne, zaburzenia elektromagnetyczne.**

Keywords: **electromagnetic disturbances, electromagnetic modeling, HPM pulses.**

1. WSTĘP

Zagadnienie przewidywania skutków oddziaływania zaburzeń elektromagnetycznych na obiekty zawierające urządzenia elektroniczne ma znaczenie w wielu obszarach działalności człowieka. W ostatnich dwóch dekadach można zauważyć wyraźny wzrost świadomości decydentów w zakresie zagrożeń ze strony intencjonalnie wytwarzanych zaburzeń EM [1]. Badania w zakresie technik generowania takich zaburzeń oraz technik ochrony i obrony przed nimi wyraźnie przyspieszyły na całym świecie (w Polsce również [2]-[3]). Skuteczna ochrona przed intencjonalnymi zaburzeniami EM wymaga nie tylko świadomości istnienia takich zagrożeń, ale również skutecznego przewidywania skutków, które mogą one wywołać w różnych scenariuszach. W dalszej części skupimy się na oddziaływaniu zaburzeń EM w postaci impulsów HPM [4].

Ocena odporności obiektów z istotnymi komponentami infrastruktury krytycznej (np. takimi jak systemy łączności) na zaburzenia HPM powinna bazować na odpowiedniej metodyce modelowania i symulacji (M+S). Przykładowy komponent infrastruktury systemu łączności pokazano na rys. 1. Efektem działania impulsów HPM na elementy składowe komponentu jest uszkodzenie lub tymczasowa degradacja parametrów użytkowych tych elementów. Biorąc pod uwagę znane z literatury metody stosowania intencjonalnych zaburzeń HPM [1]-[3], ocena odporności obiektów infrastruktury powinna być przeprowadzana w różnych scenariuszach użycia tych zaburzeń.



Rys. 1. Przykładowy element infrastruktury systemu łączności – kontener telekomunikacyjny z typowymi urządzeniami teleinformatycznymi

Istnieje wiele narzędzi komercyjnych do modelowania, projektowania i optymalizacji systemów łączności oraz oceny kompatybilności międzysystemowej (np. patrz oprogramowanie firmy Altari® Feko®, Altair® WinProp™ i Altair® WRAP™ [5]). Niestety, według wiedzy autora, dostępne na rynku oprogramowanie nie umożliwia dedykowanej analizy oddziaływania zaburzeń HPM na obiekty infrastruktury. Oprogramowanie takie jest natomiast stosowane przez siły zbrojne i agencje bezpieczeństwa państwowego. Niestety informacje o jego możliwościach są mocno ograniczone lub wręcz niedostępne. W następnej części artykułu krótko opisano takie narzędzia w oparciu o publicznie dostępne informacje. Dalej przedstawiono elementy modelowania oddziaływania zaburzeń HPM, opracowane w ramach projektu realizowanego na rzecz obronności i bezpieczeństwa naszego państwa [6].

2. METODY I NARZĘDZIA MODELOWANIA ODDZIAŁYWANIA ZABURZEŃ HPM

Pierwsze metody ochrony urządzeń elektronicznych przed oddziaływaniem zaburzeń elektromagnetycznych o dużej energii zostały opracowane w odniesieniu do impulsów typu NEMP i LEMP. Metodyka ochrony jest rozwijana nieustannie od końca II Wojny Światowej. Pełną wiedzę na temat skutków oddziaływania impulsów NEPM (HEMP) uzyskano na początku lat 60 ubiegłego wieku (USA i ZSRR) [1]. Świadomość realnego zagrożenia systemów elektronicznych ze strony celowo wytwarzanych impulsów mikrofalowych o dużej mocy (HPM)

zrodziła się w latach 80. ubiegłego wieku. W wyniku tego, na przełomie lat 80. i 90. Departament Obrony USA opracował metodykę oceny podatności urządzeń i systemów elektronicznych na impulsy HPM [7]. W tym samym czasie biuro do spraw badań naukowych Sił Powietrznych USA uruchomiło multidyscyplinarny program badawczy dotyczący źródeł mikrofalowych dużej mocy (program MURI) [8]. Od połowy lat 90. Departament Obrony USA dokonuje regularnych analiz i oceny stopnia rozwoju różnych technologii związanych z budową broni opartej na energii skierowanej oraz oceny stopnia powszechności i dostępności tych technologii na świecie [1], [9]-[10]. Jednocześnie Departament Obrony USA przeznacza znaczne środki finansowe na rozwój broni opartej na energii skierowanej, rozwój metod ochrony i obrony przed atakiem przy użyciu takiej broni oraz rozwój narzędzi do przewidywania i oceny skutków użycia energii skierowanej zarówno w scenariuszach militarnych jak i cywilnych.

Opracowana blisko 30 lat temu metodyka oceny podatności urządzeń i systemów elektronicznych na impulsy HPM była stosunkowo prosta [7]. Opierała się na uproszczonych metodach analizy i modelowania propagacji impulsów HPM i ich oddziaływania na urządzenia elektroniczne [11]. Obecnie stosowana przez Siły Zbrojne USA metodyka jest dużo bardziej zaawansowana. Obejmuje szersze spektrum zagadnień i wykorzystuje bardziej zaawansowane metody i narzędzia. Metodyka ta jest oparta na czteropoziomowej piramidzie modelowania i symulacji skutków oddziaływania impulsów HPM [11]-[12], przy czym podstawę piramidy tworzą deterministyczne metody modelowania zjawisk fizycznych (oddziaływanie HPM na elementy, podzespoły i urządzenia elektroniczne). Praktyczne wykorzystanie TEJ metodyki opiera się obecnie na komputerowych platformach/modułach symulacyjnych, które w zależności od poziomu modelowania i symulacji pobierają/generują stosowane dane/wyniki. Poniżej przedstawiono krótką charakterystykę narzędzi komputerowych wykorzystywanych przez różne rodzaje sił zbrojnych USA (według wiedzy autora są to jedyne upublicznione informacje).

Pierwszym znanym narzędziem było programowanie DREM, opracowane w połowie lat 90. ubiegłego wieku na zlecenie Wojsk Lądowych USA (oprogramowanie I generacji). DREAM jest akronimem nazwy Directed Radio Frequency Energy Assessment Model. Jest to pierwsze znane oprogramowanie, w którym zaimplementowano metodykę oceny podatności sprzętu wojskowego na zaburzenia HPM, opracowaną przez Departament Obrony USA [7]. Moduł DREAM oferował/oferuje jedynie bardzo prostą analizę propagacyjną (wykorzystuje model wolnej przestrzeni). Obszerniejsza charakterystyka modułu została przedstawiona w [11]. Kolejne oprogramowanie to RF-PROTEC. Oprogramowanie powstało prawdopodobnie również w połowie lat 90. ubiegłego wieku na zlecenie Sił Powietrznych (Lotnictwa) USA. Moduł RF-PROTEC charakteryzuje wyższy poziom zaawansowania i z tego powodu można go traktować jako oprogramowanie wyższej generacji (pomiędzy I i II). RF-PROTEC jest akronimem nazwy RF Propagation and Target Effects Code. Oprogramowanie zostało opracowane w celu szacowania skuteczności użycia broni HPM podczas ataku infrastruktury sieci komputerowych z powietrza. Moduł RF-PROTEC został wyposażony w bardziej zaawansowane modele źródeł HPM oraz modele do analizy propagacji zaburzenia HPM (np. metodę GTD – propagacja wewnątrz budynków) [11]. Przez kolejne dwie dekady było rozwijane następne oprogramowanie, tj. JREM (lata 2000-2010). Jest to II generacja oprogramowania [11]. Akronim JREM pochodzi od nazwy Joint RF Effectiveness Model. Moduł JREM łączy najlepsze cechy dwóch poprzednich modułów i dodatkowo może być uruchamiany na komputerze osobistym. Modele źródeł HPM oraz modele propagacyjne wykorzystywane w module zostały eksperymentalnie zweryfikowane w Laboratoriach Badawczych Sił Powietrznych USA.

Najnowsze i obecnie najbardziej zaawansowane oprogramowanie do oceny skutków oddziaływania zaburzeń HPM nosi nazwę HPM LAVA. Jest to pełny moduł drugiego poziomu piramidy M+S, ponieważ służy głównie do modelowani i symulacji efektywności sprzężenia zaburzenia HPM z obiektem/celem ataku. Oprogramowanie powstało w latach 2012-2016 na zamówienie Marynarki Wojennej USA. Akronim LAVA pochodzi od nazwy Lethality and Vulnerability Assessment, co można tłumaczyć jako szacowanie stopnia śmiercionośności i podatności. Jest to III generacja oprogramowania i wykorzystuje zaawansowane modele symulacyjne dla źródeł HPM oraz środowiska propagacyjnego i ścieżek wnikania zaburzenia HPM do skomplikowanych, zamkniętych obiektów. Obecnie HPM LAVA jest najbardziej wiarygodnym narzędziem modelowania i symulacji sprzężenia zaburzeń HPM z obiektami/platformami wojskowymi w USA [11].

W naszym kraju również stosowane są narzędzia do symulacji działań na polu walki oraz sytuacji kryzysowych na wyższych poziomach piramidy M+S. Jako przykład można tu wymienić system symulacyjny JTLS (Joint Theater Level Simulation), wykorzystywany w Centrum Symulacji i Komputerowych Gier Wojennych w Akademii Obrony Narodowej [13]-[14]. Niestety wymieniony system nie umożliwia oceny zagrożenia i skutków związanych z celowym użyciem zaburzeń elektromagnetycznych.

3. ELEMENTY MODELOWANIA ODDZIAŁYWANIA ZABURZEŃ HPM

Osiągniecie pełniej elastyczności w ocenie skutków oddziaływania zaburzeń HPM na obiekty wymaga modelowania na pierwszym i drugim poziomie piramidy M+S.

3.1. Pierwszy poziom piramidy M+S

W tym przypadku modelowanie musi mieć ścisły związek z ogólnym schematem oddziaływania zaburzenia HPM na urządzenia i systemy elektroniczne. Model pełnej ścieżki oddziaływania impulsów HPM powinien obejmować cztery obszary, a mianowicie [11]-[12], [15]: źródła impulsów HPM; propagację zaburzeń HPM w kierunku obiektu; charakterystykę samego obiektu (jego podatność na zaburzenia HPM); ocenę skutków oddziaływania zaburzeń HPM na elementy składowe obiektu (tj. elementy tworzące tzw. ścieżki oddziaływania). Wymienione obszary modelowania zilustrowano na schemacie blokowym przedstawionym na rys. 2.

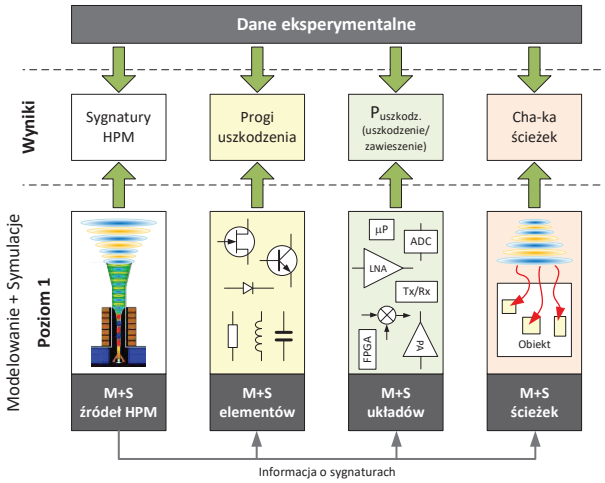
Modelowanie źródeł impulsów HPM ma na celu określenie parametrów impulsów pojawiających się na wyjściu źródła w zależności od zakresu częstotliwości, typu źródła, zadanej mocy wyjściowej, itd. Charakteryzacja impulsów HPM powinna być przeprowadzona zarówno w dziedzinie czasu jak i częstotliwości. Ponieważ źródła impulsów HPM są zwykle zintegrowane z określonego typu antenami, to opis parametrów źródła może obejmować również parametry polowe (falowe) zastosowanej anteny.

Z kolei modelowanie propagacji impulsów HPM ma na celu określenie parametrów zaburzenia HPM w obszarach, miejscach lub punktach ich wejścia do obiektów. W tym miejscu należy zaznaczyć, że punkty wejścia są cechą charakterystyczną danego obiektu/celu i są charakteryzowane w oderwaniu od propagacji. Natomiast sam model propagacyjny uwzględnia elementy scenariusza wystąpienia zaburzenia HPM.

Jeśli idzie o charakteryzację obiektu to ma ona na celu określenie krytycznych punktów wejścia zaburzenia HPM do jego wnętrza oraz krytycznych ścieżek propagacji tego zaburzenia we wnętrzu obiektu do krytycznych obwodów i elementów elektronicznych. Jednym z najtrudniejszych zagadnień w modelowaniu jest opisanie stopnia sprzężenia punktów wejścia zaburzenia HPM z krytycznymi ścieżkami jego propagacji wewnątrz obiektu. Charakteryzacja obiektu powinna również określać związek punktów wejścia i ścieżek propagacji zaburzenia HPM z krytycznymi funkcjami obiektu w odniesieniu do konkretnego przeznaczenia obiektu. W tym miejscu należy podkreślić, że wymieniony obszar modelowania dotyczy ilościowego określania zjawisk fizycznych i wynikających z nich progów podatności/odporności krytycznych elementów obiektu.

Ostatni etap modelowania dotyczy skutków oddziaływania impulsów HPM na cały obiekt. Określenie wypadkowej podatności/odporności obiektu na zaburzenie HPM powinno się odbywać w odniesieniu do konkretnego scenariusza wystąpienia danego zaburzenia. Nie należy tu mylić wymienionej wcześniej podatności/odporności obiektu z jego podatnością/odpornością określoną w warunkach laboratoryjnych według określonej normy. Generalnie normy dotyczą szczególnego scenariusza badań i w praktyce mogą nie uwzględniać rzeczywistych przypadków oddziaływania zaburzenia HPM na dany obiekt. O wpadkowych skutkach oddziaływania decyduje scenariusz wystąpienia zaburzenia HPM oraz charakterystyka obiektu i jego bieżący stan. Ze względu na losową zmienność danych wejściowych do tego obszaru modelowania, wynik końcowy powinien mieć postać łącznego warunkowego prawdopodobieństwa wyeliminowania obiektu z pełnionej roli/funkcji w ramach infrastruktury.

Na zakończenie należy podkreślić, że opisane powyżej metody i zasady powinny również uwzględniać wyniki badań eksperymentalnych. Metodyka musi umożliwiać włączanie do opisu obiektów wyników eksperymentalnych, uzyskanych w laboratoriach lub na poligonach doświadczalnych (w ten sposób modele obiektów są automatycznie poddane praktycznej weryfikacji). Dopiero taki zestaw wyników modelowania i symulacji



Rys. 2. Ogólny schemat blokowy obszarów modelowania na pierwszym poziomie piramidy M+S

może służyć do modelowania skutków oddziaływania zaburzeń HPM na drugim poziomie piramidy M+S.

3.2. Drugi poziom piramidy M+S

Jak już wcześniej wspomniano, na drugi poziom nie są przekazywane żadne modele fizyczne ani żadne skomplikowane metody analizy. Modelowanie opiera się na wynikach uzyskanych na pierwszym poziomie i wykorzystuje np. tablice parametryczne lub uproszczone metody analizy z niezbędnymi danymi, np. elementami geometrii obiektów. Zasady modelowania ograniczają się tu głównie do użycia sparametryzowanych modeli zastępczych. Na tym etapie wykorzystuje się predefiniowane dla danego obiektu ścieżki wnikania zaburzenia HPM, które składają się z kaskady sparametryzowanych modeli zastępczych, począwszy od źródła aż po krytyczne elementy elektroniczne obiektu. To właśnie te elementy decydują o głównych cechach funkcjonalnych obiektów i/lub urządzeń. Parametryzacja może obejmować urządzenia, całe obiekty oraz schematy i metody obliczeniowe. Jej celem jest umożliwienie oceny wpływu zaburzenia HPM przy zmiennych (dynamicznych) warunkach scenariusza oddziaływania. Sparametryzowane modele zastępcze powinny się opierać na: analizie funkcjonalnej obiektów w odniesieniu do funkcji pełnionej w infrastrukturze krytycznej; identyfikacji elementów i urządzeń elektronicznych, krytycznych ze względu na rolę obiektu pełniona w infrastrukturze krytycznej; określeniu krytycznych ścieżek propagacji zaburzenia HPM wewnątrz obiektu lub urządzenia; określeniu progów podatności i prawdopodobieństwa uszkodzenia lub zmiany parametrów roboczych krytycznych elementów elektronicznych; określeniu wypadkowego prawdopodobieństwa uszkodzenia lub zmiany parametrów roboczych całego obiektu/urządzenia w zależności od scenariusza wystąpienia zaburzenia HPM. Opisane zasady zilustrowano na rys. 3.

3.3. Ocena skutków odziaływania HPM

Ocena skutków odziaływania HPM na dany obiekt infrastruktury może się opierać na różnych skalach. W ogólnym przypadku eliminacja obiektu z roli pełnionej w infrastrukturze krytycznej nie musi się wiązać z całkowitym uszkodzeniem obiektu. Spowodowanie czasowej

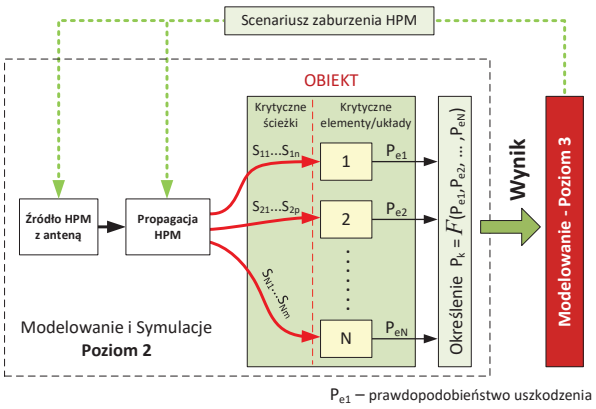
niezdolności obiektu do pełnienia dedykowanej funkcji może również skutkować dotkliwymi stratami. Takie zdarzenie może się pojawić w skrajnie innych scenariuszach zaburzenia HPM niż te, które są wymagane do całkowitego uszkodzenia obiektu. Docelową ocenę skutków odziaływania impulsów HPM można przeprowadzać wg. skali zaproponowanej w pracy [12] (patrz rys. 4). Skala zastała opracowana w celu ujednolicenia oceny skutków oddziaływania impulsów HPM pomiędzy różnymi organizacjami standaryzującymi.

4. PODSUMOWANIE

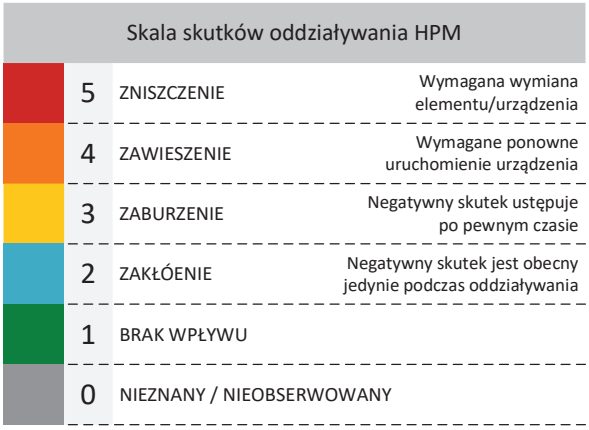
Skuteczność opisanej metodyki M+S w przewidywaniu skutków oddziaływania zaburzeń HPM na obiekty IK zależy przede wszystkim od dokładności modeli opisujących poszczególne ścieżki oddziaływania impulsów HPM. Modelowanie samych ścieżek jest odrębnym zagadnieniem, które może być rozwiązane za pomocą oprogramowania do analizy obwodowej i pełnofalowej. Lepszym rozwiązaniem jest jednak, podobnie jak w przypadku metodyki stosowanej prze Siły Zbrojne USA, implementacja wybranych modeli z pierwszego i drugiego poziomu piramidy M+S w dedykowanych symulatorach.

Jeśli idzie o same modele ścieżek oddziaływania impulsów HPM, to muszą być one spójne i kompatybilne z modelami zaburzeń HPM. Muszą też umożliwiać przekazywanie odpowiedzi pomiędzy modelami składowymi tworzącymi daną ścieżkę oddziaływania zaburzenia HPM. Przedstawienie tego zagadnienia wykracza jednak daleko poza ramy niniejszego artykułu i będzie przedmiotem kolejnych publikacji.

Artykuł zawiera wyniki realizacji projektu finansowanego przez NCBR (DOB-BIO2/PS/5/2/2016).



Rys. 3. Ogólny schemat blokowy elementów modelowania na drugim poziomie piramidy M+S



Rys. 4. Przykładowa skala oceny skutków oddziaływania impulsów HPM [12]

LITERATURA

[1] Wilson C. 2008. „High Altitude Electromagnetic Pulse (HEMP) and High Power Microwave (HPM) Devices: Threat Assessments". *CRS report for Congress*, RL32544, Congressional Research Service, The Library of Congress, USA.

[2] Szubrycht T., Szymański T. 2005. „Broń elektromagnetyczna jako nowy środek walki w erze informacyjnej". *Zeszyty naukowe Akademii Marynarki Wojennej*, AMW, Gdynia.

[3] Wnuk M., Matuszewski J., Chudy Z. 2015. „Nowe technologie i urządzenia rażenia elektromagnetycznego w dziedzinie walki elektronicznej". *Przegląd Elektrotechniczny*, 91 (3) : 92–95.

[4] Benford J., Swegle J., Schamiloglu E. 2015. „High Power Microwaves". CRC Press, Boca Raton.

[5] Altair Feko Applications. Opis na stronie internetowej: https://www.altair.com/feko-applications/.

[6] Projekt. „Badania i symulacje skutków oddziaływania impulsów HPM". DOB-BIO2/PS/5/2/2016 (NCBR). Politechnika Wrocławska (Lider).

[7] Chesser N., Pesta A. 1989. „Department of Defense Methodology Guidelines for High Power Microwave (HPM) Susceptibility Assessments". *HPM Methodology Panel Report, Draft*. Office of the Secretary of Defense, U.S.

[8] Abrams M. 2003. „Dawn of the E-Bomb". *IEEE Spectrum*, 40 (11) : 24–30.

[9] United States. 1998. „Radio frequency weapons and proliferation: potential impact on the economy". *Hearing before the Joint Economic Committee*. Congress of the United States, One Hundred Fifth Congress. Second session. U.S. Government Printing Office, Washington.

[10] Thompson L., Goure D. 2003. „Directed-Energy Weapons: Technologies, Applications and Implications". *Directed-Energy Program*. Lexington Institute, VA, USA.

[11] Tatum J. 2014. „High-Power Microwave Directed Energy Weapons: A Model and Simulation Toolbox". *DSIAC Journal*, 1 (2) : 20–24.

[12] Tatum J. 2017. „HPM DEWs and Their Effects on Electronic Targets". *DSIAC Journal*, 4 (3) : 33–42.

[13] Pietrzak M. 2014. „Wykorzystanie systemu symulacyjnego JTLS i aplikacji JEMM do prowadzenia ćwiczeń z dowództwami oddziałów artylerii". *Szybkobieżne Pojazdy Gąsienicowe*, 2 (35) : 57–88.

[14] Sołoducha M. 2016. „Wykorzystanie systemu symulacyjnego JTLS w ćwiczeniach dowódczo-sztabowych wspomaganych komputerowo". *Zeszyty Naukowe AON*, 3 (104) : 18–34.

[15] Scannell E. 2003. „Progress in Directed Energy Weapons Part II: High Power Microwave Weapons," *WSTIAC Newsletter*, 4 (3) : 1–10.

MODELOWANIE I BADANIE IMPULSÓW HPM

MODELING AND TESTING OF HPM PULSES

Rafał Królikowski[1]; Piotr Słobodzian[2]
[1] Politechnika Wrocławska, Wrocław, rafal.krolikowski@pwr.edu.pl
[2] Politechnika Wrocławska, Wrocław, piotr.slobodzian@pwr.edu.pl

DOI: 10.15199/59.2022.4.110

Streszczenie: **W artykule przedstawiono model analityczny typowego impulsu HPM uzyskiwanego za pomocą wirkatora (ang. Virtual Cathode Oscillator). Opisano również eksperymentalną weryfikację zaproponowanego modelu w oparciu o pomiary impulsów HPM wygenerowanych za pomocą generatora arbitralnego.**

Abstract: **This article presents an analytical model of a typical HPM impulse obtained with a vircator (Virtual Cathode Oscilla-tor). An experimental verification of the proposed model based on measurements of HPM pulses generated by means of an arbitrary generator is also described.**

Słowa kluczowe: **impulsy HPM, modelowanie sygnałów, pomiary EMC.**

Keywords: **HPM pulses, signal modelling, EMC testing.**

1. WSTĘP

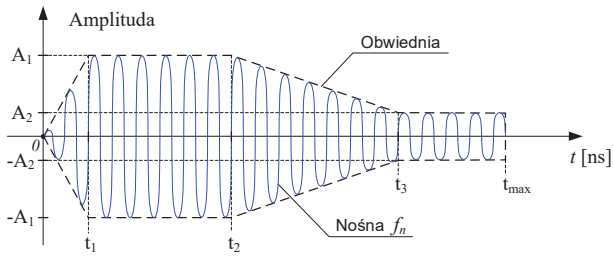
Impuls HPM (ang. High Power Microwave) jest impulsem mikrofalowym o specyficznych własnościach. Jego częstotliwość nośna zwykle mieści się w zakresie od kilkuset MHz do 10 GHz, czas trwania jest rzędu ns, moc szczytowa wynosi od kilkuset MW do kilku GW, a jego widmo jest relatywnie szerokie [1]. W związku z tym impulsy HPM mogą oddziały-wać na różnorodne urządzenia elektroniczne. Obec-nie wytwarzane układy półprzewodnikowe są podat-ne na bardzo niewielkie zakłócenia (nawet o napięciu o amplitudzie 1-2 V). W związku z tym, szybko zau-ważono potencjał impulsów HPM w zastosowaniach militarnych, związanych z celowym zakłócaniem urządzeń elektronicznych. Przykładem takiego zastosowania jest bomba elektromagnetyczna (ang. E-Bomb) [2]. Obecnie na świecie obserwuje się coraz większe zainteresowanie bronią elektromagnetyczną jako nowymi środkami walki w erze informacyjnej [3], [4]. Świadomość i wiedza w tym zakresie drastycznie wzrosły w okresie dwóch ostatnich dekad obecnego wieku.

Znanych jest wiele rozwiązań technicznych, za pomocą których można wytwarzać impulsy HPM. Do najbardziej znanych należą: generator Marksa (MG) [4], [5], generator z kompresją strumienia magnetycznego (FCG) [5], wirkator (VCO) [5], [6]. Parametry impulsów uzyskiwanych za pomocą wymienionych urządzeń są opisywane w specjalistycznej literaturze (np. [7]), jednak dostępność samych urządzeń, ze względu na ich militarne zastosowanie, jest mocno ograniczona. W związku z tym w badaniach oddziaływania impulsów HPM na urządzenia elektroniczne są stosowane generatory o obniżonych parametrach (stosowane w badaniach laboratoryjnych) [8] lub generatory arbitralne, w przypadku których impulsy HPM są syntezowane według znanego modelu. Modelowanie impulsów HPM jest zatem ważnym elementem oceny odporności urządzeń elektronicznych [9].

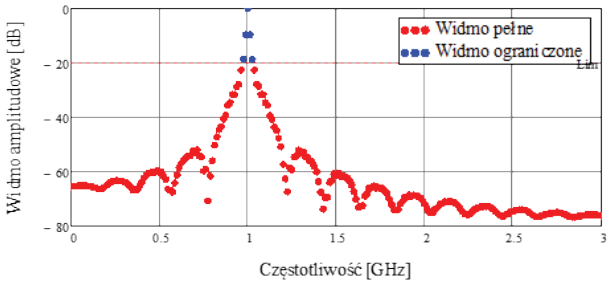
2. MODELE IMPULSÓW HPM

Modele typowych impulsów wysokoenergetycznych (ESD, HRF, LEMP, NEMP i HPM), wykorzystywane w badaniach EMC, są scharakteryzowane w odpowiednich normach cywilnych (np. IEC 61000-2-13) i wojskowych (np. MIL-HDBK-235-8, norma niejawna). Niestety modele sygnatur impulsów opisanych w normach cywilnych nie uwzględniają charakterystyk wszystkich rodzajów źródeł. Z kolei modele opisane w normach wojskowych są niedostępne. W związku z tym, w celu umożliwienia badań z impulsami o sygnaturze zbliżonej do rzeczywistych impulsów HPM zaproponowaliśmy własny model impulsu HPM generowanego przez wirkator. Model został opracowany na podstawie analizy sygnatur i/lub widm impulsów HPM opublikowanych w literaturze specjalistycznej, tj. głównie w czasopiśmie IEEE Transactions on Plasma Science (np. patrz [6]). Wirkator jest typem lampy mikroflowej (podobnej do diody), która generuje ciąg impulsów harmonicznych o określonej częstotliwości i zmiennej w czasie obwiedni. Czas trwania pojedynczego impulsu wynosi zwykle od kilkudziesięciu do kilkuset nanosekund, a okres powtarzania waha się (w zależności od konstrukcji wirkatora) od kilkuset mikrosekund do kilkuset milisekund.

Zaproponowany przez nas model sygnatury impulsu HPM pokazano na rys. 1. Parametry modelu wymienione w podpisie rysunku należy uzupełnić o okres powtarzania impulsów T_0 oraz liczbę powtórzeń N. Z kolei na rys. 2 pokazano widmo zaproponowanej sygnatury HPM.



Rys. 1. Model impulsu HPM opisany za pomocą: częstotliwości nośnej f_n, amplitud A_1 i A_2 oraz chwil czasu t_1, t_2, t_3 i t_{max}



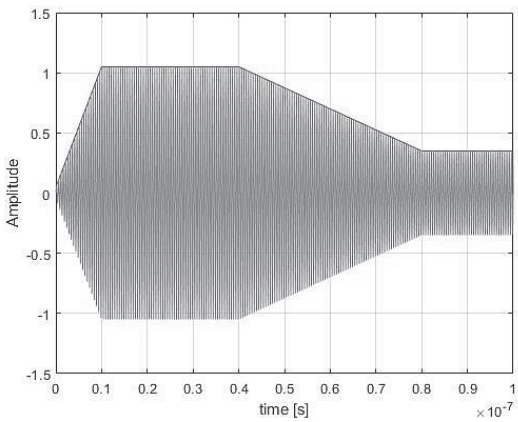
Rys. 1. Unormowane widmo amplitudowe impulsu HPM o mocy szczytowej 600 MW. Parametry sygnatury: f_n=1 GHz, A_1=44,78 kV, A_2=4,48 kV, t_1=5 ns, t_2=50 ns, t_3=70 ns i t_{max}=80 ns, T_0=20 μs, N=10

Rys. 2 przedstawia dyskretne widmo sygnatury HPM (otrzymane dla zbioru próbek sygnatury). Jak widać, przy parametrach wymienionych w podpisie rysunku, widmo jest stosunkowo wąskie. W analizie częstotliwościowej wystarczy uwzględnić jedynie kilka prążków widma w otoczeniu f_n (dla poziomów >-20 dB).

3. POMIARY IMPULSÓW HPM

3.1. Generator impulsów testowych

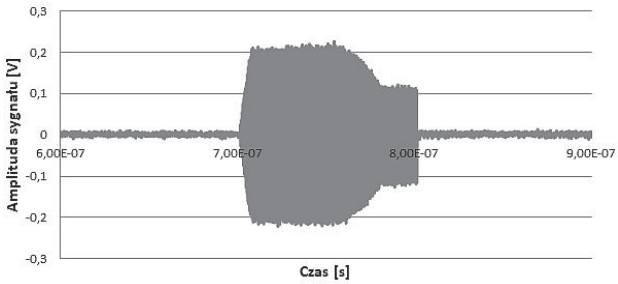
W celu weryfikacji możliwości generowania impulsów HPM o parametrach zgodnych z zaproponowanym modelem przeprowadzono pomiary impulsu testowego, wygenerowanego przy użyciu generatora sygnału arbitralnego. W pomiarach zastosowano sygnaturę HPM o następujących parametrach: częstotliwość nośna f_n=2 GHz, A_1=1 V, A_2=0,35 V, t_1=10 ns, t_2=40 ns, t_3=80 ns, t_{max}=100 ns, okres powtarzania T_0=800 ns, liczba impulsów w serii nieskończona. Parametry sygnatury zostały dobrane tak, aby umożliwić niezawodną synchronizację generatora z oscyloskopem cyfrowym, za pomocą którego rejestrowano wygenerowany sygnał. W pomiarach wykorzystano generator AWG0002A firmy Tektronix, który syntezuje zadany sygnał na podstawie zbioru próbek. Próbki sygnatury HPM przygotowano za pomocą środowiska MATLAB; próbki pobrano z częstotliwością f_p=16 GHz. Otrzymany sygnał pokazano na rys. 3.



Rys. 2. Sygnatura impulsu HPM o czasie trwania 100 ns wygenerowana przy użyciu środowiska MATLAB

3.2. Pomiar sygnatury impulsu HPM

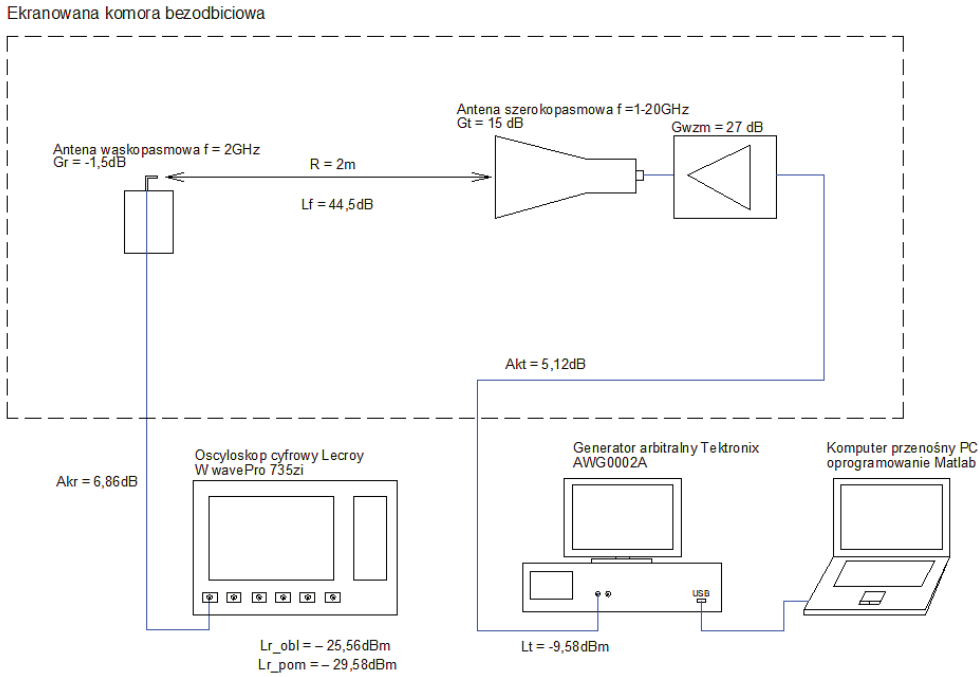
Pomiary wstępne wykonano za pomocą oscyloskopu cyfrowego LeCroy W wavePro735Zi (40Gs/s, pasmo pracy 3,5 GHz). Oscyloskop podłączono bezpośrednio do generatora AWG0002A za pomocą kabla koncentrycznego o długości 3m i tłumieniu ok 2dB przy częstotliwości nośnej impulsu HPM. Przykładowy wynik pomiaru pokazano rys. 5. Jak widać, zmierzona sygnatura odzwierciedla kształt sygnatury zadanej zgodnie z zaproponowany modelem. Różnice są widoczne głównie w długości zbocza opadającego obwiedni impulsów (przy ocenie wizualnej należy wziąć pod uwagę względne skalowanie osi wykresu, które wpływa na proporcje sygnatury).



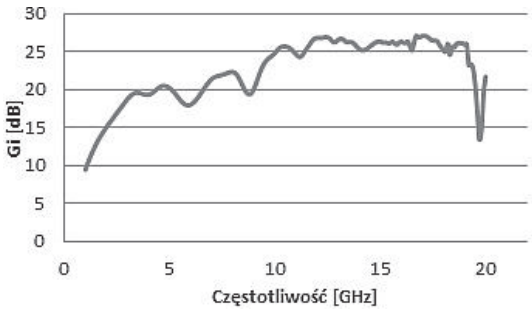
Rys. 3. Sygnatura impulsu HPM o czasie trwania 100 ns zmierzona w układzie generatora połączonego z oscyloskopem za pomocą kabla koncentrycznego

W kolejnym kroku przeprowadzono pomiar impulsu HPM po przejściu przez kanał propagacyjny w postaci wolnej przestrzeni. Wymieniony pomiar symuluje realne warunki oddziaływania impulsów HPM na obiekty. W praktyce, wyemitowane ze źródła impulsy HPM propagują w kierunku obiektu poprzez kanał propagacyjny o określonej funkcji transmitancji. Anteny nadawcza i odbiorcza również posiadają określoną funkcję. Wypadkowa transmitancja ścieżki oddziaływania impulsu HPM (od nadajnika do odbiornika) wpływa na postać rejestrowanego impulsu HPM. Badanie tego wpływu przeprowadzono na stanowisku pomiarowym w komorze bezodbiciowej. Schemat stanowiska zilustrowano na rys. 5 (patrz następna strona). Układ pomiarowy składał się z następujących elementów: generator arbitralny Tektronix AWG0002A; antena nadawcza firmy Zondas (szerokopasmowa antena tubowa o zysku energetycznym 15dB przy 2 GHz, zysk anteny pokazano na rys. 6); anteny odbiorczej (zastosowano wąskopasmową antenę drutową o zysku energetycznym −1,5 dB przy 2 GHz, zysk anteny pokazano na rys. 7); oscyloskop cyfrowy LeCroy W wavePro735Zi; komputer PC z oprogramowaniem Matlab.

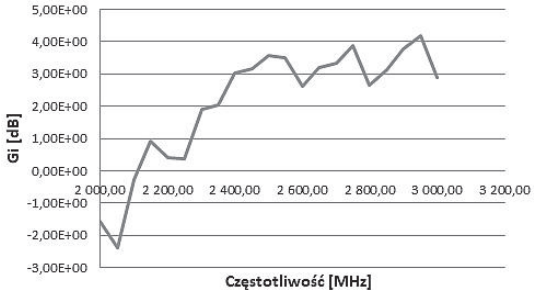
Anteny nadawcza i odbiorcza były połączone z urządzeniami za pomocą kabli koncentrycznych (tłumienie kabli przy częstotliwości pomiaru podano na schemacie). System pomiarowy pracował przy impedancji odniesienia Z_0=50 Ω. Odległość pomiędzy anteną nadawczą i odbiorczą na stanowisku pomiarowym wynosiła 2 m. Dodatkowo, w celu zwiększenia dynamiki pomiaru, zastosowano w układzie szerokopasmowy wzmacniacz sygnału po stronie nadawczej (wzmacniacz o wzmocnieniu 27dB przy 2 GHz).



Rys. 4. Schemat stanowiska pomiarowego do badania impulsów HPM transmitowanych poprzez swobodną przestrzeń



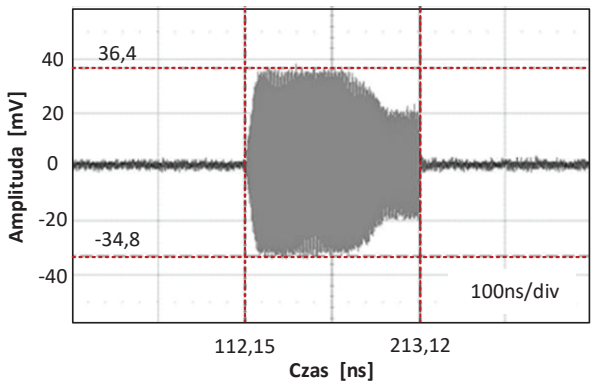
Rys. 5. Przebieg zysku energetycznego w funkcji częstotliwości dla anteny nadawczej firmy Zondas



Rys. 6. Przebieg zysku energetycznego w funkcji częstotliwości dla anteny odbiorczej (antena drutowa)

Dzięki tym zabiegom zwiększono dynamikę pomiaru o ok. 22dB względem układu bez wzmacniacza. Spodziewaną dynamikę systemu pomiarowego oszacowano na podstawie bilans mocy w układzie pomiarowym. Na schemacie układu (rys. 5) oznaczono wielkość tłumień oraz poziomy mocy sygnału umożliwiające przeprowadzenie bilansu (oszacowano wypadkowe tłumienie ścieżki oddziaływania na poziomie 44,5 dB). Dodatkowo, oszacowano również poziom mocy szumu (tła) rejestrowanego na wyjściu toru pomiarowego (wejście oscyloskopu). Pomiar wykonano za pomocą szerokopasmowego miernika mocy HP E4419A (pomiar w paśmie od 50 MHz do 26,5 GHz). W celu zapewnienia synchronizacji oscyloskopu z generatorem przed każdym właściwym pomiarem wykonywano pomiar kontrolny poprzez szerokopasmowy kabel współosiowym. Takie podejście, przy odpowiednim uwzględnieniu opóźnień propagacyjnych, umożliwiało prawidłowe wyodrębnienie impulsu HPM po stronie odbiorczej.

Przykładowy wynik pomiaru sygnatury HPM (o zadanych wcześniej parametrach) pokazano na rys. 8. Jak widać, podobnie jak poprzednim przypadku, zmierzona sygnatura odzwierciedla kształt sygnatury zadanej zgodnie z zaproponowany modelem. Różnice są widoczne głównie w długości zbocza opadającego obwiedni.

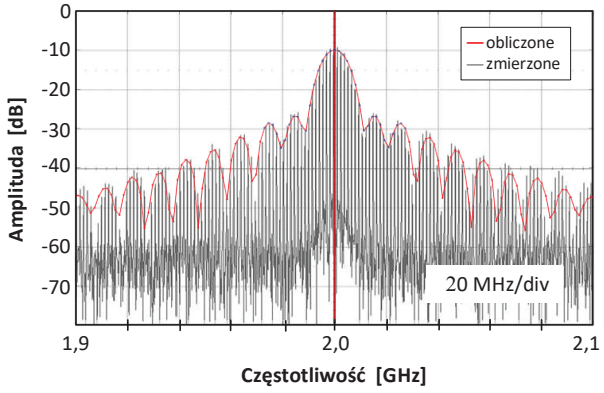


Rys. 7. Sygnatura impulsu HPM o czasie trwania 100 ns zmierzona w układzie pomiarowym pokazanym na rys. 5

Widać również duże podobieństwo do sygnału zarejestrowanego poprzez sam kabel koncentryczny. Pokazuje to, że transmitancja zamodelowanej ścieżki oddziaływania (jej przebieg w funkcji częstotliwości) nie ma tutaj dużego wpływu. Ten wynik jest spójny z własnościami widmowymi zamodelowanej sygnatury HPM (taki wynik jest spodziewany, gdy widmo sygnatury jest skupione w otoczeniu częstotliwości nośnej). W celu potwierdzenia tego spostrzeżenia w kolejnym kroku wykonano pomiar widma wygenerowanego impulsu HPM.

3.3. Pomiar widma sygnatury impulsu HPM

Pomiar widma sygnatury HPM, podobnie jak w przypadku pomiarów w dziedzinie czasu, rozpoczęto od pomiaru poprzez kabel współosiowy. Pomiary wykonano za oscyloskopu cyfrowego LeCroy W wavePro735Zi (wykorzystano tryb analizatora widma). Po ustaleniu nastaw oscyloskopu wykonano ponownie pomiar w układzie z rys. 5. Przykładowy wynik pomiaru pokazano na rys. 9. Otrzymane widmo jest zbieżne z widmem, które zostało wyznaczone na podstawie modelu sygnatury HPM opisującej impuls wykorzystany w eksperymencie. Wymieniona obserwacja ostatecznie potwierdza możliwość modelowania i generowania impulsów HPM na potrzeby badania skutków ich oddziaływania na urządzenia elektroniczne. Zatem cel pracy został w pełni osiągnięty.



Rys. 8. Zmierzone widmo sygnatury impulsu HPM badanej wcześniej w dziedzinie czasu

4. PODSUMOWANIE I WNIOSKI

W artykule opisano model sygnatury impulsy HPM generowanego przez wirkator. W celu weryfikacji zaproponowanego modelu wykonano stosowne pomiary w dziedzinie czasu i częstotliwości. W celu uzyskania odpowiedniej dynamiki pomiaru (powyżej 30dB) konieczne było zastosowanie szerokopasmowego, mikrofalowego wzmacniacza mocy i anteny szerokopasmowej o dużym zysku energetycznym. Zgodność widma teoretycznego i zmierzonego (rys.9) pozwala stwierdzić, że wybrana metoda pomiaru i dobór sprzętu pozwalają na efektywne pomiary impulsów HPM. Zarówno pomiary w dziedzinie czasu jak i częstotliwości potwierdziły możliwość generowania i badania sygnatur impulsów HPM w warunkach laboratoryjnych. Zestawiony system pomiarowy nadaje się do pomiarów sygnałów impulsowych o zaproponowanej sygnaturze – transmitancja ścieżki pomiarowej nie wpływa w znaczący sposób na parametry badanego impulsu.

LITERATURA

[1] Wnuk M., Matuszewski J., Chudy Z. 2015. „Nowe technologie i urządzenia rażenia elektromagnetycznego w dziedzinie walki elektronicznej". *Przegląd Elektrotechniczny*, 91 (3) : 92–95.

[2] Abrams M. 2003. „Dawn of the E-Bomb," *IEEE Spectr.*, 40 (11): 24–30.

[3] Wilson C. 2008. „High Altitude Electromagnetic Pulse (HEMP) and High Power Microwave (HPM) Devices: Threat Assessments". *CRS report for Congress*, RL32544, Congressional Research Service, The Library of Congress, USA.

[4] Szubrycht T., Szymański T. 2005. „Broń elektromagnetyczna jako nowy środek walki w erze informacyjnej". *Zeszyty naukowe Akademii Marynarki Wojennej*, AMW, Gdynia.

[5] Benford J., Swegle J., Schamiloglu E. 2015. „High Power Microwaves". CRC Press, Boca Raton.

[6] Baryshevsky V., et al. 2013. „2-D Simulation and Experimental Investigation of an Axial Vircator". *IEEE Transactions on Plasma Science*, 41 (10) : 2712–2716.

[7] Choi E. H., et al. 2004. „Efficiency Enhancement of High-Power Microwave Generation from a Cylindrical Virtual Cathode Oscillator with Ring-Type Reflector". *Journal of the Korean Physical Society*, 44 (5) :1049–1053.

[8] Białek R., Kuchta M., Kubacki R. 2017. „Metody pomiaru wysokomocowych ipulsów elektromagnetycznych HPEM". *Przegląd Elektrotechniczny*, 93 (1) : 125–128.

[9] Piwowarczyk K., Wnuk M., Nowosielski L. 2014. „Model analityczny impulsu HPEM". *Przegląd Elektrotechniczny*, 90 (7) : 243–246.

[10] Wnuk M., Chudy Z. 2014. „Pomiar mocy impulsu elektromagnetycznego z zakresu mikrofal". *Przegląd Elektrotechniczny*, 90 (8) :239–242.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

MODELOWANIE I POMIARY POLA OŚWIETLAJĄCEGO OBIEKTY IK W PAŚMIE MIKROFALOWYM

MODELING AND MEASUREMENT OF AN INCIDENT FIELD ON CI OBJECTS AT MICROWAVE FREQUENCIES

Paweł Bieńkowski[1]; Piotr Słobodzian[2]; Przemysław Sobkiewicz[3]; Bartłomiej Zubrzak[4]

[1] Politechnika Wrocławska, Wrocław, pawel.bienkowski@pwr.edu.pl
[2] Politechnika Wrocławska, Wrocław, piotr.slobodzian@pwr.edu.pl
[3] Politechnika Wrocławska, Wrocław, przemyslaw.sobkiewicz@pwr.edu.pl
[4] Politechnika Wrocławska, Wrocław, bartlomiej.zubrzak@pwr.edu.pl

DOI: 10.15199/59.2022.4.111

Streszczenie: **W pracy opisano ogólne zasady modelowania oświetlenia obiektów infrastruktury krytycznej (IK) przez źródła za-burzeń HPM. Opracowany model oświetlenia obiektu jest podstawowym elementem wszystkich ścieżek oddziaływania zaburzenia HPM na obiekt. W artykule przedstawiono również przykład eksperymentalnej weryfikacji opracowanego modelu.**

Abstract: **This paper describes general principles for modeling the incident field on objects of critical infrastructure (CI) that is generated by HPM disturbance sources. The developed model is a basic element of all paths of HPM disturbance impact on an CI object. The paper also presents an exam-ple of experimental verification of the model.**

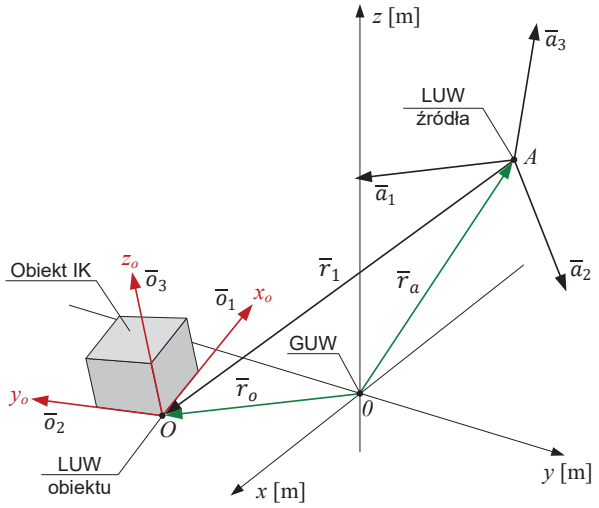
Słowa kluczowe: **impulsy HPM, modelowanie elektromagnetyczne, pomiar pola oświetlającego, zaburzenia elektromagnetyczne.**

Keywords: **electromagnetic disturbances, electromagnetic modeling, HPM pulses, incident field measurement.**

1. WSTĘP

Modelowanie pola EM oświetlającego obiekty infrastruktury krytycznej (IK) narażone na promieniowane zaburzenia EM stanowi fundamentalny element modelowania pełnych ścieżek odziaływania takich zaburzeń. Model pola oświetlającego jest nieodzownym elementem opisu ścieżek oddziaływania zaburzeń HPM [1]-[2]. Modelowanie pola oświetlającego można przeprowadzić różnymi metodami, np. metodą Ray-Tracing [3]. Jednak metody stosowane w analizie propagacji są zbyt wymagające pod względem obliczeniowym i z tego powodu nie mogą być wykorzystywane w narzędziach do szybkiego modelowania (prognozowania) skutków zaburzenia HPM, niezbędnych w systemach wsparcia decyzji (np. w centrach zarządzania kryzysowego).

Zaburzenia typu HPM, w przeciwieństwie do zaburzeń LEMP/NEMP, mają z natury charakter lokalny. W związku z tym globalna analiza propagacyjna (na dużym obszarze terenu) nie jest potrzebna. Dodatkowo, impulsowy charakter zaburzenia HPM powoduje, że istotny obszar propagacji wielodrogowej wokół obiektu jest niewielki. W takim przypadku bardzo dobrym przybliżeniem



Rys. 1. Ilustracja elementów geometrii scenariusza oświetlenia obiektu z anteny źródła impulsów HPM

pola oświetlającego obiekt jest pole padające wykorzystywane w zagadnieniu równoważnym wynikającym z twierdzenia o indukcji [3], a mianowicie (dla składowej **E**):

$$\mathbf{E} = \mathbf{E}^i + \mathbf{E}^s \tag{1}$$

gdzie **E** oznacza pole wypadkowe w otoczeniu obiektu, $\mathbf{E}^i$ jest polem oświetlającym obiekt (w układzie bez obiektu), natomiast $\mathbf{E}^s$ oznacza pole rozproszone przez obiekt.

2. MODELOWANIE POLA OŚWIETLAJĄCEGO

W rzeczywistym scenariuszu oddziaływania zaburzenia HPM na obiekt IK, zaburzenie jest emitowane przez antenę źródła HPM. Obiekt znajduje się zwykle na tyle daleko od źródła HPM, że możemy założyć jego oświetlenie w postaci jednorodnej fali płaskiej (obiekt znajduje się w polu dalekim źródła HPM). W opisanym przypadku modelowanie pola oświetlającego obejmuje dwa elementy: modelowanie geometrii oświetlenia oraz wyznaczanie natężenia pola oświetlającego $\mathbf{E}^i$.

2.1. Modelowanie geometrii oświetlenia

W celu wyznaczenia pola oświetlającego wymagane jest opisanie konfiguracji geometrycznej układu obiekt-antena źródła HPM. Model geometrii oświetlenia można opisać za pomocą wektorów pozycji $\overline{r}$ oraz różnych układów współrzędnych, zilustrowanych na rys. 1.

W przyjętej metodyce modelowania założono, że scenariusz oświetlenia obiektu jest definiowany w globalnym układzie współrzędnych kartezjańskich (GUW), natomiast własności obiektów IK oraz źródeł HPM są zdefiniowane w ich lokalnych kartezjańskich układach współrzędnych (LUW). Pozwala to na niezależne modelowanie obiektów i źródeł HPM oraz elastyczne wykorzystywanie ich modeli w dynamicznych scenariuszach oddziaływania zaburzeń HPM. Definiowanie geometrii scenariusza polega na określeniu współrzędnych wszystkich wektorów w kolejnych chwilach czasu scenariusza. Dane te mogą być pobierane np. z modeli dynamiki ruchu źródeł i obiektów lub mogą być wprowadzane przez operatorów systemu symulacyjnego podczas budowania scenariusza oddziaływania zaburzenia HPM.

2.2. Wyznaczanie pola oświetlającego

Pole oświetlające jest generowane przez antenę źródła HPM. Do wyznaczenia składowych pola oświetlającego (E_x, E_y, E_z) potrzebne są następujące dane: wektor wycelowania anteny źródła HPM w kierunku obiektu (wektor $\overline{r}_1$ w LUW źródła); zysk energetyczny oraz pełna (sferyczna) charakterystyka promieniowania anteny źródła HPM (moduł i faza składowych E_θ i E_φ oraz zysk G_z); częstotliwość i moc źródła HPM (f_n i P_z). Obliczenie natężenia pola elektrycznego w punkcie obserwacji wymaga następujących działań: wyznaczania kątów promieniowania z anteny źródła HPM oraz kątów oświetlenia obiektu w układach współrzędnych sferycznych LUW; wyznaczenia pola **E** promieniowanego z anteny źródła HPM w środku LUW obiektu, w tym transformacji pola **E** z LUW anteny źródła HPM do LUW obiektu. Pole promieniowane z anteny źródła HPM można opisać następującą zależnością (wzór dla składowej E_θ) [5]:

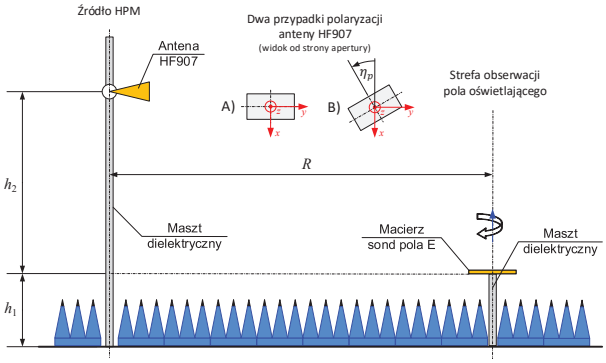
$$E_\theta^a(\theta,\varphi) = \sqrt{2\eta_0 \cdot \frac{P_Z G_z}{4\pi R_0^2} \cdot F_\theta^n(\theta,\varphi)} \cdot \cdot e^{j\cdot\left[\psi_\theta^m(\theta,\varphi) - \frac{2\pi}{\lambda}\cdot(R_0 - R_m)\right]} \tag{2}$$

przy czym, $F_\theta^n(\theta,\varphi)$ i ψ_θ^m oznaczają odpowiednio charakterystykę amplitudową i fazową anteny, zmierzoną lub obliczoną w odległości R_m (wzór dla składowej pola E_φ jest analogiczny). Opisane powyżej działania wykonujemy dla każdej częstotliwości, którą zdefiniowano w modelu źródła HPM.

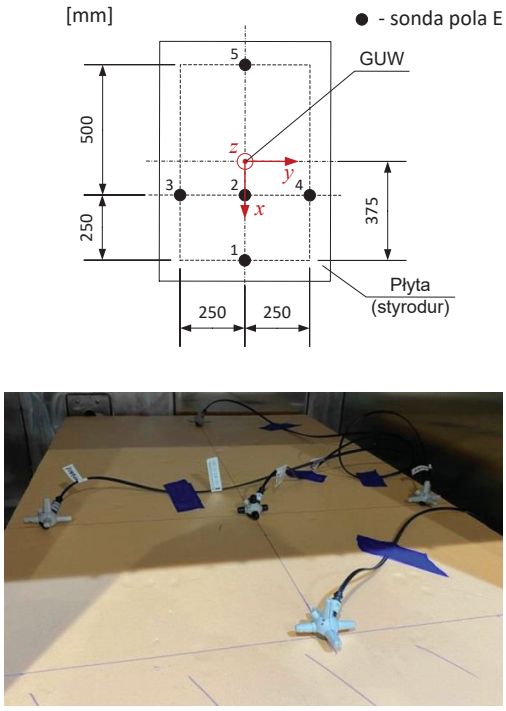
3. WERYFIKACJA POMIAROWA POLA OŚWIETLAJĄCEGO W WARUNKACH LABORATORYJNYCH

3.1. Obiekt badań

Przedmiotem części badań przedstawionych w pracy było wyznaczenie natężenia pola oświetlającego obiekt metaliczny umieszczony w polu kierunkowej anteny nadawczej, w odległości zbliżonej do granicy pola dalekiego



Rys. 2. Schemat układu pomiarowego w komorze bezodbiciowej typu FAC



Rys. 3. Geometria i widok macierzy sond pola elektrycznego (z serii EP-600)

tej anteny. Całe stanowisko było umieszczone w komorze bezodbiciowej typu FAC Laboratorium Techniki Antenowej Katedry Telekomunikacji i Teleinformatyki Politechniki Wrocławskiej. Schemat układu pomiarowego przedstawiono na rys. 2.

Jako źródło PEM zastosowano antenę tubową R&S®HF907. Antena HF907 została zawieszona na maszcie dielektrycznym na wysokości h2 względem poziomu h1. W odległości R od anteny HF907, tj. w strefie obserwacji pola oświetlającego, umieszczono macierz aktywnych sond pola elektrycznego, za pomocą których wykonano pomiar trzech składowych pola, tj. E_x, E_y i E_z. Macierz sond została umieszczona na stoliku obrotowym z masztem dielektrycznym na wysokości h1. Macierz składała się z pięciu sond firmy Narda z serii EP-600, każda z własnym łączem światłowodowym (specyfikację użytych sond przedstawiono w tab. 1). Sondy zostały wywzorcowane w akredytowanym laboratorium wzorcującym LWIMP AP-078. Geometrię zastosowanej macierzy sond pokazano na rys. 3. Z macierzą sond pola związano

globalny układ współrzędnych (GUW) scenariusza pomiarowego. Parametry zasilania anteny (tab. 2) dobrano tak, żeby uzyskać odpowiednią dynamikę sygnału na sondach zarówno dla pola oświetlającego jak i pola wnikającego do obiektów IK, które miały być przedmiotem dalszych badań.

Tab. 1. Parametry robocze aktywnych sond pola elektrycznego z serii EP-600.

Parametr	Sonda	
	EP-600	EP-601
Zakres częstotliwości pracy [GHz]	0,1 – 9,25	0,01 – 9,25
Dynamika [V/m]	0,14 – 140	0,5 – 500
Wymiary (dł. dipola)	53 mm	53 mm

Tab. 2. Parametry źródła PEM

Dana wejściowa	Wartość		
Częstotliwość źródła PEM (f_n) [GHz]	2,5	3,5	4,5
Moc źródła PEM (P_z) [W]	10,0	4,90	3,31
Zysk energetyczny anteny (G_0) [dB]	8,90	9,52	9,81

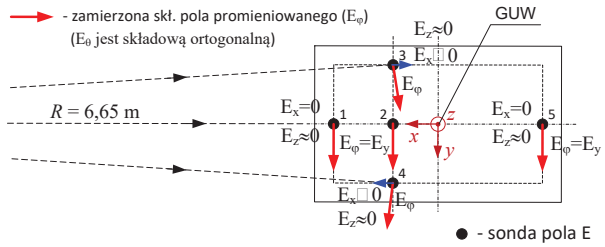
3.2. Scenariusze pomiarowe

W celu weryfikacji pomiarowej modelowania, przygotowano zestaw scenariuszy pomiarowych, różniących się wysokością zawieszenia h2 anteny oświetlającej, jej polaryzacją i kątem obrotu stolika z macierzą sond. Wykorzystane konfiguracje zebrano w tab. 3.

Tab. 3. Konfiguracje pomiarowe anteny i macierzy sond

Nr konfig.	R [m]	h_2 [m]	Kąt stolika φ	Polaryzacja anteny źródła HPM (rys. 0.1)		
1	6,65	0,38	0°	η_p= 0° (pol.V)	η_p= 45° (pol.X)	η_p= 90° (pol.H)
2		1,48				
3		2,98				
4		0,38	45°			
5		1,48				
6		2,98				

Kombinacja wysokości h2, kąta obrotu stolika i polaryzacji anteny daje w sumie 18 przypadków. Zanim przeprowadzono pomiary, dokonano analizy parametrów składowych PEM w przyjętych scenariuszach pomiarowych pod kątem wyboru reprezentatywnych konfiguracji, które pozwalają na uzyskanie jak największej liczby mierzalnych wyników z sond pomiarowych. Ponieważ charakterystyka promieniowania zastosowanej anteny nadawczej przy każdej częstotliwości pomiarowej ma symetryczny listek główny (symetria względem kierunku maksymalnego promieniowania w granicach błędu pomiarowego), a pozycja anteny w GUW jest taka, że jej kierunek maksymalnego promieniowania leży na płaszczyźnie XZ, to pole oświetlające rozkłada się symetrycznie na macierzy sond pola elektrycznego względem płaszczyzny XZ GUW. Oznacza to, że zarówno wyniki symulacji jak i wskazania sond S_3 i S_4 powinny być identyczne. Z kolei,



Rys. 4. Ilustracja składowych pola oświetlającego macierz sond pomiarowych przy orientacji anteny źródła PEM zgodnej z konfiguracją nr 1 przy η_p=90°

ze względu na niewielkie przesunięcie przestrzenne sond S_1 i S_2 ich wskazania powinny być również zbliżone. Po pomiarach okazało się, że sonda S_5 została niewłaściwie skonfigurowana w systemie transmisji danych i jej wyniki nie były miarodajne – została pominięta w analizie danych. Na rys. 4 przedstawiono przykładową analizę składowych PEM dla konfiguracji nr 1 przy polaryzacji anteny η_p=90°. W tym przypadku antena nadawcza R&S®HF907 została ustawiona tak, że polaryzacja pola **E** jest równoległa do płaszczyzny *XY* GUW. A zatem dominującą składową pola elektrycznego, w LUW anteny nadawczej jest składowa E_φ. Przy małej wysokości zawieszenia anteny (h_2 = 0,38 m), wymieniona składowa daje wkład prawie wyłącznie do składowej E_y w GUW scenariusza. Tak powinno być w przypadku sond S_1, S_2 i S_5, tj. składowa E_x powinna być prawie równa 0 V/m. Jednak niewielkie odchylenie położenia sondy względem wzdłużnej osi symetrii stanowiska pomiarowego (patrz sondy S_3 i S_4) powoduje wzrost składowej E_x (pojawia się mała składowa E_x). Analogiczne analizy przeprowadzono również dla pozostałych konfiguracji. Zestawienie uzyskanych wyników przedstawiono w tab. 4, przy czym symbolami P1, P2 i P3 oznaczono względne narastające natężenia pola dla danej składowej PEM w analizowanej konfiguracji.

Tab. 4. Względne poziomy poszczególnych składowych PEM w wybranych konfiguracjach (φ= 0° i η_p= 90°)

Nr konfig.	h_2 [m]	Składowe pola oświetlającego														
		E_x					E_y					E_z				
	SONDA:	S_1	S_2	S_3	S_4	S_5	S_1	S_2	S_3	S_4	S_5	S_1	S_2	S_3	S_4	S_5
1	0,38	≈0	≈0	P1	P1	≈0	P3	P2	P2	P2	P1	P1	P1	P1	P1	P1
2	1,48	P1	P1	P2	P2	P1	P3	P2	P2	P2	P1	P1	P1	P1	P1	P1
3	2,98	P1	P1	P2	P2	P1	P3	P2	P2	P2	P1	P1	P1	P1	P1	P1

3.3. Wyniki weryfikacji pomiarowej

Analizując tabelę 4, można zauważyć, że dla zadanych konfiguracji, największych natężeń PEM i jednocześnie największej zmienność poziomów spodziewać się należy dla składowej E_y i to ta składowa została poddana szczegółowej analizie. Wyniki dla pozostałych składowych były w większości na granicy lub nawet poniżej czułości sond pomiarowych, a tym samym były obarczone bardzo dużym błędem. Niepewność pomiaru uwzględniającą niepewność wzorcowania i ustawienia geometrycznego sond oraz wzajemnych sprzężeń oszacowano na 30%. W poniższych tabelach (tab. 5, 6 i 7) dokonano porównania wyników pomiarów i symulacji dla

składowej E_y przy trzech częstotliwościach, tj. 2,5 GHz, 3,5 GHz i 4,5 GHz. Literą P w tabelach oznaczono wynik pomiaru, S – symulacji, a $\Delta = (\mathrm{P} - \mathrm{S})/\mathrm{S} \cdot 100\%$ oznacza względną różnicę między pomiarem a symulacją.

Na różnicę między wynikami pomiarów i symulacji składa się niepewność pomiaru natężenia PEM, niepewność modeli obliczeniowych oraz niepewność związana z wyznaczeniem pobudzenia anteny nadawczej i niejednorodność generowanego PEM związana z nieidealnym odwzorowaniem swobodnej przestrzeni w FAC (odbicia od ścian) oraz odległością od anteny nadawczej do punktu pomiaru. Dla zastosowanej anteny i częstotliwości pomiarowych szacunkowa niepewność symulacji i niejednorodności pola wynosi nie więcej niż 15%, co daje całkowitą niepewność metody ok. 35%. Analiza wyników z tabeli pozwala określić, że 30 na 36 – a więc ponad 80% wyników nie przekroczyło oszacowanej niepewności metody. Potwierdza to prawidłowość przyjętych modeli, a także przydatność zastosowanego eksperymentu do weryfikacji modeli obliczeniowych.

4. PODSUMOWANIE

Przedstawiona w pracy koncepcja potwierdzenia modelu obliczeniowego metodą pomiarową jest elementem większego projektu. Znajomość natężenia PEM w określonym obszarze jest typową potrzebą przy badaniach związanych z techniką antenową, kompatybilnością elektromagnetyczną, czy oddziaływaniem PEM na biosferę. Metoda pomiaru bezpośredniego jest stosunkowo prosta w realizacji, ale w przypadku konieczności wyznaczenia PEM w wielu punktach i dla wielu częstotliwości może być bardzo pracochłonna, a także obarczona dużą niepewnością, zwłaszcza przy analizach rozkładu PEM w otoczeniu elementów przewodzących. W szczególności niepewność rośnie podczas badań wnikania PEM do obiektów metalowych oraz podczas analizy rozkładu pola wewnątrz takich obiektów – stąd dla optymalizacji nakładu pracy w takich badań wykorzystuje się obliczeniowe modelowanie rozkładu PEM dla zadanej konfiguracji źródła.

Przedstawiona w pracy koncepcja potwierdzenia modelu obliczeniowego metodą pomiarową jest elementem większego projektu – obejmującego również modelowanie i analizy rozkładu PEM wnikającego do „prawie szczelnych" obiektów metalowych. Zastosowanie matrycy sond PEM pozwala na uproszczoną weryfikację modelu nie tylko fali oświetlającej – ale także pola wewnątrz takich obiektów. Dodatkowo, pomiary pola oświetlającego można potraktować jako referencję dla pomiarów pola w z obiektem, minimalizując w ten sposób błędy pomiaru związane z niepewnością wzorcowania sond oraz niejednorodnością generowanego przez antenę nadawczą PEM. Dokonuje się tego przez korektę macierzy pobudzenia źródła o macierz poprawek pomiarowych będących różnicą między wartościami z symulacji i pomiarów w danym punkcie i konfiguracji. Przeprowadzone doświadczenia, których przedstawienie przekracza ramy niniejszej pracy wykazały, że wyniki pomiarów porównane z symulacją skorygowaną o macierz poprawek pobudzenia zbiegają się z niepewności nieprzekraczającą kilku % - co dla zakresu częstotliwości wykorzystywanych w eksperymencie jest wynikiem praktycznie nieosiągalnym metodami pomiarowymi w warunkach pola w otoczeniu obiektów przewodzących.

Tab. 5. Porównanie wyników pomiarów i symulacji f=2,5GHz

Sonda	$\overline{E}_y$ [V/m] (wartość skuteczna)								
	Konfig. nr 1			Konfig. nr 2			Konfig. nr 3		
	P	**S**	**Δ**	**P**	**S**	**Δ**	**P**	**S**	**Δ**
S_1	6,63	7,68	*-13,7*	5,07	6,89	*-26,4*	3,27	4,82	*-32,2*
S_2	7,33	7,39	*-0,8*	5,42	6,63	*-18,3*	2,88	4,80	*-40,0*
S_3	8,18	7,34	*11,4*	4,52	6,60	*-31,5*	3,66	4,79	*-23,6*
S_4	5,09	7,34	*-30,7*	3,16	6,66	*-52,6*	2,64	4,81	*-45,1*

Tab. 6. Porównanie wyników pomiarów i symulacji f=3,5GHz

Sonda	$\overline{E}_y$ [V/m] (wartość skuteczna)								
	Konfig. nr 1			Konfig. nr 2			Konfig. nr 3		
	P	**S**	**Δ**	**P**	**S**	**Δ**	**P**	**S**	**Δ**
S_1	6,32	5,54	*14,1*	5,72	5,06	*13,0*	2,88	3,45	*-16,5*
S_2	6,53	5,33	*22,5*	7,61	4,87	*56,3*	4,21	3,46	*21,7*
S_3	5,38	5,34	*0,7*	7,11	4,88	*45,7*	4,38	3,45	*27,0*
S_4	3,94	5,34	*-26,2*	4,69	4,91	*-4,5*	3,37	3,50	*-3,7*

Tab. 7. Porównanie wyników pomiarów i symulacji f=4,5GHz

Sonda	$\overline{E}_y$ [V/m] (wartość skuteczna)								
	Konfig. nr 1			Konfig. nr 2			Konfig. nr 3		
	P	**S**	**Δ**	**P**	**S**	**Δ**	**P**	**S**	**Δ**
S_1	4,47	4,70	*-4,9*	3,95	4,47	*-11,6*	2,51	3,13	*-19,8*
S_2	7,38	4,52	*63,3*	5,22	4,31	*21,1*	3,89	3,15	*23,5*
S_3	5,64	4,53	*24,5*	3,81	4,32	*-11,8*	2,82	3,16	*-10,8*
S_4	3,73	4,54	*-17,8*	2,82	4,32	*-34,7*	2,13	3,14	*-32,2*

Artykuł zawiera wyniki realizacji projektu finansowanego przez NCBR (DOB-BIO2/PS/5/2/2016).

LITERATURA

[1] Tatum J. 2014. „High-Power Microwave Directed Energy Weapons: A Model and Simulation Toolbox". *DSIAC Journal*, 1 (2) : 20–24.
[2] Tatum J. 2017. „HPM DEWs and Their Effects on Electronic Targets". *DSIAC Journal*, 4 (3) : 33–42.
[3] Yun Z., Iskander M. 2015. „Ray Tracing for Radio Propagation Modeling: Principles and Applications". *IEEE Access*, 3 : 1089–1100.
[4] Harrington R. F. 2001. „Time-Harmonic Electromagnetic Fields". Wiley-IEEE Press, New York.
[5] Balanis C. A. 2005. „Antenna Theory: Analysis and Design" (3rd ed.). John Wiley & Sons, New Jersey.

METODYKA BADANIA WNIKANIA IMPULSÓW HPM DO OBIEKTÓW EKRANOWANYCH

METHODOLOGY OF TESTING THE PENETRATION OF HPM IMPULSES INTO SHIELDED OBJECTS

Leszek Nowosielski[1]; Bartosz Dudziński[2]; Rafał Przesmycki[3]; Jan Kelner[3]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, email: Leszek.Nowosielski@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, email: Bartosz.Dudzinski@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, email: Rafal.Przesmycki@wat.edu.pl
[4] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, email: Jan.Kelner@wat.edu.pl

DOI: 10.15199/59.2022.4.112

Streszczenie: **W artykule przedstawiony został opis schematu blokowego oraz metodyki badania wnikania impulsów wysokoenergetycznych HPM (High Power Microwave) do wnętrza obiektów ekranujących. Przedstawiono opis aparatury pomiarowej wykorzystanej w metodyce badawczej do generacji oraz rejestracji impulsów HPM oraz opis zestawionego stanowiska badawczego w komorze SAC (Semi Anechoic Chamber). Przedstawiono i przeanalizowano wyniki pomiarów propagacji impulsów HPM do wnętrza przykładowej obudowy ekranującej.**

Abstract: **The article presents a description of the block diagram and the methodology of testing the penetration of HPM (High Power Microwave) pulses into the interior of shielding objects. A description of the measuring apparatus used in the test methodology for the generation and recording of HPM pulses and a description of the assembled test stand in the SAC chamber (Semi Anechoic Chamber) are presented. With the help of a pledged test stand, measurements of the propagation of HPM pulses to the inside of an exemplary shielding casing were carried out. The obtained measurement results were presented and analyzed.**

Słowa kluczowe: **HPM, sonda D-dot, EMC.**

Keywords: **HPM pulse, D-dot probe, EMC.**

1. WSTĘP

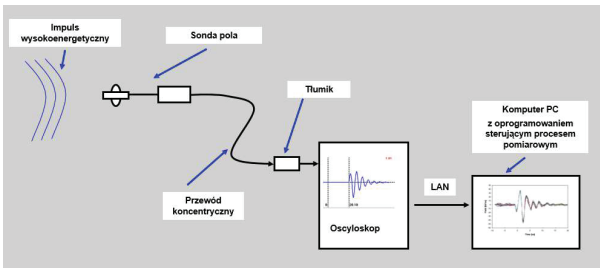
Jedną z metod weryfikacji poprawności działania oprogramowania przeznaczonego do analizy wnikania impulsów wysokomocowych HPM (High Power Microwave) do wnętrza obudów ekranujących jest porównanie wyników uzyskanych metodą symulacji komputerowej z wynikami uzyskanymi metodą badań laboratoryjnych. W artykule przedstawiony został opis schematu blokowego oraz metodyki badania wnikania impulsów wysokoenergetycznych HPM do wnętrza obiektów ekranujących.

W rozdziale drugim artykułu przedstawiono opis aparatury pomiarowej wykorzystanej w metodyce badawczej do generacji oraz rejestracji impulsów HPM oraz opis zestawionego stanowiska badawczego w komorze SAC (Semi Anechoic Chamber). Schemat blokowy stanowiska do badania wnikania impulsów HPM do obiektów ekranowanych przedstawiono w rozdziale trzecim. W rozdziale czwartym przedstawiono opis zestawionego stanowiska badawczego na bazie aparatury pomiarowej znajdującej się na ukompletowaniu Laboratorium Kompatybilności Elektromagnetycznej Wojskowej Akademii Technicznej. Przy pomocy zastawionego stanowiska badawczego przeprowadzono pomiary propagacji impulsów HPM do wnętrza przykładowej obudowy ekranującej. W rozdziale piątym przedstawiono i przeanalizowano otrzymane wyniki pomiarów.

2. GENERATOR ORAZ REJESTRATOR IMPULSÓW HPM

Przykładowym rozwiązaniem konstrukcyjnym generatora sygnału probierczego, który został wykorzystany w zaproponowanej metodzie badania wnikania impulsów HPM do obiektów ekranowanych jest generator Marksa impulsów wysokoenergetycznych HPEM RF DS110 produkowany przez firmę Diehl [6]. Zestaw generatora HPEM RF DS110 wraz z anteną nadawczą i akumulatorem zabudowany jest w przenośnej walizce o wymiarach zewnętrznych 55 cm × 45 cm × 25 cm. Generator ten emituje impulsy w formie tłumionej fali sinusoidalnej o czasie trwania pojedynczego impulsu mniejszym niż 5 ns i poziomie natężenia pole elektrycznego w odległości 2 m około 100 kV/m.

W proponowanym rozwiązaniu stanowiska do badania wnikania impulsów HPM do obiektów ekranowanych wykorzystano rejestrator impulsów, którego schemat blokowy przedstawiono na rys. 1.



Rys. 1. Schemat blokowy rejestratora impulsów HPM

Do monitorowania generowanego poziomu natężenia pola elektrycznego wykorzystywana jest sonda pola D-dot typu AD 70 [1][2][3][4] wraz z układem symetryzatora podłączona przez tłumiki do wejścia oscyloskopu przy pomocy przewodu koncentrycznego. Wygląd zewnętrzny sondy przedstawiono na rys. 2.

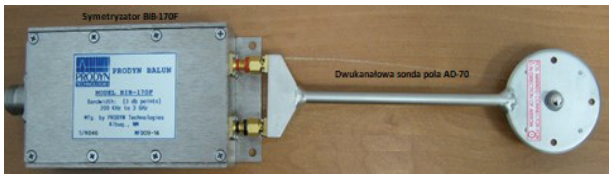
Sonda pola D-dot wraz z układem symetryzatora jest sensorem, na wyjściu którego napięcie *u(t)* jest proporcjonalne do pochodnej funkcji zmian poziomu mierzonego natężenia pola elektrycznego *e(t)*:

$$u(t) = 2 \cdot R_s \cdot A_e \cdot \varepsilon_0 \cdot \frac{\partial e(t)}{\partial t}, \qquad (1)$$

gdzie:

R_s – rezystancja widziana przez pojedynczy kanał sensora [Ω],
A_e – powierzchnia elektryczna pojedynczego kanału sondy pola [m^2],
ε_0 – przenikalność elektryczna próżni.

Według zależności (1) sonda pola elektrycznego D-dot transformuje mierzone natężenie pola elektrycznego $e(t)$ na napięcie pojawiające się na wyjściu sondy $u(t)$, którego wartość jest wprost proporcjonalna do wartości pochodnej zmian mierzonego natężenia pola elektrycznego w funkcji czasu $\frac{\partial e(t)}{\partial t}$. W celu odtworzenia zmian mierzonego natężenia pola elektrycznego $e(t)$ należy sygnał napięciowy $u(t)$ pojawiający się na wyjściu symetryzatora scałkować po czasie.



Rys. 2. Wygląd zewnętrzny sondy pola elektrycznego AD-70

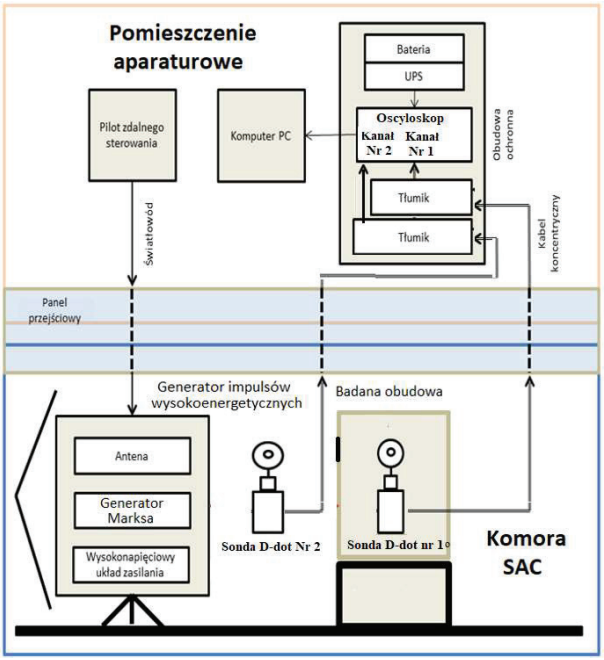
W mierniku poziomu natężenia pola elektrycznego wyjście sondy pola AD 70 oraz symetryzatora BIB 170F jest podłączone poprzez tłumiki do wejścia oscyloskopu, przy pomocy którego zmiany napięcia *u(t)* proporcjonalne do pochodnej zmian poziomu mierzonego natężenia pola elektrycznego *e(t)* są zapamiętywane w pamięci wewnętrznej oscyloskopu. Zarejestrowane próbki sygnału *u(t)* są następnie przesyłane poprzez interfejs zdalnego sterowania do komputera sterującego procesem pomiarowym. Oprogramowanie sterujące procesem pomiarowym wylicza wartość całki zarejestrowanego sygnału, która odpowiada funkcji *e(t)* mierzonego poziomu natężenia pola elektrycznego. Do wyliczenia widma częstotliwościowego zarejestrowanego sygnału *e(t)* wykorzystywana jest transformata Fouriera.

Oprogramowanie sterujące pracą miernika natężenia pola elektrycznego posiada poniższe funkcjonalności:

- zdalne sterowanie pracą oscyloskopu rejestrującego przebieg *u(t)* impulsu wysokoenergetycznego w dziedzinie czasu,
- odczyt z pamięci oscyloskopu próbek zarejestrowanego przebiegu czasowego *u(t)* wysokoenergetycznego impulsu proporcjonalnego do pochodnej zmian mierzonego natężenia pola elektrycznego,
- realizacja obliczeń wartości całki zarejestrowanego przebiegu czasowego wysokoenergetycznego impulsu, która po uwzględnieniu powierzchni elektrycznej A_e sondy pola odpowiada funkcji *e(t)* poziomu mierzonego natężenia pola elektrycznego,
- obliczanie widma częstotliwościowego zarejestrowanego impulsu wysokoenergetycznego za pomocą transformaty Fouriera,
- zobrazowanie graficzne reprezentacji czasowej i częstotliwościowej zarejestrowanego impulsu wysokoenergetycznego.

3. SCEMAT BLOKOWY STANOWISKA POMIAROWEGO ORAZ METODYKA BADAŃ

Schemat blokowy stanowiska do badania wnikania impulsów HPM do obiektów ekranowanych, w którym jako nadajnik wykorzystywany jest generator Marksa impulsów wysokoenergetycznych typu HPEM-RF DS110, a funkcję dwukanałowego rejestratora natężenia pola elektrycznego pełni oscyloskop z sondami pola elektrycznego AD-70, przedstawiono na rys. 3.

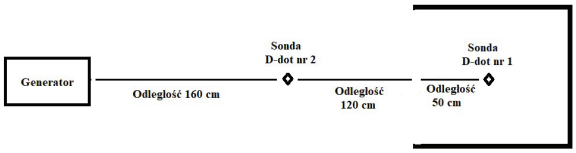


Rys. 3. Schemat blokowy stanowiska pomiarowego wykorzystywanego do badania wnikania impulsów HPM do obiektów ekranowanych

Podczas procesu pomiarowego generator impulsów wysokoenergetycznych generuje impuls HPM o czasie trwania mniejszym niż 5 ns oraz amplitudzie natężenia pola elektrycznego 100 kV/m. Oprogramowanie sterujące pracą miernika natężenia pola elektrycznego ma za zadanie zarejestrowanie wygenerowanego impulsu w dziedzinie czasu oraz wyliczenie natężenia pola elektrycznego zgodnie z zależnością (1).

Rejestrację wygenerowanego impulsu HPM przy pomocy rejestratora natężenia pola elektrycznego w dziedzinie czasu przeprowadzono w układzie pomiarowym przedstawionym na rys. 3. Dokładny opis miernika poziomu impulsów HPM jest przedstawiony w pozycjach literaturowych [5][6][7]. Wynikiem końcowym pomiarów są zarejestrowane dwa przebiegi wartości natężenia pola elektrycznego w dziedzinie czasu. Pierwszy zarejestrowany jest za pomocą sondy pola D-dot o numerze 2 usytuowanej przed badaną obudową ekranującą. Natomiast drugi przebieg jest zarejestrowany za pomocą sondy pola D-dot o numerze 1 usytuowanej wewnątrz badanej komory ekranującej. Rejestracja przebiegów wartości natężenia pola elektrycznego w dwóch punktach pomiarowych jednocześnie przy pomocy dwóch sond pomiarowych ma celu określenie czasu opóźnienia impulsu probierczego, wynikającego z opóźnienia wnoszonego drogą

propagacji impulsu probierczego lub jego odbić od elementów konstrukcyjnych badanej komory ekranującej, względem czasu rejestracji impulsu probierczego zarejestrowanego przez sondę D-dot o numerze 2 zlokalizowaną bezpośrednio przed badaną obudową ekranującą. Uproszczony schemat blokowy układu pomiarowego przedstawiono na rys. 4.

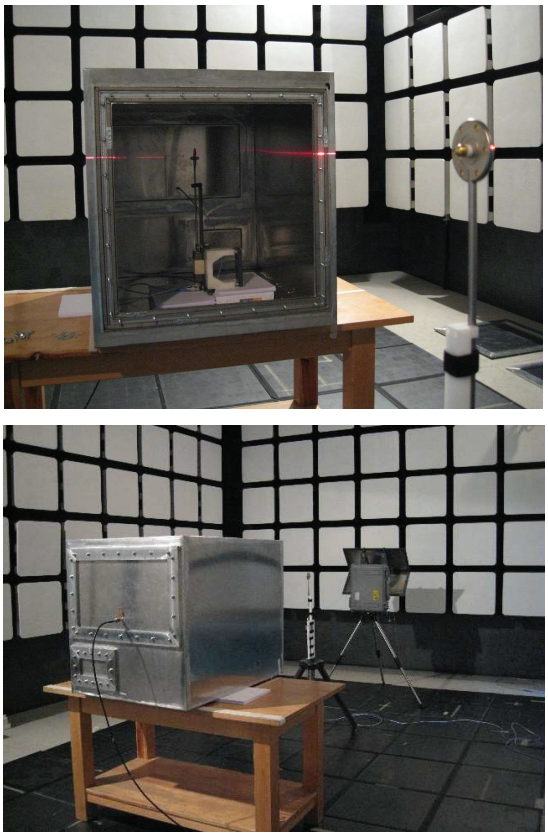


Rys. 4. Uproszczony schemat blokowy układu pomiarowego podczas pomiaru z badaną komorą ekranującą z usuniętą jej przednią pokrywą

Ze względu na to, że generator HPEM RF DS110 generuje stosunkowo wysokie poziomy natężenia pola elektrycznego do 100 kV/m w szerokim zakresie częstotliwości, koniecznym jest przeprowadzanie badań z zastosowaniem opisanego powyżej stanowiska pomiarowego wewnątrz dużej komory SAC.

4. ZESTAWIONE STANOWISKO BADAWCZE

Zdjęcie obiektu badań, sond D-dot oraz generatora impulsów HPM zlokalizowanych wewnątrz komory SAC przedstawiono na rys. 5.



Rys. 5. Zdjęcie obiektu badań, sond D-dot oraz generatora impulsów HPM zlokalizowanych wewnątrz komory SAC

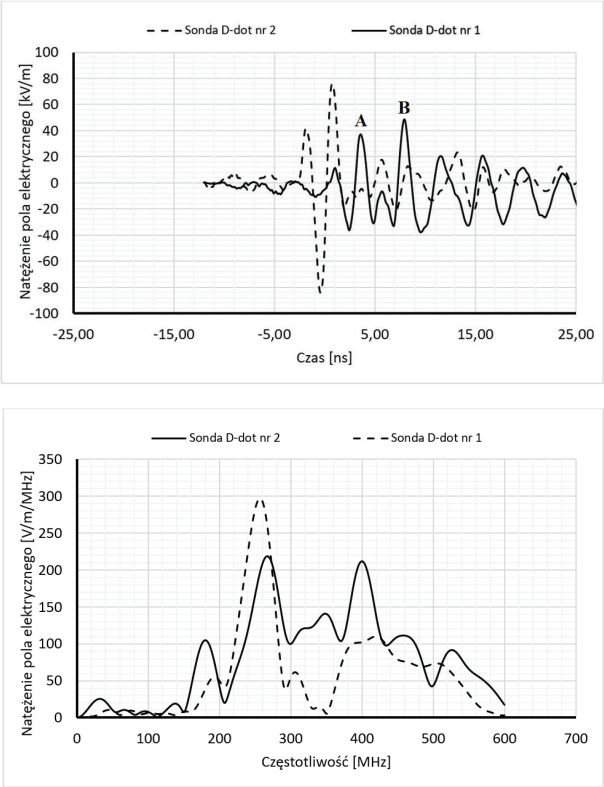
W celu przebadania wpływu elementów konstrukcyjnych komory ekranującej na amplitudę oraz opóźnienia czasowe impulsu probierczego HPM przeprowadzono wstępne badania propagacji impulsu wewnątrz komory ekranującej wykonanej ze stopu aluminium o wymiarach: wysokość 80 cm, szerokość 75 cm, głębokość 100 cm z otwartym otworem przeznaczonym na instalację przedniej pokrywy o wymiarach 80 cm na 80 cm. Sygnał probierczy zaindukowany w sondzie D-dot nr 1 umiejscowionej wewnątrz badanej obudowy ekranującej jest wyprowadzony do wejścia oscyloskopu za pośrednictwem adaptera panelowego SMA/SMA umiejscowionego w tylnej ścianie badanej obudowy ekranującej.

Urządzenia pomiarowe takie jak oscyloskop oraz komputer sterujący procesem pomiarowym zlokalizowano w przyległym do komory SAC pomieszczeniu aparaturowym.

5. WYNIKI POMIARÓW

W celu określenia różnicy poziomu amplitudy oraz opóźnienia czasowego impulsu wysokoenergetycznego zarejestrowanego przez dwie sondy D-dot umieszczone w linii propagacji impulsu wysokoenergetycznego na kierunku od sondy D-dot nr 1 do sondy D-dot nr 2 wykonano pomiar z badaną komorą ekranującą bez przedniej pokrywy. Schemat drogi propagacji badanego impulsu wysokoenergetycznego przedstawiono odpowiednio na rys. 4. Przebiegi czasowe oraz widmo impulsu wysokoenergetycznego zarejestrowanego przez sondę D-dot nr 1 oraz sondę D-dot nr 2 umieszczone w linii propagacji impulsu wysokoenergetycznego przedstawiono na rys. 6.

Na przebiegu czasowym przedstawionym na rys. 6 można zaobserwować wyraźne echa impulsu zarejestrowanego przez sondę D-dot nr 1. Pierwszy impuls „A" (rys. 6) zarejestrowany przez sondę D-dot nr 1 jest opóźniony względem impulsu zarejestrowanego przez sondę D-dot nr 2 o czas wynikający z propagacji fali elektromagnetycznej na odcinku pomiędzy dwiema sondami (170 cm). Drugi impuls „B" (rys. 6) zarejestrowany przez sondę D-dot nr 1 (echo impulsu wynikające z jego odbicia od tylnej ściany badanej komory ekranującej) jest opóźniony względem pierwszego impulsu o czas niezbędny na pokonanie w obu kierunkach odcinka pomiędzy miejscem lokalizacji sondy D-dot nr 1 a tylną ścianą badanej obudowy ekranującej. Przeprowadzona analiza opóźnień czasowych impulsu probierczego oraz jego odpić potwierdza poprawność działania stanowiska pomiarowego.



Rys. 6. Przebiegi czasowe oraz widmo impulsu wysokoenergetycznego zarejestrowanego przez sondę D dot nr 1 oraz sondę D dot nr 2

6. WNIOSKI

W ramach artykułu przedstawiono:

- metodykę badań penetracji impulsu wysokoenergetycznego HPM do wnętrza obiektów ekranujących,
- budowę stanowiska badawczego w komorze SAC do badań penetracji impulsów wysokoenergetycznych do wnętrza obiektów ekranujących,
- wyniki pomiarów poziomu oraz opóźnienia impulsów wysokoenergetycznych zarejestrowanych przez dwie sondy D-dot umieszczone w linii propagacji impulsu wysokoenergetycznego na kierunku od generatora impulsu do wnętrza badanej komory ekranującej.

Przedstawiona w opracowaniu: metodyka badawcza, aranżacja stanowiska pomiarowego oraz wstępne wyniki pomiarów mogą być wykorzystane jako dane wejściowe procedury walidacyjnej mającej na celu sprawdzenie poprawności procedur symulacyjnych, których zadaniem jest symulacja komputerowa wnikania impulsów HPM do wnętrza obudów ekranujących.

Praca została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu nr UGB-22-741/2022 pt.” Pomiar parametrów elektrycznych materiałów stałych metodą odbiciową i metodą wolnej przestrzeni w paśmie milimetrowym”.

LITERATURA

[1] Edgel R. W. „ Prodyn aplication note 1195”.
[2] Edgel R. W. „ Prodyn aplication note 606”.
[3] Edgel R. W. „ Prodyn aplication note 192”.
[4] Edgel R. W. „ Prodyn aplication note 895”.
[5] Nowosielski Leszek. 2015. “Measurement of the Pulses Generated by the High Power Electromagnetic Pulse Generator”. *Proceedings of PIERS 2015*, Praga, Czechy, 6-9.07.2015, s. 297-301, ISSN 15599450.
[6] “Operating and Maintenance Manual for DIEHL, DS110.KS2.F.MP1.B (HPM Case System DS110), Fixed Frequency High-Power RF Source”. 2007.
[7] Przesmycki Rafał, Nowosielski Leszek, Bugaj Marek, Wnuk Marian, Piwowarczyk Kazimierz. 2012. „The Shielding Effectiveness Measurement Using High Voltage Pulse Generator”. *Proceedings of PIERS 2012*, Moskwa, Rosja, 19-23.08.2012 r. s. 1449-1452, ISBN 978-193414222-6.

PROBLEMATYKA ODDZIAŁYWANIA IMPULSÓW HPM NA WYBRANE UKŁADY ELEKTRONICZNE

Józef Małecki[1]; Andrzej Żak[1]; Ryszard Studański[2]

[1] Akademia Marynarki Wojennej
[2] Uniwersytet Morski w Gdyni

…

DOI: 10.15199/59.2022.4.113

Streszczenie: **Elementem poprawnego działania urządzeń cyfrowych takich jak urządzenia teletransmisyjne jest prawidłowe działanie podstawowych elementów, z których te urządzenia zbudowano. Najczęściej do budowy tych urządzeń wykorzystuje się cyfrowe układy logiczne. Działanie tego typu układów można zakłócić poprzez przewodzone zakłócenia impulsowe wysokiej częstotliwości. zakłócenia, które mogą być generowane w wyniku pracy innych urządzeń lub być generowane sztucznie. Ostatnimi czasami szczególne znaczenie nabrało generowanie impulsów elektromagnetycznych w ramach tak zwanej broni nieśmiercionośnej (impuls HPM). Aby móc oszacować odporność urządzenia na takie zakłócenia należy określić progi, przy których działanie układów cyfrowych będzie nieprawidłowe. W związku z powyższym celem przeprowadzonych badań jest określenie odporności układów CMOS na przewodzone zakłócenia impulsem wysokiej częstotliwości. W artykule skupiono się na badaniach układów serii HC.**

Abstract: **An element of the correct operation of digital devices such as teletransmission devices is the correct operation of the basic elements from which these devices are built. Most often, digital logic circuits are used to build these devices. The operation of such systems can be disturbed by conducted high-frequency impulse noise. disturbances that may be generated by the operation of other devices or generated artificially. Recently, the generation of electromagnetic pulses as part of the so-called non-lethal weapon (HPM pulse) has become particularly important. In order to be able to estimate the resistance of the device to such interferences, it is necessary to determine the thresholds at which the operation of digital circuits will be incorrect. In connection with the above, the aim of the research is to determine the resistance of CMOS systems to conducted disturbances by high-frequency impulse. The article focuses on the research of HC series systems.**

Słowa kluczowe: **atak elekroniczny (EA), celowe oddziaływanie EM (IEMI), impuls elektromagnetyczny HPM, układ elektroniczny.**

Keywords: **electronic attack (EA), intentional EM interaction (IEMI), electromagnetic pulse HPM, digital electronics..**

1. WSTĘP

Jednym z newralgicznych elementów poprawnego działania teleinformatycznych urządzeń cyfrowych takich jak: komputery, routery, przełączniki, punkty dostępowe, itp., jest prawidłowe działanie podstawowych elementów, z których te urządzenia zbudowano. Najczęściej do budowy tych urządzeń wykorzystuje się mikroprocesory i ogólniej dyskretne układy logiczne [6]. Działanie takich układów można zakłócić poprzez przewodzone zakłócenia impulsowe wysokiej częstotliwości. Najczęściej zakłócenia te powstają w wyniku rozchodzenia się fali elektromagnetycznej i indukowania się napięć na płytkach poszczególnych modułów. Fala elektromagnetyczna może być generowana w wyniku pracy innych urządzeń (z tym wiąże się między innymi kompatybilność elektromagnetyczna urządzenia), mogą być generowane sztucznie lub powstawać z przyczyn naturalnych np. w wyniku wyładowań atmosferycznych. W ostatnich czasach szczególne znaczenie nabrało generowanie impulsów elektromagnetycznych w ramach tak zwanej broni nieśmiercionośnej (impuls HPM). Aby móc oszacować odporność urządzenia na takie zakłócenia należy określić progi, przy których działanie układów cyfrowych będzie nieprawidłowe [4, 5]. W związku z powyższym celem przeprowadzonych badań było określenie odporności układów CMOS na przewodzone zakłócenia impulsem wysokiej częstotliwości. W tym artykule i badaniach skupiono się na układach serii HC.

Zakłócenia elektromagnetyczne (EMI), które oddziaływują na urządzenia teletransmisyjne zawirające sprzęt półprzewodnikowy i systemy elektroniczne mogą być spowodowane zjawiskami naturalnymi. Do takich zaliczyć należy błyskawice i elektryczność statyczną. Oprócz niezamierzonego oddziaływania EMI, zakłócenia mogą wystąpić celowo [4, 5]. Celowe zaklócenia EMI (IEMI) to spowodowana przez człowieka ingerencja mająca na celu zakłócenie lub zniszczenie sprzętu elektronicznego.

Zastosowanie IEMI w kontekście wojskowym to nic innego jak w najszerszym tego słowa znaczeniu znaczeniu atak elektroniczny (EA), który jest jednym ze składników walki elektronicznej (EW). Działanie EA rozciąga się od zagłuszania sygnałów elektronicznych do niszczenia sprzętu elektronicznego [1, 2, 3 - do zniszczenia komponentów elektronicznych i sprzętu koniecznym jest użycie generatorów mikrofalowych o dużej mocy (HPM).

2. WARUNKI POMIARÓW - STANOWISKO POMIAROWE

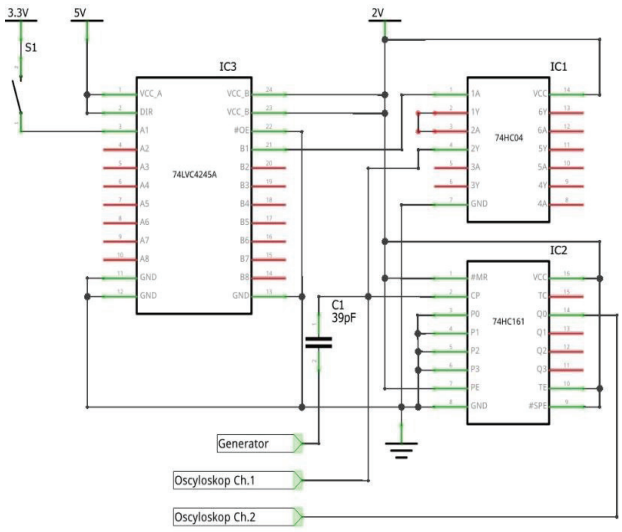
W celu zbadania odporności układów CMOS na przewodzone zakłócenia impulsem wysokiej częstotliwości zestawiono stanowisko laboratoryjne składające się z:

- Układu elektronicznego zbudowanego na układach CMOS;
- Oscyloskopu cyfrowego Hewlett Packard Infinium 500MHz 1GS/s;
- Generatora sygnałów Rohde & Schwartz SMBV 100A / SMJ100A;
- Zasilacza laboratoryjnego Power Lab 3010D-II.

Badany układ elektroniczny składał się z podzespołów wykonanych w technologii CMOS:

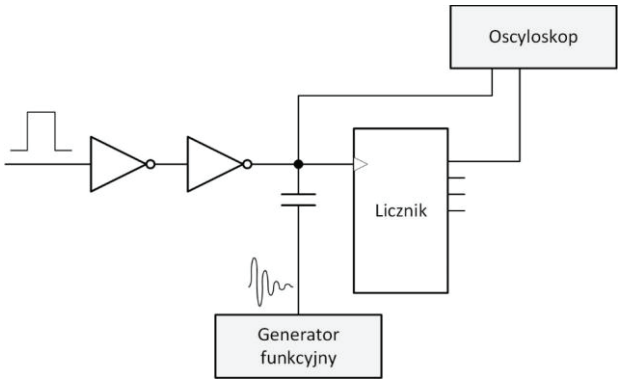
- Inwertera 74HC04;
- 4 bitowego licznika 74HC161;
- Konwertera napięć 74LVC4245A;
- Kondensatora 39pF;
- Przełącznika monostabilnego.

Schemat elektroniczny połączeń dla testowanego układu przedstawiono na rysunku poniżej. Przełącznik monostabilny jest wykorzystywany do wytworzenia impulsowego zbocza narastającego, które było zliczane przez 4-bitowy licznik (74HC161). W celu odseparowania przełącznika od układu sygnał z przełącznika monostabilnego był podawany na dwa inwertery połączone szeregowo (74HC04). Aby umożliwić przeprowadzenie badań dla różnych napięć zasilania zastosowano konwerter napięć (74LVC4245A).



Rys.1. Schemat ideowy badanego układu

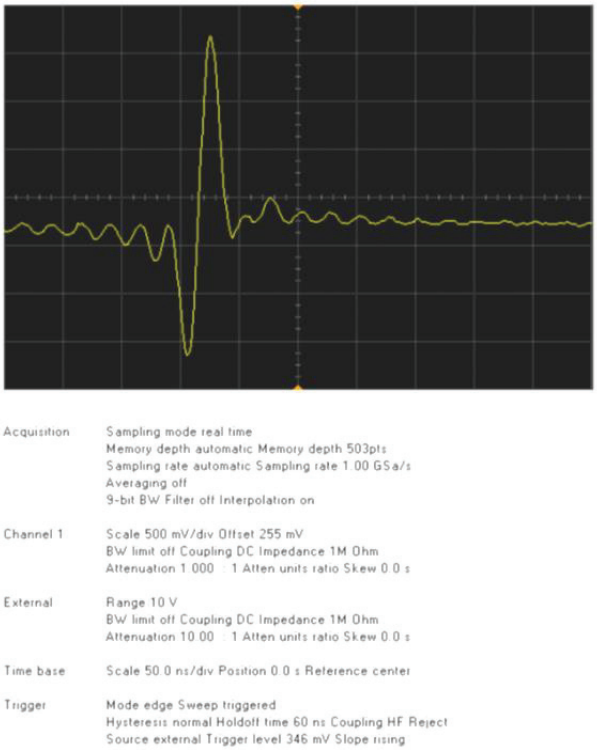
Logiczny schemat funkcjonalny opracowanego układu, który został poddany badaniom testowym przedstawiono na rysunku poniżej.



Rys. 2. Schemat funkcjonalny stanowiska pomiarowego

Z generatora funkcyjnego SMBV podawano sygnał na wejście licznika o szybkości modulacji 20MBd (maksymalna szybkość modulacji zapewniana przez generator). Ciąg modulujący miał postać 1 jedynki i 40 zer i był podawany na modulator w sposób ciągły. Sygnał modulujący filtrowano za pomocą filtru podniesionego cosinusa o współczynniku roll-off 0,5. Natomiast częstotliwość sygnału nośnego zmieniano od 10 MHz aż do górnego zakresu pracy generatora (3,3 GHz). W praktyce uzyskano falę impulsów w.cz o czasie trwania 50 ns i przerwie 2 μs. Efektywne zakłócenie stanowił impuls w.cz., którego przykładowe realizacje przedstawiono na rysunku poniżej (Rys. 3).

Na prezentowanych w artykule rysunkach kolorem zielonym reprezentowano zakłócenie łącznie z wartością logiczną na wejściu układu, kolorem żółtym postać sygnału na jego wyjściu.



Rys. 3. Przykładowy sygnał zakłócający użytych w badaniach o częstotliwości nośnej 20MHz

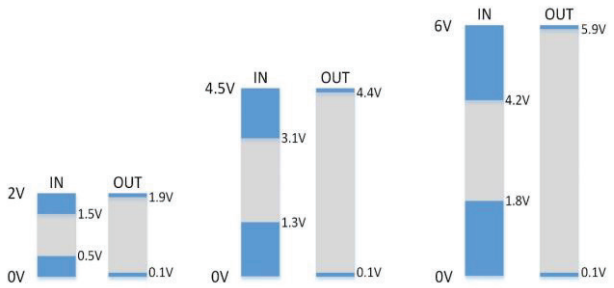
Sygnał z generatora poprzez kondensator o pojemności 39pF był wprowadzany na wejście zliczające

licznika. Na oscyloskopie został zwizualizowany sygnał podawany na wejście zliczające licznika oraz sygnał z wyjścia Q_0 (najmłodszy bit) licznika. Obraz na oscyloskopie rejestrowano podczas pojawienia się zbocza narastającego na wejściu przed dwiema bramkami (wyjście przełącznika S1), którym sterowano zewnętrzne wejście synchronizacji oscyloskopu. W ten sposób istniała możliwość obserwacji wpływu zakłócenia nie tylko w warunkach statycznych (gdy nie dokonywano zmian na wejściu dynamicznym licznika), ale również w trakcie przełączania układu, gdy zakłócenie występowało w trakcie zbocza narastającego. W przypadku powstania zakłócenia na wyjściu tym obserwowano zmiany stanu niezgodne z oczekiwanym.

Głównym ograniczeniem zestawu pomiarowego była możliwość wygenerowania sygnału o odpowiednio dużych wartościach napięciowych wobec powyższego badania miały charakter ograniczony.

3. WYNIKI POMIARÓW

Pomiarów dokonano przy trzech napięciach zasilania układów cyfrowych tj. 2V, 4.5V oraz 6V. Wybrane napięcia zasilania obrano na podstawie parametrów z karty katalogowej zastosowanego licznika 74HC161. Poziom napięcia zasilania decydował o poziomach sygnałów stanowiących logiczną 1 i logiczne 0. Poziomy te, wraz z marginesami błędów, przedstawiono na rysunku poniżej. Poniżej zestawiono uzyskane wybrane wyniki z wykonanych badań pomiarowych.



Rys. 4. Poziomy napięć wejść i wyjść układów CMOS dla różnych wartości napięcia zasilania

3.1. Napięcie zasilania 2V

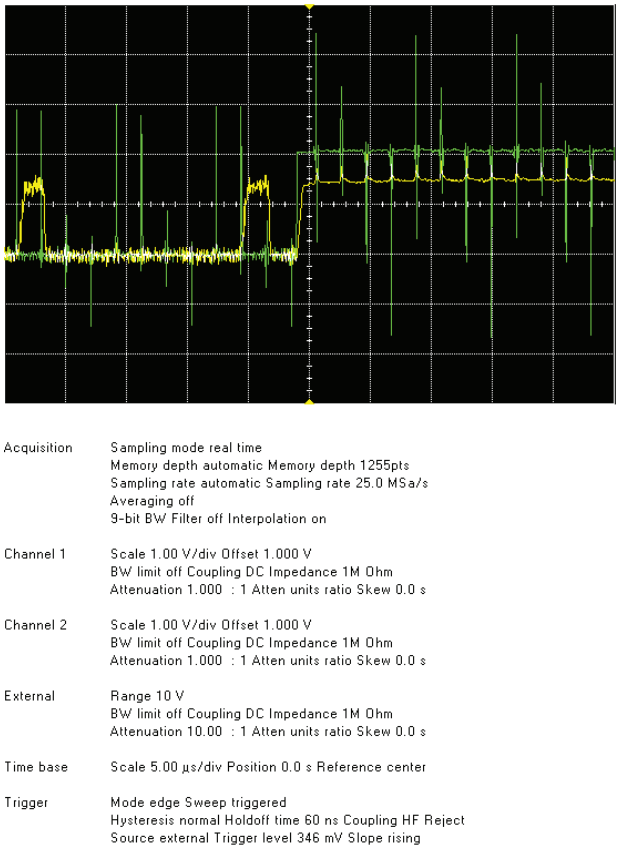
Układ został zasilony napięciem 2V. Dokonując zmian częstotliwości nośnej sygnału zakłócającego w zakresie 20MHz do 70MHz oraz zmieniając międzyszczytową wartość napięcia sygnału zakłócającego dokonano sprawdzenia w jakich przypadkach zachodzi nieprawidłowe działanie układu licznika (74HC161). W Tabeli 1 zestawiono graniczne międzyszczytowe wartości napięcia sygnału zakłócającego dla poszczególnych częstotliwości nośnych, dla których dochodziło do nieprawidłowego działania badanego układu.

Tabela 1. Graniczne międzyszczytowe wartości napięcia sygnału zakłócającego dla Vcc=2V.

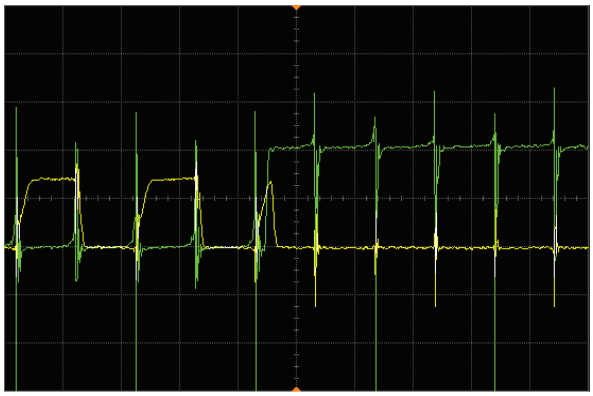
Częstotliwość nośna [MHz]	20	30	40	50	60	70
Wartość międzyszczytowa [V]	4.73	4.21	3.92	4.41	7.55	7.77

Dla częstotliwości nośnej poniżej 20MHz (ze względu na szybkość modulacji) oraz powyżej 70 MHz, ze względu na ograniczenia aparaturowe, nie można było wygenerować sygnału o właściwej międzyszczytowej wartości napięcia, który mógłby doprowadzić do zakłócenia działania układu.

Na rysunkach (Rys. 4 i 5) poniżej przedstawiono przykłady powstawania zakłóceń dla dwóch różnych częstotliwości nośnych i międzyszczytowej wartości napięcia sygnału zakłócającego.



Rys. 4. Przykład powstawania błędów w pracy układu logicznego dla częstotliwości nośnej sygnału zakłócającego f=20MHz



Acquisition Sampling mode real time
Memory depth automatic Memory depth 1004pts
Sampling rate automatic Sampling rate 50.0 MSa/s
Averaging off
9-bit BW Filter off Interpolation on

Channel 1 Scale 1.00 V/div Offset 1.000 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 1.000 : 1 Atten units ratio Skew 0.0 s

Channel 2 Scale 1.00 V/div Offset 1.000 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 1.000 : 1 Atten units ratio Skew 0.0 s

External Range 10 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 10.00 : 1 Atten units ratio Skew 0.0 s

Time base Scale 2.00 µs/div Position 0.0 s Reference center

Trigger Mode edge Sweep triggered
Hysteresis normal Holdoff time 60 ns Coupling HF Reject
Source external Trigger level 346 mV Slope rising

Rys. 5. Przykład powstawania błędów w pracy układu logicznego dla częstotliwości nośnej sygnału zakłócającego f=70MHz

3.2. Napięcie zasilania 6V

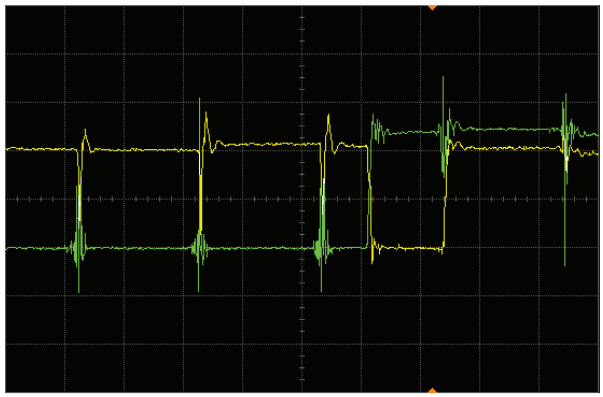
Następnie badany układ zasilono napięciem 6V. Dokonując zmian częstotliwości nośnej sygnału zakłócającego w zakresie 20MHz do 70MHz oraz zmieniając wartość międzyszczytową napięcia sygnału zakłócającego dokonano sprawdzenia warunków w jakich zachodzi nieprawidłowe działanie układu licznika (74HC161). W tabeli 3 zestawiono graniczne wartości międzyszczytowe napięcia sygnału zakłócającego dla poszczególnych częstotliwości nośnych, dla których dochodziło do nieprawidłowego działania licznika.

Tabela 2. Graniczne wartości międzyszczytowe napięcia sygnału zakłócającego dla Vcc=6V.

Częstotliwość nośna [MHz]	20	30	40	50	60	70
Wartość międzyszczytowa [V]	4.36	4.19	3.06	4.55	6.46	6.93

Dla częstotliwości nośnej powyżej 70 MHz, ze względu na ograniczenia aparaturowe, nie można było wygenerować sygnału o właściwej wartości międzyszczytowej napięcia, który mógłby doprowadzić do zakłócenia działania układu.

Na rysunkach (Rys. 6. i 7) poniżej przedstawiono przykłady powstawania zakłóceń dla różnych częstotliwości nośnej i wartości międzyszczytowych napięcia sygnału zakłócającego.



Acquisition Sampling mode real time
Memory depth automatic Memory depth 1004pts
Sampling rate automatic Sampling rate 100 MSa/s
Averaging off
9-bit BW Filter off Interpolation on

Channel 1 Scale 2.00 V/div Offset 1.99 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 1.000 : 1 Atten units ratio Skew 0.0 s

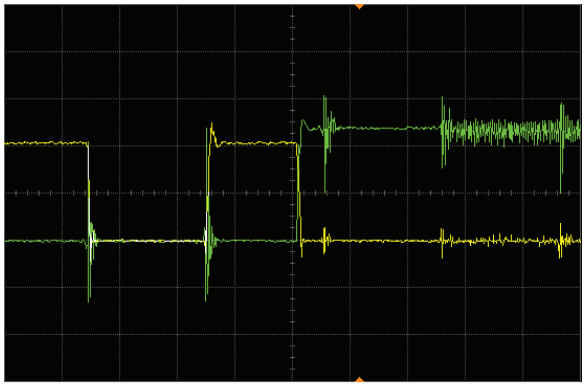
Channel 2 Scale 2.00 V/div Offset 2.02 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 1.000 : 1 Atten units ratio Skew 0.0 s

External Range 10 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 10.00 : 1 Atten units ratio Skew 0.0 s

Time base Scale 1.00 µs/div Position -2.195000 µs Reference center

Trigger Mode edge Sweep triggered
Hysteresis normal Holdoff time 60 ns Coupling HF Reject
Source external Trigger level 346 mV Slope rising

Rys. 6. Przykład powstawania błędów w pracy układu dla częstotliwości nośnej sygnału zakłócającego f=30MHz



Acquisition Sampling mode real time
Memory depth automatic Memory depth 1004pts
Sampling rate automatic Sampling rate 100 MSa/s
Averaging off
9-bit BW Filter off Interpolation on

Channel 1 Scale 2.00 V/div Offset 1.99 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 1.000 : 1 Atten units ratio Skew 0.0 s

Channel 2 Scale 2.00 V/div Offset 2.02 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 1.000 : 1 Atten units ratio Skew 0.0 s

External Range 10 V
BW limit off Coupling DC Impedance 1M Ohm
Attenuation 10.00 : 1 Atten units ratio Skew 0.0 s

Time base Scale 1.00 µs/div Position -1.165000 µs Reference center

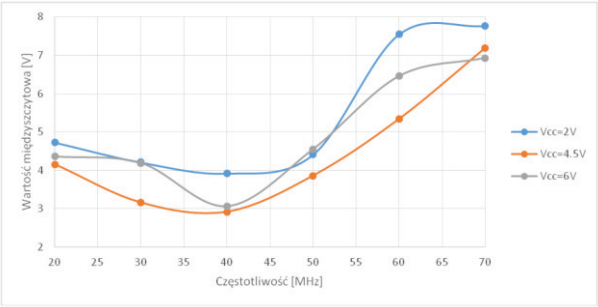
Trigger Mode edge Sweep triggered
Hysteresis normal Holdoff time 60 ns Coupling HF Reject
Source external Trigger level 346 mV Slope rising

Rys. 7. Przykład powstawania błędów w pracy układu dla częstotliwości nośnej sygnału zakłócającego f=70 MHz

Zbadano również wpływ na pracę układu przy podawaniu zakłócenia na wejście zasilające Vcc=6V. Zakłócenie wystąpiło przy częstotliwości nośnej sygnału 10MHz i wartości międzyszczytowej napięcia 3.76V oraz częstotliwości nośnej 20MHz i wartości międzyszczytowej napięcia równej 4.03V. Na rysunku poniżej przedstawiono przykład błędnej pracy układu dla tego przypadku zakłócenia.

4. WNIOSKI

Na rysunku poniżej (Rys. 8) przedstawiono wykres obrazujący zależność granicznego napięcia międzyszczytowego zakłócenia od częstotliwości nośnej sygnału zakłócającego, przy których zachodziło do błędnej pracy układu dla różnych napięć zasilania Vcc.



Rys. 8. Zależność granicznego napięcia międzyszczytowego w funkcji częstotliwości nośnej sygnału zakłócającego, przy którym zachodziło do błędnej pracy układu dla różnych napięć zasilających.

Z przeprowadzonych badań wynika, że krótkotrwałe impulsy w.cz. mogą powodować błędną pracę układu logicznego. Zakłócenie podawane na wejście dynamiczne jest szczególnie skuteczne. Zakłócenia podawane na wejście zasilające a zarazem na wejścia informacyjne jest mniej efektywne (nie dochodziło do błędów w pracy układu dla częstotliwości nośnej zakłócenia powyżej 20 MHz dla miedzyszczytowych wartości napięcia poniżej 7,5 V).

Z przeprowadzonych badań wynika, że wzrost częstotliwości nośnej zakłócenia wymaga większych wartości amplitud napięcia tego sygnału by wywołać błędną prace układu.

Należy podkreślić, że obserwowane zakłócenie pojawia się na wszystkich wyprowadzeniach układu lecz nie zawsze powoduje błędy w logice jego działania.

Zakłócenia w każdym przypadku są zdecydowanie większe niż marginesy zakłóceń. Układ badany przy niższych wartościach napięć zasilania pomimo mniejszych wartości marginesów zakłóceń wykazywał większą odporność na zakłócenia. Należy zauważyć, że przy niższych napięciach zasilania występują większe wartości pojemności złączy PN spolaryzowanych zaporowo. Większe pojemności układu stanowią mniejszą impedancję dla sygnałów w.cz.

W dalszej kolejności należałoby przeprowadzić badania w większym zakresie częstotliwości. W tym celu należy zastosować inną aparaturę pomiarową a w szczególności generator lub wzmacniacz umożliwiający wygenerowanie sygnałów o wyższej wartości międzyszczytowej napięcia. Należy również przeprowadzić podobne badania dla innych rodzin układów cyfrowych.

Artykuł zawiera wyniki realizacji projektu finansowanego przez NCBR (DOB-BIO2/PS/5/2/2016).

LITERATURA

[1] ”Electrical and electromagnetic environmental conditions”, AECTP 250, edition 2, Leaflet 257 ”HPM”, NATO standardization agency, January 2011.
[2] ”Electromagnetic Compatibility (EMC) – Part 2-13: Environment – High-power electromagnetic (HPEM) environments – Radiated and conducted”, IEC 61000-2-13, 1st edition, International Electrotechnical Commission, March 2005.
[3] ”Electromagnetic environmental effects requirements for systems”, MIL-STD-464C, Department of Defense Interface Standard, 1 December 2010.
[4] ”Requirements for the control of electromagnetic interference characteristics of subsystems and equipment”, MIL-STD-461F, Department of Defense Interface Standard, 10 December 2007.
[5] Hurtig T., Adelöw L., Akyuz M., Elfsberg M., Larsson A., ”Methodology for destructive HPM testing”, Electronic Environment, #1.2015, pp 22-27 (2015).
[6] Suhrke M., “The threat of high power microwaves to infrastructure” presentation at the Directed Energy Systems 2013, London, United Kingdom, 6-7 March 2013.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

STRATEGIE TRENINGU NEURONOWEGO ESTYMATORA CZĘSTOTLIWOŚCI TONU KRTANIOWEGO Z UŻYCIEM GENERATORA SYNTETYCZNYCH SAMOGŁOSEK

TRAINING STRATEGIES OF NEURAL FUNDAMENTAL FREQUENCY ESTIMATOR USING A SYNTHETIC VOWELS GENERATOR

Marek Blok[1]; Jan Banaś[2]; Mariusz Pietrołaj[3]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, marek.blok@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jan.banas@pg.edu.pl
[3] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, mariusz.pietrolaj@pg.edu.pl

DOI: 10.15199/59.2022.4.114

Streszczenie: **W wielu zastosowaniach telekomunikacyjnych pojawia się problem przetwarzania lub analizy sygnału mowy, w ramach którego, często w obszarze podstawowych algorytmów, stosuje się estymator częstotliwości tonu krtaniowego. Estymator rozpatrywany w tej pracy bazuje na neuronowym klasyfikatorze podejmującym decyzje na podstawie częstotliwości oraz mocy chwilowej wyznaczanych w podpasmach analizowanego sygnału mowy. W pracy rozważamy problematykę treningu tego estymatora, gdy trening odbywa się z użyciem sygnałów generowanych syntetycznie.**

Abstract: **In many telecommunication applications there is a need for a speech signal processing or analysis, within which the pitch tone frequency estimator is one of the common basic algorithms. The estimator considered in this paper is based on a neural classifier, whose decisions are driven by the instantaneous frequency and power determined in the sub-bands of the analyzed speech signal. In the paper, we consider the problems of selecting a training strategy for this estimator, when training is carried out with synthetically generated vowels.**

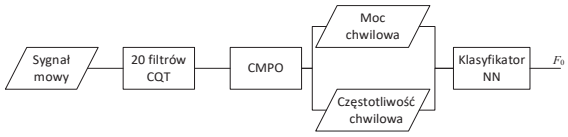
Słowa kluczowe: **estymacja częstotliwości tonu krtaniowego, klasyfikator neuronowy, częstotliwość chwilowa, trening sieci neuronowej.**

Keywords: **pitch estimation, neural classifier, instantaneous frequency, neural network training.**

1. WSTĘP

Jednym z dobrze rozpoznanych, lecz wciąż nierozwiązanych problemów w obszarze przetwarzania sygnałów mowy jest estymacja częstotliwości podstawowej tonu krtaniowego. Niezawodny algorytm realizujący to zadanie miałby szansę wspomóc rozwiązanie wielu problemów z zakresu przetwarzania sygnałów dźwiękowych, takich jak redukcja wymaganej przepustowości kanału transmisji sygnału poprzez kodowanie i resyntezę; poprawa jakości życia osób niesłyszących poprzez korektę intonacji, zniekształcanej przez aparat słuchowy bądź implant ślimakowy; separacja mówców w sygnale mowy dla zastosowań multimedialnych, takich jak telekonferencje. Niniejszy artykuł stanowi opis usprawnionego algorytmu IFE [1] - nowatorskiego podejścia do rozwiązania problemu estymacji częstotliwości podstawowej tonu krtaniowego. Opiera się ono o dwustopniowe przetwarzanie sygnałów dźwiękowych, którego schemat przedstawiono na rys. 1.

Pierwszy stopień stanowi przetwarzanie wstępne (*front-end*), realizujące ekstrakcję cech sygnału z rozdzielczością czasową rzędu pojedynczej próbki; następnie wyodrębnione cechy, zgrupowane w segmenty, stanowią wejście klasyfikatora neuronowego (*back-end*), realizującego przypisanie każdej próbki do jednej z predefiniowanych klas, odpowiadających spodziewanym wartościom częstotliwości tonu krtaniowego sygnału F_0. Klasyfikator został zaimplementowany jako sztuczna sieć neuronowa. Modyfikacją względem rozwiązania zaproponowanego w pracy [1] jest dodanie mechanizmów, które w ramach treningu pozwalają na generację syntetycznych danych treningowych na żądanie. Takie podejście umożliwia teoretycznie nieskończenie długi trening, co jest istotną zmianą, zwłaszcza w stosunku do treningów korzystających z rzeczywistych nagrań mowy, dla których istotnym problemem jest pozyskanie adnotacji dotyczącymi częstotliwości podstawowej tonu krtaniowego. Co więcej, ten sposób daje pewność co do referencyjnej wartości częstotliwości podstawowej tonu krtaniowego - w przeciwieństwie do istniejących korpusów z nagraniami, których adnotacje, powstałe na podstawie sygnału elektroglotograficznego, często obarczone są błędami wynikającymi również z samego mechanizmu powstawania dźwięcznych głosek [7].



Rys. 1. Schemat przetwarzania IFE

2. GENEROWANIE DANYCH

2.1. Synteza samogłosek

Kluczowym usprawnieniem IFE przedstawionym w niniejszym artykule jest metoda dostarczania danych treningowych do sieci neuronowej. Dzięki zastosowaniu syntezy samogłosek, dane mogą być generowane w czasie treningu, co pozwala na dostarczenie tak dużej ilości danych, jak jest żądane oraz eliminuje konieczność korzystania z dużych zasobów pamięci w czasie treningu.

$$A[k] = \left\{ \sin\left[\left(2\pi \frac{k \cdot F_0}{F_s} \right)^v \right] \right\}^{s^4} \quad (1)$$

Synteza samogłosek jest dwuetapowym procesem, który rozpoczyna się od poliharmonicznego generatora amplitudy, opisanego powyższym wzorem (1), w którym k opisuje indeks harmonicznej $F_k = k \cdot F_0$, ograniczonej w zakresie od losowo wygenerowanej F_0 do częstotliwości Nyquista $F_N = \frac{1}{2} F_s$, zależnej od szybkości próbkowania F_s. Dodatkowo wprowadzono losowe czynniki kształtujące rezonans samogłoski, v oraz s, których zakresy losowania zdefiniowano jako $v \in [0.15, 0.4]$ i $s \in [1.4, 2.5]$. Ostatecznie przebieg czasowy syntetycznej samogłoski jest generowany zgodnie ze wzorem

$$y[n] = \sum_{k=1}^{K} A[k] e^{j(k\hat{\varphi}[n] + \varphi_0[k])}, \quad (2)$$

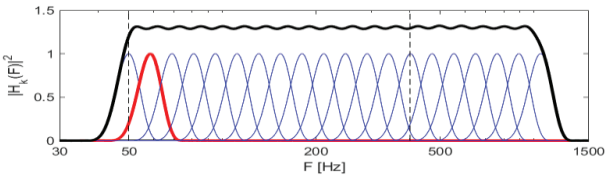
gdzie

$$\hat{\varphi}[n] = \frac{2\pi}{F_s} \sum_{m=0}^{n} F[m], \quad (3)$$

początkowa faza $\varphi_0[\cdot]$ jest losowana a chwilowa częstotliwość podstawowa $F[\cdot]$ ma losową wartość początkową i nieliniowo zmienia się w czasie. Kolejny etap addytywnej syntezy to dodanie zniekształceń typu shimmer i jitter [6] oraz dodanie szumu białego o kontrolowanym poziomie, reprezentującego naturalne zakłócenia powstające w ramach wypowiedzi jak i przy rejestracji sygnału, pozwalającym na regulację wynikowego stosunku poziomu sygnału użytecznego do szumu (SNR).

2.2. Przetwarzanie wstępne - ekstrakcja cech

Zarówno syntetyczne samogłoski wygenerowane w sposób opisany powyżej, jak i nagrania rzeczywistej mowy, zostają podane wstępnemu przetwarzaniu, którego celem jest ekstrakcja cech sygnału na potrzeby sieci neuronowej. Celem jest uzyskanie wartości chwilowej mocy i częstotliwości dla każdej kolejnej próbki analizowanego sygnału, w każdym z 20 podpasm częstotliwościowych. Przetwarzanie to obejmuje dwa etapy – filtrację oraz obliczanie zespolonej mocy wzajemnej.



Rys. 2. Bank filtrów przetwarzania wstępnego

Pierwszy etap polega na podaniu sygnału wejściowego na 20 filtrów o stałej dobroci, częstotliwościach środkowych rozłożonych w zakresie od 50Hz do 1050Hz. Są to filtry Hilberta o skończonej odpowiedzi impulsowej, generujące na wyjściu sygnały o wartościach zespolonych. Filtry te zostały zaprojektowane metodą jądra CQT (*Constant-Q Transform*) [5] z użyciem okna Blackmana, jak pokazano na rys. 2.

W każdym z podpasm wydzielonych w wyniku powyższej filtracji, stosowany jest operator zespolonej mocy wzajemnej (*complex mutual power operator - CMPO*) zdefiniowany jako:

$$y_{\mathrm{CMPO}}[n] = x[n] \cdot x^*[n-1], \quad (4)$$

gdzie $x[n]$ próbką wyjściową z jednego z filtrów, a $x^*[n]$ jej wartością sprzężoną. Na podstawie CMPO można łatwo obliczyć moc chwilową

$$P_i[n] = |y_{\mathrm{CMPO}}[n]| \quad (5)$$

i częstotliwość chwilową

$$F_i[n] = \arg(y_{\mathrm{CMPO}}[n]). \quad (6)$$

Tak obliczone wartości nie oferują wystarczającej dokładności by stanowić samodzielny estymator częstotliwości podstawowej tonu krtaniowego, niosą one jednak wystarczająco dużo informacji dotyczącej dominujących zakresów częstotliwości (moce chwilowe) i prawdopodobnych chwilowych wartości częstotliwości (częstotliwości chwilowe), żeby użyć ich do dalszego przetwarzania z użyciem sztucznej sieci neuronowej.

3. SIEĆ NEURONOWA

Przedstawiona metoda określania podstawowej częstotliwości tonu krtaniowego bazuje na algorytmie sztucznej inteligencji zaimplementowanym w postaci sieci neuronowej. Trenowany model wykorzystuje architekturę wielowarstwowego perceptronu, zawierającego dwie ukryte warstwy liczące odpowiednio 250 i 500 neuronów [1]. Topologia warstwy wejściowej jest podyktowana rozmiarem danych zwracanych przez generator danych treningowych, zakłada on 40 wartości odpowiadającym częstotliwości i energii każdej próbki sygnału.

3.1. Trening i klasyfikacja

Inferencja sieci neuronowej przyporządkowuje każdą z analizowanych próbek sygnału jako jedną z 351 predefiniowanych wartości częstotliwości. Odpowiadają one przedziałom o logarytmicznej dystrybucji w zakresie od 50 do 400 Hz. Decyzja o zastosowaniu skali logarytmicznej została podyktowana charakterystyką sygnału mowy i większym skupieniu F_0 w niższych przedziałach częstotliwości [1]. Ważnym wyróżnikiem prezentowanej metody jest użycie wyłącznie syntetycznie generowanych danych do treningu i testowania sieci neuronowej. Tylko ostateczna ewaluacja i porównanie zastosowanego rozwiązania bazuje na rzeczywistych sygnałach mowy takich jak baza danych Keele Pitch [2]. Zastosowany mechanizm generacji danych pozwala na wykorzystanie nowego zestawu treningowego na każdą z epok treningu.

W przypadku prezentowanych wyników poziom SNR generowanych sygnałów był analogiczny do danych wejściowych zastosowanych w IFE i wynosił 40 dB [1]. W zależności od wariantu eksperymentu, unikalny zestaw treningowy pojedynczej epoki zawierał od 5000 do 50000 próbek, co w najdłuższym wariancie odpowiada około 7 godzinom ciągłego sygnału audio. Przed podaniem parametrów na warstwę wejściową sieci były one poddawane procesowi normalizacji w zakresie od 0 do 1 w dziedzinie częstotliwości i energii. Syntetyczny zestaw walidacyjny obejmował 1000 syntetycznych głosek i był używany wyłącznie do zweryfikowania dokładności trenowanej sieci.

3.2. Parametryzacja sieci

Proces treningu obejmował zróżnicowaną liczbę epok z uwzględnieniem podziału danych wejściowych na zestawy po 128 próbki (*batch size*). W celu aktywacji wykorzystano funkcję ReLU. Obliczenie funkcji straty opierało się na entropii skrośnej (*cross entropy*) i klasyfikacji przy pomocy funkcji softmax. Zastosowany parametr szybkości uczenia (*learning rate*) wynosił 10^{-4}. Omawiane rozwiązanie z uwzględnieniem treningu, walidacji i ewaluacji zostało zaimplementowane przy pomocy języka Python i środowiska PyTorch [3].

4. EKSPERYMENTY

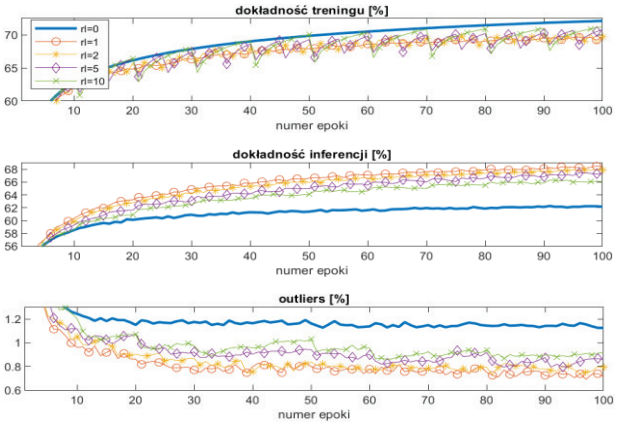
Wykorzystanie syntetycznych danych treningowych generowanych na żądanie pozwala na użycie w każdej kolejnej epoce nowego ich zestawu. Pozwala to uniknąć przeuczenia sieci dla danego zestawu treningowego. Problem ten jest szczególnie istotny, gdy dostępne dane treningowe są ograniczone. Przykładowo użyta do walidacji metody IFE baza Keele Pitch zawiera 10 wypowiedzi, 5 żeńskich i 5 męskich, o łącznym czasie trwania 337 sekund, z czego 159 sekund to segmenty dźwięczne opatrzone informacją o częstotliwości tonu krtaniowego. Dla porównania do treningu sieci w [1] użyto 30 000 półsekundowych segmentów (syntetycznych głosek).

W ramach prac nad metodą IFE zauważyliśmy, że chociaż wielokrotne podawanie tego samego zestawu danych prowadzi do przetrenowania sieci, to jednak, jeżeli liczba segmentów jest dostatecznie duża, kilkukrotne wykorzystanie tego samego zestawu danych pozwalało na uzyskanie poprawy jakości inferencji. Maksymalny rozmiar zestawu treningowego ogranicza jednak rozmiar dostępnej pamięci, zwłaszcza, że na wyjściu przetwarzania wstępnego na każdą próbkę sygnału treningowego otrzymujemy 40 wartości analizowanych przez sieć.

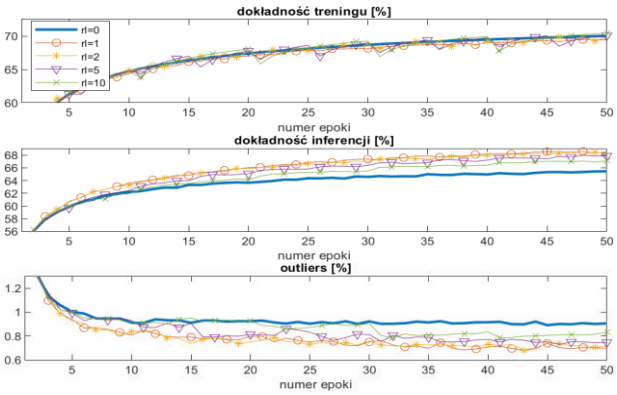
Problem ten skłonił nas do przeanalizowania efektywności rozwiązania, w ramach którego syntetyczne dane treningowe są generowane w mniejszych segmentach na potrzeby treningu kolejnych epok. W trakcie eksperymentów zmienialiśmy zarówno liczbę syntetycznych głosek używanych do treningu w ramach danej epoki N_{seg} (5000, 10000, 25000 oraz 50000), jak i częstość ponownego użycia danych treningowych w kolejnych epokach rl, gdzie 0 oznacza użycie tych samych danych we wszystkich epokach, a większa wartość określa, co ile epok generowany jest nowy sygnał. Do walidacji wytrenowanej sieci użyto danych syntetycznych złożonych z 1000 półsekundowych głosek o SNR równym 40dB.

Jak można zauważyć na rys. 3-6 dla małego N_{seg} (rys. 3 i 4) zmiana zestawu danych treningowych skutkuje skokowym pogorszeniem dokładności obserwowanej podczas treningu oraz drobnemu skokowi w górę w dokładności inferencji. Ponadto wielokrotne użycie tego samego zestawu danych treningowych skutkuje zwiększeniem się udziału w estymatach F_0 wartości odchylonych od wartości referencyjnej o więcej niż 20% (*outliers*), który przy przełączaniu zestawu skokowo spada.

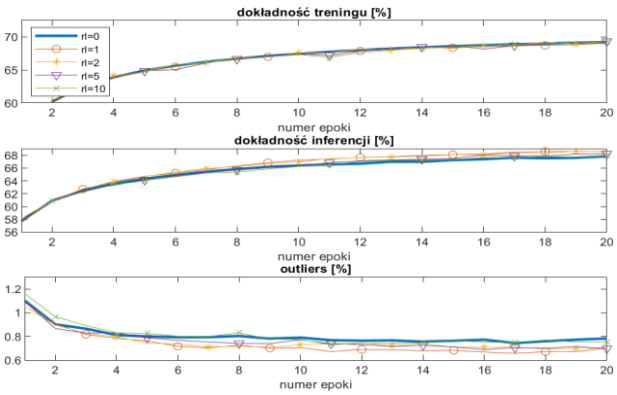
Biorąc pod uwagę jakość inferencji wytrenowanej sieci, to najlepsze efekty, dla krótkich zestawów danych treningowych, daje ciągła wymiana danych treningowych. Różnice pomiędzy strategiami treningu zacierają się wraz ze wzrostem wielkości zestawu treningowego (rys. 5 i 6). Dla N_{seg} = 50 000 co prawda ponownie najlepsza strategia to generowanie nowych danych co epoka (rl = 1), ale różnice nie są tym razem tak duże w porównaniu do innych strategii. Ta wielkość zestawu jest bliska maksymalnej wielkości segmentu pozwalającej na trening na komputerze z 64GB RAMu, a jednocześnie zbliżone wyniki uzyskano dla innych wartości N_{seg}, w tym 5000, dla rl = 1, jednak z nieco zwiększoną liczbą outlierów.



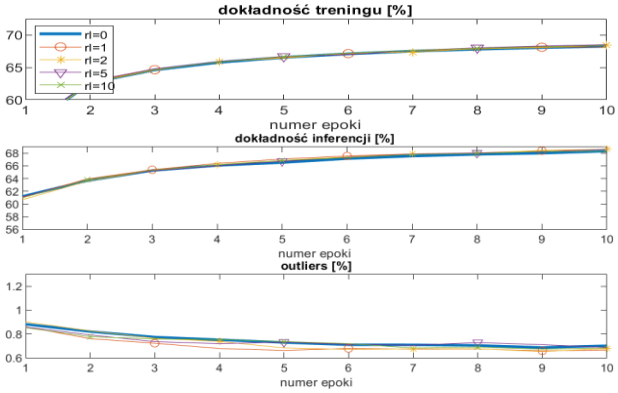
Rys. 3. Dokładność treningu i inferencji oraz procent znacznych błędów dla treningu z N_{seg}= 5 000



Rys. 4. Dokładność treningu i inferencji oraz procent znacznych błędów dla treningu z N_{seg}= 10 000



Rys. 5. Dokładność treningu i inferencji oraz procent znacznych błędów dla treningu z N_{seg}= 25 000



Rys. 6. Dokładność treningu i inferencji oraz procent znacznych błędów dla treningu z N_{seg}= 50 000

Lekka przewaga treningu z największym N_{seg} prawdopodobnie wynika z tego, że tasowanie danych treningowych silniej zaburza korelację pomiędzy danymi dla kolejnych próbek sygnału. Sygnały te co prawda cechują się zmiennymi parametrami, ale zachowują ciągłość tych zmian w ramach poszczególnych głosek, na które składa się aż 4000 próbek.

Trening z N_{seg} = 5 000 przy ciągłej resyntezie danych pozwala na trening ze znacznie mniejszymi wymaganiami pamięciowymi z kolei trening z $N_{seg} \geq$ 50 000 pozwala na wielokrotne wykorzystanie raz wygenerowanych danych treningowych.

W tab. 1 i 2 przedstawiono wyniki pomiaru dokładności estymacji F_0 dla wypowiedzi z bazy Keele Pitch. Jest to procent estymat mieszczących się w 5% przedziale wokół wartości referencyjnej. W zestawieniu uwzględniono cztery strategie treningu: (a) N_{seg} = 5 000 oraz rl = 0, (b) N_{seg} = 5 000 oraz rl = 1, (c) N_{seg} = 50 000 oraz rl = 0 i (d) N_{seg} = 50 000 oraz rl = 1.

Tab. 1. Procent poprawnych estymat F_0 dla tolerancji ±5% dla wypowiedzi żeńskich

	f1	*f2*	*f3*	*f4*	*f5*
(a)	91,1	95,4	93,2	91,3	94,5
(b)	91,8	95,6	93,0	92,4	96,4
(c)	91,2	95,6	93,5	92,3	94,5
(d)	91,1	95,5	92,7	92,1	94,0

Tab. 2. Procent poprawnych estymat F_0 dla tolerancji ±5% dla wypowiedzi męskich

	m1	*m2*	*m3*	*m4*	*m5*
(a)	69,2	85,9	91,6	87,5	85,0
(b)	71,6	87,1	93,5	88,4	86,9
(c)	71,5	86,2	92,6	89,3	87,9
(d)	71,8	86,1	92,5	89,8	86,8

W tym przypadku można zauważyć, że trening dla krótkiego segmentu z ciągłą generacją nowego zestawu (b) poza sygnałem f3, dla którego odnotowano niewielkie pogorszenie, skutkuje poprawą w porównaniu do ciągłego wykorzystania tego samego zestawu danych treningowych (a). Średnio wyniki są lepsze o 0,7% dla głosów żeńskich oraz 1,7% dla głosów męskich, z najlepszą poprawą dla głosu m1 (najtrudniejszego przypadku) wynoszącą 2,4%. Z kolei dla długich segmentów (c i d) różnice są wyraźnie mniejsze, a wyniki są zbliżone (wahania w zakresie 1%) do wyników z wariantu (b), z wyróżniającymi się wynikami dla wypowiedzi f5, dla której strategia (b) dała prawie 2% poprawę.

5. PODSUMOWANIE

W ramach przedstawionych w pracy badań zweryfikowano różne strategie treningu sztucznej sieci neuronowej estymującej częstotliwość tonu krtaniowego z wykorzystaniem możliwości generowania danych treningowych na żądanie. Warto zwrócić uwagę na to, że wykorzystanie w treningu syntetycznych głosek, z jednej strony jest skuteczne, co potwierdzają przedstawione wyniki, a z drugiej strony jest konieczne, bo trening sieci wymaga kilku godzin dźwięcznych segmentów mowy opatrzonych wiarygodną wartością F_0, co stanowi dużym problem. Jednocześnie nawet korzystając z dużego zestawu danych treningowych problem stanowią wymagania pamięciowe, które można ograniczyć realizując trening dla mniejszych zestawów danych generowanych na bieżąco w trakcie treningu. Jak pokazano, mały zestaw danych treningowych może być skutecznie użyty do treningu, o ile wymiana danych treningowych następuje dla każdej epoki. Taka strategia pogarsza osiągi obserwowane bezpośrednio w ramach treningu, ale poprawia wyniki inferencji, które są porównywalne z wynikami uzyskiwanymi dla treningu z wykorzystaniem dużych segmentów danych.

LITERATURA

[1] Blok, Marek, Jan Banas, and Mariusz Pietrolaj. 2021.“IFE: NN-Aided Instantaneous Pitch Estimation.”. 14th International Conference on Human System Interaction (HSI).

[2] “Keele Pitch Database”. https://lost-contact.mit.edu/afs/nada.kth.se/dept/tmh/corpora/KeelePitchDB/. Accessed April 23, 2022.

[3] “Pytorch.” PyTorch. https://pytorch.org/. Accessed April 23, 2022. – C

[4] Rosenblatt, F. 1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. Psychological Review 65, no. 6: 386–408.

[5] Schörkhuber, Christian, and Anssi Klapuri. 2010. “Constant-Q Transform Toolbox for Music Processing”. 7th Sound and Music Computing Conference.

[6] Teixeira, João Paulo, Carla Oliveira, and Carla Lopes. 2013. “Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters”. Procedia Technology 9: 1112–22.

[7] Veprek, Peter, and Michael S. Scordilis. 2002. "Analysis, enhancement and evaluation of five pitch determination techniques". Speech Communication 37.3-4: 249-270.

SUBIEKTYWNY POMIAR JAKOŚCI PROGRAMÓW RADIOWYCH STRUMIENIOWANYCH W SIECI METODĄ CROWDSOURCINGU

SUBJECTIVE QUALITY EVALUATION OF RADIO PROGRAMS STREAMED ONLINE USING CROWDSOURCING

Przemysław Falkowski-Gilski

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki,
ul. Narutowicza 11/12, 80-233 Gdańsk, przemyslaw.falkowski@eti.pg.edu.pl

DOI: 10.15199/59.2022.4.115

Streszczenie: **Obecnie słuchacze mają dostęp do swoich ulubionych programów i audycji radiowych za pośrednictwem naziemnego standardu analogowego FM (Frequency Modulation) oraz cyfrowego DAB+ (Digital Audio Broadcasting plus). Należy podkreślić, że ten sam materiał nadawany jest jednocześnie w kilku technikach (tzw. *simulcast*), a znaczna większość rozgłośni udostępnia swoje programy także online. Niniejsza praca przedstawia wyniki badań dotyczących jakości transmisji programów radiowych strumieniowanych w sieci. Kampanię pomiarową przeprowadzono metodą tzw. crowdsourcingu przy udziale 45 słuchaczy.**

Abstract: **Currently, listeners have access to their favorite radio programs and broadcasts via terrestrial analog FM (Frequency Modulation) and digital DAB+ (Digital Audio Broadcasting plus) radio. It should be emphasized that the same content is emitted simultaneously in various techniques (so-called simulcast), whereas most broadcasters offer their stations also in the web. This paper presents results of a quality evaluation of radio programs streamed online. The measurement campaign was carried out using crowdsourcing, and involved 45 listeners.**

Słowa kluczowe: **pomiary jakości, radiofonia, simulcast, streaming, transmisja multimediów.**

Keywords: **broadcasting, multimedia transmission, quality evaluation, simulcast, streaming.**

1. WSTĘP

Radiofonia jest obecna w naszym życiu już od ponad 100 lat. Transmisja sygnałów mowy i muzyki towarzyszy nam od najmłodszych lat, umilając m.in. podróż czy wypoczynek. Audycje radiowe dostarczają najnowszych informacji z kraju i ze świata. Kształtują także gusta muzyczne i pozwalają wielu artystom dzielić się swoją twórczością.

Współczesna radiofonia obejmuje transmisję w szeregu systemów naziemnych. Do najpopularniejszych należy analogowy FM (Frequency Modulation) oraz cyfrowy DAB+ (Digital Audio Broadcasting plus) [5]. Warto podkreślić, że ten sam materiał nadawany jest jednocześnie w kilku technikach. Określane jest to angielskim terminem *simulcast*. Dodatkowo znaczna większość rozgłośni radiowych emituje ten sam materiał także w sieci. Ich sygnał dostępny jest z poziomu dedykowanej aplikacji mobilnej na popularne urządzenia z systemem operacyjnym Android lub iOS. Z kolei inni nadawcy udostępniają swoje audycje poprzez własną stronę internetową lub portale zrzeszające kilka programów radiowych. Podobną sytuacją obserwujemy zarówno na rynku lokalnym (regionalnym) jak i ogólnokrajowym w Polsce, Europie i na świecie.

2. KODOWANIE, KOMPRESJA I FORMOWANIE SYGNAŁU

W przeciwieństwie do analogowej radiofonii FM, sygnały cyfrowe grupowane są w zbiór, określany mianem *ensemble*. Obecnie słuchacze w Polsce mają dostęp do dwóch multipleksów cyfrowych: nadawcy publicznego (Polskie Radio) oraz komercyjnego (szereg podmiotów).

W multipleksie DAB+ publicznego nadawcy dostępne są główne i najpopularniejsze programy radiowe, dobrze znane z klasycznej radiofonii FM. Z kolei na multipleksie komercyjnym pojawiły się zarówno stacje radiowe o ugruntowanej pozycji na rynku, jak i nowi gracze zainteresowanie procesem cyfryzacji tego medium. Należy podkreślić, że koszty związane z prowadzeniem regularnej emisji programów radiowych są znacznie niższe w przypadku standardu cyfrowego DAB+.

Cały proces oraz opłaty związane z pobieraniem i przetwarzaniem sygnałów fonii oraz stowarzyszonych danych dodatkowych, a także formowaniem i emisją sygnału zbiorczego multipleksu, może zostać rozdzielony pomiędzy kilku lub nawet kilkunastu udziałowców. Sumaryczna przepływność netto wynosząca 1152 kbit/s może być elastycznie rozdysponowana w zależności od profilu stacji, typu realizowanej audycji, emitowanego gatunku muzycznego, itp., wedle ustalonych potrzeb.

Innym podejściem związanym z cyfrową emisją sygnału radiowego jest udostępnianie audycji i treści w formie tzw. *podcastu*. Słuchacz otrzymuje możliwość odsłuchania uprzednio zarejestrowanego nagrania z poziomu przeglądarki internetowej lub aplikacji mobilnej. Biorąc pod uwagę liczbę dostępnych programów radiowych oraz bogactwo tworzonych przez nie treści, ta forma jest jak najbardziej uzasadniona.

W przypadku emisji audycji w sieci w czasie rzeczywistym, tzw. *streaming*, sygnały mowy i muzyki oraz stowarzyszone z nimi multimedia (np. okładka płyty, informacje o autorze, itp.) są dostępne na żywo i online. Oczywiście odmienne jest medium transmisyjne.

Z tego względu, biorąc także pod uwagę wszechobecność urządzeń przenośnych i bezprzewodowych, postanowiono przeprowadzić badania dotyczące subiektywnego pomiaru jakości programów radiowych strumieniowanych w sieci.

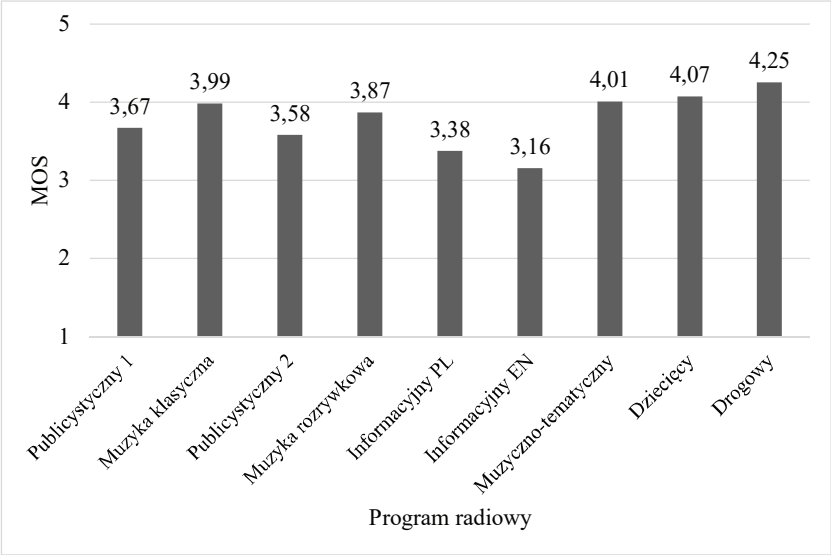
3. BADANIA

Badania przeprowadzono w kwietniu 2022 roku przy udziale 45 uczestników w wieku 20-35 lat zgodnie z metodą tzw. *crowdsourcingu* [1, 6, 7]. Podejście to bazuje na fakcie rozproszenia użytkowników, którzy mogą konsumować treści i korzystać z dostępnych zasobów przy pomocy własnego urządzenia. Podczas testów nie wyznaczono konkretnego sprzętu, o danych parametrach czy danej klasie, do odbioru i odsłuchu programów radiowych. Nie narzucono także konkretnej godziny, np. godziny najwyższej słuchalności, tzw. *primetime*. Wszystkich uczestników poproszono natomiast o poświęcenie kilkunastu minut dla każdego wskazanego programu radiowego. Listę wytypowanych programów i ich profili, dostępnych z poziomu dedykowanej platformy Player Polskiego Radia, przedstawiono w Tabeli 1.

Tab. 1. Badane programy radiowe

Lp.	*Profil programu*	*Emisja*
1	Publicystyczny 1	FM/DAB+/streaming
2	Muzyka klasyczna	FM/DAB+/streaming
3	Publicystyczny 2	FM/DAB+/streaming
4	Muzyka rozrywkowa	FM/DAB+/streaming
5	Informacyjny PL	FM/DAB+/streaming
6	Informacyjny EN	DAB+/streaming
7	Muzyczno-tematyczny	DAB+/streaming
8	Dziecięcy	DAB+/streaming
9	Drogowy	DAB+/streaming

Portal ten udostępnia 9 programów radiowych, prowadzących regularne emisje. Część z nich dostępnych jest równolegle w standardzie FM i/lub DAB+. Użytkowników poproszono o zaznajomienie się z możliwościami usługi Player Polskiego Radia, a następnie o dokonanie kilkunastominutowego odsłuchu każdego programu radiowego. Celem było określenie ogólnego poziomu jakości emitowanych audycji oraz profilu danej stacji. Wyniki badań zostały zapisane w formularzu online, a następnie przetworzone w celu określenia oceny w 5-stopniowym wariancie skali MOS (Mean Opinion Score). Rezultaty przedstawiono na rys. 1.



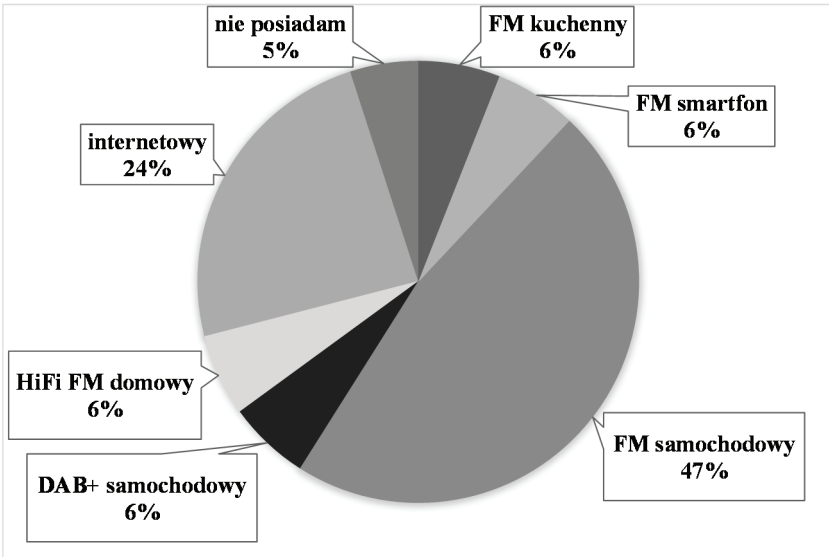
Rys. 1. Wyniki badań subiektywnych dotyczących ogólnej jakości programów radiowych emitowanych w sieci

Na podstawie otrzymanych wyników można określić, że średni poziom realizowanych audycji w przypadku pierwszych 6 programów radiowych (Publicystyczny 1, Muzyka klasyczna, Publicystyczny 2, Muzyka rozrywkowa, Informacyjny PL, Informacyjny EN), oceniono na poziomie od 3 do 4 (dość dobry do dobry). Spośród 9 badanych programów radiowych, jedynie 3 z nich (Muzyczno-tematyczny, Dziecięcy, Drogowy), przekroczyło próg 4 punktów (ocena powyżej dobrej). Powodem tego faktu był z pewnością różny profil każdej ze stacji, a także związana z tym emisja odmiennego sygnału fonii. W przypadku niektórych dominowały sygnały mowy w języku polskim (Informacyjny PL) lub języku angielskim (Informacyjny EN). Z kolei w innych audycje słowne przeplatane były muzyką klasyczną i/lub rozrywkową.

Należy zaznaczyć, że w przypadku niektórych wywiadów goście, w przeciwieństwie do prowadzącego, rzadko byli obecni w studiu radiowym. Rozmowa realizowana była przy użyciu telefonii stacjonarnej, komórkowej lub łącza internetowego. Ich jakość niekiedy pozostawiała wiele do życzenia. Jednak w przypadku zdalnej komunikacji najistotniejszym czynnikiem jest prowadzenie wymiany informacji w czasie rzeczywistym, przy możliwie

najmniejszym opóźnieniu. Można również przyjąć, że gusta muzyczne, specyficzne dla grupy uczestników w wieku 20-35 lat, również miały wpływ na ich ogólną ocenę.

Po zakończeniu odsłuchu uczestników poproszono o odpowiedzenie na szereg pytań, w tym m.in. wskazanie liczby i/lub typu posiadanego odbiornika radiowego. Wyniki przedstawiono na rys. 2.



Rys. 2. Typ posiadanego odbiornika radiowego

Na podstawie udzielonych odpowiedzi można wskazać, że standard analogowy FM jest wciąż dominujący w naszym kraju. Ponad połowa badanych posiada taki odbiornik w kuchni (6%), w swoim smartfonie (6%), w salonie zintegrowany z zestawem domowym HiFi (również 6%) lub w samochodzie (aż 47%).

Jak się okazało, większość badanych to ludzie zmotoryzowani, przy czym tylko 6% użytkowników posiada odbiornik DAB+ w kokpicie pojazdu. Należy jednak zaznaczyć, że większość nowych sprzedawanych aut posiada odbiornik kompatybilny z cyfrowym standardem. W celu odblokowania funkcji odbioru sygnału DAB+ konieczne jest wykupienie odpowiedniego pakietu w salonie samochodowym.

Blisko jedna czwarta (24%) osób korzysta z tzw. radia internetowego, czy to z poziomu aplikacji mobilnej, przeglądarki internetowej, czy też zewnętrznego portalu lub usługi. Wśród dominujących odpowiedzi przeważał Spotify oraz YouTube.

Otrzymane wyniki wyraźnie wskazują, że możliwe jest polepszenie jakości świadczonych usług strumieniowania sygnałów audycji radiowych w sieci. Wskazane zdają się natomiast dalsze działania promujące standard cyfrowy DAB+, np. poprzez szersze wspieranie badań naukowych, działania i kampanie promocyjne oraz zacieśnienie współpracy z branżą samochodową.

4. DYSKUSJA

Biorąc pod uwagę zainteresowania naukowe oraz wykształcenie techniczne i muzyczne autora, postanowiono zapytać słuchaczy o ich odczucia po przeprowadzonych badaniach.

Według wielu kluczowe znaczenie miała jakość pierwotnego (oryginalnego) materiału dźwiękowego. W przypadku muzyki stwierdzono zależność pomiędzy wiekiem utworu a jego jakością. Jako ocenili słuchacze, im starszy był dany kawałek muzyczny, tym miał gorszą jakość. Z pewnością wpływ miała metoda (stratna, bezstratna) i algorytm kompresji materiału dźwiękowego. Kilku słuchaczy było na tyle dociekliwych, że porównali usłyszane w radiu utwory, dobrze im znane, z własną biblioteką muzyczną obejmującą pliki w formacie bezstratnym ALAC (Apple Lossless Audio Codec). Odnotowano znacznie lepszą dynamikę na korzyść ALAC.

Innym aspektem była kwestia jakości sygnałów mowy. Jak wiemy, sygnał ten ma za zadanie dostarczać informacje, jednak jego jakość w wielu przypadkach okazała się ogromnym rozczarowaniem. Dodatkowo bywała znacząco odmienna pomiędzy różnymi programami radiowymi i/lub audycjami w ramach tej samej stacji. Zauważalne było m.in.: wyraźne odcięcie pasma, kiepska akustyka pomieszczeń (pogłosy a nawet echa), dźwięki tła i świsty powietrza, czy też szumy i zanikający dźwięk (związany z jakością połączenia w sieci). W kwestii jakości mowy pozytywnie wyróżniały się programy dla grupy docelowej dzieci i kierowców. Natomiast w czasie emisji reklam, niezależnie od programu radiowego, jakość była perfekcyjna.

Spora grupa osób wspomniała, że woli korzystać z radia internetowego z uwagi na dostępną szerszą ofertę programową, ze szczególnym uwzględnieniem *podcastów*. Wielu z nich ceni sobie muzyczne audycje tematyczne oraz wywiady z ciekawymi gośćmi, w tym te w językach obcych, jak np. angielski, niemiecki, czy języki

skandynawskie. Sam odsłuch traktują jako udział w rozmowie z drugim człowiekiem, podczas której ćwiczą i trenują swoje umiejętności lingwistyczne.

Co ciekawe, część osób nie odnosiła swojej oceny do audycji dostępnych w naziemnej radiofonii FM lub DAB+. Wykazywali natomiast podobieństwa i różnice w stosunku do popularnych aplikacji, jak np. Spotify czy YouTube.

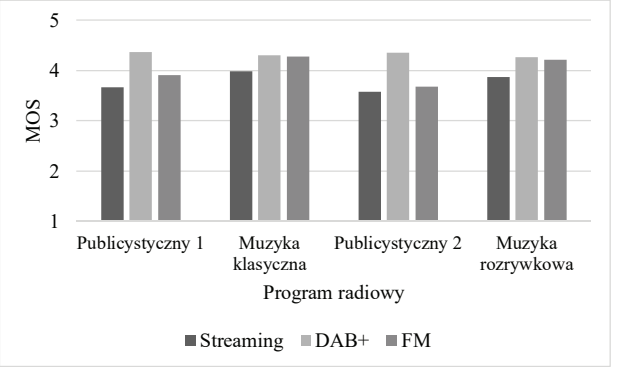
Kierowcy samochodowi wskazali na wystarczającą jakość zintegrowanych odbiorników FM. Jak nadmienili, w czasie jazdy główna uwaga skupiona jest na drodze, a przy typowym wygłuszeniu pojazdu i tak odnotowywane są szumy i zakłócenia z otoczenia. Z kolei ci sami ludzie, którzy konsumują treści w domu za pomocą sprzętu wysokiej jakości, byli w stanie opisać wiele różnic w dźwięku pomiędzy platformą Player PR, radiofonią naziemną FM i DAB+, czy też szeregu popularnych usług strumieniowych.

Postanowiono także przygotować zestawienie wyników ze wcześniejszymi badaniami opublikowanymi w roku 2021 [2]. Porównano oceny jakości wybranych programów radiowych nadawanych w sieci z tą samą grupą stacji w standardach FM oraz DAB+. W Tabeli 2 przedstawiono przepływności wybranych programów radiowych, które emitowane są również w standardzie DAB+.

Tab. 2. Przepływność programów radiowych emitowanych w standardzie DAB+

Lp.	*Profil programu*	*Przepływność DAB+ [kbit/s]*
1	Publicystyczny 1	112
2	Muzyka klasyczna	112
3	Publicystyczny 2	96
4	Muzyka rozrywkowa	96

Wyniki zestawienia, dla 4 programów radiowych, w tym: Publicystyczny 1, Muzyka klasyczna, Publicystyczny 2, Muzyka rozrywkowa, przedstawiono na rys. 3.



Rys. 3. Porównanie wyników subiektywnej oceny wybranych programów radiowych emitowanych w szeregu standardach

Jak pokazują wyniki, wyraźnie zauważalna jest wyższa jakości emisji cyfrowej w standardzie DAB+. Należy podkreślić, że przepływność bitowa przypadająca na poszczególne programy radiowe jest znacząco odmienna w porównaniu do strumieniowanych danych w sieci.

5. PODSUMOWANIE

Wyniki analizy w sposób ciekawy i praktyczny opisują zmiany jakie zaszły w przypadku transmisji programów radiowych oraz stowarzyszonych audycji słownych i muzycznych. Wszechobecność sieci bezprzewodowych, a także dostępność urządzeń przenośnych, zmieniły potrzeby i oczekiwania użytkowników. Wzrosły wymagania słuchaczy dotyczące niezawodnych usług, świadczonych na wysokim poziomie.

Otrzymane wyniki badań wykazują możliwość dalszej poprawy i rozwoju usług transmisji radiowej i stowarzyszonych treści. Praca posiada walory praktyczne, stanowiące cenne wsparcie dla nadawców publicznych i prywatnych w kwestii optymalizacji zarządzania dostępnymi zasobami. Warto zaznaczyć, że proces cyfryzacji radiofonii nie jest zakończony. Przewiduje się uruchomienie kolejnych multipleksów oraz dalszy rozwój rynku radiowego [8]. Z drugiej strony, dzięki globalnej sieci, każdy może stać się twórcą treści i nadawcą w Internecie [4].

Przyszłe badania mogłyby uwzględniać nowych nadawców i programy radiowe, które niedawno pojawiły się jako lokalne i ogólnopolskie rozgłośnie, a także szersze spektrum aktywności użytkowników, grup wiekowych, czy też wybranych usług strumieniowych [3].

LITERATURA

[1] Assis Neto, Fabio R., Celso A. S. Santos. 2018. „Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management". *Information Processing & Management* 54 (4) : 490–506.

[2] Falkowski-Gilski, Przemysław. 2021. „Digital transformation of terrestrial radio: An analysis of simulcasted broadcasts in FM and DAB+ for a smart and successful switchover". *Applied Sciences* 11 : 11114.

[3] Falkowski-Gilski, Przemysław, Tadeus Uhl. 2020. „Current trends in consumption of multimedia content using online streaming platforms: A user-centric survey". *Computer Science Review* 37 : 100268.

[4] Gerber, Hannah R. 2017. „eSports and streaming: Twitch literacies". *Journal of Adolescent & Adult Literacy* 61 (3) : 343–345.

[5] Hoeg, Wolfgang, Thomas Lauterbach. 2009. *Digital Audio Broadcasting: Principles and Applications of DAB, DAB+ and DMB*. John Wiley & Sons.

[6] ITU-T Technical Report PSTR-CROWDS. 2008. *Subjective evaluation of media quality using a crowdsourcing approach*. ITU.

[7] Naderi, Babak, Rafael Zequeira Jiménez, Matthias Hirth, Sebastian Möller, Florian Metzger, Tobias Hoßfeld. 2021. „Towards speech quality assessment using a crowdsourcing approach: Evaluation of standardized methods". *Quality and User Experience* 6 (1) : 1–21.

[8] Polskie Radio. 2022. https://www.polskieradio.pl/ (online). dostęp: 20.04.2022.

ANALIZA PORÓWNAWCZA ZMIERZONYCH I OBLICZONYCH NATĘŻEŃ PÓL ELEKTRYCZNYCH OD NADAJNIKÓW SFN DAB+

COMPARATIVE ANALYSIS OF MEASURED AND CALCULATED ELECTRIC FIELD STRENGTHS FROM SFN DAB + TRANSMITTERS

Ryszard Zieliński
Politechnika Wrocławska, wyb. Wyspiańskiego 27, 50-370 Wrocław, ryszard.zielinski@pwr.edu.pl

DOI: 10.15199/59.2022.4.116

Streszczenie: **W artykule przedstawiono metodykę oraz wyniki pomiarów poziomu sygnału odbieranego w miejscach, w których potencjalnie będzie można zaobserwować zjawisko zaników interferencyjnych powodowanych przez sygnały nadawane przez nadajniki DAB+ sieci jednoczęstotliwościowej SFN (Single Frequency Network). Przedstawiono metodykę pomiarów, zastosowany sprzęt pomiarowy oraz uzyskane wyniki. Porównano je do wyników natężenia pola elektrycznego uzyskanych z predykcji przy użyciu metody propagacyjnej ITU-R 1546, która została w odpowiedni sposób zaimplementowana w Instytucie Łączności PiB we Wrocławiu.**

Abstract: **The article presents the methodology and measurement results of the received signal level in places where it will be possible to observe the phenomenon of interference fade caused by signals transmitted by DAB+ transmitters of the SFN (Single Frequency Network) network. The methodology of measurements, the measuring equipment used and the obtained results were presented. They were compared to the results of the electric field strength obtained from the prediction using the ITU-R 1546 propagation method, which was implemented at the National Institute of Telecommunications in Wroclaw.**

Słowa kluczowe: **DAB+, pomiary natężenia pola elektrycznego, sieć jednoczęstotliwościowa.**

Keywords: **DAB+, electric field measurements, single frequency network (SFN).**

1. WSTĘP

Jednym z silnych ograniczeń hamujących zwiększenie liczby programów radiowych i telewizyjnych i ograniczających rozwój sieci radiowych i telewizyjnych w zakresie częstotliwości VHF i UHF (30 – 300 MHz i 300 – 700 MHz) jest zasób częstotliwościowy. Został on w ostatnich latach silnie ograniczony przez wejście w zakresy częstotliwości powyżej 700 MHz systemów telefonii komórkowej. Problemy z ograniczonym widmem częstotliwości zostały częściowo rozwiązane dzięki zastosowaniu znacznie skuteczniejszych pod względem widmowym interfejsów radiowych stosujących technikę ortogonalnego zwielokrotnienia częstotliwościowego OFDM (*Orthogonal Frequency Division Multiplexing*). Umożliwia ona budowę tzw. multipleksów, w ramach których w jednym kanale częstotliwościowym możemy nadawać wiele programów radiowych (DAB+) lub telewizyjnych (DVB-T). Zastosowany interfejs radiowy ma ponadto jedną dodatkową cechę. Jakość odbieranego sygnału nie ulega pogorszeniu, gdy do odbiornika dociera wiele replik nadanego sygnału pod warunkiem, że czasy ich odbioru mieszczą się w tzw. odstępie ochronnym (prefiksie cyklicznym) ramki OFDM o czasie trwania T_G [1]. W przypadku transmisji DAB+ odstęp ochronny dla jedynego obecnie zalecanego pierwszego trybu transmisji wynosi 246 μs. Z tego względu transmisje z wykorzystaniem interfejsu OFDM są do pewnego stopnia odporne na zjawisko wielodrogowości, występujące w znacznej skali w środowiskach zurbanizowanych. Repliki nadanego sygnału, powstałe na skutek zjawiska wielodrogowości poprawiają odbiór, o ile obserwowane opóźnienia mieszczą się w czasie trwania odstępu ochronnego. Tę odporność kanału transmisyjnego OFDM na wielodrogowość wykorzystano również w inny sposób, budując sieci jednoczęstotliwościowe. Dla odstępu ochronnego o długości 246 μs wystarczy zapewnić maksymalną odległość pomiędzy jednoczęstotliwosciowymi nadajnikami nie większą od 73,83 km, by sygnały nadawane przez te nadajniki mieściły się w czasie trwania odstępu ochronnego.

Kolejną istotną cechą związaną ze stosowaniem sieci jednoczęstotliwościowej jest redukcja obszarów znajdujących się w tzw. cieniu radiowym. Ma to szczególne znaczenie w środowiskach silnie zurbanizowanych. Ten sam efekt można uzyskać stosując sieć wieloczęstotliwościową MFN (*Multi Frequency Network*), ale wtedy trzeba zająć więcej zasobów widmowych. W przypadku sieci jednoczęstotliwościowej nie są zajmowane dodatkowe zasoby częstotliwościowe. Wykorzystując te same zasoby widmowe wybrany obszar można obsłużyć przy użyciu jednego nadajnika lub sieci N nadajników SFN. Rozwiązanie bazujące na sieci nadajników SFN jest ze względu na zapotrzebowanie energetyczne bardziej ekologiczne. Wymaga ono mniejszej całkowitej mocy promieniowanej. Z tego względu generuje również mniejsze zakłócenia do innych sieci. Wielkość zysku związanego z zagęszczeniem sieci można oszacować, zakładając że moc nadajnika P wymagana do zapewnienia prawidłowego odbioru sygnału na granicy obszaru o promieniu R jest proporcjonalna do R^{α}. W [2] pokazano, że przy założeniu takiego modelu propagacyjnego stosunek wymaganej mocy dla pojedynczego nadajnika P_1 w stosunku do mocy P_{SFN} wszystkich N nadajników sieci SFN można wyznaczyć z zależności (1):

$$\frac{P_1}{P_{SFN}} = N^{\left(\frac{\alpha}{2}-1\right)} \tag{1}$$

Dla parametru α=2 odpowiadającego stratom wolnej przestrzeni nie uzyskuje się żadnego dodatkowego zysku. Ale w środowisku zurbanizowanym parametr α jest większy od 2 i wtedy zysk mocy występuje. Przykładowo dla modelu propagacji nad doskonale przewodzącą płaską ziemią parametr α=4. Wtedy, zastępując jeden nadajnik trzema nadajnikami SFN, otrzymamy trzykrotny zysk mocy.

Kolejną zaletą sieci SFN jest sumowanie mocy sygnałów pochodzących od nadajników SFN. Efekt ten powinien być dobrze obserwowany wtedy, gdy sygnały nadawane są w polaryzacji pionowej a antena odbiorcza ma dookólna charakterystykę promieniowania. Wtedy amplitudę wypadkowego natężenia pola elektrycznego E_W działające na antenę można wyznaczyć z zależności (2):

$$E_W = \left|\sum_i \overline{E_i}\right| = \left|\sum_i E_i \cdot e^{j\phi_i}\right| \qquad (2)$$

przy czym: E_i jest amplitudą natężenia pola elektrycznego od i-tego nadajnika SFN, a fazę wektora natężenia pola można wyznaczyć z zależności (3):

$$\phi_i = \left(\frac{2\pi}{\lambda}\right) \cdot d_i \qquad (3)$$

w której λ jest długością fali e-m a d_i jest odległością od i-tego nadajnika SFN.

Należy się spodziewać, że zjawisko wzrostu poziomu odbieranego sygnału będzie dobrze widoczne w miejscach, w których poziomy natężenia pola od poszczególnych nadajników SFN będą równe lub nieznacznie różniące się. Szczegółową analizę tego zjawiska przedstawiono w [3]. W tym przypadku Autor definiuje zysk sieci SFN w postaci (4):

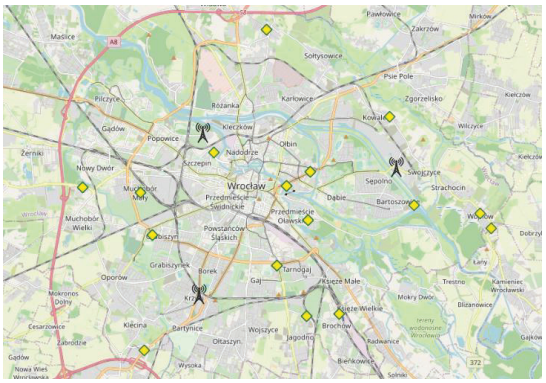
$$SFN_{zysk}=E_{SFN}-E_{MFN} \qquad (4)$$

przy czym E_{SFN} to wypadkowe natężenie pola pochodzące od sieci nadajników SFN, które może być wyrażone zależnością (2), natomiast E_{MFN} to natężenie pola pochodzące od dominującego nadajnika sieci MFN (5):

$$E_{MFN} = \max\{E_1, E_2, ...E_N\} = \max_i\{E_i\} \qquad (5)$$

Natężenia pola wyznaczone zostały przy użyciu metody ITU-R P.1546 [4] zaimplementowanej w programie PIAST opracowanym przez Instytut Łączności PiB we Wrocławiu. Analizowano zysk SFN w całym obszarze działania sieci SFN (przypadek 1) jak również dla obszaru nie obsługiwanego przez sieć MFN (przypadek 2). Największe zyski w wysokości 4 dB uzyskano dla odbioru stacjonarnego (przypadek 1) i 5,3 dB (przypadek 2). Dla odbioru przenośnego i mobilnego zyski były mniejsze i wynosiły 2,0 i 1,5 dB odpowiednio (przypadek 1) oraz 3,7 i 3,0 dB odpowiednio (przypadek 2). Jak zatem widać, zastosowanie sieci SFN zamiast sieci MFN oprócz korzyści widmowych przynosi bardzo realne korzyści związane ze zwiększeniem obszaru obsługi oraz podniesieniem poziomu odbieranego sygnału. Rodzi się zatem pytanie, czy zastosowanie sieci SFN zamiast MFN jest łatwe i czy nie ma innych zjawisk towarzyszących odbiorowi wielu sygnałów o tej samej częstotliwości. Okazuje się, że wstępne analizy tego typu odbioru wykazują występowanie zjawiska zaników selektywnych. Analiza miejsc występowania największych zaników przedstawiona została w [5]. Analizę wielkości tych zaników przedstawiono w [6] i [7], a ich selektywną naturę w [8] i [9]. W artykułach tych jako przykład do analizy wykorzystano sieć SFN składającą się z trzech nadajników, która w ramach projektu LokalDAB została zbudowana we Wrocławiu. Budowa realizowana była w latach 2015 – 2017 przez Politechnikę Wrocławską, Instytut Łączności – PIB oraz Polskie Radio Wrocław. Głównym celem projektu było opracowanie demonstratora sieci jednoczęstotliwościowej SFN opartej na zespole trzech platform nadawczych pracujących w standardzie DAB+, na bazie uniwersalnych układów programowalnych oraz oprogramowania „open source".

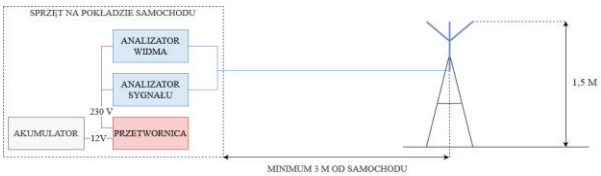
Do wyznaczenia obszarów, w których zaniki mogą być największe zastosowano metodę propagacyjną ITU-R P.1546. Od dokładności tej metody zależy precyzja ustalania miejsc występowania zaników, jak również wyniki symulacji zaników, uwzględniające predykcje poziomów sygnału w tych miejscach. Z tego względu w dalszej części artykułu przedstawiona będzie metodologia weryfikacji natężenia pola elektrycznego wyznaczonego metodą ITU-R P.1546 z wynikami pomiarów. Miejsca, w których wykonano pomiary opisano w [9] i przedstawiono na rys. 1.



Rys. 1. Rozmieszczenie punktów pomiarowych (romby) i stacji nadawczych (symbole nadajnika) na obszarze Wrocławia

2. STANOWISKO POMIAROWE

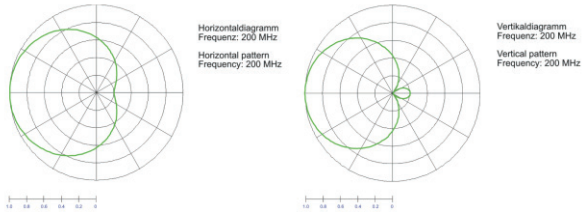
Do pomiarów użyto analizatora widma Tektronix H600 RF Hawk, analizatora sygnałów Promax Ranger HD+ oraz anteny kierunkowej firmy Rohde&Schwarz HE300 na zakres częstotliwości od 200 do 500 MHz. Schemat stanowiska pomiarowego pokazano na rys. 2.



Rys. 2. Schemat stanowiska pomiarowego

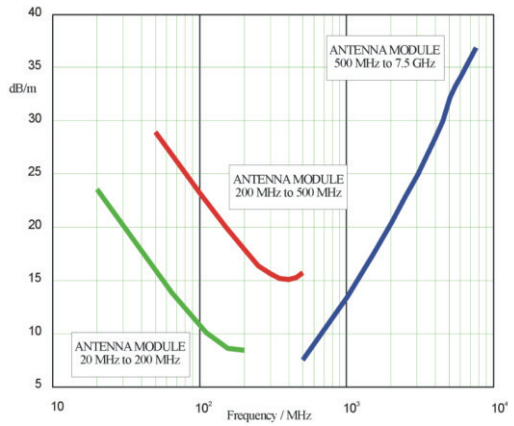
Nadajniki sieci LokalDAB emitują falę elektromagnetyczną o polaryzacji pionowej na częstotliwości 216,928 MHz. Do tej polaryzacji zostało dostosowane

ustawienie anteny pomiarowej. Antena ta ma kierunkową charakterystykę promieniowania przedstawioną na rys. 3.



Rys. 3. Charakterystyka promieniowania kierunkowej anteny pomiarowej R&S H300 [10]

Jak widać, trzydecybelowa szerokość wiązki głównej anteny wynosi ok. 120^0. Z tego względu w niektórych lokalizacjach nie wykonano pomiaru natężenie pola e-m od wszystkich nadajników, bo na danym kierunku w obszarze głównej wiązki anteny występował więcej niż jeden sygnał. Niestety Autorzy nie posiadali anteny pomiarowej o węższej wiązce głównej. Również wyłączanie niektórych nadajników na czas pomiaru nie było możliwe ze względu na emisję programów. Na rys. 4 przedstawiono przebieg współczynnika antenowego AF (*Antenna Factor*) anteny pomiarowej w funkcji częstotliwości.



Rys. 4. Współczynnik antenowy anteny pomiarowej HE 300 w trybie aktywnym [10]

Pomiary wykonano w aktywnym trybie pracy anteny (rys. 5). Na częstotliwości 216,928 MHz współczynnik AF wynosi ok. 17,5 dB.

Rys. 5. Antena odbiorcza HE 300

3. METODOLOGIA POMIARÓW

Mierzono moc sygnału na wyjściu anteny pomiarowej. Moc ta przeliczana była na natężenie pola elektrycznego wyrażone w dB[μV/m] przy założeniu impedancji systemu pomiarowego 50 Ω i AF=17,5 dB. Pomiary wykonano w 16 punktach pomiarowych, z których 7 znajduje się w obszarze zwiększonych zaników (wiersze zacienione tabeli 1). Przedstawiono je w [9] w tabelach 2 i 3. Pozostałych 9 dodatkowych punktów pomiarowych dobrano tak, by były rozmieszczone możliwie równomiernie na obszarze całego Wrocławia. Pomiary w lokalizacjach dodatkowych miały na celu pomiar natężenia pola elektrycznego od nadajników i porównanie z wynikami symulacji. W każdym z punktów pomiarowych antenę kierowano w kierunku jednego, dwóch lub trzech nadajników biorąc pod uwagę kierunki azymutów i szerokość głównej wiązki promieniowania anteny. Łącznie wykonano 27 pomiarów, których wyniki zamieszczono w Tabeli 1. W każdej lokalizacji wykonano 5 pomiarów, przemieszczając punkt pomiarowy w obszarze kwadratu o boku 5 m z pierwszym punktem w środku tego kwadratu. Mierzono moc w paśmie 1536 kHz przy ustawieniu analizatora dla SPAN=2 MHz i RBW=1kHz. Wynikiem pomiaru była uśredniona wartość z 10 pomiarów w każdym z 5 punktów pomiarowych. Uzyskany wynik można traktować jako wartość średnią natężenia pola elektrycznego E_m. Przyjmując zgodnie z Zaleceniem ITU-R P.1546 odchylenie standardowe σ=5,5 dB i rozkład logarytmicznie normalny, wyznaczono 17 wyników pomiarów, które mieszczą się w zakresie $E_s \pm \sigma$, co stanowi 63% wyników. W zakresie $E_s \pm 3\sigma$ mieszczą się 24 wyniki, co stanowi 89%.

4. WNIOSKI

Celem badań było określenie stopnia przydatności metody propagacyjnej ITU-R P.1546 do określania poziomu natężenia pola w środowisku miejskim na potrzeby wyznaczania obszarów, w których mogą występować zaniki selektywne o intensywnym charakterze. Warunkiem zaistnienia tego typu zaników jest występowanie na tym obszarze natężeń pól elektrycznych od wspólnokanałowych nadajników o zbliżonych wartościach. Tylko w takim przypadku zjawisko zaniku będzie bardzo wyraźne i łatwe do zaobserwowania. Dla dwóch sygnałów o tej samej wartości natężenia pola i fazach przesuniętych o 180^0 zanik ma nieskończoną głębokość. Jeżeli pomiędzy sygnałami jest różnica w natężeniu równa 1 dB, to zanik ma głębokość ok. 19 dB. Jeżeli różnice pomiędzy natężeniami pól e-m wynoszą 2 dB lub 3 dB, to obserwowane zaniki wyniosą odpowiednio 14 dB i 11 dB. Im większa różnica poziomów pomiędzy interferującymi sygnałami tym płytsze zaniki będą obserwowane. Z tego punktu widzenia jedynie w 5 na 27 analizowanych przypadków różnica pomiędzy wartością zmierzoną i wyznaczoną na podstawie symulacji była poniżej 1 dB, stwarzając warunki odpowiednie do analizy zaników metodą symulacyjną.

Biorąc to pod uwagę do wyznaczania obszarów występowania istotnych zaników powinny być stosowane bardziej precyzyjne metody predykcji natężenia pole e-m.

Tab 1. Zestawienie wyników pomiarów; Np. – nr punktu pomiarowego, E_m – wartość zmierzona, E_s – wartość obliczona, Δ - różnica E_s-E_m [11]

Np.	Szer. / Dł. geogr. [WGS-84]	Nadajnik mierzony	E_m [dBµV/m]	E_s [dBµV/m]	Δ [dB]
1	Hallera	PRW	57,7	61,0	3,3
1	Hallera	PWr	60,7	57,5	-3,2
2	Konduktorska	PRW	56,1	55,6	-0,5
2	Konduktorska	IL – PIB	53,8	54,8	1,0
3	Okólna	IL – PIB	40,9	61,0	20,1
3	Okólna	PWr	41,2	59,8	18,6
4	Pl. Grunwaldzki	PWr	51,2	76,0	24,8
4	Pl. Grunwaldzki	IL – PIB	59,7	60,2	0,5
6	Armii Krajowej	PWr	58,2	54,4	-3,8
6	Armii Krajowej	IL – PIB	55,9	54,6	-1,3
6	Armii Krajowej	PRW	57,6	55,5	-2,1
8	Polanowicka	IL – PIB	55,2	48,8	-6,4
8	Polanowicka	PWr	59,2	59,4	0,2
9	Szwajcarska	PWr	61,8	65,4	3,6
9	Szwajcarska	PRW	58,5	53,6	-4,9
10	51,10837/ 16,944368	PWr	48,6	59,1	10,5
10	51,10837/ 16,944368	PRW	48,0	50,6	2,6
11	51,108945/ 17,054419	IL – PIB	57,1	57,3	0,2
11	51,108945/ 17,054419	PWr	58,8	61,1	2,3
12	51,065707/ 17,082681	IL – PIB	49,0	57,4	8,4
12	51,065707/ 17,082681	PRW	54,0	51,00	-3,0
13	51,094780/ 17,164684	IL – PIB	48,1	63,9	15,8
14	51,099725/ 17,158666	IL – PIB	55,8	66,6	10,8
15	51,132399/ 17,109921	IL – PIB	75,7	70,3	-5,4
16	51,120384 /17,015212	PWr	83,1	99,4	16,3
17	51,053519/ 16,97732	PRW	56,1	54,6	-1,5
18	51,102343/ 17,122982	IL – PIB	86,3	79,3	-7,0

Charakterystyki zaników w funkcji częstotliwości dla obliczonych wartości natężenia pola elektrycznego metodą ITU-R P.1546 zostały pokazane w [9]. W dalszych badanach zostanie przeprowadzona analiza zaników obliczonych na podstawie zmierzonych wartości natężenia pola elektrycznego i porównanie ich z przebiegami zaników zmierzonymi przy użyciu analizatora widma.

LITERATURA

[1] Zieliński Ryszard, 2017, „Warunki uzyskania poprawnego odbioru sygnału DAB+ w sieci jednoczęstotliwościowej", *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, nr. 6, str. 509-512.

[2] Mattson Anders, 2005, "Single Frequency Networks in DTV", *IEEE Transactions On Broadcasting*, Dec. vol. 51, no. 4, pp. 413-422.

[3] Staniec Kamil, 2021, „Analysis of the Single Frequency Network Gain in Digital Audio Broadcasting Networks", *Sensors*, 21, 569. https://doi.org/10.3390/s21020569.

[4] International Telecommunication Union, 2013, Recommendation ITU-R P.1546-5 : Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 3 000 MHz.

[5] Zieliński Ryszard, 2018, „Rozmieszczenie obszaru zaników w jednoczęstotliwościowej sieci SFN na przykładzie sieci DAB+", *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, nr. 8-9, str. 581-584.

[6] Zieliński Ryszard, 2018, „Analiza wielkości zaników w jednoczęstotliwościowej sieci DAB+ na przykładzie sieci we Wrocławiu", *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, nr. 6, str. 320-325.

[7] Zieliński Ryszard, 2020, „Analysis and Comparision of the Fade Phenomenon in the SFN DAB+ Network With Two and Three Transmitters", *Intl Journal of Electronics and Telecommunications*, Vol. 66, No. 1, str. 85-92.

[8] Zieliński Ryszard, 2019, "Fade analysis in DAB+ SFN network in Wroclaw", *Proc. of the 2019 International Symposium on Electromagnetic Compatibility (EMC Europe 2019)*, Barcelona, Spain, September 2– 6, 2019, (978-1-7281-0594-9/19/$31.00 © 2019 IEEE).

[9] Michalski Igor, Zieliński Ryszard, 2021, „Verification Methodology of Fade Characteristics in a DAB+ SFN in Wroclaw", *Intl Journal of Electronics and Telecommunications*, Vol. 67, No. 3, str. 537-542.

[10] R&S HE300 Active Directional Antenna R&S HE300UK Upgrade Kit Manual, 2010.

[11] Michalski Igor, 2020, Praca dyplomowa magisterska, "Analiza rozkładu natężenia pola e-m i zaników od sieci SFN DAB+ we Wrocławiu", Politechnika Wrocławska.

MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

WIELOKANAŁOWY REJESTRATOR EMISJI RADIOFONICZNYCH UKF-FM WYKONANY W TECHNICE SDR

AN SDR-BASED VHF-FM RADIO BROADCAST MULTICHANNEL ACQUISITION SYSTEM

mgr inż. Radosław Helon[1]; dr inż. Przemysław Korpas[2]

[1,2]Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa
[1]radoslaw.helon.stud@pw.edu.pl, [2]przemyslaw.korpas@pw.edu.pl

DOI: 10.15199/59.2022.4.117

Streszczenie: **Niniejszy artykuł zawiera opis implementacji zdefiniowanego programowo rejestratora sygnałów z pasma UKF-FM pozwalającego na równoczesną demodulację oraz zapis do 32 niezależnych emisji radiofonicznych. Urządzenie zrealizowano na bazie zestawu deweloperskiego ZYBO-Z7 z układem Zynq-7010 z dodanym autorskim front-endem w.cz. Przetwarzanie sygnału w domenie cyfrowej wykorzystuje zaawansowane techniki zrównoleglania. W artykule przedstawiono najciekawsze z zastosowanych rozwiązań.**

Abstract: **This paper describes the implementation of a software-defined VHF-FM broadcast acquisition system that allows for the simultaneous processing and recording of up to 32 independent radio emissions. The device is based on the ZYBO-Z7 development board containing the Zynq-7010 chip equipped with a custom RF front-end module. Signal processing in the digital domain has been implemented with advanced parallelization techniques. This paper presents the most interesting of the applied solutions.**

Słowa kluczowe: **DSP, FPGA, SDR, Zynq, rejestrator.**

Keywords: **DSP, FPGA, SDR, Zynq, recorder**

1. WSTĘP

Obserwowany w ostatnich latach rozwój techniki radia programowalnego (ang. Software Defined Radio) pozwala realizować szerokopasmowe urządzenia i systemy rejestracji sygnałów radiowych. Potrzeby w tym obszarze obejmują zagadnienia takie jak bezpieczeństwo, wymogi ustawowe, czy badania i weryfikacja urządzeń radiokomunikacyjnych.

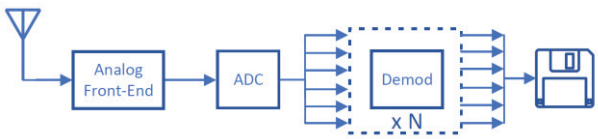
Sposoby przechowywania informacji o zarejestrowanych sygnałach radiowych można podzielić na dwie podstawowe grupy: (1) zapis wybranego wycinka widma elektromagnetycznego w postaci strumienia próbek IQ, bez uprzedniej demodulacji treści przekazów [1][2][3], oraz (2) zapis zdemodulowanych sygnałów informacyjnych [4]. Te dwa podejścia różnią się zasadniczo w wielu aspektach, gdzie do podstawowych należy zaliczyć ilość przestrzeni na nośniku danych niezbędną do przechowania wyników rejestracji o tym samym czasie trwania oraz możliwości szczegółowej analizy zarejestrowanych już danych. Podobnie dostęp do zarejestrowanych treści wymaga różnego nakładu mocy obliczeniowej w tych dwóch przypadkach – z korzyścią dla zapisu sygnałów po demodulacji.

W niniejszej pracy zaprezentowano wyniki implementacji wielokanałowego rejestratora sygnałów radiofonicznych pracujących w paśmie rozsiewczym UKF-FM. Rejestrator został zbudowany w technice SDR i wykorzystuje zapis zdemodulowanych sygnałów audio, dzięki czemu wykonane nagrania można łatwo odsłuchać. Za cel postawiono sobie optymalizację projektu pod kątem kosztów, co miało wyraz w doborze prostego układu front-endu analogowego oraz możliwie efektywnego wykorzystania zasobów układu FPGA.

Projekt stanowił przedmiot pracy magisterskiej i jako taki był motywacją do zgłębiania technik równoległego przetwarzania sygnałów oraz wydajnej implementacji algorytmów DSP w strukturach FPGA. Pod względem technicznym rejestrator demonstruje potencjalne możliwości wykorzystania techniki SDR do realizacji wielokanałowych rejestratorów sygnałów radiowych. Po wprowadzeniu odpowiednich zmian (np. po dokonaniu implementacji odpowiednich demodulatorów w strukturze układu FPGA i modyfikacji układu front-end) możliwa będzie rejestracja sygnałów w innych standardach analogowych (np. w sieciach radiotelefonicznych pracujących w pasmach VHF lub UHF) lub w systemach cyfrowych. Zrealizowany projekt może z powodzeniem znaleźć zastosowanie praktyczne w roli systemu realizującego ustawowy obowiązek utrwalania treści transmitowanych w eter przez koncesjonowanych nadawców [5].

2. ARCHITEKTURA REJESTRATORA

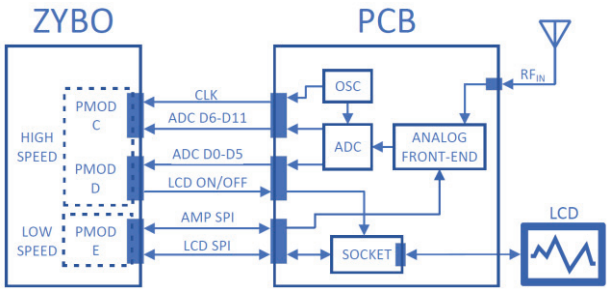
Na Rys. 1 przedstawiono schemat blokowy zdefiniowanego programowo rejestratora realizującego proces równoległego wielokanałowego przetwarzania sygnału przed zapisem danych.



Rys. 1. Schemat koncepcji rejestratora realizującego wielokanałowe przetwarzanie w domenie cyfrowej

W przedstawionej koncepcji sygnał radiowy wprowadzany jest przez antenę na analogowy front-end, gdzie poddawany jest odpowiedniej filtracji i wzmocnieniu, aby mógł być efektywnie spróbkowany przez ADC. Sygnał w domenie cyfrowej wprowadzany jest na wejście układu FPGA, gdzie następuje jego rozdział na kanały i zrównoleglony proces filtracji i demodulacji każdego z nich. Ostatecznie przetworzone dane podlegają podziałowi na pakiety i rejestracji na zewnętrznym nośniku danych.

Opisany projekt zorientowany był pod kątem optymalizacji kosztów. Z tego powodu do realizacji prototypu wykorzystano popularną płytkę deweloperską ZYBO-Z7 zawierającą układ Zynq 7010, do której dołączony został samodzielnie opracowany układ front-endu w.cz. Cechą charakterystyczną układów Zynq jest integracja funkcjonalności FPGA i procesora w ramach jednej struktury krzemowej [6]. W tym przypadku wykorzystywany układ jest jednym z najsłabszych w całej rodzinie Zynq, co wymuszało odpowiednie zaplanowanie cyfrowego toru przetwarzania sygnału. Na Rys. 2 zaprezentowano schemat blokowy zrealizowanego prototypu.

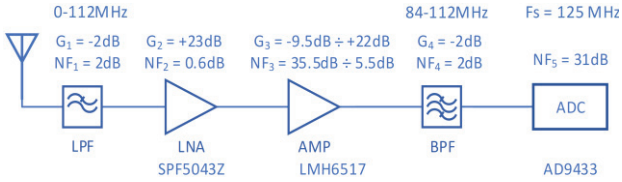


Rys. 2. Schemat blokowy prototypu rejestratora

3. ANALOGOWY FRONT-END

W celu minimalizacji kosztów projektu zdecydowano się na wybór architektury z próbkowaniem bezpośrednim w 2. strefie Nyquista. Taka architektura może być zrealizowana przy pomocy relatywnie prostego front-endu analogowego a zarazem pozwala na zastosowanie tańszego przetwornika ADC o niższej częstotliwości próbkowania, niż miałoby to miejsce w przypadku próbkowania w 1. Strefie Nyquista. Z kolei próbkowanie bezpośrednie eliminuje niedoskonałości znane z popularnej architektury z zerową częstotliwością pośrednią, wynikające ze skutków niedokładności zrównoważenia kanałów kwadraturowych.

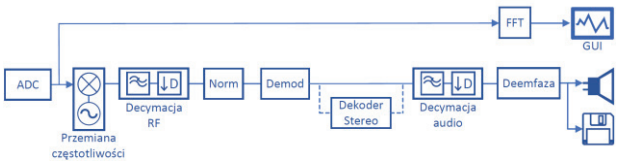
W skład zaprojektowanego front-endu weszły: eliptyczny filtr dolnoprzepustowy, wzmacniacz niskoszumny SPF5043Z, wzmacniacz LMH6517 o regulowanym wzmocnieniu oraz eliptyczny filtr pasmowoprzepustowy. Odfiltrowany i wzmocniony sygnał z pasma 87,5 ÷ 108 MHz był próbkowany przetwornikiem AD9433 ze stałą częstotliwością 125 MHz. Dobrane elementy pozwalały oczekiwać czułości odbiornika na poziomie 1 µV oraz tłumienia niepożądanych częstotliwości na poziomie minimum 70 dB. Wszystkie elementy front-endu analogowego zostały przedstawione na Rys. 3. Jak widać, układ pozbawiony jest mieszacza analogowego dzięki zastosowaniu techniki próbkowania bezpośredniego.



Rys. 3. Schemat blokowy front-endu analogowego

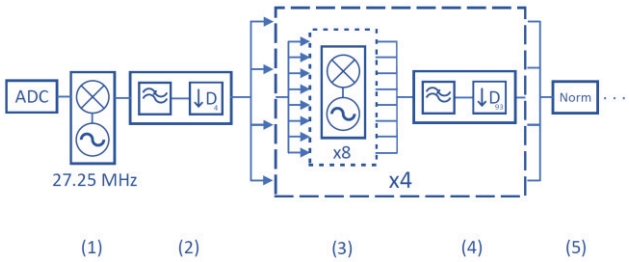
4. CYFROWY TOR PRZETWARZANIA SYGNAŁU

Głównym zadaniem realizowanym w domenie cyfrowej było przetworzenie spróbkowanego sygnału do postaci zdemodulowanego sygnału audio, osobno dla każdego z N kanałów. Zakładano wstępnie uzyskanie rejestracji przynajmniej N=30 kanałów z pasma UKF-FM, co koresponduje z liczbą rozgłośni dostępnych w warszawskim eterze. Przetwarzanie sygnału dla pojedynczego kanału obejmuje następujące operacje: przemianę częstotliwości, decymację, normalizację, demodulację, decymację audio oraz deemfazę. Kompletny schemat toru przetwarzania realizowanego z perspektywy pojedynczego kanału FM został przedstawiony na Rys. 4.



Rys. 4. Schemat funkcjonalny toru przetwarzania dla pojedynczego kanału FM

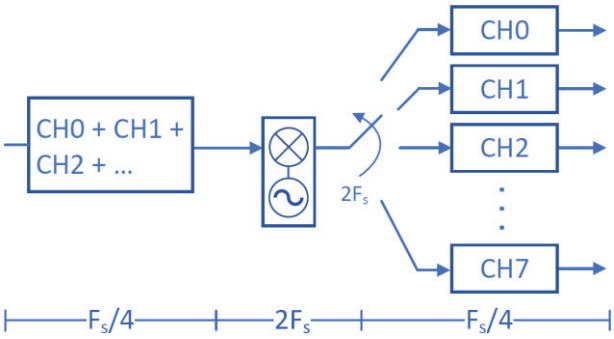
Ze względu na wyraźnie ograniczone zasoby zastosowanego układu FPGA zdecydowano się na operowanie wyłącznie na arytmetyce stałoprzecinkowej oraz ograniczenie liczby wykonywanych operacji mnożenia i dzielenia. Fragment toru przetwarzania realizowany równolegle został zilustrowany na Rys. 5. Wstępny etap decymacji (2) przeprowadzany był przed rozdzieleniem sygnału na poszczególne kanały. W celu uzyskania jak najwyższego stopnia decymacji konieczne było uprzednie sprowadzenie pasma zainteresowania do częstotliwości bazowej (1).



Rys. 5. Schemat funkcjonalny sekcji zrównoleglonego przetwarzania kanałów

Obniżenie częstotliwości próbkowania sygnału będące wynikiem decymacji pozwalało na wykorzystanie tych samych zasobów logiki programowalnej do naprzemiennego przetwarzania sygnału pochodzącego z różnych kanałów. Ponadto część zasobów FPGA mogła być taktowana z częstotliwością wyższą od częstotliwości

próbkowania oryginalnego sygnału, co pozwalało uzyskać dalszą oszczędność w ich wykorzystaniu. Na Rys. 6 zaprezentowano koncepcję współdzielenia pojedynczego modułu mieszacza pomiędzy 8 niezależnymi kanałami.



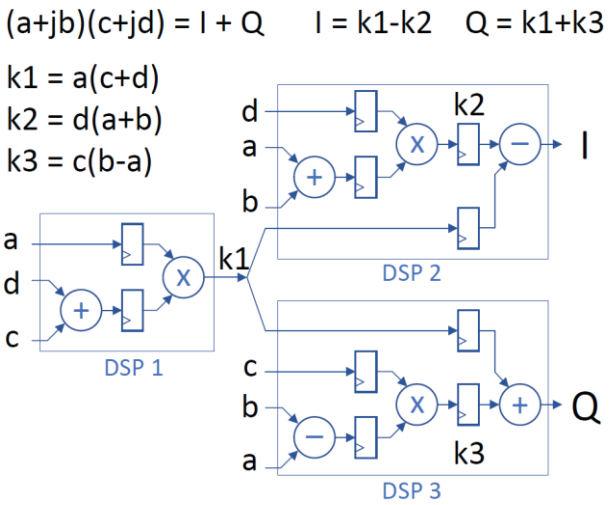
Rys. 6 Schemat koncepcji przetwarzania wielu kanałów przy użyciu tych samych zasobów sprzętowych

Zastosowanie filtrów polifazowych [7] w procesie decymacji pozwalało na efektywne współdzielenie zasobów pomiędzy kilkoma kanałami również na samym etapie decymacji. Ostatecznym zyskiem z tej operacji (4) była możliwość realizacji dalszej części toru przetwarzania bez konieczności zrównoleglenia (5) – wykorzystując te same zasoby sprzętowe pomiędzy wszystkimi kanałami.

Kolejne etapy przetwarzania również zostały poddane stosownej optymalizacji. Stosując algorytm opisany w [8] dokonano implementacji procesu demodulacji wykorzystującego zaledwie dwie operacje mnożenia. W przypadku algorytmu normalizacji sygnału, część pracochłonnych operacji takich jak pierwiastkowanie zostało zastąpione przybliżeniami o akceptowalnej dokładności [9], natomiast operację dzielenia zrealizowano przy użyciu dzielnika o iteracyjnej architekturze charakteryzującej się niskim wykorzystaniem zasobów kosztem wydłużonego czasu przetwarzania danych [10]. Układy FPGA Xilinx wyposażone są w dedykowane bloki DSP przeznaczone do wykonywania operacji arytmetycznych. Ze względu na szeroką funkcjonalność bloków DSP odpowiedni dobór kolejności przeprowadzania operacji w algorytmie może pozwolić na efektywne użycie pojedynczego bloku DSP do wykonania kilku różnych działań arytmetycznych bez wykorzystywania dodatkowych zasobów logiki programowalnej. Ta własność była szczególnie użyteczna w przypadku operacji mnożenia liczb zespolonych zaimplementowanej zgodnie z własnością opisaną w [8]. Na Rys. 7 przedstawiono architekturę mieszacza zrealizowanego wyłącznie przy użyciu 3 bloków DSP.

5. BADANIA

Algorytm przetwarzania sygnałów zaimplementowany w FPGA był w stanie równocześnie demodulować do 32 niezależnych transmisji radiowych. Głównym ograniczeniem w przetwarzaniu większej liczby kanałów okazała się niemożność poprowadzenia połączeń pomiędzy



Rys. 7 Schemat budowy mieszacza zrealizowanego przy pomocy bloków DSP

pozostałymi zasobami wynikająca z wysokiego wykorzystania dostępnych modułów Slice. Wykorzystanie poszczególnych zasobów logiki programowalnej zestawiono w Tab. 1.

Tab. 1. Wykorzystanie zasobów logiki programowalnej układu XC7Z7010

Rodzaj zasobu	*Wykorzystanie*	*Dostępność*
LUT	12342	17600
FF	16149	35200
DSP	58	80
BRAM	44.5	60
Slice	4272	4400

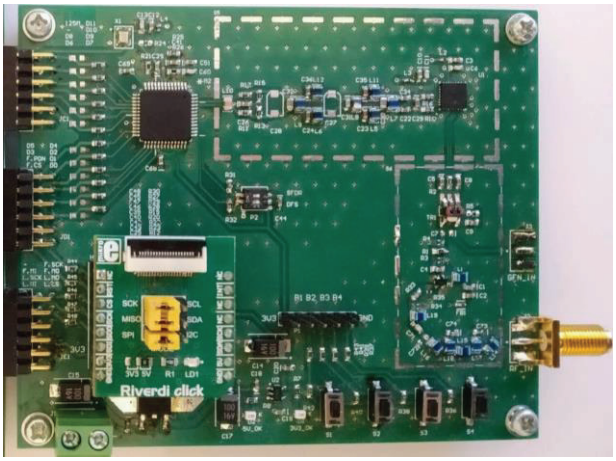
Ostatecznie dane przetworzone do postaci zdemodulowanych kanałów audio były przekazywane w postaci pakietów z logiki programowalnej do systemu procesorowego układu Zynq. Tam mogły być poddane procesowi zapisu na kartę microSD lub zostać przesłane po łączu sieciowym do aplikacji działającej na PC. System procesorowy układu Zynq był odpowiedzialny również za obsługę graficznego interfejsu użytkownika składającego się m.in. z ekranu dotykowego. Interfejs pozwala na analizę częstotliwościową odbieranych sygnałów oraz regulację wybranych parametrów rejestratora. Alternatywnie rejestrator może być sterowany z poziomu aplikacji PC. W celu uatrakcyjnienia projektu rejestrator został wyposażony w dodatkowe funkcjonalności: możliwość odsłuchu wybranego kanału audio w czasie rzeczywistym (dźwięk mono lub stereo), czy opcjonalną kompresję MP3 sygnału przed zapisem.

W Tab. 2. zestawiono uzyskane maksymalne czasy rejestracji dla nośnika pamięci o pojemności 128 GB. Wartości te warto odnieść do czasów zapisu sygnału RF bez demodulacji, tj. w postaci próbek IQ. Przykładowo rejestracja całego pasma UKF-FM o szerokości około 20 MHz wymaga próbkowania kwadraturowego na poziomie 25 MS/s. Dla zapisu 16-bitowego uzyskiwany jest strumień danych o przepływności 100 MB/s, co zapełni rzeczony nośnik jedynie w około 21 minut.

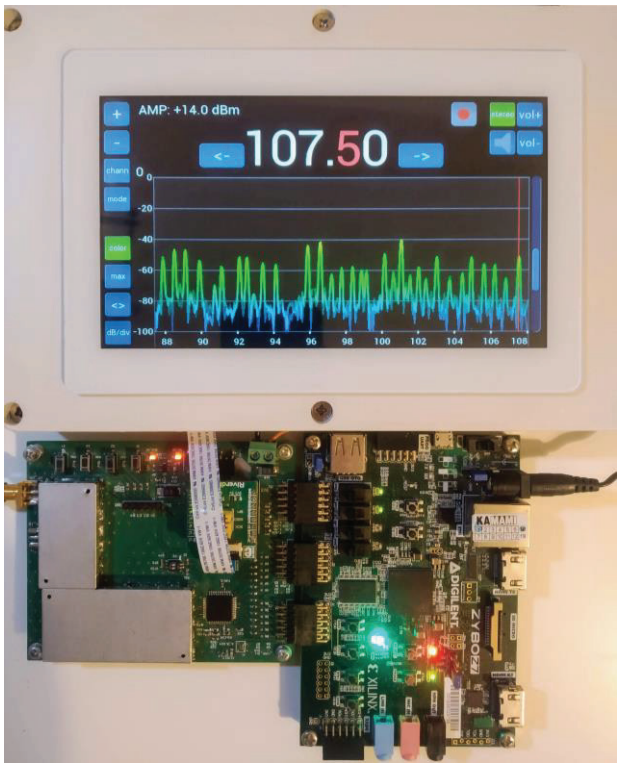
Tab. 2. Maksymalne czasy rejestracji danych dla nośnika pamięci o rozmiarze 128 GB

Metoda rejestracji	*Czas rejestracji*
32 kanały mono	13 godzin
1 kanał mono (bez kompresji MP3)	15 dni 10 godzin
1 kanał mono (z kompresją MP3)	185 dni 4 godziny

Na Rys. 8 umieszczono fotografię wykonanego front-endu z widocznymi sekcjami zasilania, filtrów i wzmacniaczy oraz przetwornikiem analogowo-cyfrowym. Natomiast na Rys. 9 zaprezentowano zdjęcie kompletnego zestawu projektu rejestratora.



Rys. 8. Moduł front-endu radiowego zawierający przetwornik A/C z niezbędnymi filtrami i wzmacniaczami



Rys. 9. Kompletny rejestrator podczas pracy

6. PODSUMOWANIE

W artykule przedstawiono implementację zdefiniowanego programowo wielokanałowego rejestratora FM będącego w stanie przetwarzać równocześnie do 32 niezależnych emisji radiofonicznych. Uzyskany rezultat pozwala na rejestrację emisji wszystkich rozgłośni dostępnych w warszawskim paśmie UKF-FM. Odbiornik umożliwia akwizycję zdemodulowanych sygnałów na zewnętrznym nośniku pamięci lub ich strumieniowanie za pomocą łącza sieciowego. Sterowanie odbiornikiem może odbywać się przy użyciu graficznego interfejsu użytkownika lub za pomocą programu komunikującego się po łączu sieciowym.

Realizacja cyfrowego przetwarzania sygnałów przez układ z logiką programowalną zapewnia możliwość dalszego rozwoju projektu poprzez skalowanie rozwiązania lub dokonanie implementacji dodatkowych funkcjonalności. Projekt może z powodzeniem być zaadaptowany o kolejne funkcjonalności pod warunkiem modyfikacji układu front-end.

LITERATURA

[1] Rosolowski D., Gryglewski D., Korpas P., Wojtasiak W., Modelski J. 2016. “An ultrawideband 1 to 6 GHz 0-IF radio receiver with 500 MHz of instantaneous bandwidth”. *21st International Conference on Microwave, Radar and Wireless Communications (MIKON)*, s. 1–4.

[2] Grayver E., Davidson B., Lee C. 2018. “Ultrawideband (5 GHz) mixed-signal front end and recorder for software defined radio”. *IEEE Aerospace Conference*, s. 1-8.

[3] Lehmann A. E., Kirichenko D. E., Kirichenko A. F., Sarwana S., Gupta D. 2013. “Superconducting digital-RF receiver with high-speed data recording capability". *IEEE 14th International Superconductive Electronics Conference (ISEC)*, s. 1-3.

[4] Bano R. S., Ramzan R. 2013. “COTS based multichannel FM receiver and recorder using SDR concept”. *IEEE 20th International Conference on Electronics, Circuits, and Systems (ICECS)*, s. 213-216.

[5] Art. 20 ustawy z dnia 29 grudnia 1992 r. o radiofonii i telewizji, (Dz.U.2020.805).

[6] Xilinx. 2018. “DS190 Zynq-7000 SoC Data Sheet: Overview”, s. 2-3.

[7] Harris F., Dick Ch. 2003. „Digital Receivers and Transmitters Using Polyphase Filter Banks for Wireless Communications”, *IEEE Transactions on Microwave Theory and Techniques*, t. 51, nr. 4, s. 1395–1412.

[8] Lyons R. 2011. „Understanding Digital Signal Processing”. *Upper Saddle River*, NJ: Prentice Hall.

[9] Frerking M. 1994. „Digital Signal Processing in Communication Systems”.

[10] Xilinx. 2021. “PG151 Divider Generator v5.1”, s. 7-11.

ZASTOSOWANIE AUTOMATYCZNEJ OPTYMALIZACJI ORAZ UPROSZCZONEGO MODELU CIAŁA DO PROJEKTOWANIA WĄSKOPASMOWEJ STRUKTURY EKRANUJĄCEJ POLE EM

APPLICATION OF AUTOMATIC OPTIMIZATION AND A SIMPLIFIED BODY MODEL FOR DESIGNING A NARROWBAND SCREENING STRUCTURE FOR EM FIELD

Jarosław Kawecki[1]; Łukasz Januszkiewicz[2]; Paolo Di Barba[3]

[1] Politechnika Łódzka, Instytut Elektroniki, Łódź, jaroslaw.kawecki@p.lodz.pl
[2] Politechnika Łódzka, Instytut Elektroniki, Łódź, lukasz.januszkiewicz@p.lodz.pl
[3] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pawia, Włochy, paolo.dibarba@unipv.it

DOI: 10.15199/59.2022.4.118

Streszczenie: **W artykule przedstawiono zastosowanie algorytmu automatycznej optymalizacji oraz uproszczonego modelu ciała człowieka do projektowania wąskopasmowej struktury ekranującej pole elektromagnetyczne pochodzące od nadajnika systemu piątej generacji, pracującego w paśmie 3,5 GHz. Opracowana struktura ogranicza gęstość mocy w obszarze gałek ocznych i wykonana jest z elementów prętowych. Została zaprojektowana z wykorzystaniem symulacji komputerowych metodą różnic skończonych w dziedzinie czasu oraz algorytmu optymalizacyjnego wykorzystującego strategię ewolucyjną. Aby skrócić czas wymagany do przeprowadzenia symulacji, zastosowano zredukowany model twarzy i gałek ocznych. Zaprojektowana struktura około siedmiokrotnie zmniejsza gęstość mocy w obszarze gałki ocznej oraz może zostać wykonana w postaci przezroczystej osłony oczu.**

Abstract: **The article presents the use of an automatic optimization algorithm and a simplified model of the human body to design a narrow-band structure that is shielding the electromagnetic field coming from the fifth-generation transmitter operating in the 3.5 GHz band. The structure is designed to limit the power density in the eyeball area and is made of rod elements. It was designed using computer simulations utilizing the finite difference time domain method and an optimization algorithm that is based on an evolutionary strategy. In order to shorten the time required for the simulation, a reduced model of the face and eyes was used. The designed structure reduces the power density approximately sevenfold in the eyeball area and can be made in the form of an eye shield.**

Słowa kluczowe: **ekranowanie, FDTD, modelowanie ciała, optymalizacja komputerowa, systemy 5G.**

Keywords: **5G systems, computer optimization, FDTD, human body models, shielding.**

1. WSTĘP

W ostatnim czasie można zaobserwować rosnące zainteresowanie zagadnieniami związanymi z wpływem fal elektromagnetycznych (EM) na organizm człowieka. Temat ten jest szczególnie ważny dla osób, które odczuwają lęk przed skutkami promieniowania emitowanego przez systemy komunikacji bezprzewodowej, co dotyczy w szczególności systemów piątej generacji (5G) [1]. Jest to spowodowane głównie łatwą do zaobserwowania zmianą w infrastrukturze telekomunikacyjnej, która wymaga umieszczenia nowego wyposażenia stacji bazowych dla nowej generacji systemu bezprzewodowego. Sytuacje te mogą w pewnej grupie osób powodować rozwój irracjonalnych zachowań, które są przedmiotem badań psychologów. Wiele z tych osób uważa się za szczególnie wrażliwych na promieniowanie i szuka sposobu na zminimalizowanie ich ekspozycji na pola EM.

W literaturze opisane są struktury ekranujące ciało człowieka w kontekście tzw. efektu termicznego oddziaływania fal na ciało. Ten efekt może powodować zwiększenie temperatury tkanek pod wpływem energii pochłoniętej przez nie, a struktury ekranujące mają za zadanie ograniczenie gęstości energii fali EM w obszarze ciała. Z opisanych w literaturze badań oddziaływania fal EM na organizm człowieka wynika, że zarówno konstrukcja, jak i materiały użyte do wykonania osłony ludzkiego ciała mają wpływ na jej działanie.

Typowym rozwiązaniem w przypadku projektowania osłon ciała jest zastosowanie specjalnych materiałów tekstylnych o kontrolowanej przewodności, które przylegają do ciała. Są to materiały zawierające nano-dodatki [2] lub w prostszych wariantach, kompozyty zawierające płytki metalowe [3] lub materiały wielowarstwowe o kontrolowanej przewodności [4]. Tego rodzaju struktury ekranujące nie mogą zostać zastosowane do ekranowania oczu ze względu na ich nieprzezroczystość.

W literaturze opisane są różne struktury periodyczne do ochrony selektywnej częstotliwościowo. Układy te nie wymagają pokrycia całej powierzchni materiałem przewodzącym a zatem mogą zostać zaadaptowane do wykonania struktur ekranujących oczy. Stosowany jest w ich przypadku dielektryczny materiał bazowy na którym osadzone są liczne struktury elementarne wykonane z materiałów przewodzących [5,6]. Można w tym celu zastosować technologie opracowane do realizacji struktur przewodzących o różnych kształtach z materiałów włókienniczych [7]. Stosowane mogą być również technologie druku atramentem przewodzącym lub haftu przewodzącymi przędzami stosowane na anteny tekstylne [8]. Szczególnie ciekawe wydaje się być zastosowanie technologii fizycznego naparowywania metali na podłoża dielektryczne, które pozwala na wytwarzanie bardzo cienkich (nawet częściowo przezroczystych)

struktur przewodzących o skomplikowanych kształtach, a więc może być również stosowane do wykonania osłon oczu.

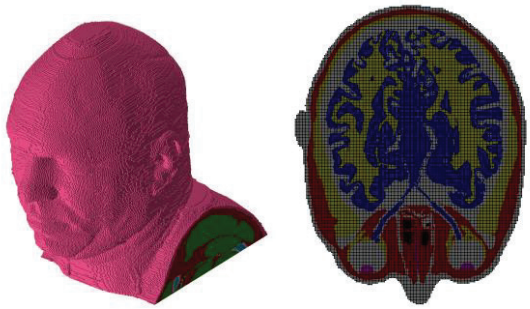
W artykule opisano wstępny projekt wąskopasmowej struktury ekranującej oczy, wykorzystującej elementy prętowe. Na obecnym etapie prac zbadano możliwość jej opracowania dla częstotliwości 3,5 GHz. Dzięki prostocie konstrukcji możliwej jest jej wykonanie w postaci częściowo przezroczystej dzięki czemu może w minimalnym stopniu zaburzać pole widzenia.

2. UPROSZCZONY MODEL GŁOWY CZŁOWIEKA

Opisane w artykule badania przeprowadzono wykorzystując symulacje numeryczne w metodzie różnic skończonych w dziedzinie czasu (metoda FDTD). Zastosowany został w tym celu program XFdtd firmy Remcom, w którym dostępny jest heterogeniczny model głowy człowieka, złożony z 24 tkanek i podzielony na komórki elementarne (woksele) o rozmiarze 1 mm. Widok modelu i jego przekrój płaszczyzną poziomą przedstawiono na rys. 1. Symulacje komputerowe ekspozycji ciała na fale EM o częstotliwości 3,5 GHz z wykorzystaniem tego modelu wymagają użycia 1,8 GB pamięci komputera. Są ponadto dość czasochłonne i trwają około 6 minut na komputerze wyposażonym w kartę graficzną posiadającą jednostkę obliczeniową GPU o pamięci 16GB. Tak długi czas symulacji może być ograniczeniem w przypadku procesu projektowania, wymagającego wielu iteracji.

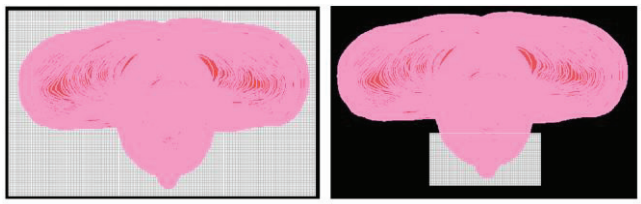
W badaniach dotyczących projektowania osłon oczu, przyjęto, że największa gęstość mocy w obszarze gałek ocznych występuje w przypadku ekspozycji ciała na falę pochodzącą od źródła umieszczonego po stronie twarzy. Aby badać to zjawisko konieczne jest zatem opracowanie modelu odwzorowującego przede wszystkim gałki oczne. Ze względu na to, że model heterogeniczny całej głowy zawiera bardzo wiele wokseli odwzorowujących również tylną część głowy, zbadano w jaki sposób ograniczenie objętości tego modelu jedynie do obszaru twarzy wpłynie na wyniki symulacji. Na rys. 2. przedstawiono model heterogeniczny całej głowy oraz model heterogeniczny ograniczony jedynie do twarzy. Model całej głowy zawierał 52 mln wokseli i obejmował obszar o wymiarach 510 x 340 x 300 mm. Model twarzy ograniczono do obszaru o wymiarach 185 x 240 x 87 dzięki czemu zawierał jedynie 3,9 mln wokseli. Pozwoliło to o 40% zredukować czas wykonywania symulacji. Niestety wyniki symulacji gęstości mocy w obszarze gałek ocznych uzyskane za pomocą modelu całej głowy różniły się nawet 30 krotnie od wyników uzyskanych za pomocą modelu twarzy.

Kolejnym etapem badań było opracowanie modelu uproszczonego twarzy i oczu, który zbudowany byłby z mniejszej liczby wokseli w porównaniu z modelem całej głowy i jednocześnie umożliwiałby otrzymanie zbliżonych wartości gęstości mocy. W tym celu opracowano model, który przedstawiono na rys. 3. Twarz odwzorowana jest za pomocą wycinka elipsoidy, a oczy za pomocą kuli i jej koncentrycznych wycinków.



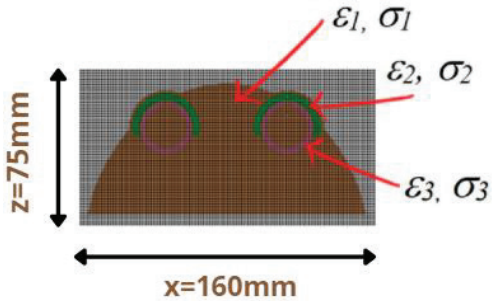
a) b)

Rys. 1. Model heterogeniczny głowy: a – widok, b – przekrój w płaszczyźnie poprzecznej, na wysokości oczu

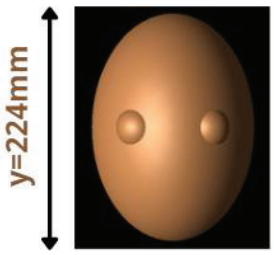


a) b)

Rys. 2. Model heterogeniczny całej głowy (a) oraz model heterogeniczny twarzy (b). Biały obszar wyznacza część modelu, dla której przeprowadzono symulacje.



a)



b)

Rys. 3. Model uproszczony głowy: a – przekrój w płaszczyźnie poprzecznej, na wysokości oczu, b – widok

Wymiary modelu dobrane zostały w taki sposób aby jak najlepiej odwzorować obszar oczu w stosunku do modelu heterogenicznego.

Dobór parametrów materiałowych modelu uproszczonego przeprowadzony został z wykorzystaniem programu Remcom XFdtd oraz algorytmu automatycznej

optymalizacji EStra [9]. Jest to algorytm wykorzystujący strategię ewolucyjną w wariancie najniższego rzędu. Jako operator ewolucyjny zastosowano mutację, która wprowadza do genotypu danego osobnika losowe zmiany (bez krzyżowania genotypów kilku osobników). Niewiadomymi w procesie identyfikacji były ε_k - przenikalności elektryczne k-tej warstwy modelu, σ_k - przewodność elektryczna k-tej warstwy modelu.

Funkcją celu f_1, którą minimalizowano, była największa z różnic pomiędzy danymi uzyskanymi z modelu odniesienia (model heterogeniczny całej głowy) i uproszczonego. Porównywano moduły wektora Poyntinga uzyskane dla 4 punktów usytuowanych w obszarach różnych tkanek oka.

$$f_1(\varepsilon_n, \sigma_n) = \max_{n=1,4} ||\left|S_{ref}(n)\right| - |S_{simp}(n, \varepsilon_k, \sigma_k)|| \quad (1)$$

gdzie:

S_{ref} – wartości Wektora Poyntinga uzyskane z modelu odniesienia,

S_{simp} - *wartości* Wektora Poyntinga uzyskane z modelu uproszczonego

Wartości parametrów materiałowych wyznaczone w procesie optymalizacji przedstawiono w tabeli 1. Wartości początkowe były identyczne z parametrami rzeczywistych tkanek. Początkowa wartość funkcji celu wynosiła 5,64, a końcowa uzyskana w wyniku 47 iteracji wyniosła 1,58. Ze względu na ograniczony obszar modelu uproszczonego, wykorzystuje on jedynie 2,68 mln wokseli dzięki czemu czas obliczeń zredukowano o około 50 % w stosunku do modelu całej głowy. Opracowany w ten sposób model uproszczony został zastosowany do zaprojektowania struktury ekranującej.

Tab. 1. Parametry materiałów tworzących model uproszczony

Parametr	*Wartość*	*Jednostka*
ε_1	8,93	-
σ_1	0,57	[S/m]
ε_2	7,09	-
σ_2	0,44	[S/m]
ε_3	2,01	-
σ_3	1,59	[S/m]

3. PROJEKT STRUKTURY EKRANUJĄCEJ

Opracowana struktura ekranująca ma za zadanie ograniczenie gęstości mocy w obszarze oka w przypadku ekspozycji ciała na falę o częstotliwości 3,5 GHz o polaryzacji liniowej pionowej. Aby nie ograniczała widoczności, taka osłona nie może zostać wykonana z litego materiału przewodzącego. W tym przypadku zastosowano układ prostoliniowych cienkich przewodów [10], których przewodność oraz długość dobrano tak aby częściowo absorbowały energie fali. Takie przewody można wykonać w postaci ścieżek napylanych na przezroczystym tworzywie sztucznym co w minimalnym stopniu ograniczy widoczność użytkownika.

Budowę ekranu przedstawiono na rys. 4. Składa się on z 20 przewodów oddalonych od oka o 32mm. Promień przewodów wynosi R=0,5 mm, a odległość między nimi wynosi $D=4$ mm. Parametry przewodów wyznaczone zostały z wykorzystaniem programu Remcom XFdtd oraz algorytmu automatycznej optymalizacji EStra [9]. Funkcją celu f_2 minimalizowaną w procesie optymalizacji była maksymalna wartość gęstości mocy w obszarze oka, uzyskana podobnie jak w poprzednim przypadku dla 4 punktów we wnętrzu oka.

$$f_2(L, \sigma) = \max_{n=1,4} ||S(n)|| \quad (2)$$

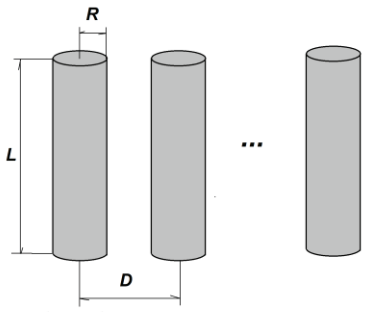
gdzie:

S – wartości Wektora Poyntinga w obszarze oka

L – długość przewodów

σ – przewodność materiału

Parametry podlegające optymalizacji to wysokość każdego z przewodów (h) oraz przewodność elektryczna materiału, z którego zostały wykonane(σ).



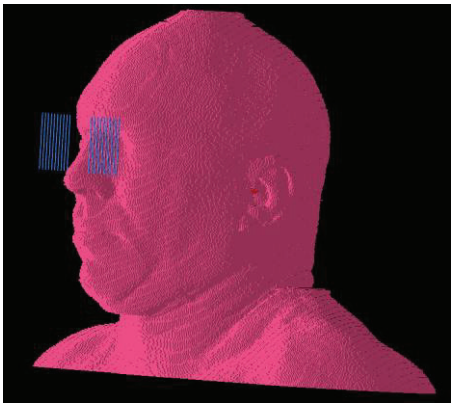
Rys. 4. Wymiary struktury ekranującej

Wartości początkowe wyniosły L=0 mm oraz $\sigma=300\ \frac{S}{m}$. W wyniku 161 iteracji wartość funkcji celu wyniosła 0,83 W/m^2. Czas optymalizacji wyniósł 12 godzin 20 minut. Parametry zoptymalizowanej struktury przedstawiono w tabeli 2.

Tab. 2. Parametry materiałów tworzących strukturę ekranującą

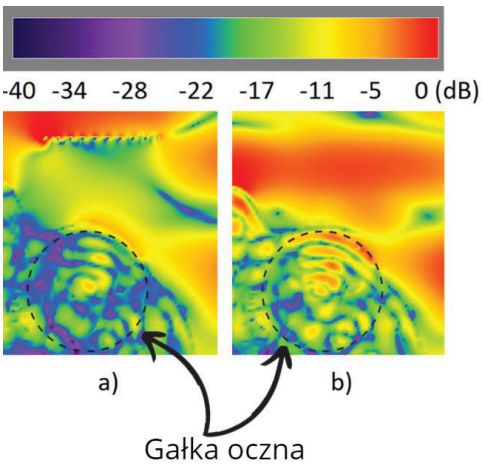
Parametr	*Wartość*	*Jednostka*
ε	1	-
σ	13,83	[S/m]
L	76	[mm]

Opracowaną strukturę poddano weryfikacji za pomocą modelu heterogenicznego całej głowy. Na rys. 5 przedstawiono model heterogeniczny ze struktura ekranującą. Maksymalna wartość gęstości mocy w obszarze gałki ocznej bez zastosowania ekranu wyniosła 11,97 W/m^2. Po zastosowaniu struktury została zmniejszona do 1,04 W/m^2. Wartości gęstości mocy uzyskane w tym eksperymencie wynikają z przyjętej amplitudy fali padającej (**87** V/m).



Rys. 5. Model heterogeniczny ze strukturą ekranującą.

Wyniki symulacji komputerowych przedstawiających rozkład gęstości mocy podczas ekspozycji modelu heterogenicznego na falę elektromagnetyczną przedstawiono na rys. 6. Wykonano symulację dla modelu wraz ze strukturą ekranującą oraz bez niej. Symulacje wykonano dla modelu obejmującego całą głowę, a na rysunku przedstawiono wycinek modelu.



Rys. 6. Rozkład gęstości mocy dla modelu heterogenicznego – przekrój w płaszczyźnie poziomej, obszar prawego oka: z wykorzystaniem struktury ekranującej (a) oraz bez wykorzystania struktury ekranującej (b). Pobudzenie falą o częstotliwości 3.5GHz.

4. PODSUMOWANIE

W artykule przedstawiono sposób projektowania prostej struktury ograniczającej gęstość mocy fali EM dla częstotliwości 3,5 GHz. Ze względu na prostotę budowy, można ją zastosować do ochrony oczu. Zaproponowana struktura może być zintegrowana jako dodatkowa warstwa z typowymi okularami ochronnymi i wykorzystana np. przez techników pracujących w pobliżu masztów sieci 5G.

Aby zwiększyć efektywność iteracyjnego procesu projektowania struktury ekranującej, opracowano zredukowany model twarzy i oczu. Umożliwił on przeprowadzanie symulacji metodą FDTD w czasie krótszym o 50% w stosunku do modelu dokładnego. Opracowany model uproszczony umożliwił skrócenie czasu optymalizacji struktury o około 3,5 godziny.

Podczas dalszych badań planowane jest również przeprowadzenie optymalizacji dla większej liczby parametrów takich jak np. odległość pomiędzy przewodami, w celu uzyskania jeszcze lepszej jakościowo struktury ekranującej. Planowane jest też prowadzenie dalszych prac z wykorzystaniem opracowanej metodyki, których celem będzie opracowanie struktur ekranujących fale o różnych polaryzacjach. Ponadto zbadana zostanie skuteczność zaproponowanej struktury dla częstotliwości odpowiadających wszystkim kanałom rozważanego systemu.

LITERATURA

[1] Pretz K., 2019. „*Will 5G Be Bad for Our Health? IEEE Spectr*”. Available online: https://spectrum.ieee.org/news-from-around-ieee/the-institute/ieee-member-news/will-5g-be-bad-for-our-health (accessed on 20 February 2021).

[2] Tomovski B., Gräbner F., Hungsberg A., Kallmeyer C., Linsel M., 2011. „*Effects of Electromagnetic Field Over a Human Body, Sar Simulation with and Without Nanotextile in the Frequency Range 0.9-1.8GHZ*” J. Electron. Eng., 62, 349–354.

[3] Li Y., Zhang X., Yu J., Wang Q., Tan B., 2008. „*Simulation of EM Field in Head Model and Shielding Effectiveness for Cellular Handset with PIFA*”. In Proceedings of the 2008 World Automation Congress, Waikoloa, HI, USA, pp. 1–4.

[4] Dutta P.K., Jayasree P.V.Y., Baba V.S.S.N.S., 2016„*SAR reduction in the modelled human head for the mobile phone using different material shields*”. Hum. Cent. Comput. Inf. Sci., 6, 44.

[5] Sivasamy R., Kanagasabai M., Baisakhiya S., Natarajan R., Pakkathillam J.K., Palaniswamy S.K. 2013. „*A Novel Shield for GSM 1800 Mhz Band Using Frequency Selective Surface*”. Prog. Electromagn. Res. Lett 38: 193–199.

[6] Farooq U., Iftikhar A., Shafique M., Khan M., Fida A., Mughal M., Anagnostou D., 2021. „*C-Band and X-Band Switchable Frequency-Selective Surface* . Electronics 10: 476.

[7] Mohamadzade B., Hashmi R.M., Simorangkir R.B.V.B., Gharaei R., Rehman S.U., Abbasi Q.H. 2019. „*Recent Advances in Fabrication Methods for Flexible Antennas in Wearable Devices: State of the Art.*”. Sensors 19 : 2312.

[8] Ali S.M., Sovuthy C., Imran M.A., Socheatra S., Abbasi Q.H., Abidin Z.Z. 2020. „*Recent Advances of Wearable Antennas in Materials, Fabrication Methods, Designs, and Their Applications: State-of-the-Art.*”. Micromachines 11 : 888.

[9] Di Barba P. 2010. „*Multiobjective Shape Design in Electricity and Magnetism*”. Springer: Berlin/Heidelberg, Germany.

[10] Januszkiewicz Ł. 2021. „*Analysis of Shielding Properties of Head Covers Made of Conductive Materials in Application to 5G Wireless Systems*”. Energies 14 : 7004.

LOKALIZACJA BEZPRZEWODOWYCH ENDOSKOPÓW KAPSUŁKOWYCH Z WYKORZYSTANIEM ALGORYTMU WYBORU ODBIORNIKÓW

WIRELESS CAPSULE ENDOSCOPE LOCALIZATION WITH RECEIVER SELECTION ALGORITHM

Paweł Oleksy[1]; Łukasz Januszkiewicz[2];

[1] Politechnika Łódzka, Instytut Elektroniki, Łódź, pawel.oleksy@p.lodz.pl
[2] Politechnika Łódzka, Instytut Elektroniki, Łódź, lukasz.januszkiewicz@p.lodz.pl

DOI: 10.15199/59.2022.4.119

Streszczenie: **W artykule przedstawiono modyfikację algorytmu bezprzewodowej lokalizacji endoskopów kapsułkowych, opartego o analizę fazy i mocy sygnału odebranego. Opisany wariant wykorzystuje nadmiarową liczbę odbiorników, dzięki czemu możliwe jest wybranie sygnałów gwarantujących najlepszą zbieżność algorytmu, a co za tym idzie największą dokładność estymacji położenia. Przedstawiony algorytm wyboru odbiorników zewnętrznych w oparciu o kryterium mocy odebranej, pozwolił zmniejszyć błąd lokalizacji o 22%.**

Abstract: **The article presents a modification of the algorithm for the wireless location of capsule endoscopes based, on the received signal phase and power analysis. The described variant uses an redundant number of receivers, thanks to which it is possible to select signals that guarantee the best convergence of the algorithm, and thus the highest accuracy of position estimation. The presented algorithm for selecting external receivers based on the received power criterion allowed to reduce the location error by 20%.**

Słowa kluczowe: **Radiolokacja, FDTD, Modele ciała człowieka, Endoskopy bezprzewodowe.**

Keywords: **Radiolocation, FDTD, Human body models, Radiolocation, Wireless endoscopes.**

1. WSTĘP

Diagnostyka obrazowa układu pokarmowego człowieka może być wykonywana z wykorzystaniem endoskopów. W odróżnieniu od urządzeń przewodowych, ich bezprzewodowe odpowiedniki umożliwiają diagnostykę całego przewodu pokarmowego, która może trwać około 8 godzin. Kapsułki bezprzewodowe wyposażone są w małe kamery, które pomagają wykryć zmiany chorobowe układu pokarmowego w sposób nieinwazyjny i komfortowy dla pacjenta. Pełne wykorzystanie tego typu urządzeń w diagnostyce medycznej wymaga, aby zarejestrowany przez kamerę obraz był uzupełniony przez informację o jej dokładnym położeniu. Jest to bardzo ważne zwłaszcza, gdy dalsze leczenie pacjenta niesie konieczność przeprowadzenia operacji lub precyzyjnej aplikacji środków farmakologicznych. Istotne jest nie tylko zidentyfikowanie położenia kapsułki wzdłuż przewodu pokarmowego, ale również w przestrzeni trójwymiarowej tak, aby możliwe było powiązanie danych zebranych za pomocą kapsułki z wynikami zebranymi za pomocą metod diagnostyki obrazowej takich jak np. rezonans magnetyczny.

Dotychczas opracowano wiele metod lokalizacji endoskopów kapsułkowych, które można podzielić na te, które bazują na analizie pola elektromagnetycznego [4,5,7,8,10], analizie obrazu [1,2] oraz śledzeniu i estymacji położenia na podstawie informacji z czujników inercyjnych [2]. Metody lokalizacji endoskopów kapsułkowych można sklasyfikować również ze względu na to w jaki sposób definiowane jest położenie obiektu w stosunku do ciała. Wyróżnić można metody pozwalające wyznaczyć położenie endoskopu wzdłuż przewodu pokarmowego oraz te, które identyfikują położenie w trójwymiarowym układzie odniesienia. Pierwsza grupa metod bazuje na śledzeniu przemieszczeń i orientacji kapsułki. W przypadku drugiej grupy, najczęściej stosuje się metody bazujące na analizie pola elektromagnetycznego. Przykładowo, rozwiązania wykorzystujące informację o rozkładzie natężenia pola magnetycznego dają wyniki obarczone błędem nie przekraczającym 20 mm [4,7], co jest wielkością porównywalną z rozmiarem endoskopu. Inne metody (do których należy opisana w tym artykule) wykorzystują algorytmy bazujące na parametrach sygnału odebranego, takich jak poziom mocy odebranej, czas propagacji sygnału, różnica fazy sygnału odebranego. Ze względu na złożony kanał propagacyjny jakim jest ciało człowieka, w tym przypadku konieczne jest zastosowanie modelu uwzględniającego parametry elektryczne poszczególnych tkanek.

Wszystkie wspomniane metody posiadają pewne ograniczenia, które znacząco wpływają na dokładność lokalizacji. W przypadku algorytmów bazujących na analizie pola elektromagnetycznego bardzo duży wpływ na parametry sygnału odebranego ma niejednorodny kanał propagacyjnego jakim jest ciało człowieka. W rezultacie nie jest możliwe jednoznaczne odwzorowanie zmierzonych wielkości na odległość nadajnika od poszczególnych węzłów odbiorczych zlokalizowanych na ciele pacjenta. W tym przypadku zastosowanie dokładnych modeli ciała wykonanych np. przy pomocy skanera NMR (ang. Nuclear Magnetic Resonance) pozwala znacząco poprawić dokładność lokalizacji jednak jest to rozwiązanie kosztowne i czasochłonne. Z drugiej strony, w przypadku metod bazujących na analizie obrazu lub śledzeniu ruchu kapsułki z wykorzystaniem czujników inercyjnych, otrzymane wyniki mogą być obarczone błędem wynikającym z ruchu pacjenta czy też zmieniającej się prędkości przemieszczania się kapsułki w poszczególnych odcinkach układu pokarmowego.

Metody bazujące na analizie sygnału odebranego w poszczególnych węzłach odbiorczych są wrażliwe na rozmieszczenie odbiorników na ciele pacjenta. Zaproponowany dotychczas algorytm może wykorzystać informację o fazie sygnału pochodzącą od dowolnej liczby odbiorników zewnętrznych [9]. Jak pokazały przeprowadzone badania, niektóre odbiorniki nie dostarczają istotnej poprawy końcowej estymaty, i można pochodzące od nich dane odrzucić. W związku z tym opracowano metodę, w której z zestawu kilkunastu odbiorników umieszczonych na ciele pacjenta algorytm automatycznie wybiera tylko 5 w oparciu o kryterium mocy odebranej.

2. ALGORYTM LOKALIZACJI

Opracowany dotychczas algorytm lokalizacji endoskopu kapsułkowego umożliwia automatyczne dopasowanie współczynnika skrócenia fali dla danego położenia [9]. W zaproponowanym podejściu współrzędne geometryczne kapsułki w przestrzeni trójwymiarowej wyznaczane są z wykorzystaniem metody PDoA (ang. *Phase Difference of Arrival*). Wymagane są w związku z tym informacje o różnicy fazy sygnałów o częstotliwościach f_1 i f_2 odebranych w poszczególnych odbiornikach, współrzędnych odbiorników zewnętrznych oraz początkowej wartości przenikalności elektrycznej modelu. Dokładność lokalizacji z wykorzystaniem takiego podejścia zależy w dużym stopniu od rozmieszczenia odbiorników zewnętrznych. W pierwszym etapie algorytm przedstawiony na rysunku 1 dokonuje selekcji 5 odbiorników zewnętrznych. Dla uproszczenia przyjęto, że antenami odbiorników są skrócone dipole ustawione pionowo. Dostępnych jest 9 odbiorników, jednak biorąc pod uwagę fakt, że kapsułka endoskopowa zmienia swoją orientację podczas przemieszczania się w układzie pokarmowym, wprowadzono dodatkowe odbiorniki wykorzystujące anteny odbiorcze w postaci dipoli ustawionych poziomo, tak jak to zaprezentowano na rysunku 2. Do wyznaczenia położenia kapsułki wybierane są te odbiorniki dla których poziom mocy odebranej jest największy, przy czym informacja o fazie sygnału nie może być wykorzystana jednocześnie z anten A_n i B_n.

W kolejnym kroku, z równania 1 na podstawie zmierzonej różnicy faz dla wybranych odbiorników zewnętrznych wyznaczane są odległości od kapsułki endoskopowej. Zastosowany model ciała charakteryzuje się taką samą uśrednioną wartością przenikalności elektrycznej w kierunku do każdej anteny odbiorczej, jednak wartość ta może być różna dla kolejnych pozycji kapsułki. W środowisku rzeczywistym uśrednione wartości przenikalności będą się różnić, dlatego wyznaczone w tym kroku odległości są jedynie przybliżone.

$$d_n = \frac{c}{2\pi * \sqrt{\varepsilon}} * \frac{\Delta\varphi_n}{\Delta f} \qquad (1)$$

d_n - odległość kapsułki od N-tej anteny umieszczonej na ciele człowieka,

c - prędkość światła w próżni,

ε - przenikalność dielektryczna modelu,

$\Delta\varphi_n$- różnica fazy sygnału w odbiorniku wysyłanego na dwóch częstotliwościach f_1 i f_2,

Δf - różnica pomiędzy częstotliwością f_1 i f_2.

Następnie do wyznaczenia współrzędnych geometrycznych kapsułki wykorzystany jest algorytm Gaussa-Newtona. Celem algorytmu jest znalezienie takich współrzędnych, które pozwolą zminimalizować funkcję S (2).

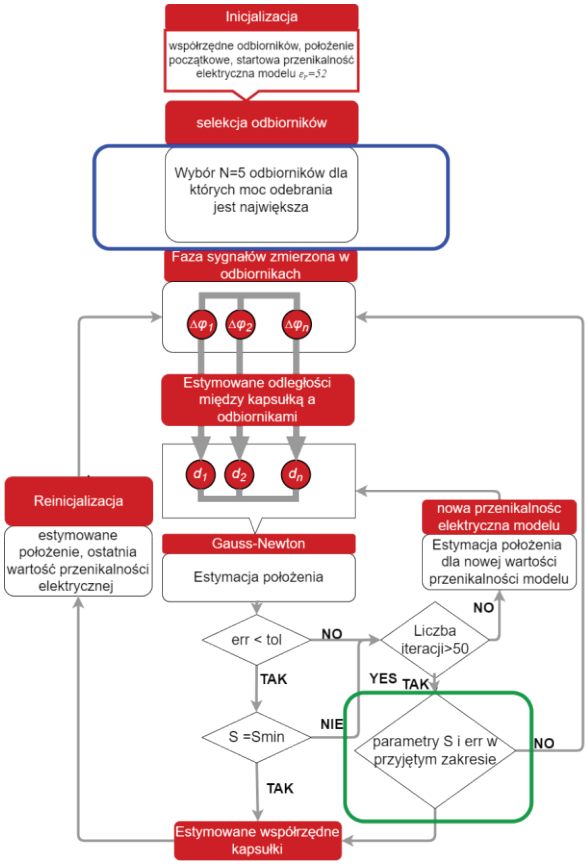
$$S = \sum_{n=1}^{N} r_n^2 \qquad (2)$$

N - liczba anten odbiorczych umieszczonych na ciele,

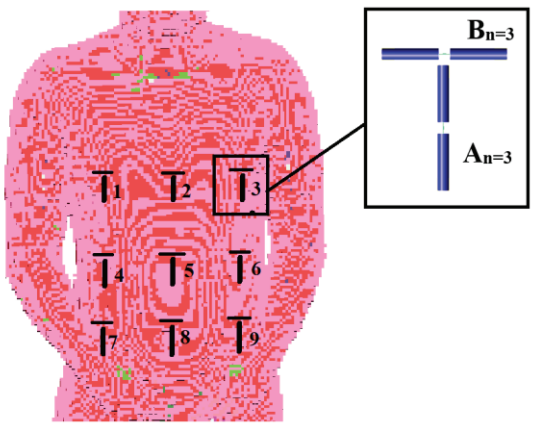
$$r_n = d_n - \sqrt{(x-u_n)^2 + (y-v_n)^2 + (z-w_n)^2},$$

x, y, z - współrzędne kapsułki,

u_n, v_n, w_n współrzędne kolejnych anten umieszczonych na ciele.



Rys. 1. Algorytm lokalizacji kapsułki endoskopowej



Rys. 2. Ułożenie anten odbiorników zewnętrznych na ciele

Dla każdego estymowanego położenia kapsułki wyznaczana jest funkcja błędu, która jest różnicą odległości dwóch ostatnich iteracji algorytmu Gaussa-Newtona. Przyjęto, że aby warunek zbieżności algorytmu został osiągnięty wartość ta musi być mniejsza niż czterokrotność wysokości anteny kapsułkowej. Dla początkowej wartości przenikalności elektrycznej ε_r=52 warunek ten bardzo często nie jest spełniony, dlatego w kolejnym kroku algorytmu poszukiwana jest taka wartość przenikalności, która pozwoli go osiągnąć. Kolejnym krokiem mającym na celu poprawę dokładności algorytmu jest minimalizacja funkcji S, zgodnie z równaniem 3, poprzez zmianę wartości przenikalności w zakresie, dla którego osiągnięto zbieżność metody w poprzednim kroku. W rzeczywistości jest to wartość z zakresu od 1 do 60.

$$S_{min} = \min_{\varepsilon_r \in <1,60>} |S(\varepsilon_r)| \tag{3}$$

Ostatni krok algorytmu stanowi klasyfikacja otrzymanych wyników na podstawie oszacowanych wartości funkcji S i e_{rri} zgodnie z poniższymi zależnościami:

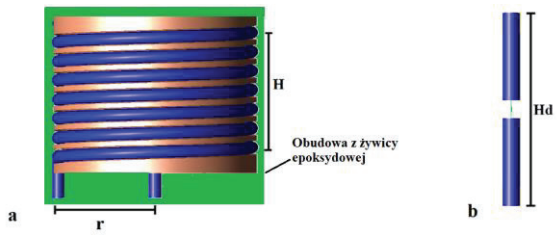
$$S(\varepsilon_{r,}, \Delta\varphi) \leq N * 30 \tag{4}$$

$$e_{rri}(\varepsilon_{r,}\Delta\varphi) < 4 * H \tag{5}$$

H – wysokość anteny kapsułkowej

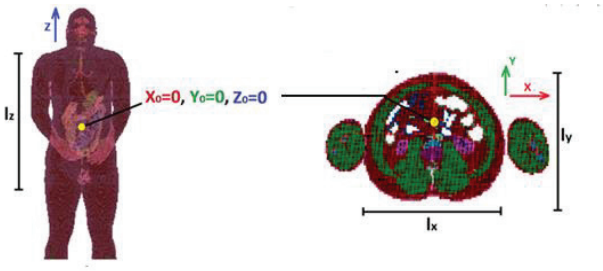
3. WERYFIKACJA ALGORYTMU

Ze względu na rozmiar kapsułki endoskopowej, której długość wynosi około 2 cm koniecznym jest zminimalizowanie rozmiarów zastosowanej w niej anteny. Z drugiej jednak strony rozmiar anteny jest odwrotnie proporcjonalny do częstotliwości sygnału, która w przypadku pasma MICS wynosi 401 do 406 MHz. Dlatego aby zmniejszyć jej rozmiar konieczne było zastosowanie konstrukcji spiralnej. W przeprowadzonych symulacjach wykorzystano antenę przedstawioną na rysunku 3a [8]. Jej konstrukcja uwzględnia karkas wykonany w technologii druku 3D, który pozwolił na wykonanie fizycznego odpowiednika z zachowaniem wszystkich wymiarów. Dodatkowo antena została umieszczona w obudowie z żywicy epoksydowej o przenikalności elektrycznej 3,6, która zapobiega znacznemu odstrojeniu impedancyjnemu, podczas gdy znajduje się ona we wnętrzu układu pokarmowego. Jako odbiorniki zewnętrzne zastosowano skróconą antenę dipolową przedstawioną na rysunku 3b.



Rys. 3. wykorzystane anteny: a. antena kapsułkowa [8] H=5,6 mm, r=4 mm, antena zewnętrzna Hd = 22 mm.

Weryfikację algorytmu lokalizacji przeprowadzono w oparciu o symulacje w środowisku XFdtd Remcom z wykorzystaniem heterogenicznego modelu ciała człowieka NMR Hershey o wymiarach *lx* =330 mm, *ly* = 290 mm, *lz* = 650 mm przedstawionego na rysunku 4. Przyjmując układ odniesienia przedstawiony na rysunku 4, w przeprowadzonych symulacjach zmieniano zarówno pozycję kapsułki w zakresie x = $x_0 \pm$ 70 mm, y = $y_0 \pm$ 40 mm, z = $z_0 \pm$ 120 mm oraz jej orientacje dokonując obrotu wokół osi x i y o 90˚



Rys. 4. Wykorzystany model numeryczny wraz z przyjętym okładem współrzędnych lx=330 mm, ly=290 mm, lz=650 mm

Na rysunku 5 przedstawiono wyniki analizy błędu lokalizacji z wykorzystaniem adaptacyjnego modelu ciała oraz algorytmu doboru odbiorników zewnętrznych. Rysunek 6 przedstawia dystrybuantę rozkładu prawdopodobieństwa błędu lokalizacji. Biorąc pod uwagę fakt, że orientacja kapsułki endoskopowej ma wpływ na błąd lokalizacji, porównano dwie wersje ułożenia odbiorników: odbiorniki ułożone pionowo (A_n) oraz odbiorniki ułożone pionowo i poziomo (A_n i B_n). Przedstawiony błąd lokalizacji opisany równaniem 6 został zdefiniowany jako odległość pomiędzy rzeczywistym położeniem kapsułki a pozycją estymowaną.

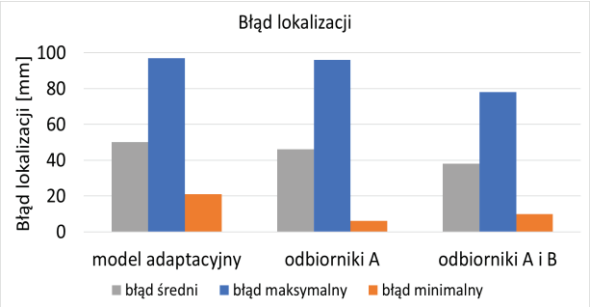
$$Le = \sqrt{(x_t - x_{est})^2 + (y_t - y_{est})^2 + (z_t - z_{est})^2} \tag{6}$$

x_t, y_t, z_t – współrzędne rzeczywistego położenia kapsułki.

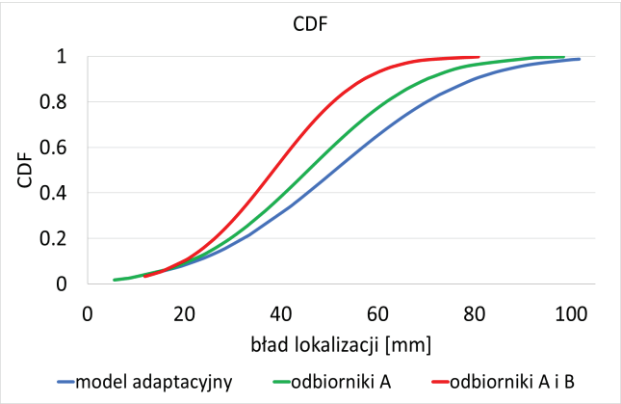
x_{est}, y_{est}, z_{est} - współrzędne estymowanego położenia kapsułki.

Przeprowadzone symulacje z wykorzystaniem heterogenicznego modelu ciała człowieka pokazały, że adaptacyjny dobór odbiorników zewnętrznych pozwala znacząco zmniejszyć błąd lokalizacji względem metody wykorzystującej jedynie adaptacyjny model ciała człowieka [9].

Średni błąd lokalizacji oryginalnego algorytmu wynosi 50 mm. Stosując jedynie odbiorniki wyposażone w anteny ułożone pionowo, średni błąd lokalizacji wyniósł 46 mm, korzystając dodatkowo z odbiorników z antenami ułożonymi poziomo była to wartość 39 mm . Ponadto zaznaczony zieloną ramką na rysunku 2 algorytm oceny otrzymanych wyników wskazał, że estymacja położenia kapsułki z wykorzystaniem adaptacyjnego doboru odbiorników jest możliwa dla około 10% więcej przypadków.



Rys. 5. Porównanie błędu lokalizacji dla poszczególnych wersji algorytmu



Rys. 6. Dystrybuanta rozkładu prawdopodobieństwa błędu lokalizacji

4. WNIOSKI

W artykule zaprezentowano rozszerzoną wersję algorytmu lokalizacji endoskopów kapsułkowych z wykorzystaniem adaptacyjnego modelu ciała człowieka. Zastosowana metoda doboru odbiorników zewnętrznych pozwala zmniejszyć błąd o około 22 % w stosunku do algorytmu oryginalnego, również w przypadku, gdy antena kapsułkowa zmienia swoją orientację względem anten odbiorczych. Istotne jest również to, że zmodyfikowany algorytm skutecznie wyznaczył położenie kapsułki dla większej liczby przypadków. Z pewnością zastosowanie większej liczby anten odbiorczych zwiększy koszt i poziom złożoności całego systemu, z drugiej strony poprawa dokładności na poziomie kilkunastu mm znacząco podnosi użyteczność zaproponowanego rozwiązania.

Dalsze prace nad opisanym algorytmem poświęcone będą optymalizacji anten odbiorczych stosowanych do lokalizacji endoskopu.

LITERATURA

[1] W. Khan, S. M. Lutful Kabir, H. A. Khan, A. Al Helal, M. A. Mukit and R. Mostafa, 2016 A localization algorithm for capsule endoscopy based on feature point tracking, International Conference on Medical Engineering, Health Informatics and Technology (MediTec), pp. 1-5, doi: 10.1109/MEDITEC.2016.7835361.

[2] Vedaei, S.S., Wahid, K.A. 2021. A localization method for wireless capsule endoscopy using side wall cameras and IMU sensor. Sci Rep 11, 11204. https://doi.org/10.1038/s41598-021-90523-w

[3] C. Hekimian-Williams, B. Grant, Xiuwen Liu, Zhenghao Zhang and P. Kumar, 2010. Accurate localization of RFID tags using phase difference, IEEE International Conference on RFID (IEEE RFID 2010), 2010, pp. 89-96, doi: 10.1109/RFID.2010.5467268.

[4] V. Cavlu and P. Brennan, "Determining the Position and Orientation of In-body Medical Instruments Using Near-Field Magnetic Field Mapping," in IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology.doi: 10.1109/JERM.2019.2914402.

[5] Chandra, R., Johansson, A. J., &Tufvesson, F. (2013). Localization of an RF source inside the Human body for Wireless Capsule Endoscopy. 48-54. Paper presented at 8th International Conference on Body Area Networks

[6] H. S. Savci, A. Sula, Z. Wang, N. S. Dogan and E. Arvas, 2005. MICS transceivers: regulatory standards and applications [medical implant communications service], Proceedings. IEEE SoutheastCon, 2005., Ft. Lauderdale, FL, USA, pp. 179-182.

[7] Kim, M.-C.; Kim, E.-S.; Park, J.-O.; Choi, E.; Kim, C.-S. 2020 Robotic Localization Based on Planar Cable Robot and Hall Sensor Array Applied to Magnetic Capsule Endoscope. Sensors, 20, 5728. https://doi.org/10.3390/s20205728

[8] Sang Il Kwak, Kihun Chang and Young Joong Yoon, 2005. The helical antenna for the capsule endoscope, IEEE Antennas and Propagation Society-International Symposium, Washington, DC, USA, 2005, pp. 804-807 vol. 2B, doi: 10.1109/APS.2005.1552139..

[9] Oleksy, Paweł, and Łukasz Januszkiewicz. 2022. "Wireless Capsule Endoscope Localization with Phase Detection Algorithm and Adaptive Body Model" Sensors 22, no. 6: 2200. https://doi.org/10.3390/s22062200

[10] Dey, Nilanjan&Ashour, Amira S. &Fuqian, Shi &Sherratt, Robert. 2017. Wireless Capsule Gastrointestinal Endoscopy: Direction-of-Arrival Estimation Based Localization Survey. IEEE Reviews i Biomedical Engineering. PP. 10.1109/RBME.2017.2697950

POSZUKIWANIE OBIEKTYWNYCH ILOŚCIOWYCH PARAMETRÓW OPISUJĄCYCH WPŁYW POLA ELEKTROMAGNETYCZNEGO NA ORGANIZM CZŁOWIEKA

SEARCHING FOR OBJECTIVE QUANTITATIVE PARAMETERS DESCRIBING THE INFLUENCE OF THE ELECTROMAGNETIC FIELD ON HUMANS

Gabriela Kanclerz[1]; Jakub Popielak[1]; Urszula Sołdaj[1]; Sylwia Lefek[1]; Artur Kacprzyk[2]; Tomasz Rok[3]; Rafał Pawlak[4]; Grzegorz Tatoń[3]

[1] Students' Scientific Group of Medical Biophysics, Jagiellonian University Medical College, Kraków, zelazny.kanclerz@student.uj.edu.pl
[2] Doctoral School of Medical and Health Sciences, Jagiellonian University Medical College , Kraków
[3] Department of Biophysics, Faculty of Medicine, Jagiellonian University Medical College, Kraków
[4] National Institute of Telecommunications, Warszawa

DOI: 10.15199/59.2022.4.120

Streszczenie: **Pola elektromagnetyczne obecne w otaczającym nas środowisku, prawdopodobnie nie pozostają obojętne dla naszego zdrowia. Ilościowy pomiar wpływu pola elektromagnetycznego na podstawie zmian w funkcjonowaniu czy morfologii ciała człowieka nie był do tej pory możliwy ze względu na brak miarodajnych metod. Potencjalnym narzędziem mogącym służyć do pomiaru efektów promieniowania elektromagnetycznego może być sakadometr.**

Abstract: **Electromagnetic fields, although present in our environment, are possibly not indifferent to our health. Quantitative measurement of the effect of electromagnetic radiation on the basis of changes in the functioning or morphology of the human body has not been possible so far due to the lack of reliable methods. A potential tool that can be used to measure the effects of electromagnetic radiation may be the saccadometer.**

Słowa kluczowe: **centralny układ nerwowy, pola elektromagnetyczne, sakady, sakadometr**

Keywords: **central nervous system, electromagnetic fields, saccade, saccadometer**

1. WSTĘP

Doniesienia literaturowe na temat wpływu pola elektromagnetycznego (electromagnetic field – EMF) na zdrowie są niejednoznaczne. Nie udało się potwierdzić żadnego z mechanizmów potencjalnych oddziaływań na człowieka poza oddziaływaniami termicznymi, ale postulowane są takie efekty wewnątrzkomórkowe jak: modulacja ekspresji genów, wpływ na zmianę morfologii komórek i błony komórkowe, wpływ na wewnątrzkomórkowe szlaki sygnałowe, homeostazę jonową i poziom wolnych rodników tlenowych [1]. W literaturze dyskutowana jest korelacja między ekspozycją na EMF a karcynogenezą, czy procesami neurodegeneracyjnymi [2]. Kluczowe wydaje się wytypowanie mierzalnego parametru efektu wpływu EMF na zdrowie.

W krakowskim Zakładzie Biofizyki Wydziału Lekarskiego UJ CM podejmowano próby określania obiektywnych i ilościowych parametrów opisujących wpływ EMF na ludzi. Badano na przykład jego wpływ na zmysł słuchu stosując tympanometrię, otoemisję wymuszoną i audiometrię [3]. Określano zmiany temperatury ciała pod wpływem ekspozycji na EMF z zastosowaniem termografii [4], czy też próbowano określić skutki ekspozycji dla autonomicznego układu nerwowego analizując zapisu ciągły aktywności elektrycznej serca i zmian ciśnienia w układzie krwionośnym. Żadne z tych badań nie przyniosło oczekiwanych rezultatów.

Sakadometr to urządzenie pozwalające obserwować i ilościowo opisywać mimowolne ruchy gałek ocznych (sakady). Ruchy sakadowe to proces fizjologiczny, świadczący o prawidłowym funkcjonowaniu ośrodkowego układu nerwowego. Nieprawidłowe mimowolne ruchy gałek ocznych obserwowano w niektórych chorobach neurologicznych, jak otępienie [5], choroba Huntingtona [6] czy stwardnienie zanikowe boczne [7], ale też np. w przypadku uszkodzeń mózgu w wyniku wypadku [8]. Donoszono o istnieniu związku między ekspozycją na EMF a zmianami w aktywności elektrycznej mózgu badanej przy pomocy elektroencefalografii [9]. Badania mimowolnych ruchów gałek ocznych za pomocą sakadometru znajdują zastosowanie w praktyce neurologicznej i mogą odzwierciedlać zaburzenia funkcjonowania ośrodkowego układu nerwowego.

2. MATERIAŁ I METODY

Celem projektu było badanie użyteczności sakadometru jako urządzenia pozwalającego ilościowo opisać wpływ ekspozycji na EMF o częstotliwości radiowej (radio frequency EMF – RF-EMF) wywołanej standardową rozmową telefoniczną na działanie centralnego układu nerwowego. Miał on być odzwierciedlany zmianami w ilościowych parametrach opisujących sakady.

2.1. Sprzęt użyty w badaniu

W badaniu użyty został sakadometr Advanced (Ober Consulting, Polska) [10]. Sakadometr to urządzenie, które umieszcza się na pasku w ten sposób, że znajduje się ono między oczami osoby badanej u nasady nosa. Wykonuje on pomiar ruchu oka w technologii bliskiej podczerwieni za pomocą fotodetektorów Cyclops-ODS. W skład sakadometru wchodzą też lasery (zielony i czerwony). W momencie wykonywania testu osoba badana znajduje się naprzeciwko ściany (optymalnie 1-5m), na której za pomocą

laserów wyświetlane są punkty. Z zależności od typu testu uczestnik ma podążać za zmieniającym się wychyleniem punktu w prawo lub w lewo (np. test dynamiczny) lub w przypadku pojawienia się punktu po jednej stronie jego zadaniem jest skierowanie wzroku w stronę przeciwną (test antysakadyczny). Dwa wspomniane typy testów (dynamiczny i antysakadyczny) zostały w omawianym badaniu użyte równolegle. W analizę zostały włączone trzy parametry ruchu gałki ocznej mierzone w obu testach: latencja (opóźnienie pomiędzy wyświetleniem bodźca, a reakcją oka), czas trwania oraz prędkość szczytowa (maksymalna prędkość kątowa ruchu gałki ocznej). Miejsce wykonywania pomiarów zostało odpowiednio przygotowane (krzesło z zagłówkiem, ekran na ścianie, odległość między krzesłem a ścianą 2m, ujednolicone oświetlenie, opaska fiksująca telefon przy uchu osoby badanej).

Telefon komórkowy użyty w ekperymencie został skonfigurowany w ten sposób, by łączył się jedynie z siecią 3G. Zastosowano aparat Huawei P20 Lite (Huawei, Korea). Najbliższa stacja bazowa obsługiwała sieć 3G za pomocą fali o częstotliwości 900 MHz. Poziom sygnału od stacji bazowej podczas pomiarów był na poziomie RSSI od -88 bB do -92 dBm.

Parametry pola były badane ekspozymetrem ExpoM-RF (Fields at Work GmbH, Szwajcaria). Rejestrował on natężenie EMF pochodzącego do 16stu predefiniowanych zakresów częstotliwości, podawał również całkowite natężenie pola. Pomiar pozwalał więc wyznaczyć zarówno poziom tła, jak i szacować zgrubnie ekspozycję generowaną przez telefon zastosowany w eksperymencie. Próbkowanie odbywało się co trzy sekundy.

2.2. Protokół badania

Każdy uczestnik badania po zapoznaniu z protokołem eksperymentu wykonywał dwa testy z użyciem sakadometru: dynamiczny i antysakadyczny. Następnie odbywała się 15-minutowa symulowana rozmowa telefoniczna, w której przez pierwsze 10 minut uczestnik słuchał nagranego wcześniej tekstu, a przez kolejne 5 minut czytał na głos kolejne rozdziały. Po skończeniu rozmowy osoba badana poddawana była ponownie testom sakadycznym. Ekspozymetr był włączony w trakcie całego badania (tj. od początku pierwszego do końca ostatniego testu z użyciem sakadometru). Z pomieszczeniu w którym odbywały się badania wyeliminowano wszystkie potencjalne źródła EMF, mogące wpłynąć na pozyskane pomiary, a telefon używany do badania był włączany bezpośrednio przed i wyłączany bezpośrednio po symulowanej rozmowie. Pomieszczenie nie było co prawda ekranowane od źródeł zewnętrznych, ale poziom tła był zawsze monitorowany i był na poziomie 0,016 (±0,007) V/m .

W analizie danych zastosowano program TIBCO Statistica 13.3 (StatSoft, Polska). Do porównania wyników testów sakadycznych zastosowano test t-Studenta. Normalność rozkładów badano testem Kołmogorowa-Smirnova.

3. WYNIKI

W badaniu wzięło udział 15 osób w wieku od 18 do 25 lat, w tym 8 kobiet (53,3%). Średnie wyniki testów sakadycznych zostały przedstawione w tabeli 1.

Tab. 1. Uśrednione wyniki testów sakadycznych

Test dynamiczny			
Wynik testu	*Pomiar 1*	*Pomiar 2*	*p*
Latencja	155 (±21)	165 (±46)	0,21
Czas trwania	64 (±8)	66 (±9)	0,08
Max prędkość	470 (±90)	450 (± 100)	0,12
Test antysakadyczny			
Parametr	*Pomiar 1*	*Pomiar 2*	
Latencja	255 (±42)	250 (±40)	0,18
Czas trwania	62 (±10)	62 (±9)	1,00
Max prędkość	366 (±72)	410 (±290)	0,55

W tabeli podane zostały wartości średnich (±odchylenie standardowe).

Analizowano różnice wyników opisujących ilościowo ruchy sakadyczne przed wykonaniem rozmowy telefonicznej i po jej wykonaniu. Wartości p testu t-Studenta dla zmiennych zależnych nie pokazały istotnych statystycznie różnic w żadnym z rozważanych przypadków. Zaobserwować można jednak spodziewane tendencje, np. zwiększenie latencji czy zwiększenie czasu trwania w teście dynamicznym.

Średnie natężenie pola elektromagnetycznego o częstotliwości 900 MHz pochodzących od źródeł zewnętrznych w trakcie badań z użyciem sakadometru wyniosło 0,008 (±0,006) V/m, a w trakcie symulowanej rozmowy telefonicznej 0,020 (±0,011) V/m. Średnie natężenie składowej elektrycznej pola związane z ekspozycją na RF-EMF emitowane przez telefon wzrosło około 2.5 krotnie w stosunku do tła. Należy jednak zaznaczyć, że ekspozymetr znajdował się w odległości około 60 cm od smartfonu, więc w zakresie pola pośredniego.

4. WNIOSKI I DYSKUSJA

W opisywanym eksperymencie nie udało się jednoznacznie potwierdzić użyteczności pomiaru mimowolnych ruchów gałek ocznych z użyciem sakadometru jako jednoznacznej, ilościowej metody do określania wpływu RF-EMF o niskim natężeniu na ośrodkowy układ nerwowy.

W badaniu Terao i wsp. [11] po raz pierwszy użyto pomiaru mimowolnych ruchów gałek ocznych w relacji do rozmowy telefonicznej. W badaniu typu podwójnie ślepej próby wzięło udział 10 uczestników. Wykonywane były następujące testy: *visually guided saccade (VGS), gap saccade (GAP) and memory guided saccade (MGS).* Dodatkowo badany był czas reakcji dłoni w odpowiedzi na sygnał wzrokowy. Testy były wykonywane dwukrotnie, przed i po 30-minutowej ekspozycji na pole elektromagnetyczne wytwarzane przez telefon komórkowy lub pozornej ekspozycji. W badaniu nie udało się uzyskać statystycznie istotnych różnic między grupą, która była eksponowana na działanie pola elektromagnetycznego a pozostałą częścią badanych, jak również pomiędzy testami wykonanymi przed i po ekspozycji.

Ta sama grupa badaczy przeprowadziła ponownie to samo badanie po 3 latach [12], w którym wykonano 4 testy : antysakadowy, *cued saccade* oraz dwa testy z nakła-

dającymi się sakadami. Powtórnie nie uzyskano statystycznie istotnych różnic między testami pre- i postekspozycyjnymi oraz pomiędzy grupą badawczą i kontrolną.

Dotychczasowe badania z użyciem pomiaru mimowolnych ruchów gałek ocznych z użyciem sakadometru nie wykazały istotnej statystycznie zmiany w pomiarach przed i po narażeniu na promieniowanie elektromagnetyczne w trakcie rozmowy przez telefon. Może to wynikać z kilku powodów:

i) Grupy, które wzięły udział w badaniu, były za mało liczne;

ii) Ekspozycja na promieniowanie elektromagnetyczne była zbyt krótka;

iii) Zmiany funkcjonalne ośrodkowego układu nerwowego mogły pojawiać się po dłuższym czasie po ekspozycji, a nie w momencie pomiaru;

iv) Dodatkowe pole elektromagnetyczne emitowane przez telefon komórkowy było zbyt słabe w stosunku do pola tła;

v) Nie uwzględniono krzywej uczenia - za pierwszym razem wykonanie testu jest utrudnione ze względu na brak doświadczenia, co może obniżać wyniki podczas tej fazy eksperymentu;

vi) Pole elektromagnetyczne może wpływać na układ nerwowy w sposób niezaburzający odpowiednich mimowolnych ruchów gałek ocznych;

vii) EMF przy poziomie natężenia obserwowanego podczas standardowej rozmowy telefonicznej nie ma żadnego wpływu na osobę poddawaną ekspozycji.

Mimo wymienionych wątpliwości wydaje się jednak, że pomiar mimowolnych ruchów gałek ocznych mógłby być jest przydatny w ocenie wpływu EMF na funkcjonowanie ośrodkowego układu nerwowego. Kolejne badania eksperymentalne w tej kwestii, szczególnie z uwzględnieniem znacznie większej grupy badawczej, wydają się potrzebne.

W kwestii obiektywnych ilościowych parametrów opisujących wpływ EMF na organizm człowieka wciąż nie udało się odnaleźć jednoznacznie potwierdzonej naukowo metody ani też innego poza termicznymi mechanizmów oddziaływania pola [13] W dobie powszechnie występującego w środowisku pola elektromagnetycznego, poszukiwanie takiej metody powinno być kontynuowane nie tylko ze względów poznawczych, ale też ze względu na zdrowie całej populacji.

LITERATURA

[1] Piszczek Piotr, Wójcik-Piotrowicz Karolina, Gil Krzysztof, Kaszuba-Zwoińska Jolanta. 2021. „Immunity and electromagnetic fields". Environmental Research (200): 111505.

[2] Murat Terzi, Berra Ozberk, Omur Gulsum Deniz, Suleyman Kaplan. 2016. „The role of electromagnetic fields in neurological disorders". Journal of Chemical Neuroanatomy (75): 77-84.

[3] Kacprzyk Artur, Kocoń Sebastian, Składzień Jacek, Rokita Eugeniusz, Pawlak Rafał, Kwiecień Jakub, Tatoń Grzegorz. 2020. „Does the short-term exposure to radiofrequency electromagnetic field originating from mobile phone affect auditory functions as measured by Acoustic Admittance and Evoked Otoacoustic Emission tests?". Electromagnetic Biology and Medicine 39(4):411-418

[4] Rok Tomasz, Basta-Klonowska Karolina, Lisowski Bartosz, Tatoń Grzegorz, Rokita Eugeniusz. 2020. „Termiczne efekty oddziaływania smartfonu na małżowinę uszną". Przegląd Elektrotechniczny 96(1):198-201

[5] Burrell JR, Hornberger M, Carpenter RH, Kiernan MC, Hodges JR. 2012. "Saccadic abnormalities in frontotemporal dementia". Neurology (78):1816-1823.

[6] Grabska N, Rudzińska M, Wójcik-Pędziwiatr M, Michalski M, Sławek J, Szczudlik A. 2014. „Saccadic eye movements in juvenile variant of Huntington disease". Neurologia i Neurochirurgia Polska. (48):236-241.

[7] Burrell JR, Carpenter RH, Hodges JR, Kiernan MC. 2013. "Early saccades in amyotrophic lateral sclerosis". Amyotrophic lateral sclerosis & frontotemporal degeneration (14):294-301.

[8] Mullen SJ, Yücel YH, Cusimano M, Schweizer TA, Oentoro A, Gupta N. 2014. "Saccadic eye movements in mild traumatic brain injury: a pilot study". The Canadian Journal of Neurological Sciences. (41):58-65.

[9] Loughran SP, Verrender A, Dalecki A, et al. 2019. "Radiofrequency Electromagnetic Field Exposure and the Resting EEG: Exploring the Thermal Mechanism Hypothesis". International Journal of Environmental Research and Public Health 16(9):1505.

[10] http://ober-consulting.com/product/saccadometer/

[11] Terao Yasuo, Tomoko Okano, Toshiaki Furubayashi, Akihiro Yugeta, Satomi Inomata-Terada, and Yoshikazu Ugawa. 2007. "Effects of Thirty-Minute Mobile Phone Exposure on Saccades". Clinical Neurophysiology 118: 1545–56.

[12] Okano T, Terao Y, Furubayashi T, Yugeta A, Hanajima R, Ugawa Y. 2010. „The effect of electromagnetic field emitted by a mobile phone on the inhibitory control of saccades". Clinical Neurophysioly. 121(4):603-611.

[13] Bortkiewicz Anna. 2019. "Health effects of Radiofrequency Electromagnetic Fields (RF EMF)." Industrial health, 57(4), 403–405.

NAJNOWSZE BADANIA DOTYCZĄCE CZĘSTOŚCI WYSTĘPOWANIA NADWRAŻLIWOŚCI NA PROMIENIOWANIE ELEKTROMAGNETYCZNE W POLSCE

PAPER TITLE IN ENGLISH (TITLE LENGTH SHOULD NOT EXCEED 2 LINES)

Grzegorz Tatoń[1]; Artur Kacprzyk[2]; Gabriela Kanclerz[3]; Tomasz Rok[1]; Rafał Pawlak[4]; Eugeniusz Rokita[1]

[1] Jagiellonian University Medical College (CM UJ), Kraków, g.taton@uj.edu.pl
[2] Doctoral School of Medical and Health Sciences, CM UJ, Kraków
[3] Students' Scientific Group of Medical Biophysics, CM UJ, Kraków
[4] National Institute of Telecommunications, Warszawa

DOI: 10.15199/59.2022.4.121

Streszczenie**: Celem projektu było zbadanie częstotliwości występowania nadwrażliwości elektromagnetycznej w Polsce i zdefiniowanie obiektywnych kryteriów klasyfikacji osób badanych jako nadwrażliwe. Badanie przeprowadzono na grupie reprezentatywnej liczącej 2000 respondentów pod koniec roku 2020. Maksymalną częstotliwość występowania nadwrażliwości w Polsce oszacowano na poziomie 1,9% stosując dwustopniowe kryterium.**

Abstract**: The aim of the project was to investigate the prevalence of electromagnetic hypersensitivity in Poland and to define objective criteria that classify the examined people as hypersensitive. The survey was conducted at the end of 2020 in a representative group of 2,000 respondents. The maximum prevalence of hypersensitivity in Poland was estimated at 1.9% using a two-stage criterion.**

Słowa kluczowe**: EHS, IEI-EMF, nadwrażliwość elektromagnetyczna, działanie pola elektromagnetycznego na ludzi**

Keywords**: EHS, electromagnetic hypersensitivity, effects of electromagnetic fields, IEI-EMF.**

1. WSTĘP

Jednym z postulowanych problemów zdrowotnych związanych z wpływem pola elektromagnetycznego (electromagnetic field - EMF) o niskiej gęstości mocy jest tzw. idiopatyczna nietolerancja środowiskowa przypisywana polom elektromagnetycznym (idiopathic environmental intollerance attributed to electromagnetic fields, IEI-EMF), potocznie nazywana „nadwrażliwością elektromagnetyczną" (electromagntic hypersensitivity, EHS). Wiąże się ona licznymi niespecyficznymi objawami, które przez osoby dotknięte tym zespołem (EHS+) przypisane są ekspozycji na EMF różnorodnego pochodzenia [1]. Etiologia zjawiska nie jest znana, stąd określenie „idiopatyczna". Przyczyną EHS może być realny wpływ EMF na organizm, ale wiele wskazuje na to, że u jego podstaw leży mechanizm psychologiczny. Może on polegać na tym, że osoby przekonane o negatywnym wpływie EMF przypisują mu odczuwane dolegliwości [2], a u podstaw takiego przekonania leży obawa przed niezrozumianym zjawiskiem. Ostatecznie stres związany z lękiem przed EMF prowadzi do objawów psychosomatycznych.

Nie ma wiarygodnych metod diagnozy EHS, a kwalifikacja pacjentów do grupy EHS+ odbywa się na zasadzie samookreślenia. Stanowi to problem w diagnostyce i potencjalnym leczeniu takich pacjentów, ale również w badaniach naukowych EHS. Podstawowym problemem jest zdefiniowanie grupy osób nadwrażliwych na EMF i grupy kontrolnej (EHS-) [3]. Może to powodować dysproporcje w określaniu częstotliwości występowania EHS i wątpliwości co do wartości tego rodzaju badań. Dotychczasowe badania częstotliwości występowania EHS prowadzone w innych krajach pozwalają ją oszacować na poziomie pomiędzy 0,7% [4], a 8,6% [5], jednak niektóre badania wskazywały częstość na poziomie ponad 14% [6]. Rozbieżności mogą wynikać z rzeczywistych różnic pomiędzy badanymi populacjami, ale mogą być efektem przyjętej w badaniach metodologii, a w szczególności kryteriów identyfikacji osób EHS+. Kwestia ta została opisana szczegółowo w innej pracy [7], tutaj skoncentrujemy się na opisie specyfiki problemu EHS w naszym kraju. Według wiedzy autorów problemy te nie były badane w ostatnim czasie w Polsce w badaniach z wykorzystaniem licznej i reprezentatywnej grupy.

Celem projektu było ustalenie metody kwalifikacji pacjentów do grupy osób nadwrażliwych na EMF, która byłaby pewna i obiektywna. Drugim celem było ustalenie częstotliwości występowania EHS w społeczeństwie polskim i scharakteryzowanie niektórych aspektów związanych z tym zjawiskiem.

2. MATERIAŁ I METODA

Przeprowadzono badanie ankietowe z udziałem 2000 osób stanowiących grupę reprezentatywną dla społeczeństwa polskiego pod względem płci, wieku i miejsca zamieszkania (wieś/miasto). Wykorzystano metodę CATI (Computer Assissted Telephone Interviewing). Dobór respondentów przeprowadzony został losowo, przy użyciu komercyjnej bazy danych w celu zagwarantowania reprezentatywności próby. Niespełnienie tego warunku było główną przyczyną przeszacowania populacji EHS w Polsce (37%) we wcześniejszych badaniach [8].

Analizę statystyczną przeprowadzono o oparciu o pakiet Statgraphics Centurion 19 (Statgraphics Technologies, Inc.). Zależności zmiennych kategoryzowanych badano testem Chi^2. Przyjmowano, że zmienne są zależne na poziomie ufności 95,0% dla $p < 0,05$. W przypadku zmiennych ilościowych wyniki podawane są w postaci średnia (± odchylenie standardowe).

3. WYBRANE WYNIKI I DYSKUSJA

Badanie objęło 2000 pełnoletnich obywateli Polski. Udział kobiet i mężczyzn był zbliżony (p=0,011) do udziału obydwu płci w społeczeństwie i wynosił odpowiednio 52,8% oraz 47,2%. Według danych Głównego Urzędu Statystycznego (GUS) za 2020 rok udział ten wynosił odpowiednio 51,6% i 48,4% [9]. Rozkład wieku badanych przedstawia tabela 1.

Tab. 1. Struktura wieku respondentów w porównaniu do struktury wieku społeczeństwa polskiego (GUS, 2020)

Wiek	*Badanie [%]*	*Dane GUS [%]*
18-29	13,4	16,3
30-39	17,4	19,3
40-49	17,8	18,4
50-59	17,4	14,7
>59	34,1	31,3

Aby określić wiarygodne i obiektywne kryteria klasyfikacji osób EHS+, zadano w kwestionariuszu szereg pytań o samodeklarację nadwrażliwości. Pytania były formułowane na różne sposoby, co pozwoliło zbadać wpływ konstrukcji pytań i kryteriów klasyfikacji do grupy EHS+ na określenie częstotliwości zjawiska. W zależności od tego, jak skonstruowano kryterium klasyfikacji uzyskiwano częstotliwość występowania EHS w Polsce na poziomie od 1,9 % do 21,9 %. Nawet subtelne zmiany w sformułowaniu pytań ankietowych mogło zmieniać to oszacowanie o kilka procent.

Rozsądnym wydaje się założenie, że osoba może być klasyfikowana jako EHS+, gdy jej objawy wymagają pomocy lekarskiej w celu złagodzenia lub wyeliminowania dolegliwości, aktywnie szuka wyjaśnienia swojego problemu, albo gdy dolegliwości znacząco wpływają na jej codzienne życie. Kryteria oparte na tych założeniach zaproponowali Schröttner i Leitgeb [10]. Osoba nadwrażliwa powinna również odczuwać objawy, które nie występują u innych osób w podobnych warunkach ekspozycji, a nadwrażliwość powinna byś potwierdzona przy pomocy więcej niż jednego pytania. Kryteria stosujące jedynie deklarację przynależności do EHS+ dają oszacowania częstości występowania EHS na poziomie od kilku do kilkunastu procent, zastosowanie silniejszych kryteriów ogranicza tę liczbę do kilku procent [3], [10], [11].

Kryterium, uwzględniające wszystkie wyżej wymienione warunki najsilniej ogranicza szacowaną wielkość populacji EHS+, a zdaniem autorów, zapewnia również najbardziej wiarygodne oszacowanie. Osoba może zostać zakwalifikowana jako EHS+ jeśli pozytywnie odpowie na dwa pytania: (1) "Czy doświadczasz objawów zdrowotnych związanych z używaniem urządzeń elektrycznych, których twoim zdaniem nie doświadczają inni ludzie?" i (2) "Czy twoje dolegliwości są na tyle poważne, aby wpłynąć na twoje codzienne życie lub zmusiły cię do skonsultowania się z lekarzem z ich powodu?" Przy stosowaniu takich ograniczeń górna szacunkowa częstość występowania EHS w Polsce wynosi 1,9 % (38 osób na 2000 ankietowanych). Bardziej szczegółowa analiza stanu zdrowia tych osób być może pozwoliłaby wyjaśnić go innymi schorzeniami, więc oszacowanie to stanowi górną granicę częstotliwości EHS w Polsce.

W dalszej części skoncentrujemy się na szczegółowej analizie charakterystyki tych 38miu osób, które według niestosowanego wcześniej bardzo silnego kryterium mogą być uważane istotnie za osoby EHS+.

Wśród osób EHS+ było 15 kobiet i 23 mężczyzn (odpowiednio 39% i 61%). Ze względu na większy odsetek mężczyzn wynik ten na pierwszy rzut oka odbiega od doniesień literaturowych, panuje pogląd, że przeciętna osoba nadwrażliwa, to kobieta w średnim wieku [11], [12]. Zależność statystyczna pomiędzy elektrowrażliwą i płcią w prezentowanych wynikach nie została jednak potwierdzona (p = 0,0966).

Kolejnym problemem jest kwestia, czy częstotliwość występowania EHS+ ma związek z wiekiem. Wyniki pokazujące udział osób EHS+ w przyjętych kategoriach wiekowych pokazuje tabela 2. Pokazano w niej również liczebność populacji poszczególnych kategorii wiekowych i procentowy udział EHS+ w każdej z nich. Liczebność EHS+ w poszczególnych zakresach jest różna, ale jeśli odniesiemy ją do liczebności tych kategorii, to udział procentowy EHS+ w każdej z nich jest podobny. Dla omawianej zależności p = 0,9923, co sugeruje, że częstotliwość występowania EHS nie zależy od wieku.

Tab. 2. Wyniki analizy powiązania wieku osób badanych z nadwrażliwością na EMF (p = 0,9923)

Wiek	*EHS+ N (%)*	*Całkowita (T) N (%)*	*EHS+ / T [%]*
18-29	5 (13)	267 (13,6)	1,9
30-39	6 (16)	348 (17,7)	1,7
40-49	6 (16)	356 (8,6)	1,7
50-59	7 (18)	348 (17,7)	2,0
>59	14 (37)	681 (34,7)	2,0

Tab. 3. Wpływ miejsca zamieszkania na częstotliwość występowania EHS (p=0,3228)

Miejsce zamieszkania	*EHS+ N (%)*	*Całkowita (T) N (%)*	*EHS+ / T [%]*
Obsz. wiejskie	13 (34)	586 (29,8)	2,2
m. <20 tys.	5 (13)	301 (15,3)	1,7
m. 20-100 tys.	11 (29)	396 (18,8)	2,8
m. 100-500 tys.	7 (18)	412 (21,0)	1,7
m. >500 tys.	2 (5)	305 (15,5)	0,7

Zbadano związek pomiędzy miejscem zamieszkania badanych w sensie wielkości miejscowości zamieszkania i częstotliwością EHS (tabela 3). Zastosowano cztery kategorie wielkości miast pod względem liczby ludności: >500 tys., 100-500 tys., 20-100 tys. i <20 tys., piątą kategorią były obszary wiejskie. Analiza statystyczna wykazała, że częstotliwość występowania EHS może nie mieć związku z miejscem zamieszkania (p = 0,3228). W przypadku ograniczenia liczby kategorii miejsca zamieszkania do dwóch (miasto i wieś) wnioski są podobne (p = 0,5019). To sugeruje z kolei, że częstotliwość występowania EHS może nie mieć związku z ekspozycją na wyższe natężenia RF-EMF i EMF w środowisku, jako że obszary miast charakteryzują się większymi wartościami tła.

Ciekawe wyniki uzyskano w odpowiedzi na pytanie o czynniki inne niż EMF, na które badani są uczuleni lub reagują alergicznie. Do wyboru było 9 czynników: futro zwierząt, pyłki, kurz, pleśń, roztocza, gluten, wypełnienia

dentystyczne, nikiel i kosmetyki (tabela 4). Obliczono liczbę czynników wskazywanych przez każdego z ankietowanych. W grupie osób nadwrażliwych 3 osoby (8 %) nie wskazały żadnego czynnika, w pozostałej części populacji odsetek ten wynosił 37,7 %. Średnia liczba wskazywanych alergenów dla EHS+ wyniosła 3,6 (± 2,3), a dla pozostałych 1,6 (± 1,8). Różnica ta jest statystycznie istotna ($p < 0,0001$), co każe odrzucić hipotezę o niezależności tych dwóch zmiennych.

Obserwację tą można tłumaczyć dwojako. Obecność alergii na inne czynniki niż EMF może się wiązać ze zwiększoną podatnością na działanie EMF. Możliwe jest również, że osoby cierpiące na alergie są bardziej czułe na punkcie swojego zdrowia i obawa przed kolejnymi czynnikami, które w ich mniemaniu mogą pogarszać ich samopoczucie prowadzi do objawów psychosomatycznych wywołanych zagrożeniem ze strony EMF.

Drugą hipotezę potwierdza fakt, że osoby EHS+ postrzegają również inne czynniki środowiskowe częściej jako groźne dla ich zdrowia niż pozostali. Badanych poproszono o wskazanie, który z wymienionych czynników wpływa na ich samopoczucie: pogoda, zanieczyszczenie powietrza, stres i hałas (tabela 4). Pierwsze trzy były częściej wybierane przez EHS+ niż przez pozostałych. Tylko w przypadku hałasu procentowa liczba wskazań w obydwu populacjach jest podobna. Osoby EHS+ średnio wybierały 3,5 (±0,8) czynniki, pozostali badani 2,9 (±1,3).

Tab. 4. Ocena wpływu czynników środowiskowych innych niż EMF na osoby nadwrażliwe.

Czynnik	*EHS+ N (%)*	*Całkowita N (%)*	*p*
Pogoda	36 (95)	1389 (71)	0,0012 *
Zan. powietrza	33 (92)	1345 (69)	0,0158 *
Stres	36 (95)	1583 (80)	0,0334 *
Hałas	28 (74)	1352 (69)	0,5285

* parametry wykazujące zależność od EHS+

Tab. 5. Urządzenia wskazywane najczęściej jako źródła EMF powodujące objawy u osób nadwrażliwych

Urządzenie	*N (%)*
Telefon komórkowy	13 (34)
Odbiornik telewizyjny	13 (34)
Komputer	11 (29)
Router WiFi	10 (26)
Laptop	7 (18)
SBTK	4 (11)
Kuchenka mikrofalowa	4 (11)
Linie wysokiego napięcia (LWN)	3 (8)

Kolejnym zagadnieniem były urządzenia stanowiących źródła EMF związane z objawami EHS. Zadano dwa pytania w tym kontekście. Pierwsze było kierowane tylko do populacji EHS+. Osoby nadwrażliwe miały do wyboru 14 urządzeń z listy: telefon komórkowy, komputer, odbiornik telewizyjny, router WiFi, laptop, stacja bazowa telefonii komórkowej (SBTK), kuchenka mikrofalowa, linia wysokiego napięcia (LWN), tablet, monitor/ekran/wyświetlacz, kuchnia indukcyjna, żarówka LED, żarówka energooszczędna, telefon bezprzewodowy (tabela 5). Istniała możliwość zdefiniowania dodatkowych urządzeń, w kategorii „inne". Znalazły się w niej m.in. takie urządzenia jak telewizor (pomimo, że odbiornik telewizyjny był na liście), radio, sieć 5G (w czasie przeprowadzania ankiety standard ten był w fazie testów), nie wiem / trudno powiedzieć. Średnio wskazywano 2,0 (± 2,0) urządzenia.

Na szczycie listy znalazły się telefon komórkowy i odbiornik telewizyjny. Ten ostatni jest w większym stopniu odbiornikiem niż źródłem EMF, więc wskazywanie go wynika raczej niewiedzy na temat działania urządzeń elektronicznych. Warto zauważyć, że zarówno telefon komórkowy, jak i odbiornik telewizyjny plasują się znacznie wyżej niż SBTK. Jest to zgodne są z wcześniej uzyskanymi polskimi wynikami [13], nie jest natomiast zgodne z doniesieniami z innych krajów, gdzie zwykle SBTK postrzegana jest jako większe zagrożenie niż telefon komórkowy [6], [10], [14], [15]. Wydaje się, że postrzeganie telefonów jako większe zagrożenie niż SBTK jest specyfiką Polaków, chociaż wniosek ten tylko częściowo przystaje do wyników analizy kolejnego pytania.

Drugie pytanie związane ze źródłami EMF było kierowane do całej grupy badanej. Proszono o wskazanie dowolnej liczby urządzeń wywołujących u badanego objawy nadwrażliwości lub alergii. Do wyboru były: telefon komórkowy, komputer, LWN, SBTK i jakiekolwiek urządzenie elektryczne (tabela 6). Średnia liczba wskazywanych urządzeń ma związek z klasyfikacją respondentów jako EHS+ ($p < 0,0001$). Osoby wrażliwe wskazywały średnio 2,7 (±1,7) urządzeń, a pozostali 0,4 (±1,0). W ich populacji SBTK wyprzedza telefony komórkowe. W przypadku pozostałych osób telefon komórkowy pozostaje wciąż powyżej SBTK. Osoby EHS+ znacznie chętniej wybierają „jakiekolwiek inne" co może być konsekwencją poszukiwań źródeł ich problemów zdrowotnych.

Tab. 6. Ranking urządzeń stanowiących źródła EMF w odpowiedzi na drugie pytanie dotyczące źródeł

Urządzenie	*EHS+ N (%)*	*Całkowita N (%)*
SBTK	23 (61)	156 (7,9)
Telefon komórkowy	22 (58)	169 (8,6)
Jakiekolwiek inne	21 (55)	120 (6,0)
LWN	19 (50)	227 (11,4)
Komputer	18 (47)	157 (8,0)

Tab. 7. Objawy nadwrażliwości na EMF

Objaw ekspozycji na pole	*N (%)*
Ból głowy	25 (66%)
Ból oczu	8 (21%)
Bezsenność	4 (11%)
Stres	4 (11%)
Problemy z sercem	4 [11%]

Tabela 7 przedstawia wyniki związane z objawami EHS. W pytaniu o objawy ankietowany wybierał je z listy 32 pozycji, a w kategorii „inne" mógł ją uzupełnić. W kategorii tej część osób opisało objawy spoza listy, które po szczegółowej analizie okazywały się jedną z wymienionych opcji. W tabeli pokazano najczęściej wybierane objawy, do których należy dodać ból stawów, problemy z zasypianiem, problemy z koncentracją, lęk i zmęczenie

(po dwie osoby). Duszności, podwyższone ciśnienie, swędzenie lub zaczerwienienie skóry wybierały pojedyncze osoby. W stosunku do doniesień [8], [15]–[16] zestaw objawów najczęściej wskazywanych przez osoby EHS+ jest krótka. Efekty zdrowotne stosunkowo często raportowane w literaturze wskazywane były przez pojedyncze osoby, albo wcale. W innych badaniach najczęściej pojawia się ból głowy (co pokrywa się z prezentowanymi wynikami), ale zwykle wysoko w rankingu jest ogólne zmęczenie, albo przemęczenie. I w prezentowanym zestawieniu i w literaturze zwykle wysoką pozycję zajmuje ból oczu, postrzegany jako objaw związany z ekspozycją na EMF, a który w istocie wynika ze zmęczenia wzroku monitorami, lub wyświetlaczami urządzeń multimedialnych.

4. WNIOSKI

Poprzez zastosowanie bardzo wymagających kryteriów kwalifikujących ankietowanych do grupy osób EHS+ prawdopodobnie udało się uzyskać informacje od osób, które naprawdę cierpią na objawy związane (ich zdaniem) z ekspozycją na EMF. Według wiedzy autorów jest to najbardziej aktualne badanie EHS w naszym kraju z zaangażowaniem tak licznej grupy reprezentatywnej.

Górna granica częstotliwości występowania nadwrażliwości na EMF w Polsce została oszacowana jako 1,9%. Wydaje się, że przynależność do grupy osób postrzegających się jako nadwrażliwe nie zależy od płci, wieku, jak również od miejsca zamieszkania, czyli od ekspozycji na EMF tła.

Osoby zidentyfikowane jako nadwrażliwe deklarują związek swojego samopoczucia z innymi czynnikami środowiskowymi, jak zanieczyszczenie powietrza, pogoda i stres. Częściej niż ogólnej populacji występują u nich alergie/nadwrażliwości na inne czynniki niż pola. Można tłumaczyć to dwojako. Osoby takie istotnie mają predyspozycje do podlegania takim oddziaływaniom, albo bardziej troszczą się o swoje zdrowie i wszelkiego rodzaju czynniki potencjalnie szkodliwe wywołują u nich lęk, prowadzący do objawów psychosomatycznych.

LITERATURA

[1] Baliatsas Christos, Irene Van Kamp, Mariette Hooiveld, Joris Yzermans, and Erik Lebret. 2014. "Comparing Non-Specific Physical Symptoms in Environmentally Sensitive Patients: Prevalence, Duration, Functional Status and Illness Behavior." Journal of Psychosomatic Research 76 (5): 405–13.

[2] Dieudonné Maël. 2020. "Electromagnetic Hypersensitivity: A Critical Review of Explanatory Hypotheses." Environmental Health: A Global Access Science Source 19 (1): 1–12.

[3] Szemerszky Renáta, Z. Dömötör, and Ferenc Köteles. 2019. "One Single Question Is Not Sufficient to Identify Individuals With Electromagnetic Hypersensitivity." Clinical Psychology in Europe 1 (4).

[4] Korpinen Leena H., and Rauno J. Pääkkönen. 2009. "Self-Report of Physical Symptoms Associated with Using Mobile Phones and Other Electrical Devices." Bioelectromagnetics 30 (6): 431–37.

[5] Röösli Martin, Evelyn Mohler, and Patrizia Frei. 2010. "Sense and Sensibility in the Context of Radiofrequency Electromagnetic Field Exposure." Comptes Rendus Physique 11 (9–10): 576–84.

[6] Mei-Chih Meg Tseng, L. Yi-Ping, and C. Tsun-Jen. 2011. "Prevalence and Psychiatric Comorbidity of Self-Reported Electromagnetic Field Sensitivity in Taiwan: A Population-Based Study." Journal of the Formosan Medical Association 110 (10): 634–41.

[7] Tatoń Grzegorz, Artur Kacprzyk, Tomasz Rok, Monika Pytlarz, Rafał Pawlak, and Eugeniusz Rokita. 2021. "A Survey on Electromagnetic Hypersensitivity: The Example from Poland." Electromagnetic Biology and Medicine 00 (00): 1–8.

[8] Tatoń Grzegorz, Piotr Kuterba, Bartosz Lisowski, Tomasz Rok, and Eugeniusz Rokita. 2019. "Nadwrażliwość Elektromagnetyczna w Polsce-Badania Ankietowe (The Influence of Electromagnetic Fields on the Human Health-Survey Studies)." Przegląd Elektrotechniczny 12: 224–28.

[9] Statistics Poland. 2021. "Population. State, Structure and Natural Movement in the Territorial Section in 2021." https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/ludnosc-stan-i-struktura-ludnosci-oraz-ruch-naturalny-w-przekroju-terytorialnym-stan-w-dniu-31-grudnia-2015-roku,6,19.html.

[10] Schröttner Joerg, and Norbert Leitgeb. 2008. "Sensitivity to Electricity - Temporal Changes in Austria." BMC Public Health 8: 1–7.

[11] Huang Po-Chang, Meng-Ting Cheng, and How-Ran Guo. 2018. "Representative Survey on Idiopathic Environmental Intolerance Attributed to Electromagnetic Fields in Taiwan and Comparison with the International Literature." Environmental Health 17 (1): 5.

[12] Gruber Michael J., Eva Palmquist, and Steven Nordin. 2018. "Characteristics of Perceived Electromagnetic Hypersensitivity in the General Population." Scandinavian Journal of Psychology 59 (4): 422–27.

[13] Kacprzyk, Artur, Gabriela Kanclerz, Eugeniusz Rokita, and Grzegorz Tatoń. 2021. "Which Sources of Electromagnetic Field Are of the Highest Concern for Electrosensitive Individuals?–Questionnaire Study with a Literature Review." Electromagnetic Biology and Medicine 40 (1): 33–40.

[14] Röösli Martin, Mirjana Moser, Yvonne Baldinini, Martin Meier, and Charlotte Braun-Fahrländer. 2004. "Symptoms of Ill Health Ascribed to Electromagnetic Field Exposure--a Questionnaire Survey." International Journal of Hygiene and Environmental Health 207 (2): 141–50.

[15] Kato Yasuko, and Olle Johansson. 2012. "Reported Functional Impairments of Electrohypersensitive Japanese: A Questionnaire Survey." Pathophysiology : The Official Journal of the International Society for Pathophysiology 19 (2): 95–100.

[16] Kacprzyk Artur, Gabriela Kanclerz, Urszula Sołdaj, Eugeniusz Rokita, Bartosz Lisowski, and Grzegorz Tatoń. 2020. "Symptoms Related to Idiopathic Environmental Intolerance Attributed to Electromagnetic Fields (IEI-EMF) - a Questionnaire Study." Przegląd Elektrotechniczny 96 (12): 186–89.

Instytut Radioelektroniki i Technik Multimedialnych
Politechnika Warszawska

Instytut Radioelektroniki i Technik Multimedialnych jest jednym z 6 instytutów działających na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej.

Prace naukowo-badawcze koncentrują się w trzech głównych obszarach:

- radioelektroniki,
- technik multimedialnych,
- elektroniki jądrowej i medycznej.

ul. Nowowiejska 15/19, 00-665 Warszawa
+48 22 234 7742
www.ire.pw.edu.pl

Radioelektronika

Radioelektronika łączy teoretyczne i praktyczne aspekty przesyłania i odbioru sygnałów radiowych. Badane i tworzone są systemy radiokomunikacyjne, systemy lokalizacji radiowej i radionawigacji pracujące w różnych zakresach częstotliwości (do teraherców) i pasmach pracy (od systemów wąskopasmowych do ultraszerokopasmowych

Mamy wieloletnie doświadczenie w projektowaniu nowoczesnych, gotowych do użycia systemów i urządzeń radiowych, w tym różnego typu anten, wzmacniaczy mocy, modulatorów, a nawet całych modułów nadawczo-odbiorczych do systemów radarowych.

Ważny obszar naszych badań jest związany z programowalnymi systemami radiowymi, w tym przed wszystkim z inteligentnymi i konfigurowalnymi antenami .

Specjalizujemy się również w badaniach fizycznych aspektów propagacji fal elektromagnetycznych. Zazwyczaj fale są wykorzystywane jako środek transmisji informacji, ale mogą być stosowane również do przekazywania energii. Stosowane są także w badaniach właściwości materiałów (w tym półprzewodników, ferroelektryków, grafenu, kompozytów itp.).

Elektronika jądrowa i medyczna

W zakresie elektroniki jądrowej Instytut uczestniczy w projektowaniu i budowie detektorów oraz elektroniki „front-end" do eksperymentów fizyki wysokich energii, takich jak: COMPASS (CERN), T2K (Japonia) czy ICARUS (Włochy). Instytut bierze też udział w projektowaniu nowych eksperymentów neutrinowych New-PRISM i Hyper-K w Japonii.

W zakresie elektroniki i informatyki medycznej istotnym obszarem badań i nauczania jest rozwój oprogramowania i sprzętu dla systemów wykorzystujących technikę rezonansu magnetycznego (MRI), w tym – opracowywanie nowych środków kontrastowych, m.in. wykorzystujących technikę hiperpolaryzacji (DNP). Badania dotyczą również przetwarzania danych z badań neuroobrazowych MRI (w tym czynnościowych i strukturalnych). Zajmujemy się także systemami akwizycji danych w medycznych urządzeniach obrazujących, rekonstrukcją i przetwarzaniem obrazów, w szczególności w elektrycznej tomografii pojemnościowej oraz systemami informatycznymi w medycynie. Opracowujemy także nowe algorytmy przetwarzania danych dla potrzeb proteomiki. Prowadzone są również prace związane z badaniami populacyjnymi wykorzystującymi zaawansowane algorytmy eksploracji danych.

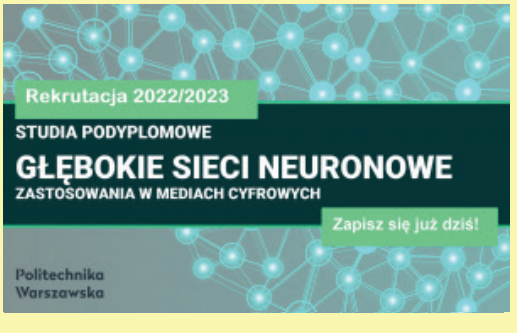
Multimedia

Prace Instytutu dotyczące multimediów łączą różne obszary badań w dziedzinie technik obrazowych oraz dźwiękowych. Obejmują tematy kompresji, rozpoznawania obiektów w obrazach oraz przetwarzania cyfrowych sygnałów wizyjnych. Obszar badań i nauczania obejmuje również prace z zakresu bezpieczeństwa mediów, tworzenia modeli trójwymiarowych, rzeczywistości rozszerzonej oraz nowatorskich aplikacji multimedialnych opartych na głębokich sieciach neuronowych: splotowych i rekurencyjnych.

Grupa elektroakustyczna w swoich badaniach i nauczaniu koncentruje się na zagadnieniach związanych z akustyką, projektowaniem systemów dźwiękowych, przetwarzaniem sygnałów fonicznych z różnymi aspektami percepcji dźwięku oraz ochroną przed hałasem i jego wpływem na ludzi. Również w tych obszarach są wykorzystywane metody uczenia maszynowego.

Prowadzimy studia podyplomowe

siam.ire.pw.edu.pl



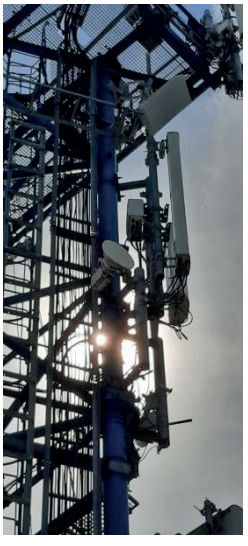
deeplearning.ire.pw.edu.pl

Wszystkich tych, którzy są zainteresowani współpracą i korzystaniem z potencjału badawczego oraz dorobku naukowego naszego Instytutu, zachęcamy do zapoznania się z ofertą zawartą w Katalogu Zespołów Badawczych WEiTI PW:

www.elka.pw.edu.pl/Media/Files/KATALOG_Zespoly-badawcze_WEITI-PW

Instytut Telekomunikacji Politechniki Warszawskiej

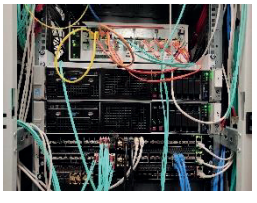
https://secure.tele.pw.edu.pl



Instytut Telekomunikacji funkcjonuje pod tą nazwą od 1975 roku, obejmując w swej strukturze zakłady dydaktyczno-naukowe i zespoły badawcze. Instytut kształci w obszarze sieci, systemów telekomunikacyjnych i teleinformatycznych oraz cyberbezpieczeństwa na studiach inżynierskich, magisterskich, doktoranckich i podyplomowych (w języku polskim i angielskim).

Działalność naukowo-badawcza Instytutu obejmuje m.in. zagadnienia projektowania i analizy sieci teleinformatycznych, sieci następnej generacji, w tym sieci komórkowych 5G/6G, sieci programowalnych, cyberbezpieczeństwa sieci i usług, Internetu rzeczy, przesyłania i przetwarzania informacji z wykorzystaniem fal świetlnych, sprzętowych systemów cyfrowych przetwarzania sygnałów i danych, a także projektowania i wdrażania nowych rozwiązań w obszarze infrastruktury sieciowo-obliczeniowej.

Pracownicy Instytutu są również zaangażowani w realizację praktycznych rozwiązań, wspólnie z dostawcami urządzeń i usług, a także w działalność doradczą zarówno na poziomie krajowym, jak i europejskim (w tym międzynarodowe ciała standaryzacyjne).



Instytut Telekomunikacji dysponuje nowoczesnymi laboratoriami wyposażonymi w zaawansowaną aparaturę technologiczną i pomiarową, odpowiednie narzędzia programistyczne oraz wysoko wykwalifikowany personel.

Struktura

- Zakład Cyberbezpieczeństwa
- Zakład Sieci i Usług Teleinformatycznych
- Zakład Systemów Telekomunikacyjnych
- Zespół Architektur i Zastosowań Internetu

Zespoły badawcze

- Zespół Architektur i Zastosowań Internetu
- Zespół Komunikacji Fotonicznej
- Zespół Optymalizacji i Analizy Sieci
- Zespół Projektowania Systemów Cyfrowych, Przetwarzania Sygnałów i Informacji w Układach Programowalnych
- Zespół Sieci Komórkowych Następnej Generacji
- Zespół Systemów i Usług Internetu Rzeczy
- Zespół Systemów i Sieci Czujnikowych

Faculty of Electronics and Information Technology
WARSAW UNIVERSITY OF TECHNOLOGY

Politechnika Warszawska

Instytut Telekomunikacji
Wydział Elektroniki i Technik Informacyjnych PW
ul. Nowowiejska 15/19
00-665 Warszawa

Tel.: (+48) 22 234 7868
E-mail: sekretariat@tele.pw.edu.pl