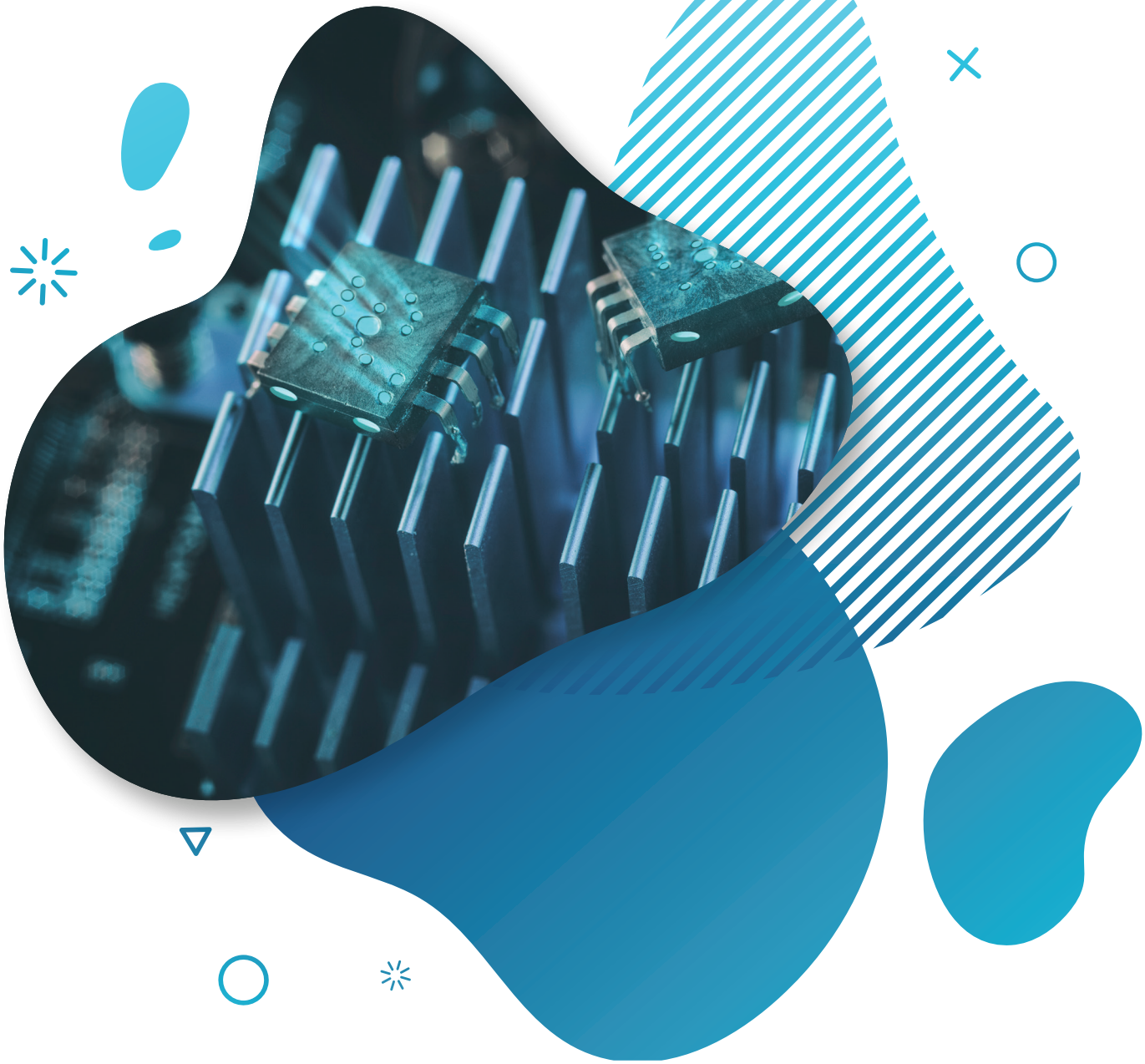
ITU Journal

Future and evolving technologies

FREE I FAST I FOR ALL



Volume 1 (2020), Issue 1

ITU Journal
Future and evolving technologies

Volume 1 (2020), Issue 1

Inaugural issue

ITU Journal
Future and evolving technologies

FREE I FAST I FOR ALL

The ITU Journal on Future and Evolving Technologies (ITU J-FET) is an international journal providing complete coverage of all communications and networking paradigms, free of charge for both readers and authors.

The ITU Journal considers yet-to-be-published papers addressing fundamental and applied research. It shares new techniques and concepts, analyses and tutorials, and learnings from experiments and physical and simulated testbeds. It also discusses the implications of the latest research results for policy and regulation, legal frameworks, and the economy and society. This publication builds bridges between disciplines, connects theory with application, and stimulates international dialogue. Its interdisciplinary approach reflects ITU's comprehensive field of interest and explores the convergence of ICT with other disciplines.

The ITU Journal welcomes submissions at any time, on any topic within its scope.

Publication rights

SUGGESTED CITATION: ITU Journal on Future and Evolving Technologies, Volume 1 (2020), Issue 1.

ADDITIONAL INFORMATION
Please visit the ITU Journal website at:
https://www.itu.int/en/journal/j-fet/Pages/default.aspx

Inquiries should be addressed to
Alessia Magliarditi at: journal@itu.int

Foreword

The new ITU Journal on Future and Evolving Technologies will deliver unique value to academia and the broader ITU community.

It seeks global representation in its published papers and teams of editors and reviewers, and it welcomes an interdisciplinary approach to the broad scope of topics addressed by ITU.

The ITU Journal embodies the inclusive character of ITU and it is this inclusivity that will be the defining factor in the Journal's path to global impact.

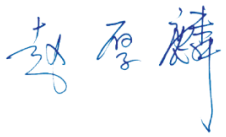
ITU is the United Nations specialized agency for information and communication technologies (ICTs). Our global membership includes 193 Member States and over 900 companies, universities, and international and regional organizations.

In contributing to the Journal and the work of ITU, leading minds in science and engineering are providing decision-makers in the public and private sector with unique insight into the latest developments in ICT and prospects for future innovation.

ITU works in service of the public interest, aiming to ensure that all the world's people share in the benefits of the ICT advances changing our world. This mission is very much aligned with the mission of the global academic community. Therefore, our Journal is free of charge for both readers and authors.

I express my gratitude to all contributors to this first issue of the ITU Journal and I would especially like to thank our Editor-in-Chief, Professor Ian F. Akyildiz, for the great dedication and conviction that define his leadership.

I am very much looking forward to our work together to build a journal focused on ensuring that breakthroughs in research impact our lives for the better, on a global scale.



Houlin Zhao
Secretary-General
International Telecommunication Union

Foreword

The ITU Journal on Future and Evolving Technologies is the latest initiative to contribute to the growing strength of ITU's relationship with academia.

Researchers participate alongside policymakers and industry-leading engineers in ITU expert groups responsible for radiocommunication, standardization and development. Contributions from research communities bring greater strength to the work of ITU, and participation in ITU helps these communities to increase the impact of their research.

Researchers play a key part in ITU's Primetime Emmy winning standardization work for video coding. They are also an important driving force in ITU standardization work in fields such as network orchestration, AI and machine learning, video gaming, blockchain, digital finance, digital health, autonomous driving, and quantum information technology.

We are entering new frontiers in information and communication technology (ICT) and supporting ITU standardization work. Research communities are integral to this movement and this movement is certain to grow stronger with the support of the ITU Journal.

I would like to thank all contributors to this first issue of the ITU Journal as well as our Editor-in-Chief, Ian F. Akyildiz, for creating a first issue that addresses topics of great strategic importance to ITU.

I am very glad to welcome Professor Akyildiz to the ITU community. We share a vision focused on the future and a firm belief that the ITU Journal will deliver unique value to readers and authors worldwide.

Isaac Newton famously said that if he had seen further, it was by standing on the shoulders of giants. Academia is helping the ITU community to see further. I thank you with all my heart for your support.

Chaesub Lee
Director
ITU Telecommunication Standardization Bureau

Editor-in-Chief's Message

On behalf of the International Telecommunication Union (ITU), I welcome you to the inaugural issue of the ITU Journal on Future and Evolving Technologies (ITU J-FET), an international, archival and open access journal providing a publication vehicle for complete coverage of all topics of interest to those involved in all aspects of communications and networking from academia, industry and governments, highlighting standardization perspectives.

The ITU Journal will share new techniques and concepts, analyses and tutorials, and learnings from experiments and physical and simulated testbeds and discuss the implications of the latest research results for policy and regulation, legal frameworks, and the economy and society. As a globally representative publication with a broad scope of interest and prospective impact, the Journal is well positioned to deliver unique value to academia and industry and presents a unique opportunity to share their work with an international audience invested in the broad scope of issues under the remit of ITU.

Free, fast, for all, the Journal aims to promote accessibility of research to academics and industry researchers across the world. The publication is free of charge for both readers and authors, highlighting the true sense of the term, 'open access'.

ITU J-FET will be committed to the timely publication of very high quality, peer-reviewed, original papers that advance the state of the art and applications of communications and networking paradigms. Survey and roadmap papers reviewing the state of the art of relevant topics will also be considered.

A strong editorial board is the key to success. We have created an Editorial Board with truly outstanding, experienced and distinguished people who are in the forefront of the telecommunications research world. We are determined to give detailed, constructive feedback on submitted papers, as well as a fast turn-around time.

All submissions will be handled electronically and must be in PDF format. Please submit your paper (no page limits) through EDAS here.

The submitted papers should be original, unpublished work, and not currently under review for any conference or journal. We understand how much effort goes into creating papers and, accordingly, we aim at a careful, fair treatment of all papers. All of us are committed to doing everything to satisfy our authors' expectations! Regular issues will be accompanied by special issues focusing on particular topics of broad interest within the telecommunications field. If you have any ideas for timely special issues please do not hesitate to contact us.

I want to express my sincere thanks to all the authors of papers who accepted our invitation and published their papers in this inaugural issue which includes ten papers that provide an in-depth analysis of evolving technologies.

Examining the full potential of backscattering in the realization of Internet of Things, "*Backscatter communication with passive receivers: From fundamentals to applications*" presents an overview of recent innovations in hardware architecture for backscatter modulation. "*Non-coherent massive MIMO-OFDM for communications in high mobility scenarios*" proposes the use of non-coherent demodulation schemes to improve performance under scenarios of high mobility while also highlighting new potential areas for research.

The paper on "*MSICA: Multi-Scale Signal decomposition based on independent component analysis with application to denoising and reliable multi-channel transmission*" examines the invaluable tools needed in digital signal processing while "*SDN-based Sociocast group communications in the Internet of Things*" looks into the literature around introducing new disruptive network-layer solutions to address the challenges related to traditional group communication solutions which tend to lack control policies on involved endpoints. Introducing a new paradigm, "*The Internet of Metamaterial Things (IoMMT) and other software enablers*" explores how artificial materials with real time tunable physical properties will significantly enrich the Internet of Things ecosystem. "*Design and analysis of a reconfigurable intelligent meta-surface for vehicular networks*" also introduces a new paradigm, this time for vehicular communications, through the manipulation of electromagnetic waves.

With the COVID-19 pandemic currently ongoing, "*A blueprint for effective pandemic mitigation*" offers a timely focus on contact tracing as a traditional method for mitigating pandemics and the associated challenges. This paper provides a framework on how to contain the spread of a pandemic through the use of wireless technologies, providing numerical results to show the efficacy of the testing strategy.

Considering the central feature in 5G and beyond systems, network slicing, the paper "*Machine learning-assisted cross-slice radio resource optimization: Implementation framework and algorithmic solution*" provides a description of a feasible implementation framework for deploying ML-assisted solutions for cross-slice radio resource optimization. Looking towards future networks including 6G, authors of the paper "*6G vision: An ultra-flexible perspective*" provide an overview of the potential 6G key enablers from the flexibility perspective and give a general framework to incorporate these enablers into future networks. The paper further considers the role of artificial intelligence and integrated sensing as key enablers within this framework.

Finally, "*On the evolution of infrastructure sharing in mobile networks: a survey"* provides a complete picture of infrastructure sharing both over time and in terms of research branches that have stemmed from it such as performance evaluation and resource management. This survey also highlights the relation between infrastructure sharing and the decoupling of infrastructure from services, wireless network virtualization and multi-tenancy in 5G networks. Such relation reflects the evolution of infrastructure sharing over time and how it has become a commercial reality in the context of 5G.

We do hope that readers will enjoy these papers in the inaugural issue. More details about the journal can be found at: https://www.itu.int/en/journal/j-fet/Pages/default.aspx

Many individuals have contributed to the launching of this journal and the preparation of the inaugural issue. Primarily, I would like to express my gratitude to the ITU Secretary-General, Houlin Zhao for entrusting me with the responsibility to lead this Journal as Editor-in-Chief. I would also like to thank the Director of the ITU Telecommunication Standardization Bureau, Chaesub Lee for his support, and Reinhard Scholl and Alessia Magliarditi who believed in my vision and keenly supported this journal from day one. The incredible efforts of Alessia cannot be described in words. Her dedication and very hard work helped to launch this journal successfully. Other team members, Erica Campilongo and Simiso Dlodlo are the engine of the entire operation and their diligent work is always appreciated.

Our journal's administration truly reflects the solid base for the success of this new journal.

Our objective is to become the premier international forum for addressing all aspects of evolving and future technologies in the telecommunications field. The ITU Journal will continue to publish online all year round, welcoming papers at any time, on all topics within its scope with the aim of build bridges between disciplines, connect theory with application, and stimulate international dialogue around the future and evolution of the digital transformation underway across our economies, with five special issues already underway for 2021. Achieving a significant impact factor is also our goal for this Journal and will derive from the relevance of journal papers to the priorities of academia, industry and governments, leading the way to new frontiers in research.

We all look forward to serving you and again special thanks for your interest to make this journal the premier journal in our research community.

Dr. Ian F. Akyildiz
Ken Byers Chair Professor in Telecommunications Emeritus
Director of Broadband Wireless Networking Lab
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 300332, USA
Ian.akyildiz@itu.int

EDITORIAL BOARD

TABLE OF CONTENTS

Page

LIST OF ABSTRACTS

Backscatter communications with passive receivers: From fundamentals to applications

Pages 1-11

Milutin Stanaćević, Akshay Athalye, Zygmunt J. Haas, Samir R. Das, Petar M. Djurić

The principle of backscattering has the potential to enable a full realization of the Internet of Things. This paradigm subsumes massively deployed things that have the capability to communicate directly with each other. Based on the types of excitation and receivers, we discriminate four types of backscattering systems: (i) Dedicated Exciter Active Receiver systems, (ii) Ambient Exciter Active Receiver systems, (iii) Dedicated Exciter Passive Receiver systems, and (iv) Ambient Exciter Passive Receiver systems. In this paper, we present an overview of bacskscattering systems with passive receivers which form the foundation for Backscattering Tag-to-Tag Networks (BTTNs). This is a technology that allows tiny batteryless RF tags attached to various objects to communicate directly with each other and to perform RF-based sensing of the communication link. We present an overview of recent innovations in hardware architectures for backscatter modulation, passive demodulation, and energy harvesting that overcome design challenges for passive tag-to-tag communication. We further describe the challenges in scaling up the architecture from a single link to a distributed network. We provide some examples of application scenarios enabled by BTTNs involving object-to-object communication and inter-object or human-object dynamic interactions. Finally, we discuss key challenges in present-day BTTN technology and future research directions.

View Article

Non-coherent massive MIMO-OFDM for communications in high mobility scenarios

Pages 13-24

Kun Chen-Hu, Yong Liu, Ana Garcia Armada

Under scenarios of high mobility, the traditional coherent demodulation schemes (CDS) have a limited performance, due to the fact that reference signals cannot effectively track the variations of the channel with an affordable overhead. As an alternative solution, non-coherent demodulation schemes (NCDS) based on differential modulation have been proposed. Even in the absence of reference signals, they are capable of outperforming the CDS with a reduced complexity. The literature on NCDS laid the theoretical foundations for simplified channel and signal models, often single-carrier and spatially uncorrelated flat-fading channels. In this work, the most recent results assuming orthogonal frequency division multiplexing (OFDM) signaling and realistic channel models are explained, and the impact of some hardware impairments such as the phase noise (PN) and the non-linear high power amplifier (HPA) are also considered. Moreover, new potential research lines are also highlighted.

View Article

MSICA: Multi-scale signal decomposition based on independent component analysis with application to denoising and reliable multi-channel transmission

Pages 25-35

Abolfazl Hajisami, Dario Pompili

Multi-scale decomposition is a signal description method in which the signal is decomposed into multiple scales, which has been shown to be a valuable method in information preservation. Much focus on multi-scale decomposition has been based on scale-space theory and wavelet transform. In this article, a new powerful method to perform multi-scale decomposition exploiting Independent Component Analysis (ICA), called MSICA, is proposed to translate an original signal into multiple statistically independent scales. It is proven that extracting the independent components of the even and odd samples of a digital signal results in the decomposition of the same into approximation and detail. It is also proven that the whitening procedure in ICA is equivalent to a filter bank structure. Performance results of MSICA in signal denoising are presented; also, the statistical independency of the approximation and detail is exploited to propose a novel signal-denoising strategy for multi-channel noisy transmissions aimed at improving communication reliability by exploiting channel diversity.

View Article

SDN-based Sociocast group communications in the Internet of Things

Pages 37-54

Luigi Atzori, Claudia Campolo, Antonio Iera, Giuseppe Massimiliano Milotta, Giacomo Morabito, Salvatore Quattropani

The new applications populating the Future Internet will increasingly rely on the exchange of data between groups of devices, dynamically established according to their profile and habits (e.g., a common interest in the same software updates and services). This will definitely challenge traditional group communication solutions that lack the necessary flexibility in group management and do not support effective control policies on involved endpoints (i.e., authorized senders and intended receivers). To address the cited issues, the idea of introducing new disruptive network-layer solutions has emerged from recent literature. Among them, Sociocast has been theorized as an enabler of flexible interactions between groups of devices tied by social relationships. In this paper we start from the concept of Sociocast and propose a solution based on Software Defined Networking (SDN) for its implementation at the network layer in the Internet of Things. The performance of Sociocast is studied and compared to methods running at the application layer that provide similar features. Experimental results, achieved through an emulation-based playground, confirm that the Sociocast approach allows for significant reduction of signaling and data packets circulating in the network with respect to traditional approaches.

View Article

The Internet of MetaMaterial Things and their software enablers

Pages 55-77

Christos Liaskos, Georgios G. Pyrialakos, Alexandros Pitilakis, Ageliki Tsioliaridou, Michail Christodoulou, Nikolaos Kantartzis, Sotiris Ioannidis, Andreas Pitsillides, Ian F. Akyildiz

A new paradigm called the Internet of MetaMaterial Things (IoMMT) is introduced in this paper where artificial materials with real-time tunable physical properties can be interconnected to form a network to realize communication through software-controlled electromagnetic, acoustic, and mechanical energy waves. The IoMMT will significantly enrich the Internet of Things ecosystem by connecting anything at any place by optimizing the physical energy propagation between the metamaterial devices during their lifetime, via “eco-firmware” updates. First, the means for abstracting the complex physics behind these materials are explored, showing their integration into the IoT world. Subsequently, two novel software categories for the material things are proposed, namely the metamaterial Application Programming Interface and Metamaterial Middleware, which will be in charge of the application and physical domains, respectively. Regarding the API, the paper provides the data model and workflows for obtaining and setting the physical properties of a material via callbacks. The Metamaterial Middleware is tasked with matching these callbacks to the corresponding material-altering actuations through embedded elements. Furthermore, a full stack implementation of the software for the electromagnetic metamaterial case is presented and evaluated, incorporating all the aforementioned aspects. Finally, interesting extensions and envisioned use cases of the IoMMT concept are discussed.

View Article

Design and analysis of a Reconfigurable Intelligent Meta-surface for vehicular networks

Pages 79-88

Mohammad Ojaroudi, Valeria Loscrí, Anna Maria Vegni

In this work, a new paradigm for vehicular communications based on Reconfigurable Intelligent Meta-surfaces (RIMs) is presented. By using the proposed RIM, we are able to manipulate electromagnetic waves in the half-space, since the element is reflective. The unit cell consists of a U-shaped designed microstrip structure equipped with a pin diode and via a hole. In this study, two different reflection modes are achieved for 1-bit data transferring in each state. By incorporating these two different configurations together, the reflected phases in the proposed RIM surface can be controlled respectively in 0∘ and 180∘. The proposed unit cell can provide a usable double negative functional characteristic around 5.3 GHz. The main goal of this paper is the use of a multifunctional behavior RIM for vehicular communications to code the transmitted wave. A novel phase distribution diagram is generated to propagate in each angle. Moreover, two major electromagnetic modulation functions, beam forming and space coding have been demonstrated. Finally, we show how the RIM can be employed for vehicular communications, acting as a coated access point along the street. We derive the instantaneous data rate at the receiver node, the outage probability and the channel capacity, as affected by different beam widths, distances and vehicle speed.

View Article

A blueprint for effective pandemic mitigation

Pages 89-101

Rahul Singh, Wenbo Ren, Fang Liu, Dong Xuan, Zhiqiang Lin, Ness B. Shroff

Traditional methods for mitigating pandemics employ a dual strategy of contact tracing plus testing combined with quarantining and isolation. The contact tracing aspect is usually done via manual (human) contact tracers, which are labor-intensive and expensive. In many large-scale pandemics (e.g., COVID-19), testing capacity is resource limited, and current myopic testing strategies are resource wasteful. To address these challenges, in this work, we provide a blueprint on how to contain the spread of a pandemic by leveraging wireless technologies and advances in sequential learning for efficiently using testing resources in order to mitigate the spread of a large-scale pandemic. We study how different wireless technologies could be leveraged to improve contact tracing and reduce the probabilities of detection and false alarms. The idea is to integrate different streams of data in order to create a *susceptibility* graph whose nodes correspond to an individual and whose links correspond to spreading probabilities. We then show how to develop efficient sequential learning based algorithms in order to minimize the spread of the virus infection. In particular, we show that current contact tracing plus testing strategies that are aimed at identifying (and testing) individuals with the highest probability of infection are inefficient. Rather, we argue that in a resource constrained testing environment, it is instead better to test those individuals whose expected impact on virus spread is the highest. We rigorously formulate the resource constrained testing problem as a sequential learning problem and provide efficient algorithms to solve it. We also provide numerical results that show the efficacy of our testing strategy.

View Article

Machine learning-assisted cross-slice radio resource optimization: Implementation framework and algorithmic solution

Pages 103-120

Ramon Ferrús, Jordi Pérez-Romero, Oriol Sallent, Irene Vilà, Ramon Agustí

Network slicing is a central feature in 5G and beyond systems to allow operators to customize their networks for different applications and customers. With network slicing, different logical networks, i.e. network slices, with specific functional and performance requirements can be created over the same physical network. A key challenge associated with the exploitation of the network slicing feature is how to efficiently allocate underlying network resources, especially radio resources, to cope with the spatio-temporal traffic variability while ensuring that network slices can be provisioned and assured within the boundaries of Service Level Agreements / Service Level Specifications (SLAs/SLSs) with customers. In this field, the use of artificial intelligence, and, specifically, Machine Learning (ML) techniques, has arisen as a promising approach to cater for the complexity of resource allocation optimization among network slices. This paper tackles the description of a feasible implementation framework for deploying ML-assisted solutions for cross-slice radio resource optimization that builds upon the work conducted by 3GPP and O-RAN Alliance. On this basis, the paper also describes and evaluates an ML-assisted solution that uses a Multi-Agent Reinforcement Learning (MARL) approach based on the Deep Q-Network (DQN) technique and fits within the presented implementation framework.

View Article

6G vision: An ultra-flexible perspective

Ahmet Yazar, Seda Doğan Tusha, Huseyin Arslan

The upcoming sixth generation (6G) communications systems are expected to support an unprecedented variety of applications, pervading every aspect of human life. It is clearly not possible to fulfill the service requirements without actualizing a plethora of flexible options pertaining to the key enabler technologies themselves. At that point, this work presents an overview of the potential 6G key enablers from the flexibility perspective, categorizes them, and provides a general framework to incorporate them in the future networks. Furthermore, the role of artificial intelligence and integrated sensing and communications as key enablers of the presented framework is also discussed.

View Article

On the evolution of infrastructure sharing in mobile networks: A survey

Lorela Cano, Antonio Capone, Brunilde Sansò

Infrastructure sharing for mobile networks has been a prolific research topic for more than three decades now. The key driver for Mobile Network Operators to share their network infrastructure is cost reduction. Spectrum sharing is often studied alongside infrastructure sharing although on its own it is a vast research topic outside the scope of this survey. Instead, in this survey we aim to provide a complete picture of infrastructure sharing both over time and in terms of research branches that have stemmed from it such as performance evaluation, resource management etc. We also put an emphasis on the relation between infrastructure sharing and the decoupling of infrastructure from services, wireless network virtualization and multi-tenancy in 5G networks. Such a relation reflects the evolution of infrastructure sharing over time and how it has become a commercial reality in the context of 5G.

View Article

BACKSCATTER COMMUNICATIONS WITH PASSIVE RECEIVERS: FROM FUNDAMENTALS TO APPLICATIONS

Milutin Stanaćević[1], Akshay Athalye[1], Zygmunt J. Haas[2,3], Samir R. Das[4], Petar M. Djurić[1]
[1]Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794, [2]Computer Science, University of Texas at Dallas, Richardson, TX 75080, [3]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, [4]Computer Science, Stony Brook University, Stony Brook, NY 11794,

NOTE: Corresponding author: Milutin Stanaćević, milutin.stanacevic@stonybrook.edu

Abstract – *The principle of backscattering has the potential to enable a full realization of the Internet of Things. This paradigm subsumes massively deployed* things *that have the capability to communicate directly with each other. Based on the types of excitation and receivers, we discriminate four types of backscattering systems: (i) Dedicated Exciter Active Receiver systems, (ii) Ambient Exciter Active Receiver systems, (iii) Dedicated Exciter Passive Receiver systems, and (iv) Ambient Exciter Passive Receiver systems. In this paper, we present an overview of bacskscattering systems with passive receivers which form the foundation for Backscattering Tag-to-Tag Networks (BTTNs). This is a technology that allows tiny batteryless RF* tags *attached to various objects to communicate directly with each other and to perform RF-based sensing of the communication link. We present an overview of recent innovations in hardware architectures for backscatter modulation, passive demodulation, and energy harvesting that overcome design challenges for passive tag-to-tag communication. We further describe the challenges in scaling up the architecture from a single link to a distributed network. We provide some examples of application scenarios enabled by BTTNs involving object-to-object communication and inter-object or human-object dynamic interactions. Finally, we discuss key challenges in present-day BTTN technology and future research directions.*

Keywords – Backscatter-based communication, baterryless tags, Internet of Things, protocols, tag-to-tag networks

1. INTRODUCTION

The promise of the Internet of Things (IoT) has fomented research into a wide array of wireless technologies and devices capable of providing the required ubiquitous connectivity at a very large scale. In order to maximize the everywhereness and scalability of the IoT, such devices should satisfy the following two key requirements: (i) very low power consumption allowing for batteryless operation and (ii) direct communication with one another and networking without the need for a central master controller. It is in this context that *backscattering* technology has seen a rapid emergence in recent years, beyond its traditional uses in radar and more recently in Radio Frequency Identification (RFID). Backscattering is a form of wireless transmission based on modulated reflection of external RF signals. Since the source of the RF signal is external, such transmission does not require an 'active' radio transceiver, allowing devices to function in an extremely low power regime (under 10 μW). The power needed to operate the transmitter can be harvested from the external RF signal itself, and thus it is possible for such devices to be batteryless.

While the backscattering transmitter is a necessary constituent of every backscatter system, the type of RF excitation source and the type of receiver can vary, giving rise to different classes of systems and networks. In the broad literature, backscatter systems are classified based on the source of excitation into two types:

1. Dedicated exciter (DE) systems: a source of excitation is deployed specifically for the purpose of enabling backscatter transmissions, and

2. Ambient exciter (AE) systems: backscatter transmissions leverage preexisting sources of excitation in the environment such as TV towers, WiFi APs and cell phone towers.

Independent of the excitation source, we posit that an equally important classifying feature of backscatter systems is the type of receiver. Based on this, we identify the following two subclasses of backscatter systems:

1. Active receiver (AR) systems: the receiver is a device with an on-board radio transceiver capable of IQ demodulation and carrier cancellation resulting, typically, in a very high sensitivity (down to -110 dBm for data rate of 20 kb/s and 30 mW power consumption for commercially available transceivers [1]), and

2. Passive receiver (PR) systems: the receiver is a radio-less passive device using an envelope detector for signal demodulation resulting in a much lower sensitivity (-56 dBm for data rate of 8 kb/s

and 236 nW power consumption as in a Bluetooth wakeup receiver [2]).

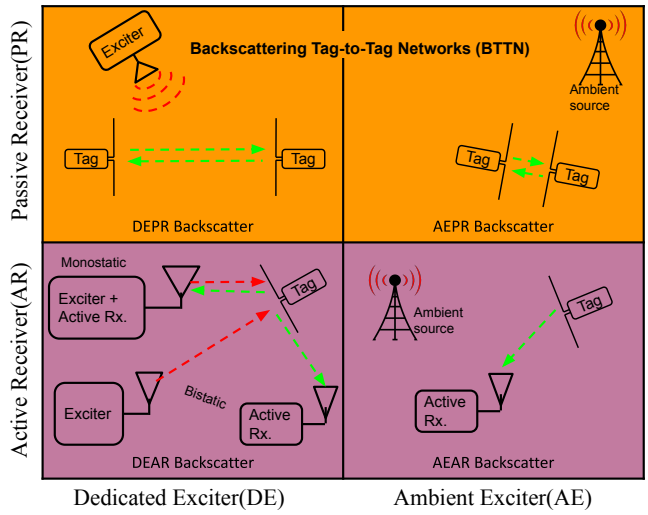


Fig. 1 – Classification of backscatter systems. All such systems contain a backscattering transmitter (Tag). They are classified based on the source of excitation signal and the type of receiver employed.

Combining the above criteria, we classify all backscatter systems into four types as follows:

(i) Dedicated Exciter Active Receiver (DEAR) [3],

(ii) Ambient Exciter Active Receiver (AEAR) [4, 5, 6, 7],

(iii) Dedicated Exciter Passive Receiver (DEPR) [8, 9, 10], and

(iv) Ambient Exciter Passive Receiver (AEPR) [11].

This classification is shown in Fig. 1. As pointed out in [12], DEAR systems can be either monostatic where the exciter and receiver functions are on the same device (e.g., traditional RFID) or bistatic with these functions being on different devices.

We focus specifically on passive receiver (PR) backscatter systems which form the basis of the so-called Backscattering Tag-to-Tag Networks or (BTTNs). PR backscatter systems present fundamentally different challenges than AR backscatter systems. The challenges stem from the need to passively receive the backscatter signals in the presence of the interfering excitation with only an envelope detector. However, if these challenges are overcome, then BTTN can fundamentally transform the capabilities of the IoT by enabling all passively tagged "things" to talk directly with each other without any central active controller or master. In AR systems whether of DEAR or AEAR variety, this capability is impractical since the high cost and high power-requirement of active receivers means that they cannot be used as devices to ubiquitously tag "things" on a large scale.

In summary, the BTTN paradigm utilizing DE or AE can provide a common medium or language for direct peer-to-peer communication between all constituent objects of the IoT irrespective of cost, volume or density of object population. In a sense, the excitation source in BTTN can be thought of as simply illuminating an area, and the tags can "see" each other in this "light." Whether the illumination comes from a natural source (AE approach) or a light bulb (DE approach), the communication between the tags is unaltered.

1.1 Scalability and practical implications of the AE approach

Generally, AE backscatter systems are considered to be extremely scalable because they can theoretically enable communication without any deployment. However, there are important practical considerations that can adversely affect this in a real-world IoT deployment.

- AEAR systems: There are widespread efforts in the literature on these systems. The systems are based on the idea of building passive tags that can use ambient excitation to synthesize backscatter packets that are compatible with commodity standards such as WiFi, Bluetooth or ZigBee [4, 5, 7]. A corresponding commodity receiver can then receive this signal and communicate with the tag. This approach significantly complicates the transmit circuitry on the tag. Further, such tags can only be built to synthesize one kind of backscatter packets using one kind of excitation signal. Under this approach, the tags cannot communicate with each other.
- AEPR systems: These systems fall under the BTTN umbrella. While this, in theory, enables maximum scalability, it is important to note the ambient power level requirement. As shown in [13], in order to enable a practical link distance, the required power level is of the order of -25 dBm. Most ambient excitation signals in general indoor environments from various sources including TV towers, cell phone towers or WiFi APs are far below this value.

Thus, a practical implementation of BTTN might invariably call for a dedicated exciter. However, such an exciter is simply an autonomous, RF transmitter that is not part of the communication network and does not centralize the tag-to-tag communications (refer to the light bulb analogy above). Furthermore, the BTTN backscatter modulator is designed for tag-to-tag communication as opposed to tag to commodity receiver communication. This keeps the "language" and thereby the design of the transmitter much simpler. Unlike the

AEAR approach where backscattering is tuned to one format of the excitation signal, BTTN tags can talk to each other irrespective of the excitation signal format whether it is a CW or a modulated signal. The only requirement is that in the case of a modulated exciter signal, the bandwidth of the tag-to-tag backscatter is lower than the bandwidth of the excitation signal.

Most of the review papers in the literature on the passive backscatter communication focus on the backscatter systems with active receiver(AR) [14] or group both the AR and PR systems in the same category [12] without addressing specific issues that exist in PR systems. In enabling a single BTTN link and further scaling up to a larger network, a vast array of challenges needs to be overcome. These stem mainly from having to process received signals and mitigate interference in passive receivers, to operate in an extreme low power regime (e.g., [15]), and to communicate in inherently high volume and high density networks. In the rest of this paper, we provide an overview of recent advances in BTTNs, challenges, applications of BTTNs, and future directions for research.

2. THE FUNDAMENTALS

One of the most challenging requirements of a PR is to receive the inherently weak backscatter signal in the presence of a much stronger interfering excitation without IQ demodulation or carrier cancellation capability. We illustrate this challenge with a basic BTTN link consisting of two tags in an area that sees a sufficient level of excitation signal whether DE or AE. In this basic link, at any given time, one of the tags transmits (Tx) and the other one receives (Rx). All BTTN tags are identical, and they switch between Tx and Rx roles based on the MAC-layer and network-layer protocols.

The Tx tag generates the modulated backscatter signal by altering the antenna's reflection cross section. The signal seen at the Rx tag is a superposition of the excitation signal and the modulated backscatter signal from the Tx tag. In the absence of an on-board radio transceiver, the Rx tag has to demodulate the backscatter signal using envelope detection. The received signal has a very low modulation index due to the small amplitude of the backscatter signal combined with the much larger magnitude of the exciter signal. Additionally, as the two signals combine at the Rx tag, the modulation index is significantly impacted by the relative phase difference between the excitation signal and the backscatter signal.

In Fig. 2, we see the characteristic of the received baseband backscatter signal amplitudes as a function of the Tx to Rx distance for an ideal link and a PR backscatter link which constitutes BTTNs. In the PR BTTN link, the relative phase difference between the received

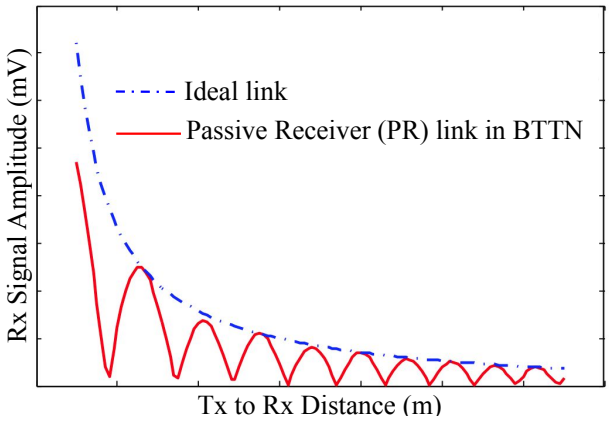


Fig. 2 – Characteristic of backscatter signal amplitude with increasing Tx to Rx distance in BTTN (PR) link compared to an ideal link without phase cancellation.

backscatter signal and the excitation signal causes a phenomenon referred to as *phase cancellation*. This is seen in Fig. 2 (the red solid line), where the received signal instead of monotonically decreasing with distance undergoes alternating peaks and nulls, with decreasing peak values [16, 17]. We note that this phenomenon will also occur in so-called bistatic AR systems [12] where the exciter and receiver are separate. On the other hand, in monostatic AR systems like traditional RFID, the receiver (reader) is able to cancel out the excitation signal and use IQ demodulation for the received backscatter. In this case the received signal amplitude decreases monotonically with distance (the blue dash-dot line). Phase cancellation and low modulation index are two of the most fundamental challenges in enabling basic communication in a PR BTTN link. The phase cancellation can be addressed using a multi-phase backscatter modulator, while signals with low modulation indices are processed with demodulators with innovative architectures [17, 16]. We note that in a link with AR, there are alternative ways to avoid phase cancellation – by providing a frequency shift to the backscattered signal, thereby avoiding interference with the excitation signal altogether [14].

The maximum communication range of a BTTN link (Tx to Rx distance) depends critically on the excitation power available at the Tx tag regardless of whether the excitation source is DE or AE. Distances up to about 3 m have been reported with −20 dBm power available at the Tx tag (5 kbps, BER below 10^{-3}) [16]. Further improvement is possible using coding techniques such as CDMA, but at the expense of data rate. For example, [18] has reported distance up to about 10 m with similar power levels but providing much slower bit rates, in the order of 100 bps. Innovations in the demodulator design can improve the distance and/or improve the data rate. Other innovations are also possible including multiple antennas on the tag [18] or beamforming using multiple tags via a collaborative arrangement. The con-

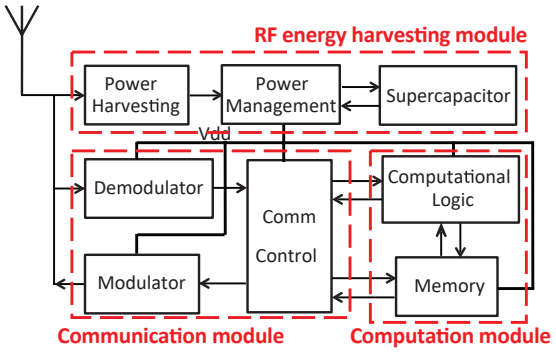


Fig. 3 – The architecture of a backscattering RF tag.

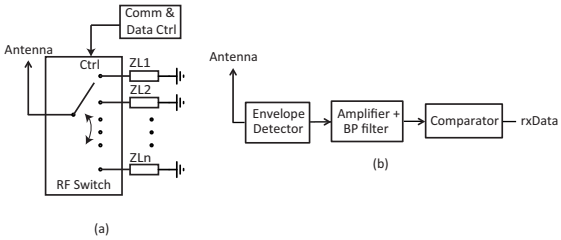


Fig. 4 – Circuit implementation of (a) modulator and (b) demodulator backscattering-based RF tag.

tributions to the backscatter PR tag-to-tag (T2T) links are summarized in Table 1.

3. TAG HARDWARE

The overall architecture of the BTTN tag is shown in Fig. 3. It has three modules, a communication module, an energy harvesting module, and a computation module. The tag optionally interfaces to an external near-zero power sensor. While the sensing and computational module greatly depend on the application, the energy harvesting and communication modules are similar across a wide range of BTTN tags and will be described in greater detail. The control logic manages the operation of the tag while the computational logic, based on the collected data, deduces information on the tag's environment. The power consumption of the BTTN tag is on the order of a few μW as the operating frequency typically does not need to exceed 1 MHz due to a 10s of kbps data rate in a tag-to-tag communication link. The critical resource that requires careful optimization on the system level is memory, both volatile and non-volatile.

3.1 Communication Module

The communication module of the BTTN tag incorporates the passive backscattering transmitter and the PR. These operations are implemented, respectively, by the modulator and demodulator sections.

3.1.1 Modulator architecture

The modulator of the BTTN tag generates the backscatter signal by varying the impedance of the tag antenna between different values (or states). This in turn changes the amplitude and/or the phase of the reflected signal in accordance with the data to be transmitted. This is the conventional *backscatter modulation* process. In a monostatic AR backscattering system like standard RFID, the impedance is typically varied between two values selected so as to maximize the modulation index of the received signal at the reader in the two states. In a BTTN link on the other hand, the backscatter modulation index depends on the relative phase difference between the exciter signal and the backscatter signal seen at the Rx tag. In order to overcome this problem, it was proposed in [8, 17] that the backscatter modulator has the ability to introduce a variable phase offset into the backscattered signal. At some value of the phase offset, the backscatter signal and the excitation signal will be in phase at the Rx tag resulting in the maximum received backscatter amplitude. When the phase offset is shifted by $\pi/2$ from this value, the received backscatter amplitude is minimum. The variable phase offset is achieved by switching the tag antenna impedance between a range of systematically designed values; each such impedance corresponds to one phase in a set of phases that span the range from $-\pi/2$ to $\pi/2$, as illustrated in Fig. 4(a). The number of different phases is a trade-off between the achieved voltage difference in the received signal, communication data rate and the tag form factor.

3.1.2 Demodulator architecture

Demodulating the weak backscatter signal is a fundamental challenge in PR backscatter systems because in the absence of an active radio, the tags need to rely on a passive envelope detector for demodulation. The Rx tag must resolve a weak backscatter signal from the presence of a much stronger external excitation signal resulting in a low modulation index input signal to the demodulator. The communication distance of the BTTN link is directly related to the modulation index that a demodulator can resolve [13]. The demodulator uses an envelope detector that serves as an analog front-end for extraction of the baseband signal. For the detection and demodulation, this analog front-end is followed by a comparator. Because of the much smaller modulation index in the received signal, using conventional RFID tag demodulator architecture leads to short distances of communication [9]. By inserting an amplifier with high-pass filtering after the envelope detection, as illustrated in Fig. 4(b), a tailored demodulator for a tag-to-tag link can demodulate signals with a modulation index as low as 0.5% [13]. The sensitivity of this architecture is related to the power consumption of the amplifier. The ripple voltage in the baseband signal is a critical parameter that determines the performance of the demodulator. To reduce the ripple voltage, higher-order adaptable low-pass filtering could be integrated in the envelope detector prior to signal amplification at a

Table 1 – Summary of the contributions to backscatter PR tag-to-tag (T2T) links

Article	Key contribution	Frequency	Experimental Results
[9]	T2T communication concept	915 MHz	T2T link at 10 cm
[11]	T2T link with ambient exciter	539 MHz	1 kbps at 0.75 m and -8 dBm
[18]	coding technique to extend link distance multi-antenna tag for increased data rate	915 MHz 539 MHz	0.003 kbps at 6 m and -20 dBm 1000 kbps at 2.1 m and -20 dBm
[17]	phase cancellation in T2T link	915 MHz	
[13]	theoretical analysis of T2T link	915 MHz	
[16]	demonstration of multi-hop network	915 MHz	5 kbps at 3 m and -20 dBm
[19]	M-PSK for increased data rate	539 MHz	20 kbps at 0.75 m
[10]	MAC protocol	915 MHz	multi-hop T2T links reach 5.65 m

cost of chip area and power consumption.

3.2 Energy harvesting architecture

The RF energy harvesting module acquires energy from the external excitation signal. A power harvesting circuit comprises rectification of the incident AC voltage, followed by multiplication and regulation that provides stable DC supply voltage for the operation of the tag. The energy efficiency of the conventional power harvesting circuit is optimized for a certain range of input power. As the input power can exceed the power consumption of the tag, the extra energy can be stored using a supercapacitor. This enables the operation of the tag when the harvested energy is lower than the instantaneous power consumption. The size of the supercapacitor is limited by the form factor of the tag. The power management logic optimizes the charging of the supercapacitor based on the incident RF power and the power needs of the tag operation.

Based on the incident RF power, the stored energy and the operation of the tag, e.g., backscatter, receive or compute, the power management module directs the tag's operation. The operation of such tags powered by RF harvested energy and low capacity supercapacitors introduces some unique challenges compared to those of a traditional sensor node. Sensor nodes incorporate active radios that dominate the power budget. Though significant steps have been made in reducing their power consumption at the receiver end [1], the transmit power still dominates the operation as the radios must generate the RF carrier signal used for communication. The principal difference between the power budgets of conventional sensor nodes and the RF tags is that the tags operate at orders of magnitude of lower power consumption due to the low energy cost of the communication, as the energy cost of their communication can be orders of magnitude lower than for the nodes comprising active radios. This is because tags only reflect (backscatter) externally supplied RF signals and do not generate any signal on their own. However, in BTTN tags there is no such dominance – the energy costs for communication between the tags and computation are of the same order of magnitude. Further, the different energy costs for performing different operations on the tags lead to a unique power management paradigm for BTTN.

4. SCALING FROM A SINGLE LINK TO A FULL NETWORK: ROUTING FOR IOT APPLICATIONS

4.1 From a Link to a Network

Extending a single link communication to a tag network is far from trivial [20]. Two issues need to be considered: topology formation and routing. The topology formation involves selection of network links for communication based on the energy states of the individual tags. This decision typically involves tags beyond local neighborhoods and may require dynamic operation as the tags' energy states continuously vary.
For communication across a tag-to-tag link, there must be enough RF power reaching the Rx tag to power up the receiving tag for effective demodulation and then to do any needed post-demodulation computation (e.g., MAC, routing decisions). The power needed for effective demodulation is dependent on the modulation index, which in turn depends on the wireless channel conditions that determine the powers reaching the tags. This is heavily influenced by the tag and exciter locations. As mentioned in Section 3.2, the energy management module decides the power split among the various operations. The Rx tag in a weak link may have to decide whether to receive a packet at all if it may not be able to forward it immediately for a lack of enough available power. Similarly, a more "energy-rich" tag may be able to take up more responsibilities for routing or MAC protocol decisions.

4.2 Routing and MAC

The challenges of designing routing and MAC protocols for BTTN arise from the unique characteristics of the backscattering environment, including the extreme low-power operation and from the intended BTTN applica-

tions [21]. In this section, we discuss these challenges and propose some solutions that have been considered in this field.

First, in BTTNs there is strong dependence of the range over which a tag can communicate on the distance between the tag and the source of its backscattered RF energy, whether an ambient source or an RF exciter is used. In other words, a tag which is located close to the RF energy source will be able to backscatter over a larger distance, than the same tag if it is placed further away from the RF energy source. This is, of course, different than a typical sensor/ad hoc networks, where the length of a link depends on the node itself and does not change with the location of the node. This has a number of implications in the design of the routing and MAC protocols, including the fact that the set of destination of a node in a BTTN depends on the location of the RF energy source relative to the node. Because of this phenomenon, there is also a larger likelihood of unidirectional links being present between two communicating tags, when the two tags are at different distances from the RF source. (In general, this likelihood depends on the tags and the distribution of RF sources.) As BTTNs tend to be distributed in their operation (i.e., there is no central element that coordinates the MAC access or the routing discovery operations) the need to perform these operations over unidirectional links is a more difficult problem than in undirected networks ([22]), often leading to network partitions in the unidirectional graph type. Of course, preservation of network integrity is critical for most networking environments. This is unlike other typical wireless networks where each node is powered by its own battery, thus creating links with similar capabilities in the two directions, a fact that is quite often relied upon in the design of the protocols. (E.g., if node A sends a message to node B, it is given that node B expects to be able to reply to node A on the link in the reverse direction.)

Second, as multiple tags are usually powered by a single RF source, any increase or decrease in the source's RF power is likely to drastically affect large portions of the network topology. Thus, movements of nodes (of the RF power sources) or changes in the RF propagation impairments of the RF sources could significantly, and more problematically nearly instantaneously, affect large portions of the network topology. Similarly, a movement of another RF source into the network coverage area would increase the lengths of at least some of the network links – and typically of all the links in a particular area – thus, creating a topology with richer connectivity. This is unlike other wireless networks (e.g., typical ad hoc or sensor networks), where the changes of topology caused by a movement of a single node are mostly limited to individual nodes in the neighborhood of the moving node only. Such sizable changes of BTTN topology require a much more robust and adaptable routing approach to preserve connectivity and to maintain optimal routes. Furthermore, the fact that these changes occur with little delay, and thus little advanced warning, and the fact that these changes may occur frequently, even more exacerbate the problem.

Third, the backscattering tags experience high level of interference at various protocol layers. Interference is generally not a problem in sparse networks or networks with infrequent communications among the network nodes. But in the envisioned applications of BTTN ([21]), such as those for densely deployed IoT systems, even a simple query might cause at least some message flooding among the tags, significantly affecting the throughput of big portions of the network. This problem is further intensified in real-time IoT applications.

Fourth, in some configurations where the RF exciter is tasked with at least part of the computational functions, some part of the routing and MAC processing could be done by the exciter [20], offloading some of the complexity from the tags. On the other hand, in the case of a zero-intelligence exciter or when an ambient RF source is utilized, all the computations need to be performed distributively by the tags themselves. Thus, to adjust to different operational scenarios, the routing and MAC protocols may need to adapt to the division of processing between the tags and the exciters. Furthermore, in the case of a zero-intelligence exciter, distributed processing is especially a challenge, since the BTTN tags operate at extremely low energy levels, significantly limiting their processing capabilities. Depending on the limited processing capabilities of the tags and their extreme low-power operation, there is a need for new approaches to design very simple MAC and routing protocols [20, 21], such as by trading the protocols' performance for processing complexity.

In applications such as IoT, the network of tags should facilitate interactions among smart objects, each tagged with a passive tag that stores information about the object, such as the object's identity, its capabilities, attributes, and past history of interactions with other objects. As an example, if the BTTN is designed to track infectious contacts among individuals, as to alert them of possible infection ([21]), the lists of contacts need to be stored and maintained in the tags. Routing among such passive tags, each being associated with a particular object, requires creation of a suitable routing infrastructure and appropriate protocols. More specifically, the routing functions consist mainly of: finding paths between specific tags or among related tags; ensuring that the communication among the tags is expedited and takes information priority into consideration; maximizing the network throughput, i.e., concurrent communications among the tags; and reducing the interference among the selected paths. Although the basic operation of the routing protocol is to facilitate communication among the tags, i.e., finding multi-hop routes among

the tags, routing should also facilitate higher-level operations, e.g., searching for a particular object, such as other associated/related tags in the network (e.g., for all the tags that were in close contact with a tag carried by an infected individual); querying to identify all the objects with certain attributes or certain historical values, thus creating "communities of interest" among objects to facilitate interactions and information exchange among such member objects, etc.

As an example of an approach to routing in BTTN, we now briefly discuss how to address two of the specific challenges of routing in BTTN: (a) *routing scalability* in a densely-deployed network and (b) route discovery in the presence of *unidirectional links* in the network.

In a massively deployed network, such as is envisioned for IoT applications, it is difficult to discover whether a particular tag is reachable by another tag. To combat this problem, the tags can establish loose associations, creating *communities of interest* – a collection of related objects, which are interspersed by other objects in the network. For example, all books in a library by a particular author could be an example of a community of interest. In this way, as further explained below, rather than routing a message to a particular book (i.e., a particular tag), a message is anycasted to the "community of books by the author," rather than unicasted to a specific tag. Routing in the network is then performed based on the attributes of a community. When a node moves away or changes its attributes, it removes itself from the particular community of interest. Once a message is delivered to any member of a community of interest (i.e., anycasted), based on the attribute of the community, the member will then share the message with all the other members of its community through intra-community routes. In other words, we proposed a two-level distributed routing hierarchy, where each tag maintains a route to some members of its community of interest, so that delivery to a particular member of a community of interest requires only delivery to one (i.e., any) member of the community. The notion of communities of interest addresses a major challenge in routing in the network of tags – *routing scalability*. Instead of discovering routing paths between *every pair* of tags in the network, routing within only a much smaller community of tags is needed.

We now discuss the second challenge – discovering routing paths in unidirectional graphs. One approach to path discovery in ad hoc networks and sensor networks is through broadcasting *Route Request Query* (RREQ), which is a message sent from the source node to the destination node. As the RREQ propagates through the network, the nodes append their ID to the message, until the message reaches the destination. The destination extracts the accumulated route in the RREQ and creates the *Route Reply Message* (RREP), which is then forwarded back to the source node through reversing the accumulated route. (Any node which receives the RREQ and knows the route to the destination, can create an RREP by appending the known route to the accumulated route in the RREQ and forwarding the RREP back to the source node through reversing the accumulated route.) Unfortunately, the above process will not work in a BTTN, because many links are unidirectional only, thus reversing the route will create an infeasible path. First, we note that our route discovery operates between a source tag and a community of tags, rather than a single destination node. Second, a new RREQ/RREP process could be introduced, where a message *Forward Route Request* (FREQ) is broadcast by the source and propagates (with route accumulation). When the FREQ is received by any member of the community of tags, such a node now becomes the destination node. The destination node, upon receipt of the FREQ, initiates *a new Backwards Route Request* (BREQ), by appending the forward route from the FREQ and broadcasting the BREQ back to the source. When the BREQ arrives at the source, it now contains both, the forward and the backward routes, where the routes in the two directions are not necessarily the same. The source then creates an RREP message with the backward route and uses the forward route to send the RREP to the destination.

5. APPLICATIONS OF BTTNS

In this section, we first explain a fundamental operation of two tagged objects that will facilitate many applications based on object interactions, then we describe an application that involves human interactions, and finally we list a number of possible applications of BTTNs.

5.1 Object interactions

By object interaction we mean exchange of information between two objects with attached tags that are in the proximity of each other and that is used for some purpose. For example, tagged objects can localize themselves relative to one another or even in an absolute sense if some tagged objects serve as anchors, that is, their locations are known. Tagged objects can also track other tagged objects in their neighborhood.

The central problem here is the estimation of distances between communicating tags. One technique for distance estimation is based on multiphase backscattering, where a tag changes the phase offset of the signal that is being backscattered in a systematic manner [23]. Suppose there are two tags, Tag 1 and Tag 2, where Tag 1 acts as Tx tag with different phases. It can readily be shown that the square of the estimated amplitude of the Rx Tag 2 at the output of the envelope detector is a sinusoid that is a function of the used phase offsets and a fixed parameter that carries information about the distance between the two tags. When the roles of Tag 1 and Tag 2 are reversed, i.e., Tag 1 receives and Tag

2 transmits, the same phenomenon occurs. Tag 1 now obtains a sinusoid but with a different fixed parameter. It turns out that when the respective parameters of the sinusoids are added, their sum is equal to $4\pi d/\lambda$, where d is the distance between the tags and λ is the wavelength of the excitation signal. From this relationship, the distance can readily be determined.

Further, with the same line of reasoning, the tags can estimate Doppler shifts due to moving tags. Experimental results suggest that tags can estimate Doppler shifts with about the same accuracy as that obtained by active conventional RFID readers. Also, the median tracking error based on data from two tags can be as low as 2.5 cm [23].

5.2 Human interactions

An interesting application of BTTNs is related to human interactions [8]. Here we present a setting where BTTNs serve as a 'device-free' activity recognition system [8]. Namely, when the tags in the network communicate with each other, the backscatter channel state is influenced by the surrounding environment. The channel state thus carries information that can be used for classification of dynamic activities that take place in the proximity of the tags. As explained earlier, with multiphase backscattering, the communication between two tags becomes more reliable. It turns out that this is not the only advantage of the scheme. Multiphase backscattering also helps to quantify channel state information that can serve as a unique signature of activities which in turn allows for their accurate classification.

More specifically, when a Tx tag backscatters the external signal with different phases, the Rx tag can compute features of these signals. These features vary according to the dynamic alterations of the multipath wireless channel between the tags. When there is no one near the communicating tags, the amplitudes of the received signals with different phases have features that can serve as no-activity features. Similarly, when a person performs an activity near the tags, the signature of the features takes its own value and carries information about the activity. Clearly, it is important to identify good features that allow for accurate classification. For example, it has been found that the backscatter channel phase, the backscatter amplitude, and the change in excitation amplitude between two multiphase probings have a high discriminatory power for classification [8].

Experimental results suggest that with signals provided by a BTTN, one can recognize human activities with an average error of about 6%. This was accomplished with 8 different activities and 9 individuals. Interestingly, this level of performance is similar to that achieved by systems that use powered, active radios. The classification results were obtained by convolutional neural networks (for details, see [8]). While the ability to recognize activities in such a fashion is already available in various other radio technologies, BTTN provides a unique approach due to its entirely batteryless operation, possibility of ubiquity and hence ability to measure a large number of tag-to-tag channels for very fine grain measurements.

5.3 From smart cities to biomedicine

Since the introduction of the RFID technology in the supply chain area about 15 years ago, the technical literature has provided numerous articles that promote the concept of smart homes and smart cities. One can easily imagine a smart home with BTTNs, where the tags equipped with sensors pepper the space of the home and where many of them are placed on various types of objects. The location and tracking of such objects will then readily be enabled by the functionality described in Section 5.1. Applications in smart cities include use on structures like buildings, streets, bridges, and parking spaces. The tags (with attached sensors) can be tasked to monitor air pollution, traffic, and availability of parking spaces. If the tags' density is high, these operations can be completed with high spatial resolution. The BTTNs can also be applied to perform the structural monitoring of buildings and bridges where abnormalities can be detected without actual sensing devices and instead based on the changes in the backscattered signals due to the developed abnormalities, (e.g., cracks can be found by detecting changes in distances between two tags before and after the appearance of a crack). BTTNs will also find a number of applications in medicine, environmental sensing, precision farming, and manufacturing. For more details and other applications, see a recent review on ambient backscatter communication [12].

6. FUTURE RESEARCH DIRECTIONS

BTTNs offer a unique system to enable ubiquitous massively-deployed IoT. Being batteryless and small form factor, they can easily blend with everyday objects and thus almost everything can become part of the network. Current research has successfully prototyped and evaluated single BBTN links, explored their ability to characterize the intervening wireless channel (RF sensing) with applications to localization, tracking and activity recognition. Current research has also produced theoretical studies on large-scale network routing issues. But much still needs to be done to make BTTNs practical and their applications realizable. One key issue is effective power harvesting and associated power management, so that the optimal power is allocated to activities such as communication, sensing, and computation at all times. This may limit the computation needed for routing and other application level signal processing due to a limited power budget. These are trade-offs that need to be explored in very dense deployments, e.g., tags

in a neighborhood can time multiplex various activities to achieve a network-level power management. Also, effective distributed computing techniques are needed to address processing limitation issues.

These are quite exciting times of the IoT era. The invention of tags that can form BTTNs presents a springboard for launching the concept of IoT to new heights. The possibility for connecting every tagged object in a network that turns into a part of the IoT has finally become a reality.

BTTNs offer a range of research challenges. For example, one of them is in energy harvesting and involves the design of an energy harvester with high energy efficiency over a wide input power range. Scalability and routing discovery in the presence of unidirectional links are not trivial routing tasks. Further, in security, future research should focus on balancing the security needs of BTTNs with limited resource use on the tags. Signal processing on the tags is also difficult due to the limited computing power of the tags. Future work will reveal efficient ways of processing increasing amounts of data by the tags and in a cooperative manner.

The prospects of BTTNs are quite promising, creating a driving force for their further development. The number of BTTN applications is simply staggering. In the near future, the hardware and computational aspects of the tags will continue to improve. Novel machine learning methods, possibly designed for the tags only, will be developed, and more novel techniques in the networking of the tags will be invented. With all the anticipated progress, one may argue, the BTTNs will become the true backbone of the IoT and will bring to fruition many of the benefits that have been envisioned by the IoT paradigm.

7. CONCLUSIONS

In this article, we presented an overview of backscatter-based communication among batteryless tags, the hardware of the tags, the scaling from a single link to a full network, and the signal processing that is carried out by the tags. Further, we listed a number of important applications with networks composed of such tags. We also discussed challenges that the tags and the network present, including challenges in energy harvesting, computing, networking, security, and distributed signal processing and decision making.

ACKNOWLEDGEMENT

The authors are thankful for the support of NSF under Awards CNS-1901182, CNS-1763843 and CNS-1763627.

REFERENCES

[1] J. Blanckenstein, J. Klaue, and H. Karl, "A survey of low-power transceivers and their applications," *IEEE Circuits and Systems Magazine*, vol. 15, no. 3, pp. 6–17, 2015.

[2] N. E. Roberts, K. Craig, A. Shrivastava, S. N. Wooters, Y. Shakhsheer, B. H. Calhoun, and D. D. Wentzloff, "A 236nw -56.5 dbm-sensitivity Bluetooth low-energy wakeup receiver with energy harvesting in 65nm CMOS," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2016, pp. 450–451.

[3] J. F. Ensworth and M. S. Reynolds, "BLE-Backscatter: Ultralow-Power IoT Nodes Compatible With Bluetooth 4.0 Low Energy (BLE) Smartphones and Tablets," *IEEE Trans. on Microwave Theory and Techniques*, 2017.

[4] P. Zhang, D. Bharadia, K. Joshi, and S. Katti, "Hitchhike: Practical backscatter using commodity WiFi," in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, ser. SenSys '16, New York, NY, USA, 2016, p. 259–271.

[5] D. Bharadia, K. R. Joshi, M. Kotaru, and S. Katti, "Backfi: High throughput WiFi backscatter," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 283–296, 2015.

[6] V. Iyer, V. Talla, B. Kellogg, S. Gollakota, and J. Smith, "Inter-technology backscatter: Towards internet connectivity for implanted devices," in *Proceedings of the 2016 ACM SIGCOMM Conference*, 2016, pp. 356–369.

[7] P. Zhang, C. Josephson, D. Bharadia, and S. Katti, "Freerider: Backscatter communication using commodity radios," in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, 2017, pp. 389–401.

[8] J. Ryoo, Y. Karimi, A. Athalye, M. Stanaćević, S. R. Das, and P. M. Djurić, "BARNET: Towards activity recognition using passive backscattering tag-to-tag network," in *16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2018, pp. 414–427.

[9] P. V. Nikitin, S. Ramamurthy, R. Martinez, and K. Rao, "Passive tag-to-tag communication," in *2012 IEEE International Conference on RFID (RFID)*. IEEE, 2012, pp. 177–184.

[10] A. Y. Majid, M. Jansen, G. O. Delgado, K. S. Ytidtnm, and P. Pawetczak, "Multi-hop backscatter tag-to-tag networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 721–729.

[11] V. Liu, A. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. R. Smith, "Ambient backscatter: wireless communication out of thin air," in *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4. ACM, 2013, pp. 39–50.

[12] N. Van Huynh, D. T. Hoang, X. Lu, D. Niyato, P. Wang, and D. I. Kim, "Ambient backscatter communications: A contemporary survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2889–2922, 2018.

[13] Y. Karimi, A. Athalye, S. Das, P. M. Djurić, and M. Stanaćević, "Design of backscatter-based tag-to-tag system," in *IEEE International Conference on RFID (RFID)*, 2017.

[14] C. Xu, L. Yang, and P. Zhang, "Practical backscatter communication systems for battery-free internet of things: A tutorial and survey of recent research," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 16–27, 2018.

[15] Z. J. Haas and Z. Zheng, "Waveform design for RF power transfer," *U.S. Patent Application, 15-959,917*, 2018.

[16] J. Ryoo, J. Jian, A. Athalye, S. R. Das, and M. Stanaćević, "Design and evaluation of "BTTN": A backscattering tag-to-tag network," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2844–2855, 2018.

[17] Z. Shen, A. Athalye, and P. M. Djurić, "Phase cancellation in backscatter-based tag-to-tag communication systems," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 959–970, 2016.

[18] A. N. Parks, A. Liu, S. Gollakota, and J. R. Smith, "Turbocharging ambient backscatter communication," in *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4. ACM, 2014, pp. 619–630.

[19] J. Qian, A. N. Parks, J. R. Smith, F. Gao, and S. Jin, "Iot communications with m-psk modulated ambient backscatter: Algorithm, analysis, and implementation," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 844–855, 2018.

[20] C. Liu and Z. J. Haas, "Multi-hop routing protocols for rfid systems with tag-to-tag communication," in *Proceedings of the 36th IEEE Military Communications Conference*. IEEE, 2017, pp. 563–568.

[21] C. Liu, Z. J. Haas, and Z. Tian, "On the design of multi-hop tag-to-tag routing protocol for large-scale networks of passive tags," in *IEEE Open Journal of the Communications Society*. IEEE, 2020, pp. 1035–1055.

[22] P. Sinha, S. Krishnamurthy, and S. Dao, "Scalable unidirectional routing with zone routing protocol (ZRP) extensions for mobile ad-hoc networks." IEEE, Sept. 23-28, 2000, pp. 1329–1339.

[23] A. Ahmad, Y. Huang, X. Sha, A. Athalye, M. Stanaćević, S. R. Das, and P. M. Djurić, "On measuring Doppler shifts between tags in a backscattering tag-to-tag network with applications in tracking," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 9055–9059.

AUTHORS

Milutin Stanaćević received the B.S. degree in Electrical Engineering from the University of Belgrade, Serbia and Ph.D. degree in Electrical and Computer Engineering from Johns Hopkins University, Baltimore, MD. He is currently an Associate Professor in the Department of Electrical and Computer Engineering at Stony Brook University, Stony Brook, NY. His research interests include mixed-signal and RF circuit and system design. Dr. Stanaćević is a recipient of the National Science Foundation CAREER award.

Akshay Athalye is a Co-Founder of Scandent LLC, New York, NY, USA, where he currently serves as a Chief Technology Officer. He is also an Adjunct Professor with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY, USA. He has been involved in RFID, Backscatter Communications and related research more than a decade. He has received the IEEE Region 1 Technical Excellence Award.

Zygmunt J. Haas has been with AT&T Bell Laboratories from 1988 until 1995, and since 1995 with Cornell University, Ithaca, NY, USA. Since 2013, he also holds the title of a Professor and Distinguished Chair at the University of Texas at Dallas, Richardson, TX, USA. He is a recipient of numerous awards and distinctions, including IEEE Fellow, IET Fellow, EAI Fellow, and Best Paper awards, the 2012 IEEE ComSoc WTC Recognition Award, and the 2016 IEEE ComSoc AHSN Recognition Award.

Samir R. Das is Professor in the Department of Computer Science at Stony Brook University. His research interests are in wireless networking and mobile computing, focusing on protocols, systems and performance evaluation. He co-chaired the technical program committees of premier mobile networking conferences, including ACM MobiHoc and ACM MobiCom. He served on the editorial boards of IEEE/ACM Transactions on Networking and IEEE Transactions on Mobile Computing.

Petar M. Djurić received the B.S. and M.S. degrees in electrical engineering from the University of Belgrade, Belgrade, Yugoslavia, respectively, and the Ph.D. degree in electrical engineering from the University of Rhode Island, Kingston, RI, USA. He is a SUNY Distinguished Professor and currently a Chair of the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY, USA. His research has been in the area of signal and information processing with primary interests in the theory of Monte Carlo-based methods; Bayesian machine learning; signal modeling, detection, and estimation; signal and information processing over networks; RFID and the IoT. Recently, his research has been applied to problems related to machine learning methods for intrapartum fetal monitoring and brain signals. He has been invited to lecture at many universities in the United States and overseas. Prof. Djurić was a recipient of the IEEE Signal Processing Magazine Best Paper Award in 2007 and the EURASIP Technical Achievement Award in 2012. In 2008, he was the Chair of Excellence of Universidad Carlos III de Madrid-Banco de Santander. From 2008 to 2009, he was a Distinguished Lecturer of the IEEE Signal Processing Society. He has been on numerous committees of the IEEE Signal Processing Society and of many professional conferences and workshops. He was Editor-in-Chief of the IEEE Transactions on Signal and Information Processing over Networks. Prof. Djurić is a Fellow of IEEE and EURASIP.

NON-COHERENT MASSIVE MIMO-OFDM FOR COMMUNICATIONS IN HIGH MOBILITY SCENARIOS

Kun Chen-Hu[1], Yong Liu[2], Ana Garcia Armada[1]
[1]Deparment of Signal Theory and Communications of Universidad Carlos III of Madrid (Spain). E-mails: {kchen, agarcia}@tsc.uc3m.es, [2]Shanghai Research Center, Huawei Technologies (China). E-mail: liu.liuyong@huawei.com.

Abstract – *Under scenarios of high mobility, the traditional coherent demodulation schemes (CDS) have a limited performance, due to the fact that reference signals cannot effectively track the variations of the channel with an affordable overhead. As an alternative solution, non-coherent demodulation schemes (NCDS) based on differential modulation have been proposed. Even in the absence of reference signals, they are capable of outperforming the CDS with a reduced complexity. The literature on NCDS laid the theoretical foundations for simplified channel and signal models, often single-carrier and spatially uncorrelated flat-fading channels. In this work, the most recent results assuming orthogonal frequency division multiplexing (OFDM) signaling and realistic channel models are explained, and the impact of some hardware impairments such as the phase noise (PN) and the non-linear high power amplifier (HPA) are also considered. Moreover, new potential research lines are also highlighted.*

Keywords – 5G, channel estimation, MIMO, non-coherent, OFDM.

1. INTRODUCTION

The Fifth Generation of mobile communications (5G) [1] is the global standard for a unified wireless air interface, which is capable of providing a great flexibility for a multitude of use cases. The three main requirements of those services are enhanced mobile broadband (eMBB), massive machine type communications (mMTC) and ultra reliable low-latency communications (URLLC). Therefore, the peak data-rate is not the only feature to be improved, but also an enormous number of connected devices and the latency-sensitive services are taken into account. Also there is an increasing interest in providing an adequate service in high mobility scenarios [2] - [4].

Orthogonal frequency division multiplexing (OFDM) with multiple-input multiple-output (MIMO) [5] have been recently set as the radio techniques for the physical layer in 5G [1]. New frequency bands are proposed to be exploited to obtain more available bandwidth, such as 3.5 GHz and millimeter-waves (mm-Wave) [6], and thus, the existing services can be improved and new ones can be implemented. The integration of massive MIMO is a must, not only to improve the average capacity of the link, but also for the implementation of beam-steering and beamforming to mitigate propagation losses in these new higher bands. Furthermore, the complexity of the signal processing techniques need to be bounded to reduce the cost of the devices and the delay of the required operations. As an alternative to classical coherent demodulation schemes (CDS), non-coherent demodulation schemes (NCDS) [7–9] have been proposed recently to be combined with massive MIMO systems [10] - [19]. They are capable of avoiding the overhead produced by the reference signals due to the fact that the transmitted symbols can be recovered without the knowledge of channel state information (CSI). This overhead can be excessively high for very fast time-varying channels. In such cases, a significant number of reference signals is required for the continuous tracking of the channel estimation [20].

The works in the literature have provided the theoretical foundations to understand NCDS and point to some cases when they can outperform the traditional CDS, in particular in scenarios with high mobility [20]. Also, recent works show some combinations of NCDS with MIMO-OFDM for the uplink (UL) and downlink (DL). In the present work, convinced that NCDS is an idea whose time has come, we discuss the implementation of the NCDS in practical MIMO-OFDM communication systems, assuming some realistic channels characterized by high mobility. We provide the details of how to integrate the differential modulation [21] in the two-dimensional resource grid (time and frequency) provided by the OFDM. Additionally, we also show the performance of this combination under the effects of the phase noise (PN) [22], [23] or high power amplifier (HPA) [24], for both UL and DL, and its benefits as compared to the traditional CDS. Finally, a discussion related to challenges and opportunities is provided together with some concluding remarks, in order to stimulate the research on this promising topic.

The remainder of the paper is organized as follows. Section 2 introduces the main differences between CDS and NCDS, especially for high mobility scenarios. Section 3 and Section 4 provide the details of how to integrate the NCDS with MIMO-OFM for the UL and DL, respectively. Section 5 presents several numerical results to evaluate the proposed scheme under some realistic channel models, providing an assessment of the achieved system performance. Finally, in Section 6, the conclusions follow.

2. BENEFITS AND WEAKNESSES OF THE COHERENT AND NON-COHERENT DEMODULATION SCHEMES

2.1 Coherent demodulation schemes (CDS)

Well-known coherent detection requires a replica of the carrier at the receiver, with frequency and phase synchronized, with the transmitted one, and an estimation of the channel attenuation and phase. Then, the received signal and a replica of the received version of all possible transmitted signals can be cross-correlated to make a decision. CDS are widely used by many communication systems. In particular, they are used in 5G [1], where the advantages of MIMO-OFDM are exploited, providing a high throughput through the use of the well-known M-ary quadrature amplitude modulation (QAM). With this modulation format, the information is transported in both the amplitude and phase of the carrier, making an efficient use of the transmission channel. However, these benefits come at the expense of transmitting some reference signals in order to obtain accurate enough CSI, so that the effects produced by the propagation channel to the received symbols can be equalized before a decision. When the channel is frequency-selective, OFDM facilitates the implementation of CDS due to the fact that each subcarrier can be considered as having an independent flat-fading channel, reducing the complexity of the equalization.

The need to obtain accurate enough CSI is accepted in most communication systems, in particular when the channel impulse response remains quasi-static for a certain period of time and the number of antennas is not very large. Under these conditions, a reduced amount of reference signals are used in order to track the channel variations in time, frequency and space dimensions. On the other hand, if we would like to provide communications in new challenging environments, such as high speed trains, autonomous vehicles, etc. these are mainly characterized by a significant Doppler spread due to the high mobility. In these situations, the traditional CDS requires an enormous amount of reference signals in order to continuously and accurately track the variations of the channel, reducing considerably the overall efficiency of the system, as pointed out in [10], [11], [20]. Otherwise, if the CSI is not properly estimated, the performance of the CDS is also seriously compromised. Moreover, when massive MIMO is considered, the procedures of channel estimation and the computation of the pre/post-coding matrices may increase the complexity of the system. For example, the channel inversion of large dimension matrices for each subcarrier may be prohibitive for some real-time applications when a zero-forcing (ZF) criterion is chosen.

2.2 Non-Coherent demodulation schemes (NCDS)

Looking back in history, NCDS are older than CDS. In [7], a comparison is made of the output spectra comprising signal and low-frequency noise when a sinusoidal signal plus noise is applied to several types of detector. It is shown that a considerable gain may be obtained by using the (new at that time) coherent detector as compared to the non-coherent square-law detector when the input signal to noise ratio (SNR) is low. In [8], a complete theory of detection is presented for threshold reception, which requires either a suitably weighted cross-correlation of the received data with the a priori known signal (CDS), or a suitably weighted autocorrelation of the received data with itself (NCDS). The Kineplex system developed by Collins Radio Company introduced the technique of differential phase shift keying (DPSK), as described in [9].

Today, NCDS have been re-proposed as an alternative to the traditional CDS due to the fact that they are able to recover the transmitted symbols without any CSI, that is, knowledge of the amplitude and phase of the carrier is not required. Hence, reference signals are no longer needed, reducing the undesirable signalling overhead. This effect is more relevant for high mobility and/or very frequency-selective scenarios. Additionally, the complexity of the transceivers is significantly reduced. Typical approaches involve the detection of the signal energy of phase differences. Despite its simplicity, non-coherent detection usually implies a 3 dB loss in SNR as compared to CDS. For this reason, it has just been used in a few communication systems where low complexity was a primary requirement. Examples of application are Bluetooth [25], with a non-coherent frequency shift keying (FSK) receiver or Zigbee [26], using DPSK.

Differential modulation is one of the most frequently used techniques for NCDS [12] - [19]. In this case the information is encoded by computing the phase difference between the current complex data symbol and the previously transmitted symbol. At the receiver, a simple differential decoder is required, detecting the phase difference between two contiguous symbols. To apply this technique, the constellation is constrained to have a constant modulus, such as DPSK, and a single reference symbol is needed at the beginning of each stream to have an initial phase reference. This means a negligible overhead to the system. It is also required that the channel response of every two contiguous differential symbols should be very similar, otherwise the differential decoder is not able to successfully recover the transmitted information data. This condition is usually met, even in fast varying channels.

Recently, the combination of NCDS with massive MIMO has been proposed in order to improve its performance leveraging the high number of antennas. In the context of UL, [10], [11] showed that asymptotically NCDS can achieve the same performance as CDS. Nevertheless, the proposed technique that illustrated this idea required a very large number of antennas to get an acceptable performance. Then, [12] - [16] proposed the use of DPSK together with an averaging process performed at the base station (BS) over the spatial domain after non-coherent demodulation, in order to mitigate the effects of the chan-

nel and noise. Moreover, they proposed the idea of multiplexing the data of each user equipment (UE) in the constellation domain based on a joint-symbol, which is a superposition of the symbols sent by several UEs. In the case of DL, the combination of NCDS with MIMO has been until now based on block codes [27] - [31]. However, their application requires that the channel response remains quasi-static during the transmission of a block code, and they also need a high SNR in order to provide an acceptable performance. Moreover, they have the problem that they are not scalable and when the number of antennas at the BS is very large, the design of these block codes becomes unaffordable. Typically, only two and four transmit antennas are taken into account [27] - [31]. More recently, the combination of beamforming and NCDS has been proposed in order to exploit the high number of antennas through compensating the path loss and enhancing the quality of the link, and spatially multiplexing the different UEs [18], [19]. In these cases a certain channel knowledge is needed to point the beam towards the UE through the beam-management procedure, and the signal is processed non-coherently in each beam afterwards. Even though the overhead is not completely eliminated, the savings are considerable.

3. NCDS WITH MASSIVE MIMO FOR THE UPLINK

We describe in this section how to integrate the NCDS based on [12] - [17] in a realistic communication system for the particular scenario of UL. We consider one BS equipped with V antennas, which is simultaneously serving U UEs. These UEs are constrained to have a reduced number of antennas, typically single-antenna devices. Let us assume that the U UEs are simultaneously transmitting N OFDM symbols. The OFDM signal has K subcarriers, and the length of the cyclic prefix (CP) is long enough to absorb the effects of the multi-path channel. At the receiver side, after removing the CP and performing a fast-Fourier transform (FFT) to each block at each antenna of the BS, we can process each subcarrier as one of a set of K independent subchannels.

3.1 Integration of differential encoding in OFDM for high mobility scenarios

Similar to CDS, the NCDS can be also implemented in an OFDM system [17], suitable for dealing with a doubly-dispersive channel. The stream of differential symbols produced by the differential encoding can be mapped in the two-dimensional resource grid provided by the OFDM (time and frequency). According to [17], the way this mapping is performed will significantly impact on the overall system, especially for high mobility scenarios.

Typically, NCDS based on differential modulation is performed using the time domain scheme. This scheme is represented in Fig. 1, where the red arrows indicate the direction in which differential modulation and demodulation is performed, in this case between resources that belong to the same frequency and contiguous symbols in the time domain. The differential encoding can be described as

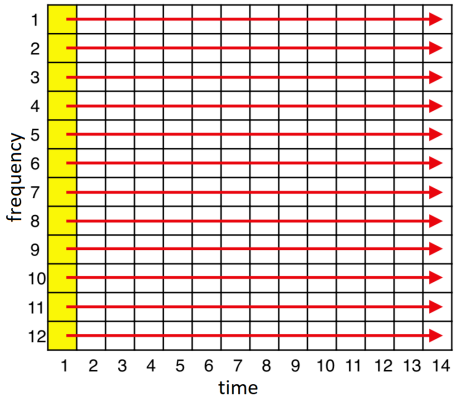


Fig. 1 – Time domain scheme in the OFDM resource grid when $K = 12$ and $N = 14$. The yellow box represents a reference symbol required by the differential modulation.

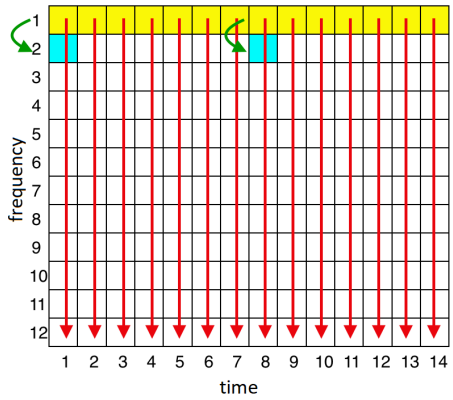


Fig. 2 – Frequency domain scheme in the OFDM resource grid when $K = 12$, $N = 14$ and $\mathcal{I}_N = \{1, 8\}$. The yellow and blue boxes denote the reference symbols required by the differential modulation and phase difference estimation, respectively.

$$x^u_{k,n} = \begin{cases} r^u_{k,n}, & n = 1 \\ x^u_{k,n-1} s^u_{k,n-1}, & 2 \le n \le N \end{cases}, \qquad (1)$$

$$1 \le u \le U, \quad 1 \le k \le K,$$

where $r^u_{k,1}$ is the reference symbol transmitted at the k-th subcarrier of the first OFDM symbol by the u-th UE, $s^u_{k,n}$ and $x^u_{k,n}$ are the complex data and differential symbol, respectively, transmitted at the k-th subcarrier and n-th OFDM symbol by the u-th UE. The data symbol $s^u_{k,n}$ belongs to a PSK constellation due to the fact that the differential encoding can only transmit information in the phase component. However, this time-domain implementation has the drawback of an increased latency and memory consumption, since this mapping scheme requires waiting for the reception of two complete OFDM symbols in order to obtain $s^u_{k,n}$, due to the fact that it performs a differential decoding of two contiguous symbols in the time domain. Also, it cannot be exploited when the Doppler shift is very high, since any two consecutive OFDM symbols will not face a similar channel response.

Alternatively, the OFDM frame enables exploiting the frequency dimension, and hence, the differential modulation technique can be also implemented using the frequency domain scheme (see Fig. 2). According to [17], the differential symbols are mapped into contiguous frequency resources of the same OFDM symbol as

$$x^u_{k,n} = \begin{cases} r^u_{k,n}, & k = 1, \\ x^u_{k-1,n} p^u_{k,n}, & k = 2, \ n \in \mathcal{J}_N \\ x^u_{k-1,n} s^u_{k-1,n}, & \text{otherwise} \end{cases}, \qquad (2)$$

$$1 \leq u \leq U, \quad 1 \leq n \leq N,$$

where $r^u_{1,n}$ and $p^u_{2,n}$ are two reference symbols for different purposes, and the set $\mathcal{J}_N$ contains the indexes that correspond to those OFDM symbols which carry $p^u_{2,n}$. The first kind of reference symbol is required for the differential demodulation as explained before. The second one is required for the estimation of the phase difference between two subcarriers, consequence of the frequency-domain mapping; see [17] for more details. We can see that this scheme has a reduced latency and is robust against high Doppler shifts. Furthermore, it is reasonable to assume that the channel responses of any two contiguous subcarriers are similar due to the fact that the number of subcarriers is always designed to be much larger than the number of taps of the channel. However, these benefits come at the expense of an additional phase estimation and compensation procedure. This additional phase component is very small and consequently can be neglected for channels that are not very frequency-selective. On the other hand, this phase must be compensated for strong frequency-selective channels. However, when diversity is exploited, only an additional reference pilot is required for all OFDM symbols within the coherence time ($p^u_{2,n}$), which produces a negligible impact on overhead.

Both time and frequency domain schemes, presented in [17], may introduce a significant overhead, if the number of allocated resources is reduced ($K \downarrow$ and/or $N \downarrow$). For example, in scenarios of mMTC, the machine devices are designed to send short packets of just a few bytes. The adoption of any of the two presented schemes implies to send a significant amount of reference symbols. Hence, we propose a new mapping scheme named as mixed domain scheme (see Fig. 3). Firstly, we differentially encode the data symbols as

$$x^u_j = \begin{cases} r^u_j, & j = 1 \\ x^u_{j-1} p^u_j, & j = 2 \\ x^u_{j-1} s^u_{j-1}, & 3 \leq j \leq KN \end{cases}, \quad 1 \leq u \leq U \qquad (3)$$

where the j denotes the resource index. Then, the differential symbols x^u_j are allocated to the two-dimensional resource grid as

$$x^u_{k,n} = x^u_j \mid (k,n) = f(j), \quad 1 \leq j \leq KN, \qquad (4)$$

where $f(\bullet)$ is the resource mapping policy function. Fig. 3 shows a recommended example of a mapping policy

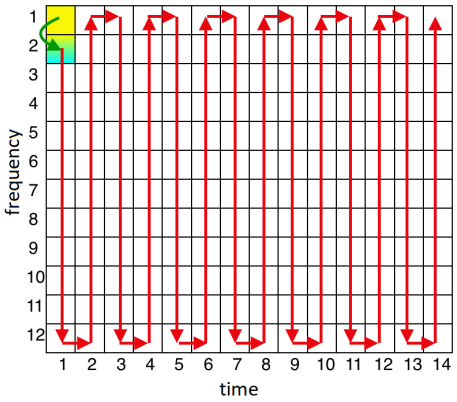


Fig. 3 – Mixed domain scheme in the OFDM resource grid when $K = 12$ and $N = 14$. The yellow and blue boxes denote the reference symbols required by the differential modulation and phase difference estimation, respectively.

function, where the dramatic reduction of reference signals can be observed. This policy mainly follows the frequency domain scheme, except for the edge subcarriers of the block, that follow a time domain scheme. This proposal cannot only significantly reduce the number of reference symbols, but it is also capable of taking all advantages of a frequency domain scheme. Moreover, in the case of time-varying channels, only those complex symbols placed at both edge subcarriers may suffer from an additional degradation, that can be easily mitigated by using some channel coding [16], [32] or spreading [33] techniques.

For the sake of conciseness and to ease the notation, the frequency domain scheme is the chosen one for the rest of the paper. Note that any of the presented techniques in the following sections can be straightforwardly adopted for both time and mixed domain schemes.

3.2 Multi-user multiplexing in the constellation domain

For a single-user case, the use of a constant modulus constellation, such as DPSK [12] - [15], is the only requirement for the non-coherent demodulation based on differential detection. However, when a multi-user scenario is considered, if we would like that all independent transmit sources are transmitting in the same time-frequency resource (to increase the spectral efficiency), the received signals from these independent sources are summed up and need to be conveniently separated [12]. Then, the choice of the constellation for each individual UE is crucial in order to produce joint-symbols that belong to a joint-constellation from which it is possible to unambiguously recover the transmitted data of all UEs.

At the BS, after removing the CP and performing the DFT, the received signal at the k-th subcarrier, n-th OFDM symbol and v-th antenna can be expressed as

$$y^v_{k,n} = \sum_{u=1}^{U} \sqrt{\beta_u} h^{u,v}_{k,n} x^u_{k,n} + w^v_{k,n}, \qquad (5)$$

$$1 \leq v \leq V, \quad 1 \leq k \leq K, \quad 1 \leq n \leq N,$$

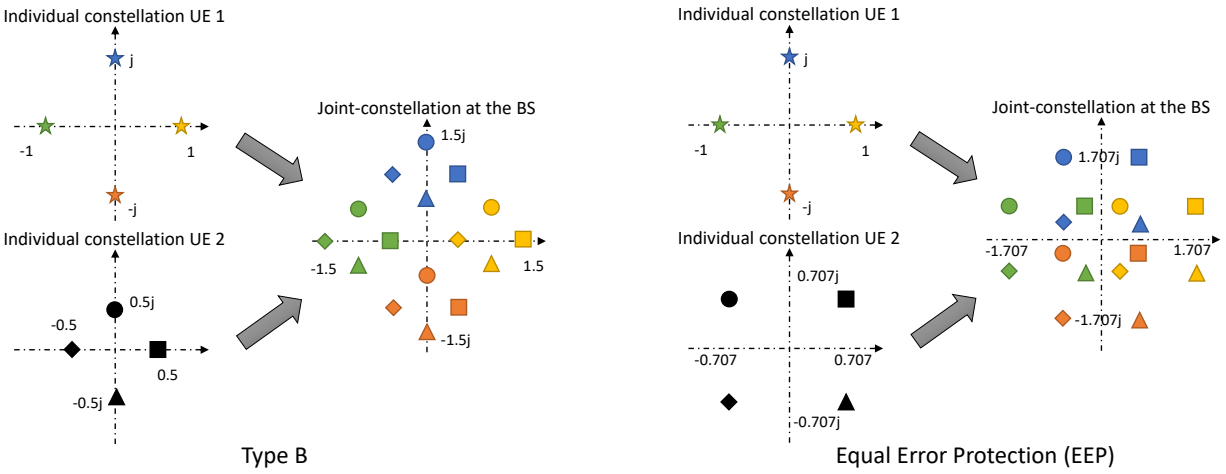


Fig. 4 – Example of two joint-constellations, for two UEs and four symbols each. The symbols of UE_1 are shown using different colours and the symbols of UE_2 are plotted using different markers.

where β_u is the average power of the signal of the u-th UE, $w^v_{k,n}$ denotes the additive white Gaussian noise (AWGN) at k-th subcarrier, n-th OFDM symbol and v-th antenna, distributed as $\mathcal{CN}(0, \sigma^2_w)$; and $h^{u,v}_{k,n}$ corresponds to the channel frequency response between the u-th UE and the v-th antenna at the k-th subcarrier and n-th OFDM symbol, distributed as $\mathcal{CN}(0, 1)$. For simplicity, we assume here that the channel response is spatially uncorrelated, while we will use more realistic channel models for performance evaluation. Besides, note that the difference in β_u among different UEs may be due to the constellation design or to different propagation path loss. In the latter case, an accurate power control must be implemented to compensate this difference.

According to [12], $y^v_{k,n}$ is fed to the differential decoder and averaged over the spatial dimension as

$$z_{k,n} = \frac{1}{V}\sum_{v=1}^{V}\left(y^v_{k-1,n}\right)^* y^v_{k,n}, \qquad (6)$$

$$2 \le k \le K, \quad 1 \le n \le N,$$

where $(\bullet)^*$ is the complex conjugate operation and $z_{k,n}$ denotes the received joint-symbol at the k-th subcarrier and n-th OFDM symbol. When the number of antennas is large enough and making use of the Law of Large Numbers, $z_{k,n}$ can be approximated as

$$z_{k,n} \xrightarrow{V\to\infty} s_{k,n} = \sum_{u=1}^{U}\beta_u s^u_{k,n} \qquad (7)$$

$$2 \le k \le K, \quad 1 \le n \le N,$$

where $s_{k,n}$ is the joint-symbol at the k-th subcarrier and n-th OFDM symbol. Note that the interference and noise terms are averaged out thanks to the large number of antennas at the BS, otherwise the performance may be degraded. More details are given in [12] - [17].

The performance of the overall multi-user systems depends on the constellation of the joint-symbol. This joint-symbol must be properly designed to enable the demodulation of the transmitted information by all the UEs. Consequently, the choice of the individual constellation is crucial to produce a robust joint-constellation against interference and noise effects. The most used constellations are the Type B [12] and equal error protection (EEP) [13]. The constellation of the u-th UE can be expressed as

$$\mathcal{M}^u_B = \left\{\sqrt{\beta_u}\exp\left(i\frac{2\pi}{M}m\right) \mid 0 \le m \le M-1\right\}, \qquad (8)$$

$$\mathcal{M}^u_E = \left\{\exp\left(i\left(\frac{2\pi}{M}m + \frac{\pi}{2U}u\right)\right) \mid 0 \le m \le M-1\right\}, \qquad (9)$$

for Type B and EEP, respectively, where M denotes the number of symbols in the constellation. Fig.4 shows an illustrative example of these two types of joint-constellations, designed for the particular case of UL with only two UEs, each of them using a 4-DPSK. In the first case, all UEs have the same 4-DPSK constellation and are distinguished with a different amplitude. This produces the joint-constellation also shown in the same figure, where we can see that all symbols are equally spaced providing a robustness against possible interference and noise terms. However, those UEs with a lower amplitude will obtain a worse performance for the same noise conditions as compared to the stronger ones. Indeed, the average distance of the symbols of UE $u = 1$ (shown in different colours) is much larger than the distance of UE $u = 2$ (plotted in different markers). In EEP, both UEs have the same amplitude, and then the same performance. Their constellations differ in a rotation of 45°. However, this option presents several symbols of the joint-constellation (those placed in the middle) that are too close to each other, degrading the performance.

The design of optimal individual constellations for multi-user NCDS that work well in realistic channel conditions is still a very challenging topic, due to the diverse effects of the channel impairments and interference and the difficulty to analyse them.

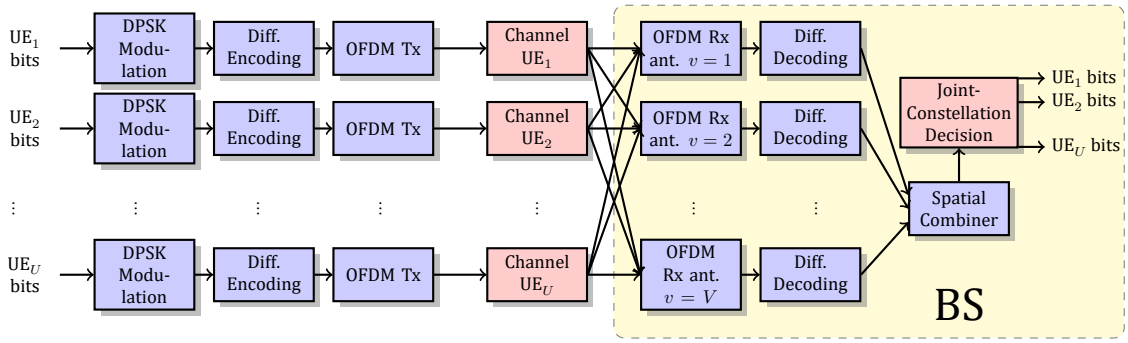


Fig. 5 – Block diagram for UL, where different UEs are multiplexed in the constellation domain.

The block diagram of a UL system addressing a more general multi-user case is shown in Fig. 5, from [17]. In this system, several UEs map their data bits into complex symbols that belong to a specific DPSK constellation. Then, the differentially encoded symbols are transmitted into an OFDM signal through a multi-path channel. The received signal at each antenna of the BS after the OFDM receiver corresponds to a superposition of different signals coming from each UE. Later, these differential symbols are non-coherently combined by using a differential decoder and averaged over the spatial domain in order to obtain the desired joint-symbol. Exploiting the spatial diversity in order to reduce the noise and multi-user interference (MUI) is crucial to obtaining a right decision from the joint-symbol [12]. Additionally, the constellation design can be combined with channel coding, where the soft information can either improve the performance or reduce the number of required antennas at the BS [16] [32].

4. NCDS WITH MASSIVE MIMO FOR THE DOWNLINK

Now turning to the DL, the massive number of antennas at the BS are used for transmission and we can only count on a few antennas at the UE receivers, just one per UE in many cases. The mapping schemes and multi-user constellations proposed for the UL are still valid, while we need a few more ingredients to make these schemes suitable for the DL. We explain in the following these key ingredients for a combination of NCDS with beamforming based on [19], where the BS is simultaneously transmitting the U data streams through its V antennas, while the UEs are receiving with their single-antenna device.

4.1 Beamforming

The exploitation of the diversity from the transmitter without the knowledge of the CSI is still a challenge. Due to the fact that techniques based on block codes [27] failed to exploit the large number of antennas at the transmitter, we propose the use of beamforming in order to take advantage of the massive number of antennas of the BS at the expense of using some (reduced) channel knowledge. Then, it is assumed that the angular positions of each UE are obtained through a beam management procedure to point the beams towards each UE. Once this is achieved, the data information is sent over a non-coherently processed link. Reference [18] proposed a similar idea assuming, however, an ideal case where the MUI is completely mitigated by the beamforming. Meanwhile, the combination of NCDS with a practical beamforming technique is proposed in [19], taking into account the residual MUI.

In [19] the beam-management procedure defined in 5G [1] is suggested to be performed as a first step. This procedure is responsible for accurately determining the angle of the spatial clusters of the propagation channel contributing to the signal of each UE, by transmitting some reference signals. These reference signals are specifically the synchronization signals (SS) and channel state information-reference signals (CSI-RS). The former are used when a UE would like to enter the system for the first time, while the latter are exploited for updating the angular position of an existing UE in the system.

Then, the BS transmits one or several differential data streams to each UE by using beamforming [34] as

$$x_{k,n}^{u,v} = b_{k,n}^{u,v} x_{k,n}^{u}, \quad 1 \le v \le V, \qquad (10)$$
$$1 \le u \le U, \quad 1 \le k \le K, \quad 1 \le n \le N,$$

where $b_{k,n}^{u,v}$ is the precoding coefficient for the u-th UE and v-th antenna of the BS, placed at the k-th subcarrier and n-th OFDM symbol. This precoding coefficient is obtained according to the estimated angular positions of each UE, and thus, it is in charge of focusing the energy in the obtained specific directions. In this way, the path loss is compensated and the MUI that results from spatially multiplexing the UEs in different beams is avoided. Similarly, beamforming can be used in the UL for the BS to receive the signal coming from these spatial directions.

Depending on the angular position of different UEs, it may be difficult to completely remove the MUI by exploiting the beamforming. Therefore, the overall performance critically depends on the scheduler which is capable of properly selecting those UEs to be simultaneously served in the same time and frequency resources and minimizing the negative impact of the mentioned MUI. Even though we are making use of some reference signals to perform

the beamforming, any other additional overhead used in the CDS, such as the demodulation-reference signals (DM-RS), is avoided, increasing the spectral efficiency.

4.2 Frequency diversity

Due to the usually limited number of antennas at the UE, averaging in any dimension other than space (e.g. in time or frequency) is proposed in [19] to provide an additional source of diversity. This diversity is needed in order to obtain the required SINR gain for a good performance of the NCDS [12]. It is particularly needed if we want to multiplex several UEs in the constellation domain or enable services that are critical in terms of performance. The use of the frequency dimension is described in [17], with the advantage that each OFDM symbol can be independently processed. The proposed scheme can be easily extended to averaging either in time (processing several consecutive OFDM symbols) or space (increasing the number of receive antennas of the UE, when possible).

The way to leverage frequency diversity consists in transmitting the same differential complex symbol in several frequency resources. At the transmitter, after performing the differential encoding, the Q differential symbols are repeated as

$$x^u_{k,n} = x^u_{q,n} \mid q = \text{mod}\,(k-1, Q) + 1, \quad K = Q \times F, \tag{11}$$
$$1 \le u \le U, \quad 1 \le k \le K, \quad 1 \le n \le N,$$

where F is the frequency repetition/averaging factor.

At the receiver, analogously to (*(6)*), the frequency diversity is exploited in the non-coherent detection, where the received data in the subcarriers that carry the same transmitted data are averaged as

$$z_{q,n} = \frac{1}{F} \sum_{f=0}^{F-1} \left(y^v_{q-1+fQ,n}\right)^* y^v_{q+fQ,n}, \tag{12}$$
$$2 \le q \le Q, \quad 1 \le n \le N.$$

With this scheme there is a trade-off between overhead and robustness. We will see that for some particular scenarios with high mobility, even with the added overhead, this scheme will outperform the CDS in terms of throughput.

In Fig. 6 the block diagram of the system proposed in [19] is shown, combining the beamforming with the NCDS. At the receiver, assuming single-antenna devices, only frequency diversity is exploited in order to reduce the noise and MUI terms.

5. PERFORMANCE DISCUSSION

In this section we illustrate the performance of the combination of NCDS with a large number of antennas by discussing some numerical results. A comparison with some CDS counterparts is also provided.

5.1 Simulation parameters

To show some illustrative results, the numerology of the OFDM signal is chosen according to 5G [1]. The carrier spacing is $\Delta f = 30$ KHz, which is the most frequent value in different bands [35]. The bandwidth of the system is $BW = 100$ MHz and the carrier frequency is $f_c = 3.5$ GHz. The BS is equipped with a uniform linear array (ULA) of $V = 128$ antenna elements, which are simultaneously serving two UEs ($U = 2$) in the whole bandwidth. Their angular separation corresponds to 72° and the path loss is not considered, since the power control is assumed to work perfectly. We adopt a geometric channel model, which corresponds to a spatially correlated channel, where the power delay profile corresponds to the Type B given in [35], the delay spread is $DS = 16$ ns and the angular spread is $AS = 5°$. We assume that there is a Doppler shift of $f_d = 1.6$ KHz which corresponds approximately to a speed of $v = 500$ km/h at the mentioned carrier frequency. We assume a perfect time-frequency synchronization and power control at the receiver. For the sake of space we do not provide any results for a higher carrier frequency. However, the chosen delay and angular spread can be also representative of the propagation at higher frequencies, and the same Doppler frequency would correspond to a smaller speed. Hence, the conclusions obtained with these numerical results, in particular those including beamforming (which would be mandatory to compensate the path loss), can be extrapolated to other higher frequency bands, such as mm-Wave [36]. The SNR is conventionally defined as the ratio of the received signal power over the noise power at each antenna of the receiver.

For a baseline CDS system to compare the performance, we adopt the pilot configuration specified for the demodulation reference signals (DM-RS) in 5G [1]. In the time domain, due to the high mobility, we set four reference symbols for each slot, which corresponds to the maximum pilot density allowed by the standard. In the frequency domain, we assume the configuration Type-1, where each half of the subcarriers are allocated to each UE: the even subcarriers are used for the channel estimation of the UE_1 and the odd subcarriers are for UE_2. At the receiver, the channel estimation is firstly obtained at the pilot symbol resources by applying least squares (LS) [37]. Then, these estimates are interpolated to the entire resource grid by using spline interpolation.

Moreover, some hardware impairments are also considered, such as the effects of PN and the non-linear HPA. The effect of the PN is due to the instability of the local oscillators, that can only be reduced by making a more expensive one. The PN is typically modelled according to a classical Wienner random walk process given in [23], where the system performance is related to the phase noise increment variance σ^2_η over the sample period T, where $\sigma^2_\eta = 2\pi B_\eta T$, with B_η equal to the 3-dB bandwidth of the Lorentzian power density carrier spectrum. The negative effect of the PN not only will degrade the received

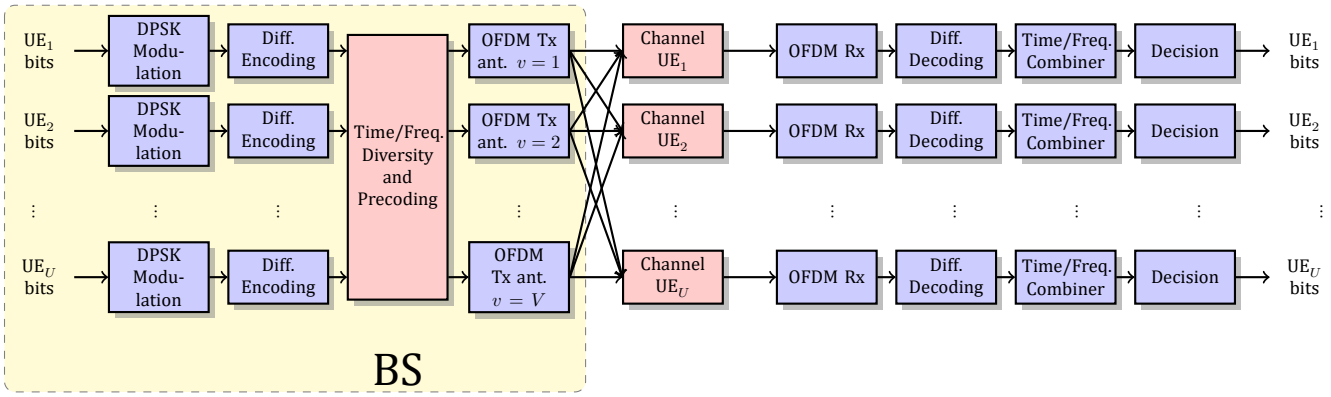


Fig. 6 – Block diagram for DL, where the BS uses a beamforming technique and all UEs are single-antenna devices.

symbols, but it will also add a common phase error [23]. The variance of the PN corresponds to $\sigma_\eta^2 = 10^{-5}$ rad^2. On other hand, the realistic transfer function of the HPA is not a linear function for all possible input values. This implies that the output might be saturated for those input values that are higher than the saturation point. This non-linear effect will not only degrade the quality of the received signal, but it will also enhance the out-of-band emissions. According to [24], we consider a solid state power amplifier whose output back-off (OBO) is OBO $= 8$ dB.

5.2 Numerical results

In Fig. 7, we show the SER comparison between CDS and NCDS for the UL. The constellations of the two UEs are QPSK for CDS and EEP for NCDS, both using two bits per symbol. The CDS performs a post-equalizer at the BS based on a ZF criterion. In the absence of PN and HPA, the NCDS outperforms the traditional CDS by almost two orders of magnitude of SER for moderate and high SNR scenarios. When hardware impairments are considered, the PN and HPA effects do not significantly degrade the performance of NCDS. On the other hand, the performance of CDS with and without the effect of the HPA is very poor, and it is even worse with the PN. The PN does not affect our proposed system due to the use of a differential modulation and the fact that the phase noise does not change between two contiguous subcarriers [23]. The negative effect of the HPA is negligible in both systems because the OBO is enough, in view of the robustness of the PSK signals, which are amplified separately at the transmitter of each UE.

In Fig. 8, we plot the SER comparison between CDS and NCDS for the DL. The same beamforming is considered for CDS and NCDS to spatially multiplex the two UEs. Additionally, a frequency averaging of factor $F = 16$ is performed in both schemes to leverage diversity and improve the overall performance, which would be otherwise compromised. Again, in the absence of PN and HPA, the NCDS outperforms the CDS by several orders of magnitude, showing that the frequency averaging is able to effectively reduce its SER, while it is not enough for CDS to work properly. When hardware impairments are considered, the performance of NCDS is degraded by both HPA and PN effects. In the same way as for the UL, we can see that NCDS is very robust to the PN effects due to the differential modulation. However, the non-linear HPA significantly degrades its performance. In this case the BS is simultaneously transmitting the signals of the two UEs and, consequently, the constant envelope characteristic of each of the PSK signals is lost when they are combined. It turns out that now the OBO is not enough and some of the signal peaks are clipped. This affects equally to both NCS and CDS. In Fig. 9, a comparison in terms of throughput is provided for the DL, whose expression is given by

$$T_r = \log_2(M)\,(1 - SER)\frac{BW}{F}. \qquad (13)$$

We can see that even with the overhead due to a very high frequency averaging factor ($F = 16$), the NCDS still outperforms the traditional CDS. This difference is even higher when either PN or HPA effects are considered. Therefore, the throughput reduction due to the overhead produced by the frequency diversity is negligible as compared to the small throughput achieved by the CDS due to a poor performance obtained even with a large overhead.

6. CONCLUSION

We have provided a detailed description of the novel combination of NCDS and multi-user MIMO-OFDM based on a differential modulation scheme. Both DL and UL scenarios are considered and the performance is analyzed for realistic channel conditions including the effect of the PN and HPA.

It is shown that for channels with high mobility, the NCDS outperforms the traditional CDS obtaining a better performance, even more noticeable when PN and non-linear HPA effects are taken into account. NCDS does not require any additional PN estimation and equalization since it is inherently robust to these effects. Moreover, it shows a similar degradation with the non-linear effects of the HPA to that suffered by CDS, since they share the sensitivity of OFDM to these effects.

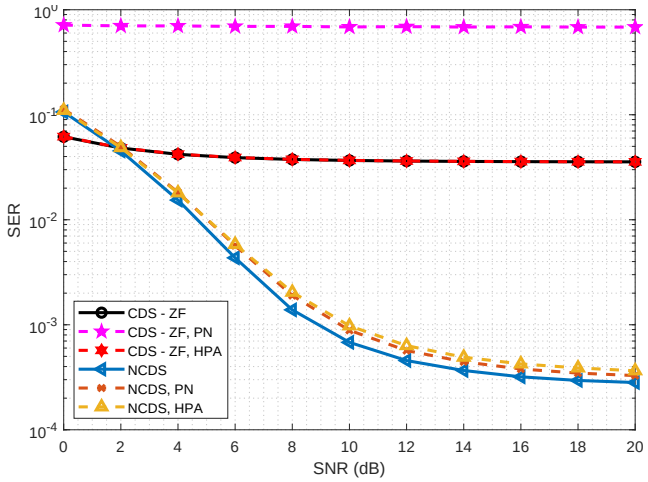


Fig. 7 – SER comparison for UL.

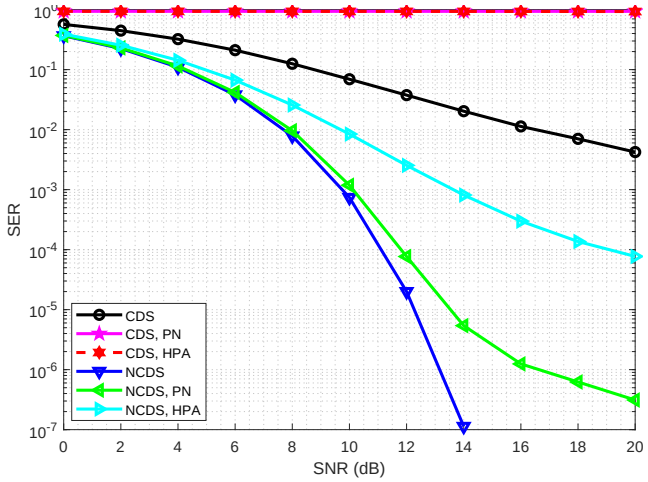


Fig. 8 – SER comparison for DL.

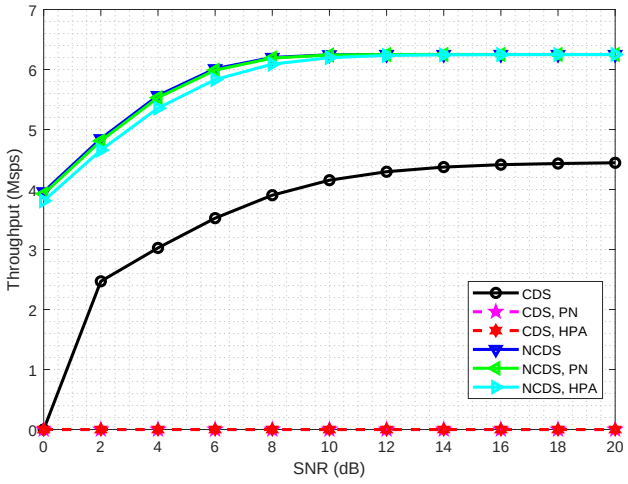


Fig. 9 – Throughput comparison for DL.

The NCDS combined with multi-user MIMO-OFDM is shown to be a feasible and very attractive technique which substantially improves the performance of the coherent systems, especially in challenging scenarios such as systems with realistic propagation channels and high mobility. However, there is still need for new research and ideas to improve the NCDS. In particular, the multiplexing of UEs in the constellation domain is not efficient, so an excessively high number of antennas is currently needed to multiplex more than two UEs [12]. There is a need to find new constellation designs that overcome this limitation. A possible way to obtain these optimal designs is to use some artificial intelligence techniques in order to automatically deal with the channel and multi-user interference effects. Not only the phases, but also the constant amplitudes of the constellations are possible values to select, giving a more complex search space, where artificial intelligence can help.

Additionally, we have seen that the performance is very sensitive to the spatial separation of the UEs that are multiplexed, either in constellation or space. Therefore, scheduling algorithms that take this into account and optimize a particular performance metric are also crucial.

The advantages of NCDS with respect to CDS vanish when the channel is quasi-static and with high SNR. Then, it is advisable to find hybrid schemes, such as [20] where the best of both paradigms is used according to the communication scenario and needs.

Finally, it is foreseen that in the future communications will be tightly integrated with sensing, which is one of the main objectives of the the Sixth Generation (6G) of mobile communications [38]. In these new systems, the efficient exploitation of CSI under a variety of scenarios will play an important role, and hence, we forecast that the exploitation of non-coherent techniques will be an interesting alternative, in order to increase the efficiency of the overall system. We hope that this review of the NCDS characteristics, feasible implementation and performance will stimulate new research and advances in this topic.

ACKNOWLEDGEMENT

This work has been funded by project TERESA-ADA (TEC2017-90093-C3-2-R) (MINECO/AEI/FEDER, UE). The authors would like to thank Ignasi Piqué-Muntané for his help in the elaboration of some figures.

REFERENCES

[1] NR; Physical channels and modulation (Release 16). Technical report, 3GPP, France, 2020.

[2] W. Guo, W. Zhang, P. Mu, F. Gao, and H. Lin. High-mobility wideband massive MIMO communications: Doppler compensation, analysis and scaling laws. *IEEE Transactions on Wireless Communications*, 18(6):3177–3191, June 2019.

[3] Y. Ge, W. Zhang, F. Gao, S. Zhang, and X. Ma. Beamforming network optimization for reducing channel time variation in high-mobility massive MIMO. *IEEE Transactions on Communications*, 67(10):6781–6795, Oct. 2019.

[4] M. Gao, B. Ai, Y. Niu, W. WU, P. Yang, F. Lyu, and X. Shen. Efficient hybrid beamforming with anti-

blockage design for high-speed railway communications. *IEEE Transactions on Vehicular Technology*, 2020.

[5] G. L. Stuber, J. R. Barry, S. W. McLaughlin, Ye Li, M. A. Ingram, and T. G. Pratt. Broadband MIMO-OFDM wireless communications. *Proceedings of the IEEE*, 92(2):271–294, Feb. 2004.

[6] B. Yang, Z. Yu, J. Lan, R. Zhang, J. Zhou, and W. Hong. Digital beamforming-based massive MIMO transceiver for 5G millimeter-wave communications. *IEEE Transactions on Microwave Theory and Techniques*, 66(7):3403–3418, July 2018.

[7] R. A. Smith. The relative advantages of coherent and incoherent detectors: a study of their output noise spectra under various conditions. *Proceedings of the IEE - Part III: Radio and Communication Engineering*, 98(55):401–406, Sep. 1951.

[8] D. Middleton. Statistical theory of signal detection. *Transactions of the IRE Professional Group on Information Theory*, 3(3):26–51, March 1954.

[9] M. L. Doelz, E. T. Heald, and D. L. Martin. Binary data transmission techniques for linear systems. *Proceedings of the IRE*, 45(5):656–661, May 1957.

[10] A. Manolakos, M. Chowdhury, and A. J. Goldsmith. CSI is not needed for optimal scaling in multiuser massive SIMO systems. In *2014 IEEE International Symposium on Information Theory*, pages 3117–3121, June 2014.

[11] M. Chowdhury, A. Manolakos, and A. Goldsmith. Scaling laws for noncoherent energy-based communications in the simo mac. *IEEE Transactions on Information Theory*, 62(4):1980–1992, April 2016.

[12] A. G. Armada and L. Hanzo. A non-coherent multi-user large scale SIMO system relaying on M-ary DPSK. In *2015 IEEE International Conference on Communications (ICC)*, pages 2517–2522, June 2015.

[13] V. M. Baeza, A. G. Armada, M. El-Hajjar, and L. Hanzo. Performance of a non-coherent massive SIMO M-DPSK system. In *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pages 1–5, Sep. 2017.

[14] V. M. Baeza, A. G. Armada, W. Zhang, M. El-Hajjar, and L. Hanzo. A non-coherent multiuser large-scale SIMO system relying on M-ary DPSK and BICM-ID. *IEEE Transactions on Vehicular Technology*, 67(2):1809–1814, Feb. 2018.

[15] V. M. Baeza and A. G. Armada. Non-coherent massive SIMO system based on M-DPSK for Rician channels. *IEEE Transactions on Vehicular Technology*, 68(3):2413–2426, March 2019.

[16] V. M. Baeza and A. G. Armada. Noncoherent massive MIMO. In *Wiley 5G Ref: The Essential 5G Reference Online*, chapter 10, pages 266–290. John Wiley & Sons, Ltd., 2019.

[17] K. Chen-Hu and A. G. Armada. Non-coherent multiuser massive MIMO-OFDM with differential modulation. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2019.

[18] S. Bucher, G. Yammine, R. F. H. Fischer, and C. Waldschmidt. A noncoherent massive MIMO system employing beamspace techniques. *IEEE Transactions on Vehicular Technology*, 68(11):11052–11063, Nov. 2019.

[19] K. Chen-Hu, Y. Liu, and A. G. Armada. Non-coherent massive MIMO-OFDM down-link based on differential modulation. *IEEE Transactions on Vehicular Technology*, 2020. (In press).

[20] M. J. Lopez-Morales, K. Chen-Hu, and A. Garcia-Armada. Differential data-aided channel estimation for up-link massive SIMO-OFDM. *IEEE Open Journal of the Communications Society*, 1:976–989, 2020.

[21] F. Adachi. Adaptive differential detection for M-ary DPSK. *IEE Proceedings - Communications*, 143(1):21–28, Feb. 1996.

[22] R. Corvaja and A. G. Armada. Phase noise degradation in massive MIMO downlink with zero-forcing and maximum ratio transmission precoding. *IEEE Transactions on Vehicular Technology*, 65(10):8052–8059, Oct. 2016.

[23] H. Ghozlan and G. Kramer. Models and information rates for wiener phase noise channels. *IEEE Transactions on Information Theory*, 63(4):2376–2393, April 2017.

[24] E. Costa, M. Midrio, and S. Pupolin. Impact of amplifier nonlinearities on ofdm transmission system performance. *IEEE Communications Letters*, 3(2):37–39, Feb. 1999.

[25] L. Lampe, R. Schober, and M. Jain. Noncoherent sequence detection receiver for Bluetooth systems. *IEEE Journal on Selected Areas in Communications*, 23(9):1718–1727, Sep. 2005.

[26] C. Wang, C. Huang, J. Huang, C. Chang, and C. Li. Zigbee 868/915-mhz modulator/demodulator for wireless personal area network. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16(7):936–939, July 2008.

[27] J. Cabrejas, S. Roger, D. Calabuig, Y. M. M. Fouad, R. H. Gohary, J. F. Monserrat, and H. Yanikomeroglu. Non-coherent open-loop MIMO communications over temporally-correlated channels. *IEEE Access*, 4:6161–6170, 2016.

[28] B. M. Hochwald and W. Sweldens. Differential unitary space-time modulation. *IEEE Transactions on Communications*, 48(12):2041–2052, Dec. 2000.

[29] V. Tarokh and H. Jafarkhani. A differential detection scheme for transmit diversity. *IEEE Journal on Selected Areas in Communications*, 18(7):1169–1174, July 2000.

[30] M. Beko, J. Xavier, and V. A. N. Barroso. Noncoherent communication in multiple-antenna systems: Receiver design and codebook construction. *IEEE Transactions on Signal Processing*, 55(12):5703–5715, Dec. 2007.

[31] R. H. Gohary and T. N. Davidson. Noncoherent MIMO communication: Grassmannian constellations and efficient detection. *IEEE Transactions on Information Theory*, 55(3):1176–1205, March 2009.

[32] F. Adachi. Reduced-state Viterbi differential detection using a recursively estimated phase reference for M-ary DPSK. *IEE Proceedings - Communications*, 142(4):263–270, Aug. 1995.

[33] N. Prasad, S. Wang, and X. Wang. Efficient receiver algorithms for DFT-spread OFDM systems. *IEEE Transactions on Wireless Communications*, 8(6):3216–3225, June 2009.

[34] J. Lota, S. Sun, T. S. Rappaport, and A. Demosthenous. 5G uniform linear arrays with beamforming and spatial multiplexing at 28, 37, 64, and 71 ghz for outdoor urban communication: A two-level approach. *IEEE Transactions on Vehicular Technology*, 66(11):9972–9985, Nov. 2017.

[35] Study on channel model for frequencies from 0.5 to 100 GHz (Release 16). Technical report, 3GPP, France, 2019.

[36] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock. *Millimeter wave wireless communications*. Prentice Hall, 2015.

[37] J. Lin. Least-squares channel estimation for mobile ofdm communication on time-varying frequency-selective fading channels. *IEEE Transactions on Vehicular Technology*, 57(6):3538–3550, Nov. 2008.

[38] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang. The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine*, 57(8):84–90, Aug. 2019.

AUTHORS

Kun Chen-Hu received his Ph.D. degree in Multimedia and Communications in 2019 from Universidad Carlos III de Madrid (Spain). Currently, he is a post-doctoral researcher in the same institution. He was awarded by UC3M in 2019 in recognition of his outstanding professional career after graduation. He visited Eurecom (France) and Vodafone Chair TU Dresden (Germany), both as guest researcher. He also participated in different research projects in collaboration with several top companies in the area of mobile communications. His research interests are related to signal processing techniques, such as waveforms design, non-coherent massive MIMO and channel estimation.

Yong Liu received a Ph.D in electronic engineering from the Department of Electric Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2012. He is now with the wireless network RAN research department in Huawei Technologies Co.,Ltd., Shanghai, China. His current research interests lie in the area of 5G and 5G+ MIMO communication and AI assisted wireless networks..

Ana García Armada (S'96-A'98-M'00-SM'08) received a Ph.D. degree in electrical engineering from the Polytechnical University of Madrid in February 1998. She is currently a Professor at University Carlos III of Madrid, Spain. She is leading the Communications Research Group at this university. She has been a visiting scholar at Stanford University, Bell Labs and University of Southampton. She has participated (and coordinated most of them) in more than 30 national and 10 international research projects, as well as 20 contracts with the industry, all of them related to wireless communications. She is the co-author of eight book chapters on wireless communications and signal processing. She has published around 150 papers in international journals and conference proceedings and she holds four patents. She has contributed to international standards organizations, such as ITU and ETSI, and is a member of the expert group of the European 5G PPP and member of the advisory committee 5JAC of the ESA as an expert appointed by Spain on 5G. She has served on the editorial boards of Physical Communication (2008-2017), IET Communications (2014-2017). She has been serving on the editorial board of IEEE Communications Letters since 2016 (Editor until Feb 2019, Senior Editor from Mar 2019, Exemplary Editor Award 2017 and 2018) and IEEE Transactions on Communications since 2019. She has served on the TPC of more than 40 conferences and she has been a member of the organizing committee of IEEE Globecom 2019, IEEE Vehicular Technology Conference (VTC) Fall 2018, Spring 2018 and 2019 and IEEE 5G Summit 2017, among others. She will be the General Chair of Globecom 2021. She was the Newsletter Editor of the IEEE ComSoc Signal Processing and Consumer Electronics Committee (2017-2018) and is now the Secretary of this committee (since 2019). She has been the Secretary of the IEEE ComSoc Women in Communications Engineering Standing Committee (2016-2017) and the Chair of this committee (2018-2019). She has received the Young Researchers Excellence Award, the Award to Outstanding achievement in research, teaching and management and the Award to Best Practices in Teaching, all from University Carlos III of Madrid, She was awarded the third place Bell Labs Prize 2014 for shaping the future of information and communications technology. She received the Outstanding service award from the IEEE ComSoc Signal Processing and Communications Electronics (SPCE) technical committee in 2019. Her main interests are multicarrier and multi-antenna techniques and signal processing applied to wireless communications.

MSICA: MULTI-SCALE SIGNAL DECOMPOSITION BASED ON INDEPENDENT COMPONENT ANALYSIS WITH APPLICATION TO DENOISING AND RELIABLE MULTI-CHANNEL TRANSMISSION

Abolfazl Hajisami and Dario Pompili
Dept. of Electrical and Computer Engineering, Rutgers University–New Brunswick, NJ, USA

Abstract – *Multi-scale decomposition is a signal description method in which the signal is decomposed into multiple scales, which has been shown to be a valuable method in information preservation. Much focus on multi-scale decomposition has been based on scale-space theory and wavelet transform. In this article, a new powerful method to perform multi-scale decomposition exploiting Independent Component Analysis (ICA), called MSICA, is proposed to translate an original signal into multiple statistically independent scales. It is proven that extracting the independent components of the even and odd samples of a digital signal results in the decomposition of the same into approximation and detail. It is also proven that the whitening procedure in ICA is equivalent to a filter bank structure. Performance results of MSICA in signal denoising are presented; also, the statistical independency of the approximation and detail is exploited to propose a novel signal-denoising strategy for multi-channel noisy transmissions aimed at improving communication reliability by exploiting channel diversity.*

Keywords – Channel Diversity, Independent Component Analysis, Multi-scale Decomposition, Wavelet Transform.

1. INTRODUCTION

Overview: Multi-scale decomposition is an invaluable tool in digital signal processing with applications such as those in [1, 2, 3, 4, 5], where an original signal is decomposed into a set of signals, each of which provides information about the original signal at a different *scale*. A major signal-processing task where multi-scale decomposition has been shown to be very useful is denoising, based on the intuition that information pertaining to the noise would be accurately characterized in certain scales that are separate from the scales of the signal. The main literature works in multi-scale decomposition have focused on scale-space decomposition [6, 7, 8, 9, 10], empirical mode decomposition [11, 12, 13], and wavelet transform [14, 15, 16, 17, 18].

In scale-space theory [19], a signal is decomposed into a single-parameter family of n signals with a progressive decrease in fine scale signal information between successive scales. This allows analyzing signals at coarser scales without the influence of finer scales such as those pertaining to noise. Knowing this, one can employ scale-space theory to suppress noise by performing scale-space decomposition on the signal and then treating one of signals at a coarser scale as the noise-suppressed signal. However, selecting the scale that represents the noise-suppressed signal can be challenging. Moreover, noise suppression using scale-space theory does not facilitate the fine-grained noise suppression at the individual scales, which limits its overall flexibility in striking a balance between noise suppression and signal structural preservation [20].

In Discrete Wavelet Transform (DWT), the original signal is decomposed into *approximation* and *detail* by passing the signal through a low-pass filter and high-pass filter, respectively, followed by a downsampling by a factor of 2. This results in a decomposition of the signal into different scales, which can be considered as low and high frequency bands. Multi-scale decomposition by wavelet transforms has a number of advantages over the scale-space decomposition and empirical mode decomposition [20]. First, since the signal information at one scale is not contained in another scale, signal information at different scales are better separated in the wavelet domain. Second, scale selection when performing noise suppression using wavelet decomposition is less critical than that for noise suppression using scale-space decomposition as all the scales are considered in noise suppression using wavelet decomposition as opposed to a single scale as done in scale-space decomposition. However, there are a number of limitations pertaining to noise suppression using wavelet transform [20]. For instance, signals processed using wavelet transforms can exhibit oscillation artifacts related to wavelet basis functions used in the wavelet transform, which is particularly noticeable in low Signal-to-Noise Ratio (SNR) regimes. Moreover, in DWT the approximation and detail are not statistically independent, which may cause a poor performance in signal denoising.

Motivation and Approach: Given these limitations of both space-scale and wavelet decomposition in terms of signal denoising, we were motivated to explore alternative approaches. We investigate the problem of decomposing a signal into multiple scales from a different point of view, i.e., we propose a new approach that takes a statistical perspective on multi-scale decomposition according to which a signal is considered as a mix-

ture of statistically independent signals, each characterizing signal information at a different scale. Having this perspective in multi-scale decomposition can be beneficial in signal denoising for two reasons. First, since most of the signal information in one scale is not included in the other scales, such decomposition provides the advantage of noise suppression at the individual scales in order to trade off noise suppression for signal-quality preservation. Second, since the noise signal is statistically independent from the original signal, by decomposing the noisy signal into statistically independent scales, the noise is expected to be separated in finer scales.

Our Contribution: Given this motivation and perspective, we propose a new method for Multi-Scale decomposition exploiting Independent Component Analysis (ICA), called MSICA, in which the original digital signal is decomposed into approximation and detail with statistically independent components. Specifically, we extract two correlated signals from the original signal and apply a linear transformation to the extracted signals so as to decompose the original signal into multiple scales. Since we need a suitable transform to decompose the original signal into statistically independent components, we consider our problem as a Blind Source Separation (BSS) problem in which the extracted signals are considered as the *observations* of the source separation problem. To relate this problem with the concept of multi-scale decomposition, we introduce an equivalent filter-bank structure for the proposed method, which is similar to the structure introduced in [14] for the DWT implementation. We also propose a method for multi-channel transmission in which MSICA outperforms common wavelet transforms in denoising of the received signal. We show that if the even and odd samples of the original signal are transmitted through two Additive White Gaussian Noise (AWGN) channels, MSICA is able to extract and filter out the noise from the noisier channel. This key property of MSICA—which exploits channel diversity and generalizes to the case in which more than two channels are available—can be used to increase the transmission efficiency in noisy communication channels, although the marginal return diminishes as the number of channels increases. It should be noted that, although single-channel ICA has been used in previous works [21, 22, 23, 24, 25] (including the spatial case of using even and odd samples), in this work single-channel ICA has been studied as a technique for signal decomposition into statistically independent approximation and detail and its performance in denoising has been compared with other wavelet transforms.

Article Organization: In Section 2, we provide some background on BSS and ICA. In Section 3, we propose our ICA-based transform for multi-scale decomposition. In Section 4, we examine the performance of MSICA in signal denoising and show how to increase transmission efficiency when multiple (noisy) channels are available. Finally, in Section 5, we draw the main conclusions and wrap up the article by discussing future work.

2. BLIND SOURCE SEPARATION

In BSS, a set of mixtures of different source signals is available and the goal is to separate the source signals when we have no information about the mixing system or the source signals (hence the name "blind") [26, 27].

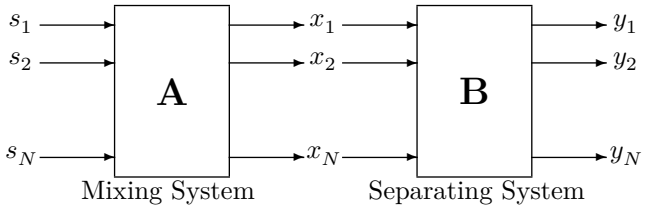


Fig. 1 – Mixing and separating systems in Blind Source Separation (BSS).

As in Fig. 1, the mixing and separating systems can be represented as,

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{A}\mathbf{s}(n), \\ \mathbf{y}(n) &= \mathbf{B}\mathbf{x}(n), \end{aligned} \tag{1}$$

where $\mathbf{s}(n) = [s_1(n), \dots, s_N(n)]^T$ is the vector of sources that are mixed by the mixing matrix $\mathbf{A}$ and create the observations vector $\mathbf{x}(n) = [x_1(n), \dots, x_N(n)]^T$. Let $\mathbf{A}$ be a square matrix ($N \times N$) of full column rank, which means that the number of sources is equal to the number of observations and that the sources are linearly independent. The goal of BSS is to find the separating matrix $\mathbf{B}$ such that $\mathbf{y}(n) = [y_1(n), \dots, y_N(n)]^T$ is an estimation of the sources.

A method to solve the BSS problem is via ICA, which exploits the assumption of source independence and estimates $\mathbf{B}$ such that the outputs y_i's be statistically independent [28]. As studied in [28, 29], the ICA can be performed by two steps: 1) whitening (or decorrelating) and 2) rotation. To illustrate the ICA model, we consider two independent components, s_i, $i = 1, 2$, with a uniform distribution,

$$p(s_i) = \begin{cases} 1 & \text{if } |s_i| \leq 0.5, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where the joint density of s_1 and s_2 is uniform on a square, as illustrated in Fig. 2(a). This follows from the definition that the joint density of two independent variables is the product of their marginal densities. Let us now mix these two independent components by the following mixing matrix $\mathbf{A}$,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}, \tag{3}$$

where the mixed variables x_1 and x_2 have a uniform distribution on a parallelogram, as shown in Fig. 2(b). Note that x_1 and x_2 are not independent anymore. To show this consider whether it is possible to predict the value of one of them, say x_2, from the value of the other; it is clear that if x_1 attains one of its maximum or minimum values, then this completely determines the value

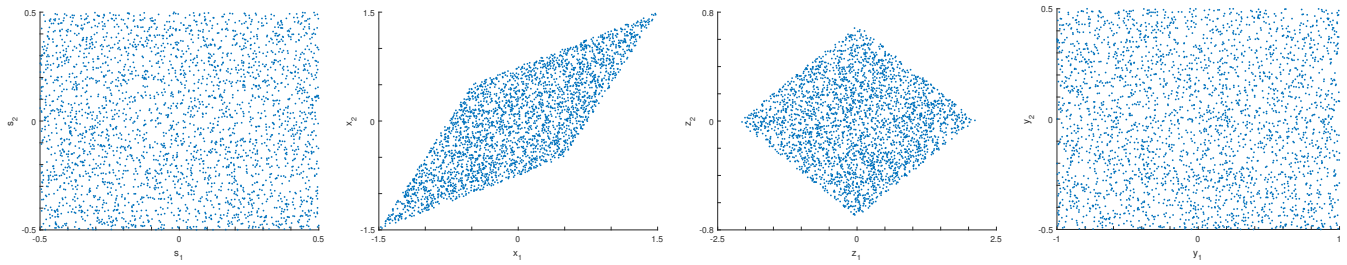


Fig. 2 – Illustration of Independent Component Analysis (ICA) algorithm. (a) Joint distribution of the independent components s_1 and s_2 with uniform distribution; (b) Joint distribution of the observed mixtures x_1 and x_2; (c) Joint distribution of the whitened mixtures, z_1 and z_2; (d) Joint distribution of the independent output components, y_1 and y_2, as determined by the ICA.

of x_2. However, the situation for variables s_1 and s_2 is different: from Fig. 2(a) it is clear that knowing the value of s_1 does not help predict the value of s_2.
The problem of source separation is now to estimate the mixing matrix $\mathbf{A}$ and multiply its inverse ($\mathbf{B} = \mathbf{A}^{-1}$) to the vector of the mixtures to retrieve the sources (s_1 and s_2). As studied in [28, 29], the ICA can be considered as a two-step procedure where, in the first step, the mixed data gets whitened (uncorrelated) by multiplying the whitening matrix by the vector of mixtures, i.e.,

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \mathbf{W} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \tag{4}$$

and then, in the second step, the independent components are separated by applying a rotation matrix, i.e.,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{R} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}. \tag{5}$$

Fig. 2(c) illustrates the effect of whitening in which the data in Fig. 2(b) has been whitened. Also, Fig. 2(d) shows how rotating the whitened data can return the statistical independency to the outputs and recover the independent components. Therefore, the separation matrix $\mathbf{B}$ can be considered as the product of the whitening and rotation matrices, i.e.,

$$\mathbf{B} = \mathbf{R}\mathbf{W}, \tag{6}$$

where $\mathbf{W}$ is the whitening matrix and $\mathbf{R}$ is the rotation matrix. Note that in the case whitened components are statistically independent, the rotation matrix $\mathbf{R}$ will be the identity matrix and no rotation will be needed.

3. MSICA: MULTI-SCALE DECOMPOSITION BY INDEPENDENT COMPONENT ANALYSIS

Generally, neighboring/consecutive samples of a common signal are highly correlated and differ slightly from each other. This slight difference of neighboring samples comes from the details of the signal. If we consider the detail scale of the original signal to be statistically independent from the approximation scale, it is expected that decomposing the neighboring samples of the signal into their independent components would decompose the signal into its approximation and detail. Given this motivation, we propose an ICA-based method for multi-scale decomposition in which the approximation and details are statistically independent. Our algorithm consists of two steps: 1) extracting the observation signals from the original signal and 2) decomposing the original signal into approximation and detail by applying a linear transformation to the observation signals.
Suppose that $x(n)$ is a Wide Sense Stationary (WSS) signal. We consider even and odd samples of $x(n)$ as $x_1(n)$ and $x_2(n)$, respectively, i.e.,

$$x_1(n) = x(2n), \quad x_2(n) = x(2n-1). \tag{7}$$

We prove that, if $x_1(n)$ and $x_2(n)$ are the observations signals (mixtures) of the ICA, the outputs of the ICA will be the approximation and detail of $x(n)$, which are statistically independent. We define the observation vector $\mathbf{x}$ as,

$$\mathbf{x} = \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} x(2n) \\ x(2n-1) \end{bmatrix}, \tag{8}$$

where the even and odd samples of the original signal are the first and second rows of the observation vector, respectively. If we apply the ICA to the observation vector, the output is,

$$\mathbf{y} = \mathbf{B}\mathbf{x} = \begin{bmatrix} y_1(n) \\ y_2(n) \end{bmatrix}, \tag{9}$$

where $\mathbf{B}$ is the separation matrix estimated by the ICA, and $y_1(n)$ and $y_2(n)$ are statistically independent.
To prove our claim, we need to show that $y_1(n)$ and $y_2(n)$ are the approximation and detail of the original signal. As explained in Section 2, ICA involves two steps: 1) whitening (or decorrelating) and 2) rotation, and the separation matrix $\mathbf{B}$ can be considered as the product of the whitening and rotation matrices (i.e., $\mathbf{B} = \mathbf{R}\mathbf{W}$). Now, let us consider the whitened vector $\mathbf{z}$ and the whitening matrix $\mathbf{W}$ as follows,

$$\mathbf{z} = \begin{bmatrix} z_1(n) \\ z_2(n) \end{bmatrix} = \mathbf{W}\mathbf{x}, \quad \mathbf{W} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}. \tag{10}$$

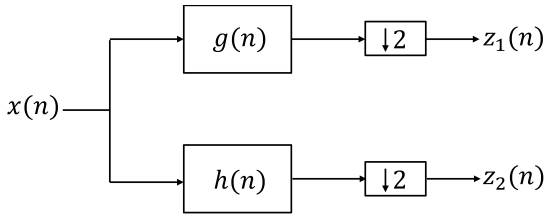


Fig. 3 – Whitening process in ICA as a filter-bank structure.

Using *(8)* and *(10)*, we can write,

$$\begin{aligned} z_1(n) &= w_{11}x_1(n) + w_{12}x_2(n), \\ z_2(n) &= w_{21}x_1(n) + w_{22}x_2(n). \end{aligned} \tag{11}$$

Also, $x_1(n)$ and $x_2(n)$ can be written as,

$$\begin{aligned} x_1(n) &= x(2n) = [x(n) * \delta(n)] \downarrow 2, \\ x_2(n) &= x(2n-1) = [x(n) * \delta(n-1)] \downarrow 2, \end{aligned} \tag{12}$$

where $*$ and $\downarrow 2$ denote convolution and downsampling by a factor of 2, respectively, and $\delta(n)$ is the unit impulse signal. Using *(11)* and *(12)*, we can rewrite $z_1(n)$ and $z_2(n)$ as,

$$\begin{aligned} z_1(n) &= w_{11}x(2n) + w_{12}x(2n-1) \\ &= w_{11}[x(n) * \delta(n)] \downarrow 2 + w_{12}[x(n) * \delta(n-1)] \downarrow 2 \\ &= [x(n) * (w_{11}\delta(n) + w_{12}\delta(n-1))] \downarrow 2, \\ z_2(n) &= w_{21}x(2n) + w_{22}x(2n-1) \\ &= w_{21}[x(n) * \delta(n)] \downarrow 2 + w_{22}[x(n) * \delta(n-1)] \downarrow 2 \\ &= [x(n) * (w_{21}\delta(n) + w_{22}\delta(n-1))] \downarrow 2. \end{aligned} \tag{13}$$

Note that, if we consider the low-pass filter $g(n)$ and high-pass filter $h(n)$ as,

$$\begin{aligned} g(n) &= w_{11}\delta(n) + w_{12}\delta(n-1), \\ h(n) &= w_{21}\delta(n) + w_{22}\delta(n-1), \end{aligned} \tag{14}$$

(13) can be rewritten in a simpler form as,

$$\begin{aligned} z_1(n) &= [x(n) * g(n)] \downarrow 2, \\ z_2(n) &= [x(n) * h(n)] \downarrow 2. \end{aligned} \tag{15}$$

From *(15)*, we observe that the whitening process can be modeled as a filter-bank structure, as shown in Fig. 3. Now, we need to show that $g(n)$ and $h(n)$ are indeed low-pass and high-pass filters. In order to do so, we consider the covariance matrix of $\mathbf{x}$ as follows,

$$\mathbf{C}_x = E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T, \tag{16}$$

where $\mathbf{Q}$ is an orthogonal matrix of eigenvectors and $\mathbf{D}$ is a diagonal matrix of eigenvalues. Interestingly, the covariance matrix of $\mathbf{z} = \mathbf{Q}^T\mathbf{x}$ can be written as,

$$\mathbf{C}_z = E\{\mathbf{z}\mathbf{z}^T\} = E\{\mathbf{Q}^T\mathbf{x}\mathbf{x}^T\mathbf{Q}\} = \mathbf{Q}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{Q}. \tag{17}$$

Given *(16)* and *(17)*, we can write,

$$\mathbf{C}_z = \underbrace{\mathbf{Q}^T\mathbf{Q}}_{\mathbf{I}}\mathbf{D}\underbrace{\mathbf{Q}^T\mathbf{Q}}_{\mathbf{I}} = \mathbf{D}. \tag{18}$$

Since $\mathbf{D}$ is a diagonal matrix, we conclude that multiplying $\mathbf{Q}^T$ by the observation vector $\mathbf{x}$ results in uncorrelated outputs. Hence, the whitening matrix $\mathbf{W}$ can be considered to be equal to $\mathbf{Q}^T$.

Now, we need to find the elements of matrix $\mathbf{W}$ so as to determine finally the $g(n)$ and $h(n)$ filters. Since $x(n)$ is a WSS signal, we have,

$$\begin{aligned} R_x(0) = E\{x^2(2n)\} = E\{x^2(2n+1)\} = \sigma_x^2, \\ R_x(1) = E\{x(2n)\,x(2n-1)\} = \sigma_x^2\rho. \end{aligned} \tag{19}$$

Hence, with regard to *(8)*, we can recast $\mathbf{C}_x$ as,

$$\mathbf{C}_x = E\{\mathbf{x}\mathbf{x}^T\} = \begin{bmatrix} \sigma_x^2 & \sigma_x^2\rho \\ \sigma_x^2\rho & \sigma_x^2 \end{bmatrix} = \sigma_x^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{20}$$

where the eigenvalues and eigenvectors of $\mathbf{C}_x$ are,

$$\begin{aligned} \lambda_1 = \sigma_x^2(1+\rho) \quad \lambda_2 = \sigma_x^2(1-\rho), \\ \mathbf{q}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \mathbf{q}_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}. \end{aligned} \tag{21}$$

Hence, the whitening matrix $\mathbf{W}$ can be presented as,

$$\mathbf{W} = \mathbf{Q}^T = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{bmatrix}^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{22}$$

Comparing *(14)* and *(22)*, $g(n)$ and $h(n)$ can be written as,

$$\begin{aligned} g(n) &= \tfrac{1}{\sqrt{2}}\delta(n) + \tfrac{1}{\sqrt{2}}\delta(n-1), \\ h(n) &= -\tfrac{1}{\sqrt{2}}\delta(n) + \tfrac{1}{\sqrt{2}}\delta(n-1). \end{aligned} \tag{23}$$

From *(23)*, it is clear that $g(n)$ and $h(n)$ are a low and high-pass filter, respectively, as we wanted to demonstrate. Hence, we can conclude that the whitening process in the ICA (presented as a filter-bank structure in Fig. 3) decomposes the observation signals into uncorrelated approximation and detail. Moreover, the rotation process in the ICA makes sure that the approximation and the detail are statistically independent[1]. Hence, if the even and odd samples of a one-dimensional signal are considered as the observations of the ICA, we ensure the approximation and detail to be statistically independent. As we will see in the next section, the statistical independency between the approximation and detail can be very beneficial in signal denoising, especially when the even and odd samples are transmitted through different (noisy) channels.

Fig. 4 showcases a signal decomposition by different wavelet transform and MSICA, where a Piece-Regular signal is decomposed into the approximation and detail. As shown in Fig. 4(h), like all the other wavelet transforms, MSICA is also able to decompose the original signal into approximation and detail, where the approximation and detail coefficients contain the low and high-frequency components, respectively. As said earlier, to

[1]Based on our simulations, the separation matrix was always close to *(22)*, which means that the rotation matrix was always close to the identity matrix.

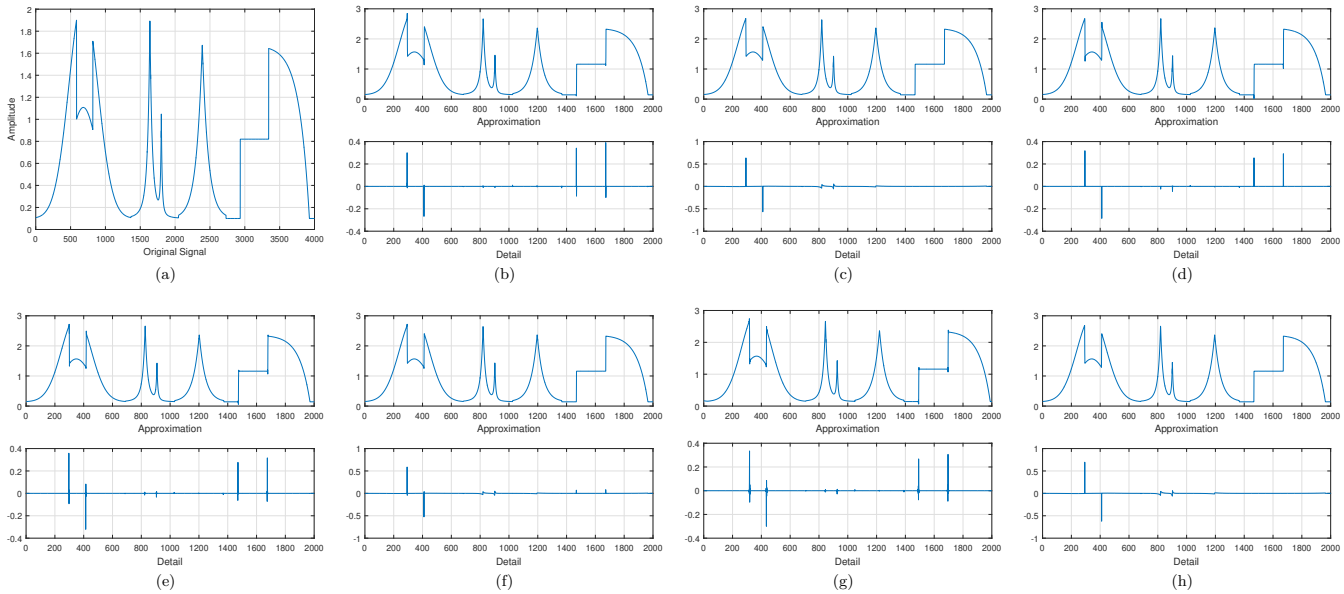


Fig. 4 – Comparison of MSICA with different wavelet transforms in decomposing a Piece-Regular signal. (a) Original Signal. Approximation and detail by (b) Daubechies 3 wavelet; (c) Haar wavelet; (d) Biorthogonal 2.2 wavelet; (e) Coiflets 4 wavelet; (f) Fejer-Korovkin 4 wavelet; (g) discrete Meyer wavelet; and (h) our proposed MSICA.

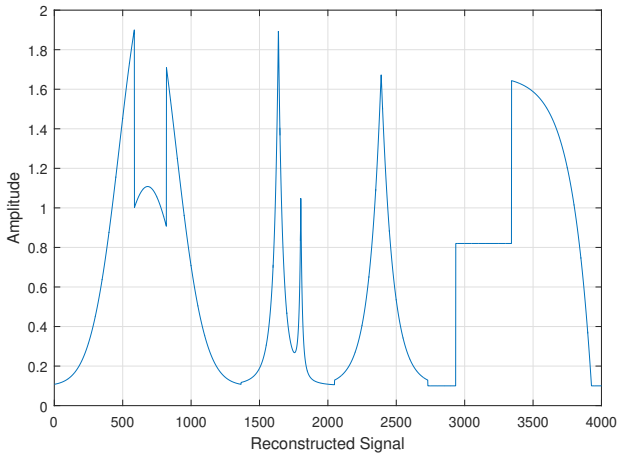


Fig. 5 – Reconstructed signal by MSICA using the approximation and detail in Fig. 4(h).

reconstruct the original signal we need to multiply the inverse of the separation matrix ($\mathbf{B}^{-1}$) by the vector of approximation and detail $\mathbf{y}$ so as to obtain the observation vector $\mathbf{x}$ and reconstruct the original signal $x(n)$ from $\mathbf{x}$, i.e.,

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \mathbf{x} = \mathbf{B}^{-1}\mathbf{y}, \qquad (24)$$

$$x(n) = (x_1(n) \uparrow 2) + (x_2(n) \uparrow 2) * \delta(n+1), \qquad (25)$$

where $x_i(n) \uparrow 2$, $i = 1, 2$, denotes the upsampling of $x_i(n)$ by a factor of 2. Fig. 5 shows the reconstructed signal using *(24)* and *(25)*. This figure shows that MSICA can successfully reconstruct the original signal from the approximation and detail obtained in the decomposition procedure.

4. MSICA FOR SIGNAL DENOISING

We discuss now how MSICA can be beneficial in signal denoising. Specifically, we compare MSICA with the other wavelet transforms and show how MSICA can suppress the noise via a simple wavelet thresholding. We also show that, in the case of multi-channel transmission, MSICA outperforms the other wavelet transforms and is able to extract and filter out the noise of the noisier communication channel by exploiting channel diversity.

Let us assume that the original signal is passed through an AWGN channel, the noisy output signal is then,

$$r(n) = x(n) + w(n), \quad n = 1, \ldots, N, \qquad (26)$$

where $x(n)$ is the original signal and $w(n)$ is the AWGN with zero mean and variance of σ_w^2. The goal of signal denoising is to remove the noise and obtain an estimate $\hat{x}(n)$ of $x(n)$ that minimizes the Mean Squared Error (MSE), defined as,

$$\mathrm{MSE}(\hat{x}) = \frac{1}{N}\sum_{n=1}^{N}(\hat{x}(n) - x(n))^2. \qquad (27)$$

Note that the model presented in *(26)* is not general since the noise may be non-additive, and the relation between the noisy observed signal and the original signal may be stochastic. Nevertheless, *(26)* is a widely used model in many practical situations as it serves well as a motivating example to give a good sense as to what happens in more realistic channels.

Let us emphasize that there are many existing approaches in the literature to perform signal denoising, which can be roughly divided into two main categories: 1) denoising in the original signal domain (e.g., in time

or space [30]) and 2) denoising in the transform domain (e.g., using Fourier or a wavelet transform [31, 32]). Since the wavelet transform provides information in both the time and frequency domain, and the information in one scale is not contained in another scale, approaches in this second category can strike a balance between noise suppression and signal structural preservation, and have therefore been used widely in signal denoising. It is interesting to note that, usually, the detail coefficients of a noiseless signal are sparse. This means that in the wavelet transform most of the detail coefficients of a noiseless signal are very small/close to zero (as can be seen, for instance, in Fig. 4). So, the detail coefficients with a small magnitude can be considered as a noise component and can be set to zero. This is the basic idea of wavelet thresholding approaches, which are employed in signal denoising where the coefficients are compared with a threshold to determine whether they contain only noise or not.

It should be noted that since the approximation coefficients contain the low-frequency/important information of the signal, the thresholding is usually applied only to the detail coefficients (high-frequency components). The thresholding methods retain the significant components by setting to zero only the detail coefficients whose absolute values are less than a certain threshold. Most of the thresholding approaches try to find the optimal threshold value. The SureShrink method [33], proposed by Donoho and Johnstone, is an effective wavelet thresholding method for signal denoising that fits a level-dependent threshold to the wavelet coefficient. To show the performance of MSICA with respect to the other wavelet transforms, we extract the first and second-level detail coefficients of the noisy signal and use the SureShrink method for signal denoising. In our experiments, we also employ the FastICA method [34] and use the different standard test signals given in [35], i.e., Piece-Regular, Blocks, Doppler, and HeaviSine.

We consider now two scenarios for signal transmission and compare the signal denoising performance of MSICA in single and multi-channel transmissions.

Single-channel Transmission: In the first scenario, i.e., in the case of single-channel transmission as described in *(26)*, if we divide the noisy signal into the even and odd samples, we obtain the vector of noisy observations as,

$$\mathbf{r} = \begin{bmatrix} r_1(n) \\ r_2(n) \end{bmatrix} = \begin{bmatrix} r(2n) \\ r(2n-1) \end{bmatrix} = \begin{bmatrix} x(2n)+w(2n) \\ x(2n-1)+w(2n-1) \end{bmatrix}, \tag{28}$$

where $w(2n)$ and $w(2n-1)$ have the same variance, σ_w^2. The covariance matrix of $\mathbf{r}$ can be written as,

$$\mathbf{C}_r = E\{\mathbf{r}\mathbf{r}^T\} = \sigma_r^2 \begin{bmatrix} 1 & \rho' \\ \rho' & 1 \end{bmatrix}, \tag{29}$$

where,

$$\sigma_r^2 = \sigma_x^2 + \sigma_w^2, \quad \rho' = \frac{\rho\sigma_x^2}{\sigma_x^2+\sigma_w^2}. \tag{30}$$

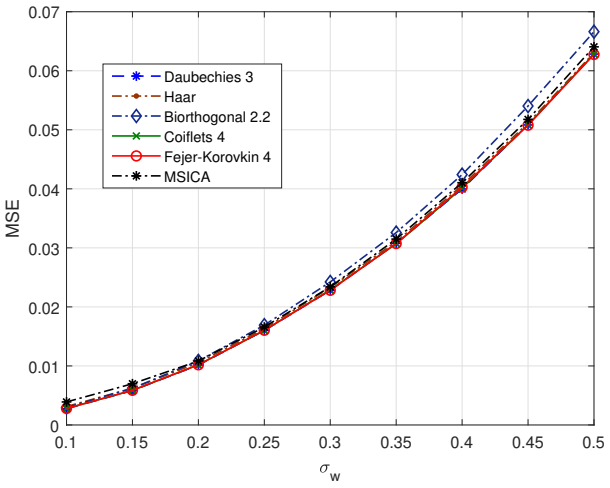


Fig. 6 – MSICA Mean Square Error (MSE) compared to well-known wavelet transforms.

Comparing *(29)* and *(20)*, it is clear that the whitening matrix for the vector of noisy observations $\mathbf{r}$ is like the one in *(22)*,

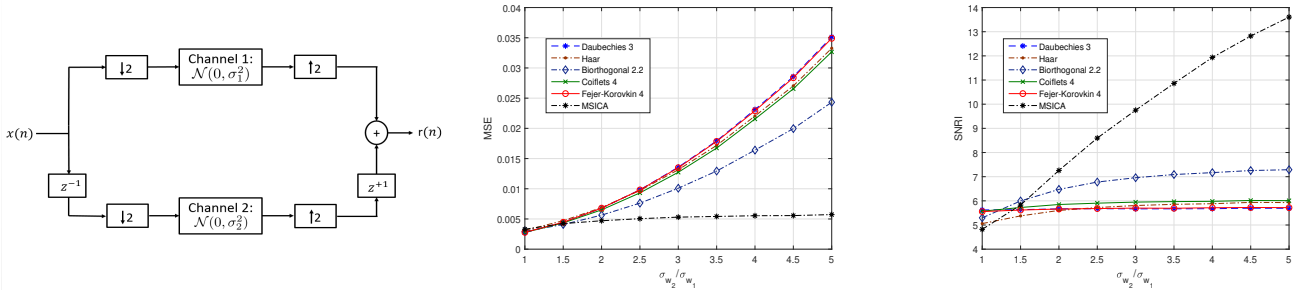
$$\mathbf{W} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{31}$$

Fig. 6 shows the performance of MSICA in the first scenario in terms of MSE against well-known wavelet transforms. As expected, MSICA shows similar performance compared to the other wavelet transform in suppressing the noise level and enhancing the quality of the signal as it is able to decompose the signal into approximation and detail. However, in the following we will explain how MSICA can have extraordinary performance when the odd and even samples of the original signal are passed through different channels.

Multi-channel Transmission: Let us consider now the second scenario where we investigate the performance of MSICA in a multi-channel transmission. Assume that two AWGN channels, named CH1 and CH2, are available to transmit the signal where, for example, we assume the variance of the noise in CH1 to be smaller than in CH2, which means that CH1 is more reliable and has better quality than CH2. In this case, if we transmit the even and odd samples through different channels, see Fig. 7(a), no matter through which one, then MSICA shows an extraordinary performance as it is able to filter out the noise of a noisier channel. In such a scenario, the even and odd samples of the noisy signal are,

$$r_1(n) = x(2n) + w_1(n), \tag{32}\text{a}$$
$$r_2(n) = x(2n-1) + w_2(n), \tag{32}\text{b}$$

where $w_1(n)$ and $w_2(n)$ are the AWGN in CH1 and CH2, respectively, and $\sigma_{w_2}^2 = K\sigma_{w_1}^2$ $(K>1)$. In this case, the covariance matrix of the vector of noisy observations $\mathbf{r}$



(a) (b) (c)

Fig. 7 – (a) Proposed multi-channel transmission where the even and odd samples of the original signal are transmitted through different channels and the receiver reconstructs the signal from the two outputs (z^{-1} and z^{+1} indicates time shift by n=1 to the right and left, respectively); (b) Mean Squared Error (MSE) vs. $\sigma_{w_2}/\sigma_{w_1}$ ($\sigma_{w_1} = 0.1$); (c) Performance of MSICA in terms of Signal-to-Noise Ratio Improvement (SNRI).

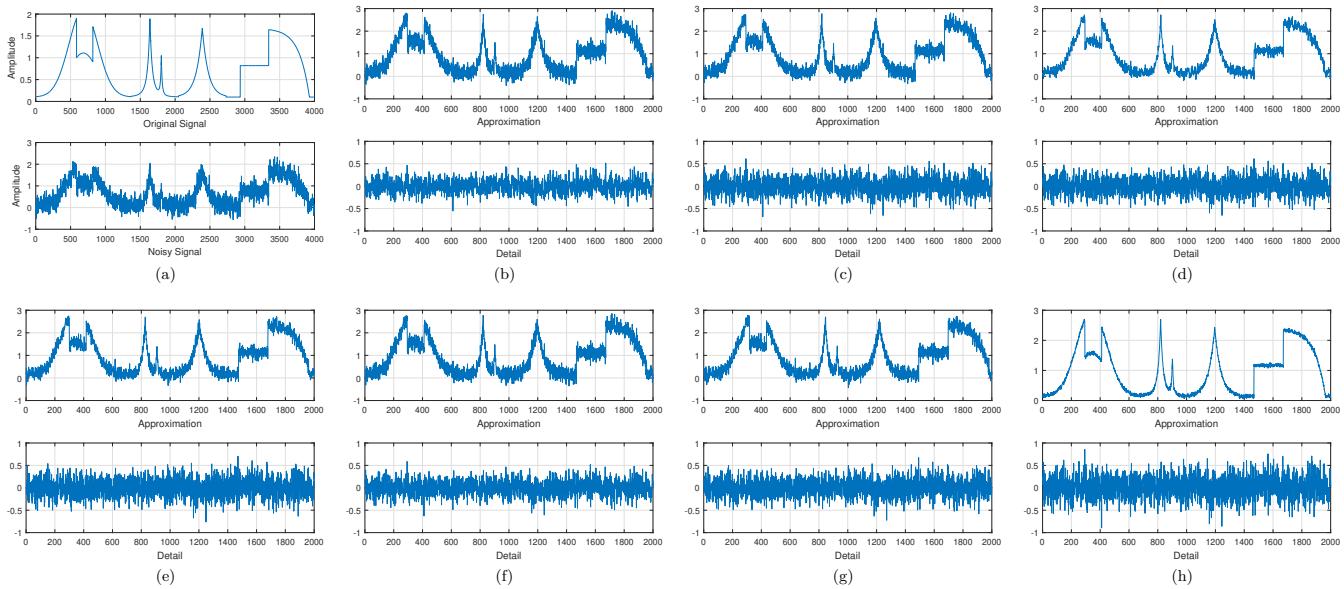


Fig. 8 – Comparison of MSICA with different wavelet transforms in decomposing a noisy PieceRegular signal. (a) Original Signal; (b) Approximation and detail by Daubechies 3 wavelet; (c) Approximation and detail by Haar wavelet; (d) Approximation and detail by Biorthogonal 2.2 wavelet; (e) Approximation and detail by Coiflets 4 wavelet; (f) Approximation and detail by Fejer-Korovkin 4 wavelet; (g) Approximation and detail by discrete Meyer wavelet; (h) Approximation and detail by MSICA.

is,

$$\mathbf{C}_r = \sigma_r^2 \begin{bmatrix} 1 & \rho' \\ \rho' & 1 + \frac{(K-1)\sigma_{w_1}^2}{\sigma_r^2} \end{bmatrix}, \qquad (33)$$

where,

$$\sigma_r^2 = \sigma_x^2 + \sigma_{w_1}^2, \quad \rho' = \frac{\rho\sigma_x^2}{\sigma_x^2 + \sigma_{w_1}^2}. \qquad (34)$$

As it is clear from *(33)*, in the case that the variance of the noise in CH1 and CH2 are different, the eigenvalues and eigenvectors of $\mathbf{C_r}$ are dependent to the parameter K. This means that the low-pass and high-pass filters in the whitening process will be adaptive to the parameter K. In the following we will show that this adaptive filter is able to reject the effect of CH2 almost entirely. Figure 7(b) shows the performance of MSICA with respect to the other wavelet transforms when the original signal is passed through two different channels. As shown in Fig. 7(b), MSICA performance does not depend on the variance of CH2, which means that MSICA is able to reject the AWGN of the CH2 from the noisy signal.

Moreover, to evaluate better the noise suppression performance, we have also examined the performance of MSICA in terms of Signal-to-Noise Ratio Improvement (SNRI),

$$\text{SNRI} = \text{SNR}_{out} - \text{SNR}_{in} = 10\log\left(\frac{\sum_{n=1}^{N}(r(n)-x(n))^2}{\sum_{n=1}^{N}(\hat{x}(n)-x(n))^2}\right), \qquad (35)$$

where SNR_{out} and SNR_{in} are the SNR of the denoised signal (output) and of the noisy signal (input), respectively. As shown in Fig. 7(c), the wavelet transforms have almost a fixed SNRI, whereas MSICA shows higher SNRI when the CH2 becomes noisier. This is because in MSICA the approximation and detail are statisti-

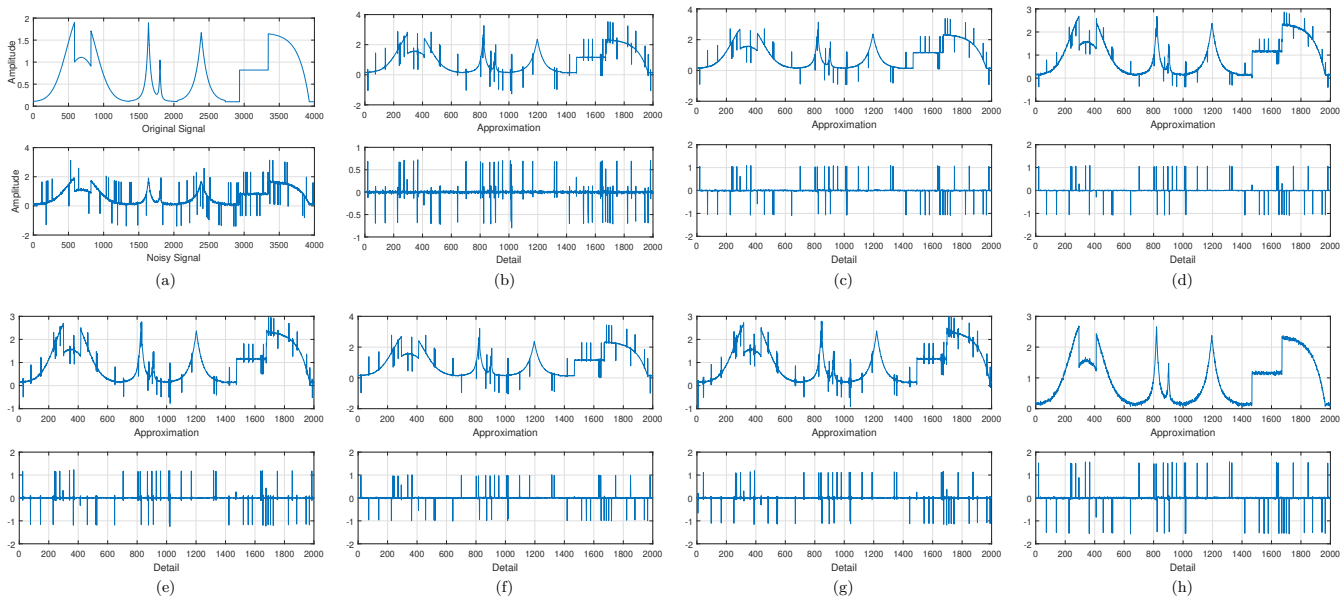


Fig. 9 – Comparison of MSICA with different wavelet transforms in decomposing a PieceRegular signal corrupted by impulse noise. (a) Original and Noisy Signal; (b) Approximation and detail by Daubechies 3 wavelet; (c) Approximation and detail by Haar wavelet; (d) Approximation and detail by Biorthogonal 2.2 wavelet; (e) Approximation and detail by Coiflets 4 wavelet; (f) Approximation and detail by Fejer-Korovkin 4 wavelet; (g) Approximation and detail by discrete Meyer wavelet; (h) Approximation and detail by MSICA.

cally independent; hence, MSICA is able to extract the noise signal from CH2 (via channel diversity), while the wavelet transforms are not able to do so.

Fig. 8 shows a signal decomposition where $\sigma_{w_2} = 0.2$, $\sigma_{w_1} = 0.05$. As it can be seen, the approximation obtained using MSICA is less noisy than the one obtained using the other wavelet transforms (Daubechies 3, Haar, Biorthogonal 2.2, Coiflets 4, Fejer-Korovkin 4, and Meyer). This result confirms our statement and shows that, because of the statistical independenc between the approximation and detail, MSICA is able to extract the AWGN from the noisier channel.

In the other experiment, in order to show visibly that MSICA is able to extract the noise of CH2, we have explored its performance when the odd samples, passed through CH2, are corrupted by impulse noise. The Probability Density Function (PDF) of the impulse noise is given as,

$$P(w) = \begin{cases} P_a & w = a, \\ P_a & w = -a, \\ 1 - 2P_a & w = 0, \end{cases} \qquad (36)$$

where $2P_a$ is the probability of existence of impulse noise in the received samples. In Fig. 9(a), the noisy signal is obtained by passing the even samples of the original signal through CH1 with AWGN with zero mean and $\sigma^2_{w_2} = 0.004$, while the odd samples were passed through CH2 with impulse noise ($P_a = 0.01$ and $a = 1.5$). Fig. 9(b)-(h) show the performance of MSICA compared to a number of well-known wavelet transforms. As it is clear from Fig. 9(b)-(g), the traditional wavelet transforms (i.e., Daubechies 3, Haar, Biorthogonal 2.2, Coiflets 4, Fejer-Korovkin 4, Meyer) are not able to extract accurately the impulse components from the noisy signal. However, as it is shown in Fig. 9(h), MSICA is successful as the detail contains all the impulse components. This is because in MSICA the approximation and detail are statistically independent and, since the impulse noise is statistically independent from the original signal, MSICA can extract it in the detail coefficients.

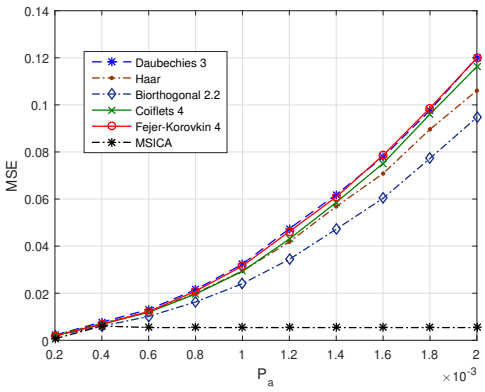


Fig. 10 – Impulse noise rejection in terms of Minimum Square Error (MSE); MSICA performance does not depend on P_a, whereas the performance of the other transforms decreases when P_a increases.

Fig. 10 shows the performance of MSICA compared with different wavelet transforms when various values of P_a, as in *(36)*, are considered. Here, the detail coefficients obtained by different methods have been set to zero to denoise the noisy signal. Since MSICA is able to extract the impulse noise, its performance does not depend on P_a, whereas the performance of the other transforms decreases when P_a increases.

To show that MSICA works on real signals too, we

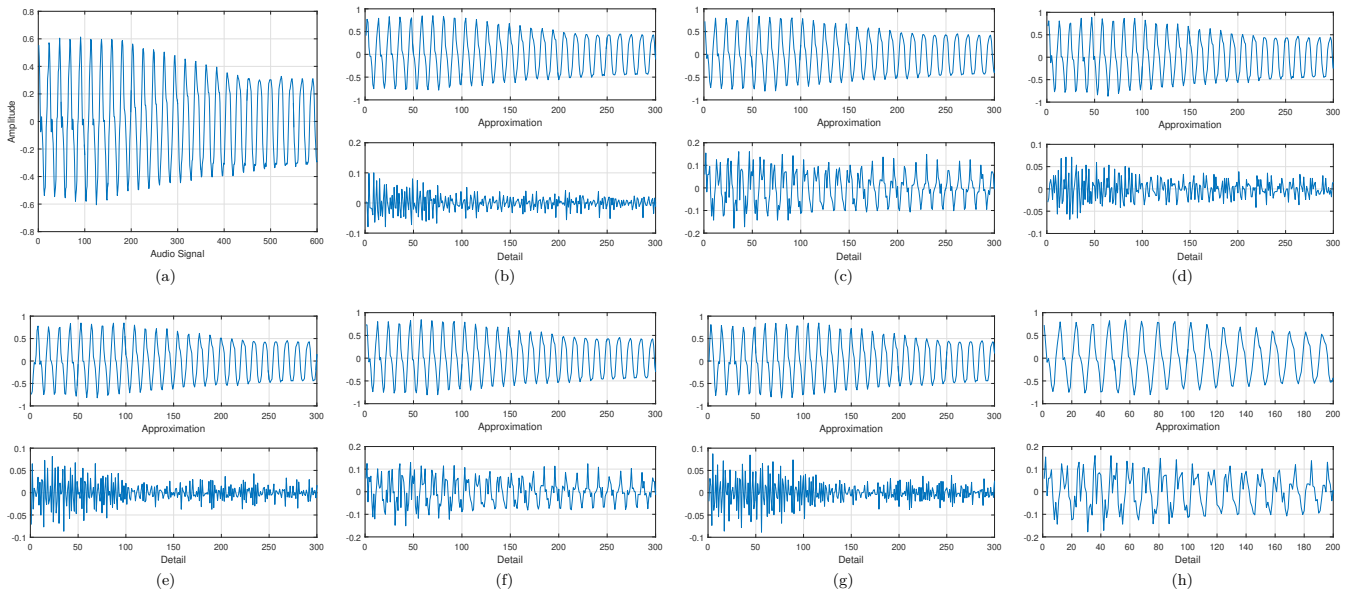


Fig. 11 – Approximation and detail coefficients obtained from an audio signal by some well-known wavelet transforms and MSICA. (a) Original audio Signal. Approximation and detail by (b) Daubechies 3; (c) Haar; (d) Biorthogonal 2.2; (e) Coiflets 4; (f) Fejer-Korovkin 4; (g) discrete Meyer; (h) MSICA.

have examined its performance on the signal depicted in Fig. 11(a), which is an audio signal with a sampling frequency equal to 16 KHz. Fig. 11(b)-(g) show the performance of the considered wavelet transforms in decomposing this signal into approximation and detail. As it is clear from Fig. 11(h), like the other transforms, MSICA is also able to decompose this audio signal into approximation and detail.

5. CONCLUSIONS

We presented MSICA, a new method for Multi-Scale decomposition based on Independent Component Analysis (ICA), where the approximation and detail are statistically independent. First, we extracted two correlated signals from the original digital signal by separating their even and odd samples; then, we proved that extracting the independent components of the correlated signals leads to the decomposition of the original signal into the approximation and detail. We showed that MSICA outperforms well-known wavelet transforms in signal denoising when transmitting a signal over multiple (noisy) channels as it exploits channel diversity. This property makes MSICA useful in many critical scenarios such as transmitting multimedia content reliably through underwater acoustic channels. Since these channels may vary quickly over time, it is difficult to estimate them, which makes transmitting multimedia content underwater a very challenging task.

REFERENCES

[1] B. Ophir, M. Lustig, and M. Elad, "Multi-scale dictionary learning using wavelets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1014–1024, 2011.

[2] Y. Kopsinis and S. McLaughlin, "Development of emd-based denoising methods inspired by wavelet thresholding," *IEEE Transactions on Signal Processing*, vol. 57, no. 4, pp. 1351–1362, 2009.

[3] A. F. Laine, S. Schuler, J. Fan, and W. Huda, "Mammographic feature enhancement by multiscale analysis," *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 725–740, 1994.

[4] F. M. Bayer, A. J. Kozakevicius, and R. J. Cintra, "An iterative wavelet threshold for signal denoising," *Signal Processing*, vol. 162, pp. 10–20, 2019.

[5] P. G. Bascoy, P. Quesada-Barriuso, D. B. Heras, and F. Argüello, "Wavelet-based multicomponent denoising profile for the classification of hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 2, pp. 722–733, 2019.

[6] A. Wong and A. K. Mishra, "Generalized probabilistic scale space for image restoration," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2774–2780, 2010.

[7] G. Gilboa, "Nonlinear scale space with spatially varying stopping time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2175–2187, 2008.

[8] A. Mishra, A. Wong, D. A. Clausi, and P. W. Fieguth, "Quasi-random nonlinear scale space," *Pattern Recognition Letters*, vol. 31, no. 13, pp. 1850–1859, 2010.

[9] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *Journal of applied statistics*, vol. 21, no. 1-2, pp. 225–270, 1994.

[10] F. Jager, I. Koren, and L. Gyergyek, "Multiresolution representation and analysis of ecg waveforms," in *Proceedings Computers in Cardiology*. IEEE, 1990, pp. 547–550.

[11] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 112–114, 2004.

[12] B. Weng, M. Blanco-Velasco, and K. E. Barner, "ECG denoising based on the empirical mode decomposition," in *International Conference of the Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 1–4.

[13] M. Blanco-Velasco, B. Weng, and K. E. Barner, "ECG signal denoising and baseline wander correction based on the empirical mode decomposition," *Computers in biology and medicine*, vol. 38, no. 1, pp. 1–13, 2008.

[14] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[15] S. Sardy, P. Tseng, and A. Bruce, "Robust wavelet denoising," *IEEE Transactions on Signal Processing*, vol. 49, no. 6, pp. 1146–1152, 2001.

[16] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[17] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[18] G. Chen, T. D. Bui, and A. Krzyżak, "Image denoising with neighbour dependency and customized wavelet and threshold," *Pattern Recognition*, vol. 38, no. 1, pp. 115–124, 2005.

[19] A. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9. IEEE, 1984, pp. 150–153.

[20] A. Wong and X. Y. Wang, "A bayesian residual transform for signal processing," *IEEE Access*, vol. 3, pp. 709–717, 2015.

[21] B. Mijovic, M. De Vos, I. Gligorijevic, J. Taelman, and S. Van Huffel, "Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 9, pp. 2188–2196, 2010.

[22] S. Wang and C. J. James, "On the independent component analysis of evoked potentials through single or few recording channels," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 5433–5436.

[23] M. E. Davies and C. J. James, "Source separation using single channel ica," *Signal Processing*, vol. 87, no. 8, pp. 1819–1832, 2007.

[24] G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "Single-channel signal separation using time-domain basis functions," *IEEE Signal Processing Letters*, vol. 10, no. 6, pp. 168–171, 2003.

[25] J. Miettinen, M. Matilainen, K. Nordhausen, and S. Taskinen, "Extracting conditionally heteroskedastic components using independent component analysis," *Journal of Time Series Analysis*, vol. 41, no. 2, pp. 293–311, 2020.

[26] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[27] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ilrma originating from ica and nmf," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[28] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, 2000.

[29] P. Comon, "Independent component analysis, a new concept?" *Transactions on Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[30] I. Kauppinen and K. Roth, "Improved noise reduction in audio signals using spectral resolution enhancement with time-domain signal extrapolation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1210–1216, 2005.

[31] S. Durand and J. Froment, "Artifact free signal denoising with wavelets," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 6, 2001, pp. 3685–3688.

[32] Z. A. A. Alyasseri, A. T. Khader, M. A. Al-Betar, A. K. Abasi, and S. N. Makhadmeh, "Eeg signals denoising using optimal wavelet transform hybridized with efficient metaheuristic methods," *IEEE Access*, vol. 8, pp. 10 584–10 605, 2019.

[33] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[34] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[35] R. R. Coifman and D. L. Donoho, *Translation invariant denoising*. Springer, 1995.

AUTHORS

Abolfazl Hajisami received his PhD degree in Electrical and Computer Engineering (ECE) from Rutgers University, NJ, USA, in 2018. He received his MSc and BSc degrees from Sharif University of Technology and from Shahid Beheshti University (Tehran, Iran), in 2010 and 2008, respectively. His research interest lies in wireless communication, vehicle-to-everything (V2X) communication, and signal processing. His PhD thesis dealt with spectral and energy efficiency in cloud radio access networks. He is now a Staff Research Engineer at Qualcomm, where he works on V2X technologies.

Dario Pompili is an Assoc. Prof. with the Dept. of ECE at Rutgers, where he directs the Cyber-Physical Systems Laboratory (CPS Lab), which focuses on mobile edge computing, wireless networking, acoustic communications, and sensor networks. He received his PhD in ECE from the Georgia Institute of Technology in 2007. He had previously received his 'Laurea' (combined BS and MS) and Doctorate degrees in Telecommunications and System Engineering from the U. of Rome "La Sapienza," Italy, in 2001 and 2004, respectively. He has received a number of awards including the NSF CAREER'11, ONR Young Investigator Program'12, and DARPA Young Faculty'12 awards. In 2015, he was nominated Rutgers-New Brunswick Chancellor's Scholar. He has served in various roles on many international conferences. He published about 150 refereed scholar publications, some of which selected to receive best paper awards: with about 11,600 citations, Dr. Pompili has an h-index of 42 and an i10-index of 101 (Google Scholar, Oct'20). He is a Senior Member of the IEEE Communications Society (2014) and a Distinguished Member of the ACM (2019). He is an Area Editor for IEEE Transactions on Mobile Computing (TMC).

SDN-BASED SOCIOCAST GROUP COMMUNICATIONS IN THE INTERNET OF THINGS

Luigi Atzori[1,5], Claudia Campolo[2,5], Antonio Iera[3,5], Giuseppe Massimiliano Milotta[2,5], Giacomo Morabito[4,5], Salvatore Quattropani[5]
[1]DIEE, University of Cagliari, Italy, [2]DIIES, University Mediterranea of Reggio Calabria, Italy, [3]DIMES, University of Calabria, Italy, [4]DIEEI, University of Catania, Italy, [5]CNIT - National Inter-University Consortium for Telecommunications, Italy,

NOTE: Corresponding author: Giacomo Morabito, giacomo.morabito@unict.it

Abstract – *The new applications populating the Future Internet will increasingly rely on the exchange of data between groups of devices, dynamically established according to their profile and habits (e.g., a common interest in the same software updates and services). This will definitely challenge traditional group communication solutions that lack the necessary flexibility in group management and do not support effective control policies on involved endpoints (i.e., authorized senders and intended receivers). To address the cited issues, the idea of introducing new disruptive network-layer solutions has emerged from recent literature. Among them, Sociocast has been theorized as an enabler of flexible interactions between groups of devices tied by social relationships. In this paper we start from the concept of Sociocast and propose a solution based on Software Defined Networking (SDN) for its implementation at the network layer in the Internet of Things. The performance of Sociocast is studied and compared to methods running at the application layer that provide similar features. Experimental results, achieved through an emulation-based playground, confirm that the Sociocast approach allows for a significant reduction of signaling and data packets circulating in the network with respect to traditional approaches.*

Keywords – Communication Primitives, Internet of Things, Social Internet of Things, Software Defined Networks

1. INTRODUCTION

The Internet is experiencing a rapid transformation pushed by the growing need to overcome its intrinsic limitations and ossification, which challenge network practitioners and researchers. The pressing need to come to the definition of a new Internet of the future is also motivated by the multitude of Internet of Things (IoT) applications that are recently emerging in various vertical markets [1]. Such applications are increasingly characterized by *group-based* (i.e., one-to-many, many-to-many) communications established between large sets of devices in need of simultaneously exchanging data, e.g., in the case of sensors' software updates, service advertisements, device configurations.

In human-centric communications, frequent instant messaging occurs within communities of users sharing similar interests and people largely interacting with their friends, and friends of their friends, over social networks. Similarly, groups of IoT devices are likely to interact with each other, especially if they are located in the same place (e.g., sensors/actuators in the same building), are owned by the same user (e.g., consumer devices and home appliances), share similar profiles (e.g., the same brand and type), or frequently meet each other (e.g., vehicles on a given road segment).

Support of interactions between devices raises outstanding challenges for network operators. First, IoT applications require the *dynamic and flexible management* of group-based interactions, whose scope is decided according to a given topic and to the ties existing between involved endpoints (e.g., co-locality, similarity of devices, etc.).

Second, *the communication endpoints should have the power to control* data exchanges. Indeed, a control of the enabled data receivers is strongly desired by the source device, due to the potentially confidential nature of exchanged data.

Moreover, the *massive* presence of group-based communications established by billions of IoT devices, expected to increase even at a higher pace in the near future, can cause network congestion and waste device and network resources, unless proper countermeasures are taken.

A solution is required to allow nodes to flexibly specify how to *prioritize* (*filter*) the nodes from which they want (or they do not want) to receive data, and the network to react accordingly, so to prevent the threats of Denial of Service (DoS) attacks.

Conventional multicast-based approaches [2], being mainly designed to simultaneously transmit data from

one or multiple senders to a group of (unknown) receivers, fail in *natively* achieving such objectives and in ensuring the required flexibility in group establishment and management. Clumsy patches to existing multicast solutions may further complicate their design and hinder their (already limited) deployment.
This is why in [3] authors argue in favor of a novel and future-proof comprehensive solution, named *Sociocast*, encompassing both a communication method and a data delivery scheme, going well beyond Internet Protocol (IP)-based multicast. Sociocast is theorized as a means for identifying, in a flexible manner, the intended endpoints (senders/receivers) of data exchange sessions. Groups are dynamically created according to the mutual position of endpoints in a *social network of devices* and the type of relationships among them, by means of properly defined filtering rules and policies.
This work treasures the theoretical analysis in the cited vision paper and takes a significant step forward both in terms of *practical design* and *experimental evaluation*. Herein, we argue about the actual possibility of implementing the conceived Sociocast primitive as a *network-layer* solution in IoT domains, wherein switches and routers are responsible for the efficient delivery of packets issued by IoT devices. In particular, the reference network infrastructure is deployed according to the Software-Defined Networking (SDN) technology [4].
SDN has been introduced to address a typical issue in traditional IP networks: the lack of programmability in network management and configuration. Thanks to its peculiarities, it can play a crucial role to bring the social dimension into the group data delivery procedures enforced at the network layer.
The main contributions of the work can be summarized as follows.

- The design of an architectural framework encompassing all the entities and functionalities supporting Sociocast, according to a software-defined network approach.
- The definition of the main procedures for the creation of the Sociocast packets, their forwarding and filtering, and the subscription of devices to Sociocast groups.
- The performance assessment through the widely known Mininet network emulator [5], when dealing with push-based data dissemination and deploying the Sociocast network application into the ONOS SDN controller [6]. The impact of different endpoint distribution patterns and different involved social relationships on the performance is evaluated by comparing our proposal to an alternative approach where the groups are created at the application layer. Results show that the Sociocast approach allows for a reduction of signalling and data packets by a factor of 10 and 5, respectively, in the scenario where the number of recipients is high and are close to each other.

The remainder of this paper is organized as follows. In Section 2 we survey the related literature in the field of group-based communications. In Section 3 we introduce the major Sociocast concepts and discuss the design guidelines we have considered. Section 4 discusses the conceived architectural framework along with the envisioned entities and their main role and functionalities. In Section 5 we describe the main procedures to enable the treatment (i.e., forwarding, dropping, modifying) of Sociocast packets. Then, in Section 6 we describe the playground for the evaluation, before discussing achieved results in Section 7. Finally, in Section 8 we draw some concluding remarks.

2. BACKGROUND AND MOTIVATIONS

In this section we will first overview how group communications are traditionally supported in the Internet (see Section 2.1); then, we will discuss the drawbacks of such solutions (see Section 2.2); finally, the advantages of exploiting social relationships between IoT nodes are summarized (see Section 2.3).

2.1 Multicast approaches in the literature

A large number of different applications rely on one-to-many and many-to-many data traffic exchange. They range from live video streaming, audio/video conferencing [7] and multiplayer games [8] to communications between groups of servers within data centers [9] and wide-area control in smart grids [10]. Multicasting functionality is typically leveraged in such contexts, which can be performed either at the network (IP) layer or at the application layer [2], [11] and also with the support of SDN [12], [13].
IP-based multicasting. Traditional multicast routing and management protocols, such as Protocol-Independent Multicast (PIM) [14] and Internet Group Management Protocol (IGMP) [15], effectively establish and maintain multicast communication paths between sources and receivers to enable the forwarding of packets to a multicast group. Each group is assigned a unique class D IP address. A host can send data to a multicast group by using the local network multicast capability to transmit the packet. A multicast router, upon reception of a packet, looks up its routing table and forwards the packet to the appropriate outgoing interface. Group membership is managed at the network level through

routers. When a host decides to join/leave a particular multicast group, it sends the request to the local multicast router, through IGMP [15].

IP multicast allows data to be distributed in such a way that the least amount of replicas of the same packet is placed into the network.

In its recent version, v3, IGMP allows to specify the set of senders from which a node wants to receive, in agreement with the Source-Specific Multicast (SSM) protocol [16]. In other words, the only packets that are delivered to a receiver are those originating from a specific source address requested by the same receiver. Hence, SSM is particularly well-suited to dissemination-style applications with one or more senders whose identities are known before the application begins.

Non-IP multicasting. The design of multicast solutions has also been investigated beyond IP. In application-layer solutions, group membership, multicast delivery structure construction, and data forwarding are exclusively controlled by participating end hosts, thus, the support of network nodes is not needed [11].

In the clean-slate future Internet MobilityFirst architecture [17], a context-aware delivery primitive is proposed, which generalizes multicast to groups established on the basis of attribute-based descriptors. The name service, in charge of resolution procedures between global unique identifier (GUID) and network addresses, maintains a membership set that consists of all GUIDs of devices that are subscribed to the multicast group. The sender is responsible for sending data to each of the returned addresses.

SDN-based multicasting. SDN can simplify multicast traffic engineering thanks to the centralized nature of the network control plane. Current multicast solutions employ a shortest-path tree to connect the source to the receivers which is built according to local information. Traffic engineering is difficult to be supported in a shortest-path tree. By utilizing the global view of the SDN controller, in [18] all the possible routes between the sources and each host of the multicast group are calculated in advance. In contrast with IP multicast, there are no de facto standards for SDN multicast routing. Different approaches targeting different optimization objectives can be targeted in a flexible manner and it is unlikely that a given approach is going to be dominant. SDN multicast is enabled by writing an application for the SDN controller that optimizes the traffic flows to meet the particular needs of the end user [12]. The SDN controller can build the multicast tree to meet link constraints (bandwidth consumption) or path constraints (end-to-end delay) [13]. Hence, it is a valuable solution when Quality of Service (QoS) requirements need to be ensured to a multicast flow, e.g., in the case of a multi-party video-conferencing service [19].

2.2 Weaknesses and open issues

The use of the traditional IP multicast is prone to multiple issues:

- Without the explicit join to the multicast group, a router will not forward multicast IP packets destined to end hosts. This process implies the distribution of the consent to join the multicast group among devices, increasing the signaling overhead.
- There is no way for the sender to control who subscribes to a multicast group.
- It prevents the creation of discrimination policies based on the destinations of the information within the same multicast group. Therefore, when a limitation to the distribution of packets to some entities of the same multicast group is needed, another multicast group must be created, with a consequent increase in the number of signaling packets in the network.
- All routers must be replaced with multicast-enabled routers, which could be expensive and hardly viable for the network operator, raising interoperability issues.

The poor flexibility of the IP-based multicast discourages the pursuit of such an approach for the wide variety of sender-initiated dynamic group-based communications, as demanded by future IoT deployments.
On the other hand, application-layer solutions have the drawback of a definitely worse performance in terms of end-to-end latency and efficiency compared to IP multicast. This is because end hosts have little or no knowledge of the underlying network topology.
Thanks to its programmability and global knowledge of the topology, SDN can make more efficient the creation of the multicast tree, improving forwarding procedures. However, to the best of our knowledge, the flexibility of SDN has not been investigated to manage dynamic group formation.
These issues have motivated the theorizing of a new communication method and data delivery scheme [3], able to better fit the nature of upcoming group-based communications: *Sociocast*.
This is introduced as a novel and flexible solution that allows *group-based communications in the IoT enhanced with the notion of social ties*. It inherits the strengths of IP multicast, in that it lets network nodes disseminate packets in an efficient manner: sociocast packets are assigned an IP address to facilitate their forwarding. In

addition, the proposal in [3] enables a *mutual control* of the endpoints. Not only the receiver can filter different senders, as in SSM, but also the sender can (implicitly) decide which node should belong to the set of intended receivers, by specifying the features (in terms of social relationship) of such receivers. The above capabilities are disruptive when compared to conventional IP-based multicast. Sociocast relieves the burden of group management from network nodes and of explicit join procedures from devices.
Moreover, SDN is chosen to facilitate the implementation of multicast groups with a social flavour directly at the network layer.

2.3 Advantages of a "social-oriented" approach

The use of social links to support network functionality is not new in the telecommunications landscape.
Several routing protocols in wireless ad hoc [20], mobile opportunistic and delay-tolerant networks [21, 22, 23, 24], have been designed to build upon the key concepts of social network analysis, i.e., *small world phenomenon* [25] and *centrality*. The former one, a.k.a. *community*, captures the fact that actors within a social network are separated from each other by an average number of fairly *limited hops*. The latter one shows that *some nodes in a community are the common acquaintances of other nodes*.
In the aforementioned works, the knowledge of social characteristics (e.g., node centrality, in-betweenness) is used to make better forwarding decisions and assist the relay selection when delivering data to the intended destination(s).
Many of the studied approaches involved unicast or multicast communications [26, 27, 28]. The issue of data broadcasting in a Mobile Social Network, where mobile social users physically interact with each other, is analyzed in [29].
The objective of this work is to exploit similar concepts but under a different perspective. We aim not to improve forwarding decisions by leveraging social network properties, but to better disseminate data at the network layer within dynamically created groups of socially connected devices.
The proposal has the potential of a real game changer in view of the creation of the future Internet of Things, by providing superior advantages compared to what has been done so far in the literature.
In fact, social bonds not only ensure minimum separation distances between actors, crucial for efficient and fast data propagation, but may enable data exchange within trusted groups and creation of groups that include actors belonging to different communities. In Sociocast this translates into the possibility of efficient and flexible group end-points discovery, an intrinsic possibility of implementing policies for creating trusted groups of end-points directly at the network level, and the ability to effectively and simply deal with the problem of interoperability among different IoT platforms.
Obviously, to do this we need to start from a paradigm that can provide for the establishment of pseudo-social ties between devices (to operate at the network layer). This is already available in solutions of "social networks of IoT devices", such as the Social Internet of Things (SIoT) [30] for example. However, they need to be moved from the application layer, wherein they have been initially conceived, down to the control plane of the network layer. In so doing, group establishment and data exchange among members of such groups can be managed in a tighter way, with inherent flexibility and efficiency in terms of network resource usage.

3. SOCIOCAST: OBJECTIVES AND DESIGN PRINCIPLES

In this Section we describe how we can achieve a real implementation of the Sociocast concept by relying on the capabilities of the Software Defined Networking paradigm. The resulting solution is an enabler for group communications based on social notions at the network layer.
Social ties among devices. Devices are likely to interact with other devices with similar profiles and habits, e.g., those located in the same place, owned by the same user, produced in the same company branch.
Such ties are well captured by the SIoT paradigm in [30]. There, a few basic types of social relationships, defined according to user-defined policies, are introduced: *co-ownership object relationship* (OOR), created between devices that belong to the same owner; *co-location object relationship* (CLOR), created between stationary devices located in the same place; *parental object relationship* (POR), created between devices of the same model, producer and production batch; *co-work object relationship* (CWOR), created between moving devices that meet each other at the owners' workplace; *social object relationship* (SOR), created as a consequence of frequent interactions between moving devices. The framework is quite flexible and other types of relationships can be easily added on a per use-case basis.
Applications requiring data dissemination to a social group of devices are, for instance, *software updates*: a given software patch needs to be safely delivered to all the devices or sensors of the same brand, model, batch. For this, POR relationships should be exploited. Similarly, some data needs to reach all other devices belonging to the same owner in the case of *personal bubbles*: the

OOR relationship is appropriate in this scenario. Business services may be advertised to all devices that either are currently in the same area (CLOR) or often visited the same place (SOR).

Targeted data delivery schemes. Sociocast aims to enable:

- a given sender to disseminate data in a *push*-like manner to specific nodes, which are friends over a social network of devices, according to properly defined filters and policies (i.e., the social relationship type);
- a node *to subscribe* to specific social-based topics (i.e., to receive data from friends of a given type);
- a node *to prioritize* (and not to receive) data from particular senders, e.g.: to enforce QoS; to identify the more suited and trusted communication endpoints for security reasons; to save resources.

Deployment options. To target the aforementioned objectives, the envisioned framework has *(i)* to enable nodes to indicate in an agile manner the features of the endpoints of data flows (i.e., the set of intended recipients and/or the authorized senders) based on the distance in a social network of devices, *(ii)* to properly and dynamically identify them, *(iii)* to forward data packets accordingly.

A straightforward approach to accomplish the first two features could be to rely on an *application-layer* solution. For instance, the intended set of receivers can be specified by a given sender at a high-level, e.g., through metadata. Then, the resulting request can be sent to a purpose-built proxy which is in charge of mapping such data onto IP addresses of the receivers, similar to [17]. Despite the virtue of simplicity, such an approach has the drawback of poor performance in terms of efficiency in the usage of network resources, since data forwarding to each intended destination is performed at the underlying network layer in a myopic manner.

Thus, our interest is on a *network-layer* approach, according to which the features of the intended set of receivers of a given data packet (or of a sender of unwanted data packets) can be translated into a network-layer IP address, hence treated (forwarded/dropped) by network nodes, accordingly. The type of proposed approach is inspired by the traditional IP-based multicast, with which it shares a few aspects, such as the routing of packets with a multicast address (a Sociocast address, in our case). However, multicast lacks the flexibility necessary to implement the aforementioned critical functionalities for the future Internet of billions of devices, while meeting the requirements of the end users and those of the network operators.

By overstepping the agnosticism about Sociocast traffic at the network layer, the following advantages are expected:

- data forwarding can occur in an efficient manner, e.g., by reducing the number of duplicated packets, and saving bandwidth accordingly;
- filtering procedures can be enforced in-network, as requested by potential data recipients, to limit the massive amount of generated traffic;
- network operators can benefit from traffic reduction, which is particularly crucial for their infrastructures expected to be largely overwhelmed in the near future.

Programmable packet treatment. Recent advancements in networking technologies make the deployment of Sociocast at the network layer even more viable. We identify SDN as the key enabler for Sociocast. Thanks to its programmability, which reduces the complexity of network elements, SDN can inject forwarding/dropping rules and properly manipulate headers of packets to make more efficient their forwarding.

Such policies can be defined in a network application, with no need to modify the data plane of the underlying network infrastructure.

4. THE ARCHITECTURAL FRAMEWORK

The main entities of the envisioned framework are: the *Sociocast nodes*, the *SDN network* (encompassing both switches and controller), augmented with the notion of Sociocast, and the *Sociocast Relationship Service*, as shown in Figure 1 and detailed in the following sections.

4.1 The Sociocast nodes

The *Sociocast nodes* are the endpoints of a Sociocast communication. They are legacy IoT devices (e.g., smartphones, sensors) augmented with the *Sociocast Support Layer* (ScSL) running on top of the transport layer, through which they are enabled to create, send and/or receive Sociocast packets. The ScSL exposes the Sociocast Application Programming Interfaces (APIs) to the applications that want to use the Sociocast communication configuration for data delivery. It is through this layer that Sociocast packets are created and received by the end devices.

4.2 The SDN network

The SDN network is composed of three different planes, according to the legacy deployment.

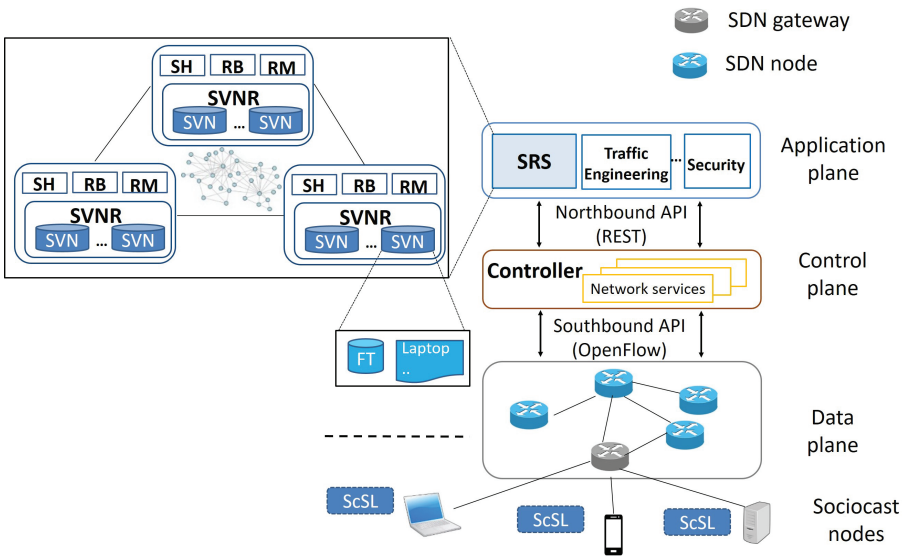


Fig. 1 – Sociocast architectural framework.

The *data plane* encompasses the *SDN switches*. They are SDN-enabled network nodes which are connected to each other and interact with the SDN controller. Between them, the *SDN gateways* are the ingress/egress nodes of the SDN network. SDN nodes interact with the SDN controller through the OpenFlow (OF) southbound interface.
The *control plane* includes the *SDN controller*, which oversees the SDN nodes, according to specific orchestration policies defined at the *application plane*. It tracks the graph of the network topology in the *Network Information Base* (NIB). According to information in the NIB and policies defined by network applications, it injects rules in the so-called *flow tables* of SDN nodes to enable the forwarding of sociocast packets through OF messages [31].

4.3 The Sociocast Relationship Service

The Sociocast Relationship Service (SRS) is implemented at the application plane, next to conventional SDN applications, and it provides the following main functionalities:

1. establishing social relationships between nodes. Without loss of generality, we inherit concepts and methodologies regarding the policies for the establishment of the social links between nodes from the well-accepted SIoT paradigm [30];

2. keeping track of the established social relationships;

3. providing interfaces towards the SDN network and to navigate the social network so to identify the nodes that belong to the set of the potential recipients/authorized senders of a Sociocast packet.

Herein, a major element is the *Social Virtual Node (SVN)*, which represents the digital counterpart of a physical device. It stores some metadata providing information about the nature of the device and a list of friends, which is organized in a table named *Friends Table*. For each friend in the table, the SVN records the type(s) of friendship(s), defined according to the SIoT paradigm and the trust level associated with each friend. The *Social Virtual Node Repository (SVNR)* stores all SVNs associated to the physical devices in a given area. Indeed, one SVNR is responsible for provididing the described services for the objects in a given area; more SVNRs are then interconnected in a distributed system. The following modules are associated to the SVNR.

- The *Relationship Manager* (RM) is responsible for the relationships' lifecyle management, i.e., detecting, creating, updating and deleting relationships[1].

- The *Relationship Browser* (RB) navigates the Friends Table to find potential recipients of a Sociocast packet, according to their position in the social network. Policies for the social network navigation are discussed in [32].

- The *Sociocast Handler (SH)*, whenever queried by the SDN controller, provides the members of a Sociocast group, after querying the RB module,

[1]For a detailed description of relationships management, the reader is referred to [32].

through a Representational State Transfer (REST) API.

SVNRs, along with relevant functionalities (i.e., RM, RB and SH), can be deployed as a peer-to-peer system, for instance building upon the one described in [32].
Our design choice is aimed at providing an implementation of SDN-based group communications based on a de-facto global IoT resource directory, which is distributed and without a single player in control of the system. Digital representations of physical IoT devices will run in distributed servers and can create autonomously social-like relationships with each other. Based on such a distributed resource directory, interactions (both point-to-point and point-to-multipoint) between IoT resources belonging to different platforms can be straightforwardly enabled. Each SVNR (or group of SVNRs) could, in fact, contain the images of the devices belonging to a given platform, it can be owned and maintained by the owner of the platform (or even the owner of the group of IoT devices), and it interacts in a peer-to-peer fashion with other SVNRs constituting the SRS.

5. SOCIOCAST IN ACTION

In the following, we detail the main steps for the creation of a Sociocast packet. Then, we describe Sociocast data delivery according to a push-based dissemination, publish/subscribe procedures to sociocast groups, as well as filtering according to sociocast rules.

5.1 Creating a Sociocast packet

A Sociocast packet is created whenever a device needs the services offered by the Sociocast framework, which are intended to: *(i)* disseminate data in a push-like manner; *(ii)* indicate the subscription to a Sociocast group; *(iii)* or to filter/prioritize data from particular senders. Whenever a packet is created, it has to indicate which one of these three types of service is requested. The above are the types of Sociocast services supported in the current implementation, but the set of Sociocast services can be easily extended in the future.
Let us consider a device, say A, which creates a packet with data to be sent to a Sociocast group. The application in A makes a request to the ScSL via the available APIs, providing the following information: *(i)* the type of requested Sociocast service; *(ii)* the social relationship (e.g., OOR, CLOR) according to which the Sociocast group has to be formed; *(iii)* the social distance (number of hops over the social network), which represents the scope of the Sociocast group.
The ScSL reacts to the incoming request by creating an IP packet with the following header fields:

- Source IP address: the IP address of the source device.
- Destination IP address: a fixed IP address, identified in this paper as IP_{SC}, assigned to Sociocast that allows SDN gateways to identify Sociocast packets.
- Sociocast Tag: a 2-bytes field that is carried inside the transport-layer destination port and is used to uniquely identify the type of social relationship and other appropriate filters (e.g., number of hops, possible application of Sociocast, etc.). The encoding is as follows:
  - Metadata: device metadata available for future applications.
  - RelationFilter: type of relationship (e.g., OOR, SOR, CLOR, etc.).
  - FeatureGroup: type of Sociocast services needed by the application (e.g., GroupCreation, SourceFiltering, Pub/Sub).
  - Radius: maximum distance, in number of hops, from the source.

  Fig. 2 shows some examples of Sociocast Tag configuration.

Being Sociocast packets identified through conventional layers 3 and 4 header fields, legacy matching rules can be applied, with no need to resort to OF *experimenter fields* [33]. Such design choices would facilitate the deployment of Sociocast, which candidates itself as a short-term solution to be exploited by network operators.
For the sake of simplicity, the encoding described above refers to the case where the IPv4 is used. Similar considerations hold for IPv6 packets, for which matching fields can be handled by OF since version 1.2 [33].
For those constrained IoT devices belonging to Low power and Lossy Networks (LLNs), 6LoWPAN (IPv6 over Low-Power Wireless Personal Area Networks) header compression methods can be used [34] over the link interconnecting the devices to the SDN gateway. For the IPv6 headers, compression methods may also affect source and destination addresses, and they vary according to the fact that the source is communicating with nodes either within or outside the WPAN. In the latter case, a 50 percent compression ratio can be still achieved by letting the full destination address, carrying the Sociocast address, be transmitted.
TCP header compression for IoT scenarios [35] is still an open issue at the standardization level [36], not part of RFC 6282 [34]. The compression foresees to avoid sending the port numbers in each packet, which however does not affect the Sociocast communications as

the port number with the SOCIOCAST TAG is reconstructed at the gateway. Indeed, decompression occurs at the SDN gateway letting Sociocast packets travel with conventional IP header fields in the SDN network. Similar operations are performed at the SDN gateways the destinations are attached to, if the latter ones belong to a WPAN.

5.2 Push-based data dissemination

Once the Sociocast packet is created with data to be disseminated, it is sent by the source device and treated in the network through the following steps.

1. The Sociocast packet reaches the SDN gateway, which the source device is connected to. Since, initially, a forwarding rule is not set in the flow table of the SDN gateway, the GOTOCONTROLLER rule applies for it. Hence, a OF PACKET_IN message is issued to be transmitted to the SDN controller.

2. Upon reading the header of the Sociocast packet[2], the SDN controller realizes that a Sociocast group must be created (FEATURE field set to GROUPCREATION). Thus, it issues a request to the SRS, to retrieve the set of devices, intended to act as recipients of the Sociocast packet.

3. The SH triggers the browsing of the social network, as specified before, and returns to the SDN controller the addresses of the set of devices of the Sociocast group.

4. The SDN controller retrieves from the NIB the SDN nodes in the shortest paths towards the intended receivers of the Sociocast group. Then, it builds the routing paths by ensuring that SDN nodes belonging to the path towards multiple receivers receive a single rule and forward the Sociocast packet only once. Hence, it injects forwarding rules in the flow table of involved SDN nodes accordingly, by sending OF FLOW_MOD messages. In particular, the SDN gateways which the Sociocast destinations are attached to, will be instructed by the SDN controller with a rule that: *(i)* matches the Sociocast-related header fields that identify the Sociocast communication and *(ii)* foresees to forward the packet to the correct physical port after changing the destination Sociocast IP address with the IP destination (unicast) address as action. This is to ensure that all devices belonging to the Sociocast group correctly receive the Sociocast packet. Other SDN nodes, instead, are instructed to forward the Sociocast packet to the physical correct ports by matching the Sociocast fields values.

[2]The entire Sociocast packet is sent by the SDN gateway, hence a PACKET_OUT is transmitted by the controller, back to the SDN gateway [33].

Once the Sociocast group is created, subsequent Sociocast packets transmitted by the source device may be handled by the SDN gateway with no need to contact the SDN controller, but rather forwarded according to rules already available in the flow table. According to the legacy SDN implementation, a timeout is applied to rules injected by the controller into SDN nodes, to prevent a rule to stay in the table for a long time and unnecessarily occupy space in the flow table [33]. Within our framework, such a timeout can be set to reflect the lifetime and frequency of interactions within the Sociocast group, the mobility patterns of nodes.

5.3 Publish/subscribe

Sociocast can be exploited to support a publish/subscribe interaction model as well. In fact, a device can *subscribe* to receive packets *published* by devices identified by their position in the social network. For example, assume that device B wants to subscribe to receive packets generated by its friends of type OOR. If this is the case, it will generate a Sociocast packet with the FEATUREGROUP field set to PUB/SUB and the RELATIONFILTER field identifying an OOR.
Such an information will reach the SDN controller which will perform the following operations:

1. It sends a query to the SH and receives the identities of the devices with position in the social network consistent with the request by device B.

2. It adds this information in a pending interest table which tracks all subscriptions received by devices. Whenever a device begins to disseminate data, the SDN controller will check whether there are devices that have subscribed to its updates (e.g., B).

3. If this is the case, the SDN controller will instruct the SDN nodes in the path to B to forward the data packets to it.

5.4 Source filtering

Sociocast allows a device to select those that are entitled to send packets to it, based on their position in the social network. Such a feature can be used both in a proactive and a reactive way. More specifically,

- *Proactive*: a device might decide to receive packets by its *friends* only, for security reasons or to save energy, computational and communication resources.

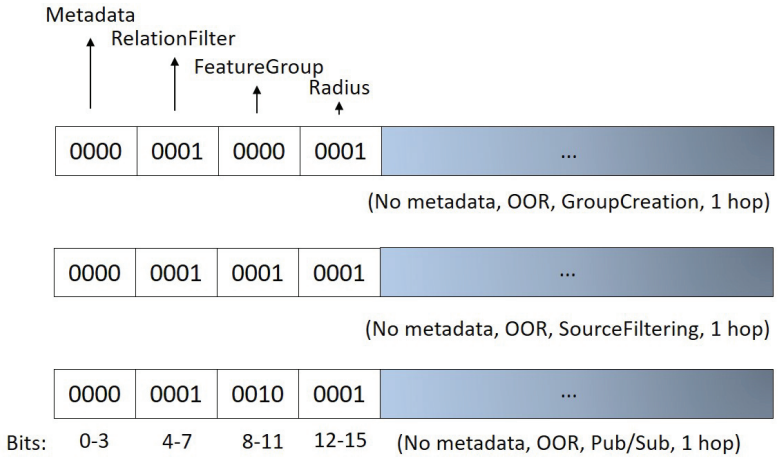


Fig. 2 – Examples of Sociocast Tag configuration.

- *Reactive*: the computational or communication load for a device may become too high, e.g., because of a DoS attack. If this is the case, the device might decide to accept packets by a subset of devices, based on their position in the social network. In this way Sociocast can be exploited to realize a firewall the policies of which change depending on the current load.

A device, say C, wishing not to receive packets from nodes with certain social properties sends a Sociocast packet by specifying in the FeatureGroup field SourceFiltering. Once the packet reaches the SDN controller, the latter one will query the SH, which will reply with the list of authorized IP addresses. Accordingly, the SDN controller will insert entries in the flow table of the SDN gateway which C is attached to, to specify the forwarding rule for packets destined to it sent from authorized senders and the dropping rule for those which are not allowed.

6. PERFORMANCE EVALUATION

In this section we describe the environment for the performance evaluation. More specifically, in Section 6.1 we describe the tools utilized for the performance evaluation, and in Section 6.2 we discuss the scenarios. The benchmark utilized for comparison purposes is presented in Section 6.3, whereas the considered performance metrics are identified in Section 6.4.

6.1 Tools and reference topology

The focus of the performance evaluation is to assess Sociocast in the case of *push-based data dissemination towards a group of devices*.

To this purpose, we built an emulation playground. In particular, the Mininet network emulator [5] has been used, it allows the creation of a network with thousands of nodes on the limited resources of a single (virtual) machine. In particular, it enables fast prototyping and experimental evaluation of OF-enabled networked systems. The experimental setting consists of the network topology depicted in Fig. 3. A full-mesh interconnects the core SDN nodes, which are the roots of a three-layers fat-tree topology. Up to 21 devices are attached to each SDN gateway (not all the devices are shown in the figure). ONOS has been considered as a reference SDN controller in the context of this work, due to its scalability properties and its highly modular architecture [6].

The ONOS controller interacts with an external SRS, which establishes social relationships among emulated devices, and manages them.

The ONOS controller and the Mininet network emulator are both running on the same virtual machine, while the SRS runs in a different one. Both these virtual machines are located in a physical server with an Intel Xeon(R) CPU E5-2630C v3 1.80 GHz x32 processor and 377,8 GiB of memory.

6.2 Social relationships settings and traffic patterns

The performance of the proposed solution has been evaluated with a set of representative IoT test configurations properly designed to take into account different numbers and distributions of nodes in the emulated topology, different physical distances between sources and destinations, and different types of service. This is aimed at making the obtained results as generalizable as possible and having a clear idea of the potential and limits of

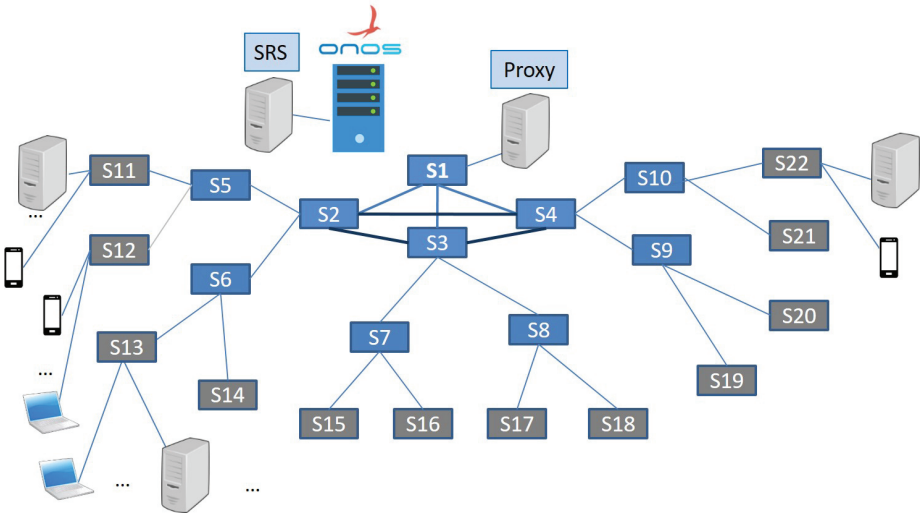


Fig. 3 – Reference topology.

Sociocast in multiple scenarios. Each of the test configurations has been mapped onto a use case characterized by the exploitation of a particular type of social relationship between the devices involved. In this way, helpful guidelines can be provided about the suitability of the proposed solution in the context of different application scenarios and, at the same time, of the effectiveness of communications based on each of the possible social-like relationships established among IoT devices. Details are given in the following subsections. Table 1 also summarizes the major features characterizing each scenario, which are: the types of social relationship (shortened as Rel.), the number of destinations (shortened as DSTs) for each communication, their distance from the source (shortened as SRC), and their position with reference to the considered network topology.

6.2.1 Scenario A: Smart industrial plant

Group communication needs: an industrial plant is equipped with several connected devices (sensors and actuators) and one of these (randomly selected) belonging to the emulated topology issues a Sociocast packet destined to *all the devices connected to the same gateway*. The group can be created, for instance, for the dissemination of alarms, for group configuration and reconfiguration, for functional testing.

Involved relationship type: CLOR.

Endpoint distribution profile: all endpoints clustered in the same area.

6.2.2 Scenario B: Smart home monitoring

Group communication needs: a randomly selected device in the emulated topology, resembling a smartphone of a user currently at office, acts as a sender and issues a Sociocast packet to create a group of recipients made of *all the smart devices connected to the (same) home gateway, which is different from the one the user's smartphone is attached to*. The group can be created, for instance, to notify devices to configure a warm welcome for the user.
Involved relationship type: OOR (ownership).
Endpoint distribution profile: sender in a location and all destinations clustered in a different (potentially) remote location.

6.2.3 Scenario C: Wireless Sensor Network (WSN) management

Group communication needs: a randomly selected device in the emulated topology acts as a sender and issues a Sociocast group creation destined to *all the devices of the same brand, uniformly distributed in the topology* to disseminate a new configuration for the device, a software update, or a new driver version.
Involved relationship type: POR (parental).
Endpoint distribution profile: uniform distribution of endpoints.

6.2.4 Scenario D: Smart mobility

Group communication needs: we assume mobile devices (e.g., smartphones, laptops) carried by people moving in a *smart city/smart campus* and interacting with

other devices met either in the neighborhood or close offices/classrooms. The type of the data exchanged within the group includes: information related to mobility applications, tourist information, data for the implementation of any Intelligent Transportation Systems application.
Involved relationship type: SOR.
Endpoint distribution profile: variable location of endpoints in the group.

6.2.5 Relationships creation

As to the creation of the relationships, these have been set in deterministic way except for the SOR. In particular, different groups of devices linked with POR, OOR, and POR relationships are created so as to have from 5 to 20 recipients for each simulated communication. However, the CLOR relationship has been created between devices that are connected to the same gateway as the co-location has to be assured. As to the SOR relationships used in Scenario D, these are established between devices in the emulated topology according to their physical distance and follow a simple probabilistic model. The principle adopted is such that the closer the devices, the higher the probability that the two devices have established a SOR relationship. Accordingly, devices attached to the same SDN gateway (i.e., an Access Point) have the highest probability to establish it. These devices are characterized by sharing the same path to reach the root node ($s1$ in Figure 3), which is made of 4 SDN nodes. We base on this number to define the notation to denote the relevant probability to create a SOR between them: $p_{soc,4}$. Following the same principle, devices sharing three, two, or one SDN nodes in the path to reach $s1$, establish a relationship with probability $p_{soc,3}$, $p_{soc,2}$, and $p_{soc,1}$, respectively. The higher j the higher the probability $p_{soc,j}$, with $j \in \{1, 2, 3, 4\}$. The setting of $p_{soc,j}$ used in the performed simulations is reported in Table 2; different configurations have been considered to evaluate the impact of different numbers of friends and their distribution in the considered topology.

6.3 Benchmark scheme

The performance of the proposal has been compared against an application-layer solution we refer to as *multiple unicast* (labeled in the plots as *M-Unicast*). Note that also for this benchmark scheme, we are focusing on the push-based data dissemination scenario. The choice of this benchmark is meant to *quantitatively* estimate the benefits of the Sociocast proposal against an application-layer solution. In the latter one, the network layer is agnostic about the communicating group, but it offers the same features in terms of sender-initiated and dynamic Sociocast group creation, hence ensuring a fair comparison. Specifically, the source node contacts a proxy in charge of interacting with a SIoT-like platform to get the set of intended destinations belonging to the Sociocast group. The latter one is described through attributes/metadata defined at the application layer, similarly to the information encoded in the tags in Sociocast packets. After retrieving the list, the proxy forwards it to the source node which sends the packet to the destinations through multiple unicast exchanges. In other words, the controller sets up distinct routing paths for each destination and some links can be shared by multiple paths towards destinations belonging to the same group. Without losing generality, we assume that the proxy is attached to the root node of the topology (i.e., $s1$ in Fig. 3).

6.4 Metrics

The following metrics have been considered to evaluate the performance of the compared schemes in the creation of a Sociocast group and data exchange among its members:

- the *number of OF signaling packets* exchanged between SDN nodes and controller to build routing paths towards the intended Sociocast destinations. The metric only refers to the control packets exchanged to process incoming requests from sociocast nodes at the SDN gateway, namely PACKET_IN, PACKET_OUT and FLOW_MOD. The background (periodic) signaling exchanged between the controller and the SDN nodes is not considered;

- the *number of data packets* exchanged into the network to reach all the intended destinations of the communicating group, once it has been created; the metric considers the number of transmitted packets per link and are represented by either Sociocast or M-Unicast packets.

For the benchmark scheme, the request packets issued by the source towards the proxy as well as the signaling messages required to instruct the relevant SDN nodes towards it are also considered.
The above metrics have been measured through the well-known Wireshark network protocol analyzer[3].
Comparison experiments have been conducted when varying the number of destinations (or relevant probability settings) and are averaged over 20 runs.

[3]Please notice that the analysis of the signaling incurred for the creation of social relationships between devices is outside the scope of this paper and is peculiar of the conceived SIoT implementation. An interested reader is referred to [37].

Table 1 – Summary of the main social relationships settings.

Scenario	Use case	Rel.	#DSTs	SRC-DSTs Distance	Position of DSTs
A	Smart industry	CLOR	5-20	1 hop for all destinations	Attached to the same SDN gateway (= sender)
B	Smart home	OOR	5-20	Fixed for a given set of destinations	Attached to the same SDN gateway (≠ sender)
C	WSN management	POR	5-20	1-7 hops	Uniformly distributed in the topology
D	Smart mobility	SOR	*	*	*

Table 2 – Probabilities of SOR establishment.

Sim-ID	#Destinations	$p_{soc,1}$	$p_{soc,2}$	$p_{soc,3}$	$p_{soc,4}$
1	4.5	0.1	0.2	0.3	0.4
2	5.5	0.2	0.3	0.4	0.5
3	8.4	0.3	0.4	0.5	0.6
4	11.2	0.4	0.5	0.6	0.7
5	12.7	0.5	0.6	0.7	0.8
6	15.3	0.6	0.7	0.8	0.9
7	18.4	0.7	0.8	0.9	1

7. EXPERIMENTAL RESULTS

In this section we show the performance results. More specifically, in Section 7.1 we assess Sociocast in terms of generated signaling packets; whereas in Section 7.2 we will focus on data packets.

7.1 Signaling packets

The first set of results aims to analyze the control plane signaling footprint incurred by the proposal w.r.t. the benchmark scheme. Fig. 4 reports the number of exchanged OF packets when varying the number of destinations of the Sociocast group for the scenarios A-C, whereas the results for scenario D are shown in Fig. 5(a). It can be clearly observed that for the M-Unicast approach the metric significantly increases with the number of destinations, in all the considered scenarios. Such a trend is due to the fact that the end-to-end communication path towards each *single* destination needs to be discovered with the support of the SDN controller. In other words, an SDN node receives a number of M-Unicast packets to forward equal to the number of destinations it allows to reach. For each of them, it contacts the controller by generating a Packet_In message and waits for the corresponding Packet_Out and Flow_Mod with instructions about the forwarding behaviour.

For a given number of destinations, the highest number of OF packets are exchanged in case of Scenario C. In the latter one, indeed, the destinations are spread over the topology and the routing path towards them may involve several SDN nodes (and gateways). Scenario B follows with a lower number of exchanged OF packets. In Scenario A, instead, only a single SDN gateway is in charge of Sociocast packet forwarding. It is the only SDN node transmitting and receiving OF packets.

In the proposed Sociocast solution, the controller is in charge of building routing paths towards them so to avoid the forwarding of the same Sociocast packet over the same link.

Hence, unlike the benchmark scheme, in our proposal, those SDN nodes which belong to the paths towards different destinations receive only a single Sociocast packet to forward and a single FLOW_MOD from the controller. The gain of Sociocast w.r.t. M-Unicast in terms of exchanged OF packets gets more remarkable as the number of destinations increases. For instance, in Scenario C, it passes from a factor of around 6 for five destinations to a factor of more than 14 for twenty destinations.

It is worth observing that, in Sociocast, a single FLOW_MOD message may convey multiple rules to be injected into an SDN node. In particular, Table 3 reports, for Scenario A, the size of the FLOW_MOD message, as measured at the SDN gateway, which the source and the destinations are both attached to. For the Sociocast proposal, the size reasonably increases with the number of destinations to accommodate the action rule for each of them. The rule specifies the physical output port as well as the change of the IP address from Sociocast to unicast. For M-Unicast, each FLOW_MOD carries a single rule, since its injection is issued per each M-Unicast packet traversing an SDN node. The size increases of less than a factor of 3 for the Sociocast approach compared to M-Unicast, in the case of twenty destinations.

Despite the larger size of FLOW_MOD packets, it can be easily inferred that, overall, the OF signaling footprint of the proposal, in terms of number of exchanged bytes, is significantly lower than M-Unicast. Also, the proposal better scales with the size of the Sociocast group.

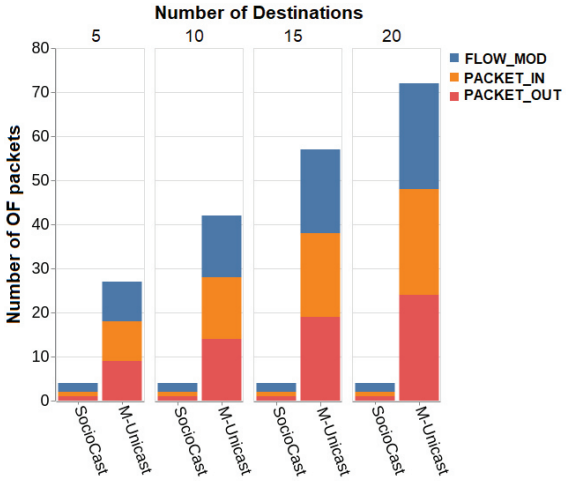
Similar to the benchmark scheme, the proposal experiences the largest signaling in Scenario C, wherein multiple SDN nodes, involved in forwarding Sociocast packets to destinations, spread over the topology, need to be instructed.

Similar considerations hold for Scenario D, Fig. 5(a). Also in such a case, the proposed Sociocast solution is less sensitive to the simulation settings (i.e., size of Sociocast group and its configuration in terms of proximity of destinations w.r.t. the source) than the benchmark.

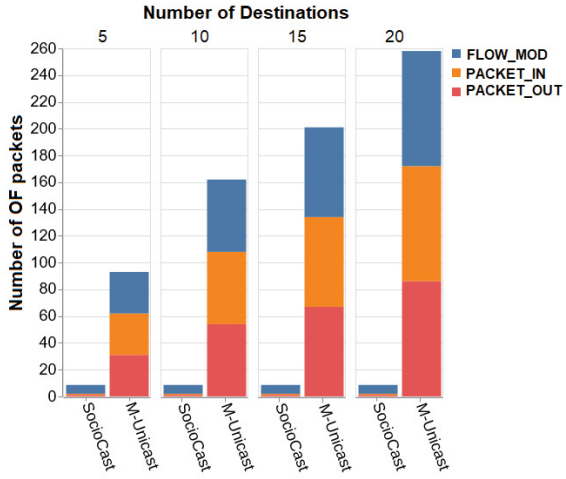
7.2 Data packets

Results in Fig. 6 shed further light into the efficiency of the compared schemes in delivering the data packets. Similar to the OF signaling, also the number of exchanged Sociocast packets increases with the number of destinations; the highest values are experienced for Scenario C and the lowest ones in Scenario A.

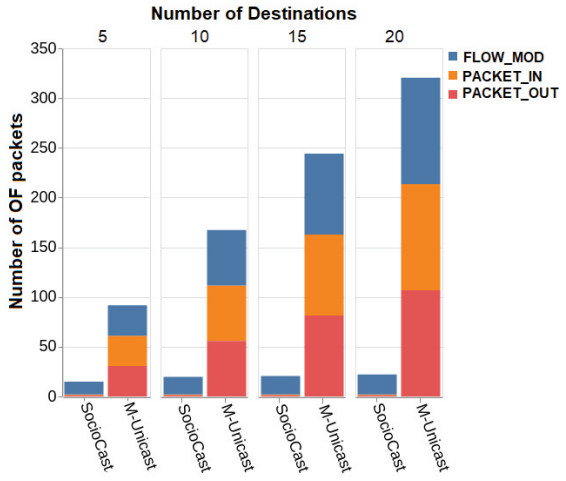
As a general remark, it can be observed that the proposal is less sensitive to increases in the number of destinations when compared to the benchmark. This happens because the controller builds the routing paths to avoid that packets are redundantly transmitted over a given



(a) Scenario A

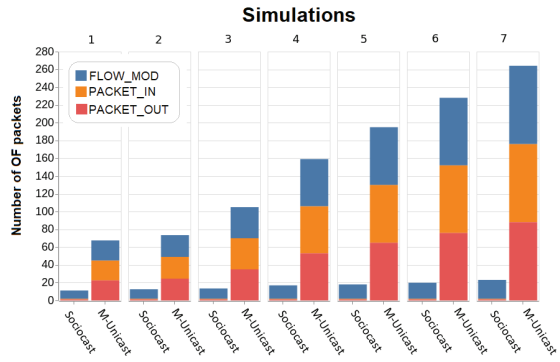


(b) Scenario B

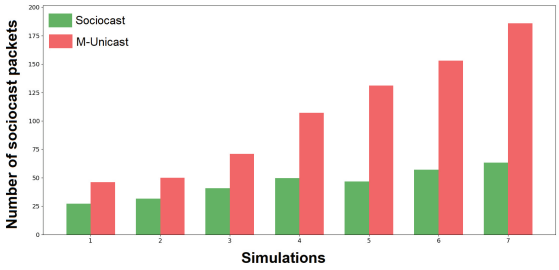


(c) Scenario C

Fig. 4 – Sociocast Vs. Multiple Unicast: Exchanged OF packets for Sociocast group creation when varying the number of destinations, in different scenarios.



(a) OF packets exchanged to create the Sociocast group



(b) Exchanged data plane packets in the topology

Fig. 5 – Sociocast Vs. Multiple Unicast: performance metrics under different simulation settings for Scenario D. The different simulation runs (from 1 to 7) correspond to the different Sim-IDs of Table 2.

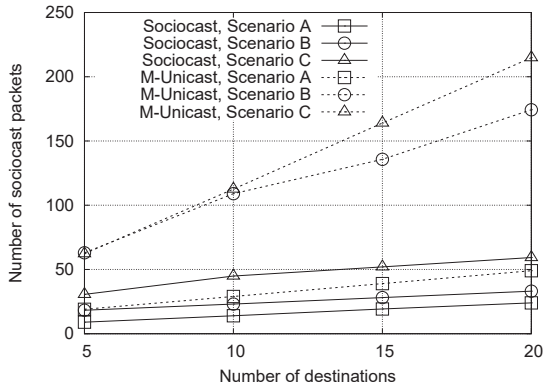


Fig. 6 – Sociocast Vs. Multiple Unicast: Exchanged data plane packets when varying the number of receivers, in different scenarios.

Table 3 – Size (in bytes) of the FLOW_MOD packet for Scenario A.

#Destinations	M-Unicast	Sociocast
1	172	172
5	172	252
10	172	332
15	172	412
20	172	492

link shared by more destinations.
This is not the case for the M-Unicast solution where forwarding decisions are separately taken for each data packet, according to the address of the intended destination.
When referring to Scenario A, the M-Unicast approach always sends twice as many data packets as the proposal. This is an obvious consequence of the fact that, after receiving the destinations list, for the M-Unicast approach there are two packets, for each destination, traveling into the topology. One packet travels from the source to the SDN gateway, and the other one from the SDN gateway to the corresponding destination. This does not apply for the Sociocast approach, where there is only the data packet from the SDN gateway to each destination.
Improvements get larger for other scenarios.
In Scenario B, more SDN nodes are involved in the routing path, despite the fact that all the destinations are connected to the same SDN gateway. Hence, more data packets travel into the network, especially for the M-Unicast solution. Such a trend is more remarkable in Scenario C, due to the larger spread of destinations over the topology. A similar trend is observed for Scenario D in Fig. 5(b).
Not surprisingly, improvements of Sociocast w.r.t. M-Unicast are greater in Scenario B compared to Scenario C. Indeed, in Scenario B the path towards all intended destinations is the same from the source to the SDN gateway. Hence, in Sociocast, the SDN controller judiciously issues rules that prevent from forwarding duplicated packets over the same links.

8. DISCUSSION AND CONCLUSION

In this paper we have proposed and analyzed the behaviour of an architectural framework encompassing all the entities, functionalities, and procedures that support a fresh new network-layer group dissemination method, i.e., Sociocast, by leveraging a software-defined network approach.
Results achieved through an emulation testbed show the better scalability of the proposal in terms of OF signaling and data packet redundancy when compared to an application-layer benchmark scheme, under different representative IoT scenarios.

Improvements are achieved by leveraging a purpose-built network application in the controller *(i)* in charge of identifying the set of Sociocast destinations by interacting with an external SIoT platform (feature implemented at the application layer by the benchmark scheme) and *(i)* responsible for smartly building routing paths towards multiple receivers so as to avoid packet duplication over links. SDN allows to manage the implementation of such functionalities at the control plane in a flexible and programmable manner, with no changes in the forwarding elements, hence making the devised framework practically viable at a low implementation cost.

Benefits of the proposal are definitely large when big groups of destination devices are clustered together, as witnessed by results referring to Scenario B: the OF signaling is reduced by a factor higher than 10 and the number of exchanged data packets shrinks by more than a factor of 5 (for twenty destinations). The lower gains for Sociocast packets w.r.t. OF signaling are due to the fact that Sociocast also resorts to multiple unicasts forwarding in the last hop from the SDN gateway towards the intended destinations, to ensure successful reception at the application layer. It can be further easily inferred (although not shown in results) that improvements get even larger as the distance between the source and the set of destinations increases.

Overall, the proposal is especially suited for push-based data dissemination to large Sociocast groups highly clustered and far from the source, which well resembles the case of multiple devices of a smart home (e.g., appliances) to be remotely configured by the user's smartphone.

In the other cases, the gains are also significant and always higher than a factor of 2.

The achieved encouraging results motivate us to further explore this fertile research area which has large room for improvements. The effectiveness of the proposal in handling other Sociocast features, like source filtering and publish/subscribe, needs to be practically explored. As a further challenge, IoT devices belonging to Sociocast groups may move long distances between different access points. Hence, tracking their positions at the virtual counterparts (SVN and SVNR), as well as managing the forwarding rules associated to them in the SDN nodes, become very difficult and entail proper workarounds which will be a subject matter of future investigations.

ACKNOWLEDGEMENT

This work was partially supported by the European Union's Horizon 2020 research and innovation program under the COG-LO project (grant agreement no. 769141).

REFERENCES

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[2] C. Diot, B. N. Levine, B. Lyles, H. Kassem, and D. Balensiefen, "Deployment issues for the IP multicast service and architecture," *IEEE network*, vol. 14, no. 1, pp. 78–88, 2000.

[3] L. Atzori, A. Iera, and G. Morabito, "Sociocast: A new network primitive for IoT," *IEEE Communications Magazine*, vol. 57, no. 6, pp. 62–67, 2019.

[4] D. Kreutz, F. M. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.

[5] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: rapid prototyping for software-defined networks," in *Proc. of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, p. 19, ACM, 2010.

[6] "ON.LAB, "Introducing ONOS - a SDN network operating system for service providers," 2014.

[7] S.-H. Shen, "Efficient SVC multicast streaming for video conferencing with SDN control," *IEEE Transactions on Network and Service Management*, 2019.

[8] B. Knutsson, H. Lu, W. Xu, and B. Hopkins, "Peer-to-peer support for massively multiplayer games," in *IEEE INFOCOM 2004*, vol. 1, IEEE, 2004.

[9] X. S. Sun, Y. Xia, S. Dzinamarira, X. S. Huang, D. Wu, and T. E. Ng, "Republic: Data multicast meets hybrid rack-level interconnections in data center," in *2018 IEEE 26th International Conference on Network Protocols (ICNP)*, pp. 77–87, IEEE, 2018.

[10] X. Li, Y.-C. Tian, G. Ledwich, Y. Mishra, X. Han, and C. Zhou, "Constrained optimization of multicast routing for wide area control of smart grid," *IEEE Transactions on Smart Grid*, 2018.

[11] M. Hosseini, D. T. Ahmed, S. Shirmohammadi, and N. D. Georganas, "A survey of application-layer multicast protocols," *IEEE Communications Surveys and Tutorials*, vol. 9, no. 1-4, pp. 58–74, 2007.

[12] S. Islam, N. Muslim, and J. W. Atwood, "A survey on multicasting in software-defined networking," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 355–387, 2017.

[13] Z. AlSaeed, I. Ahmad, and I. Hussain, "Multicasting in software defined networks: A comprehensive survey," *Journal of Network and Computer Applications*, vol. 104, pp. 61–77, 2018.

[14] A. Adams, J. Nicholas, and W. Siadak, "RFC 3973, protocol independent multicast-dense mode (PIM-DM): Protocol specification (revised)," tech. rep., 2005.

[15] B. Cain, S. Deering, I. Kouvelas, B. Fenner, and A. Thyagarajan, "RFC 3376, Internet Group Management Protocol, version 3," tech. rep., August 2006.

[16] H. Holbrook and B. Cain, "RFC 4607, Source-Specific Multicast for IP," tech. rep., August 2006.

[17] A. Venkataramani *et al.*, "MobilityFirst: a mobility-centric and trustworthy internet architecture," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 3, pp. 74–80, 2014.

[18] C. A. Marcondes, T. P. Santos, A. P. Godoy, C. C. Viel, and C. A. Teixeira, "Castflow: Clean-slate multicast approach using in-advance path processing in programmable networks," in *2012 IEEE Symposium on Computers and Communications (ISCC)*, pp. 000094–000101, IEEE, 2012.

[19] M. Zhao, B. Jia, M. Wu, H. Yu, and Y. Xu, "Software defined network-enabled multicast for multi-party video conferencing systems," in *2014 IEEE International Conference on Communications (ICC)*, pp. 1729–1735, IEEE, 2014.

[20] D. Katsaros, N. Dimokas, and L. Tassiulas, "Social network analysis concepts in the design of wireless ad hoc network protocols," *IEEE network*, vol. 24, no. 6, pp. 23–29, 2010.

[21] Y. Zhu, B. Xu, X. Shi, and Y. Wang, "A survey of social-based routing in delay tolerant networks: Positive and negative social effects," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 387–401, 2012.

[22] K. W. *et al.*, "Exploiting small world properties for message forwarding in delay tolerant networks," *IEEE Transactions on Computers*, vol. 64, no. 10, pp. 2809–2818, 2015.

[23] K. Wei, X. Liang, and K. Xu, "A survey of social-aware routing protocols in delay tolerant networks: applications, taxonomy and design-related issues," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 556–578, 2013.

[24] M. Xiao, J. Wu, and L. Huang, "Community-aware opportunistic routing in mobile social networks," *IEEE Transactions on Computers*, vol. 63, no. 7, pp. 1682–1695, 2013.

[25] D. J. Watts, "Networks, dynamics, and the small-world phenomenon," *American Journal of sociology*, vol. 105, no. 2, pp. 493–527, 1999.

[26] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *Proceedings of ACM MobiHoc*, pp. 299–308, ACM, 2009.

[27] W. Gao, Q. Li, B. Zhao, and G. Cao, "Social-aware multicast in disruption-tolerant networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 20, no. 5, pp. 1553–1566, 2012.

[28] X. Hu, T. H. Chu, V. C. Leung, E. C.-H. Ngai, P. Kruchten, and H. C. Chan, "A survey on mobile social networks: Applications, platforms, system architectures, and future research directions," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1557–1581, 2015.

[29] J. Fan, J. Chen, Y. Du, W. Gao, J. Wu, and Y. Sun, "Geocommunity-based broadcasting for data dissemination in mobile social networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 4, pp. 734–743, 2013.

[30] L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The social internet of things (SIoT)–when social networks meet the internet of things: Concept, architecture and network characterization," *Computer networks*, vol. 56, no. 16, pp. 3594–3608, 2012.

[31] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: enabling innovation in campus networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.

[32] L. Atzori, C. Campolo, B. Da, R. Girau, A. Iera, G. Morabito, and S. Quattropani, "Enhancing identifier/locator splitting through social internet of things," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2974–2985, 2018.

[33] "Openflow switch specification - version 1.3.1 Open Networking Foundation (ONF)," September 2012.

[34] J. Hui and P. Thubert, "Compression format for IPv6 datagrams over ieee 802.15.4-based networks, RFC 6282," September 2011.

[35] A. Ayadi, D. Ros, and L. Toutain, "TCP header compression for 6LoWPAN," *Internet Draft (draft-aayadi-olowpan-tcphc-00), work in progress*, 2010.

[36] C. Gomez, A. Arcia-Moret, and J. Crowcroft, "TCP in the Internet of Things: from ostracism to prominence," *IEEE Internet Computing*, vol. 22, no. 1, pp. 29–41, 2018.

[37] L. Atzori, C. Campolo, B. Da, R. Girau, A. Iera, G. Morabito, and S. Quattropani, "Smart devices in the social loops: Criteria and algorithms for the creation of the social links," *Future Generation Computer Systems*, vol. 97, pp. 327–339, 2019.

AUTHORS

Luigi Atzori (PhD, 2000) is professor of Telecommunications at the University of Cagliari, where he leads the activities of the MCLab laboratory (Multimedia Communications) with around 20 affiliated researchers. Since 2018, he has been the coordinator of the master degree course in Internet Technology Engineering at the University of Cagliari. His research interests fall in the area of Internet of Things (IoT), with particular reference to the design of effective algorithms for the realization of social networks among connected devices to create the Social IoT paradigm. His interests also falls in the area of Quality of Experience (QoE), with particular application to the management of services and resources in new generation networks for multimedia communications. Lately, he also applies the study of QoE to IoT services. He serves regularly in the conference organizing committee of the sector and as associate and guest editor of several international journals (IEEE IoT journal, Ad Hoc Networks, IEEE IEEE Open Journal of the Communications Society, IEEE Communications Magazine, etc.).

Claudia Campolo (Senior Member, IEEE) received the master's and Ph.D. degrees in telecommunications engineering from the Mediter- ranea University of Reggio Calabria, in 2007 and 2011, respectively. In 2008, she was a Visiting Ph.D. Student with the Politecnico di Torino and a DAAD Fellow with the University of Pader- born, Germany, in 2015. She is currently an Associate Professor of telecommunications with the Mediterranea University of Reggio Calabria. Her main research interests include vehicular networking, 5G, and edge computing.

Antonio Iera is professor of Telecommunications at the University of Calabria, Italy. He graduated in Computer Engineering at the University of Calabria, Italy, and received a Master Diploma in Information Technology from CEFRIEL/Politecnico di Milano, Italy, and a Ph.D. degree from the University of Calabria. From 1994 to 1995 he has been with Siemens AG in Munich, Germany, and from 1997 to 2019 with the University of Reggio Calabria. He has published more than 300 papers in high-quality journals and conferences and has given several Tutorials and invited speeches at international events on the topics of IoT, Social-IoT, and 5G networks. He is also serving as Editor in Chief for Computer Networks, Elsevier.His research interests include wireless and mobile 5G networks and Internet of Things.

Giuseppe Massimiliano Milotta was born in Catania, Sicily (Italy) on May 07, 1986. He received his master degree at "Università degli studi di Catania" in 2017 after a six-month collaboration with "Télécom Paris Tech", Paris(France) in 2016, which led to the realization of his master degree thesis. He is currently enrolled at the last year of PhD in Information Engineering at "Università Mediterranea di

Reggio Calabria", Reggio Calabria (Italy) where is still cooperating with the CNIT research unit of Catania. His main research activities focus on SDN and the IoT's world, with a particular interest in the UAVs and the security systems.

Giacomo Morabito received the laurea degree cum laude in Electronic Engineering and the Ph.D from the University of Catania (Italy) in 1996 and 2000, respectively. Since 1999 to 2001 he was a research engineer at Georgia Tech (Atlanta, USA). Since 2001 he is with the Department of Electric, Electronic, and Computer Engineering of the University of Catania where he is currently full professor of telecommunications. His research interests focus on analysis and design of wireless networks and Internet of Things.

Salvatore Quattropani received the BSc. and MSc. degrees in Computer Engineering in 2016 and 2017, respectively, from University of Catania, Italy. Since August 2017 he is with the Consorzio Nazionale Interuniversitario per le Telecomunicazioni as a Research Engineer. His research interests focus on IoT connectivity and computing approaches.

THE INTERNET OF METAMATERIAL THINGS AND THEIR SOFTWARE ENABLERS

Christos Liaskos[1,2], Georgios G. Pyrialakos[2,3], Alexandros Pitilakis[2,3], Ageliki Tsioliaridou[2], Michail Christodoulou[2,3], Nikolaos Kantartzis[2,3], Sotiris Ioannidis[4,2], Andreas Pitsillides[5], Ian F. Akyildiz[5]
[1]Computer Science Engineering Dept., University of Ioannina, Ioannina, Greece, [2]Foundation for Research and Technology - Hellas, Heraklion, Greece, [3]Electrical and Computer Engineering Dept., Aristotle University, Greece, [4]Technical University of Chania, Crete, Greece, [5]Computer Science Dept., University of Cyprus, Nicosia, Cyprus,

NOTE: Corresponding author: Christos Liaskos, cliaskos@uoi.gr

Abstract *– A new paradigm called the Internet of MetaMaterial Things (IoMMT) is introduced in this paper where artificial materials with real-time tunable physical properties can be interconnected to form a network to realize communication through software-controlled electromagnetic, acoustic, and mechanical energy waves. The IoMMT will significantly enrich the Internet of Things ecosystem by connecting anything at any place by optimizing the physical energy propagation between the metamaterial devices during their lifetime, via "eco-firmware" updates. First, the means for abstracting the complex physics behind these materials are explored, showing their integration into the IoT world. Subsequently, two novel software categories for the material things are proposed, namely the metamaterial Application Programming Interface and Metamaterial Middleware, which will be in charge of the application and physical domains, respectively. Regarding the API, the paper provides the data model and workflows for obtaining and setting the physical properties of a material via callbacks. The Metamaterial Middleware is tasked with matching these callbacks to the corresponding material-altering actuations through embedded elements. Furthermore, a full stack implementation of the software for the electromagnetic metamaterial case is presented and evaluated, incorporating all the aforementioned aspects. Finally, interesting extensions and envisioned use cases of the IoMMT concept are discussed.*

Keywords – Internet of Things, metamaterials, programming interface, software-defined networking.

1. INTRODUCTION

Recent years have witnessed the advent of the Internet-of-Things (IoT), denoting the interconnection of every electronic device and the smart, orchestrated automation it entails [1]. Vehicles, smart phones, sensors, home and industrial appliances of any kind expose a functionality interface expressed in software, allowing for developers to create end-to-end workflows. As an upshot, smart buildings and even smart cities that automatically adapt, e.g., power generation, traffic and heat management to the needs of residents, have been devised in recent years. This current IoT potential stems from exposing and controlling a high-level functionality of an electronic device, such as turning on/off lights and air-conditioning units based on the time of day and temperature. This paper proposes the expansion of the IoT to the level of physical material properties, such as electrical and thermal conductivity, mechanical elasticity, and acoustic absorption. This novel direction is denoted as the *Internet of MetaMaterial Things (IoMMT)*, and can have groundbreaking potential across many industrial sectors, as outlined in this paper. There are two key enablers for the proposed IoMMT:

Key Enabler 1:

The first key enabler of the proposed IoMMT are the metamaterials, the outcome of recent research in physics that has enabled the creation of artificial materials with real-time tunable physical properties [2, 3]. Metamaterials are based on the fundamental idea stating that the physical properties of matter stem from its atomic structure. Therefore, one can create artificially structured materials (comprising sufficiently small elementary "units" of composition and geometry) to yield any required energy manipulating behavior, including types not found in natural materials. Metamaterials manipulating electromagnetic (EM) energy were the first kind of metamaterials to be studied in depth, mainly due to the relative ease of manufacturing as low-complexity electronic boards [3].

Going beyond EM waves, the collectively termed elastodynamic metamaterials can manipulate acoustic, mechanical and structural waves, whereas thermodynamic and quantum-mechanic metamaterials have also been postulated [4]. Elastodynamic metamaterials, empowered by recent advances in nano- and micro-fabrication (e.g. additive manufacturing/3D printing), can exhibit effective/macroscopic nonphysical properties such as tunable stiffness and absorption/reflection, extreme mass-volume ratios, negative sonic refraction, etc [5]. Their cell-size spans several length scales, depending on the application: acoustic cloaking/anisotropy/isolation, ultra-lightweight and resilient materials, devices for medical/surgical applications and food/drug adminis-

tration, MEMS, anti-seismic structures, etc. Tunability of elastodynamic metamaterials can be achieved with electric, magnetic, optical, thermal or chemical stimuli. In a nutshell, their operation is as follows: Impinging EM waves create inductive currents over the material, which can be modified by tuning the actuator elements within it (e.g., simple switches) accordingly. The Huygens principle states that any EM wavefront departing from a surface can be traced back to an equivalent current distribution over a surface [3]. Thus, in principle, metamaterials can produce any custom departing EM wave as a response to any impinging wave, just by tuning the state of embedded switches/actuators. Such EM interactions are shown in Fig. 1 (on the right side). The same principle of operation applies to mechanical, acoustic and thermal metamaterials [6].

Key Enabler 2:

The second key enabler of the IoMMT is the concept of networked metamaterials. These will come with an application programming interface (API), an accompanying software middleware and a network integration architecture that enable the hosting of any kind of energy manipulation over a metamaterial in real time (e.g., steering, absorbing, splitting of EM, mechanical, thermal or acoustic waves), via simple software callbacks executed from a standard PC (desktop or laptop), while abstracting the underlying physics. The goal is to constitute the IoMMT directly accessible to the IoT and software development industries, without caring for the intrinsic and potentially complicated physical principles. Regarding the IoMMT potential, large scale deployments of EM metamaterials in indoor setups have introduced the groundbreaking concept of programmable/intelligent wireless environment (Fig. 2) [7]. By coating all major surfaces in a space (e.g., indoors) with EM metamaterials, the wireless propagation can be controlled and customized via software. As detailed in [7] this can enable the mitigation of path loss, fading and Doppler phenomena, while also allowing waves to follow improbable air-routes to avoid eavesdroppers (a type of physical-layer security). In cases where the device beamforming and the EM metamaterials in the space are orchestrated together, intelligent wireless environments can attain previously unattainable communication quality and wireless power transfer [7]. Extending the EM case, we envision the generalized IoMMT deployed as structural parts of products, as shown in Fig. 3:

- EM interference and unwanted emissions can be harvested by IoMM-coated walls and be transformed back to usable EM or mechanical energy.
- Thermoelectric and mechanical metamaterials can micro-manage emanated heat and vibrations from devices, such as any kind of motor, to recycle it as energy while effectively cooling it. The same principle can be applied to a smart household or a noisy factory.

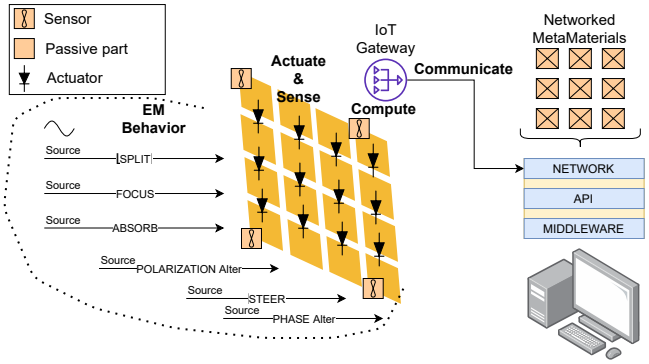


Fig. 1 – Networked metamaterial structure and possible energy wave interactions [8].

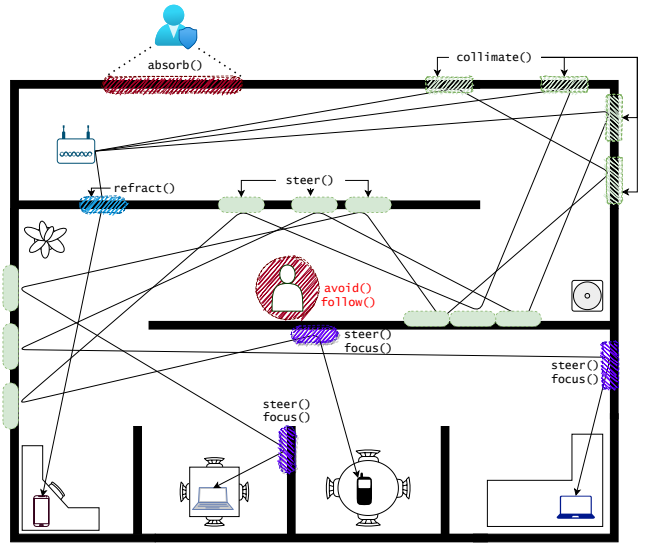


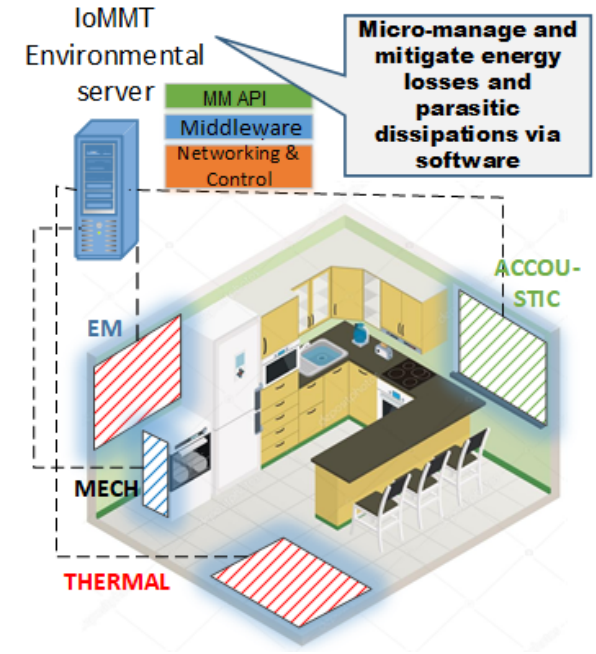
Fig. 2 – The programmable wireless environment introduced in [7], is created by coating walls with networked metamaterials. This allows for customized wireless propagation-as-an-app per communicating device pair, introducing novel potential in data rates, communication quality, security and wireless power transfer.

- The acoustic metamaterials can surround noisy devices or be applied on windows to provide a more silent environment, but to also harvest energy which can be added to a system such as a smart-household.

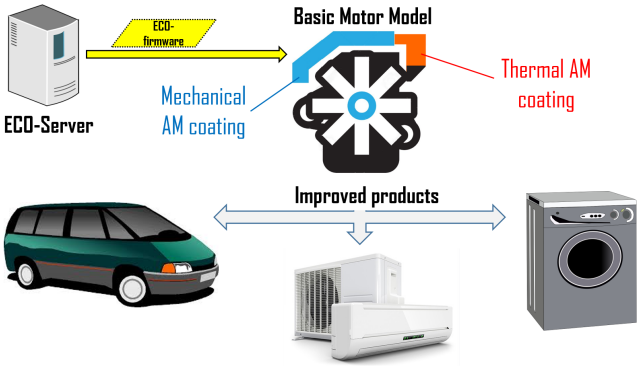
Assuming a central controller to optimize a given IoMMT deployment allows for further potential. For instance, one can allow for quickly "patching" of overlooked physical aspects (e.g., poor ecological performance) of IoMM-enabled products during operation, without overburdening the product design phase with such concerns. The "patching" may also be deferred in the form of "eco-firmware", distributed via the Internet to ecologically tune a single product or horizontal sets of products.

In this context, the principal contributions of the paper are as follows:

- We propose the concept of the IoMMT and discuss



(a) Conceptual IoMMT deployment within a smart home.



(b) Conceptual IoMMT deployment within single products.

Fig. 3 – Envisioned applications of the IoMMT in smart houses and products.

its architecture and interoperability with existing network infrastructures.

- We define two novel categories of software: the *Metamaterial API* and the *Metamaterial Middleware*, which enable any software developer to interact with a set of networked metamaterials, in a physics-agnostic manner. We establish the data models, workflows, and test bed processes required for profiling and, subsequently, componetizing metamaterials.
- We present an implemented and experimentally verified version of the metamaterial API and the Metamaterial Middleware for the EM case.
- We highlight promising, new applications empowered by the featured IoMMT concept.

In this aspect, the potential of our IoMMT paradigm is the first to offer true control over the energy propagation within a space, in every physical domain, i.e., for any physical material property and corresponding information-carrying wave. For instance, control over the equivalent RLC parameters of an electric load controls the power that can be delivered to it by an EM wave. Moreover, the presented software is a mature prototype platform for the development of IoMMT applications. This constitutes a major leap towards a new research direction. On the other hand, other research directions have proposed and explored the Internet of NanoThings [9]. Although similarly named, these directions are not related to the IoMMT, as they are about embedding nano-sized computers into materials in order to augment the penetration level of applications (e.g., sense structural, temperature, humidity changes within a material, rather than just over it, etc.), and not to control the energy propagation within them.

The remainder of this paper is organized as follows, devoting a section to each of the principal contributions of our work. In Section 2 we provide the related work overview and the necessary prerequisite knowledge for networked metamaterial. In Section 3 we present the architecture for integrating the IoMMT in existing Software-Defined Networks (SDNs) and systems. In Section 4 we present the novel metamaterial API, and Section 5 follows with the description of the Metamaterial Middleware and its assorted workflows. In Section 6 we present the implemented version of the software for the EM metamaterial case, along with a description of the employed evaluation test bed. Finally, novel realistic applications enabled with our new paradigm are discussed in Section 7, and we conclude the paper in Section 8.

2. PREREQUISITES AND RELATED WORK

Metamaterials are simple structures that are created by periodically repeating a basic structure, called a *cell* or a *meta-atom* [3]. Some examples across physical domains are shown in Fig. 4. The planar (2D) assemblies of meta-atoms, known as *metasurfaces*, are of particular interest currently [10, 11]. For instance, EM are currently heavily investigated by the electromagnetic/high-frequency community, for novel communications, sensing and energy applications. [12–14].

A notable trait of metamaterials is that they are simple structures and, therefore, there exists a variety of techniques for generally low-cost and scalable production [3]. The techniques such as printed circuit broads, flexible materials such as Kapton, 3D printing, Large Area Electronics, bio-skins and microfluidics have been successfully employed for manufacturing [3].

In each physical domain, a properly configured metamaterial has the capacity to steer and focus an incoming energy wave towards an arbitrary direction or even completely absorb the impinging power. In the EM case this capability can be exploited for advanced wireless com-

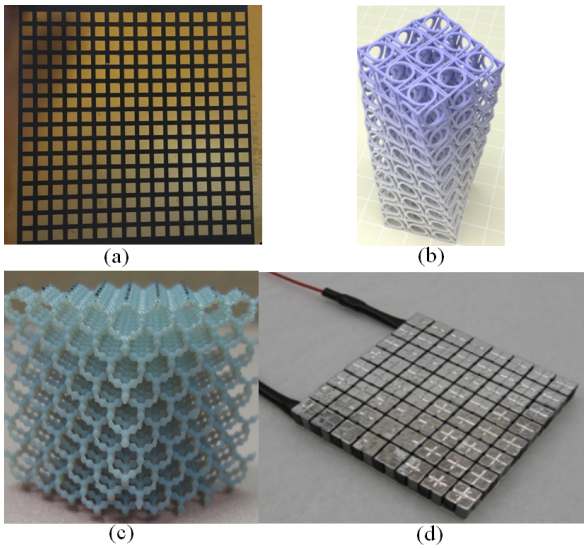


Fig. 4 – Energy manipulation domains of artificial materials: (a) Electromagnetic [20] (b) Mechanical [21] (c) Acoustic [22] (d) Thermoelectric [23].

munications [7, 15–19], offering substantially increased bandwidth and security between two communicating parties.

The potential stemming from interconnected metamaterials has begun to be studied only recently [8]. The perspective networking architecture and protocols [7,8], metamaterial control latency models [24], and smart environment orchestration issues have been recently studied for the EM case [25,26].

Notably, a similarly named concept, i.e., the Internet of NanoThings [9], was recently proposed to refer to materials with embedded, nano-sized computing and communicating elements. In general, these materials are derived from miniaturizing electronic elements and placing them over or embedding them into fabrics and gadgets, to increase their application-layer capabilities. For instance, this could make a glass window become a giant, self-powered touchpad for another IoT device. Originally. the concept of software-defined metasurfaces wa based on the nano-IoT as the actuation/control enabler [17]. Nano-devices can indeed act as the controllers governing the state of the active cells, offering manufacturing versatility and extreme energy efficiency. Nonetheless, until nano-IoT becomes a mainstream technology, other approaches can be adopted for manufacturing software-defined metasurfaces, as reported in the related physics-oriented literature [6]. It is also noted that nano-IoT as a general concept is about embedding nano-sized computers into materials in order to augment the penetration level of applications (e.g., sense structural, temperature, humidity changes within a material, rather than just over it, etc.), and not specifically to control the energy propagation within them.

In contrast, our work refers specifically to the case of metamaterials and the capabilities they offer for the manipulation of energy across physical domains. Moreover, our paper introduces the software enablers for this direction, which has not been proposed before. Additionally, our paper focuses more on the networking approaches for metamaterials, which has only been treated in our previous work [8], and only for the EM case. Finally, the work of Chen et al. [27] also advocates for the use of metamaterial in any physical domain for distributed energy harvesting, e.g., in a smart house or a city. However, software enablers and networking considerations are not discussed or solved in [27]. Moreover, the energy manipulation type is restricted to harvesting which can be viewed as a subset of our proposed IoMMT potential.

Metamaterials: Principles of Operation, Classification and Supported Functionalities

A conceptual metamaterial is illustrated in Fig. 5 [3]. Basically, a metamaterial consists of periodically repeated meta-atoms arranged in a 3D grid layout, with the metasurfaces being a sub-case. In particular, unit cells comprise passive and tunable parts, required in reconfigurable metamaterials as well as optional integrated sensory circuits, which can extract information of the incident energy wave. Furthermore, tunable parts are crucial for metamaterials, as they enable reconfigurability and switching between different functions. For illustration, in EM metamaterials at microwave frequencies, the tunable parts embedded inside the unit cells can be voltage-controlled resistors (varistors) and/or capacitors (varactors), micro-electromechanical switches (MEMS), to name a few [3, 19].

On the other hand, in mechanical and acoustic metamaterials, the tunable parts can be micro-springs with a tunable elasticity rate [28, 29]. The meta-atoms may also form larger groups, called *super-atoms* or *super-cells*, repeated in specific patterns that can serve more complex functionalities, as discussed later in this paper. Lastly, the software-defined metamaterials include a *gateway* [7], i.e., an on-board computer, whose main tasks are to: i) power the whole device and ii) control (get/set) the state of the embedded tunable elements, iii) interoperate with the embedded sensors, and iv) interconnect with the outside world, using well-known legacy networks and protocol stacks (e.g. Ethernet).

The relative size of a meta-atom compared to the wavelength of the excitation (impinging wave) defines the energy manipulation precision and efficiency of a metamaterial. For example, EM metasurfaces share many common attributes with classic antenna-arrays and reflect-arrays. Antenna arrays can be viewed as independently operating antennas, being very effective for coarse beam steering as a whole. Reflect-arrays typically consist of smaller elements (still subwavelength), permitting more fine-grained beam steering and a very coarse polarization control. Metamaterials comprise orders of magnitude smaller meta-atoms, and may also include tunable elements and sensors. Their meta-atoms are generally considered tiny with regard to the exciting wavelength, hence allowing full control over the form of the departing

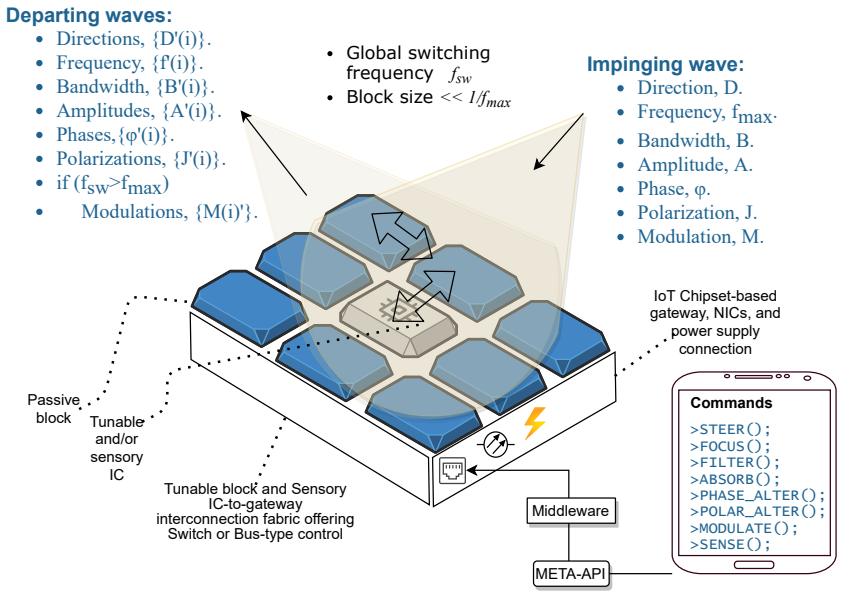


Fig. 5 – Overview of the metasurface/metamaterial structure and operating principles.

energy wave.

Regardless of their geometry and composition, the operating principle of metamaterials remains the same. As depicted in Fig. 5, an impinging wave of any physical nature (e.g., EM, mechanical, acoustic, thermal) excites the surface elements of a metamaterial, initiating a spatial distribution of energy over and within it. We will call this distribution "exciting-source". On the other hand, well-known and cross-domain principles state that any energy wavefront, which we demand to be emitted by the metamaterial as a response to the excitation, can be traced back to a corresponding surface energy distribution denoted as "producing-source" [3, 6]. Therefore, a metamaterial configures its tunable elements to create a circuit that morphs the exciting-source into the producing-source. In this way, a metamaterial with high meta-atom density can perform any kind of energy wave manipulation that respects the energy preservation principle. Arguably, the electromagnetism constitutes a very complex energy type to describe and, as a consequence, manipulate in this manner, as it is described by two dependent vectors (electric and magnetic field) as well as their relative orientation in space, i.e. polarization (mechanical, acoustic and thermal waves can be described by a single scalar field in space). As such, incoming EM waves can be treated in more ways than other energy types. The common types of EM wave manipulation via metamaterials, reported in the literature [3], can designate a set of high-level functionality types as follows:

- Amplitude: Filtering (band-stop, -pass), absorption.
- Polarization: Waveplates (polarization conversion, modulation).
- Wavefront: Steering (reflecting or refracting), splitting, focusing, collimating, beamforming, scattering.
- Bandwidth: Filtering.
- Modulation: Requires embedded actuators that can switch states fast enough to yield the targeted modulation type [30].
- Frequency: Filtering, channel conversion.
- Doppler effect mitigation and non-linear effects [8].

Additionally, sensing impinging waves may be considered one of the above functionalities and, as an outcome, the embedded sensors can extract information of any of the above parameters related to the incident wave.

In this aspect, the role of the contributed metamaterial API is to model these manipulation types into a library of software callbacks with appropriate parameters. Then, for each callback and assorted parameters, the Metamaterial Middleware produces the corresponding states of the embedded tunable elements that indeed yield the required energy manipulation type. In other words, a metamaterial coupled with an API and a Metamaterial Middleware *can be viewed as a hypervisor that can host metamaterial functionalities* upon user request [8].

In the following, we focus on EM metamaterials which, as described, yield the richest API and most complex Metamaterial Middleware. The expansion to other energy domains is discussed via derivation in Section 7.

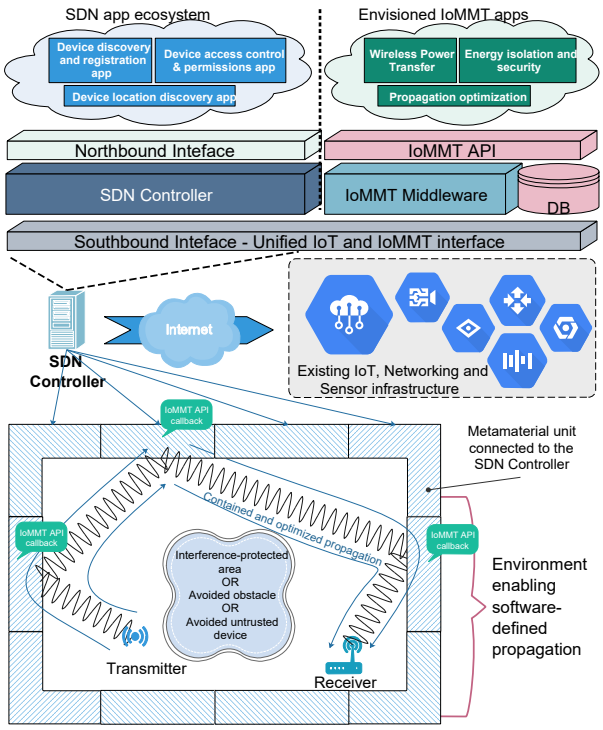


Fig. 6 – SDN schematic display of the system model and the entire workflow abstraction.

3. NETWORKED METAMATERIALS AND SDN WORKFLOWS

Many metamaterials deployed within an environment can be networked through their gateways. This means that they may become centrally monitored and configured via a server/access point in order to serve a particular end objective.

An example is given in Fig. 6, where a set of metamaterials is designed with the proper commands for energy wave steering and focusing, in order to route the energy waves exchanged between two wireless users, thus avoiding obstacles or eavesdroppers. Other applications include wireless power transfer and wireless channel customization for an advanced quality of service (QoS) [7, 13]. Such a space, where energy propagation becomes software defined via metamaterials is called a programmable wireless environment (PWE) [8].

As shown in [7], the PWE architecture is based on the software-defined networking (SDN) principles. The PWE server is implemented within an SDN controller [31]; the southbound interface abstracts the metamaterial hardware, treating metamaterial devices as networking equipment that can route energy waves (e.g. similar to a router, albeit with a more extended and unique parameterization). Thus, the metamaterial API constitutes a part of the northbound SDN interface, atop of which the security, QoS and power transfer concepts can be implemented as SDN controller applications. On the other hand, the Metamaterial Middleware is part of the SDN middleware, translating metamaterial API callbacks into metamaterial hardware directives.

A notable trait of the Metamaterial Middleware is that it is divided into *two parts,* in terms of system deployment [32]:

1. The metamaterial manufacturing stage component, a complex, offline process requiring special metamaterial measurement and evaluation setups (discussed in Section 5), and

2. The metamaterial operation stage component, which operates in real time based on a codebook. This codebook is a database populated once by the manufacturing stage component and contains a comprehensive set of configurations for all metamaterial API callbacks, supported by a given metamaterial.

The operation stage component simply retrieves configurations from the codebook and optionally combines them as needed, using an interleaving process described in Section 5.

Notably, other studies propose the use of online machine learning as a one-shot process, which can be more practical when response time is not a major concern [33]. However, in this work we propose the aforementioned separation in deployment, to ensure the fastest operation possible overall, thus covering even the most demanding cases.

It is noted that SDN is not a choice due to restrictions, but rather a choice due to compatibility. In the software-defined metasurfaces presented in this paper, a key point is the abstraction of physics via an API that allows networking logic to be reused in PWEs, without requiring a deep understanding of physics. SDN has (among other things) already introduced this separation of control logic from the underlying hardware and its administrative peculiarities. Therefore, we propose an integration of PWE within SDN to better convey the logical alignment of the two concepts.

4. APPLICATION PROGRAMMING INTERFACE FOR METAMATERIALS

In the following, we consider a metamaterial in the form of a rectangular *tile*. The term tile is used to refer to a practical metamaterial product unit, which can be used to cover large objects such as walls and ceilings in a floorplan.

A software process can be initiated for any metamaterial tile supporting a unique, one-to-one correspondence between its available switch element configurations and a large number of metamaterial functionalities. The metamaterial tiles in this work incorporate tunable switch elements, which dictate the response of each individual cell, locally. In this way, providing an arrangement of all the tile cells allows the tuning of the "concerted" metamaterial response of the entire tile.

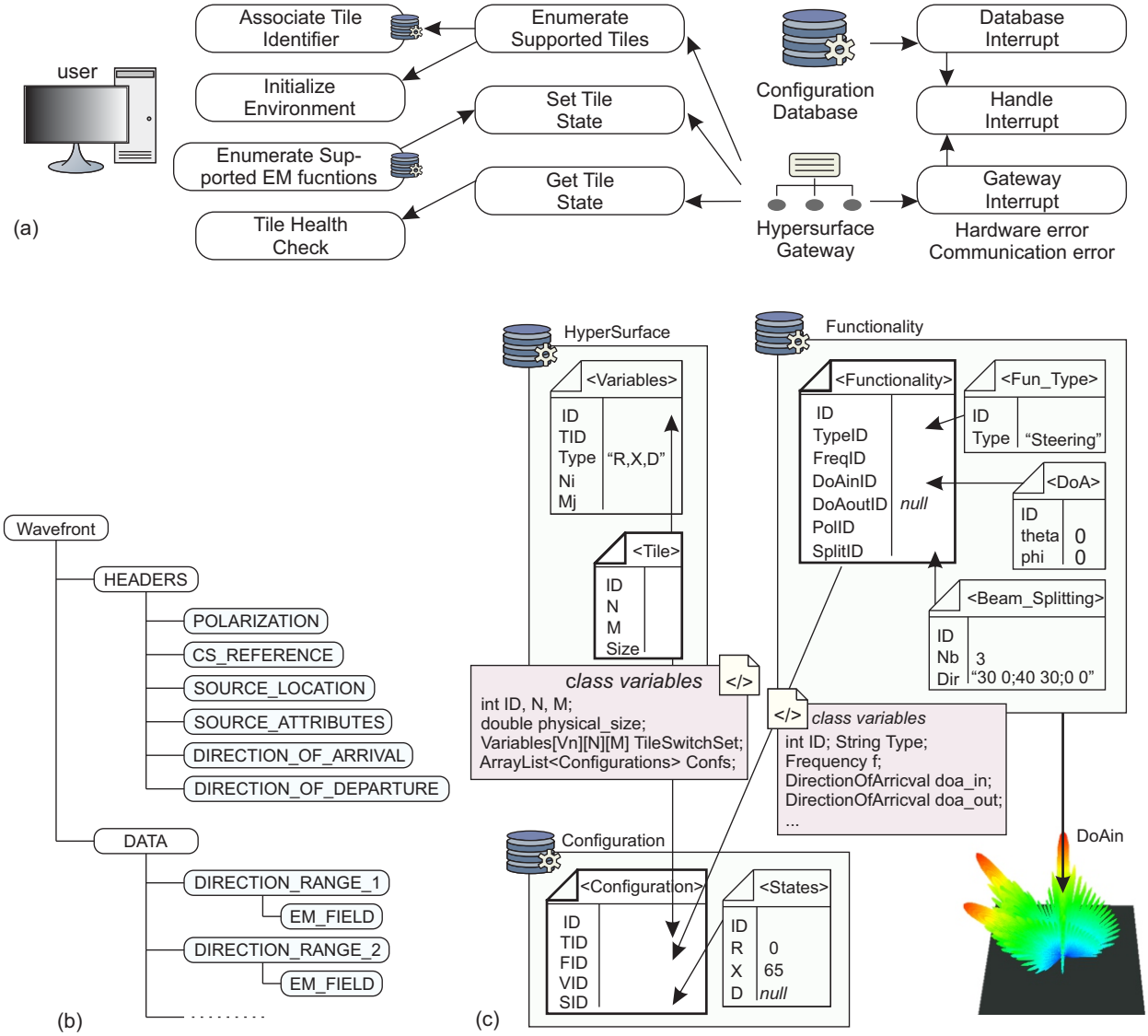


Fig. 7 – (a) Case diagram of the main functions supported by the three basic entities. Tasks highlighted with the database icon indicate that a set of data is to be retrieved from the Configuration Database. (b) Wavefront description in data object format. (c) A simplified overview of the structure of the Configuration Database. The ¡Tile¿ table hosts all information regarding a tile's physical implementation. The ¡Functionality¿ table combines a set of metamaterial parameters to define new functionalities. Both tables are combined in the ¡Configuration¿ table with a set of entries from the ¡States¿ table to compose a new configuration that supports the functionality FID on tile TID (VID refers to an entry in the table ¡Variables¿).

In this section, we present the API that grants access to the tile's metamaterial applications by defining an abstract representation of the metamaterial, its switch element configurations and their respective functionalities. Specifically, the API resides between a user, operating a common PC (desktop, smartphone, etc.) and a tile gateway, linked to the network of switch element controllers. The case diagram of the proposed concept, presented in Fig. 7(a), involves the following main entities:

- A Configuration Database which stores all information regarding the tiles, the switch element configurations, and their corresponding functionalities.
- A User which initiates all API callbacks though either the source code or button click events in the Graphical User Interface (GUI).
- The HyperSurface Gateway which represents the electronic controller of the hardware.
- An Interrupt handling service which acts as a persistent daemon, receiving and dispatching commands to the Hypersurface Gateway.

4.1 Data Structures of the Metamaterial API

In the configuration Database (DB), each tile is associated with an element array S that represents all possible arrangements of switch element states on the metamaterial under study. Each switch element is represented by either a discrete or a continuous variable, creating a mathematical space of $V_1 \times N \times M + \cdots + V_n \times N \times M$ dimensions, where V_i is the number of elements of the same type (e.g. capacitors) and N, M the number of unit cells towards the two perpendicular directions. Furthermore, every object in this space corresponds to a different state of S and therefore a different configuration. As an example, a tile with two controllable resistive and one diode elements per unit cell is parameterized by a $2 \times N \times M + 1 \times N \times M$ array, where the first and second sets span a continuous $[R_{\min}, \ldots, R_{\max}]$ and a discrete $[0, 1]$ range, respectively. In this case, 0 and 1 correspond to the OFF and ON states of the diode. This representation will then acquire the following form

$$F \longleftarrow [(d_1, d_2, i_1)_{n=1}, \ldots, (d_1, d_2, i_1)_{n=N\times M}] \qquad (1)$$

where d_1, d_2 are double-type variables, i_1 is an integer variable, and F is the appointed functionality. The primitive data types of all variables should be selected so as to minimize the total parameter space of combined states without any loss of relevant information. This lays a better optimized communication and computational burden to both the API and the Compiler, especially during the compilation process where a sizable amount of mathematical computations is required. Accordingly, all functionalities are, also, associated with their own representation and classified pertinent to their own type and defining parameters. For instance, a complex beam-splitting and polarization control operation is parameterized by a discrete variable corresponding to the number of outgoing beams, their directivity amplitudes, and an appropriate number of (θ, φ) pairs, indicating the steering angles. The most complex functionality can be generally described by a custom scattering pattern and represented herein by a collection of variables that indicate the reflected power towards all directions within the tile's viewing area.

It is, also, worth noting that the data objects being passed as arguments in the callbacks are primarily descriptions of wavefronts. A simplified data structure is illustrated in Fig. 7(b). Hence, a wavefront is described by a type (string identifier), such as "Planar", "Elliptic", "Gaussian", "Custom", etc. For each type, a series of headers defines the location and attributes of the creating source (for impinging wavefronts only) as well as the coordinate system origin with respect to which all distances are measured. Moreover, the direction of arrival and departure are arrays that can be used to define multiple impinging or departing wavefronts at the same time. Notably, the information within the headers may be sufficient to produce any value of the wavefront via simply analytical means. In such cases, the data part can be left empty. In custom wavefronts, the data is populated accordingly. A mechanism for defining periodicity is supplied via the notion of ranges (i.e. coordinate ranges where the energy field is approximately equal), to potentially limit the size of the overall data object.

The parameters that represent the functionalities and configurations of a tile constitute the set of variables that are exposed to the programmer through the metamaterial API. They are organized in a unified manner within the Database, as shown in Fig. 7(c) which provides an illustration of the unique association between all primary tables. Particularly, the ¡Tile¿ table stores all information of a tile's hardware implementation, such as the number of variables per unit cell and the type of switch elements. The ¡Functionality¿ table stores the representation scheme described in subsection 2 for all available metamaterial functionalities. Each parameter associated with a functionality is organized in a separate table, including a table that stores an identification variable representing the type of functionality. This table ID enumerates all possible operations supported by the tile, including full power absorption, wavefront manipulation (steering, splitting, etc.), and wavefront sensing. Finally, the ¡Configuration¿ table combines, in an exclusive manner, both primary tables (¡Tile¿ and ¡Functionality¿) to link each stored functionality with a specific set of switch element states, acquired from the pool of available entries in the secondary table ¡States¿.

4.2 API Callbacks and Event Handling

Using the Database as a reference point, the API is responsible for interpreting a configuration array to the proper set of hardware commands, when a suitable call-

back is executed. In general, an API callback can refer to a number of common requests such as:

- Detect the number and type of accessible tiles in the environment.
- Get the current state of all switch elements or set them to a specific configuration.
- Check the health status and handle interrupts from the tiles or the Database.

Prior to any other callback, the API follows an initiation process, while the software detects all presently active and connected (discoverable) tiles by broadcasting a corresponding network message. The tiles report their location and a unique identifier, e.g. a fixed value, that associates all tiles with the same hardware specifications. The API validates the support of the active tiles by checking if the identifier exists in the tile list present in the database. It, then, retrieves the switch element arrays that correspond to these tiles and remains idle until a new "get" or "set" request is received for a currently active or new configuration, respectively. This means that the API is now open to receive new functionality requests from a user, physically operating the software, or generate its own requests by reacting to unexpected changes in the environment of devices linked to the MS network. When a new functionality request is received, the API retrieves one of the available configurations from the Database and translates it to a proper set of element states on an active metamaterial tile. The corresponding API callback process is illustrated in Fig. 8. In particular, the Caller (user) executes a metamaterial Function Deployment request, which, in turn, invokes the Configuration Resolver, identifying a tile that supports the requested functionality. Next, the resolver queries the Database and returns a configuration that matches the intended metamaterial application, looking for a proper entry in the ¡Configuration¿ table. The API creates a string command, using the tile identifier and a hardware representation of the element state variables, which is, then, conveyed to the tile Gateway using the

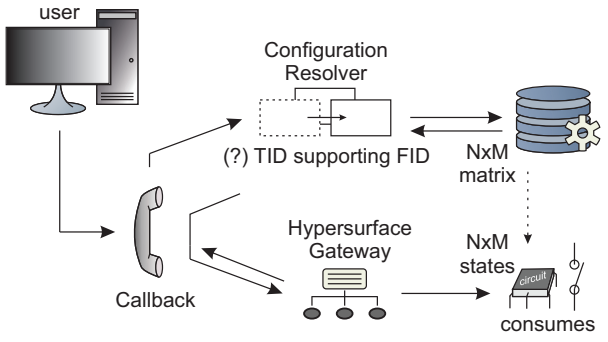


Fig. 8 – A metamaterial Function Deployment request initiates an API callback on the tile (TID). The Configuration Resolver seeks an appropriate configuration that supports the selected functionality (FID). The matrix of states is conveyed to the HyperSurface Gateway, where it is translated to a set of corresponding hardware states.

corresponding protocol. This notifies the intra-tile control network to assign the switch element states to their suitable values. Finally, feedback from a successful or failed configuration setup is received from the tile, notifying the user. The state of the newly set configuration is evaluated through either the identification of failed or unresponsive switches, or by activating the sensing app, in controlled conditions, as a self-diagnosing tool for the tile.

A more advanced API callback may involve the assignment of a secondary or supplementary functionality, on top of an already existing operation. For instance, the Metamaterial Middleware may receive independent requests from different users to steer the wavefront of several point power sources (i.e. the users' cellphones) towards the direction of a single nearby network hotspot. This can be handled by the API in many ways. In the case where several tiles are present in the environment, each tile can be repurposed to host a separate functionality, distributing all users to their own active tiles. When this is not feasible, the API can divide a single tile into separate areas and associate the respective element switches to different configurations (the division is usually performed in equal-sized rectangular patterns, but interlacing can, also, be used). Lastly, two functionalities can be combined into a single one, when a corresponding physical interpretation exists. For example, two separate steering operations, from the same source, may be combined into a single dual-splitting operation, expressed by a single unified pattern on the metamaterial.

In other cases, the configuration resolver may need to combine several functionalities to produce a special new application. This occurs when a functionality is parameterized by a continuous variable (e.g. a steering angle), while the Metamaterial Middleware can evaluate and store only a finite number of entries in the database. In such a scenario, the resolver seeks the two closest matching entries through an appropriate minimization function (e.g. for a beam-steering operation, the minimum distance between the requested and currently stored steering angles), whereas performing an interpolation of the switch state values.

To ensure a seamless operation, the API is reinforced with a set of specialised algorithms to handle unexpected failures in the hardware or communication network. So, a tile must include all necessary identification capabilities (e.g. the ability to identify power loss in its switch elements), and be able to notify the Metamaterial Middleware in case of failure. The API is, then, responsible for handling these errors by ensuring no loss of the current functionality [32]. For illustration, if a group of switch elements is stuck to an unresponsive state, the Metamaterial Middleware may instantly seek the closest matching functionality with a fixed configuration for the faulty elements. In more severe cases of demanding human intervention, like an unresponsive network (after a certain timeout), the API is also responsible for

informing an available user.

5. THE METAMATERIAL MIDDLEWARE

For the metamaterial to be reconfigured between different functionalities, a physical mechanism for locally tuning each unit cell response must be infused [8]. In the context of the present work, we assume that the response of the unit cells is controlled by variable *impedance loads* connected to the front side metallization layer of the metamaterial, where structures such as the resonant patch pair resides [3]. The loads are complex valued variables, comprising resistors and capacitors or inductors. The value of the i-th load, $Z_i = R_i + jX_i$, comprises two parameters: its resistance ($R_i > 0$) and reactance ($X_i = -(\omega C_i)^{-1}$ or $X_i = +\omega L_i$), for capacitive and inductive loads, respectively. The loads are, thus, electromagnetically connected to the surface impedance of the "unloaded" unit cell and by tuning their values we can regulate the unit cell response, e.g. the amplitude and phase of its reflection coefficient. The latter is naturally a function of frequency and incoming ray direction and polarization. When the metamaterial unit cells are properly "orchestrated" by means of tuning the attached (R_i, X_i) loads, the desired functionality (global response) of the metamaterial is attained.

In the most rigorous approach, the metamaterial response can be computed by full-wave simulations, which implement Maxwell's laws, given the geometry and metamaterial properties of the structure as well as a complex vector excitation, i.e. the impinging wave polarization and wavefront shape (phase and amplitude profile). The full-wave simulation captures the entire physical problem and, hence, does not require a metamaterial-level abstraction for the structure. Frequency-domain solvers, which assume linear media and harmonic excitation (i.e. the same frequency component in both the excitation and the response), are the prime candidates for full-wave simulation. They typically discretize the structure's volumes or surfaces at a minimum of $\lambda/10$ resolution, formulate the problem with an appropriate method (e.g. the finite-element or the boundary-element method) and, then, numerically solve a large sparse- or full-array system to compute the response, in our case, the scattered field. Conversely, time-domain solvers assume a pulsed excitation, covering a predefined spectral bandwidth and iteratively propagate it across the structure, solving Maxwell's equations to compute its response; they, typically, require a dense discretization of the structure, e.g. a minimum of $\lambda/20$ resolution. From this process, it becomes evident that metamaterial with a wide aperture, i.e. spanning over several wavelengths along the maximum dimension, require high computational resources in the full-wave regime, scaling linearly, when parametric simulations need to be performed to optimize the structure and/or the response. For instance, broadband simulation for the response of a unit cell of volume $(\frac{\lambda}{5})^3$ on a contemporary desktop computer could take several minutes, especially if the cell includes fine subwavelength features. The memory and CPU resources scale-up linearly for metasurfaces comprising hundreds of thousands of unit cells. Moreover, full-wave simulations do not explicitly unveil the underlying principles that govern the metamaterial functionality.

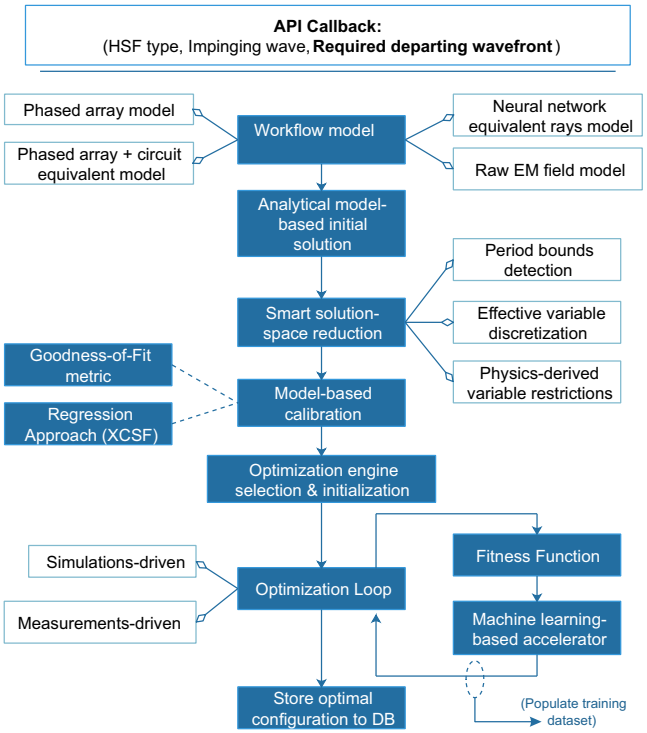


Fig. 9 – The Metamaterial Middleware functionality optimization workflow. The workflow seeks to match an analytical metamaterial model and its parameters to a specific parameterized API callback. A selected analytical model is first calibrated. Then, an iterative process (simulation or measurement-based) optimizes the input parameters of the model that best yield the API callback.

5.1 Functionality Optimization Workflow: Metamaterial Modeling and State Calibration

In this section we establish the optimization workflow of Fig. 9 that drives the calibration process of the metamaterial via an appropriate approximation model. Here calibration denotes the matching of actual active element states (e.g., the states of a tunable varactor) to the corresponding model parameter values (e.g., phase difference per cell in the reflectarray model). In this workflow, the optical scattering response is initially investigated and, then, the solution is hill-climbed via an optimization loop relying on either field measurements or full-wave simulations. The approximate models are as follows.

The simplest model is the phased antenna-array analysis, where each single unit cell is treated as an independent antenna, excited by a single impinging ray and emitting a single ray in response, with a local phase and

amplitude alteration. Assuming a metamaterial consisting of $M \times N$ unit cells, the scattered E-field complex amplitude pattern at a given frequency can be calculated by the envelope (coherent superposition) of all rays scattered from the metamaterial [2]

$$E(\theta,\varphi) = \sum_{m=1}^{M}\sum_{n=1}^{N} A_{mn} e^{j\alpha_{mn}} f_{mn}(\theta_{mn},\varphi_{mn}) \cdot \Gamma_{mn} e^{j\gamma_{mn}} f_{mn}(\theta,\varphi) e^{j\Phi_{mn}(\theta,\varphi)}. \quad (2)$$

In *(2)*, φ and θ are the azimuth and elevation angles in the scattering direction, $(\theta_{mn},\varphi_{mn})$ denotes the direction of the wavefront 'ray' incident on the mn-th cell, A_{mn} and α_{mn} are the amplitude and phase of the incident wavefront on the mn-th cell, Γ_{mn} and γ_{mn} form the reflection coefficient (amplitude and phase) of the mn-th cell, while f_{mn} defines the scattering pattern of the mn-th cell, which, according to reciprocity, is identical for the incident and scattered direction, and, in this work, is assumed that $f_{mn}(\theta,\varphi) = \cos(\theta)$. Finally, $\Phi_{mn}(\theta,\varphi)$ is the phase shift in the mn-th cell stemming from its geometrical placement, as

$$\Phi_{mn}(\theta,\varphi) = k\sin\theta\,[d_x m\cos\varphi + d_y n\sin\varphi] + \phi_0(\theta,\varphi), \quad (3)$$

where $d_{x,y}$ are the rectangular unit-cell lateral dimensions, $k = 2\pi/\lambda$ is the wavenumber in the medium enclosing the metamaterial, and ϕ_0 is the reference phase denoting the spherical coordinate system center, typically in the middle of the metamaterial aperture. Given a uniform, single-frequency impinging wave, any departing wavefront is essentially a Fourier composition of the individual meta-atom responses. Thus, we can, also, calculate the meta-atom amplitudes Γ_{mn} and phases γ_{mn} that yield a desired departing wavefront, by applying an inverse Fourier transform, as elaborately discussed in [2]. The calculated Γ_{mn} and γ_{mn} values must be mapped to the R_i and X_i values that generate them, since the latter are the actual tunable metamaterial parameters. This process requires a set of simulations yet it can be automated: existing model calibration techniques, such as the Regression and Goodness of Fit can be employed [34].

The shortcoming of the antenna-array approach is that the coupling between adjacent unit cells (e.g., compare against Fig. 5) is not properly accounted for, which can result to model imprecision [2]. To this aim, the Metamaterial Middleware user is presented with an alternative model. It utilizes the phased array and equivalent circuit model, which assumes not only the transmitting-responding antenna per meta-atom, but, also, circuit elements that interconnect them and account for the cross-meta-atom metamaterial interactions. The disadvantage of this approach is that an expert needs to define this circuit model, that is generally unique per metamaterial design [3]. Once this model has been selected and provided in the proper format, the optimization workflow of Fig. 9 continues, once again, with the calibration phase, which is identical as before. The key difference and merit is that the calibration is, now, extremely precise with regard to the full-wave simulations, while it takes much less time to complete, as detailed in the corresponding study of [2].

An intermediate solution, combining the precision of the circuit model and the automation of the antenna-array model, is an equivalent propagation model, mentioned here for the sake of completion. The main idea is to introduce a generic mechanism to capture the cross-interactions among meta-atoms (as opposed to the strict, physics-derived nature of the circuit model) and then proceed with automatic model calibration, avoiding the need for expert input. The equivalent ray model uses a neural network approach as the generic cross-talk descriptor [35]. A short summary is as follows. Each meta-atom is mapped to a neural network node, and the locally impinging wave amplitude and phase are its inputs. Then, we clone this layer (omitting the inputs) and form a number of intermediate, fully connected layers (usually 3-5), thereby emulating a recurrent network with a finite number of steps. We define links per node (shared among all node clones), which define an alteration of the local phase and amplitude, and its distribution to other neighboring meta-atoms/nodes. Next, we proceed to calibrate the model via feed-forward/back-propagation, thereby obtaining a match between R, X, Γ_{mn}, and γ_{mn} values. Nonetheless, despite its automated nature, a major drawback of this model is the need for considerable computational resources, without which the model loses its value, since it becomes restricted only to very simple metamaterial designs.

Since computational complexity is a concern regardless of the chosen model, the Metamaterial Middleware workflow allows the user to define solution reduction across three directions. First, meta-atoms may be grouped into periodically repeated super-cells. Thus, the optimization workflow needs only to optimize the configuration parameters of a super-cell, as opposed to optimizing the complete metamaterial. Second, the range of possible R and X values per meta-atom can be discretized into regular or irregular steps, reducing the solution space further[1]. Finally, some R and X values or ranges can be discarded due to the physical nature of the optimization request. For instance, if we seek to optimize a wave steering approach with an emphasis on minimal losses over the metamaterial (maximum reflection amplitude), the Ohmic resistance R needs to receive its boundary value. On a related track, machine learning-based approaches can quickly estimate the performance deriving from one set of R and X values, thereby discarding non-promising ones and accelerating convergence [36].

Subsequently, the Metamaterial Middleware workflow moves to the optimization stage, where it attempts to

[1] Notably, contemporary optimization engines already incorporate equivalents to this direction, as they are able to detect strongly and loosely connected inputs-outputs [34]).

hill-climb the initial solution detected via any of the described approximate models. At this point, the workflow is compatible to any modern optimization engine, which receives an input solution and outputs one or more proposed improvements upon it at each iteration. Herein, we stress the existence of engines that, also, incorporate machine learning mechanisms, to accelerate the optimization cycle [34]. The optimization can be based either on full-wave simulations or a real measurement test bed, described in Section 6. The optimization metric can be any reduction of the produced departing wavefront. Various metrics relevant to antenna and propagation theory may be extracted, namely: the number of main lobes (beam directions), the directivity of main lobes, the side (parasitic) lobes and their levels, the beam widths, etc. Such metrics can be used to quantify the metamaterial performance for the requested functionality, e.g. the main lobe directivity and beam width measures how "well" a metamaterial steers an incoming wavefront to a desired outgoing direction. Lastly, the hill-climbed metamaterial configuration pertaining to the metamaterial API callback is stored into a database for any future use by metamaterial users.
Finally, we note that multiple simultaneous functionalities can be supported by interlacing different scattering profiles across the metamaterial. In general, this is preformed by spatially mixing the profiles in phasor form

$$A_{mn}e^{j\alpha_{mn}} = \sum_{c=1}^{N_c} A_{c,mn}e^{j\alpha_{c,mn}}, \qquad (4)$$

where c iterates over single, "low-level" functionalities and n, m are the unit cell indices. Typically, low-level functionalities correspond to simple beam steering operations, which are produced exclusively by phase variations on the metamaterial ($A_{c,mn} = 1$). In this case, a "high-level" functionality will correspond to a multi-splitting operation with variable spatial distribution of A_{mn} amplitude, raising the hardware requirements for the metamaterial. Therefore, a metamaterial with no absorption capabilities (and thus no control over A_{mn}) will have limited access to high-level operations, unless a mathematical approximation is to be applied, skewing the scattering response from its ideal state. As discussed in Section 6, a method for minimizing amplitude variations has been successfully investigated by increasing the number of secondary parasitic lobes. Such a problem can be easily reformulated into an optimization task, where an optimal match to the ideal high-level operation can be pursued under specific constrains (e.g. $A_{mn} > \text{const.}, \forall m, n$).

5.2 The Metamaterial Functionality Profiler

The optimization workflow of Fig. 9 opts for the best metamaterial configuration for a given, specific pair of impinging and departing wavefronts. However, in real deployments, it is not certain that a metamaterial will always be illuminated by the intended wavefront [24]. For instance, user mobility can alter the impinging wavefront in a manner that has limited relation to the intended one and, consequently, to the running metamaterial configuration. As such, there is a need for *fully profiling a metamaterial*, i.e. calculate and cache its expected response for each intended metamaterial configuration, but, also, for each possible (matching or not) impinging wavefront of interest. This profiling process is outlined in Fig. 10.

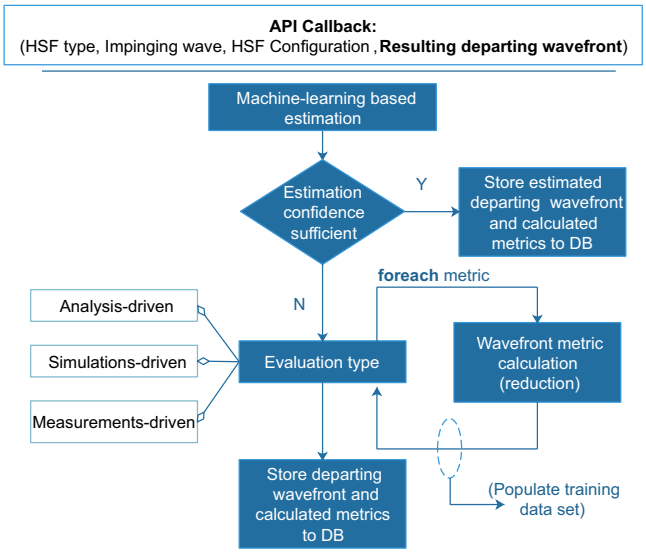


Fig. 10 – Workflow for profiling a metamaterial functionality. The workflow seeks to produce a data set that describes the metamaterial behavior for any impinging wave type that does not match the one specified in the current metamaterial configuration. An exhaustive evaluation takes place first for a wide set of possible impinging waves. For intermediate impinging wave cases, the workflow can rely on estimations produced by machine learning algorithms or simple extrapolation means, provided that it yields an acceptable degree of confidence.

The profiling process begins by querying the existing cache (part of the DB) or trained model for the given metamaterial and an estimation (or existing calculated outcome) of the expected metamaterial response for a given impinging wave. If it exists, this response is stored into a separate profile entry for the metamaterial in the Metamaterial Middleware DB[2]. If the response needs to be calculated anew, the process proceeds with either an analysis-, simulation- or measurement-driven evaluation. Therefore, the choice is given as a means to facilitate the expert into reducing the required computational time, as allowed per case. Then, the profiler proceeds to, also, calculate all possible reductions of the departing wavefront, e.g. the number of main lobes (beam directions), the directivity of the main lobes, the side (parasitic) lobes and their levels, the beam widths, etc. Finally, once all required impinging wavefronts have been successfully processed, the profiling process is con-

[2]In case of an estimated response, the user has control over the process to filter out estimations with low confidence. However, the selected estimation engine must be able to provide a confidence degree for this automation.

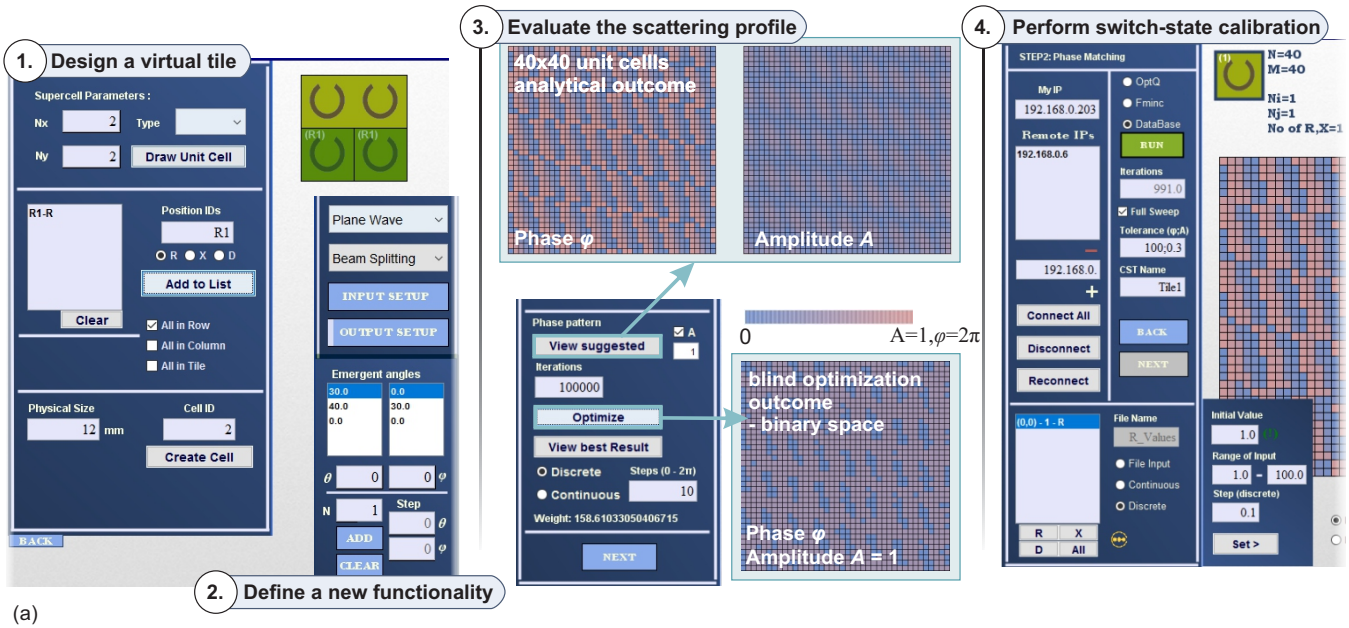
cluded.
It is clarified, that the middleware operations are one-time only, i.e., once the database containing the behavior profile of a metasurface is complete, it can be used in any application setting in the real world by any tile of the same type.

6. SOFTWARE IMPLEMENTATION AND EVALUATION

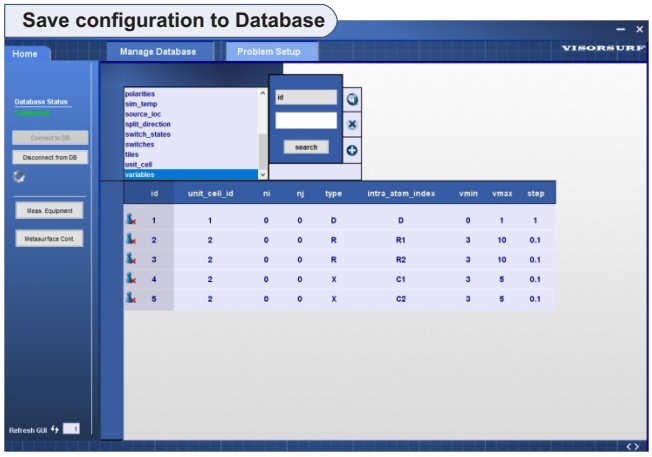
Employing the concepts of the previous sections, we developed a complete Java implementation of the described software. The software is subdivided into two integral modules: i) an implementation of the metamaterial API that handles the communication and allocation of existing configurations and ii) the Metamaterial Middleware that populates the configuration DB with new data (new tiles, configurations, and functionalities). The Metamaterial Middleware incorporates a full GUI environment, guiding the user through a step-by-step process to produce new configurations. It utilizes all available theoretical and computational tools for the accurate characterization of a metamaterial tile. Furthermore, it offers direct access to the configuration DB, manually, via a custom-made Structured Query Language (SQL) manager or through the automated process following a successful metamaterial characterization. Through this process, all the necessary data related to a newly produced configuration become explicitly available to the API.
A microwave metasurface was selected to demonstrate the capabilities of the developed concepts and methods for software-tunable metamaterials. We adapted the design of [37], where a set of RF diodes can be employed to toggle the reflection-phase of each cell between 16 states. We numerically extracted the response of the metasurface (i.e., its scattering pattern), and finally used the developed software to demonstrate how the metasurface response can be controlled. For the practical demonstration of the developed software in the same measurement environment (anechoic chamber) with a simpler, 1-bit metasurface hardware, we redirect the reader to [20,30], since the hardware manufacturing topic is quite extensive and clearly beyond the software aspects that constitute the focus of this paper.
In the following, we list and comprehensively describe the steps undertaken during a optimization process, as seen through the GUI environment of the Metamaterial Middleware. In summary, this process involves:

- The definition of a new unit cell structure and tile array (if required).
- The parameterization of a new functionality.
- The analytical evaluation of the scattering profile on the metamaterial for the selected functionality.
- The association of the metamaterial profile to the set of element states, through the use of numerical simulations.
- The experimental evaluation of the exported configuration through physical measurement of a metamaterial prototype. Notably, the presented software has been verified experimentally, and a full report can be found online [20].
- The final storing of all configuration parameters into the configuration DB to complete the function optimization process.

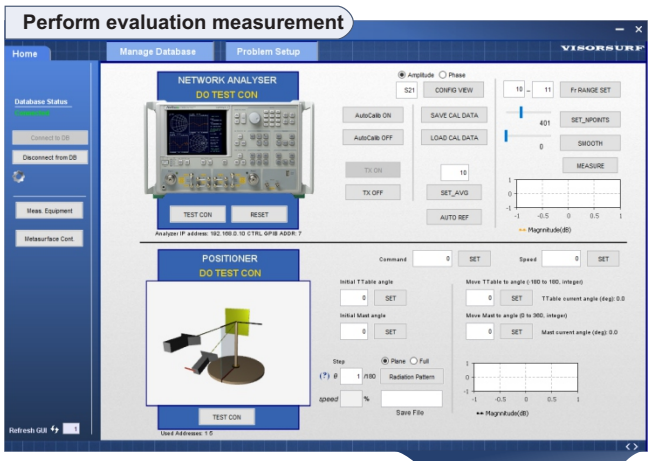
In this context, Fig. 11(a) depicts all the individual steps in separate panels. If a new configuration is to be defined for a tile already present in the configuration DB, then, the first step can be skipped. Alternatively, the user must input all essential parameters of the unit cell structure, i.e. the number and type of all variables that correspond to the sum of reconfigurable metamaterial elements. The definition of a new configuration begins with the parameterization of the desired functionality (Fig. 11(a.2)). The current implementation supports plane wave or point source inputs (for far- and near-field energy sources) and a set of output options corresponding to all basic metamaterial functionalities, discussed in Section 2. Here, we select a beam splitting operation and proceed to the first main step of the characterization process.
The analytical evaluation of the energy scattering profile is performed in the software locally and in real time. During this step, the Metamaterial Middleware calculates the proper scattered field response of the impinging wave, for each unit cell at the $N \times M$ tile, via either the analytical methods of Section 5 or through an optimization process. The scattered fields are evaluated for each unit cell as a double complex variable ($A_1 e^{i\phi_1}$ and $A_2 e^{i\phi_2}$, magnitude and phase of the reflected TE and TM polarizations, respectively), a process physically correct under the condition that the unit cell is a subwavelength entity. This simply implies that the input of the "physical size" field in Fig. 11(a.1) must comply with this specification or a warning message will appear. In our example, the selected tile consists of binary elements (D stands for diodes), which may only control the phase of the co-polarized scattered field (ϕ_1 term). By clicking the "View suggested" button, we analytically calculate and display $\phi_0 - \phi_1$ (ϕ_0 is the phase of the incoming wavefront), which should give an indication of the diode states at the metamaterial. Alternatively, a similar result can be extracted by launching the metaheuristic optimizer, either via a blind optimization process (all-0 initial solution) or an assisted optimization, using the analytically evaluated profile as an initial solution. The latter practice leads to more refined results, over an analytical evaluation, by considering the finite size of the tile and the non-infinitesimal size of the unit cell.



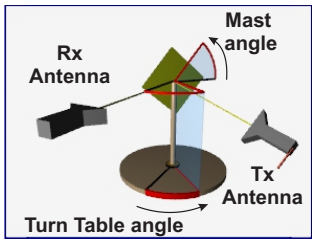
(a)

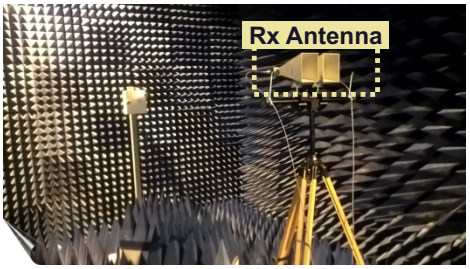


(b)



(c)





(d)

Fig. 11 – (a) Snapshots of the Metamaterial Middleware GUI during the metamaterial characterization process. The user may select an existing tile or create a new one by drawing its unit cell topology and variable elements within. A new functionality for this tile is assigned by selecting the desired parameters for the impinging and outgoing waves. The process begins by evaluating the scattering response profile, calculated as a phase and amplitude array for the selected metamaterial function. During the final step the Metamaterial Middleware seeks the set of optimal states that can realize the scattering response into an actual hardware configuration. (b) After a successful characterization process the configuration is saved to the DB. Herein, as an instance, a snapshot of the DB manager that was created during development of the software is presented. (c) The Metamaterial Middleware offers extensive experimental capabilities through a dedicated module. (d) Utilizing the hardware present in an anechoic chamber, the Metamaterial Middleware was able to acquire full scattering diagrams of a metamaterial tile, mounted on the available positioning equipment. The three-dimensional (3D) schematic, shown on the left, is updated live as the turntable or the positioner head rotates during a measurement or optimization process. On the right hand side, the Metamaterial Middleware host (white laptop) connects via Ethernet to the VNA (measuring the Rx/Tx antennas) and to the positioner controller.

In the final step, we seek to match the scattering field response (i.e. the $(A_1 e^{i\phi_1}, A_2 e^{i\phi_2})_{i,j}$ pairs calculated in the previous step) to the appropriate set of element states, such as the resistance R and reactance X (capacitance or inductance) of the loads, for all unit cells indexed by i, j. The correspondence between the scattering response of a unit cell and its physical structure constitutes a highly complex and demanding propagation problem. Actually, the search for a proper set of states for a fully defined response constitutes an inverse problem with closed-form solutions available only for very simple unit cell types. As such, it demands the use of highly efficient optimization algorithms and strong computation power to perform the necessary numerical simulations. In our implementation, we employ a cluster of interconnected computer clients (that serve as simulation nodes) receiving instructions through a TCP/IP network from the host Metamaterial Middleware running on the main machine. In a lightweight scenario, a single-cell simulation assigned to one of the clients can be completed in less than a minute but this may grow substantially when the complexity of the design is increased. An acceptable convergence is expected to be reached within a few thousands simulation trials. In perspective, the prototype studied in this work displayed a typical evaluation time of 8 to 12 hours for a full characterization of its configuration space and all possible impinging plane waves (θ,ϕ sets) at the operation frequency of the hardware. The evaluation is performed through Algorithm 1 (see next page) whose steps are described below. Specifically, the user:

- selects one of the three options: i) a gradient based optimizer, recommended for a small number of variables (less than four), ii) a metaheuristic optimizer, recommended for a higher count of variables (more than three), and iii) a database-based optimization process that configures the proper switch-states, through simulation results already stored in the configuration DB.

- suggests a convergence limit Tol to the optimizer in the Tolerance text box for all variable targets,

- fills the IP list with all available simulation nodes and initiates a connection. All nodes will launch the installed simulation software, open the corresponding geometry model, and remain idle until further instructions are received.

- modifies, if necessary, the value range, step and initial value for each variable in the unit cell.

- begins the optimization process by clicking the "RUN" button. The software will iterate over all unit cells (from top-left $(0, 0)$ to bottom-right (N, M)), seeking a set of optimal values for the variables of each cell. During this sequence, each unit cell (i, j) initiates an independent optimization process and a new series of simulations begins until a sufficiently close convergence to the targeted scattering field response S_{ij}=$(A_1 e^{i\phi_1}, A_2 e^{i\phi_2})_{i,j}$ is achieved. Prior to each simulation, the optimizer looks up all entries in the associated table of the configuration DB, in case a result (S_{db}) is already stored. If not, the simulation will start and the result will be stored afterwards. Over time, this process can populate the DB with enough results, making option (iii) of the initial step a highly efficient method for evaluating new functionalities for this particular tile.

Algorithm 1 Physical Element Calibration. Highlighted commands run in a separate thread.

```
Set index to i = 0, j = 0
while index does not exceed max do
    Get S_ij; Get Tol
    if !(exists S_db where |S_db − S_ij| < Tol) then
        while searching do
            Select node where state == free
            Set to busy
            Initiate parallel thread;
            Optimizer: Fixes (R, L, C, D) variables;
            Optimizer → Requests new simulation;
            node ← Returns S_sim
            S_db = S_sim;
            if |S_sim − S_ij| < Tol then
                searching = false
                Set (R, L, C, D) variables as optimal
            end if
            Set node state to free
        end while
        Iterate index
    end if
end while
```

By completing the previous steps, the software has successfully defined a new configuration for the chosen tile, which, now, remains to be stored in the configuration DB. Before doing so, we can apply an additional evaluation step by conducting an experimental measurement on a physical prototype (if available). Our current implementation is able to assess multi-splitting or absorption functionalities by measuring the full scattering diagram on the front metamaterial hemisphere, which can then be compared to the scattering profile produced by the software. As presented in the corresponding technical report [20], several tests with a metamaterial unit have already been successfully conducted in a fully equipped anechoic chamber (Fig. 11). This test bed incorporates a variety of algorithms, meeting individual needs for accuracy and speed for various cases of scattering response measurements (e.g. a full 3D-pattern versus a 2D-slice might be required for arbitrary lobe scattering). A simple case is outlined in Algorithm 2, where the user has previously set the appropriate parameters in the GUI (Fig. 11(c)). The GUI automates

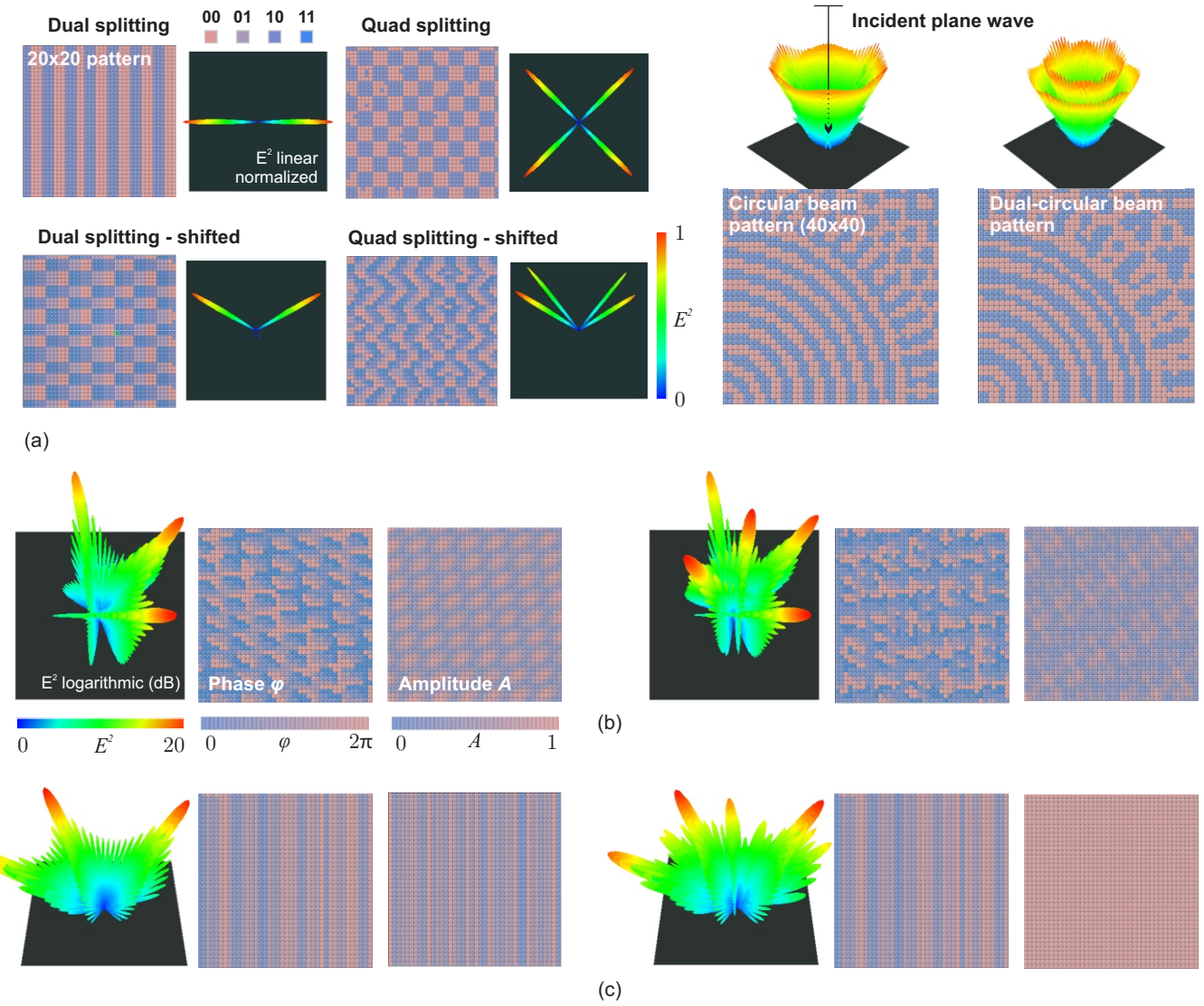


Fig. 12 – (a) Element states and far-field scattering diagram of a 4-bit metamaterial array for 6 different cases. (b) Element states and scattering diagrams for triple- and quintuple-beam scattering, assuming continuously adjustable states with absorption capabilities. (c) Comparison between a tile with resistor elements (left) and a tile without (right) for an in-plane triple-beam splitting functionality. A non-uniform amplitude pattern (A_{mn}) can eliminate all side-lobes and provide increased security to the signal. The incident field in all cases is a vertically impinging plane wave, chosen for simplicity; yet any plane wave or point source input can be considered by a corresponding shift of the phase for each unit cell.

Algorithm 2 Evaluate Scattering Response, 3D, slow rotation case. Highlighted commands refer to HW instructions.

Set $N \leftarrow$ number of equidistant points on the hemisphere
Set $P[N] \leftarrow$ Struct of (θ, ϕ) points
for all elements i in P **do**
  Send rotation command $\rightarrow$ Positioner
  (Mast - $P[i]$.theta, Head - $P[i]$.phi, Speed)
  Receive $\leftarrow$ Positioner feedback
  Refresh 3D Figure
  Send measurement command $\rightarrow$ VNA
  $E[i]$ = **Receive** $\leftarrow$ S21 power
  Refresh Plot graph
end for
Save ($P[i]$,$E[i]$) to DB

the process of measuring metamaterial devices in the anechoic chamber by supporting several communication interfaces for the following equipment:

- Vector network analyzer (VNA): produces the energy signal and receives the response (i.e. S_{21}-parameter) from the antenna setup.
- Positioner: allows the mechanical support of the metamaterial and antenna devices. Its controller can instruct the rotation of both heads (towards θ, ϕ), allowing a complete characterization of the scattering profile.
- metamaterial controllers: A metamaterial hosting reconfigurable elements incorporates a communication network for the explicit control of its element states. The Metamaterial Middleware implements the proper interface for the evaluation prototypes (serial port connection, WiFi, and Bluetooth have been integrated). The same interfaces are, also, used for the metamaterial API developed in Section 4.

A supplementary note is that the final switch-state configuration can be re-evaluated using the same metaheuristic optimizer utilized in step 3 of Fig. 12(a) via actual experimental results. The optimizer starts with the software-defined configuration as an initial solution and gradually adjusts the switch-state matrices to more optimally converge to the pursued functionality under true operational conditions. The implemented algorithm follows the template of Algorithm 2, where N correspond to the number or optimization variables (e.g. the number of scattering lobes) and a second for loop nests the existing loop, seeking to maximize the $Sum_i(E[i])$ metric.
For further evaluation purposes, we validate a number of indicative examples from the literature, based on previously measured and simulated results. Hence, Fig. 12(a) presents the optimization outcome for a 4-bit metamaterial array, which can switch over four available states per

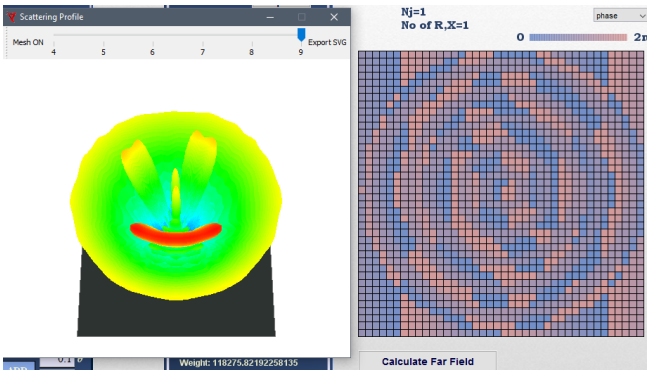


Fig. 13 – Arbitrary functionality optimization test. A smiley face-shaped scattering pattern is successfully produced (left). The corresponding metamaterial element configuration (different meta-atom states expressed in colormap) is shown to the left.

unit cell with reflection phases $(-90, 0, 90, 180)$ and full reflection amplitude $(A = 1)$ [14,38]. The outcome complies fully with the results provided by the corresponding authors, indicating that the Metamaterial Middleware may cooperate with arbitrary hardware configurations and thus be compatible with any reasonable design in a future diverse metamaterial market. Following these results, we, also, test four exclusive cases that highlight the additional capabilities of our software. In particular, Fig. 12(b) displays the results for a triple- and a quintuple-beam splitting case, while Fig. 12(c) shows the optimization outcome for an in-plane triple-beam splitting functionality. For the latter case, the integrated theoretical algorithms were able to eliminate all side lobes by suggesting a non-uniform pattern for the amplitude A of the co-polarized scattered field. This particular case demands a tile with controllable absorption elements (resistors). Finally, in Fig. 13 we proceed to showcase the optimization of an arbitrary departing wavefront formation. A smiley-face shaped scattering pattern is selected as the required energy wave response of the metamaterial to a planar impinging wave. The optimization process successfully produces the required wavefront, and the corresponding metamaterial element states are shown to the right of Fig. 13.

7. DISCUSSION: THE TRANSFORMATIONAL POTENTIAL OF THE IOMMT AND FUTURE DIRECTIONS

While the potential of the IoMMT paradigm alone may be worth the investigation, here we evaluate its practical opportunities affecting the industry, the end users and the environment, namely:

- How can the IoMMT prolong the life cycle of products across deployment scales?
- How can the IoMMT help maintain a high-speed product development pace, without sacrificing ecological concerns during the product design phase?

To these ends, we believe that the concept of Circular Economy (CE) and its associated performance indices is a fitting framework for the initial exploration and evaluation of the IoMMT paradigm [39, 40]. CE seeks to make technological products reusable, repairable and recyclable across their lifetime (i.e., development, purchase, usage and disposal) by introducing cross-product and cross-manufacturer interactions. Instead of the traditional, linear order of life cycles phases, i.e., i) raw resource acquirement, ii) processing, iii) distribution, iv) its use and v) disposal, the CE advocates to create links from disposal to all preceding phases, promoting i) reprocessing or refurbishment, ii) redeployment and redistribution, and iii) multiple uses.

However, according to the literature [41], the CE introduces a paradoxical tension in the industry: While the industry is pressed for faster growth and, hence, a faster product development rate, CE can introduce a series of design considerations that make for a slower product development rate. In this view, the paradigm of IoMMT is by its nature impactful for the energy and ecological footprint of multiple products, across disciplines and scales.

The fact that it enables the tuning and optimization of the physical properties of matter allows for a tremendous impact both in terms of quality of service per product and scale, but also for energy savings in a horizontal manner. Moreover, the IoMMT can contribute a software-driven way for optimizing material properties. Using this new technology, the industrial players can maintain a fast-paced product design, where energy efficiency and sustainability can be upgraded programmatically via "eco-firmware" during its use, thereby offloading the product design phase of such concerns.

An important future research direction of the IoMMT is to quantify the financial savings stemming from adopting this technique. While the Circular Economy-derived paradoxical tensions have been around for long and are hard to eradicate, IoMMT can facilitate their resolution by quantifying them, potentially aligning fast-paced marketing with environmental sustainability.

Apart from the CE line of work, future work will seek to provide the theoretical and modeling foundations of the IoMMT. Our vision, overviewed in Fig. 14, is for a full-stack study of this new concept, covering: the physical layer modeling, the internetworking and communications layer, and the application layer.

At the physical layer, future research needs to classify and model metamaterials in a functionality-centric way, and introduce fitting Key-Performance Indicators for each offered energy manipulation type. Metasurface-internal control variations (technologies and ways of monitoring embedded active elements) can be taken into account, and aspire to deduce models covering the aspects of control data traffic, energy expenditure and feature-based manufacturing cost estimation. This will enable the creation of the first, cross-physical domain profiling and recommendation system for metamaterials, to match the requirements and specifications of any envisioned application.

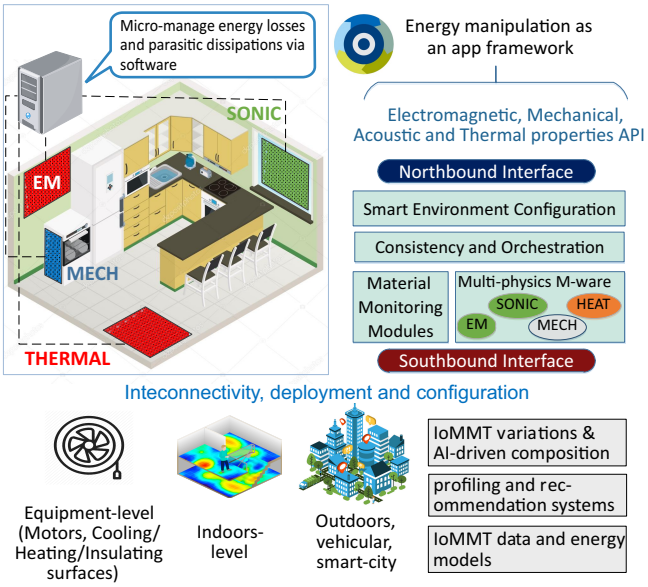


Fig. 14 – Envisioned future research directions for the Internet of MetaMaterials.

Moving to the networking layer, research can follow the proposed northbound/southbound abstraction model inspired from the SDN paradigm. This will provide the necessary platform for: i) Interconnecting the IoMMT to the vast array of existing networked devices and assorted standards, models and protocols. ii) Provide the necessary software abstractions, to open the field of energy micromanagement to software developers (i.e., without specialty in Physics), enabling the energy-propagation-as-an-app paradigm. In this aspect, we also envision the need for algorithms optimizing IoMMT deployments for minimal-investment-maximal-control, and orchestrating IoMMT deployments for any set of generic energy micro-management objective.

The control time granularity depends on the application scenario and the volatility of the factors affecting the energy propagation within an environment. In an indoors wireless communications setting, such as the one studied in [7], where the Intelligent Wireless Environment needs to continuously adapt to the position of user devices, the control granularity can be considered to be 10-25 msec (i.e., randomly walking users running a mobile application).

Finally, at the application layer research can define key-applications of the IoMMT at multiple scales, starting from the internals of equipment, such as motors, heating, cooling and insulating surfaces. This will provide the basic units for IoMMT incorporation to devices spanning home appliances (ovens, refrigerators, washers, heating and cooling units), electronics (from interference cancellation, to smart cooling) and building materials (acoustic, thermal and mechanical insulators). Various scales can be taken into consideration, from indoor (smart-house) and outdoor (city-level, ve-

hicular networks) IoMMT deployments.

8. CONCLUSION

In this paper, we pursue the systematic expansion of the concept of software-defined metamaterials by establishing the key software elements of a metamaterial network. With this goal in mind, we introduce two special categories of software, the Metamaterial Middleware, which can methodically produce novel configurations for a single metamaterial tile and the metamaterial API, which is in charge of supervising the energy propagation within a metamaterial-coated space. The two components were studied and developed autonomously, and then merged into a unified application via a common layer of abstraction.

For the API, we explored the means to interpret a metamaterial, its configuration space, and the supported energy manipulation functionalities via a group of well-defined software objects. The key objective was to successfully conceal the physical layer of the metamaterial, and only expose the essential parameters for configuring an operation. In light of this definition, the API is capable of instructing its environment to steer, focus, absorb or split the incoming signals through a simple interface, without any reference to the underlying physics.

The Metamaterial Middleware was developed by means of various theoretical and computational tools, including analytical algorithms that assess the scattering response, full-wave simulations that accurately evaluate a unit cell response, and experimental modules that can perform physical measurements of metamaterial devices. Through a step-by-step process we defined a robust characterization methodology, supporting a large set of metamaterial operations with no restriction on the specifications for the metamaterial. We also discuss the prospect of applying different optimization algorithms for each stage of the metamaterial characterization process. Finally, the evaluation of the middleware outcomes demonstrated that the IoMMT paradigm can be employed successfully within a unified framework.

ACKNOWLEDGMENT

This work was supported by the European Union's Horizon 2020 research and innovation programme-project C4IIoT, GA EU833828. The authors also acknowledge FETOPEN-RIA project VISORSURF. All presented software modules were conceived, designed and implemented in full by the Foundation for Research and Technology-Hellas (FORTH) and G. Pyrialakos served as the lead developer.

REFERENCES

[1] J. Pan, R. Jain, S. Paul, T. Vu, A. Saifullah, and M. Sha, "An internet of things framework for smart energy in buildings: Designs, prototype, and experiments," *IEEE Internet Things J.*, vol. 2, no. 6, pp. 527–537, Dec. 2015. [Online]. Available: https://doi.org/10.1109/jiot.2015.2413397

[2] F. Capolino, *Theory and phenomena of metamaterials*. CRC press, 2017.

[3] A. Li, S. Singh, and D. Sievenpiper, "Metasurfaces and their applications," *Nanophotonics*, vol. 7, no. 6, pp. 989–1011, 2018.

[4] M. Kadic, T. Bückmann, R. Schittny, and M. Wegener, "Metamaterials beyond electromagnetism," *Reports on Progress in physics*, vol. 76, no. 12, p. 126501, 2013.

[5] J. U. Surjadi, L. Gao, H. Du, X. Li, X. Xiong, N. X. Fang, and Y. Lu, "Mechanical metamaterials and their engineering applications," *Advanced Engineering Materials*, vol. 21, no. 3, p. 1800864, 2019.

[6] M. Pishvar and R. L. Harne, "Foundations for soft, smart matter by active mechanical metamaterials," *Advanced Science*, p. 2001384, 2020.

[7] C. Liaskos, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "Using any surface to realize a new paradigm for wireless communications," *Commun. ACM*, vol. 61, pp. 30–33, 2018.

[8] C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "On the network-layer modeling and configuration of programmable wireless environments," *IEEE/ACM Trans. Netw.*, vol. 27, no. 4, pp. 1696–1713, 2019.

[9] I. F. Akyildiz and J. M. Jornet, "The internet of nano-things," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 58–63, 2010.

[10] S. B. Glybovski, S. A. Tretyakov, P. A. Belov, Y. S. Kivshar, and C. R. Simovski, "Metasurfaces: From microwaves to visible," *Phys. Rep.*, vol. 634, pp. 1–72, 2016.

[11] H.-T. Chen, A. J. Taylor, and N. Yu, "A review of metasurfaces: Physics and applications," *Rep. Prog. Phys.*, vol. 79, no. 7, pp. 076 401(1–40), 2016.

[12] C. Huang, C. Zhang, J. Yang, B. Sun, B. Zhao, and X. Luo, "Reconfigurable metasurface for multifunctional control of electromagnetic waves," *Adv. Opt. Mater.*, vol. 5, no. 22, pp. 1 700 485(1–6), 2017.

[13] Ö. Özdogan, E. Björnson, and E. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 581–585, 2019.

[14] Q. Zhang, X. Wan, S. Liu, J. Yin, L. Zhang, and T. Cui, "Shaping electromagnetic waves using software-automatically-designed metasurfaces," *Sci. Rep.*, vol. 7, no. 1, pp. 3588(1–11), 2017.

[15] A. P. Mosk, A. Lagendijk, G. Lerosey, and M. Fink, "Controlling waves in space and time for imaging and focusing in complex media," *Nature photonics*, vol. 6, no. 5, pp. 283–292, 2012.

[16] X. Tan, Z. Sun, D. Koutsonikolas, and J. M. Jornet, "Enabling indoor mobile millimeter-wave networks based on smart reflect-arrays," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 270–278.

[17] C. Liaskos, A. Tsioliaridou, A. Pitsillides, I. F. Akyildiz, N. V. Kantartzis, A. X. Lalas, X. Dimitropoulos, S. Ioannidis, M. Kafesaki, and C. Soukoulis, "Design and development of software defined metamaterials for nanonetworks," *IEEE Circuits and Systems Magazine*, vol. 15, no. 4, pp. 12–25, 2015.

[18] F. Liu *et al.*, "Programmable metasurfaces: State of the art and prospects," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 2018-May. IEEE, May 2018, p. 8351817.

[19] T. Saeed, V. Soteriou, C. Liaskos, A. Pitsillides, and M. Lestas, "Toward fault-tolerant deadlock-free routing in hypersurface-embedded controller networks," *IEEE Networking Letters*, vol. 2, no. 3, pp. 140–144, 2020.

[20] The VISORSURF project consortium, "A hypervisor for metasurface functionalities: progress report," *European Commission Project VISORSURF: Public Report, Aug-2019, [Online:] https://ec.europa.eu/research/participants /documents/downloadPublic? documentIds=080166e5c560b376&appId=PPGMS*.

[21] T. Frenzel, M. Kadic, and M. Wegener, "Three-dimensional mechanical metamaterials with a twist," *Science*, vol. 358, no. 6366, pp. 1072–1074, 2017.

[22] M. R. Haberman and M. D. Guild, "Acoustic metamaterials," *Physics Today*, vol. 69, no. 6, pp. 42–48, Jun. 2016. [Online]. Available: https://doi.org/10.1063/pt.3.3198

[23] M. S. Dresselhaus, G. Chen, M. Y. Tang, R. Yang, H. Lee, D. Wang, Z. Ren, J.-P. Fleurial, and P. Gogna, "New directions for low-dimensional thermoelectric materials," *Advanced materials*, vol. 19, no. 8, pp. 1043–1053, 2007.

[24] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. F. Akyildiz, "Mobility-aware beam steering in metasurface-based programmable wireless environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 9150–9154.

[25] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "End-to-end wireless path deployment with intelligent surfaces using interpretable neural networks," *IEEE Transactions on Communications*, 2020.

[26] F. Mathioudakis, C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "Advanced physical-layer security as an app in programmable wireless environments," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[27] Z. Chen, B. Guo, Y. Yang, and C. Cheng, "Metamaterials-based enhanced energy harvesting: A review," *Physica B: Condensed Matter*, vol. 438, pp. 1–8, 2014.

[28] B. Florijn, C. Coulais, and M. van Hecke, "Programmable mechanical metamaterials," *Phys. Rev. Lett.*, vol. 113, no. 17, pp. 175 503(1–4), 2014.

[29] S. A. Cummer, J. Christensen, and A. Alù, "Controlling sound with acoustic metamaterials," *Nat. Rev. Mat.*, vol. 1, no. 3, pp. 16 001(1–13), 2016.

[30] L. Zhang, X. Q. Chen, S. Liu, Q. Zhang, J. Zhao, J. Y. Dai, G. D. Bai, X. Wan, Q. Cheng, G. Castaldi, V. Galdi, and T. J. Cui, "Space-time-coding digital metasurfaces," *Nat. Commun.*, vol. 9, no. 1, pp. 4334(1–11), 2018.

[31] Y. E. Oktian, S. Lee, H. Lee, and J. Lam, "Distributed SDN controller system: A survey on design choice," *Computer Netw.*, vol. 121, pp. 100–111, 2017.

[32] C. Liaskos, A. Pitilakis, A. Tsioliaridou, N. Kantartzis, and A. Pitsillides, "Initial UML definition of the hypersurface compiler middleware," *European Commission Project VISORSURF: Public Deliverable D2.2, 31-Dec-2017, [Online:] http://www.visorsurf.eu/m/VISORSURF-D2.2.pdf*.

[33] N. Ashraf, M. Lestas, T. Saeed, H. Taghvaee, S. Abadal, A. Pitsillides, and C. Liaskos, "Extremum seeking control for beam steering using hypersurfaces," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2020, pp. 1–6.

[34] M. Laguna and R. Marti, "The OptQuest callable library," in *Optimization Software Class Libraries*. Springer, 2003, pp. 193–218.

[35] C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "An interpretable

neural network for configuring programmable wireless environments," in *IEEE Int. Workshop Signal Processing Adv. Wireless Commun. (SPAWC 2019)*, 2019, pp. 1–5.

[36] H. Taghvaee, A. Jain, X. Timoneda, C. Liaskos, S. Abadal, E. Alarcón, and A. Cabellos-Aparicio, "Radiation pattern prediction for metasurfaces: A neural network based approach," *arXiv preprint arXiv:2007.08035*, 2020.

[37] Y. Saifullah, A. B. Waqas, G.-M. Yang, F. Zhang, and F. Xu, "4-bit optimized coding metasurface for wideband rcs reduction," *IEEE Access*, vol. 7, pp. 122 378–122 386, 2019.

[38] S. Liu, T. J. Cui, L. Zhang, Q. Xu, W. Qiu, X. Wan, J. Q. Gu, W. X. Tang, M. Q. Qi, J. G. Han, W. L. Zhang, X. Y. Zhou, and Q. Cheng, "Convolution operations on coding metasurface to reach flexible and continuous controls of terahertz beams," *Adv. Science*, vol. 3, no. 10, pp. 1 600 156(1–12), Jul. 2016.

[39] W. R. Stahel, "The circular economy," *Nature*, vol. 531, no. 7595, pp. 435–438, 2016.

[40] C. Liaskos, A. Tsioliaridou, and S. Ioannidis, "Towards a circular economy via intelligent metamaterials," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2018, pp. 1–6.

[41] T. Daddi, D. Ceglia, G. Bianchi, and M. D. de Barcellos, "Paradoxical tensions and corporate sustainability: A focus on circular economy business cases," *Corporate Social Responsibility and Environmental Management*, vol. 26, no. 4, pp. 770–780, 2019.

AUTHORS

Christos Liaskos received a Diploma in Electrical and Computer Engineering from the Aristotle University of Thessaloniki (AUTH), Greece in 2004, an MSc degree in Medical Informatics in 2008 from the Medical School, AUTH and PhD degree in Computer Networking from the Dept. of Informatics, AUTH in 2014. He has published work in several journals and conferences, such as IEEE Transactions on: Networking, Computers, Vehicular Technology, Broadcasting, Systems Man and Cybernetics, Networks and Service Management, Communications, INFOCOM. He is currently an assistant professor with the University of Ioannina, Greece, and an affiliated researcher at the Foundation of Research and Technology, Hellas (FORTH). His research interests include computer networks, security and nanotechnology, with a focus on developing nanonetwork architectures and communication protocols for future applications.

George Pyrialakos received his Diploma and Ph.D. degree from the Dept. of Electrical and Computer Engineering, Aristotle University of Thessaloniki (AUTH), Greece, in 2013 and 2019, respectively. He is currently a post-doc researcher in AUTH, affiliated with the Foundation for Research and Technology Hellas (FORTH). His research interests include computational electromagnetics, metamaterials, photonics, and condensed matter physics.

Alexandros Pitilakis received his Diploma and Ph.D. degree from the Dept. of Electrical and Computer En-

gineering, Aristotle University of Thessaloniki (AUTH), Greece, in 2005 and 2013. He, also, holds an MSc. degree from the ENST Paris, 2007. He is a post-doc researcher in AUTH, affiliated with the Foundation for Research and Technology Hellas (FORTH). His research interests include computational electromagnetics, metamaterials, and nonlinear optics.

Ageliki Tsioliaridou received a Diploma and PhD degrees in Electrical and Computer Engineering from the Democritus University of Thrace (DUTH), Greece, in 2004 and 2010, respectively. Her research work is mainly in the field of Quality of Service in computer networks. Additionally, her recent research interests lie in the area of nanonetworks, with a specific focus on architecture, protocols, security and authorization issues. She has contributed to a number of EU, ESA and National research projects. She is currently a researcher at the Foundation of Research and Technology, Hellas (FORTH).

Michail Christodoulou received his Diploma and Ph.D. degree in Electrical and Computer Engineering from the Dept. of Electrical and Computer Engineering, Aristotle University of Thessaloniki (AUTH) in 1994 and 2006, respectively. In 1995, he received an MSc. degree from the University of Bradford, UK. In 2014, he joined AUTH as a Postdoctoral Research Fellow. His research interests include RF circuits, antennas, RF MEMS, and electromagnetic measurements.

Nikolaos Kantartzis is a professor at the School of Electrical and Computer Engineering, Aristotle University of Thessaloniki (AUTH), Greece, from where he received the Diploma and Ph.D. degrees in 1994 and 1999, respectively. His primary research interests include computational electromagnetics, metamaterials, graphene, EMC/EMI problems, microwaves, and nano devices. Dr. Kantartzis is a Senior Member of IEEE, an ICS and ACES member.

Sotiris Ioannidis received a BSc degree in Mathematics and an MSc degree in Computer Science from the University of Crete in 1994 and 1996 respectively. In 1998 he received an MSc degree in Computer Science from the University of Rochester and in 2005 he received his PhD from the University of Pennsylvania. Ioannidis held a Research Scholar position at the Stevens Institute of Technology until 2007, and since then he is Research Director at the Institute of Computer Science of the Foundation for Research and Technology - Hellas. Since November 2017 he is a member of the European Union Agency for Network and Information Security (ENISA) Permanent Stakeholders Group (PSG). His research interests are in the area of systems, networks, and security. Ioannidis has authored more than 100 publications in international conferences and journals, as well as book chapters, including ACM CCS, ACM/IEEE ToN, USENIX ATC, NDSS, and has both chaired and served in numerous program committees in prestigious international conferences. Ioannidis is a Marie-Curie Fellow and has participated in numerous international and European projects. He has coordinated a number of European and National projects (e.g. PASS, EU-INCOOP, GANDALF, SHARCS) and is currently the project coordinator of the THREAT-ARREST, I-BiDaaS, BIO-PHOENIX, IDEAL-CITIES, CYBERSURE, and CERTCOOP European projects.

Andreas Pitsillides is a Professor in the Department of Computer Science, University of Cyprus, heads NetRL, the Networks Research Laboratory he founded in 2002, and is appointed Visiting Professor at the University of the Witwatersrand (Wits), School of Electrical and Information engineering, Johannesburg, South Africa. Earlier (2014-2017) Andreas was appointed Visiting Professor at the University of Johannesburg, Department of Electrical and Electronic Engineering Science, South Africa. His broad research interests include communication networks (fixed and mobile/wireless), Nanonetworks and Software Defined Metasurfaces/Metamaterials, the Internet- and Web- of Things, Smart Spaces (Home, Grid, City), and Internet technologies and their application in Mobile e-Services, especially e-health, and security. He has a particular interest in adapting tools from various fields of applied mathematics such as adaptive non-linear control theory, computational intelligence, game theory, and recently complex systems and nature inspired techniques, to solve problems in communication networks. He has published over 270 referred papers in flagship journals (e.g. IEEE, Elsevier, IFAC, Springer), international conferences and book chapters, 2 books (one edited), participated in over 30 European Commission and locally funded research projects as principal or co-principal investigator, received several awards, including best paper, presented keynotes, invited lectures at major research organisations, short courses at international conferences and short courses to industry, and serves/served on several journal and conference executive committees.

Ian F. Akyildiz received his BS, MS, and PhD degrees in Electrical and Computer Engineering from the University of Erlangen-Nï¿œrnberg, Germany, in 1978, 1981 and 1984, respectively. Currently he serves as a Consulting Chair Professor with the Computer Science Department at the University of Cyprus since January 2017.

He is also the Megagrant Leader with the Institute for Information Transmission Problems at the Russian Academy of Sciences, in Moscow, Russia, since May 2018. He serves on the Advisory Board for the newly established research center called Technology Innovation Institute (TII) in Abu Dhabi, United Arab Emirates since June 2020. He is the President of the Truva Inc. since March 1989 and Scientific Advisor for the newly established company Airanaculus since April 2020. He is a Visiting Distinguished Professor with the SSN College of Engineering in Chennai, India since October 2019. Dr. Akyildiz is an Adjunct Professor with the Department of Electrical and Computer Engineering at the University of Iceland since September 2020. He is the Ken Byers Chair Professor Emeritus in Telecommunications at the Georgia Institute of Technology, and the Director of the Broadband Wireless Networking Laboratory and Former Chair of the Telecom Group from 1985-2020. He is also former Finnish Distinguished Professor with the University of Tampere, Finland, supported by the Finnish Academy of Sciences from 2012-2016. He is the founder of NanoNetworking Center and Former Honorary Professor at the University of Politecnica de Cataluna in Barcelona from 2008-2017. Dr. Akyildiz is also former Distinguished Professor and Founder of the Advanced Wireless Networks Lab with the King Abdulaziz University in Jeddah, Saudi Arabia from 2011-2018. He is also the Founder of the Advanced Wireless Sensor Networks lab and former ExtraOrdinary Professor with the University of Pretoria, South Africa from 2007-2012.

He is the Founder and Editor in Chief of the newly established ITU J-FET (International Telecommunication Union Journal for Future and Evolving Technologies since July 2020. Dr. Akyildiz is Editor-in-Chief Emeritus of Computer Networks Journal (Elsevier) (1999-2019), the founding Editor-in-Chief Emeritus of the Ad Hoc Networks Journal (Elsevier) (2003-2019), the founding Editor-in-Chief Emeritus of the Physical Communication (PHYCOM) Journal (Elsevier) (2008-2017), and the founding Editor-in-Chief Emeritus of the Nano Communication Networks (NANOCOMNET) Journal (Elsevier) (2010-2017). Dr. Akyildiz launched many IEEE and ACM conferences. He is an IEEE Fellow (1995) and ACM Fellow (1996) and received numerous awards from IEEE and ACM and other professional organizations. His current research interests are in 6G Wireless Systems, Reconfigurable Intelligent Surfaces, Terahertz Communications, Nanonetworks, Internet of xThings (x=Underwater, Underground, Space/CubeSats, Nano and BioNano). He graduated 45 PhD students and 30 of them are in academia in very prestigious academic positions worldwide. He advised 13 Postdoctoral researchers. According to Google Scholar as of September 2020, his H-index is 124 and the total number of citations to his papers is 119+K. His rank in terms of h-index in the world is 46 and in the USA 32.

DESIGN AND ANALYSIS OF A RECONFIGURABLE INTELLIGENT META-SURFACE FOR VEHICULAR NETWORKS

Mohammad Ojaroudi[1], Valeria Loscrí[1], Anna Maria Vegni[2]
[1]Inria Lille - Nord Europe, [2]Roma Tre University, Rome, Italy

NOTE: Corresponding author: Valeria Loscrí, valeria.loscri@inria.fr

Abstract – *In this work, a new paradigm for vehicular communications based on Reconfigurable Intelligent Meta-surfaces (RIMs) is presented. By using the proposed RIM, we are able to manipulate electromagnetic waves in the half-space, since the element is reflective. The unit cell consists of a U-shaped designed microstrip structure equipped with a pin diode and via a hole. In this study, two different reflection modes are achieved for 1-bit data transferring in each state. By incorporating these two different configurations together, the reflected phases in the proposed RIM surface can be controlled respectively in* 0° *and* 180°*. The proposed unit cell can provide a usable double negative functional characteristic around* 5.3 *GHz. The main goal of this paper is the use of a multifunctional behavior RIM for vehicular communications to code the transmitted wave. A novel phase distribution diagram is generated to propagate in each angle. Moreover, two major electromagnetic modulation functions, beam forming and space coding have been demonstrated. Finally, we show how the RIM can be employed for vehicular communications, acting as a coated access point along the street. We derive the instantaneous data rate at the receiver node, the outage probability and the channel capacity, as affected by different beam widths, distances and vehicle speed.*

Keywords – Beam width, Instantaneous data rate, Reconfigurable meta-surfaces, V2X communications

1. INTRODUCTION

In recent years, special attention has been paid to driverless cars using automotive radars [7, 8]. A Vehicle-to-Everything (V2X) communication paradigm has emerged as one of the most important enabling technologies for vehicular networks. This has the potential of making streets and highways safer, the traffic more efficient and less harmful to the environment [9]. An example is the exchange of traffic information, as in the Self-Organizing Traffic Information System approach [12].

One of the most important challenges arising in the V2X paradigm is the high dynamic and the need of highly efficient communication solutions in order to "follow" the rapid changing of communication nodes. Recently, an innovative wireless communication paradigm based on the utilization of specific features of controllable meta-surfaces has been investigated in [2, 3, 4, 5, 6, 15]. With the advancement of technology in metamaterials, meta-surfaces which are a thin two-dimensional structure of metamaterials [1], due to their unique properties like the capability to provide abrupt phase shift, amplitude modulation, and polarization conversion of the electromagnetic (EM) wave, have found a variety of applications in various fields of science, such as physics, engineering, biotechnology, and telecommunications [18].

Reconfigurable Intelligent Meta-surfaces (RIMs) can be very effective for improving the performance of a V2X communication system, especially when the Line-of-Sight of the propagating signal is not available. [22] and [23] are among the first contributions of exploitation of RIM in vehicular communications. In [22], the authors focus on the physical layer security based on the exploitation of a RIM or RIS (Reconfigurable Intelligent Surface). They derive the average secrecy capacity of the system, by considering an ideal reconfigurable meta-surface. They do not give any detail about the design of a real meta-surface with the needed characteristics in order to obtain the specific behavior as described in the paper. Also in [23], the authors show the rich potentiality of the RIM integration in a vehicular communication system, by analysing a hypothetic ideal controllable meta-surface. Anyway, no detail about the features of a real meta-surface is provided in the paper. In [17], we have proposed a unit cell configuration working at millimiter-waves in order to control the phase shift from 0 to 180°. Specifically, we have investigated the unit cell design without evaluating the full structure and relate the capability to control the beam width based on the number of the unit cells of the full structure. Indeed, the higher is the number of the unit cells of the full structure, the better is the capacity to control the beam width of the radiation pattern generated as the reflected wave of an impinging signal.

As demonstrated in [11], it is possible to accurately model the number of cells as sources of secondary radiation [16] by applying the Huygens principle in the far-field limit. In practice, the higher is the number of unit cells we consider in the full structure, the better is the capability to control the phase, but above all to control the beam width of the

main beam for beam stearing objective.

In [18], it is demonstrated how this type of meta-surface allows a digital control of EM waves, by associating two coding elements with opposite reflection phases (*i.e.*, 0° and 180°) and considering them as digital bits (*i.e.*, 0 and 1 in the binary case). Reconfigurable meta-surface structures can be applied to manipulate EM waves in a simple and effective way, by changing the coding elements on a 2D plane with predesigned coding sequences [19]. In order to independently control and create different coding sequences a Field Programmable Gate Array (FPGA) is used. By changing the coding sequences stored in the FPGA, many different functionalities can be switched in real time, thereby leading to programmable meta-surfaces.

In this paper, we formulate the specific features that a meta-surface for vehicular communication applications needs to have in the frequency range of $[5, 5.9]$ GHz. In particular, we consider a tracking application with beam steering, for which is of paramount importance to control the phase and to concentrate the power in the main lobe of the reflected signal, as much as possible. In order to meet these specific requirements, we will design a unit cell and then derive a full structure constituted by the periodic repetitions of these unit cells, behaving as a reflector for a pair of transmitter receivers. We will focus on the phase shift and the main beam width in order to design a beam tracking system which is able to "follow" the mobile receiver node for improving the efficiency in terms of data rate and outage probability. Based on that, we will present a multifunctional reconfigurable meta-surface structure based on the radiation pattern modulation of the reflection coefficient. Firstly, we will design a reconfigurable U-shaped unit cell using a pin diode via a hole, which can provide 180°-phase difference between ON and OFF states. Specifically, a 10×10 meta-surface loaded with PIN diodes is designed for multifunctional behavior, such as coding and beam steering. The simulated results in both scenarios of unit cell and full structures will show the effectiveness of the proposed structure for vehicular communications.

Our main contributions can be summarized as follows:

- We design a specific meta-atom working at 5.3 GHz for automotive applications and based on this unit cell, we design a full structure and validate it to assess its suitability for the vehicular application considered;
- We design a full structure and validate it to assess its suitability for the vehicular application considered;
- We consider the integration of the designed Reconfigurable Intelligent Meta-surface (RIM) in a vehicular system and we numerically evaluate the performance; We compare the results of the system with and without RIM in terms of outage probability and capacity, by demonstrating the great potentiality of this type of structure.

The rest of the article is organized as follows. Section 2 describes the specific scenario considered and details the proposed unit cell structure. Section 3 presents simulated results for the proposed RIM unit cell in two ON and OFF states, expressed in terms of (*i*) reflection magnitude and phase, and (*ii*) effective permittivity and permeability. In Section 4 we validate the full RIM structure, while in Section 5 the performance of the proposed RIM has been exploited for vehicular communications, by considering that the RSU and the receiver vehicle are coated with the implemented 10×10 unit cells RIM. Instantaneous data rate, outage probability and channel capacity have been obtained as validation results. In Section 6, we present a few considerations related to the obtained results. Finally, conclusions are drawn at the end of the article.

2. UNIT-CELL DESIGN AND CONFIGURATION

The envisioned vehicular communications paradigm based on using RIM is shown in Fig. 1. In particular, we focus on a RIM relay based scheme. An RSU (Road Side Unit) is at the side of the road and is coated with the specific meta-surface structure we will detail later. We assume that the RSU, as represented in Fig. 1, is re-transmitting data to the receiver (*i.e.*, the green vehicle). In order to maximize the SNR to the destination, the RIM RSU will be "beam tracking" the receiver. This specific behavior can be realized by designing a unit cell with specific features as explained to the follow.

The proposed unit cell structure is shown in Fig. 2 (*a*). In this structure we proposed a U-shaped microstrip structure which is able to be reconfigurable by a PIN diode. By creating a rectangular slot in the corner of the rectangular patch a new path of surface current will be created. Hence, we can control the input impedance of the unit cell which is suitable for such an application of RIM. Moreover, the relevant equivalent circuit for this structure is shown in Fig. 2 (*b*). Regarding the ON and OFF states of the PIN diode, we put two different equivalent circuits for this section. All the dimensions are summarized in Table 1.

Table 1 – Parameters of the proposed unit cell

Parameter	Value [mm]	Parameter	Value [mm]
W_{Sub}	10	L_C	9
L_{Sub}	10	W_C	6
h_{Sub}	1.6	W_{C1}	2
L_d	1	L_{C1}	3
W_d	3	R_{via}	0.25

In meta-surface structures, beam steering can be consid-

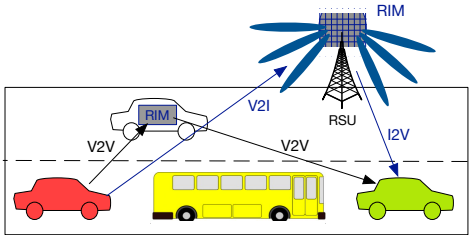


Fig. 1 – Vehicular communications paradigm based on the use of RIMs as relay node, for data transmission to a receiver node (*green vehicle*), in case of Vehicle-to-Vehicle (*black lines*) and Vehicle-to-Infrastructure/Infrastructure-to-Vehicle (*blue lines*).

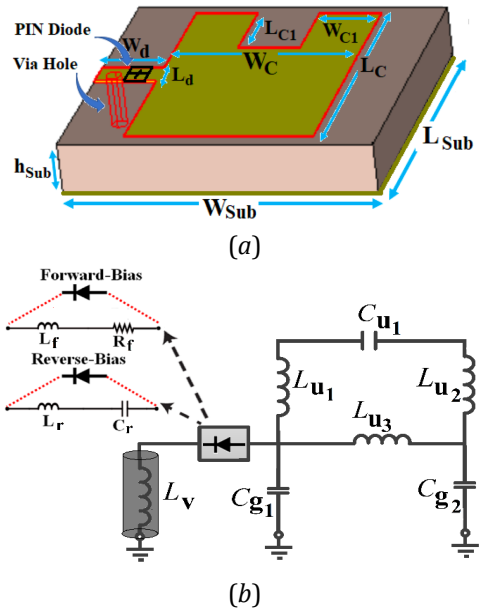


Fig. 2 – (*a*) The configuration and the geometry of the proposed unit cell with a U-shaped radiating patch, and (*b*) unit cells with PIN diode's equivalent circuit model for ON and OFF states.

ered as a particular case of wavefront manipulation that occurs in the far field. Regarding the Huygens principle, the meta-surface structures can be considered as an integrated EM radiator array [20]. Herein, in order to model the meta-antenna array, we consider a method that has been validated in several works via extensive simulations [21]. Considering each unit cell as an element of the planar array, the far field of the meta-surface can be obtained as:

$$F(\theta,\Phi) = f_E(\theta,\Phi)\cdot f_A(\theta,\Phi), \qquad (1)$$

by considering infinite sphere, θ is the elevation angle, Φ is the azimuth angle of an arbitrary direction in this coordination.

Regarding the planar array, the pattern function of each unit cell $f_E(\theta,\Phi)$ is the element factor and the pattern function of full planar configuration $f_A(\theta,\Phi)$ is the array factor. In far field region, we assume a planar wave covers the entire meta-surface. Therefore, the radiated pattern will depend only on the array factor. In this case, the far field pattern for the meta-surface with $N \times M$ unit cells, becomes

$$F(\theta,\Phi) = \sum_{m=1}^{M}\sum_{n=1}^{N} A_{mn}e^{j/\alpha_{mn}} f_{mn}(\theta_i,\Phi_i)\cdot \Gamma_{mn}e^{j/\Phi_{mn}} f_{mn}(\theta,\Phi)e^{jk_{0_{mn}}(\theta,\Phi)}, \qquad (2)$$

where A_{mn} and α_{mn} are the amplitude and phase of the wave incident to the (m,n)-th unit cell, respectively, with $m = [1,2,\dots,M]$ and $n = [1,2,\dots,N]$. In Eq. *(2)*, Γ_{mn} and Φ_{mn} are the amplitude and phase of the response of the (m,n)-th unit cell, respectively; $f_{mn}(\theta,\phi)$ denotes the scattering diagram of the (m,n)-th unit cell towards an arbitrary direction of reflection, whereas $f_{mn}(\theta_i,\Phi_i)$ denotes the response of the (m,n)-th unit cell at the direction of incidence determined by θ_i, Φ_i and $k_0 = 2\pi/\lambda_0$ is the wave number. Finally, we introduce $\zeta_{mn}(\theta,\Phi)$, which denotes the relative phase shift of the unit cells with respect to the radiation pattern coordinates, given by

$$\zeta_{mn}(\theta,\Phi) = D_u\, sin(\theta)\left[(m-\frac{1}{2})\,cos(\Phi) + (n-\frac{1}{2})\,sin(\Phi)\right], \qquad (3)$$

with D_u [m] as the unit cell size.

In order to make the model able to be calculated, we make a further assumption in the point of plane incident wave view, so that factors A_{mn}, α_{mn}, and $f_{mn}(\theta_i,\Phi_i)$ are constants for all m and n indexes. In addition, we apply the widespread assumption to the scattering pattern of the unit cell, which is modeled over the positive semisphere with the function $cos(\theta)$, which is a widespread assumption, [11]. Finally, and without loss of generality, we consider the normal incidence *i.e.*, $(\theta_i = \Phi_i = 0)$. Then, Eq. *(2)* becomes [11]

$$E(\theta,\Phi) = k\,cos(\theta)\sum_{m=1}^{M}\sum_{n=1}^{N}\Gamma_{mn}e^{j[\Phi_{mn}+k_0\zeta_{mn}(\theta,\Phi)]}, \qquad (4)$$

with k as a constant.

In order to have anomalous reflection, the main objective is controlling the phase shift of the unit cells Φ_{mn}. In particular, we manipulate the phase of the reflected waveform but not its amplitude. In this current version we do not focus on the control scheme for our system, since it is out of scope for this work. In reconfigurable meta-surface generating different coding sequence for unit cells, we are able to achieve desired functionalities such as beam steering and radiated wave modulation. In this regard, the amplitude Γ_{mn} and phase Φ_{mn} of the (m,n)-th unit cell need to be determined somehow which the entire response of the array matches with the required functionality. After this step, by mapping the required Γ and Φ to the closest available unit cell states, the desired functionality will be obtained. In the case of anomalous reflection for beam steering, analytical methods provide high accuracy.

In this study, in order to obtain beam steering functionality, the phase gradient approach is used to determine the direction of reflection [13]. Considering $\Phi(x,y)$ as the phase profile which is imposed by reconfigurable meta-surface, the virtual wave vector $\mathbf{K}_\Phi = \Phi_x\hat{\mathbf{x}} + \Phi_y\hat{\mathbf{y}}$ can be

assigned to the meta-surface. In this context, the momentum conservation law for wave vectors can be expressed as

$$k_i\, sin(\theta_i)\, cos(\Phi_i) + \frac{\partial \Phi_x}{\partial x} = k_r\, sin(\theta_r)\, cos(\Phi_r) \qquad (5)$$

$$k_i\, sin(\theta_i)\, cos(\Phi_i) + \frac{\partial \Phi_y}{\partial y} = k_r\, sin(\theta_r)\, sin(\Phi_r) \qquad (6)$$

where $\partial\Phi_x/\partial x$ and $\partial\Phi_y/\partial y$ describe the gradients along the $\hat{\mathbf{x}}$- and $\hat{\mathbf{y}}$- directions, respectively. For simplicity we consider the normal incident wave case *i.e.*, $(\theta_i = \Phi_i = 0)$ in lossless medium scenario [14]. Assuming air as the medium of the incident and reflected wave, we can simplify the formulations above as

$$\partial\Phi_x = \frac{2\pi\partial x\, cos\, \Phi_r\, sin\, \theta_r}{\lambda_0},\quad \partial\Phi_y = \frac{2\pi\partial y\, sin\, \Phi_r\, sin\, \theta_r}{\lambda_0}, \qquad (7)$$

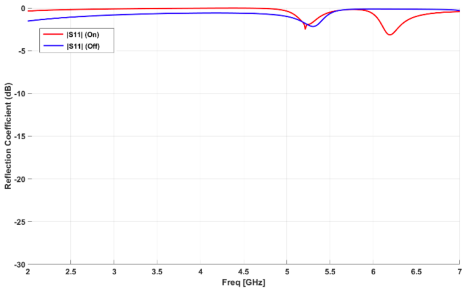
which demonstrate the phase shift Φ_x and Φ_y that need to be performed per unit of distance (*i.e.*, ∂x and ∂y) along the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ directions, respectively. Then, in Eq. *(6)* we set the unit cell size as $\partial x = \partial y = D_u$, in order to obtain the phase required at the (m, n)-th unit cell as

$$\Phi_{mn} = \frac{2\pi D_u(m\, cos\, \Phi_r\, sin\, \theta_r + n\, sin\, \Phi_r\, sin\, \theta_r)}{\lambda_0}. \qquad (8)$$

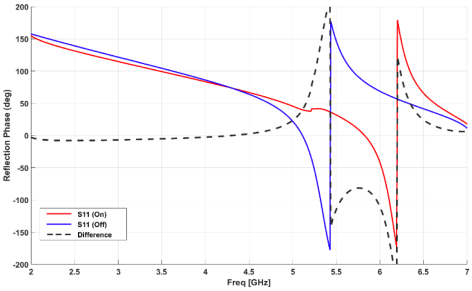
For beam-steering functionality the required phase Φ_{mn} is calculated for all the unit cells, to assign radiated states to each unit cell. Then, a closest neighbor mapping is done between the required phase and that provided by the different unit cell states.

3. SIMULATION RESULTS

In this section, we evaluate the performance of the unit cell we designed. The simulation is realized by the means of a commercial software CST studio suite. The specific configuration considered in CST is the boundary condition as unit cell in $\hat{\mathbf{x}}$- and $\hat{\mathbf{y}}$-directions with open-add space in $\hat{\mathbf{z}}$-direction. The reflectivity and reflection phases of unit cells are simulated using the frequency domain solver. The simulated reflection magnitude and phase of the unit cell are shown in Fig. 3. It is obvious that at 5.3 GHz, while the reflection magnitude is almost identical between the ON and OFF state, the reflection phase between the two cases has a 180° change. The maximum unit cell loss is around 2.5 dB for the OFF state. It is evident that the proposed 1-bit unit cell is suitable for the multifunctional meta-surface such as coding and beam steering. As the phase change between ON/OFF states is relative, we can simply state that a unit cell with an ON state corresponds to a 0° phase reflection, while one with an OFF state has a -180° phase reflection.



(*a*)



(*b*)

Fig. 3 – Simulated results for the proposed unit cell in two ON and OFF situations, expressed in terms of (*a*) reflection magnitude and (*b*) reflection phases.

In order to assess the double negative characteristics, S_{11} and S_{21} reflection and transmission coefficients are extracted from the design in CST in both magnitude and angle (expressed in rad). Then, the effective permittivity ε_{eff} and effective permeability μ_{eff} are obtained respectively as [20]:

$$\varepsilon_{eff} = \frac{\frac{1}{kd}\, cos^{-1}\left[\frac{1}{2S_{21}}(1 - S_{11}^2 + S_{21}^2)\right]}{\sqrt{\frac{(1+S_{11})^2 - S_{21}^2}{(1-S_{11})^2 - S_{21}^2}}}, \qquad (9)$$

and

$$\mu_{eff} = \frac{1}{kd}\, cos^{-1}\left[\frac{1}{2S_{21}}(1 - S_{11}^2 + S_{21}^2)\right]\sqrt{\frac{(1+S_{11})^2 - S_{21}^2}{(1-S_{11})^2 - S_{21}^2}}, \qquad (10)$$

where k denotes the wave number of the incident wave and d [mm] is the thickness of the unit cell. In this study, the meta-surface is printed on an FR-4 substrate with thickness of 1.6 mm and is designed to increase the directivity and bandwidth of the structure. Fig. 4 shows the effective permittivity and effective permeability related to the designed unit cell. It is clear that the proposed unit cell has double-negative material characteristics around 5.3 GHz for both ON and OFF states.

4. FULL STRUCTURE DESCRIPTION AND EVALUATION

In this section, in order to evaluate the performance of the proposed unit cell for multifunctional applications, a

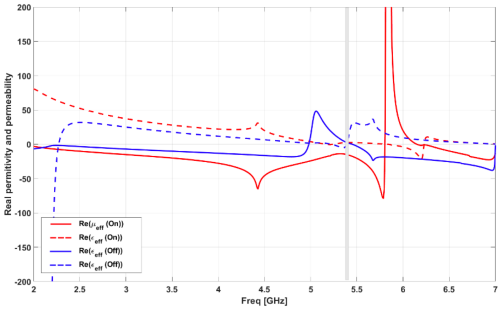


Fig. 4 – The effective permittivity and permeability of the unit cell.

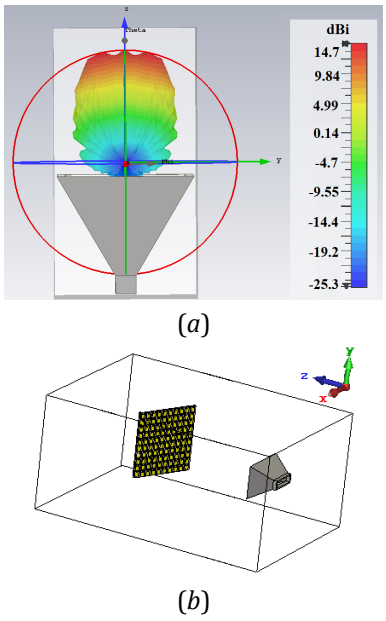


Fig. 5 – (*a*) Horn antenna as a transmitter and (*b*) the 1-bit 10×10 meta-surface simulated model.

Horn antenna is employed as a main radiator at the transmitting side, as shown in Fig. 5 (*a*). The redesigned Horn antenna is given to excite the meta-surface structure at 5.3 GHz. In addition, the 1-bit 10×10 meta-surface simulated model is shown in Fig. 5 (*b*).

4.1 Coding Meta-surface Construction

In this section we consider coding meta-surface based on the RIM structure, as the possibility to characterize the states ON and OFF as matching the bits 0 and 1. Fig. 6 (*a*) shows a random coding meta-surface with a fixed ratio and a different coding sequence, and Fig. 6 (*b*) shows the simulated 3D radiation in full structure using CST. As shown clearly, with coding sequence, the diffusion of the far-field pattern and the scattering amplitude at the normal incident angle are apparently the same. According to the code MATLAB, a 3D far-field pattern can be obtained with fixed ratio between 0 and 1 coding elements. The simulated results are demonstrated as Fig. 6 (*c*). Once ratio fixed, the scattering amplitude at the normal incident angle 0° is determined, and the efficient of coding meta-surface only depends on the uniformity of the scattering

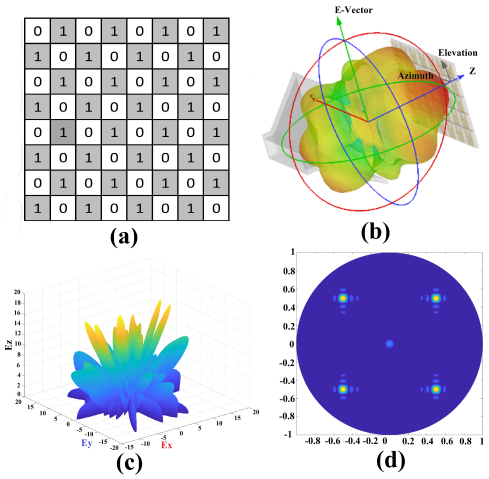


Fig. 6 – Reflection phase distribution and 3D pattern of the coding meta-surface with Horn antenna as an EM source located at (0 cm, 0 cm, 100 cm). (*a*) Ideal reflection phase distribution; (*b*) 3D pattern in full structure; (*c*) the beam pattern based on ideal phase distribution; (*d*) the beam pattern contour for coding application.

beam. Fig. 6 (*d*) demonstrates contour plot with corresponding 3D scattering pattern.

4.2 Beam-Steering Meta-surface Construction

In this section, an 8×8 beam steering meta-surface is modeled and simulated using CST Studio software to verify the beam steering capability of the RIM. In this simulation, the meta-surface is in the X-Y plane, and a horn serves as an EM source, which is located at (50 mm, 0 cm, 0 cm) with a rotation of (45°, 0°) with respect to the meta-surface. Then, an ON/OFF pattern matrix for steering to (120°, 0°) is loaded to the PIN diode of each unit cell. Finally, the simulation results are exported and shown in Fig. 7 (*a*) and (*b*). It is clearly observed that the coding meta-surface is capable of steering the beam to the desired direction with 20° angular resolution in the full structure case.

5. EVALUATION IN VEHICULAR APPLICATIONS

The proposed RIM structure is now exploited for vehicular communications, as depicted in Fig. 1. Specifically, we consider the communication link in a vehicular context, established from a source (*red vehicle*) and a receiver node (*green vehicle*). We assume a highway scenario and the transmitter and receiver vehicles have constant speed for the transmission time window. In order to characterize the impact of the controllable meta-surface in the communication system, we assume a simplified system with no interferences caused by others vehicles. In order to quantify the impact of the RIM on the performance system, we will consider (*i*) the relation between different beam width on the average rate, given a certain estimated velocity, and (*ii*) the derivation of the outage probability and the analysis of the communication system in terms of

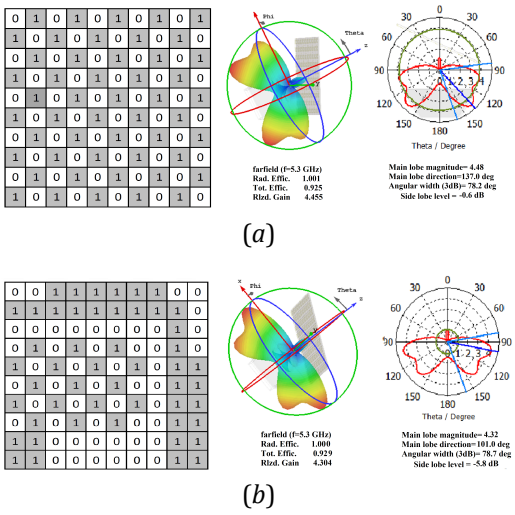


Fig. 7 – The simulation results of 3D gain total with the horn and metasurface and the 2D radiation pattern for 10×10 with 50 mm distance and 25° rotation at Y-plane at 5.25 GHz with different phase distribution, (*a*) uniform distribution and (*b*) random distribution.

both outage and capacity, in cases with and without RIM.

5.1 Beam width Impact Evaluation

A 10×10 unit cells RIM is coating both the RSU (acting as relay node) and the receiver vehicle. In particular, we compute the instantaneous data rate by assuming a perfect alignment between the beam from the RSU and the receiver node. The instantaneous data rate is derived by the Shannon capacity formula for the instantaneous rate as:

$$R(t, \theta_b) = B\, log_2(1 + SNR(t, \theta_b)), \qquad (11)$$

where B [Hz] is the bandwidth of the system, SNR [dB] is the Signal-to-Noise Ratio, t [s] is the time instant, and θ_b is the beam width computed in degrees.

We assume that our system is able to "instantaneously" switch from the $(i-1)$-th beam to the i-th one, and the "transmitter" beam and the "receiver" one are perfectly aligned. Moreover, we consider a beam model where the gain inside the beam is uniform and zero outside. In Eq. *(11)*, the expression of the SNR can be written as

$$SNR(t, \theta_b) = \frac{P_{rx}(t, \theta_b)}{P_{noise}}, \qquad (12)$$

where P_{noise} [dB] represents the thermal noise including a Noise Figure *i.e.*, NF [dB], which can be expressed as

$$P_{noise} = -174 + 10\, log_{10}\, B + NF. \qquad (13)$$

In Eq. *(12)*, the term P_{rx} is the received power, expressed in linear scale as [10]:

$$P_{rx}(t, \theta_b) = \mathcal{K} \left(\frac{\pi^2}{\theta_b^2} \right)^2 \frac{1}{\left[(vt)^2 + r^2 \right]^{n/2}}, \qquad (14)$$

where v [m/s] is the vehicle speed, n is the path loss exponent, and r [m] is the distance from the RSU and the receiver node. The term $\mathcal{K}$ is defined as

$$10\, log_{10}(\mathcal{K}) = EIRP_{dBm} - E + 10n\, log_{10}\, (\lambda/4), \qquad (15)$$

where $EIRP$ is the Equivalent Isotropic Radiated Power and λ [m] is the wavelength and E represents the shadowing margin. In order to compute $\mathcal{K}$, we consider the power associated with the reflected signal S_{21} as computed in Eqs. *(9)* and *(10)*.

Table 2 – Simulation Parameters

Parameter	Value
$EIRP$	57 dBm
B	75 MHz
f_c	5.3 GHz
r	$[0, 200]$ m
NF	6 dB
v	$[25, 50, 75]$ km/h

Notice that we assume that the transmitter node is transmitting at an EIRP that for Europe is equal to 57 dBm. The relay will not be able to retransmit at full power since a percentage of the power associated to the impinging wave will be dissipated as a reflected signal. In particular, the design of the unit cell has been optimized in order to minimize this power loss *i.e.*, a 2% of the total transmitted power will be wasted as a reflected component of the wave. Furthermore, without loss of generality, the speed of the receiver vehicle is considered constant during the transmission time, since the interval time of reception is short with respect to the variation of the vehicle speed. Table 2 collects the parameters used in the numerical results.

In Fig. 8 (*a*) we show the instantaneous and ideal rate, *i.e.*, the maximum achievable rate with the assumption of a perfect alignment and without error estimation of the velocity at the receiver by considering a narrow beam width, *i.e.*, $\theta_b = [0.1, 10]^{\circ}$. Data rate values range from $\approx [15, 25]$ Gbps, where higher values are obtained for smaller beam widths, while a data rate decrease is observed for larger beam widths.

Furthermore, we have considered the impact of different speeds on the achievable rate and we can observe that, as expected, the data rate decreases for increasing vehicle speed. Finally, the effect of the distance on the data rate is a decrease of performance when the distance increases.

In Fig. 8 (*b*) we show the impact of a wider beam width for the same scenario. Of course, wider beam widths are related with a smaller number of unit cells. Indeed, the smaller is the number of unit cells, the lower is the fine control we can realize on the RIM. As expected, we observe a decreasing of the instantaneous data rate from a maximum of 23 Gbps in Fig. 8 (*b*) to a maximum of 15 Gbps in Fig. 8 (*a*).

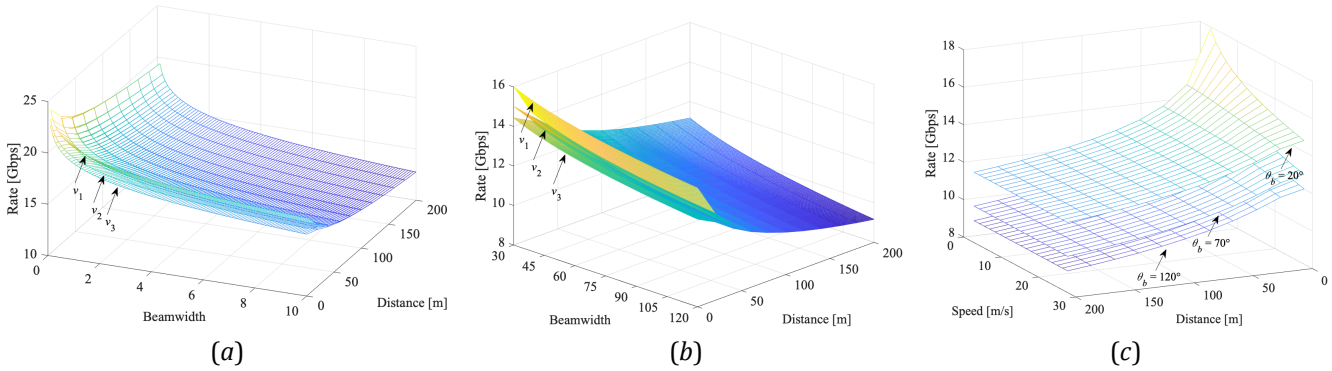


Fig. 8 – Evaluation of the instantaneous data rate [Gbps] vs. the beam width and the vehicle speed, in case of (*a*) $\theta_b = [0.1, 10]^\circ$ and (*b*) $\theta_b = [30, 120]^\circ$, both by considering different speeds *i.e.*, $v_1 = 25$, $v_2 = 50$, and $v_3 = 75$ [m/s]. (*c*) Instantaneous data rate [Gbps] vs. vehicle speed and distances, in case of $\theta_b = [20, 70, 120]^\circ$.

Finally, in Fig. 8 (*c*) we analyze the impact of the speed *i.e.*, from 10 to 100 km/h, with respect to specific values of beam width *i.e.*, $\theta_b = [20, 70, 120]^\circ$. The behavior follows the same trend as observed in Fig. 8 (*a*) and (*b*), where, as expected, for higher beam width the data rate performance is strongly reduced. Also, lower distances provide higher performance.

5.2 Outage Probability

The outage probability is a crucial metric for vehicular communication [24, 25], and can be defined as the instantaneous mutual information rate that falls below a certain threshold β, *i.e.*:

$$\underset{outage}{Pr}[SNR < \beta] = 1 - \underset{success}{Pr}, \qquad (16)$$

where $Pr_{success}$ is the success probability that SNR is higher than β, and can be expressed as

$$\underset{success}{Pr} = Pr[SNR > \beta] = exp\left(-\beta d^n \frac{P_{noise}}{P_{tx}}\right), \qquad (17)$$

with P_{tx} as the transmitting power.
Based on the outage probability, given a specific target rate as a Quality of Service (QoS) parameter, we can derive the throughput of success delivery with constrained outage probability ϵ, as [24]:

$$C = (1 - \epsilon) B log_2(1 + SNR). \qquad (18)$$

In order to evaluate the impact of the presence of the meta-surface, in Fig. 9, we consider a target rate of 1 Gbps and derive the corresponding outage probability as computed in Eq. *(16)*. In order to better appreciate the effect of the meta-surface, we consider two different beam widths of 20° and 90°. As expected, the lower is the beam width, the lower is the outage probability and the system has better performance in terms of capacity (see Fig. 10). The better performance is related to a higher precision of the system, that corresponds to an increased number of

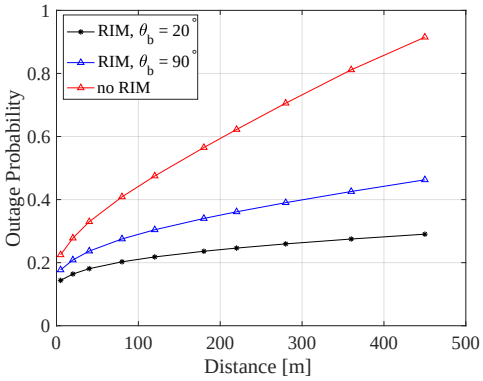


Fig. 9 – Outage Probability vs. the distance, in case of (*i*) a meta-surface with beam width $\theta_b = 20^\circ$ and $\theta_b = 90^\circ$, and (*ii*) without a meta-surface.

meta-atoms needed in the metastructure for a high tuning of the phase. In practice, the higher is the number of unit cells of the full structure, the better is the control of the whole system, but also the control logic to drive the controllable meta-surface is more sophisticated/complicated, since it is demanding a higher precision and could make the system more vulnerable to misalignment errors.

6. DISCUSSION

In order to assess the performance of the system and to evaluate the potential impact of the RIM in a vehicular context, we have considered some assumptions to simplify the analysis and give some useful insights for future work. Firstly, a perfect alignment between the relay coated with the RIM and the receiver has been considered. This assumption cannot be ensured above all in a context characterised with high speed. In order to account for that, we need to introduce an error factor of the speed to account for the misalignment. A possible approach for accounting the misalignment has been proposed in [26]. We believe that this misalignment could be translated as a reduction of the received power and then in a decreasing of the SNR.

Another simplification introduced is the ideal configura-

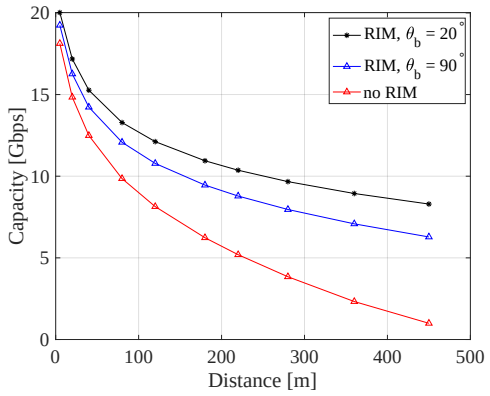


Fig. 10 – Capacity of the system with a rate set to 1 Gbps, in case of (*i*) a meta-surface with beam width $\theta_b = 20°$ and $\theta_b = 90°$, and (*ii*) without a meta-surface.

tion that is computed in real time and without error. Indeed, in order to control the RIM and make it as a tunable and intelligent meta-surface, we need to include an intelligent logic, potentially based on a machine learning (ML) approach to generate the correct configuration of the RIM when an external change impacts on the system (*e.g.*, the vehicle moves far away from the RSU and a new beam needs to be calculated). ML approaches for systems based on RIM is a topic that is gaining more and more interest in recent times. Anyway, since we considered a mobile scenario where some crucial parameters can be easily predicted (*e.g.*, the velocity of the vehicle and the direction), we trust that the computation of a new configuration can be calculated in real time and this should not impact too much on the ideal performance of the system.

7. CONCLUSIONS

In this paper, we presented a multifunctional reconfigurable meta-surface based on the radiation pattern modulation of the reflection coefficient. We designed a reconfigurable U-shaped unit cell using a pin diode via a hole, that allows to obtain a $180°$-phase difference between ON and OFF states. The proposed RIM structure has been designed for multifunctional behavior such as coding and beam steering, considering a 10×10 meta-surface loaded with PIN diodes.

Simulation results have been carried out in order to validate the proposed RIM in a vehicular scenario, for data transmission via the V2X mode. The instantaneous data rate has been obtained, as a function of the beam width, the vehicle speed and the distance. As expected, the lower the beam width is, the higher the data rate is; additionally, the lower the vehicle speed is, the higher the data rate is. Same consideration occurs for the distance. In order to assess the effectiveness of the RIM integration in the vehicular communication system, we have derived the outage probability for a specific target QoS in terms of an instantaneous rate. We also have computed the capacity based on the outage probability and we have compared both the outage and the capacity for a system equipped with RIM with two different configurations in terms of beam width, and without RIM.

As a future piece of work, a factor that we have not considered in this context is the presence of other sources and interferences. We trust the integration of the RIM in the system can improve the reduction of the interference impact by properly addressing it. In order to account for this point, we plan to extend the analytical model with the interference factor and design RIM-based approaches that reduce the impact of the interference.

ACKNOWLEDGMENT

This work is partially supported by the Exploratory Action ETHICAM.

REFERENCES

[1] A. Nemati, Q. Wang, M. Hong, and J. Teng, "Tunable and reconfigurable meta-surfaces and metadevices," in Opto-Electronic Advances, 01, 180009, 2018.

[2] C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. F. Akyildiz, "On the Network-layer Modeling and Configuration of Programmable Wireless Environments," in IEEE/ACM Transactions on Networking, vol. 27, no. 4, pp. 1696-1713, Aug. 2019.

[3] C. Liaskos, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. F. Akyildiz, "Using any Surface to Realize a New Paradigm for Wireless Communications," in Communications of the ACM, vol. 61 no. 11, pp. 30-33, October 2018.

[4] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. F. Akyildiz, "A New Wireless Communication Paradigm through Software-controlled Meta-surfaces," in IEEE Communications Magazine, vol. 56, no. 9, pp. 162-169, September 2018.

[5] L. La Spada, V. Loscrí, and A.M. Vegni, "Meta-Surface Structure Design and Channel Modelling for THz Band Communications," in Proc. of INFOCOM 2019 Workshop, 1st IEEE Workshop on Ultra-High Broadband Terahertz Communication for 5G and Beyond Networks (UBTCN 2019), April 29-May 2, Paris, France.

[6] A. Taibi, A. Durant, V. Loscrí, A.M. Vegni, and L. La Spada, "Controlling Light by Curvilinear Meta-Surfaces," in Proc. of ACM Nanocom 2019, September 25-27, 2019, Dublin, Ireland.

[7] J. Dickmann *et al.*, "Automotive radar the key technology for autonomous driving: From detection and ranging to environmental understanding," in Proc. of 2016 IEEE Radar Conference (RadarConf), Philadelphia, PA, 2016, pp. 1-6.

[8] O. Schumann, J. Lombacher, M. Hahn, C. Wöhler and J. Dickmann, "Scene Understanding With Automotive Radar," in IEEE Transactions on Intelligent Vehicles, vol. 5, no. 2, pp. 188-203, June 2020, doi: 10.1109/TIV.2019.2955853.

[9] J. Wang, Y. Shao, Y. Ge, and R. Yu, "A Survey of Vehicle to Everything (V2X) Testing," in Sensors (Basel). 2019;19(2):334. Published 2019 Jan 15.

[10] J. Kim and A. F. Molisch, "Enabling Gigabit services for IEEE 802.11ad-capable high-speed train networks," in IEEE Radio and Wireless Symposium (RWS), pp. 145-147, Jan. 2013.

[11] H. Yang, X. Cao, F. Yang, J. Gao, S. Xu, M. Li, X. Chen, Y. Zhao, Y. Zheng, and S. Li, "A programmable metasurface with dynamic polarization, scattering and focusing control," in Scientific Reports, vol. 6, no. 35692, 2016.

[12] A. Khan, F. Ullah, Z. Kaleem, S. ur Rahman and Y. Z. Cho, "Sensor-based self-organized traffic control at intersections," in Proc. of 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, 2017, pp. 634-638, doi: 10.1109/ICTC.2017.8191056.

[13] N. Yu, P. Genevet, M. a. Kats, F. Aieta, J.-P. Tetienne, F. Capasso, and Z. Gaburro, "Light Propagation with Phase Discontinuities: Generalized Laws of Reflection and Refraction," in Science, vol. 334, no. October, pp. 333–337, 2011.

[14] S. Liu, T. Jun Cui, A. Noor, Z. Tao, H. Chi Zhang, G. Dong Bai, Y. Yang, and X. Yang Zhou, "Negative reflection and negative surface wave conversion from obliquely incident electromagnetic waves," in Light: Science and Applications, vol. 7, no. 5, pp. 18 008–18 011, 2018.

[15] T. Hamidreza, A. Cabellos-Aparicio, J. Georgiou, and S. Abadal, "Error Analysis of Programmable Metasurfaces for Beam Steering," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems 10, no. 1 (2020): 62–74.

[16] H. Taghvaee et al., "Scalability Analysis of Programmable Meta-surfaces for Beam Steering," in IEEE Access, vol. 8, pp. 105320-105334, 2020, doi: 10.1109/ACCESS.2020.3000424.

[17] C. Rizza, V. Loscrí, and M. Parchin, "A Millimeter-Wave Reconfigurable Intelligent Meta-surface Design for Vehicular Networks Applications," in Proc. of 2020 IEEE Vehicular Technology Conference, Oct 2020, Victoria, Canada.

[18] L. Zhang, X. Qing Chen, S. Liu, Q. Zhang, J. Zhao, J. Yan Dai, G. Dong Bai, X. Wan, Q. Cheng, G. Castaldi, V. Galdi and T. Jun Cui, "Space-time-coding digital meta-surfaces," in Nature Communication, 9, 4334 (2018).

[19] L. Zhang, R. Y. Wu, G. D. Bai, H. T. Wu, Q. Ma, X. Q. Chen, T. J. Cui, "Transmission, Reflection, Integrated Multifunctional Coding Meta-surface for Full," in Space Controls of Electromagnetic Waves, in Adv. Funct. Mater. 2018, 28, 1802205.

[20] C. A. Balanis, Antenna Theory: Analysis and Design, 3rd ed., Wiley, Ed., 2005.

[21] E. Hosseininejad, K. Rouhi, M. Neshat, A. Cabellos-Aparicio, S. Abadal, and E. Alarcon, "Digital Metasurface Based on Graphene: An Application to Beam Steering in Terahertz Plasmonic Antennas," in IEEE Transactions on Nanotechnology, vol. 18, no. 1, pp. 734–746, 2019.

[22] A. Makarf, K. M. Rabie, O. Kaiwartya, K. Adhikari, X. Li, M. Quiroz-Castellanos, and R. Kharel, "Reconfigurable intelligent surfaces-enabled vehicular networks: A physical layer security perspective," arXiv:2004.11288, 2020.

[23] B. Masini, M., Silva, M. Cristiano, and A. Balador, "The Use of Meta-Surfaces in Vehicular Networks," in Journal of Sensor and Actuator Networks, 2020, vol. 9, no 1, p. 15.

[24] X. Wu, S. Sun, Y. Li, Z. Tan, W. Huang, and X. Yao, "A Power Control Algorithm Based on Outage Probability Awareness in Vehicular Ad Hoc Networks," in Advances in Multimedia, vol. 2018, Article ID 8729645, 8 pages, 2018. https://doi.org/10.1155/2018/8729645

[25] B. E. Y. Belmekki, A. Hamza and B. Escrig, "On the Outage Probability of Vehicular Communications at Intersections Over Nakagami-m Fading Channels," in Proc. of 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 2020, pp. 1-5, doi: 10.1109/VTC2020-Spring48590.2020.9128618.

[26] V. Va, X. Zhang and R. W. Heath, "Beam Switching for Millimeter Wave Communication to Support High Speed Trains," in Proc. of 2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall), Boston, MA, 2015, pp. 1-5, doi: 10.1109/VTC-Fall.2015.7390855.

AUTHORS

Mohammed Ojaroudi received a PhD degree at Shahid Beheshti University, Tehran, Iran, in 2014. He worked, as a visiting researcher at University of Tennessee, Knoxville, USA from 2013 to 2014. From Oct. 2014, he was a post-doctoral researcher at Middle East technical University (METU), Ankara, Turkey and from 2016 to 2017, he was an academic lecturer in Ankara University, Ankara, Turkey. From 2017-2020, he worked as a scientist researcher at XLIM research laboratory Limoges, France. From 2020 he has been working on a project in the frame of the reconfigurable intelligent meta-surface to develop short-range sensing for THz communication and imaging applications at Inria, Lille, France. From 2012, he has been a member of the IEEE-APS, IEEE-MTT, IEEE-AWPL, MAP-IET, reviewers' group. He is an author and coauthor of more than 150 journals and international conferences papers. His research interests include analysis and design of microstrip antennas, design and modeling of microwave structures, microwave imaging systems, and emerging electromagnetic technologies for THz communications.

Valeria Loscrí is a permanent researcher of the FUN Team at Inria Lille-Nord Europe since Oct. 2013. From Dec. 2006 to Sept. 2013, she was Research Fellow in the TITAN Lab of the University of Calabria, Italy. She received her MSc and PhD degrees in Computer Science in 2003 and 2007, respectively, from the University of Calabria and her HDR (Habilitation a diriger des recherches) in 2018 from Université de Lille (France). Her research interests focus on emerging technologies for new communication paradigms such as Visible Light Communication and TeraHertz bandwidth and cooperation and coexistence of wireless heterogeneous devices. She has been involved in the activity of several European Projects (H2020 CyberSANE, FP7 EU project VITAL, the FP6 EU project MASCOT, etc.), Italian and French projects. She is on the editorial board of IEEE COMST, Elsevier ComNet, JNCA, IEEE Trans. on Nanobioscience. Since 2019, she is Scientific International Delegate for Inria Lille-Nord Europe.

Anna Maria Vegni (Senior member, IEEE) is a tenure-track Assistant Professor in the Department of Engineering at Roma Tre University (Italy), since March 2020. She received the Ph.D. degree in Biomedical Engineering, Electromagnetics and Telecommunications from the Department of Applied Electronics, Roma Tre University, in March 2010. She received the 1st and 2nd level Laurea Degree cum laude in Electronics Engineering at Roma Tre University, in July 2004, and 2006, respectively. In 2009, she was a visiting researcher in the Multimedia Communication Laboratory, directed by Prof. Thomas D.C. Little, at the Department of Electrical and Computer Engineering, Boston University, Boston, MA. Her research activity focused on vehicular networking supported by heterogeneous wireless networks and optical wireless communications. She is a member of ACM and an IEEE Senior Member. In March 2018, she got the Italian Habilitation (Abilitazione Scientifica Nazionale) for Associate Professorship in Telecommunication Engineering. She is involved in the organization of several IEEE and ACM international conferences and is a member of the editorial board of IEEE Communications Magazine, Ad Hoc Networks, Journal of Networks and Computer Applications, Nanocomnet Elsevier journals, IEEE JCN, ITU J-FET and ETT Wiley journal.

A BLUEPRINT FOR EFFECTIVE PANDEMIC MITIGATION

Rahul Singh[1], Wenbo Ren[2], Fang Liu[3], Dong Xuan[4], Zhiqiang Lin[5], Ness B. Shroff[6]
[1]Department of ECE, Indian Institute of Science, Bangalore, [2]CSE Department, The Ohio State University, [3]ECE Department, The Ohio State University, [4]CSE Department, The Ohio State University, [5]CSE Department, The Ohio State University, [6]ECE and CSE Departments, The Ohio State University,

NOTE: Corresponding author: Ness B. Shroff, shroff.11@osu.edu

Abstract – Traditional methods for mitigating pandemics employ a dual strategy of contact tracing plus testing combined with quarantining and isolation. The contact tracing aspect is usually done via manual (human) contact tracers, which are labor-intensive and expensive. In many large-scale pandemics (e.g., COVID-19), testing capacity is resource limited, and current myopic testing strategies are resource wasteful. To address these challenges, in this work, we provide a blueprint on how to contain the spread of a pandemic by leveraging wireless technologies and advances in sequential learning for efficiently using testing resources in order to mitigate the spread of a large-scale pandemic.

We study how different wireless technologies could be leveraged to improve contact tracing and reduce the probabilities of detection and false alarms. The idea is to integrate different streams of data in order to create a *susceptibility* graph whose nodes correspond to an individual and whose links correspond to spreading probabilities. We then show how to develop efficient sequential learning based algorithms in order to minimize the spread of the virus infection. In particular, we show that current contact tracing plus testing strategies that are aimed at identifying (and testing) individuals with the highest probability of infection are inefficient. Rather, we argue that in a resource constrained testing environment, it is instead better to test those individuals whose expected impact on virus spread is the highest. We rigorously formulate the resource constrained testing problem as a sequential learning problem and provide efficient algorithms to solve it. We also provide numerical results that show the efficacy of our testing strategy.

Keywords – contact tracing, COVID-19, selective testing

1. INTRODUCTION

The outbreak of COVID-19 has unfolded an unprecedented worldwide health, economical, and social crisis. Today, COVID-19 has spread to 188 countries, infected nearly 30 million people globally, and resulted in close to one million deaths. The International Monetary Fund (IMF) has predicted that the global economy will shrink by 3% this year, the worst decline since the Great Depression of the 1930s [2]. Today, millions of workers have been laid off, and the tourism or hospitality industry has been hurt particularly hard.

The COVID-19 outbreak and the mixed successes that nations have had in controlling the virus has underscored the need for the development of technological tools for pandemic mitigation. This paper provides a blueprint on how technologies should be used in conjunction with smart testing techniques in order to contain a pandemic such as COVID-19. The first step for pandemic mitigation is to identify or trace the close contacts who might have been exposed to the disease from a contagious individual. Contact tracing, an old technique, has been used as effective tools to battle pandemics for many years, and some countries do use aggressive contact tracing to successfully contain COVID-19. However, traditionally, contact tracing is a manual approach, relying on a human being's memory. Such an approach cannot scale to large and rapidly moved populations today. Meanwhile, manual tracing may result in delays, which could limit its utility. Therefore, recently numerous digital contact tracing systems have been developed and deployed across the globe, by using a wide variety of sources to track "encounters" including CCTV footage, records of credit card transactions [1], locations measured using cellular networks or WiFi hotspots [26], locations via GPS, and cryptographic tokens exchanged via Bluetooth Low Energy (BLE) or acoustic channels [15]. For a recent survey of works on contact tracing and privacy-aware contact tracing, see [9, 11, 10, 23, 24, 28, 7, 19]. Also see [14, 25] for some interesting recent works in this area.

However, currently contact tracing system appears to be uncoordinated individual efforts focused on individual technologies and one type of data stream (e.g., camera based systems, phone apps, etc.). Further, a number of these solutions suffer from a variety of technological limitations including a lack of coverage, privacy and security concerns, high missed detection and/or false alarm rates. For instance, increasingly Bluetooth-based contact tracing has gained mainstream use particularly with Apple/Google's support. However, our recent analysis [29] with the released COVID-19 contact tracing apps

shows that most Bluetooth-based contact tracing apps use just the received signal strength indicator (*RSSI*) of the Bluetooth for distance measurements. Unfortunately, in practice, numerous factors can affect the *RSSI* that can make the distance measurement inaccurate, such as the power of the antenna used for broadcasting (i.e., the *TxPower*) and the obstacles blocking transmission paths. Moreover, Bluetooth-based proximity tracing can also raise false positives because of the potential misinterpretation of various scenarios. For example, a proximity tracing system may interpret two users have contact even if they are separated by a solid wall, where the risk of infection is much lower than the risk indicated by the measured distance.

Therefore, we would like to propose an improved approach by combining as many data sources as possible in an integrated way, with the key objective of minimizing false positives and false negatives in the contact tracing and meanwhile protecting user's privacy. The contact tracing data we can collect includes (1) multiple channels including both Bluetooth and ultrasound (using both microphones and speakers available in the smartphone), and multiple sources including (2) WiFi and (3) cellular networks if they are available. We show how we can use improved methodology to collect data that is privacy aware, transparent, and integrated in Section 2.

Similarly, while testing followed by quarantining/isolation is a powerful tool against a pandemic such as COVID-19, testing capacity remains an issue, especially in hard hit areas where testing results could take multiple days, even up to a week, to arrive. While traditional approaches have focused on testing individuals who exhibit symptoms or have come in contact with other infected individuals, these approaches miss out many potential areas of outbreak where asymptomatic or presymptomatic super-spreaders seed the virus, which gets detected only after it has already spread significantly. Thus, testing capacity needs to be used judiciously to prevent widespread outbreaks. In Section 4, we argue that the current myopic approach to testing focuses on identifying individuals with the highest probability of being infected, which does not help minimize the overall number of infected individuals.

Organization. The rest of this paper is organized as follows: In Section 2, we describe an improved and integrated methodology to collect contact tracing data. In Section 3, we describe techniques that enable us to efficiently integrate the data collected from various streams. This allows us to reduce the "error probabilities" associated with false alarms or missed detection of diseases, and generate a dynamic *susceptibility graph*. In Section 4, we address the practical problem of testing under constraints on resources. We perform simulations to show necessities of contact tracing and building a contact graph in Section 5, and conclude in Section 6.

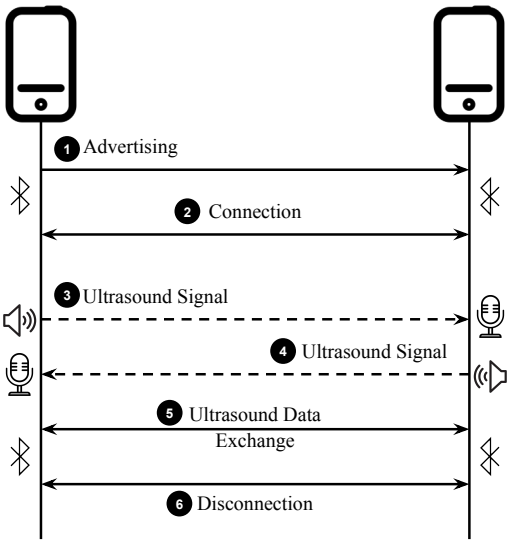


Fig. 1 – A Simplified Protocol for Improved Data Exchange.

2. IMPROVED DATA COLLECTION

There are two fundamental objectives that a good digital contact tracing system must satisfy: (1) it should be effective in tracking an individual (e.g., few false positives and missed detections), and (2) it should protect the privacy of users. For example, a CCTV footage would be highly effective if (1) were the sole objective, however it does not meet (2) since it is too much privacy-invasive. Therefore, we must look for effective and privacy-aware digital contact tracing techniques.

Since the outbreak of Covid-19, numerous techniques based on Bluetooth, WiFi, and cellular networks have been developed for contact tracing. Each technique has its own pros and cons. For instance, Bluetooth based solutions can achieve reliable communication and a low energy operation, but these suffer from a high rate of false positives due to a long communication range. WiFi based solutions do not require installation of apps on mobile phones, and they rely heavily on access point (AP) deployment, and its coverage. Therefore, in this paper, we aim to present an integrated approach that improves the accuracy of a Bluetooth-based approach with additional channels, and combines WiFi and cellular information if they are available. Furthermore by utilizing clever algorithms that are provably optimal, it aims to increase the efficiency with which infected individuals are contained early without infecting too many of their "neighbors".

Improving Bluetooth-based Contact Tracing with Ultrasound Signals. Since Bluetooth-signals can penetrate obstacles such as solid walls, and also have a long transmission range, we would like to leverage other sensors in a smartphone to improve its proximity accuracy. In particular, we can use the inaudible ultrasound generated from the speaker and recorded by the

microphone for distance measurement. This feature can be used very easily. Ultrasound-based distance measurement is promising, and it has been shown to achieve a centimeter-level accuracy, e.g. BeepBeep [16]. It can help to improve the performance of mobile contact tracing systems. According to the CDC guidelines [6], for the purpose of Covid transmission, it can be assumed that two individuals are in close contact (and hence likely to spread infection) if the distance between them is within 6 feet. Clearly, with an accuracy of a few centimeters, ultrasound is a very reliable and accurate technique for the purpose of mobile contact tracing.

Note that ultrasound cannot penetrate solid walls, and hence it can also help us to rule out those "false contacts" that have been declared by the Bluetooth connections occurring across walls.Using ultrasound for distance measurement requires a pair of devices to exchange data between each other [16]. Fortunately, Bluetooth-based mobile contact tracing systems have already implemented such an information exchange mechanism, and hence developers can easily extend these systems in order to support ultrasound distance measurement. This means that we can conveniently deploy both ultrasound and Bluetooth technologies.

An illustration of how we can integrate ultrasound and Bluetooth for improved contact tracing is presented in Fig. 1. The scheme is composed of six steps for the improved data exchange. In particular, when two smartphones encounter, they first use Bluetooth to discover each other and initiate connections. At the beginning, in Step ❶, one smartphone keeps advertising BLE packets to nearby devices. Meanwhile, another smartphone scans its environment for any possible advertising packets from its nearby devices, and initiates the connection in Step ❷. After a connection has been established, two devices emit ultrasound signals in turns (Step ❸ and ❹). When both devices have received signals from the other device, they record the information that is required by algorithms in order to measure distances, and then exchange this information amongst each other via Bluetooth in Step ❺. For example, when using the algorithm from BeepBeep [16] for distance measurement, in Step ❸, alongside emitting an ultrasound signal, the phone (P_A) will record the timestamp T_{A1} when it senses the signal sent by itself, and the other phone (P_B) will also store the timestamp T_{B1} when it receives such a signal. Similarly, in Step ❹, phone P_B records the timestamp for signal emitting as T_{B2} and phone P_A stores the timestamp T_{A2} when receiving signal. Next, in Step ❺, phone P_A has to send both T_{A1} and T_{A2} to phone P_B, and phone P_B also needs to share its two timestamps (i.e., T_{B1} and T_{B2}) with P_A.

$$Distance(P_A, P_B) = \frac{C}{2} \times ((T_{A2} - T_{A1}) - (T_{B2} - T_{B1})) + D_{B,B} + D_{A,A} \quad (1)$$

When each phone has these four timestamps, it can calculate the distance by using Equation *(1)*, where C is the sound speed, and $D_{A,A}$ or $D_{B,B}$ is the distance from one phone's speaker to its microphone respectively[1]. Finally, in Step ❻, two devices disconnect Bluetooth when the transmissions of other types of data for contact tracing finish.

Integrated Contact Tracing with WiFi.
Besides the above mobile phone based contact tracing approaches, we can also use WiFi logs. Different from the above approaches, the WiFi based solution does not require app installation on mobile phones, and relies upon widely deployed WiFi access points. It also does not require active involvement of mobile phones for exchanging the required information. The basic principle is described as follows: WiFi networks log all the associations and disconnections of devices connected to access points. We can analyze these WiFi logs to know where and when the devices (and hence their users) are close to each other. This provides information about the contacts of device users.

One of the advantages of such a WiFi-based solution is that the WiFi log data is always available as long as WiFi networks are active. Such networks allow both a) reactive and b) proactive techniques for contact tracing. Let's take the example of a university campus in order to illustrate this. In reactive contact tracing, once a student is confirmed to be infected, the university health administration can use the WiFi MAC addresses of the student's mobile phone and his/her other computing devices such as laptop and Apple watch etc. to search in the WiFi logs of campus networks, thereby determining the locations where the student visited during a certain time period, and also his/her contacts at these locations. In proactive contact tracing, the university health administration can proactively analyze WiFi logs to identify potential high risk users such as super spreaders, and hot-spots (such as big gathering) in the campus. The university health administration can proactively pull WiFi logs and determine if the number of students in a gathering exceeds the limit that social distancing allows, and take appropriate measures.

The WiFi based contact tracing technique described above has its own limitations, for example the AP association logs can generate false alarms. In order to overcome these, we might consider using Received Signal Strength Indicator (RSSI) and Channel State Information (CSI) to reduce the errors. WiFi based solution cannot be applied in areas without WiFi connectivity. We should consider enhancing contact tracing by integrating multiple solutions such as Bluetooth, WiFi and acoustic symbiotically, where one helps or replaces the other depending on user preferences, environmental dynamics, and resource availability. For example, in a WiFi-AP dense area such as a campus academic building, the WiFi-based solution can play a dominant role, while the mobile app running Bluetooth and acoustic

[1]For more details, please refer to BeepBeep [16]

based contact tracing can be automatically switched off for better energy savings. Similarly, if the density of WiFi AP is less, then Bluetooth and acoustic based solutions could be activated. They can help each other in a crowded and occluded (e.g., by walls and other types of obstacles) environment for accurate and reliable contact tracing. Note that Bluetooth gives relative location information whereas WiFi gives absolute location information. There might be a mismatch in the locations as identified by WiFi data modality in conjunction with Bluetooth data. In order to resolve this issue, we can model the location of the user using probabilistic techniques, and then use filtering techniques in order to derive a more accurate location. Readers interested in more details about these techniques can refer to Kalman filtering and related topics in [17, 13].

Integrated Contact Tracing with Cellular Network. The location of a smartphone can also be identified from its communication with nearby cell towers. Since a phone has to connect with cell towers in order to send and receive data through cellular network, it constantly searches nearby cell towers and initiates connections during movements. In each established connection, a cell tower not only knows which phone is trying (each phone has a unique identifier) to connect at which time, but can also calculate the distance from itself to such a phone (e.g., using the time elapsed between a ping command, and the corresponding reply). As such, having access to the locations of cell towers, as well as the distance of a phone from each of the involved towers, we can use a "triangulation" technique to pinpoint the location of a phone. However, in practice, such techniques often can only locate a smartphone in an area instead of an exact position.Moreover, using this technique for location tracking could raise privacy concerns, in that, it requires access to the identifier of each phone that may disclose user identify as well. Therefore, when only having the corresponding permissions, cellular network can be used for contact tracing, and meanwhile the user identity must also need to be protected.

How to collect the encounter records. Even though there are distributed models for contact tracing which allow each user to individually control whether or not to disclose its own encounter records, we advocate a centralized model in which each individual user's contact is collected by a central agency, and then stored at a central backend. This is bound to raise privacy concerns, and hence we need to introduce privacy-preserving mechanisms.To this end, we will generate pseudonyms for each user periodically and the linkage between a pseudonym and the real user is only resolved at the trust authority. The authority is only allowed to link pseudonyms to real users when the pseudonyms belong to (i) infected individuals that are confirmed by healthcare authorities or (ii) individuals who have close contact with infected ones. As such, the privacy of individuals who have no risk of infection will be preserved. A similar approached has been proposed and adopted in ROBERT [3].

Moreover, privacy concerns might arise from using the cell tower information for locating users because the identifier of a user's phone needs to be accessed. In order to mitigate these, we can also link such identifiers with pseudonyms. Similar approaches can also be applied in WiFi positioning. Therefore, the entry of data for upload involves self pseudonym, encounter pseudonym, timestamp, Bluetooth proximity, ultrasound proximity.

After processing each upload data entry, the output of this improved data collection procedure is data entries that involve pseudonyms of two encounters, the timestamp, the adjusted proximity, and the infection risk. In particular, the adjusted proximity is the weighted average from combining proximity measured from different sources (i.e., Bluetooth, ultrasound, WiFi, and cell tower), and the infection risk can be obtained by using environment detecting heuristics. For example, when there is no proximity measurement from ultrasound and the WiFi proximity indicates encounters are in different rooms, the infection risk can be adjusted to a low level. Besides the above privacy-preserving data collection methods, we can also apply tools from the field of differential privacy [8]. These utilize different kinds of data processing and noise injection methods, thereby making it difficult for any party to determine whether or not a particular individual is in the original data records and providing privacy protection to the users. Such a guarantee on the privacy would encourage more users to join the system.

3. DATA INTEGRATION AND SUSCEPTIBILITY GRAPH

Here, the goal will be to create a "susceptibility graph" that describes compactly the different ways in which disease is likely to spread. We begin by introducing this graph, and then also describe how to construct this graph by integrating the data from multiple sources. The graph would be time-variant.

3.1 Graph Structure

A basic version of the graph would contain the following components, and the designer is free to make reasonable modifications on it.

- **Nodes.** Each node represents an individual that could be potentially infected. Individuals that are isolated will be removed from the graph. Also, we can remove individuals who have recovered from the virus from the graph. However, since recovered individuals lose their antibodies for most viruses (including COVID-19), re-infections are possible after a period of time, so they would have to be reintroduced into the graph after some time. We use $\mathcal{N}$ to denote the set of nodes (individuals).

- **Node infection state.** For each individual i and time t, we use $X_i(t)$ to denote its infection state, where $X_i(t) = -1$ means that this individual does not have the disease, $X_i(t) = 0$ denotes that it has the virus but cannot spread the virus, and $X_i(t) = 1$ denotes that it has the virus and is able to spread the virus. For individual i and time t, we use $U_i(t)$ to denote its test results at time t. $U_i(t) = 0$ means individual i does not take a test at time t, $U_i(t) = -1$ means it is tested as negative at time t, and $U_i(t) = 1$ means it is tested as positive at time t.

- **Edges.** For every two nodes i and j, If persons i, j have direct contact, then there is an undirected edge (i, j) between them. We are free to choose the way in which we define "contact": for example if these people are staying less than 6 feet apart for at least a certain duration of time co-occurrence in a narrow space (e.g., a room and a bus), or participating in the same event. The contact information can be deduced by the techniques introduced in Section 2. We use $\mathcal{E}$ to denote the set of undirected edges and that (i, j) is in $\mathcal{E}$ means there is a contact between i and j.

- **Base infection probabilities.** Given the fact that we cannot test every individual, each untested individual has a base probability of being infected. This probability can be helpful for some tasks like finding a suspected infected individual. For instance, a person who contacted 500 people yesterday could be more likely to be infected than someone who was in contact with a confirmed positive person; and we can use the base infection probabilities to deduce this probability. The simulations in Section 5 also indicate that take the base infection probabilities into account can find and isolate more infected people. The base infection probability could be time-varying (e.g. abrupt changes due to certain events), and we use $p_b(t)$ to denote the base infection probability at time t.

- **Spreading probabilities.** In case two individuals i, j have been in contact, and one of them, say user i, was a positive case, then there is a chance that individual j got infected by the contact. This chance may also be time-variant. We let the spreading probability be denoted as $p_{i\to j,s}(t)$, which is the probability that j got infection from a contact with i. The calculation of the probability will be discussed later.

- **Time.** The time can either be continuous or discrete. Continuous time better fits the reality, but such an assumption also needs more storage and computation power to process the graph. Besides, given the fact that there are delays, or the occurrence time of events or contacts are not known precisely, how to construct an accurate timely graph needs to be investigated. If time is discretized, then the duration of a discrete time-slot could be anywhere from several minutes to one day. Using time slots can help reduce the storage and computation resources required.

3.2 Graph Construction

As discussed earlier, the graph $\mathcal{G}$ consists of a set of nodes and a set of edges, where each node holds a base infection probability and each edge holds a spreading probability. The nodes, edges, and the probabilities all need to be deduced from the data, and the data can be multi-sourced, for example wifi access logs of all users, CCTV cameras, or Bluetooth scanning based contact tracing. More details on how to construct such a graph are as follows:

- **Individual identification.** Identifying the individuals and avoiding duplication are necessary for the success of graph construction. How to do these may depend on the data collection methods. For instance, in the university WiFi logging system, an individual has and only has one access ID, and thus, this ID can be used to identify an individual. However, in general WiFi systems, an individual may have multiple devices, and removing the duplication is significant in this case. One method is restricting the tracking to one type of device such as mobile phones. When using the Bluetooth contact tracing, we can use the IDs of the mobile phones to identify the individuals, which is also applicable when using Bluetooth contact tracing and WiFi logging simultaneously.

- **Edge detection.** If there is a possible contact between two individuals, then an edge should be generated to connect these two individuals. The contact can have multiple types. For instance, the contact can be staying less than 6 feet, co-occurring in the same room at some time period, or connected to the same access point during some time period. This information can be deduced from the collected data.

- **Base infection probabilities.** The base infection probability can be deduced from the positive rate per test or the number of confirmed positive cases per randomly tested individuals. For instance, a university randomly tested 1,000 students and found 20 positive cases, then we can assume that each student of the university has 2% probability to be positive. If we do not have this information, we can use the number of newly detected infections in a period with a multiplier as the estimate.

- **Spreading probabilities (link probabilities).** Deducing the spreading probabilities is a relatively harder task, which can be divided into two steps.

The first step is to infer the type or level of contacts between two individuals. Have they stayed closer than 6 feet or stayed in the same room for a while? The second step is to deduce the link probability. Accurate characterizations of the link probabilities could come from exposure data studies to the virus. However, a reasonable model would be to use a concave function of time to estimate the link probability.

- **Testing results.** For an individual, if this individual has taken a virus test and got the result, then we know whether this individual has the disease or not (with a certain confidence).

3.3 Data Integration

In real-world scenarios, there are multiple data sources. For example, different contact tracking data sources as described earlier (Bluetooth or ultrasound contact tracing data, WiFi logs, GPS, etc.) could be integrated to greatly improve the quality of contact tracing. The integration could be done by using filtering techniques, in which we compute the probability of an edge conditioned on the (multi-source) information available to us. We would typically rely upon generative models of the data in order to compute these conditional probabilities. With multiple data sources, we need to deal with inconsistent data. For instance, Bluetooth gives relative location information whereas WiFi gives absolute location information, and the information of two sources may be inconsistent. We can deal with this issue by assuming that the data collections of the sources are random and independent and assign probability distributions to them. Probabilistic description allows "soft recovery" of data after we use filtering algorithms. Bayesian updates can be used to merge or pool information from various source. We can use Kalman filtering or some other filtering algorithm.
Such an integration can yield us the following kind of improvements:

- Reduced inaccuracies and better estimates of the link probabilities. Consider for example the case when people could have social contact by virtue of being located in a crowded facility such as students in the same classroom or people in the same flight. However, data sources such as building information, WIFI access might be noisy. In this case, one could combine GPS data (collected from probably smartphone usage) in order to yield an accurate estimate of social contacts.

4. TESTING UNDER RESOURCE CONSTRAINTS

The goal here is to leverage the information contained in the susceptibility graph in order to sequentially choose individuals for testing so as to minimize the spread of the pandemic. *Note that this objective is quite different from focusing on testing individuals with the highest probability of infection*, which is what current systems try to do. Rather our focus must be *on testing individuals that have the highest expected impact on viral spread.* Consider the following example as an illustration.

Example: Assume that two individuals i and j are infected with probabilities 0.1 and 0.3, respectively. However, assume that the expected number of individuals that i encounters is 50 times larger than the expected number of individuals that j comes in contact with. In this case, it makes more sense to prioritize testing individual i over individual j. This is another reason why we should test healthcare workers more often, because of their frequent contact with a large number of individuals. Based on this key insight, our goal will be to:

- Develop learning based approaches that result in smart testing capabilities which balance the exploration and exploitation subject to testing constraints. Isolate individuals who have been tested positive and quarantine their contacts.

- Our model will also incorporate practical issues such as inaccurate estimates, testing errors, pool testing, etc.

- Develop efficient rules of thumb that can be easily implemented in practice. This could mean testing asymptomatic individuals who have not encountered a confirmed infected person, but have made a large number of contacts.

4.1 Suspicious Infection Inference

One significant task is to find the most likely infected individuals from the partial observations, i.e., the test results of some individuals. To do this, one way is to interpret the probability that a person is infected given the partial observations, such as ("noisy") contact graph or test results of a few individuals from the graph etc. These algorithms could be based upon the susceptibility graph constructed by using the methods stated in Section 3.

4.1.1 Partial Observed Markov Decision Process (POMDP)

We formulate the problem of sequential testing for COVID-19 as a Markov Decision Process (MDP). Population is composed of N individuals, and the state evolves at discrete times $t \in [1, T]$. Let $X_i(t) \in \{0, 1\}$ denote the hidden state of individual i at t, where $X_i(t) = 0$ means that i is free of disease at t and $X_i(t) = 1$ indicates that i is infected. We use the vector $X(t) := (X_1(t), X_2(t), ..., X_N(t)) \in \{0, 1\}^N$ to represent the state of the entire system. Let $\mathcal{X} := \{0, 1\}^N$ denote the state-space of the network. Note that the state

vector $X(t)$ is never fully revealed to the learner[2].

Test and Quarantine: At each time $t \in [1, T]$, the learner has a unit budget to choose an individual $i \in [1, N]$ in order to "sample" (test for infection). Sampling an individual i at t reveals the state $X_i(t)$. We let $U(t) \in [0, N]$ denote the sampling decision at time t. In case no one is sampled at t, we let $U(t) = 0$. We let $Y_i(t)$ denote the test result at time t; $Y_i(t) = +1$ means the person tested positive, $Y_i(t) = -1$ means the test was negative, and $Y_i(t) = 0$ means that the individual was not tested at time t. The vector comprising of observations $Y_i(t)$ is denoted $Y(t)$.

If sampled individuals are found to be infected, then they are isolated, i.e., kept out of the population, and hence cannot spread the disease to their neighbors. We let $Q(t)$ denote the set of those individuals who are isolated at time t.

State Transition: Let us now look at the *controlled* transition probabilities of the controlled Markov process $X(t)$. We first introduce some notations. For $x, y \in \mathcal{X}$, define

$$\Delta_1(x, y) = \mathbb{1}\left\{\sum_{i=1}^{N} |x_i - y_i| = 1\right\} \quad \text{and} \tag{2}$$

$$\Delta_2(x, y) = \begin{cases} i \text{ if } x_i \neq y_i \text{ and } \Delta_1(x, y) = 1, \\ \emptyset \text{ otherwise .} \end{cases} \tag{3}$$

Clearly, $\Delta_1(y, x)$ assumes value 1 only if x and y differ in a single position; since disease can spread to only one more person during two consecutive times, this function is 0 if x cannot evolve to y in one single time-step. Δ_2 provides us with the node that "transitioned" to diseased state when the system evolved in a unit step from x to y. Thus, the single-step transition probabilities are given as

$$P_t(x, y) = \Delta_1(x, y) p \sum_{i \in \mathcal{V}'_t} w'_t(i, \Delta_2(y, x)). \tag{4}$$

Objective: Let $\mathcal{F}_t := \cup_{s=1}^{t} (U(s), Y(s), \ell(s))$ be the observation history of the learner [18]. Then, the policy π is a sampling decision at t on the basis of $\mathcal{F}_{t-1}$, i.e., $\pi : \mathcal{F}_{t-1} \mapsto U(t),\ t \in [1, T]$. Our goal is to find a policy that solves the following problem,

$$\min_{\pi}\ \mathbb{E}_{\pi}\left(\sum_{t=1}^{T} \|X(t)\|_1\right), \tag{5}$$

$$\text{s.t. } \mathbb{E}_{\pi}\left(\sum_{t=1}^{T} \mathbb{1}\left(U(t) \neq 0\right)\right) \leq C, \tag{6}$$

where $\|\cdot\|_1$ denotes the L_1 norm and C is the total testing capacity. The instantaneous cost $\|X(t)\|_1$ encourages the policy to keep the infections as low as possible in an as early as possible manner. The capacity constraints are crucial because not many testing kits are available during epidemics. An alternative, somewhat equivalent and simpler objective is to remove the capacity constraints altogether and include a cost for using testing kits,

$$\min_{\pi} \mathbb{E}_{\pi}\left(\sum_{t=1}^{T} \|X(t)\|_1 + \lambda \mathbb{1}\left(U(t) \neq 0\right)\right), \tag{7}$$

where $\lambda > 0$. We now briefly discuss how to solve the above discussed MDP. Theoretical results on the existence of optimal policies, and methods to solve constrained MDPs, or POMDPs can be found in [4, 21, 22]. In case the parameters describing the environment are unknown, we can use machine learning techniques developed in [20].

Belief State MDP: We introduce a belief state, which is a posterior distribution over the state space $\mathcal{X}$. This transforms the POMDP to a continuous-state MDP on the belief state. We denote the belief state by $\mathcal{I}(t) = \{\mathcal{I}(t, x)\}_{x \in \mathcal{X}}$, where $\mathcal{I}(t, x) := \mathbb{P}\left(X(t) = x | \mathcal{F}_t\right)$. By Bayes' Rule, the terms $\mathcal{I}_t(x)$ are computed recursively as

$$\mathcal{I}_{t+1}(x) = \sum_{y \in \mathcal{X}} \mathcal{I}_t(y) \mathbb{P}\left(Y_{U(t)} | X(t) = y\right) P_t(y, x), \tag{8}$$

where the state transition probabilities $P_t(y, x)$ are as discussed in *(4)*.

Optimal Policy: The optimal sampling policy can be obtained by solving the following set of non-linear Dynamic Programming equations [12],

$$V_t(\mathcal{I}_t) = \sum_{x \in \mathcal{X}} \|x\|_1 \mathcal{I}_t(x) + \min_{u \in [0, N]} \left(\mathbb{E}\, V_{t+1}(\mathcal{I}_{t+1}) + \lambda \mathbb{1}\{u \neq 0\}\right), \tag{9}$$

$$V_T(\mathcal{I}) = \sum_{x \in \mathcal{X}} \|x\|_1 \mathcal{I}(x),\ \forall \mathcal{I} \in \Delta(\mathcal{X}), \tag{10}$$

where $\Delta(\mathcal{X})$ denotes simplex on $\mathcal{X}$ and $\mathcal{I}_t$ denotes representative belief state at time t. Optimal sampling action at time t in state $\mathcal{I}_t$ corresponds to a minimizer of the r.h.s. in the above equation. However, equations *(9)*, *(10)* are computationally intractable because the number of required computations is $O(2^N)$. Thus, we propose tractable provably approximate solutions next.

4.1.2 Hidden Markov Model

Since we might not observe the susceptibility graph $\mathcal{G}$, we can model it as a hidden Markov process [17]. Assume that the system evolves at discrete time-instants $t = 1, 2, \ldots, T$. Each slot could be of a duration equal to half hour, one hour, or one day. For each time slot t, we use $X_i(t)$ to denote the infection state of the i-th individual. In this section we slightly tweak the binary-valued state model that was described earlier, and allow it to assume 4 values. This allows us to design an algorithm that is more accurate. Thus, we let $X_i(t) = -1$ if

[2]So this problem is a partially observable MDP (POMDP), which is non-trivial to solve in general case.

individual i is not infected by the virus at time t, and $X_i(t) = 0$ if individual i is infected by the virus but cannot spread the virus, i.e., is in incubation, $X_i(t) = +1$ if individual i is infected by the virus and can spread the virus, and finally $X_i(t) = -2$ if they have recovered from the virus or have been isolated already. We call $X_i(t) = 0$ "inactive" and $X_i(t) = 1$ active.
The rest of the discussion in this section makes the following simplifying assumptions:

- States $X_i(t)$ and decisions $U_i(t)$ do not change within a slot.
- Upon becoming an active spreader, an individual can spread the disease only after the current timeslot ends. This might seem to be restrictive, but is justifiable since our modeling procedure already introduces "noise" due to erroneous tracing and testing.
- The spreading probability, denoted as p_s, is a constant that is independent of other parameters such as the values of the states, the number of days one has been infected, etc.
- After becoming an inactive infected person, at each time slot the individual becomes active with a probability equal to $p_{0,1}$.
- For an active infected person, for every time slot, this individual has a constant probability to get removed. We use $p_{1,-2}$ to denote this probability.
- For an individual at state $X_i(t) = a$, it has a constant probability $p_{a\to b}$ to be tested to be state $U_i(t) = b$.

Let $\mathcal{F}_t$ be the filtration generated by $(X_i(s), Y_i(s), i \in \mathcal{N}, s \le t)$. With the above assumptions in place, we can write the "dynamics" or transition probabilities governing $X(t) = \{X_i(t))\}_{i=1}^N$ as follows,

$$\mathbb{P}\{X_i(t+1) = 2 \mid \mathcal{F}_t\} = \mathbb{1}_{X_i(t)=-2} + \mathbb{1}_{X_i(t)=1} \cdot p_{1,-2}, \quad (11)$$

$$\mathbb{P}\{X_i(t+1) = 1 \mid \mathcal{F}_t\} = \mathbb{1}_{X_i(t)=1}(1 - p_{1,-2}) + \mathbb{1}_{X_i(t)=0} \cdot p_{0,1}, \quad (12)$$

$$\mathbb{P}\{X_i(t+1) = 0 \mid \mathcal{F}_t\} = \mathbb{1}_{X_i(t)=0} \cdot (1 - p_{0,1}) + \mathbb{1}_{X_i(t)=-1}\Big(1 - \prod_{j:(i,j)\in\mathcal{E}, X_j(t)=1} (1 - p_s(t))\Big), \quad (13)$$

$$\mathbb{P}\{X_i(t+1) = -1 \mid \mathcal{F}_t\} = 1 - \mathbb{P}\{X_i(t+1) = 1, 0, \text{ or } -2 \mid \mathcal{F}_t\}, \quad (14)$$

and the dynamics or transition probabilities of $U(t) = \{U_i(t)\}_{i=1}^N$ as follows,

$$\mathbb{P}\{U_i(t) = 1 \mid \mathcal{F}_t\} = \mathbb{1}_{U_i(t)\neq 0}\Big(\sum_{s=-1,0,1} \mathbb{1}_{X_i(t)=s} \cdot p_{s\to 1}\Big), \quad (15)$$

$$\mathbb{P}\{U_i(t) = 1 \mid \mathcal{F}_t\} = \mathbb{1}_{U_i(t)\neq 0}\Big(\sum_{s=-1,0,1} \mathbb{1}_{X_i(t)=s} \cdot p_{s\to -1}\Big), \quad (16)$$

where whether $U_i(t) = 0$ or not is determined by the tracing or testing algorithms.
Clearly, $X(t) = \{X_i(t)\}_{i\in\mathcal{N}}$ is a Markov process, and if we are provided with the values of $U_i(0), U_i(1), U_i(2), ...$, then our goal is to find the most likely values of $X(t)$. In order to do this, we might use Markov chain Monte Carlo (MCMC) algorithms such as Gibbs Sampling. Readers may refer to [27] for a review of MCMC algorithms.

4.1.3 Graph Embedding

Computing the infection probabilities of individuals directly will be computationally cumbersome, and we can utilize graph embedding [5] techniques in order to find suspicious infected individuals. These techniques map the nodes of a graph to points in $\mathbb{R}^d$, where d is a natural number. If the graph embedding algorithm is properly chosen, then if two points are close in the space $\mathbb{R}^d$ then they are also close in the susceptibility graph, so that the probability that the virus spreads from one individual to the other is high. Note that in the graph each node may have up to $|\mathcal{N}|$ edges, but in the embedded graph, each node only has d coordinates. Since the number of edges might be much more than d, performing computations with the embedded coordinates is much more efficient than directly working with the original graph.

5. SIMULATIONS

In this section, we use a simulation to indicate the necessities of contact tracing and building a contact graph. Contact tracing is an essential technique for finding potential infectious people. A commonly employed naive contact tracing technique is to trace and test only those who have had contact with a confirmed positive person. We call this simple and intuitively appealing contact tracing policy as Policy 1. However, this method may not be optimal, especially under circumstances when a sizeable proportion of the population is infected. To see why this might be the case, consider the scenario when two people are waiting to get tested. The first person had a close contact with a confirmed infected person, while the second person did not have any such close contact with a confirmed infected person; but did closely contact 500+ untested people (for example this person works in a supermarket). Policy 1 will suggest to us to test the first person; however, when a significant proportion of the population (e.g., 1%) are positive, in

expectation, the second person would have been in close contact with more than one infected people. Thus, taking the number of contacts with all people into consideration could significantly improve the tracing. The simulation in this section also confirms our claim.
These two policies can be mathematically described as follows.

- Policy 1: Fix a time frame of a duration say 2 weeks, and call the time duration composed of the previous two weeks as the "tracing window". At any given time we only take into account those contacts that occurred during the tracing window. Let p_s be the probability that the virus spreads from an infected person to a healthy person during a contact. We assume that p_s is constant and known. For any person s, given that person s contacted m confirmed infected persons during the tracing window, we use

$$\mathbb{P}\left(\{s \text{ got infected}\}\right) = 1 - (1 - p_s)^m \qquad (17)$$

  to measure the risk that person s is infected. We then choose to test those persons who have the highest probabilities of being infected.

- Policy 2: It additionally utilizes the contact graph, and checks the number of contacts of each person $s \in \mathcal{S}$. Hence, if n is the number of contacts of s in the tracing window, we let

$$\mathbb{P}(\{s \text{ got infected}\}) = 1 - (1 - p_s)^m (1 - p_b p_s)^n \qquad (18)$$

  in order to measure the risk that person s is infected. Over here, p_b is the so-called base infection probability, which can either be a constant or depend on the proportion of confirmed cases of the population (i.e., adaptive). Note that we are assuming that the infection status of these n people are unknown.

Our simulations results are depicted in Fig. 2, and clearly show the superior performance of Policy 2 as compared to that of Policy 1. More details on the simulation setup are as follows:

- The population size is 1000 people, and a single person (that is chosen uniformly at random from the population) is infected by the virus at day zero.

- Regarding the transmission capability of the virus, we assume that a person will be able to spread the virus 1 day after getting infected. Moreover, a person remains infected for at least 7 days. After this duration, on each day the person will change his state (to either isolated due to its symptoms, recovered, or deceased) with a probability of 1/7. Thus, in expectation, the virus lasts for 14 days. A person whose state has changed to removed, will not spread the virus or get infected.

- During any particular day, any two people in the city meet each other with a probability of 0.01. Thus, on average, a person meets around 10 people per day in expectation. When two people meet, and one of them is infected while the other is not, the virus will spread with a probability of $3/(14 \times 10)$; hence an infected person spreads the virus to an average of $R = 3$ people before being removed.

- Each day the community chooses 20 people to quarantine by using its policy. If quarantined persons are found to be infected, then they will be isolated until they are removed. Otherwise, they will be quarantined for 14 days, and then will be back to the normal schedule.

- We assume that the community as a whole knows all the contacts between all of its people, and whenever a person is removed the community gets to know this information at the beginning of the next day. Also, we assume that the spreading probability p_s is known to the community. We note that, with the knowledge of p_s and assuming the value of p_b, the community is able to compute Eq. *(17)* and Eq. *(18)*.

We perform simulations for 150 consecutive days, and record the cumulative infections in the population for the following 5 policies and parameters:

- No contact tracing of any sort is utilized.

- Policy 1 (Eq. *(17)*).

- Policy 2 with $p_b = 0.02$, where 0.02 is a well tuned value.

- Policy 2 with $p_b = 0.2$, where 0.2 is an example of a not well tuned value of p_b.

- Policy 2 with adaptive $p_b = N_{\mathrm{rr}}/1000$, where "rr" denotes "recently removed" and N_{rr} means the number of people removed in the tracing window (i.e., the last two weeks).

Our simulation results are summarized in Fig. 2. We explicitly state the number of total infections in Table 1.

tracing policy	parameter p_b	total infections
no tracing	—	987
Policy 1	—	617
Policy 2	0.02	540
Policy 2	0.2	669
Policy 2	adaptive	569

Table 1 – Total number of infections of the virus under different tracing policies.

We summarize our findings as follows.

- Contact tracing and quarantine facilities are essential in order to control the spread of virus. Without these, the total number of infections are around 987,

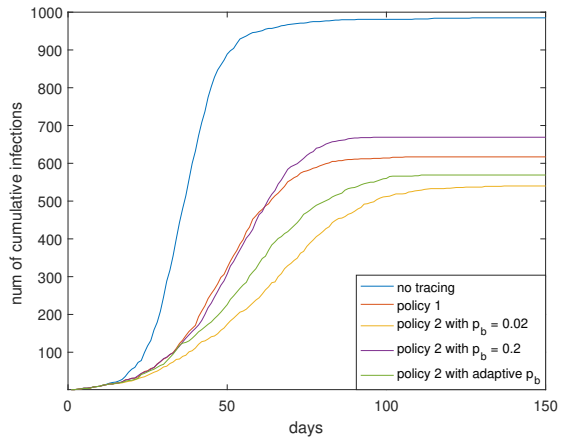


Fig. 2 – Simulation of the spreading of the virus under different tracing policies.

which is approximately the entire population. Even a simple contact tracing technique such as Policy 1, reduces the total infections to 617 (which is a 37% reduction). Thus, it is worth the effort to establish a system that can track the contacts of the people in order to better detect the infections.

- Compared to Policy 1 that only considers the contacts with confirmed infected people, Policy 2 also takes contacts with untested (but probable to be infected) people into account, and hence it has a better performance. This is clearly demonstrated in its superior performance in all of the three experiments, for example when $p_b = 0.02$, we get a gain of around 12%.

- When p_b is tuned properly (e.g., $p_b = 0.02$), Policy 2 performs better. However, the tuning effort is substantial, and might be deemed infeasible in practice. For instance, when $p_b = 0.2$, the performance of Policy 2 is worse than Policy 1. Thus, Policy 2 with adaptive values of p_b is a good option in practice.

- If we are to use only contact tracing and quarantine facilities, our performance is not very good. Even if tracing is possible for 2% of the population per day, the majority of the population will get infected after a few months. Hence, it is necessary to combine contact tracing and quarantines with other policies, e.g., avoiding contacts to reduce the number of contacts, and also wearing masks to reduce the virus spreading probability.

6. CONCLUSION

In this paper, we have provided a detailed blueprint on how to contain the spread of a pandemic by integrating the use of various wireless technologies with sequential learning based techniques. In particular, we show how different wireless technologies could be leveraged to improve contact tracing efforts and reduce the probabilities of detection and false alarms. The idea is to use possibly disparate wireless data streams for data collection, then integrate this data to improve coverage, reduce probabilities of errors and false alarms and create a *susceptibility* graph that could be used for intelligent testing. Based on this susceptibility graph, we show how to develop efficient sequential learning based algorithms in order to minimize the spread of the virus infection. Another contribution is that we develop provably optimal algorithmic solutions that rely upon the theory of partially observable Markov decision processes. In particular, we show that current contact tracing plus testing strategies that are aimed at identifying (and testing) individuals with the highest probability of infection are inefficient. Instead, we find that it is better to test those individuals whose expected impact on virus spread is the highest. We formulate the testing problem as a Partially Observable Markov Decision Process whose goal is to minimize the expected spread of the virus subject to testing capacity constraints. We provide efficient algorithmic solutions to this problem and show via numerical results that our solution substantially reduces the spread of the virus.

ACKNOWLEDGEMENT

This work was supported in part by NSF grants CNS-1618520 and CNS-2028547.

REFERENCES

[1] Covid-19 contact tracing: a briefing. `https://www.bmj.com/content/369/bmj.m1859`. (Accessed on 07/02/2020).

[2] The global economy is expected to shrink by 3% this year. `https://www.economist.com/graphic-detail/2020/04/14/the-global-economy-is-expected-to-shrink-by-3-this-year`.

[3] Robert – robust and privacy-preserving proximity tracing protocol. `https://github.com/ROBERT-proximity-tracing`. (Accessed on 09/17/2020).

[4] Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

[5] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637, 2018.

[6] CDC. How to protect yourself & others. `https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html`.

[7] Aaqib Bashir Dar, Auqib Hamid Lone, Saniya Zahoor, Afshan Amin Khan, and Roohie Naaz. Applicability of mobile contact tracing in fighting pan-

demic (covid-19): Issues, challenges and solutions. *Computer Science Review*, page 100307, 2020.

[8] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[9] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491), 2020.

[10] Stephen M Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H Grad, and Marc Lipsitch. Projecting the transmission dynamics of sars-cov-2 through the postpandemic period. *Science*, 368(6493):860–868, 2020.

[11] Mirjam E Kretzschmar, Ganna Rozhnova, Martin CJ Bootsma, Michiel van Boven, Janneke HHM van de Wijgert, and Marc JM Bonten. Impact of delays on effectiveness of contact tracing strategies for covid-19: a modelling study. *The Lancet Public Health*, 5(8):e452–e459, 2020.

[12] Vikram Krishnamurthy. *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.

[13] Panqanamala Ramana Kumar and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[14] Douglas J Leith and Stephen Farrell. Coronavirus contact tracing: Evaluating the potential of using Bluetooth received signal strength for proximity detection. 2020.

[15] Yuxiang Luo, Cheng Zhang, Yunqi Zhang, Chaoshun Zuo, Dong Xuan, Zhiqiang Lin, Adam C. Champion, and Ness Shroff. Acoustic-turf: Acoustic-based privacy-preserving covid-19 contact tracing, 2020.

[16] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*, pages 1–14, 2007.

[17] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[18] Albert N Shiryaev. *Optimal stopping rules*, volume 8. Springer Science & Business Media, 2007.

[19] Viktoriia Shubina, Sylvia Holcer, Michael Gould, and Elena Simona Lohan. Survey of decentralized solutions with mobile devices for user location tracking, proximity detection, and contact tracing in the covid-19 era. *Data*, 5(4):87, 2020.

[20] Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in Markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

[21] Richard D Smallwood and Edward J Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

[22] Edward J Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2):282–304, 1978.

[23] Qiang Tang. Another look at privacy-preserving automated contact tracing. *arXiv preprint arXiv:2010.13462*, 2020.

[24] Qiang Tang. Privacy-preserving contact tracing: current solutions and open questions. *arXiv preprint arXiv:2004.06818*, 2020.

[25] Amee Trivedi and Deepak Vasisht. Digital contact tracing: technologies, shortcomings, and the path forward. *ACM SIGCOMM Computer Communication Review*, 50(4):75–81, 2020.

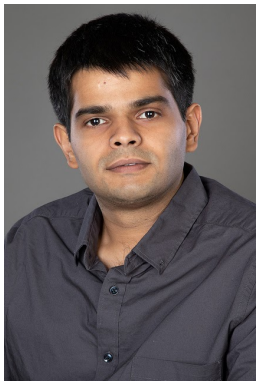
[26] Amee Trivedi, Camellia Zakaria, Rajesh Balan, and Prashant Shenoy. Wifitrace: Network-based contact tracing for infectious diseases using passive wifi sensing, 2020.

[27] Don Van Ravenzwaaij, Pete Cassey, and Scott D Brown. A simple introduction to markov chain monte–carlo sampling. *Psychonomic bulletin & review*, 25(1):143–154, 2018.

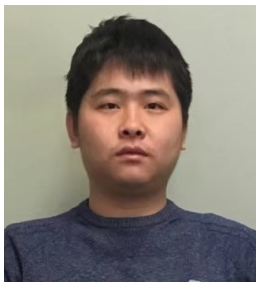
[28] Haohuang Wen, Qingchuan Zhao, Zhiqiang Lin, Dong Xuan, and Ness Shroff. A study of the privacy of covid-19 contact tracing apps. In *International Conference on Security and Privacy in Communication Networks*, 2020.

[29] Qingchuan Zhao, Haohuang Wen, Zhiqiang Lin, Dong Xuan, and Ness Shroff. On the accuracy of measured proximity of Bluetooth-based contact tracing apps. In *International Conference on Security and Privacy in Communication Networks*, 2020.

AUTHORS



Rahul Singh received his B. Tech. degree in Electrical Engineering from Indian Institute of Technology, Kanpur, India, in 2009, M.Sc. degree in Electrical Engineering from University of Notre Dame, South Bend, IN, in 2011, and his Ph.D. degree in Electrical and Computer Engineering from the Department of Electrical and Computer Engineering Texas A&M University, College Station, TX, in 2015. Currently he is an Assistant Professor at the Department of Electrical Communication Engineering, the Indian Institute of Science. His research interests include stochastic control, machine learning, applied probability, networks and large-scale complex cyber physical systems. He has earlier worked as a Postdoctoral Researcher at the Laboratory for Information Decision Systems (LIDS), Massachusetts Institute of Technology, at Intel, Santa Clara as a Machine Learning Engineer, and as a Data Scientist at Encored Technologies.



Wenbo Ren received his B.S. degree in Electronic Information Engineering from University of Science and Technology of China in 2016. Since then, he has been a Ph.D. student in Computer Science and Engineering at The Ohio State University. His research interests include machine learning theory and algorithms, with a focus on ranking, active learning, sequential decision making, recommender systems, etc. He has made publications or served as reviewers for top-tier machine learning conferences including ICML, NeurIPS, and AISTATS.

Fang Liu received his B.S. degree in Information Engineering from Shanghai Jiao Tong University, China, and his Ph.D. degree in Electrical and Computer Engineering from The Ohio State University. In 2020, he joined Facebook Inc. as a research scientist. His research interests include machine learning theory and systems, with a focus on sequential decision making under uncertainty. Dr. Liu has served as a program committee member or reviewer for many top-ranked machine learning conferences, such as ICML, NeurIPS, ICLR, and AAAI.

Dong Xuan received his B.S. and M.S. degrees in Electronic Engineering from Shanghai Jiao Tong University, China, and his Ph.D. degree in Computer Engineering from Texas A&M University. In 2001, he joined the Department of Computer Science and Engineering at The Ohio State University, where he is currently a full professor. He was on the faculty of Electronic Engineering at Shanghai Jiao Tong University from 1993 to 1998 and a visiting scholar in Computer Science at City University of Hong Kong from 1997 to 1998. He was a research assistant/associate in the Real-Time Systems group in the Department of Computer Science at Texas A&M University from 1998 to 2001. His research interests include computer networking and mobile systems. Dr. Xuan has served as an editor for IEEE TPDS and ACM ToSN, and has been a TPC member for a number of IEEE and ACM flagship conferences such as IEEE INFOCOM, ICDCS, ICNP, and ACM MobiHoc. He is a recipient of the National Science Foundation CAREER Award and the College of Engineering Lumley Research Award at The Ohio State University.

Zhiqiang Lin received his Ph.D. degree in Computer Science from Purdue University in 2011. He was a faculty member in the Computer Science Department at University of Texas at Dallas between 2011 and 2017. Since 2018, he has been a faculty member in the Department of Computer Science and Engineering at Ohio State University. His research interests are systems and software security, with an emphasis on binary analysis and vulnerability discovery. Dr. Lin is currently an associate editor of ACM TOPS, IEEE TDSC, and IEEE TMC. He has also served as a TPC member for numerous systems security conferences including IEEE S&P, ACM CCS, USENIX Security, and NDSS. Dr. Lin received the AFOSR Young Investigator award (2014) and the NSF CAREER award (2015). He is also a recipient of VMware Faculty Research Award (2012) and the Outstanding Junior Faculty Research Award at UT Dallas (2013).

Ness B. Shroff received his Ph.D. degree in electrical engineering from Columbia University in 1994. He joined Purdue University immediately thereafter as an Assistant Professor with the School of Electrical and Computer Engineering. At Purdue, he became a Full Professor of ECE and the director of a university-wide center on wireless systems and applications in 2004. In 2007, he joined The Ohio State University, where he holds the Ohio Eminent Scholar Endowed Chair in networking and communications, in the departments of ECE and CSE. He holds or has held visiting (chaired) professor positions at Tsinghua University, Beijing, China, Shanghai Jiaotong University, Shanghai, China, and IIT Bombay, Mumbai, India. He has received numerous best paper awards for his research and was listed in Thomson Reuters' on The World's Most Influential Scientific Minds, and has been noted as a Highly Cited Researcher by Thomson Reuters in 2014 and 2015. He has served on numerous editorial boards and chaired various major conferences and workshops. He currently serves as the steering committee chair for ACM Mobihoc, and Editor in Chief of the IEEE/ACM Transactions on Networking. He received the IEEE INFOCOM Achievement Award for seminal contributions to scheduling and resource allocation in wireless networks.

MACHINE LEARNING-ASSISTED CROSS-SLICE RADIO RESOURCE OPTIMIZATION: IMPLEMENTATION FRAMEWORK AND ALGORITHMIC SOLUTION

Ramon Ferrús, Jordi Pérez-Romero, Oriol Sallent, Irene Vilà, Ramon Agustí
Dept. of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), c/ Jordi Girona, 1-3, Barcelona, Spain

NOTE: Corresponding author: Ramon Ferrús (ferrus@tsc.upc.edu)

Abstract *– Network slicing is a central feature in 5G and beyond systems to allow operators to customize their networks for different applications and customers. With network slicing, different logical networks, i.e. network slices, with specific functional and performance requirements can be created over the same physical network. A key challenge associated with the exploitation of the network slicing feature is how to efficiently allocate underlying network resources, especially radio resources, to cope with the spatio-temporal traffic variability while ensuring that network slices can be provisioned and assured within the boundaries of Service Level Agreements / Service Level Specifications (SLAs/SLSs) with customers. In this field, the use of artificial intelligence, and, specifically, Machine Learning (ML) techniques, has arisen as a promising approach to cater for the complexity of resource allocation optimization among network slices. This paper tackles the description of a feasible implementation framework for deploying ML-assisted solutions for cross-slice radio resource optimization that builds upon the work conducted by 3GPP and O-RAN Alliance. On this basis, the paper also describes and evaluates an ML-assisted solution that uses a Multi-Agent Reinforcement Learning (MARL) approach based on the Deep Q-Network (DQN) technique and fits within the presented implementation framework.*

Keywords – 5G, cross-slice resource optimization, deep learning, machine learning, network slicing

1. INTRODUCTION

Network slicing allows operators to customize their networks for different applications and customers [1], [2]. Slices can differ in functionality (e.g. air interface capabilities, mobility tracking features), in performance requirements (e.g. latency, availability, reliability and data rates), or they can serve only specific users (e.g. public safety users, corporate customers, or industrial users). A network slice can provide the functionality of a complete network, including radio access network and core network functions. Support for network slicing has been introduced by the 3rd Generation Partnership Project (3GPP) as part of the first release of the Fifth Generation (5G) system specifications (Release 15), with multiple enhancements still to follow in future releases, as reflected by different study items in progress, such as [3]-[6].

The creation and management of network slices is especially challenging in the Radio Access Network (RAN), where multiple slices can be delivered over the same radio channel and the system shall guarantee that the allocation and distribution of the radio resources within the radio channel is done so that specific requirements per slice can be fulfilled (e.g. guaranteed capacity) while using radio resources efficiently [7]-[14]. Remarkably, the automation of the life-cycle management of network slices in the RAN requires two main functionalities: slice admission control and cross-slice resource optimization.

Slice admission control is needed to decide on the acceptance or rejection of a new RAN slice creation request with specific coverage, functional (i.e. features) and performance (e.g. service quality, capacity) requirements. Under Network as a Service (NaaS) business models such as neutral host services, the slice requirements will be determined by the Service Level Agreement (SLA) / Service Level Specifications (SLS) established between the service provider (e.g. the operator of a RAN infrastructure installed in a venue) and the customer (e.g. a Mobile Network Operator - MNO). The fulfillment of the RAN slice requirements may result in the need to guarantee the availability of a certain amount of radio resources to the new slice, defined in terms of, e.g. number of Resource Blocks (RBs) per cell, percentage of cell capacity, etc. Therefore, the slice admission control shall estimate the amount of radio resources required by the new slice and decide whether this can be enforced given the deployed network capacity and the amount of resources consumed by the already admitted slices.

Once multiple slices have been activated in the RAN, the cross-slice resource optimization shall ensure that the slice requirements are satisfied over time and RAN resources are efficiently utilized. This may imply a dynamic modification of the slice configurations (e.g. specifying the amount of radio resources assigned to each slice at each cell, adjustment of slice-aware scheduling settings, configuration of rate limiters, bandwidth parts, mobility load balancing parameters, access control priorities, etc.) during its lifetime in order to deal with the dynamics of the traffic load of the slice and with the random propagation effects that lead to non-deterministic mapping between radio resource consumption and performance requirements. Cross-slice resource optimization has been identified by 3GPP as a use case in the context of Self-Organizing Network (SON) feasibility studies [15], addressing not only the dynamic allocation of radio resources to slices but also the distribution of other resources such as storage and computing for virtualized implementations.

The decision-making logic for cross-slice resource optimization needs to deal with a lot of uncertainties and random processes associated with the variability in traffic generation, device mobility and radio channel conditions, so it is highly difficult to have an accurate a priori statistical knowledge of the network resource utilization and delivered performance. For this reason, model-free Machine Learning (ML)-based methods, which do not rely on predefined models but are able to learn and/or predict the particular network dynamics as well as to operate under goal-oriented policies, become adequate solutions to the problem [16]. Besides, the complexity of the problem with a huge number of variables and conditions (e.g. particular device capabilities, pending traffic, link channel conditions, resource consumption, etc.) also pushes for the introduction of these sorts of methods. As a result, the system can be in a large number of possible states in which the cross-slice resource allocation needs to determine the optimum capacity sharing among slices. In this case, among the possible ML techniques, deep reinforcement learning (RL) schemes become particularly relevant because they provide faster convergence under large state/action spaces in comparison with classical reinforcement learning.

While there is a significant amount of work addressing the cross-slice resource optimization problem from an algorithmic and performance assessment perspective, less attention has been paid to the practical implementation aspects of these solutions, as it will be further discussed in Section 2. In this respect, departing from 3GPP and O-RAN Alliance specifications, a first contribution of this paper is the delineation of the functional framework and information models to be accounted when targeting a practical realization of ML-assisted cross-slice radio resource optimization solutions for 5G and beyond systems. More specifically, the focus is put here on the identification of the specific functional components enabling the deployment of ML-based solutions for RAN management along with the set of information models that have been defined to represent SLAs, network slice instances' characteristics and slicing-related configuration parameters of 5G base stations. On this basis, a second contribution of this paper is the formulation and assessment of a plausible ML-assisted cross-slice radio resource optimization solution that fits within the delineated implementation framework. The solution makes use of Multi-Agent Reinforcement Learning (MARL) based on the Deep Q-Network (DQN) technique. Illustrative performance results of the proposed solution are provided by means of simulations.

The rest of the paper is organized as follows. Section 2 presents an overview of related works in order to position the paper in relation to the state-of-the-art. Section 3 describes the implementation framework, which is particularized to the proposed ML-assisted cross-slice optimization solution in Section 4 and Section 5 presents some illustrative proof-of-concept results. Finally, our concluding remarks are wrapped up in Section 6.

2. RELATED WORK

Artificial Intelligence (AI) and more specifically ML techniques have been applied in the literature for both slice admission control and cross-slice radio resource allocation. In the area of slice admission control, [17] studied an optimal algorithm using Semi-Markov Decision Processes (SMDP) and then proposed an adaptive algorithm based on Q-learning. Then, other works have considered deep Q-learning [18] along with variants for enhancing the training process, such as deep dueling neural networks [19]. ML tools have also been used for enhancing the slice admission control with traffic prediction, such as in [20], [21], which use Holt-Winters prediction, or [22], which uses a combination of Long Short Term Memory (LSTM) and dense neural networks for predicting the resource usage.

In the field of cross-slice optimization, different approaches exist exploiting several ML tools. Q-learning was used in [23] to design a slicing controller that decides which resource units are allocated to each slice based on requirements at the user level. Q-learning complemented with a genetic algorithm was considered in [24] for scaling down allocated resources to slices for congestion control purposes. In [25] deep deterministic policy gradient (DDPG) is used to allocate resource blocks to different tenants in a cloud RAN environment. In turn, game theory with exponential learning is proposed in [26] to divide the network resources (i.e. bandwidth) among slices using OpenFlow, being a general approach not particularized to the specificities of radio resource allocation. Recently, deep Q learning has become a quite popular tool for allocating radio resources to slices, as reflected by works [27]-[33] that include different variants of this technique and address the problem from different perspectives, such as the joint allocation of computational resources and radio resources to users in [27], the allocation of aggregate capacity per slice to multiple cells in [28], [29], the allocation of resources to slices on a single cell basis in [30], [31], [32], or the allocation of per-cell resources to the different slices jointly considering multiple cells in [33]. Finally, other works have proposed the use of traffic forecasting for cross-slice resource allocation, applying techniques such as LSTM neural networks [34], deep convolutional neural networks [35], Generative Adversarial Networks (GANs) [36], or deep neural networks [37].

This paper introduces several novelties with respect to previous works. First of all, this paper presents a functional framework aligned with current 3GPP and O-RAN specifications for implementing ML-assisted cross-slice radio resource optimization and particularizes it to a specific algorithmic solution coming from our previous work [33]. Instead, the above-mentioned works have put the focus on algorithm development but without going into detail of the mapping on existing architectures from standardization bodies. For example, some works just consider a slicing controller (e.g. [23]) or a network slicing module (e.g. [28], [29]) but without providing details of how this would be mapped on practical architectures. Only in [24] an architectural framework for slice management and orchestration that is aligned with 3GPP is presented, but without providing specific details on the algorithm implementation.

Another important novelty comes from the specification of the SLA terms for a RAN slice to be used by the ML-based solution. This paper takes as a reference the attributes defined in the GSMA Generic Slice Template considered by 3GPP to specify the SLA to be fulfilled for a RAN slice across a geographical area covering multiple cells in terms of, e.g. the total amount of capacity to be provided to each slice. Instead, other approaches such as [28]-[32] just consider the SLA specified in terms of the QoS parameters defined at the user level, but without enforcing any aggregate capacity per slice.

Finally, another difference with respect to previous works comes from the algorithmic solution considered in the proposed framework, which consists of a multi-agent DQN with one agent per slice that learns the capacity to be allocated to each slice in each cell. In contrast to single agent solutions like those of [30], [31], which jointly consider all the tenants when making decisions, the multi-agent approach has advantages such as better scalability as it allows easily adding/removing slices in the scenario simply by adding/removing the corresponding agent. Moreover, while some multi-agent approaches have already been considered in [28], [29][32], the one considered here has the advantage that an agent learns the policy for assigning capacity to be provided to the slice in each cell, in contrast to [32], which considered the capacity in a single cell, or [28], [29], which provided the aggregated capacity over all the cells.

3. ML-ENABLED CROSS-SLICE MANAGEMENT FRAMEWORK

3.1 O-RAN framework for ML-assisted solutions

As part of the specification of new interfaces and functionality for an open and intelligent RAN, the O-RAN Alliance is working on the definition of a framework for the deployment of ML-assisted solutions within the RAN (i.e. solutions that rely on the use of ML models such as supervised learning, reinforcement learning, etc.) [38].

A representation of the overall RAN functional architecture being defined by O-RAN is illustrated in Fig. 1 [39]. This constitutes a disaggregated RAN, compliant with 3GPP specifications, where the radio protocol stack is split and distributed between different RAN nodes. In particular, the O-RAN Radio Unit (O-RU) hosts the RF processing and the lower part of the PHY layer functionality (e.g. i/FFT

processing), the O-RAN Distributed Unit (O-DU) is in charge of the High-PHY layer processing (e.g. modulation, channel coding), Medium Access Control (MAC) and Radio Link Control (RLC), the O-RAN Central Unit - Control Plane (O-CU-CP) hosts the upper layers of the control plane radio protocol stack, i.e. Radio Resource Control (RRC) and control plane of Packet Data Convergence Protocol (PDCP), and the O-RAN Central Unit - User Plane (O-CU-UP) handles the upper layers of the user plane protocol stack, i.e. Service Data Adaptation Protocol (SDAP) and user plane of PDCP layers. Then, sitting on top of these RAN nodes handling the distributed radio protocol stack, there is the near-real-time RAN Intelligent Controller (near-RT RIC), which serves as the brain of the RAN by coping with the different Radio Resource Management (RRM) functions needed for overall RAN operation, such as radio connection, mobility, Quality of Service (QoS) and interference management. With respect to the interfaces between these RAN nodes, E1, F1-c and F1-u interfaces are specified by 3GPP while Open fronthaul and E2 are being specified by the O-RAN Alliance.

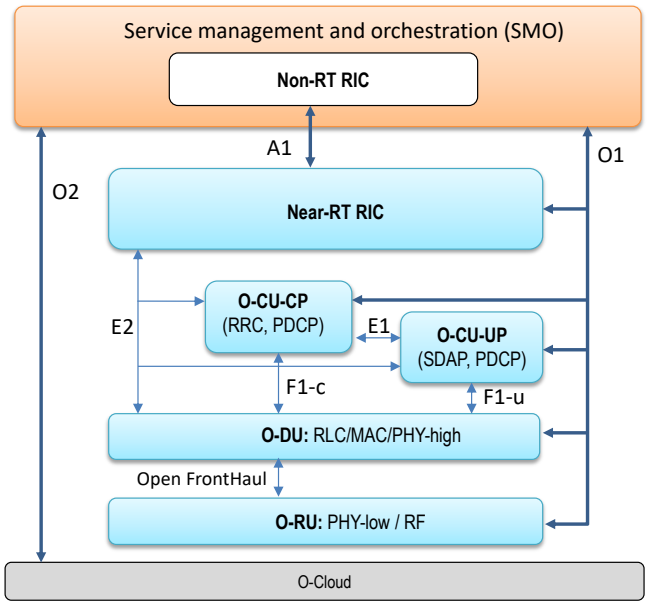


Fig. 1 – O-RAN functional architecture

Moving at the management plane, O-RAN defines the Service Management and Orchestration (SMO) layer, which actually represents the Operations Support Systems (OSS) of the MNO for the RAN domain. As part of the SMO layer, O-RAN basically defines the role of a non-real-time RAN Intelligent Controller (non-RT RIC) entity for the interaction with the near-RT RIC via the A1 interface, which is also being standardized by the O-RAN Alliance. Through the A1 interface [40], the non-RT RIC can perform policy management, ML model management (described below in more detail) and delivery of enriched information for near-RT RIC operation (e.g. RAN data analytics that could be exploited by the near-RT RIC). Furthermore, complementing the A1 interface, the interactions between the SMO and the underlying RAN nodes also rely on the adoption of other standardized interfaces named as O1 and O2 in Fig. 1. In particular, O1 refers to the set of service-based management interfaces being standardized by 3GPP for configuration, performance and fault management of the RAN functionality [41]. In turn, the O2 interface supports the management of the cloud infrastructure and resources allowing the execution of virtualized RAN functions.

Building upon such a RAN reference architecture, Fig. 2 shows the main components and relations being delineated under O-RAN for the training and deployment of ML-assisted solutions within the SMO layer and/or within the RAN nodes themselves.

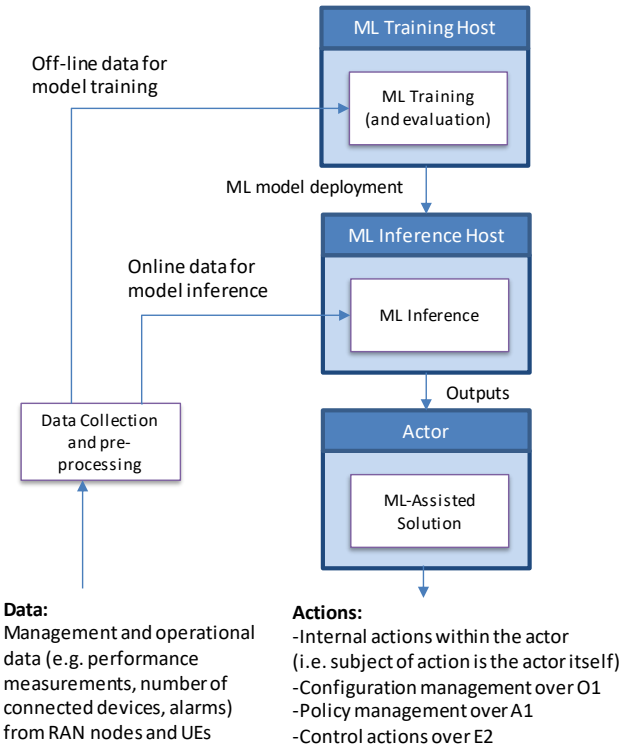


Fig. 2 – Components and relations for ML-assisted solutions within O-RAN

As shown in Fig. 2, a variety of management and operational data is collected from the different RAN nodes and User Equipment (UE) devices. Such data, properly preprocessed, is used to feed the two key components of the ML processing workflow, denoted as the ML training host and the ML inference host. The ML training host represents the

runtime environment where offline training of the ML model takes places. This refers to the training of model before being executed within the network. In addition to the data collected from the real network, offline training may also rely on synthesized data which can accurately reproduce the behavior of the real network environment. The training may include an evaluation stage to assess the performance of the model and validate that it is ready and reliable to be deployed in the live network environment. Offline training is necessary to obtain supervised learning models (e.g. deep neural networks, support vector machines, etc.) as well as reinforcement learning models (e.g. Q-learning, multi-armed bandit learning, deep RL). The training host component is likely to be part of the SMO layer.

The ML inference host represents the runtime environment where the (previously trained and validated) ML model is executed and fed with online data to produce the outputs that will be used in the network operation. Multiple ML inference hosts can be in place, whose location depends on aspects such as the purpose and type of ML models being executed, its computation complexity, the availability and the quantity of data used and the response time requirements (real-time or non-real-time) of the ML application. Hence, ML inference hosts can be placed within the SMO layer but also within the RAN nodes (i.e. near-RT RIC, O-CU, O-DU).

In turn, the actor represents the network entity (i.e. UE, O-DU, O-DU, Near-RT RIC and Non-RT RIC) that hosts the decision-making function that consumes the outputs of the ML inference host and takes actions. It is worth noting that the distinction between the ML inference host and the actor obeys the fact that these components may or may not be co-located as part of the same network entity. An example of non-co-location could be the case of a mobility prediction model executed in an inference host within the non-RT RIC that produces outputs (e.g. mobility patterns) that are retrieved and consumed by the near-RT RIC (i.e. the actor in this case) for enhanced RRM (e.g. handover decisions based on mobility patterns). In contrast, an example of co-location could be an RRM algorithm for mobility management that embeds a reinforcement learning model and is executed within the near-RT RIC, which in this case serves as both the inference host and the actor. The actions decided by the actor can be handled either internally within the actor (e.g. RL-based RRM algorithm for mobility management within the near-RT RIC) or enforced into other network components via the different specified interfaces. For example, management configuration actions from an actor within the SMO layer on any RAN node can be conducted via the O1 interface, control actions on O-CU/O-RU from an actor within the near-RT RIC can go over the E2 interface and policy management configuration actions between the non-RT RIC and the near-RT RIC can be communicated over the A1 interface.

3.2 Information models for network slice management

With regard to the management of network slicing in 5G networks, 3GPP specifications include information model definitions, referred to as Network Resource Models (NRMs), for the characterization of network slices [42] together with a set of management services (MnS) for network slice life-cycle management (e.g. network slice provisioning MnS for network slice creation, modification and termination, performance monitoring services per slice, etc.) [43]. In addition, work is being conducted at 3GPP level to support SLA/SLS management [44], as well as closed loop assurance solutions that allow a service provider to continuously deliver the expected level of communication service quality in a 5G network [45].

Fig. 3 provides an overview of the different types of information models and their relations that are relevant for network slice management. The main idea behind the overall flow of the information models, as illustrated in Fig. 3, is that a network slice is conceived as a "product" offered by a Network Slice Provider (NSP) to a Network Slice Customer (NSC). In this respect, the GSMA Generic Slice Template (GST) is used as the SLA information associated with the network slice product for the communication between the NSC and NSP through, e.g. a Business Support Systems (BSS) product order management Application Programming Interface (API).

The GSMA GST provides a standardized list of attributes (e.g. performance related, function related, etc.) that can be used to characterize different types of network slices [46]. GST is generic and is not tied to any type of network slice or to any agreement between an NSC and an NSP. A Network Slice Type (NEST) is a GST filled with (ranges of) values. There are two kinds of NESTs: Standardized NESTs (S-NEST), i.e. NESTs with values established by standards organizations, working groups, fora, etc. such as, e.g. 3GPP, GSMA, 5GAA, 5G-ACIA, etc.; and Private NESTs (P-NEST), i.e. NESTs with values

decided by the NSP. Among the attributes included in the GST there are:

- Attributes that specify the area where the terminals can access a particular network slice and the spectrum used.
- Attributes that define the services and capabilities supported in the slice (e.g. "support for non-IP traffic", "MMTel support", "NB-IoT support", "Positioning support", etc.)
- Attributes that establish the capacity and performance of the slice, including guaranteed and maximum data rates per slice and per UE, as well as maximum number of concurrent sessions and terminals in the slice.
- Attributes that define the terminal mobility conditions and density of terminals.
- Attributes that define management and operational aspects and features of the slice (e.g. performance monitoring indicators, performance prediction indicators, user management openness).
- Attributes that define the isolation level of the slice with regard to other slices (e.g. physical isolation, logical isolation).

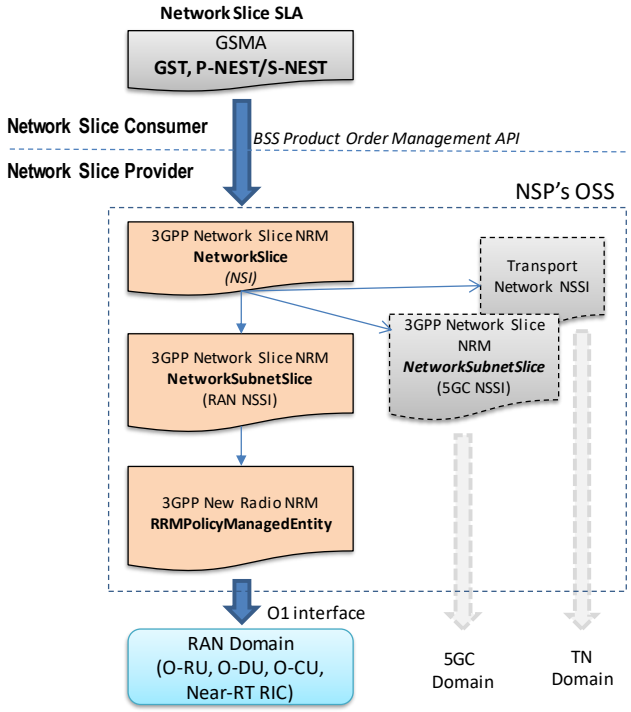


Fig. 3 – Information models for network slice management

Internally, within an NSP's OSS, the managerial representation of the network slice is realized with two Information Object Classes (IOC), named *NetworkSlice* and *NetworkSliceSubnet*, specified in the 3GPP information model definitions for "network slice NRM" [42]. The *NetworkSlice* IOC and the *NetworkSliceSubnet* IOC represent, respectively, the properties of a Network Slice Instance (NSI) and a Network Slice Subnet Instance (NSSI) in a 5G network. It is worth clarifying at this point that the realization of an NSI may be tied to the realization of several NSSIs, that is, an NSI that is composed of, e.g. a RAN NSSI, a 5G Core (5GC) NSSI and a Transport Network NSSI. However, depending on the NSP's product offering, it is also possible the realization of an NSI composed of a single domain such as, an NSI consisting of a single, RAN-only NSSI. Within the *NetworkSlice* IOC and *NetworkSubnetSlice* IOC, the attributes that are defined to encode the network slice related requirements that should be supported by the NSI and the NSSI are named, respectively, *ServiceProfile* and *SliceProfile*. Such attributes are compound data types that include attributes directly inherited from the GSMA GST template, as well as additional attributes to capture more specific requirements derived from the service performance requirements defined in [47], [48] . In particular, let us introduce here three attributes included in the *ServiceProfile* that are are directly inherited from the GSMA GST and used in the algorithm presented in Section 4:

- *dlThptPerSlice*: It defines the achievable aggregate downlink data rate of the network slice.
- *dlThptPerUe*: It defines the average data rate delivered by the network slice per UE.
- *termDensity*: It specifies the maximum user density over the coverage area of the network slice.

And last but not least, as also captured in Fig. 3, 3GPP also provides information model definitions for "New Radio NRM" [42], where different RAN management parameters are defined to configure the behaviour of the RAN nodes with regard to the operation of the established network slices. In this respect, a more detailed, though still simplified, view of the classes and attributes of the "New Radio NRM" model that allow for the characterization of the RRM policies for configuring the way that resources are allocated to the slices is provided in Fig. 4. Specifically, the *RRMPolicyManagedEntity* proxy class represents the different RAN managed components (e.g. cell resources managed at DU, cell resources managed at CU functions, DU functions,

etc.) that are subject to the RRM policies and the *RRMPolicy_* IOC represents the properties of an abstract RRMPolicy that defines two attributes: the *resourceType* attribute, used to define the type of resource (e.g. PRB, RRC connected users, etc.) and the *rRMPolicyMemberList* attribute, used to indicate the associated network slice or group of network slices that is subject to this policy. The associated network slices are specified here in terms of slice identifiers such as the Single Network Slice Assistance Information (S-NSSAI) and Public Land Mobile Network Identifier (PLMNid).

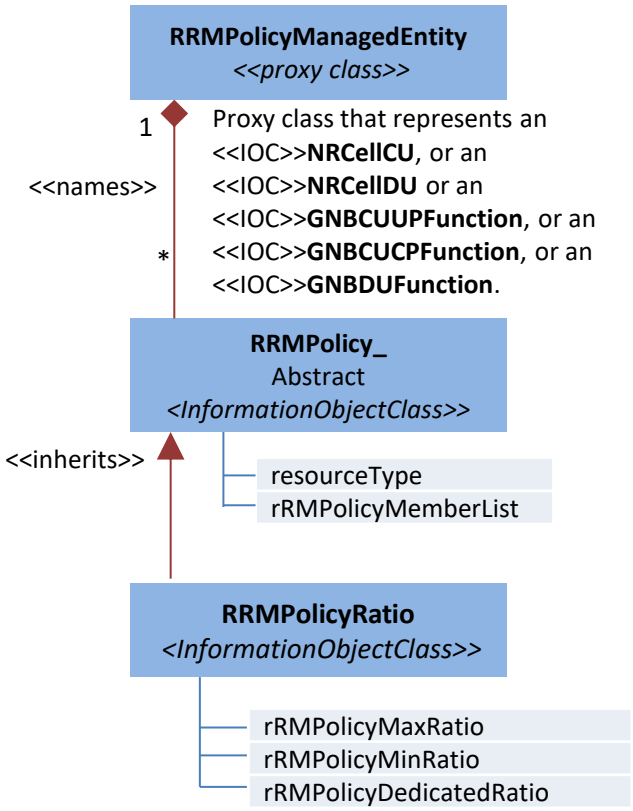


Fig. 4 – Simplified representation of the classes and attributes for configuration of RAN functions subject to RRM policies for slice management.

On this basis, the *RRMPolicyRatio* IOC represents a particular realization of an RRMPolicy. Specifically, it establishes a resource model for resource distribution among slices based on three resource categories: shared resources (shared among slices with no specific guarantees per slice), prioritized resources (guaranteed for use by associated slices but still usable for other slices when free), and dedicated resources (only used for the associated slices). Accordingly, the following attributes are included in the *RRMPolicyRatio* IOC:

- *rRMPolicyDedicatedRatio*, defines the dedicated resource usage quota for the associated network slice(s), including dedicated resources. The sum of the *rRMPolicyDedicatedRatio* values shall be less or equal than 100.
- *rRMPolicyMinRatio*, defines the minimum resource usage quota for the associated network slice(s), including prioritized resources and dedicated resources, which means the resources quota that need to be guaranteed for use. The sum of the 'rRMPolicyMinRatio' values shall be less or equal than 100.
- *rRMPolicyMaxRatio*, defines the maximum resource usage quota for the associated network slice(s), including shared resources, prioritized resources and dedicated resources. The sum of the 'rRMPolicyMaxRatio' values can be greater than 100.

4. ML-ASSISTED CROSS-SLICE OPTIMIZATION SOLUTION

This section describes an ML-assisted solution for cross-slice optimization based on the O-RAN framework and network slicing information models presented in the previous section. The solution is conceived to be deployed as part of the RAN SMO. The functional model and components of the solution are illustrated in Fig. 5. The cross-slice radio resource optimization problem considered here consists of dynamically adjusting the amount of downlink radio resources assigned to each RAN slice in each of the cells where the RAN slice is accessible in order to account for the spatio-temporal traffic variations across the cells. The solution is designed to operate with N cells and K RAN slices and keep track of the traffic variations in periods (time steps) of Δt minutes. This is achieved through the dynamic configuration of the *rRMPolicyDedicatedRatio* attribute of each cell on a per RAN slice basis. This configuration is conducted via the management provisioning services offered by the O1 interfaces, as seen in Fig. 5. In particular, since the *rRMPolicyDedicatedRatio* attribute establishes the resource usage quota assigned to the RAN slice defined in terms of the fraction of Physical Resource Blocks (PRBs) that can be used by this slice, the attribute is configured in the O-DU unit, so that the MAC layer can take this resource usage quota into account when allocating PRBs to the users of the RAN slice in the cell.

For determining the *rRMPolicyDedicatedRatio* for the n-th cell and the k-th slice, denoted as $\sigma(k,n)$, the

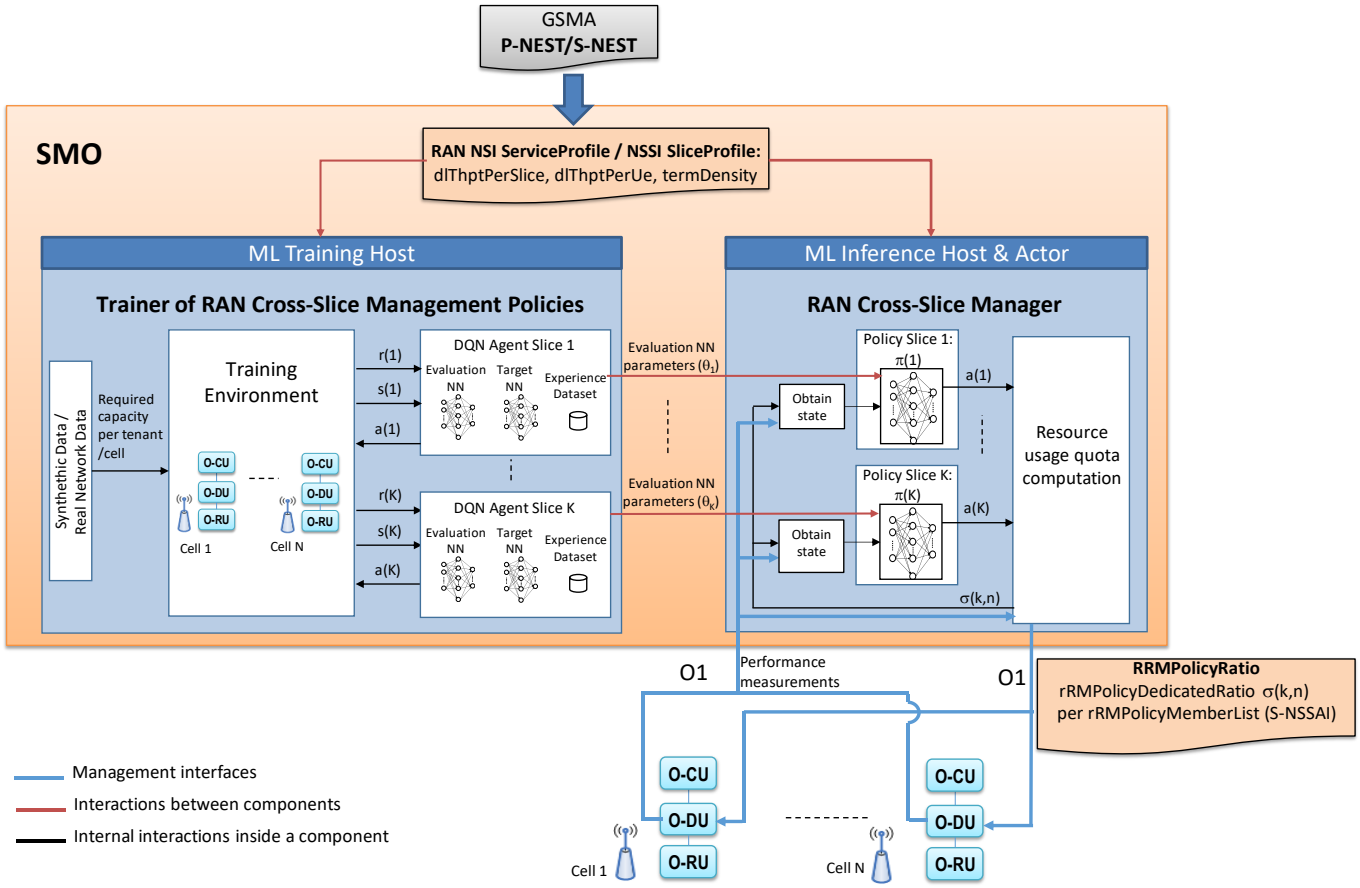


Fig. 5 – Deep Q Network-based cross-slice optimization solution

proposed ML-based solution relies on Multi-Agent Reinforcement Learning (MARL) based on Deep Q-Network (DQN) whose mathematical details can be found in [33]. An important advantage of this multi-agent scheme is that it uses slice-specific DQN agents and action selection policies for the training and inference processes and, therefore, slices can be easily added/removed in the scenario just by adding or removing the corresponding agent and action selection policy. Specifically, as seen in Fig. 5, as part of the solution there is a Resource Usage Quota Computation module that determines the values of $\sigma(k,n)$ based on the outputs obtained through the execution of K action selection policies $\pi(k)$, each one associated to one slice. Each one of these policies is specified through a deep neural network (NN) defined by a vector of parameters θ_k that have been previously learnt during the training process, as illustrated in Fig. 5.

The solution considered here assumes that the SLA specification of the RAN slice requirements is done based on three ServiceProfile parameters explained in Section 3, namely *dlThptPerSlice*, *dlThptPerUe* and *termDensity*, which are directly derived from the GSMA GST template and used as inputs for the different solution components described in more detail in the following subsections. The specific values of these parameters for the slice k are denoted as *dlThptPerSlice*(k), *dlThptPerUe*(k) and *termDensity*(k) [1].

4.1 RAN cross-slice manager

This component includes the inference part of the DQN model and the functions needed to configure the RAN nodes through the O1 interface (i.e. it takes the ML inference and actor roles of the O-RAN

[1] For the interested reader, the parameters of the algorithm described in [33] are related to the considered Service Profile attributes as follows: Scenario Aggregated Guaranteed Bit Rate (SAGBR)=*dlThptPerSlice*; Maximum Cell Bit Rate (MCBR) of slice k in cell n: $MCBR(k,n)=dlThptPerUe(k) \times termDensity(k) \times$ cell n service area.

framework). The computation of $\sigma(k,n)$ and potential update of the *rRMPolicyDedicatedRatio* attribute is done every Δt minutes. The determination of the resource usage quota $\sigma(k,n)$ is realized through the following functions:

- *Per-slice action selection policies*

The action selection policy of slice k gets the network state $\boldsymbol{s}(k)$ observed for this slice at the time when the policy is executed and determines the action $\boldsymbol{a}(k)$ to be applied for this slice. The action $\boldsymbol{a}(k)$ is composed of N per-cell actions that take one out of three possible values corresponding to: increase the resource usage quota $\sigma(k,n)$ for slice k in cell n in an amount of Δ for the next time step, maintain the same resource usage quota or decrease it in an amount of Δ.

In turn, the state $\boldsymbol{s}(k)$ includes N different per-cell components, each one given by the triple $<\rho(k,n),\sigma(k,n), \sigma_{ava}(n)>$ where $\rho(k,n)$ is the fraction of PRBs occupied by the slice k in cell n, $\sigma(k,n)$ is the current resource usage quota allocated to the slice and $\sigma_{ava}(n)$ is the total amount of resource usage quota in the cell not allocated to any slice. While the values of $\sigma(k,n)$ and $\sigma_{ava}(n)$ are directly available at the RAN cross-slice manager, the value of $\rho(k,n)$ is obtained from the different cells through the performance management (PM) services offered by the O1 interface. In particular, using the performance measurements defined in [49], it corresponds to the ratio between the "DL PRB used for data traffic", which measures the number of PRBs used in average for data traffic in a given slice and cell, and the "DL total available PRB", which measures the number of available PRBs in the cell. Both measurements are collected from the gNB-DU every time step, so their average is performed along the time step duration Δt.

Following the DQN approach, the action selection policy $\pi(k)$ of the k-th slice seeks to maximize a cumulative reward that captures the desired optimization target to be achieved. In particular, the action selection policy $\pi(k)$ for a given state $\boldsymbol{s}(k)$ is defined as $\text{argmax}_{\boldsymbol{a(k)}} Q_k(\boldsymbol{s}(k),\boldsymbol{a}(k),\theta_k)$, where $Q_k(\boldsymbol{s}(k),\boldsymbol{a}(k),\theta_k)$ is the output of a deep NN for the input state $\boldsymbol{s}(k)$ and the output action $\boldsymbol{a}(k)$, providing the maximum expected cumulative reward starting at $\boldsymbol{s}(k)$ and triggering $\boldsymbol{a}(k)$. The internal structure of the NN is specified by the vector of parameters θ_k that contains the weights of the different neuron connections. The optimum values of θ_k that determine the policies to be followed by the different slices in order to maximize the cumulative reward are learnt offline by the ML training host who provides them to the ML inference host. Further details about this training process and the reward formulation are given in Section 4.2.

- *Resource usage quota computation*

This function computes the value of the resource usage quota $\sigma(k,n)$ to be allocated to each slice and cell for the next time step by applying the increase/maintain/decrease actions provided by the action selection policies of all the slices and configures the resulting $\sigma(k,n)$ values in the O-DU through the *rRMPolicyDedicatedRatio* attribute. To make the configuration on a per-slice basis, an *rRMPolicyMemberList* is specified for each RAN slice, being composed of a single member with the S-NSSAI and PLMNid of the RAN slice. Then, the *rRMPolicyDedicatedRatio* is configured per *rRMPolicyMemberList* in each cell.

When applying the actions, this function ensures that the maximum cell bit rate value associated to the *termDensity* and *dlThptPerUe* parameters is not exceeded. Moreover, since the action selection policies for the different slices operate independently, this function also checks that the aggregated resource usage quota for all the slices in a cell after applying the actions does not exceed 1 in order not to exceed the cell capacity. If this happens, it applies first the actions of the slices involving a reduction or maintenance of the resource usage quota and the remaining capacity is distributed among the slices that have increase actions. This distribution is proportional to their *dlThptPerSlice* values, as long as their current throughput is not already higher than the *dlThptPerSlice*. For doing this adjustment, the measured throughput per slice across all the cells in the last time step is needed. It can be obtained from the PM services of the O1 interface using the "Downstream throughput for Single Network Slice Instance" Key Performance Indicator of [50].

4.2 Trainer of RAN cross-slice management policies

This component constitutes the training part of the DQN model intended to learn the NN parameters θ_k that determine the per-slice action selection policies to be used by the RAN cross-slice manager.

The training process makes use of a multi-agent DQN approach in which each DQN agent learns the optimum policy of a different RAN slice by

continuously interacting with a training environment and updating the NN parameters as a result of these interactions. The training environment considered here is a network simulator that mimics the behavior of the real network when varying the offered load of the different slices in the different cells and when modifying the resource usage quota allocated to each slice as a result of the actions made by the DQN agents. In this respect, the simulator is fed by training data consisting of multiple time patterns of the required capacity (i.e. offered load) of the slices in the different cells. This data can be either built synthetically or extracted from real network measurements. The training is assumed to be executed in a training host, located at the SMO, with the necessary libraries, supporting tools and computational capabilities for training the DQN models and running the simulator.

For carrying out the training process, each DQN agent is composed by three different elements: (i) The evaluation NN, which corresponds to the $Q_k(\boldsymbol{s}(k),\boldsymbol{a}(k),\theta_k)$ being learnt that will eventually determine the policy to be applied at the ML inference host. (ii) The target NN, which is another NN with the same structure as the evaluation NN but with weights θ_{k^-}. It is used for obtaining the so-called Time Difference (TD) target required for updating the evaluation NN. (iii) The experience data set (ED), which stores the experiences of the agent resulting from the interactions with the training environment as explained in the following.

The interactions between the DQN agent and the training environment occur in time steps of (simulated time) duration Δt. In each time step the DQN agent of the k-th slice observes the state $\boldsymbol{s}(k)$ in the training environment and selects an action $\boldsymbol{a}(k)$. Action selection is based on an ε-Greedy policy that, with probability 1-ε, chooses the action that maximizes the output of the evaluation NN, and, with probability ε, chooses a random action. As a result of applying the selected action, the training environment generates a reward value $r(k)$ that assesses how good the action was from the perspective of the desired behavior. In particular, in the considered approach the reward captures both the SLA satisfaction and the capacity utilization. In this way, the reward for slice k is defined as the weighted product of three terms given by:

$$r(k) = \gamma_{SLA}(k)^{\varphi_1} \cdot \left(\frac{1}{K-1}\sum_{\substack{k'=1 \\ k'\neq k}}^{K} \gamma_{SLA}(k')\right)^{\varphi_2} \cdot \gamma_u(k)^{\varphi_3} \quad (1)$$

where φ_1, φ_2 and φ_3 are the weights of each component.

The first and second components in (1) correspond, respectively, to the SLA satisfaction ratio $\gamma_{SLA}(k)$ of the slice k and the aggregate for the rest of slices $k'\neq k$. Specifically, $\gamma_{SLA}(k)$ is the ratio between the aggregate throughput obtained by the slice across all cells $T(k)$ and the minimum between the aggregate offered load $A(k)$ and the $dlThptPerSlice(k)$ term of the SLA and is computed as:

$$\gamma_{SLA}(k) = min\left(\frac{T(k)}{min\left(dlThptPerSlice(k),A(k)\right)}, 1\right) \quad (2)$$

where $A(k)$ is the aggregate across all the cells of the per-cell offered load $O(k,n)$ of slice k bounded by the limit established by the $TermDensity(k)$ and $dlThptPerUe(k)$ parameters of the SLA in the service area $S(n)$ of each cell n, that is:

$$A(k) = \sum_{n=1}^{N} min\big(O(k,n), dlThptPerUe(k) \cdot TermDensity(k) \cdot S(n)\big) \quad (3)$$

The third component of the reward is the capacity utilization factor, $\gamma_u(k)$, which aims at minimizing the over-provisioning of capacity and is defined as the ratio between the aggregate throughput $T(k)$ obtained by the slice and the total capacity allocated to the slice across all cells, that is:

$$\gamma_u(k) = \frac{T(k)}{\sum_{n=1}^{N} C(n)\cdot\sigma(k,n)} \quad (4)$$

where $C(n)$ is the capacity of cell n.

The reward $r(k)$ is provided by the training environment to the DQN agent at the end of each time step and, correspondingly, the $T(k)$ and $A(k)$ values correspond to average values during the time step.

As a result of the interactions between the training environment and the DQN agent, each experience of the ED is represented by a tuple that includes the state observed at the beginning of a given time step, the selected action, the obtained reward as a result of this action and the new state observed at the end of the time step duration.

The experiences stored in the ED are used by the DQN agent to progressively update the values of the weights θ_k and θ_{k^-} in the evaluation and target NNs, respectively. For each time step, the update of the weights θ_k of the evaluation NN is performed by randomly selecting a mini batch of experiences of the ED and updating the weights of the evaluation NN θ_k according to the mini-batch gradient descent

method. Moreover, the weights θ_k^- of the target NN are updated with the weights of the evaluation NN every M time steps. The reader is referred to [33] for details on the mathematical formulation of this process.

The training process stops after a sufficient number of time steps that ensures the convergence of the process. At this point, the ML training host is ready to provide the evaluation NN parameters θ_k so that the model can be applied on the real network using the ML inference host.

5. ILLUSTRATIVE SCENARIO AND EVALUATION

To illustrate the behavior the proposed cross-slice optimization framework and ML-assisted solution, let us consider a scenario with a localized RAN deployment run by an infrastructure provider, serving as an NSP, which offers a RAN slice product to a pair of MNOs, which in this case act as NSCs. This could be the case of a dense urban deployment of small cells in a business district operated under a neutral host model. Let us assume that the MNOs use the RAN slices to offer enhanced Mobile BroadBand (eMBB) services to their customers so that key parameters to include in the SLA are the number of UEs expected to be served in the area, given in terms of the maximum terminal density, the throughput guaranteed in the whole service area per slice and the expected UE experienced data rates. These SLA parameters are summarized in Table 1. On the other hand, let us assume a RAN deployment consisting of 5 small cells, which provide an aggregated capacity of 10 Gb/s in an area of 0.15 km^2. The characteristics of this deployment are captured in Table 2. Under such settings, note that the *dlThptPerSlice* values of MNO1 and MNO2 SLAs actually account for the 60% and 40% of the total capacity, respectively.

Two different cases of offered load patterns of the MNOs throughout the day are considered for evaluating the performance of the learnt policies. Case 1 shown in Fig. 6 corresponds to a situation in which the offered loads of the two MNOs exhibit a certain complementarity during the time period comprised between 900 and 1300 min, approximately, in which MNO2 exhibits a large load while the load of MNO1 is kept at low values. Instead, Case 2 shown in Fig. 7 reflects a situation in which the offered load of the two MNOs is more correlated. In both Fig. 6 and Fig. 7 the offered load corresponds to a period of one day and is represented as the average in intervals of 15 min. It is worth mentioning that the focus of the results is put on the temporal variations of the offered load so, from the spatial perspective, it is assumed, for simplicity, that the aggregate offered load is homogeneously distributed across the cells.

Table 1 – SLA parameters

GSM GST Attributes	MNO1	MNO2
dlThptPerSlice	6 Gb/s	4 Gb/s
termDensity	1000 UEs/km^2	500 UEs/km^2
dlThptPerUe	50 Mb/s	100 Mb/s

Table 2 – Cell configuration

Parameter	Value
Number of cells	5
Cell radius	100m
Cell bandwidth	100 MHz (273 PRBs with 30 kHz subcarrier spacing)
Average spectral efficiency	5.1 b/s/Hz
MIMO configuration	Spatial multiplexing with 4 layers
Total cell capacity	2 Gb/s

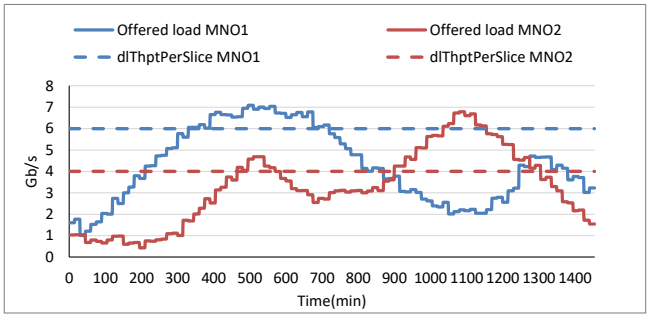


Fig. 6 – Offered load pattern of each MNO in Case 1.

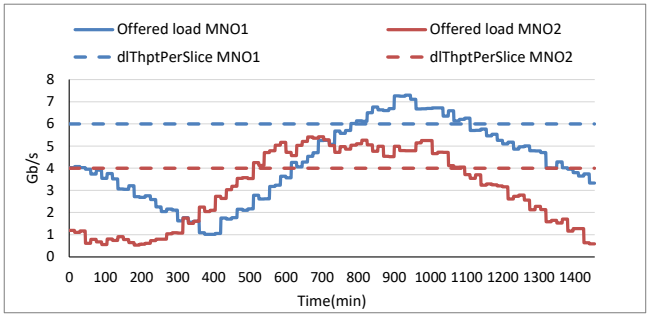


Fig. 7 – Offered load pattern of each MNO in Case 2.

The DQN-based cross-slicing solution has been implemented in Python by using the library *TF-Agents* [51]. Table 3 shows the parameters of the DQN model (see [33] for details on these parameters). To obtain the values of these parameters a prior analysis of the model behavior with different combinations of parameters has been conducted. The model has been trained using a data set composed of 140 synthetically generated

offered load patterns of the two MNOs in the different cells during one day. They capture different load levels and situations of complementarity among MNOs, in order that the DQN agents can visit multiple states during the training process.

Table 3 – DQN model parameters

Parameter	Value
Initial collect steps	5000
Number of training steps	10^6
Experience Data set maximum length	10^7
Mini-batch size	256
Learning rate	0.0001
Time steps between updates of the target NN weights (M)	1
Discount factor	0.9
ε value (ε-Greedy)	0.1
Neural network nodes	2 layers of 100 nodes
Resource quota increase (Δ)	0.1
Time step duration (Δt)	1 min
Reward weights ($\varphi_1, \varphi_2, \varphi_3$)	(0.3, 0.2, 0.5)

The training has been conducted with a system level network simulator that considers the offered load patterns of the different slices and cells as input. In every time step the DQN agents select the actions that determine the *rRMPolicyDedicatedRatio* assigned to each slice in each cell. Then, the number of PRBs that are utilized by the slice is the minimum between the assigned PRBs in accordance with *rRMPolicyDedicatedRatio* and the required PRBs, which are determined by the offered load and the spectral efficiency. Then, the throughput achieved by each slice is obtained using the number of utilized PRBs and the spectral efficiency. From this, the SLA satisfaction ratio from (2), the capacity utilization from (4) and the reward from (1) are computed. The reward, together with the selected action and the actual and previous states are stored in the experience data set and the weights of the evaluation and target NNs are updated. This process is repeated until reaching the number of training steps indicated in Table 3. At the end, the resulting weights of the evaluation NN determine the trained policy to be used during the ML inference stage.

Once the training has been completed, the ML inference stage assesses the obtained policy using the same system level network simulator of the training, but now taking as input the offered load patterns of Fig. 6 and Fig. 7 split equally among the different cells. The trained policy is executed every time step to obtain the *rRMPolicyDedicatedRatio* values, from which the SLA satisfaction ratio and capacity utilization metrics are determined.

To illustrate the operation of the considered cross-slicing solution, Fig. 8 and Fig. 9 plot the evolution of the *rRMPolicyDedicatedRatio* parameter in % configured by the algorithm for each slice in one of the cells for Cases 1 and 2, respectively. As a reference, the evolution of the offered load pattern of each MNO, measured in % of the total scenario capacity is also shown in the plots.

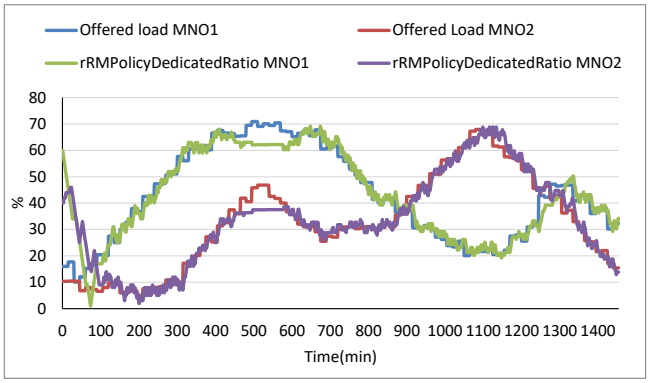


Fig. 8 – Evolution of the *rRMPolicyDedicatedRatio* for each MNO in one cell for Case 1.

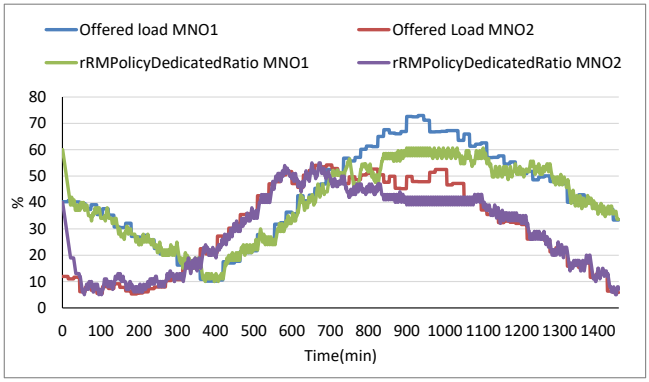


Fig. 9 – Evolution of the *rRMPolicyDedicatedRatio* for each MNO in one cell for Case 2.

Focusing on Fig. 8, it can be observed that in general the algorithm is able to modify the amount of resources assigned to each slice through the *rRMPolicyDedicatedRatio* parameter following the offered load fluctuations, so that the algorithm provides each slice with the resources it needs to support its load. Going into further details, different situations can be identified during the time evolution of Fig. 8.

Initially at time t=0 min the *rRMPolicyDedicatedRatio* is set to 60% and 40% for slice 1 and slice 2. These values correspond to the fractions of resources associated to the *dlThptPerSlice* values established in the SLA. Then, as time increases, an initial transient period of

around 100 min is observed in which the *rRMPolicyDedicatedRatio* is progressively adjusted to fit the actual resource demand. After this period, the *rRMPolicyDedicatedRatio* approximately follows the offered load of each MNO in the cell, as long as the aggregate load for all cells is below the *dlThptPerSlice* values.

A particular situation occurs between *t*=350 min and *t*=450 min, and between *t*=600 and 700 min, approximately. In these periods the offered load of MNO1 exceeds its *dlThptPerSlice* value of the SLA (60%). At the same time, the offered load of MNO2 during this period is still below its corresponding *dlThptPerSlice* value (40%) and the cell has enough resources to satisfy the demands of both MNOs. Therefore, for the sake of better resource utilisation, the algorithm allows the *rRMPolicyDedicatedRatio* of the slice of MNO1 to exceed the 60%, providing in this way all the required capacity. An equivalent situation occurs between time *t*=900 min and *t*=1300 min, but now it is the offered load of MNO2 that clearly exceeds its *dlThptPerSlice* while the load of MNO1 is substantially lower than its *dlThptPerSlice*. Again, since the cell has enough capacity to satisfy the demands of both MNOs, the *rRMPolicyDedicatedRatio* of the slice of MNO2 is increased beyond the *dlThptPerSlice* value of 40% to support all the load of this MNO.

In turn, the period between *t*=450 min and *t*=600 min corresponds to the case in which both MNOs are demanding capacity beyond their *dlThptPerSlice* values and the cell does not have sufficient resources to support all the demand (i.e. the sum of the offered loads of MNO1 and MNO2 exceeds 100%). For this reason, it is observed in Fig. 8 that the algorithm sets the *rRMPolicyDedicatedRatio* in accordance with the *dlThptPerSlice* values, i.e. 60% for MNO1 and 40% for MNO2.

Focusing now on Case 2, which reflects a larger correlation between the offered load of both MNOs, the results of Fig. 9 reveal a similar behavior than the one discussed in Fig. 8, and the algorithm is able to allocate to each slice the necessary amount of resources to support their offered load. This occurs even when the offered load of a slice is above the *dlThptPerSlice* limit as long as there are sufficient resources in the cell. In turn, when the load of both MNOs exceeds the *dlThptPerSlice* value and there are not enough resources in the cell (e.g. between *t*=850 and *t*=1150 min in Fig. 9) the algorithm sets the *rRMPolicyDedicatedRatio* in accordance with the *dlThptPerSlice* values.

Table 4 presents some indicators to quantitatively assess the performance of the cross-slicing approach. Specifically, the average SLA satisfaction obtained for each MNO is presented. This is measured as the time average of equation (2) over the whole simulation time of 1440 min (excluding the initial transient period of 100 min) and provides the percentage of time in which the *rRMPolicyDedicatedRatio* has allocated enough resources to support the offered load while this load was below the *dlThptPerSlice* value established in the SLA. It is observed that the algorithm achieves high SLA satisfaction, above 97% in all the cases. To further quantify the variability of the SLA satisfaction, the 5th percentile of this indicator is also shown in the table. The large obtained values around 91% reflect that the achieved SLA satisfaction is very good most of the time.

Similarly, to account for the degree of resource over-provisioning when allocating resources to each slice (i.e. for the situations in which the *rRMPolicyDedicatedRatio* allocated to a slice includes more resources than actually required by the MNO), Table 4 also shows the average capacity utilization of the assigned resources. This is measured as the time average of the ratio between the throughput of a slice and the amount of allocated capacity to this slice from equation (4). The average is measured along the whole simulation time of 1440 min excluding the initial transient period of 100 min. The corresponding 5th percentile is also indicated in Table 4. It is observed that the algorithm achieves high utilization, being the average above 94% and the 5th percentile approximately above 80% for both slices. This reflects that the algorithm is able to properly adjust the resource allocation to the actual needs and thus to reduce over-provisioning situations.

Table 4 – Performance indicators

Performance indicator	Case 1		Case 2	
	MNO1	MNO2	MNO1	MNO2
SLA satisfaction (average)	98.31%	97.69%	97.38%	97.14%
SLA satisfaction (perc. 5)	92.27%	90.97%	91.95%	91.75%
Capacity utilization (average)	96.44%	94.06%	96.43%	94.61%
Capacity utilization (perc. 5)	90.18%	81.79%	90.33%	79.28%

6. CONCLUDING REMARKS

ML-assisted solutions with the ability to learn particular network dynamics and operate under goal-oriented policies arise as a feasible approach to tackling the complexity of the cross-slice radio resource allocation problem. Beyond the algorithmic dimension, bringing into the equation the information models and the architectural context for the implementation of these ML-assisted solutions is also key to further progress towards their practical realization.

In this respect, this paper has outlined the building blocks of the architectural framework being established under the O-RAN Alliance for the deployment of ML-assisted solutions in the RAN along with the different types of information models developed by 3GPP for network slice management, from service characterization at the SLA level down to the specific management attributes that can be used to configure how radio resources are allocated to the slices within the RAN nodes.

Building upon this architectural framework and associated information models, the paper has presented a plausible realization of an ML-assisted cross-slice radio resource optimization solution based on the use of multi-agent DQN techniques. The presented solution is shown to be a feasible approach to dynamically adjust the resource allocation of the slices to the traffic variations in order to fulfill an SLA and achieve high resource utilization efficiency.

ACKNOWLEDGEMENT

This work has been supported by the Spanish Research Council and FEDER funds under SONAR 5G grant (ref. TEC2017-82651-R) and by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia under grant 2019FI_B1 00102.

REFERENCES

[1] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M.A. Puente, K. Samdanis, B. Sayadi, “Mobile network architecture evolution toward 5G”, *IEEE Communications Magazine*, Vol. 54, No. 5, May, 2016, pp. 84-91.

[2] 3GPP TS 23.501 v16.6.0 “System Architecture for the 5G System; Stage 2 (Release 16)”, September, 2020.

[3] RP-201612, “Study on enhancement of RAN Slicing for NR”, 3GPP TSG RAN Meeting #89e, September, 2020.

[4] SP-190931, “Feasibility Study on Enhancement of Network Slicing Phase 2”, 3GPP TSG-SA Meeting #85, September, 2019.

[5] SP-200766, “Study on network slice management enhancement”, 3GPP TSG-SA Meeting #89e, September, 2020.

[6] SP-200571, “Feasibility Study on Enhanced Access to and Support of Network Slice”, TSG SA Meeting #88e, July, 2020.

[7] R. Ferrús, O. Sallent, J. Pérez-Romero, R. Agustí, “On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration”, *IEEE Communications Magazine*, Vol. 56, No. 5, May, 2018, pp.184-192.

[8] K. Samdanis, X. Costa-Perez and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Communications Magazine*, vol. 54, no. 7, July, 2016, pp. 32-39.

[9] O. Sallent, J. Pérez-Romero, R. Ferrús, R. Agustí, “On Radio Access Network Slicing from a Radio Resource Management Perspective”, *IEEE Wireless Communications*, Vol. 24. No.5, October, 2017,pp. 166-174.

[10] A. S. D. Alfoudi, S. H. S. Newaz, A. Otebolaku, G. M. Lee and R. Pereira, "An Efficient Resource Management Mechanism for Network Slicing in a LTE Network”, *IEEE Access*, Vol. 7, July, 2019, pp. 89441-89457.

[11] P. L. Vo, M. N. H. Nguyen, T. A. Le and N. H. Tran, "Slicing the Edge: Resource Allocation for RAN Network Slicing," in *IEEE Wireless Communications Letters*, Vol. 7, No. 6, December, 2018, pp. 970-973.

[12] I. Vilà, J. Perez-Romero, O. Sallent, A. Umbert, “Characterization of Radio Access Network Slicing Scenarios with 5G QoS Provisioning”, *IEEE Access*, Vol. 6, March, 2020, pp. 51414-41430.

[13] D. Marabissi, R. Fantacci, "Highly Flexible RAN Slicing Approach to Manage Isolation, Priority, Efficiency”, *IEEE Access*, Vol. 7, July, 2019, pp. 97130-97142.

[14] R. Ferrus, O. Sallent, J. Perez-Romero, R. Agustí, "On the automation of RAN slicing provisioning: solution framework and applicability examples", *EURASIP Journal on Wireless Communications and Networking*, June, 2019

[15] 3GPP TR 28.861 v16.0.0 "Study on the Self-Organizing Networks (SON) for 5G networks (Release 16)", December, 2019.

[16] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, J. Rao, "AI-Assisted Network-Slicing Based Next-Generation Wireless Networks", *IEEE Open Journal of Vehicular Technology*, Vol. 1, January, 2020, pp. 45-66.

[17] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, X. Costa-Perez, "Optimising 5G infrastructure markets: The Business of Network Slicing", *IEEE INFOCOM,* 2017.

[18] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, X. Costa-Perez, "A Machine Learning approach to 5G Infrastructure Market optimization", *IEEE Transactions on Mobile Computing,* Vol. 19, No. 3, March, 2020, pp. 498-512.

[19] N. V. Huynh, D. T. Hoang, D. N. Nguyen, E. Dutkiewicz, "Optimal and Fast Real-Time Resource Slicing with Deep Dueling Neural Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 37, No. 6, June, 2019, pp. 1455-1470.

[20] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, A. Banchs, "Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization", *IEEE INFOCOM*, 2017.

[21] V. Sciancalepore, X. Costa-Perez, A. Banchs, "RL-NSB: Reinforcement Learning-Based 5G Network Slice Broker", *IEEE/ACM Transactions on Networking*, Vol. 27, No.4, August, 2019, pp. 1543-1557.

[22] M. Toscano, F. Grunwald, M. Richart, J. Baliosian, E. Grampin, A. Castro, "Machine Learning Aided Network Slicing", *21st Int. Conf. on Transparent Optical Networks (ICTON)*, 2019.

[23] A. Aijaz, "Hap-SliceR: A Radio Resource Slicing Framework for 5G Networks with Haptic Communications", *IEEE Systems Journal*, Vol. 12, No. 3, September, 2018, pp. 2285-2296.

[24] B. Han, A. De Domenico, G. Dandachi, A. Drosou, D. Tzovaras, R. Querio, F. Moggio, O. Bulakci, H.D. Schotten, "Admission and Congestion Control for 5G Network Slicing", *IEEE Conf. on Standards for Communications and Networking (CSCN)*, 2018.

[25] X. Foukas, M.K.Marina, K. Kontovasilis, "Iris: Deep Reinforcement Learning Driven Shared Spectrum Access Architecture for Neutral-Host Small Cells", *IEEE Journal on Selected Areas in Communications*, Vol. 37, No. 8, August, 2019, pp. 1820-1837.

[26] S. D'Oro, L. Galluccio, P. Mertikopoulos, G. Morabito, S. Palazzo, "Auction-based resource allocation in OpenFlow multi-tenant networks", *Computer Networks*, Vol. 115, March, 2017, pp. 29-41.

[27] X. Chen, Z. Zhao, C. Wu, M. Bennis, H. Liu, Y. Ji, H. Zhang, "Multi-Tenant Cross-Slice Resource Orchestration: A Deep Reinforcement Learning Approach", *IEEE Journal on Selected Areas in Communications*, Vol. 37, No. 10, October, 2019, pp. 2377-2392.

[28] G. Sun, Z. T. Gebrekidan, G. O. Boateng, D. Ayepah-Mensah, W. Jiang, "Dynamic Reservation and Deep Reinforcement Learning Based Autonomous Resource Slicing for Virtualized Radio Access Networks", *IEEE Access*, Vol. 7, April, 2019, pp. 45758-45772.

[29] G. Sun, K. Xiong, G.O. Boateng, D. Ayepah-Mensah, G. Liu, W. Jiang, "Autonomous Resource Provisioning and Resource Customization for Mixed Traffics in Virtualized Radio Access Network", *IEEE Systems Journal*, Vol. 13, No. 3, September, 2019, pp. 2454-2465.

[30] R. Li, Z. Zhao, Q. Sun, C-L. I, C. Yang, X. Chen, M. Zhao, H. Zhang, "Deep Reinforcement Learning for Resource Management in Network Slicing", *IEEE Access*, Vol. 6, November, 2018, pp. 74429-74441.

[31] C. Qi, Y. Hua, R. Li, Z. Zhao, H. Zhang, "Deep Reinforcement Learning With Discrete Normalized Advantage Functions for Resource Management in Network Slicing", *IEEE Communications Letters*, Vol. 23, No. 8, August, 2019, pp. 1337-1341.

[32] Y. Abiko, T. Saito, D. Ikeda, K. Ohta, T. Mizuno, H. Mineno, "Flexible Resource Block Allocation to Multiple Slices for Radio Access

Network Slicing Using Deep Reinforcement Learning", *IEEE Access*, Vol. 8, April, 2020, pp. 68183-68198.

[33] I. Vilà, J. Pérez-Romero, O. Sallent, A. Umbert, "A Novel Approach for Dynamic Capacity Sharing in Multi-tenant Scenarios", *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September, 2020.

[34] M. Yan, G. Feng, J. Zhou, Y. Sun and Y. Liang, "Intelligent Resource Scheduling for 5G Radio Access Network Slicing", *IEEE Transactions on Vehicular Technology*, Vol. 68, No. 8, August, 2019, pp. 7691-7703.

[35] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, X. Costa-Perez, "DeepCog: Optimizing Resource Provisioning in Network Slicing with AI-based Capacity Forecasting", *IEEE Journal on Selected Areas in Communications*, Vol. 38, No.2, February, 2020, pp. 361-376.

[36] R. Giu, J. zhang, "GANSlicing: A GAN-Based Software Defined Mobile Network Slicing Scheme for IoT Applications", *International Conference on Communications (ICC)*, 2019.

[37] S. Khatibi, A. Jano, "Elastic Slice-Aware Radio Resource Management with AI-Traffic Prediction", *European Conference on Networks and Communications (EuCNC)*, 2019.

[38] O-RAN Alliance, "Operator Defined Next Generation RAN Architecture and Interfaces", https://www.o-ran.org, Accessed December, 2020.

[39] O-RAN Alliance, "O-RAN Use Cases and Deployment Scenarios: Towards Open and Smart RAN", white paper, February, 2020.

[40] O-RAN Alliance, "O-RAN A1 interface: General Aspects and Principles Version 1.0", October, 2019.

[41] 3GPP TS 28.530 v16.0.0, "Management and Orchestration; Concepts, use cases and requirements (Release 16)", September, 2019.

[42] 3GPP TS 28.541 V16.5.0, "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3 (Release 16)", June, 2020.

[43] 3GPP TS 28.531 V16.6.0, "Management and orchestration; Provisioning; (Release 16)", July, 2020.

[44] SP-200190, "New WID Enhancement on Management Aspects of 5G Service-Level Agreement", 3GPP TSG-SA Meeting #87e, March, 2020.

[45] SP-200196, "New WID on Enhanced Closed loop SLS Assurance", 3GPP TSG-SA Meeting #87e, March, 2020.

[46] GSMA NG.116 - Generic Network Slice Template Version 2.0 (2019-10-16).

[47] 3GPP TS 22.261 v17.3.0, "Service Requirements for the 5G System; Stage 1 (Release 17)", July, 2020.

[48] 3GPP TS 22.104 v17.3.0, "Service requirements for cyber-physical control applications in vertical domains; Stage 1 (Release 17)", July, 2020.

[49] 3GPP TS 28.552 v16.6.0, "Management and Orchestration; 5G performance measurements (Release 16)", July, 2020.

[50] 3GPP TS 28.554 v16.5.0, "Management and Orchestration; 5G end to end Key Performance Indicators (KPI) (Release 16)", July, 2020.

[51] S. Guadarrama, et. al. "TF-Agents: A library for Reinforcement learning in TensorFlow.", 2018. Available at: https://github.com/tensorflow/agents

AUTHORS

Ramon Ferrús received the degrees of Telecommunications Engineering (B.S. plus M.S.) and Ph.D. from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1996 and 2000, respectively. He is currently a tenured Associate Professor with the Department of Signal Theory and Communications at UPC. His research interests include system design, functional architectures, protocols, resource optimization and network and service management in wireless communications. He has participated in 10+ research projects within the 6th, 7th and H2020 Framework Programmes of the European Commission, taking the responsibility of WP leader in H2020 VITAL and FP7 ISITEP projects. He has also participated in numerous national research projects and technology transfer projects for public and private companies. He has participated in ETSI standardisation activities. He is co-author of one book on mobile and one book on mobile broadband public safety communications. He has co-authored over 120 papers mostly in IEEE journals and conferences, with a h-index of 23 in Google Scholar.

Jordi Pérez-Romero. is a professor in the Dept. of Signal Theory and Communications of the Universitat Politècnica de Catalunya (UPC) in Barcelona, Spain, where he received a degree in telecommunications engineering in 1997 and a Ph.D. degree in 2001. He has been working in the field of wireless communication systems, with a particular focus on radio resource management, cognitive radio networks and network optimization. He has been involved in different European projects with different responsibilities, such as researcher, work package leader, and Project Lead, has participated in different projects for private companies and has contributed to the 3GPP and ETSI standardization bodies. He has published more than 250 papers in international journals and conferences, three books and has contributed to seven book chapters. He has an h-index of 31 in Google Scholar. He serves as an Associate Editor for IEEE Vehicular Technology Magazine and EURASIP Journal on Wireless Communications Networks.

Oriol Sallent is a Professor at the Universitat Politècnica de Catalunya (UPC) in Barcelona. He has participated in a wide range of European and national projects, with diverse responsibilities as Principal Investigator, Coordinator and Workpackage Leader. He regularly serves as a consultant for a number of private companies. He has been involved in the organization of many different scientific activities, such as Conferences, Workshops, Special Issues in renowned international journals, etc. He has contributed to standardization bodies such as 3GPP, IEEE and ETSI. He is co-author of 13 books and has published 250+ papers, mostly in high-impact IEEE journals and renowned international conferences. His research interests include 5G RAN (Radio Access Network) planning and management, artificial intelligence-based radio resource management, virtualization of wireless networks, cognitive management in cognitive radio networks and dynamic spectrum access and management among others.

Irene Vilà received her B.E. degree in Telecommunication Systems Engineering and her M.E. degree in Telecommunication Engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 2015 and 2017, respectively. In 2018, she joined the Mobile Communication Research Group (GRCM) of the Department of Signal Theory and Communications (TSC) at UPC where she is currently a PhD student, supported with an FI AGAUR grant by the Government of Catalunya. Her current research interests include Radio Access Network (RAN) Slicing, network virtualization and the application of artificial intelligence and, particularly, machine learning to radio resource management.

Ramón Agustí received a degree of Engineer of Telecommunications from the Universidad Politécnica de Madrid, Spain, in 1973, and a Ph.D. degree from the Universitat Politècnica de Catalunya (UPC), Spain, 1978. He became Full Professor of the Department of Signal Theory and Communications (UPC) in 1987. After graduation he was working in

the field of digital communications with particular emphasis on transmission and development aspects in fixed digital radio, both radio relay and mobile communications. For the last fifteen years he has been mainly concerned with aspects related to radio resource management in mobile communications. He has published about two hundred papers in these areas and co-authored three books. He participated in the European program COST 231 and in the COST 259 as Spanish representative delegate. He has also participated in the RACE, ACTS and IST European research programs as well as in many private and public funded projects. He received the Catalonia Engineer of the year prize in 1998 and the Narcis Monturiol Medal issued by the Government of Catalonia in 2002 for his research contributions to the mobile communications field. He is a Member of the Spanish Engineering Academy.

6G VISION: AN ULTRA-FLEXIBLE PERSPECTIVE

Ahmet Yazar[1], Seda Doğan Tusha[1], Huseyin Arslan[1,2]
[1]Department of Electrical and Electronics Engineering, Istanbul Medipol University, Istanbul, 34810, Turkey
[2]Department of Electrical Engineering, University of South Florida, Tampa, FL, 33620, USA

NOTE: Corresponding author: Ahmet Yazar (ayazar@medipol.edu.tr)

Abstract – *The upcoming sixth generation (6G) communications systems are expected to support an unprecedented variety of applications, pervading every aspect of human life. It is clearly not possible to fulfill the service requirements without actualizing a plethora of flexible options pertaining to the key enabler technologies themselves. At that point, this work presents an overview of the potential 6G key enablers from the flexibility perspective, categorizes them, and provides a general framework to incorporate them in the future networks. Furthermore, the role of artificial intelligence and integrated sensing and communications as key enablers of the presented framework is also discussed.*

Keywords – 6G, adaptive, artificial intelligence, cognitive radio, dynamic, flexibility, sensing.

1. INTRODUCTION

Following the successful standardization of the Fifth Generation (5G) networks worldwide, academia and industry have started to turn their attention to the next generation of wireless communications networks [1]. At present, there are more than 100 research papers regarding the Sixth Generation (6G) of wireless communications. There is no doubt that new visions and perspectives will continue to be developed in the coming years. However, despite all these efforts, current literature lacks gathering the distinctive features of 6G under a single broad umbrella.

The evolution of cellular communications through different generations from the Radio Access Technology (RAT) perspective is shown in Table 1. The number of capabilities for newer cellular generations increases as a result of the need to meet diversified requirements. Flexibility[1], where it is defined as the capability of making suitable choices out of available options depending on the internal and external changes, of the communications systems eventually evolves with an increasing number of new options. In this context, Fig. 2 provides a concise flexibility analysis for different generations of cellular communications considering the features in Table 1.

The Second Generation (2G) systems have paved the way for flexibility in communications systems by means of multiple frequency reuse options, adaptive equalization, and dynamic channel allocation. The Third Generation (3G) and the Fourth Generation (4G) cellular systems have incorporated voice communications with data communications. Additionally, Code Division Multiplexing (CDMA) and Orthogonal Frequency Division Multiplexing (OFDM) have provided flexibility in terms of multiplexing, rate adaptation and interference management via the exploitation of different spreading factors and the multidimensional resource utilization, respectively.

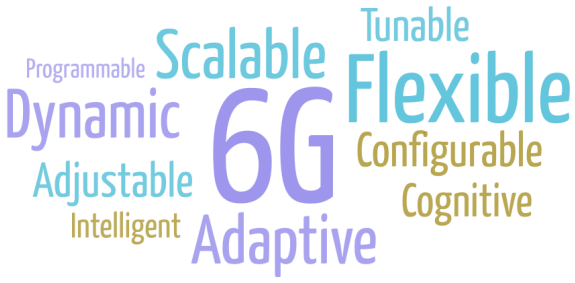


Fig. 1 – Flexibility terms.

The introduction of various services with rich requirement sets under 5G has revealed the need for a flexible network that can simultaneously meet diverse requirements. 5G has given a start for flexible wireless communications by the accommodation of different technologies. To exemplify, the coexistence of multi-numerology in a single frame has been adopted during standardization meetings. In a given network, achieving flexibility is mainly dependent on three capabilities *1) awareness, 2) availability of a rich set of technology options,* and *3) adaptation & optimization*. On this basis, although the flexibility perspective has been broadened in 5G systems with respect to previous generations, the existing technology options are not enough to reach all the goals of 5G networks. Additionally, it is expected that 6G networks will put further pressure on service providers due to emerging applications and use cases corresponding to new sets of requirements. Therefore, 6G systems need to extend the current flexibility by *(1)* exploring the awareness for the different aspects of the whole communications network and environment

[1]The other terms used interchangeably for flexibility are shown in Fig. 1.

Features/ Generations	1G			2G			3G		4G		5G	
Modulation Options	FM			GSM	GMSK		EDGE	8PSK	BPSK QPSK 16 QAM 64 QAM		NR	BPSK, QPSK, 16QAM, 64QAM, 256QAM
							CDMA 2000	QPSK, OQPSK				
				GPRS	GMSK		W-CDMA	QPSK, OQPSK				
							UMTS	QPSK				
							HSDPA	QPSK, 16 QAM				
Coding Options				Convolutional Coding			Turbo Coding		Turbo Coding		LDPC	
				Block Coding			Convolutional Coding		Convolutional Coding		Block Coding	
											Polar Coding	
Modulation and Coding Scheme (MCS) Options				Fixed			Limited MCS Options		Medium MCS Options		High MCS Options	
Waveform Options				Fixed Lattice			Fixed Lattice		Fixed Lattice		Adaptive Lattice	
				Fixed Shape (GMSK)			Fixed Shape (RC)		Windowing and Filtering		Adaptive Windowing and Filtering	
				Fixed Type (The same for uplink and downlink)			Fixed Type (The same for uplink and downlink)		Uplink	SC-FDE	Uplink	OFDM, SC-FDE
									Downlink	OFDM	Downlink	OFDM
Multiple Accessing Options	FDMA			TDMA			CDMA		Uplink	SC-FDMA	Uplink	SC-FDMA
									Downlink	OFDMA	Downlink	OFDMA
Carrier Frequency Options	microWave	AMPS	800 MHz	microWave	GSM	900 MHz	microWave	800 MHz – 2.1 GHz	microWave	600 MHz - 2.5 GHz	microWave	600 MHz – 6 GHz
		NMT	450 MHz			1800 MHz					mmWave	6 GHz – 300 GHz
Architecture Options	SISO			SISO			SISO		MIMO		mMIMO	
Cell Planning	Frequency Reuse – 7			Frequency Reuse – 3, 4, 7, 12			Frequency Reuse – 1		Frequency Reuse – 1 Fractional Frequency Reuse Soft Frequency Reuse		Frequency Reuse – 1 Fractional Frequency Reuse Soft Frequency Reuse	
User-Cell Association Options				Mobile-assisted Hand-off			Soft Hand-off		ICIC		COMP	
									Attempt to COMP		CRAN	
											Small Cell	
Diversity Options				Freq.	Frequency Hopping		Freq.	FHSS	Freq.	Multi-User Diversity	Freq.	Multi-User Diversity
											Space	MIMO
									Space	MIMO	Space	Precoding
				Time	Path Diversity		Time	DSSS	Space	Precoding	Space	Beamforming
									Space	Beamforming	Space	COMP
Receiver Types				Multi-tap TDE			Rake Receiver		A Single Tap FDE		A Single Tap FDE	
Bandwidth Options	AMPS	30 kHz		GSM	200KHz (8 slots)		CDMA	1.25 MHz	1.25 MHz to 20 MHz		BWP	
											Carrier Aggregation	
				DAMPS	30KHz (3 slots)		WCDMA	5MHz	100 MHz with Carrier Aggregation		Multi-numerology	
				PDC	25KHz (3 slots)		TD-SCDMA	1.6MHz			License Assisted Access (LAA)	

Table 1 – Increasing number of features for cellular generations.

using different sensing mechanisms including Artificial Intelligence (AI), *(2)* enriching technology options, and *(3)* providing optimum utilization of available options considering the awareness with practical sensing capabilities.

The work scopes of 6G publications in the literature are summarized in Table 2. A majority of the current 6G related studies attempt to identify the future applications and their key requirements [1,3–39]. Moreover, potential service types and application groups for 6G are analyzed in [1, 4–14, 24–28] together with the prospective key requirements of 6G networks. Several works are focused on the key enabler technologies and concepts under 6G studies in detail [1, 3] or in general [4–13, 15–21]. Furthermore, specific technologies and concepts are also being pushed for 6G as described in [24–83]. These studies are revisited in the next two sections, however, it is seen that the flexibility perspective of the key enablers is not considered as a distinguishing feature for 6G systems in the literature.

In light of the aforementioned discussions, 6G networks require the redesign of cellular communications to provide extreme flexibility in all of its building blocks. Correspondingly, this paper elaborates the example flexibility aspects of potential 6G key enablers and provides a unique categorization of the related technologies and concepts. Moreover, a novel framework is proposed to gather the said enablers under an umbrella of a single ultra-flexible framework for 6G.

The rest of the paper is organized as follows: Section 2 gives a brief overview for the initial forecasts on 6G to explain the background for the necessity of a flexible perspective without examining all potential applications, re-

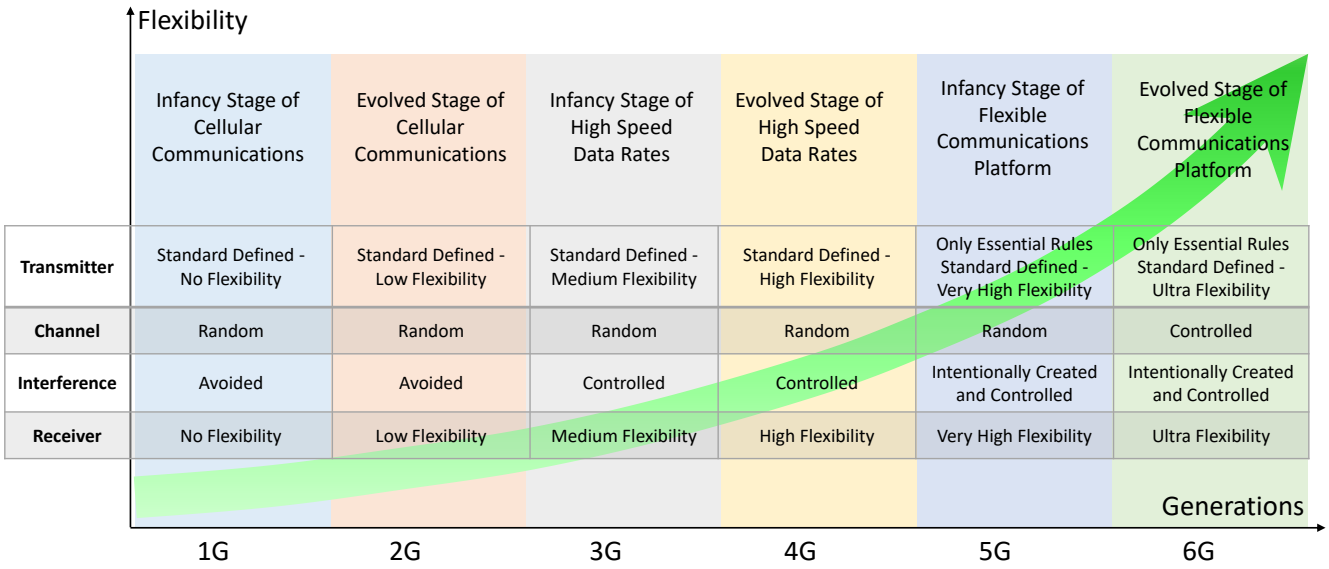


Fig. 2 – Flexibility analysis of the previous generations and 6G communications.

quirements, and service types. Flexibility discussions on the potential 6G key enablers are provided under a unique categorization rather than giving different details of these enablers in Section 3. The scope of Section 3 is limited to example flexibility aspects for the potential 6G key enablers. Next, a framework is proposed to increase interoperability of the 6G enablers in Section 4. Finally, conclusions are drawn with several open issues in Section 5.

2. A BRIEF OVERVIEW: FORECASTS ON 6G

Identification of the future applications, requirements and possible service types is one of the primary objectives of the initial 6G research studies. Fig. 3 illustrates the basic relationship between these components. Mapping the potential future applications to the several requirements with different priorities is accepted as a first step in general. Next, these requirements are grouped under the service types in a reasonable manner. At the final stage, service types have unique requirement sets for the related application groups. In 5G systems, applications are considered under three service types including enhanced Mobile BroadBand (eMBB), Ultra-Reliable and Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC) [84]. Among these, eMBB applications prioritize high throughput, capacity and spectral efficiency; mMTC prioritizes energy efficiency and massive connectivity while URLLC requires high reliability and low latency. For 6G systems, some of the initial studies inherently analyze the relations between the future applications and prioritized requirements to propose candidate service types [1, 3–38].

The following list exemplifies potential 6G applications: drone and Unmanned Aerial Vehicle (UAV) networks, drone taxi, fully automated Vehicle-to-Everything (V2X), remote surgery, health monitoring, e-health, fully sensory Virtual Reality (VR) and Augmented Reality (AR), holographic conferencing, virtual education, virtual tourism, smart city, smart home, smart clothes, disaster and emergency management, and work-from-anywhere. This list can be longer with more applications in the upcoming years. Most of the aforementioned applications were originally envisioned for 5G, however, they could not be practically realized. Therefore, it makes sense to address them first while developing the 6G networks.

General wireless communications requirements for the given application examples can be defined as: high data rate, high throughput, high capacity, high reliability, low latency, high mobility, high security, low complexity, high connectivity, long battery life, low cost, wide coverage, and more. The importance and priority of the requirements may change under different cases. Moreover, higher levels of performances need to be obtained in next generation systems while meeting the related requirements.

Since the requirement diversity is continuously increasing, more sophisticated service types are expected for 6G. Candidate service types are constituted by grouping applications with similar requirements. Examples[2] can be given as Big Communications (BigCom), secure uRLLC (SuRLLC), Three-Dimensional Integrated Communications (3D-InteCom), Unconventional Data Communications (UCDC) in [11]; ultra-High-Speed-with-Low-Latency Communications (uHSLLC) in [4]; Long-Distance and High-Mobility Communications (LDHMC), Extremely Low-Power Communications (ELPC) in [5]; reliable eMBB; Mobile Broadband Reliable Low Latency Communication (MBRLLC), massive URLLC (mURLLC),

[2]Comprehensive discussions on these potential service types can be found in the given references.

Ref.	Potential Applications and Key Requirements	Potential Service Types and Application Groups	Focusing on the Key Enablers in General	Focusing on the Key Enablers in Detail	An Inclusive Perspective for the Key Enablers	The Future of A Specific Technology or Concept
This work	✓	✓	✓		✓	
[1]	✓	✓		✓		
[3]	✓			✓		
[4–13]	✓	✓	✓			
[14]	✓	✓				
[15–21]	✓		✓			
[22, 23]	✓					
[24–28]	✓	✓				✓
[29–39]	✓					✓
[40–83]						✓

Table 2 – Scopes of 6G publications in the literature.

Human-Centric Services (HCS), Multi-Purpose Services (MPS) in [7]. As it is seen from the names, some of the service types (e.g., SuRLLC, uHSLLC, reliable eMMB, MBRLLC, mURLLC, MPS) try to be more inclusive than the 5G service types to serve target applications. It is also possible to see more specific service types (e.g., BigCom, 3D-InteCom, UCDC, LDHMC, ELPC, HCS) in comparison with 5G.

The aforementioned applications/services envisioned for 6G illustrates the expected richness of its requirements. These diverse requirements necessitate an ultra-flexible perspective for the incorporation of key enabler technologies and concepts, described below, in future networks.

3. ULTRA-FLEXIBLE PERSPECTIVE FOR 6G

In this section, an inclusive categorization of promising key enablers is presented for 6G communications and their example flexibility aspects are discussed in detail. Then, several flexibility challenges are provided for 6G. Key enabler categories and their related subcategories are shown in Fig. 4. Many of these technologies are either superficially treated or not studied during 5G standardizations, such as Integrated Sensing and Communications (ISAC) and intelligent communications.

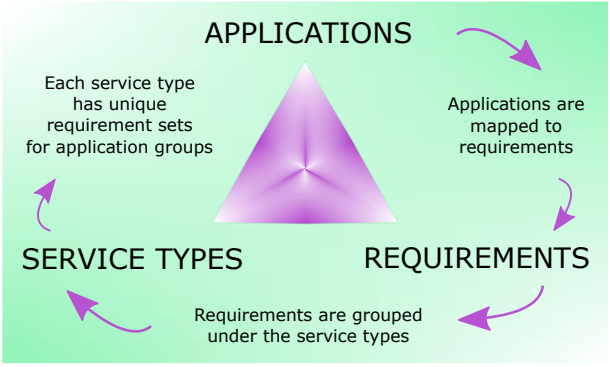


Fig. 3 – A basic relationship between the applications, requirements, and service types.

Although technologies placed in different categories can have overlapped regions, the given categorization differentiates these technologies regarding their flexibility aspects.

Table 3 provides a summary of the example flexibility options achieved by the different technologies. It is worthy to emphasize that the different key enablers have their own impact on the overall flexibility of the system. Ultimately all of them combine together to provide the complete infrastructure capable of realizing the flexible 6G vision that we aspire to achieve.

3.1 Flexible Multi-Band Utilization

The inclination of communications technologies towards high-frequency bands becomes more appealing due to the increased system capacity and throughput demands of cellular users. Furthermore, flexible usage of available frequency bands, depending on the user and service requirements, is envisioned to be an inherent characteristic of future wireless networks [74].

The millimeter Wave (mmWave) spectrum is starting to be exploited in 5G. It provides new benefits, such as multi-gigabit data rates and reduced interference, however, the use of mmWave bands in 5G is limited by the current International Mobile Telecommunications (IMT) regulations. In World Radiocommunication Conference 2019 (WRC-19), additional 17.25 GHz of spectrum is identified for IMT, where only 1.9 GHz of bandwidth was available before [85]. Therefore, it is expected that spectrum availability in these bands and consequently its flexible utilization will increase during the upcoming years [42]. Moreover, beyond 52.6 GHz communications is one of the agenda items for 3GPP Release 17 [86].

Frequency bands from 100 GHz to 3 THz are envisioned as a candidate spectrum for 6G communications [40]. If THz communications is employed in 6G, it promises a way

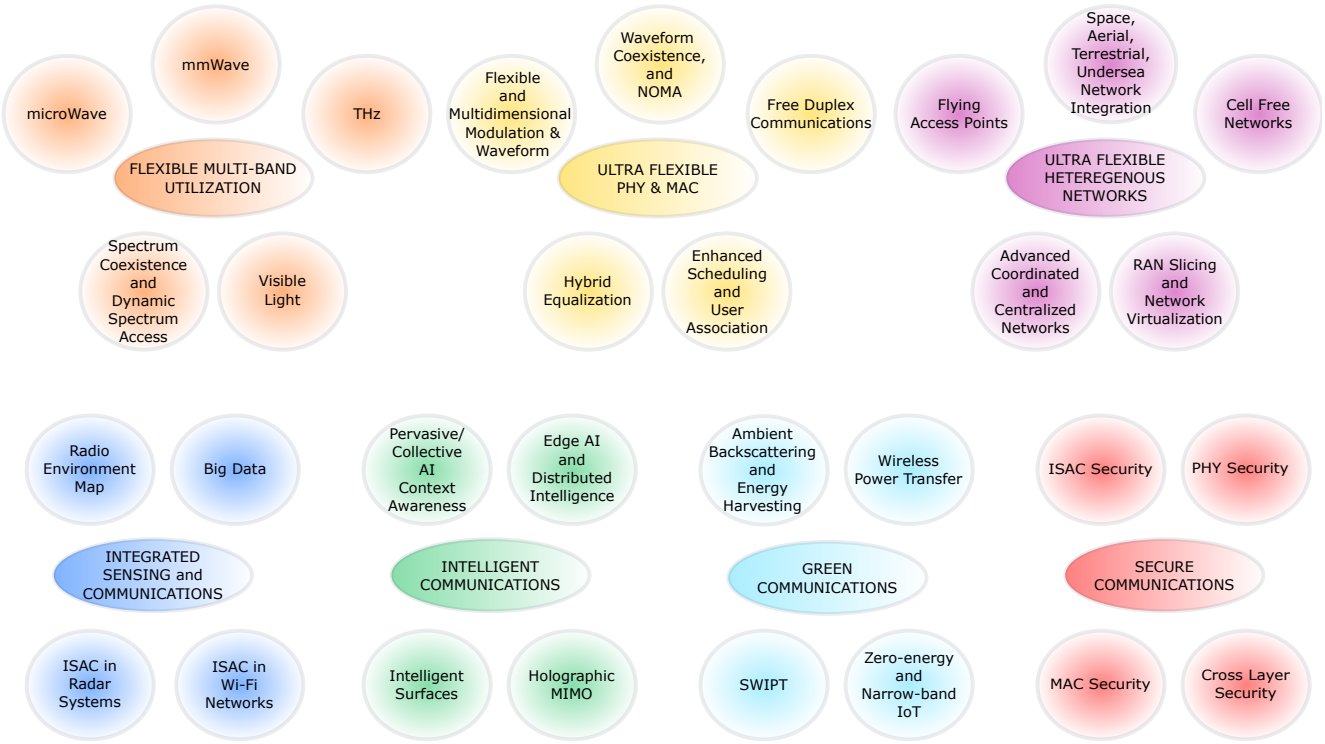


Fig. 4 – Categorization of the promising 6G key enablers under the ultra-flexible perspective.

of dealing with the spectrum scarcity issue by providing an additional degree of flexibility in assigning the most suitable frequency resources for given scenarios [46].

Apart from mmWave and THz communications, Visible Light Communications (VLC) also provides spectrum flexibility as a candidate key enabler for 6G networks [27, 52, 61]. Moreover, a new degree of freedom that is information source flexibility is exploited using visible light sources.

Spectrum coexistence is another important issue in need of flexible spectrum utilization [50, 74]. Indeed, the coexistence of cellular communications, Wi-Fi, satellite networks, and radar systems is inevitable in the future due to both scarce resources and increasing growth in user demands. To exemplify, the coexistence of radar and cellular communications in mmWave frequency bands becomes more popular nowadays [87]. Moreover, the idea of Dynamic Spectrum Access (DSA) relies on the spectrum coexistence [56].

As it is seen, there are several aspects of flexible multi-band utilization in 6G systems. Flexibility sources can be summarized under three main perspectives: 1) multi-band flexibility, 2) information source flexibility, and 3) spectrum coexistence flexibility.

3.2 Ultra-Flexible PHY and MAC

One of the unique features of 5G, specifically in the context of PHY design, is the introduction of numerology concept where different configurations of the time-frequency lattice are used to address the varying requirements [88]. While the numerology concept paves the way for flexibility in beyond 5G networks, it is rather limited considering the competing nature of requirements expected for future 6G networks [25]. In addition to the standardized activities, the use of flexible Cyclic Prefix (CP) configurations (e.g., individual CP, common CP, etc.) is explored to enhance the multi-numerology systems for 6G [89].

Taking one step beyond the use of different realizations of the same parent waveform as in 5G, multiple waveforms can be accommodated in a single frame for achieving 6G goals [49, 90]. In line with this, multi-numerology structures can be designed for promising alternative waveforms, that are more suitable for providing additional parameterization options. Having these options enhances flexibility in the PHY layer via increased adaptation capability for meeting a large number of requirements. Moreover, waveform coexistence in the same frame gives the opportunity to serve multiple networks such as radar sensing [91] and Wi-Fi communications together with 6G communications in a flexible manner. There are also several waveform-domain NOMA studies that exploit different resource utilization aspects in the literature [92–95]. Moreover,

partial and full overlapping through available resources can also be employed while designing new generation NOMA techniques [30, 96]. The waveform-domain NOMA concept provides an important flexibility by increasing the resource allocation possibilities in 6G networks [78]. Another flexibility aspect that can arise with 6G is the use of an alternate waveform domain rather than the conventional time-frequency lattice employed by 5G and older generations.

In addition to the waveform itself, there is a large number of new generation modulation options in the literature [97] and only a small set of them have appeared in the 5G standards. 6G can be enriched with the flexibility provided by these options, particularly Index Modulation (IM) based solutions [11]. This concept can even be extended to multiple domains to provide an additional degree of freedom [98]. Moreover, modulation techniques are adaptively designed considering the other key enablers such as Non-Orthogonal Multiple Access (NOMA) [99] and Reconfigurable Intelligent Surface (RIS) [45] for 6G.

Since the configuration of the PHY parameters is, to a large extent, controlled by the Medium Access Control (MAC) layer, it is imperative to develop the flexibility and adaptation capabilities of both layers simultaneously. Two important issues that require flexibility in PHY and MAC would be the "waveform parameter assignment" or "numerology scheduling" paradigm under the context of 5G multi-numerology systems [25, 100], where the MAC layer is responsible for assignment of parameters of the PHY signal. Similarly, adaptive guard utilization methods have been developed for the MAC layer [101–103] to control the new type of interferences in 5G systems. On this basis, it is expected that highly intelligent UE capabilities, and configurable network parameters, and flexible and efficient MAC designs will play a key role in 6G networks due to the expected increased diversity in service types and consequently requirements.

Example flexibility perspectives for ultra-flexible PHY and MAC technologies of potential 6G key enablers are given in Table 3.

3.3 Ultra-Flexible Heterogeneous Networks

Flying Access Points (FAPs) provide enhanced flexibility for network deployment by allowing dynamic (3-D) positioning of the nodes or even optimized trajectory planning for different objective functions [47, 55, 59]. The push in this direction occurred around the turn of the century [104], and was further empowered by projects, such as: 1) Google Loon project, 2) Facebook Aquila project, 3) ABSOLUTE project, 4) Matternet project, and 5) Thales Stratobus project. The integration of FAPs with the terrestrial network can be leveraged to provide coverage in disaster/emergency scenarios, connectivity for rural/isolated areas and capacity enhancement for temporarily crowded places (such as stadiums/concert venues) [77]. FAP-based networks are expected to be an important part of 6G not only for achieving deployment flexibility but also for having better wireless propagation provided by a high probability of Line of Sight (LOS) communications [41, 63].

In addition to the aerial and terrestrial networks, the integration of space (satellite) networks is another aspect of the flexible heterogeneous networks [54]. Space networks are also a promising solution for rural area communications [31]. They are employed for wireless backhaul communications in the previous cellular networks. However, space networks can also serve aerial user equipment such as drones and UAVs to increase coverage flexibility in 6G systems [72]. Moreover, undersea network integration with the other networks will be useful while serving naval platforms.

Although, the integration of different networks is ensured, the cell structures of these networks are changing. Cell-less or cell-free networks are one of the potential 6G concepts considering the network architecture richness [105, 106]. User equipment connects to the network via multiple small cells in the cell-less networks. Cell-centric design is transformed into the user-centric system. Hence, it provides both handover-free communications and zero inter-cell interference. Cell-less networks may exploit a new dimension of network Multi-Input Multi-Output (MIMO) flexibility in 6G. As another network MIMO example, advanced coordinated and centralized networks [107] are addressed together with NOMA schemes for 6G communications [108, 109]. These networks are called multi-cell NOMA. Flexibility comes with the number of the cells and architecture richness while exploiting other dimensions with NOMA.

From the network virtualization perspective, network slices are used in 5G to customize and optimize the network for service types or any other requirement sets [110–112]. Hence, the overall performance is increased by meeting different requirement sets with virtually privatized networks. Network slicing brings an important flexibility in 5G since it enables different network options under the same umbrella. The number of network slices can increase for 6G and there may be network slices for each user equipment. This user-centric network slicing architecture can provide full flexibility in the network layer.

The number of examples for the flexibility aspects of promising 6G heterogeneous networks can be increased with particular technologies and concepts such as blockchain systems [26, 33] and quantum communications [29] in the future.

Key Enabler Categorization	Flexibility Aspect	Example Details
Flexible Multi-Band Utilization	Multi-band flexibility	microWave, mmWave, THz, visible light
	Information source flexibility	Radio signals, visible light
	Spectrum coexistence flexibility	DSA, CR, and coexistence of cellular networks, Wi-Fi networks and radar systems
Ultra-Flexible PHY and MAC	Modulation-option flexibility	BPSK, QPSK, 16QAM, 64QAM, 256QAM, 1024QAM, etc.
	Multi-domain modulation flexibility	IM (shape, interval, position, etc.), space, time, frequency, etc.
	Multi-type coding	New types of LDPC, block coding, polar coding, etc.
	MCS option flexibility	Ultra adaptive MCS
	Multi-option waveform flexibility	Multiple numerologies for any specific waveform
	Waveform processing flexibility	Adaptive windowing/filtering and the related configurable parameters
	CP utilization flexibility	Individual and common CP utilizations
	Adaptive guard utilization flexibility	Flexible guards for multi-waveform and multi-numerology designs
	Multi-waveform flexibility	Waveform coexistence in the same frame
	Multi-network multi-waveform flexibility	Waveform coexistence for cellular and Wi-Fi networks with radar sensing
	Multi-domain NOMA flexibility	Partial and fully overlapped resources with waveform-domain NOMA
	Multi-domain waveform flexibility	Alternative lattice flexibility together with the time-frequency lattice
	Multiple access flexibility	Fully flexible, both orthogonal and non-orthogonal
	Receiver-type flexibility	Fully flexible, hybrid equalization
	Bandwidth option flexibility	BWP, carrier aggregation, LAA, DSA, etc.
	User association flexibility	Multiple options under heteregonous networks, flexible user parameter assignment
	Channel access flexibility	GB transmission, GF transmission, and their coexistence over a resource pool
Ultra-Flexible Heterogenous Networks	Positioning flexibility of the access points	Flying access points can be positioned flexibly in the sky
	Connection link and relaying flexibility	User equipment can connect to different type of access points
	Altitude-based multi-network flexibility	Coexistence of space, HAP, terrestrial and undersea networks
	Coverage flexibility	Rural area coverage with space and HAP networks
	Network architecture flexibility	Ultra massive MIMO, small cell, D2D, relaying via different networks, etc.
	Cell-free network flexibility	User-centric network designs, handover-free communications
	Multi-cell flexibility	Network MIMO solutions, multi-cell NOMA, etc.
	Network slice flexibility	Network slices for each user equipment, user-specific virtual networks
Integrated Sensing and Communications	Multi-system flexibility	All systems can collabarate with the wireless communications in different ways
	Awareness flexibility	Awareness in spectrum, location, mobility, context, user, channel, interference, etc.
Intelligent Communications	Alternative solution flexibility	No need to get stuck on conventional algorithm designs
	Edge computing flexibility	Signal and data processing at the edge nodes
	Channel control flexibility	Different types of intelligent surfaces
	Interference management flexibility	Interference management with the channel control
	Softwarization flexibility	Programmable architecture options for holographic MIMO systems
Green Communications	Battery-free implementation flexibility	Removing battery limitations and constraints
	Interference exploitation flexibility	Interference can be useful for the energy harvesting
Secure Communications	Multi-domain security flexibility	No need to get stuck on key sharing security mechanisms, complementary solutions
	Wireless channel exploitation flexibility	PHY security methods exploit the characteristics of wireless channel

Table 3 – Example flexibility aspects for the key enabler categories.

3.4 Integrated Sensing and Communications

With the emphasis on use cases such as V2X communications in recent years, sensing has attained increased importance leading to the integration of these two applications [113]. However, the use of sensing is not limited to V2X or autonomous driving. Rather, if there is any observable data that can be utilized for the optimization or enhancement of the communications systems, it should be leveraged in 6G [114]. The information pertaining to the radio environment can be utilized in improving network deployment, optimizing user association, providing secure communications and so on. Hence, one of the unique novelties of 6G systems is the integration of many different sensor hardware with the heterogeneous communications networks as exemplified in Fig. 5.

While it might sound like a novel idea to some, Integrated

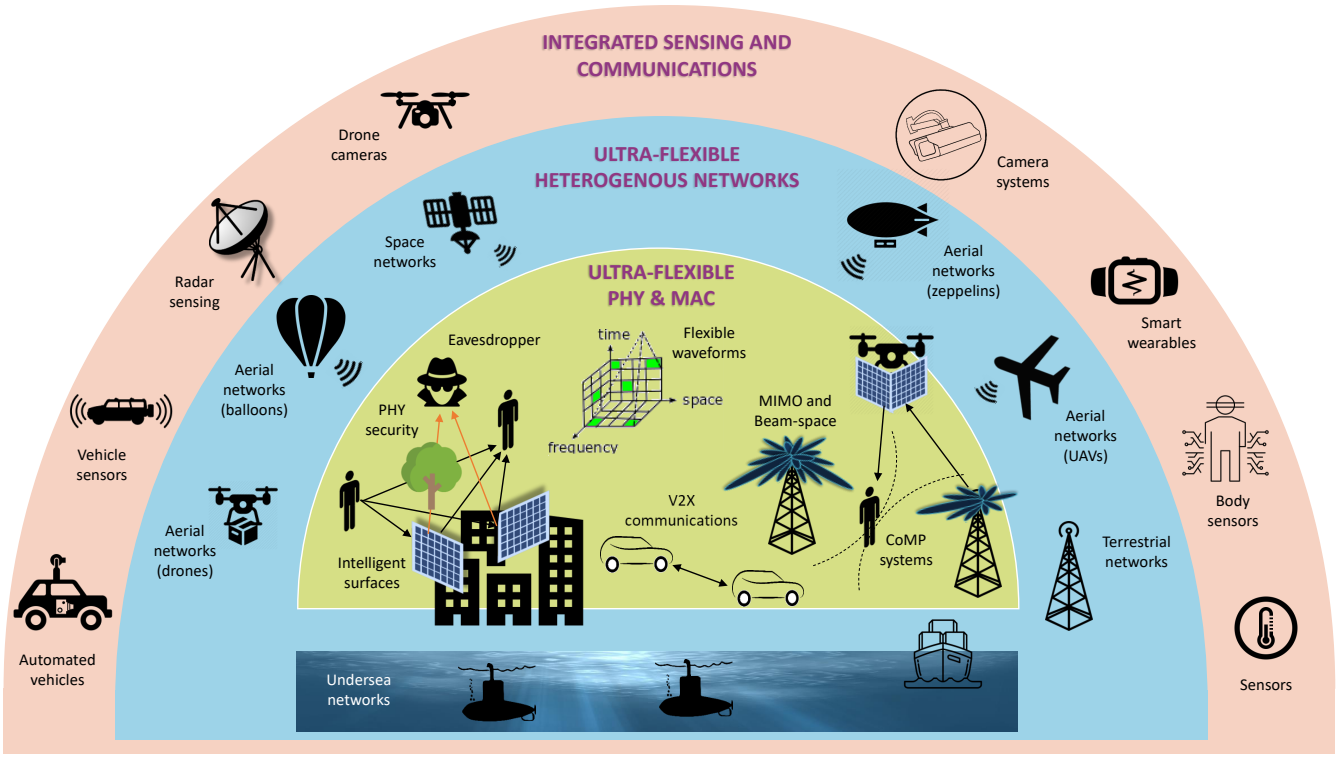


Fig. 5 – The integration of many different sensor hardware with the heterogeneous communications networks under 6G systems.

Sensing and Communications (ISAC) has been studied in different domains in the past. Cognitive Radio (CR) applications triggered the ISAC research on the last two decades. Spectrum sensing and awareness is one of the first application areas in the ISAC research [115]. Location awareness is exploited to improve the wireless communications system design in [116]. Satellite and drone images can be used to predict channel parameters [117]. Context-awareness is used to optimize network architectures in wireless communications [118]. ISAC systems are studied for radar sensing [91, 119] and Wi-Fi network coexistence [120] in the literature. However, the complete list of sensing information that can be useful for the next generation cellular communications systems from the ISAC perspective has not yet been comprehensively studied [114].

A Radio Environment Map (REM) is a realization of the ISAC concept [121]. It is mainly used to obtain environmental information in the literature, however, for the next generation systems the REM concept will be generalized from environmental-awareness to complete-awareness. REM may include all sensing information in a multi-dimensional manner for wireless communications networks. To exemplify, REM can be a specialized database for the ISAC. Therefore, the flexibility level of the ISAC systems can be determined by the dimensions in REMs. Each dimension in a REM increase the awareness, allowing better resource utilization. Moreover, control of the configurable options and parameters in different communications layers of 6G can be enhanced by more granular REM information.

The complete information and awareness of the environment comes at the cost of a high volume of data, variety of sources and significant processing [80,82]. This necessitates the use of big-data processing techniques [122]. A significant challenge, however, in this regard is the overhead of data exchange between the sensing and processing nodes. A centralized solution might not be suitable in such scenarios, rendering the use of edge-computing imperative, particularly for low-latency use cases. Moreover, the usage of Artificial Intelligence (AI) solutions can be helpful while processing big-data at the edge nodes.

3.5 Intelligent Communications

The usage of AI in the communications society has increased in recent years. Several survey and tutorial papers are published on the usage of Machine Learning (ML) for wireless communications [34, 123–129]. AI-aided design and optimization has even been leveraged for the flexible implementation options provided in 5G [25]. In many of the studies, AI is put at the center of 6G visions [6, 8, 24, 28, 32, 44, 67, 75, 76, 83] to complement the classical methods. Indeed, the use of AI is inevitable to incorporate intelligence in the future networks [130–132]. AI-aided methods can propose fast and efficient solutions in case enough data is available.

AI and ML also find a range of applications in ISAC and REM paradigms to extract information regarding the en-

vironment from sensed data. A flexible communications system needs to benefit from the advantages of popular ML approaches such as reinforcement learning, deep learning, and edge computing [37, 48, 69, 73]. Especially distributed intelligence (edge AI) with edge computing is a promising paradigm for 6G communications [36]. The management of multi-band utilization, MAC layer control, heterogeneous and cell-less networks, and the ISAC systems cannot be done in an all centralized manner. Edge computing will play an important role at that point with the help of distributed intelligence so 6G big data can be processed at the edge nodes without being collected at a centralized network.

Intelligent networks are not limited to AI-aided concepts. For example, RIS technology is one of the most popular research topics nowadays [133, 134]. Intelligent surfaces bring a new flexibility on the control of channel parameters [57]. In the past, a wireless channel was just an observable medium. However, it can be controlled at some level with new generation wireless systems. Interference management flexibility is increased by controlling capabilities of the wireless channel. These flexibility aspects also affect the technology designs in different communications layers [66, 71]. To exemplify, having a control capability in multipath propagation, such as controlling delay spread, Doppler spread and the number of multipath alleviates the constraints related to waveform design. RIS technology can also be considered as passive holographic MIMO surfaces if it is located closer to the transmitter and receiver antennas [53]. Additionally, it is possible to employ holographic MIMO surfaces as active elements. The active holographic MIMO surfaces work similar to massive MIMO but their softwarization flexibility is higher than the conventional MIMO systems [53].

3.6 Green Communications

While candidate 6G key enablers are increasing the flexibility in different domains, new architectural changes of 6G should support energy efficiency and green communications [43, 64, 70]. Zero-energy Internet of Things (IoT) is one of the most important concepts since ultra low-power wireless communications is necessary for 6G connectivity [51]. In this context, Radio Frequency (RF) energy harvesting is studied with ambient backscatter technology for 6G communications [135, 136]. Thus, low-power wireless systems can obtain their energy from the available high-power radio waves. Backscatter communications enables energy harvesting, simplifying the implementation of zero-energy IoT designs. Provision of rich options for energy-efficiency promises fulfilment of energy requirement variations belonging to different applications. Within this direction, the Symbiotic Radio (SR) concept offers highly reliable backscattering communications together with mutualism spectrum sharing [137, 138].

It is also possible to benefit from Wireless Power Transfer (WPT) while designing zero-energy IoT systems [139]. Under the WPT concept, Simultaneous Wireless Information and Power Transfer (SWIPT) is the most popular technology that may be a candidate for 6G networks [60, 140]. SWIPT designs are also used for interference exploitation purposes [141] since interference can be useful for energy harvesting. Transformation of interference into an energy source introduces another flexibility perspective.

3.7 Secure Communications

With applications such as eHealth, online banking, and autonomous driving etc., wireless communications promises to be an enabler of innumerable sensitive applications utilizing private data. However, the broadcast nature of wireless communications makes it vulnerable to several security threats such as eavesdropping, impersonation, and jamming. In order to ensure security of such applications, PHY Layer Security (PLS) is an emerging solution that has the capability to complement the conventional cryptography-based security techniques. In fact, PLS is more suited for the increased heterogeneity and power/processing restrictions of future wireless networks since it exploits the characteristics of the wireless channel and PHY properties associated with the link such as noise, fading, interference, and diversity [142]. It is also possible to increase this flexibility by designing cross-layer security algorithms with the PHY and MAC layer [143]. In several 6G papers, secure communications is discussed as one of the main topics [8, 35, 44, 81]. PHY and cross-layer security concepts are expected to play a critical role in 6G networks because of their capability to support joint design of security, reliability, and latency.

As discussed in the previous subsections, ISAC and REM concepts will be important enablers in 6G communications. However, a new security problem arises since there may be a large amount of confidential data for ISAC and REM concepts. In the literature, this problem is treated in [144] for ISAC security, and in [145] for REM security. Thus, there is a need for more secure communications options in 6G networks to meet new types of security requirements, especially for ISAC and REM concepts. Moreover, in order to tackle spoofing attacks, authentication at the physical layer by using features of channel and hardware impairments can also provide a fast, lightweight, and efficient alternative for crypto-security for authentication in future wireless networks. Furthermore, the physical layer solution will also provide efficient robustness against jamming attacks using terrestrial and flying relay and other new multi-antenna-based solution.

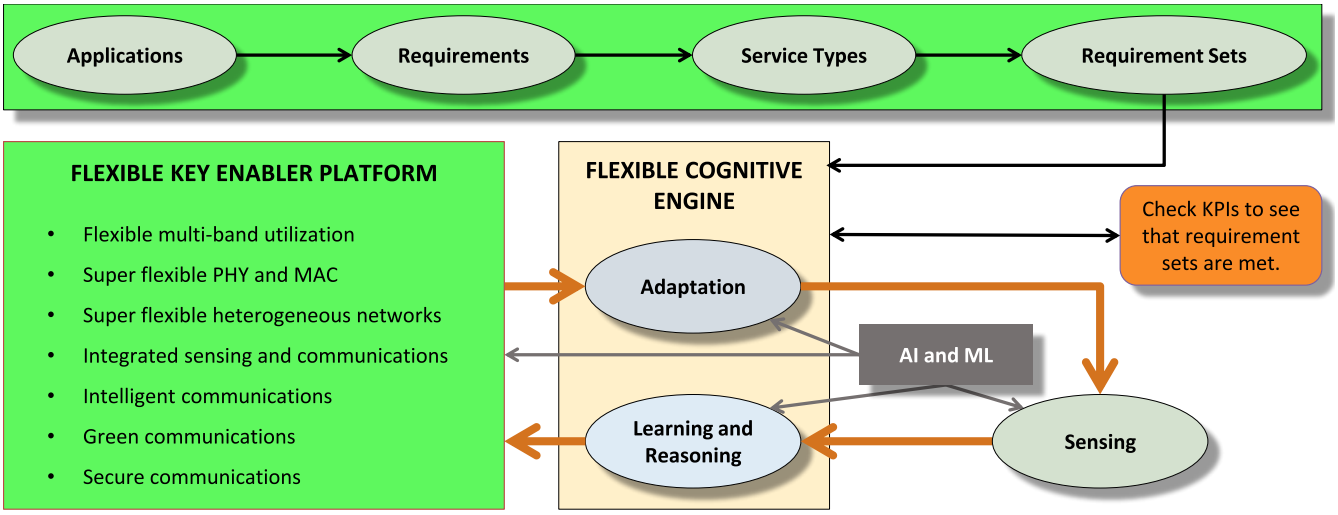


Fig. 6 – The proposed framework includes flexible key enabler platform and flexible cognitive engine. The flexible cognitive engine can be defined as a bridge between the requirements and potential technology options with the related configurations.

3.8 Flexibility Challenges and Opportunities in 6G

The exemplified key enablers show that 6G will have many different flexibility options while 5G systems have limited flexibilities. However, each flexibility is coming with unique challenges. In other words, flexibility opportunities bring new challenges for the 6G networks.

For flexible multi-band utilization, operating the cellular system at multiple frequency bands needs advanced front-end hardware. Additionally, spectrum coexistence of different networks causes new interference problems. If the flexibility challenges on the PHY and MAC layer are investigated, one of the most important problems is the necessity of a flexible waveform system. At that point, either a single but an ultra-flexible waveform can be designed or multiple waveforms can be employed in the same frame. Designing a single waveform to meet all types of requirements did not work for 5G networks. It will be more difficult for 6G with more types of requirements. Moreover, waveform coexistence in the same frame causes new interferences (like inter-numerology interference in 5G). Similarly, partial and fully overlapped NOMA systems have the same interference problem. Control and mitigation of these interferences is expected as another challenge.

Flexibility challenges for heterogeneous networks can be exemplified with the developing optimal positioning and relaying algorithms for flying access points. In addition to these algorithms, interference management during the coexistence of different networks is necessary. As another challenge, network MIMO structures provide multi-cell flexibility, however, large amounts of data need to be transferred at the backhaul systems and the amount of burden increases.

As discussed in the previous subsections, ISAC systems can include multiple systems together with the communications networks. Generally, the amount of sensing information increases in parallel to awareness capabilities. However, processing the sensing information causes computational burdens. Additionally, investigating the ways of exploiting this information to enrich the communications systems is another important challenge.

For the flexibility challenges of intelligent communications, first of all, an efficient work distribution between conventional and ML methods is required. A large data sets and useful features need to be developed to make ML mechanisms more functional. Additionally, edge computing algorithm structures should be designed to reduce the workload at transmission points.

If we summarize the challenges and opportunities, the following items can be listed:

- Need for a rich set of algorithms and techniques at different layers of the protocol stack that are optimized for different applications with their own requirements.
- Integration of these rich sets of algorithms into the flexibility framework with minimal overhead and complexity.
- Development of techniques that allows flexibility with a simple parameter change without significantly impacting the rest of the system design.
- Integration of AI and ML techniques to solve complex system problems together with the classical model based approaches. AI/ML can be applied in different parts of our proposed framework, i.e. it can be applied for better sensing and learning, or for optimal use of the given set of algorithms and approaches, or developing better solutions in the transmission, reception, and modeling of the system.

Therefore, there is a need for general frameworks and mechanisms to ease dealing with these challenges all together. Within this direction, an example framework is proposed in the next section.

4. ULTRA-FLEXIBLE 6G FRAMEWORK

Gathering together all potential 6G enablers in a flexible framework is an important challenge. Therefore, this section brings the above-mentioned flexible perspectives for the key enabler technologies and concepts under the umbrella of a single ultra-flexible framework for 6G. Here, it is important to realize that the presence of flexible options in itself is not enough to render a network intelligent. Rather, it needs the capability to make best use of the available options. Therefore, some sort of intelligence or cognition is imperative in future wireless networks. Keeping this in mind, the proposed framework has the following primary components: 1) Flexible key enabler platform (like an advanced Mitola radio), 2) flexible cognitive engine, and 3) flexibility performance indicators. Fig. 6 illustrates how these different components are interconnected within the framework. The key points of this framework can be summarized as follows:

1. New technologies should be integrated into communications standards via a **flexible key enabler platform** without waiting for ten years.

2. Key enabler technologies should work together in an optimal flexibility to meet different requirements. Therefore, a **flexible cognitive engine** can make an optimization between different flexibility aspects.

3. The amount of flexibility needs to be measured while making an optimization. Hence, developing new **flexibility performance indicators** is necessary.

The previous cellular communications generations were standardized approximately ten years apart. From a different point of view, it took about a decade for the available technologies to be included in the cellular standards. Waiting up to ten years to benefit from an available technology does not make sense if it is possible to develop a platform that hosts different technologies flexibly. For now, we need to tolerate the limited flexibility of 5G technologies for the next decade. However, an advanced Mitola radio can work like a smart phone that has installable and updateable software. We call this radio a **flexible key enabler platform**. In this concept, the platform has the ability to have new key enabler technologies by a softwarization. Thus, the flexibility level of the wireless communications system can be enhanced with new technologies and the related updates.

As it is shown in Fig. 6, each technology can bring different perspectives to the overall flexibility. There is a need for a multi-objective optimization unit to control all configurable and flexible aspects of the enablers in the flexible key enabler platform. This engine can be designed in an AI-aided manner to optimize the key enabler flexibilities jointly. An optimum work distribution should be done for the flexible configurations of key enablers to meet all the system requirements in the most efficient way. At the end, all system requirements should be met optimally. The **flexible cognitive engine** will guarantee this optimization by the help of Key Performance Indicators (KPIs) that show the success while meeting requirements. This flexibility optimizer considers also complexity requirements while operating the system.

ISAC technologies will be an important part of 6G technologies as discussed in the previous section. Any sensing information can be exploited to make the wireless communications more effective. The flexible cognitive engine can give decisions with more available information while meeting different requirements and handling with several impairments and constraints. Sensing information increases the awareness and controlling capabilities of the system. To provide these capabilities, AI tools in the flexible cognitive engine provide useful and unnoticeable relationships without heuristic designs and theoretical analysis. Hence, the flexible cognitive engine needs three important elements while optimizing the flexibility level with key enablers: 1) Sensing information to increase awareness and controlling capabilities, 2) AI tools to increase the functionality and effectiveness of sensing information, and 3) KPIs to monitor the overall system.

KPIs are needed to measure several performances of the communications system. One of these KPIs can be the **flexibility performance indicator** so that the achieved flexibility can be quantified. It is difficult to decide on a specific flexibility performance indicator because there are many different flexibility perspectives as shown in Table 3. This indicator can be technology-specific and require separate metrics for different technology categories. 6G networks will need flexibility indicators similar to the other KPIs such as spectral efficiency and reliability. Generally, the current key enabler technologies are not designed to be called flexible technologies. Flexibility aspects of these key enablers are described mostly based on the inferences. In ideal conditions, 6G technologies need to be designed considering the flexibility perspective as one of the key criteria. At that point, flexibility performance indicators should be employed to quantify the advantages and disadvantages of new designs in both the PHY and MAC layer.

5. CONCLUSION

5G systems were characterized by diverse applications and requirements. 6G is expected to continue in the same vein by enriching the application fabric even further. Fulfilling such a wide variety of use cases is not possible unless flexibility is incorporated in the promising key enabling technologies for the future networks. Driven by

this, we have presented example flexibility aspects for the potential 6G key enablers under a unique categorization.

We believe that 6G should be approached with flexibility at its primary design criterion. Flexibility aspects of the potential key enablers need to play a leading role in the design stages of 6G systems. To this end, we have presented a general framework comprising of the aforementioned flexible key enablers, empowered by a flexible cognitive engine and supported by different aspects of sensing and AI. We believe that the presence of flexible options is imperative but only that is not enough to support the future applications. The ability to extract information regarding the operating environment and making related intelligent decisions are the way forward in the wireless communications realm. The best possible utilization of the flexibility offered by the key enablers is determined with this vision.

The realization of a flexible key enabler platform like the one mentioned above is, however, not straightforward. It requires the methods capable of performing efficient multi-objective optimization to address the various competing applications requirements. Furthermore, quantifying the flexibility by proposing novel performance indicators also remains a significant challenge on the way to ensure a fully-functional flexible, cognitive wireless communications network.

ACKNOWLEDGEMENT

The authors would like to thank Muhammad Sohaib J. Solaija for his valuable comments and suggestions to improve the quality of the paper.

REFERENCES

[1] I. F. Akyildiz, A. Kak and S. Nie, "6G and Beyond: The Future of Wireless Communications Systems," in IEEE Access, vol. 8, pp. 133995-134030, 2020, doi: 10.1109/ACCESS.2020.3010896.

[2] H. Arslan and E. Basar, "Flexible and Cognitive Radio Access Technologies for 5G and Beyond," IET, 2020.

[3] L. Bariah et al., "A Prospective Look: Key Enabling Technologies, Applications and Open Research Topics in 6G Networks," in IEEE Access, vol. 8, pp. 174792-174820, 2020, doi: 10.1109/ACCESS.2020.3019590.

[4] B. Zong, C. Fan, X. Wang, X. Duan, B. Wang and J. Wang, "6G Technologies: Key Drivers, Core Requirements, System Architectures, and Enabling Technologies," in IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 18-27, Sept. 2019, doi: 10.1109/MVT.2019.2921398.

[5] Z. Zhang et al., "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," in IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 28-41, Sept. 2019, doi: 10.1109/MVT.2019.2921208.

[6] L. Zhang, Y. Liang and D. Niyato, "6G Visions: Mobile ultra-broadband, super internet-of-things, and artificial intelligence," in China Communications, vol. 16, no. 8, pp. 1-14, Aug. 2019, doi: 10.23919/JCC.2019.08.001.

[7] W. Saad, M. Bennis and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," in IEEE Network, vol. 34, no. 3, pp. 134-142, May/June 2020, doi: 10.1109/MNET.001.1900287.

[8] G. Gui, M. Liu, F. Tang, N. Kato and F. Adachi, "6G: Opening New Horizons for Integration of Comfort, Security and Intelligence," in IEEE Wireless Communications, doi: 10.1109/MWC.001.1900516.

[9] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan and M. Zorzi, "Toward 6G Networks: Use Cases and Technologies," in IEEE Communications Magazine, vol. 58, no. 3, pp. 55-61, Mar. 2020, doi: 10.1109/MCOM.001.1900411.

[10] J. F. Monserrat, D. Martin-Sacristan, F. Bouchmal, O. Carrasco, J. Flores de Valgas and N. Cardona, "Key Technologies for the Advent of the 6G," 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), Seoul, Korea (South), 2020, pp. 1-6, doi: 10.1109/WCNCW48565.2020.9124725.

[11] S. Dang, O. Amin, B. Shihada and M. Alouini, "What should 6G be?," in Nature Electronics, vol. 3, pp. 20–29, 2020.

[12] M. H. Alsharif et al., "Sixth Generation (6G) Wireless Networks: Vision, Research Activities, Challenges and Potential Solutions," in Symmetry, vol. 12, no. 4, pp. 1-21, 2020.

[13] F. Tariq, M. R. A. Khandaker, K. -K. Wong, M. A. Imran, M. Bennis and M. Debbah, "A Speculative Study on 6G," in IEEE Wireless Communications, vol. 27, no. 4, pp. 118-125, Aug. 2020, doi: 10.1109/MWC.001.1900488.

[14] G. Liu et al., "Vision, requirements and network architecture of 6G mobile network beyond 2030," in China Communications, vol. 17, no. 9, pp. 92-104, Sept. 2020, doi: 10.23919/JCC.2020.09.008.

[15] P. Yang, Y. Xiao, M. Xiao and S. Li, "6G Wireless Communications: Vision and Potential Techniques," in IEEE Network, vol. 33, no. 4, pp. 70-75, July/Aug. 2019, doi: 10.1109/MNET.2019.1800418.

[16] S. Elmeadawy and R. M. Shubair, "6G Wireless Communications: Future Technologies and Research Challenges," 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 2019, pp. 1-5, doi: 10.1109/ICECTA48151.2019.8959607.

[17] G. Wikström et al., "Challenges and Technologies for 6G," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083880.

[18] Y. Yuan, Y. Zhao, B. Zong and S. Parolari, "Potential key technologies for 6G mobile communications," in Science China Information Sciences, vol. 63, no. 183301, pp. 1-19, 2020.

[19] L. U. Khan, I. Yaqoob, M. Imran, Z. Han and C. S. Hong, "6G Wireless Systems: A Vision, Architectural Elements, and Future Directions," in IEEE Access, vol. 8, pp. 147029-147044, 2020, doi: 10.1109/ACCESS.2020.3015289.

[20] A. Dogra, R. K. Jha and S. Jain, "A Survey on beyond 5G network with the advent of 6G: Architecture and Emerging Technologies," in IEEE Access, doi: 10.1109/ACCESS.2020.3031234.

[21] S. P. Rout, "6G Wireless Communication: Its Vision, Viability, Application, Requirement, Technologies, Encounters and Research," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-8, doi: 10.1109/ICCCNT49239.2020.9225680.

[22] K. David and H. Berndt, "6G Vision and Requirements: Is There Any Need for Beyond 5G?," in IEEE Vehicular Technology Magazine, vol. 13, no. 3, pp. 72-80, Sept. 2018, doi: 10.1109/MVT.2018.2848498.

[23] A. Mourad, R. Yang, P. H. Lehne and A. de la Oliva, "Towards 6G: Evolution of Key Performance Indicators and Technology Trends," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083759.

[24] M. J. Piran and D. Y. Suh, "Learning-Driven Wireless Communications, towards 6G," 2019 International Conference on Computing, Electronics & Communications Engineering (iCCECE), London, United Kingdom, 2019, pp. 219-224, doi: 10.1109/iCCECE46942.2019.8941882.

[25] A. Yazar and H. Arslan, "A Waveform Parameter Assignment Framework for 6G With the Role of Machine Learning," in IEEE Open Journal of Vehicular Technology, vol. 1, pp. 156-172, 2020, doi: 10.1109/OJVT.2020.2992502.

[26] S. Aggarwal, N. Kumar and S. Tanwar, "Blockchain Envisioned UAV Communication Using 6G Networks: Open issues, Use Cases, and Future Directions," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.3020819.

[27] S. Ariyanti and M. Suryanegara, "Visible Light Communication (VLC) for 6G Technology: The Potency and Research Challenges," 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), London, United Kingdom, 2020, pp. 490-493, doi: 10.1109/WorldS450073.2020.9210383.

[28] K. B. Letaief, W. Chen, Y. Shi, J. Zhang and Y. A. Zhang, "The Roadmap to 6G: AI Empowered Wireless Networks," in IEEE Communications Magazine, vol. 57, no. 8, pp. 84-90, Aug. 2019, doi: 10.1109/MCOM.2019.1900271.

[29] S. J. Nawaz, S. K. Sharma, S. Wyne, M. N. Patwary and M. Asaduzzaman, "Quantum Machine Learning for 6G Communication Networks: State-of-the-Art and Vision for the Future," in IEEE Access, vol. 7, pp. 46317-46350, 2019, doi: 10.1109/ACCESS.2019.2909490.

[30] Y. Al-Eryani and E. Hossain, "The D-OMA Method for Massive Multiple Access in 6G: Performance, Security, and Challenges," in IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 92-99, Sept. 2019, doi: 10.1109/MVT.2019.2919279.

[31] E. Yaacoub and M. Alouini, "A Key 6G Challenge and Opportunity—Connecting the Base of the Pyramid: A Survey on Rural Connectivity," in Proceedings of the IEEE, vol. 108, no. 4, pp. 533-582, Apr. 2020, doi: 10.1109/JPROC.2020.2976703.

[32] Y. Chen, P. Zhu, G. He, X. Yan, H. Baligh and J. Wu, "From Connected People, Connected Things, to Connected Intelligence," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-7, doi: 10.1109/6GSUMMIT49458.2020.9083770.

[33] T. Hewa, G. Gür, A. Kalla, M. Ylianttila, A. Bracken and M. Liyanage, "The Role of Blockchain in 6G: Challenges, Opportunities and Research Directions," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083784.

[34] Y. Sun, J. Liu, J. Wang, Y. Cao and N. Kato, "When Machine Learning Meets Privacy in 6G: A Survey," in IEEE Communications Surveys & Tutorials, doi: 10.1109/COMST.2020.3011561.

[35] H. Chen, K. Tu, J. Li, S. Tang, T. Li and Z. Qing, "6G Wireless Communications: Security Technologies and Research Challenges," 2020 International Conference on Urban Engineering and Management Science (ICUEMS), Zhuhai, China, 2020, pp. 592-595, doi: 10.1109/ICUEMS50872.2020.00130.

[36] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang and D. Niyato, "Federated learning for 6G communications: Challenges, methods, and future directions," in China Communications, vol. 17, no. 9, pp. 105-118, Sept. 2020, doi: 10.23919/JCC.2020.09.009.

[37] J. Du, C. Jiang, J. WANG, Y. Ren and M. Debbah, "Machine Learning for 6G Wireless Networks: Carry-Forward-Enhanced Bandwidth, Massive Access, and Ultrareliable/Low Latency," in IEEE Vehicular Technology Magazine, doi: 10.1109/MVT.2020.3019650.

[38] C. -X. Wang, J. Huang, H. Wang, X. Gao, X. You and Y. Hao, "6G Wireless Channel Measurements and Models: Trends and Challenges," in IEEE Vehicular Technology Magazine, doi: 10.1109/MVT.2020.3018436.

[39] Y. Zhou, L. Liu, L. Wang, N. Hui, X. Cui, J. Wu, Y. Peng, Y. Qi, C. Xing,, "Service-aware 6G: An intelligent and open network based on the convergence of communication, computing and caching," in Digital Communications and Networks, vol. 6, no. 3, pp. 253-260, 2020.

[40] T. S. Rappaport et al., "Wireless Communications and Applications Above 100 GHz: Opportunities and Challenges for 6G and Beyond," in IEEE Access, vol. 7, pp. 78729-78757, 2019, doi: 10.1109/ACCESS.2019.2921522.

[41] X. Huang, J. A. Zhang, R. P. Liu, Y. J. Guo and L. Hanzo, "Airplane-Aided Integrated Networking for 6G Wireless: Will It Work?," in IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 84-91, Sept. 2019, doi: 10.1109/MVT.2019.2921244.

[42] L. Zhu, Z. Xiao, X. Xia and D. Oliver Wu, "Millimeter-Wave Communications With Non-Orthogonal Multiple Access for B5G/6G," in IEEE Access, vol. 7, pp. 116123-116132, 2019, doi: 10.1109/ACCESS.2019.2935169.

[43] T. Huang, W. Yang, J. Wu, J. Ma, X. Zhang and D. Zhang, "A Survey on Green 6G Network: Architecture and Technologies," in IEEE Access, vol. 7, pp. 175758-175768, 2019, doi: 10.1109/ACCESS.2019.2957648.

[44] F. Tang, Y. Kawamoto, N. Kato and J. Liu, "Future Intelligent and Secure Vehicular Network Toward 6G: Machine-Learning Approaches," in Proceedings of the IEEE, vol. 108, no. 2, pp. 292-307, Feb. 2020, doi: 10.1109/JPROC.2019.2954595.

[45] E. Basar, "Reconfigurable Intelligent Surface-Based Index Modulation: A New Beyond MIMO Paradigm for 6G," in IEEE Transactions on Communications, vol. 68, no. 5, pp. 3187-3196, May 2020, doi: 10.1109/TCOMM.2020.2971486.

[46] M. Yu, A. Tang, X. Wang and C. Han, "Joint Scheduling and Power Allocation for 6G Terahertz Mesh Networks," 2020 International Conference on Computing, Networking and Communications (ICNC), Big Island, HI, USA, 2020, pp. 631-635, doi: 10.1109/ICNC47757.2020.9049790.

[47] S. Zhang, H. Zhang and L. Song, "Beyond D2D: Full Dimension UAV-to-Everything Communications in 6G," in IEEE Transactions on Vehicular Technology, vol. 69, no. 6, pp. 6592-6602, June 2020, doi: 10.1109/TVT.2020.2984624.

[48] N. Kato, B. Mao, F. Tang, Y. Kawamoto and J. Liu, "Ten Challenges in Advancing Machine Learning Technologies toward 6G," in IEEE Wireless Communications, vol. 27, no. 3, pp. 96-103, June 2020, doi: 10.1109/MWC.001.1900476.

[49] X. Liu, T. Xu and I. Darwazeh, "Coexistence of Orthogonal and Non-orthogonal Multicarrier Signals in Beyond 5G Scenarios," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083780.

[50] S. Lagen, N. Patriciello and L. Giupponi, "Cellular and Wi-Fi in Unlicensed Spectrum: Competition leading to Convergence," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083786.

[51] N. H. Mahmood, H. Alves, O. A. López, M. Shehab, D. P. M. Osorio and M. Latva-Aho, "Six Key Features of Machine Type Communication in 6G," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083794.

[52] M. Katz and I. Ahmed, "Opportunities and Challenges for Visible Light Communications in 6G," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083805.

[53] C. Huang et al., "Holographic MIMO Surfaces for 6G Wireless Networks: Opportunities, Challenges, and Trends," in IEEE Wireless Communications, doi: 10.1109/MWC.001.1900534.

[54] C. Liu, W. Feng, Y. Chen, C. Wang and N. Ge, "Cell-Free Satellite-UAV Networks for 6G Wide-Area Internet of Things," in IEEE Journal on Selected Areas in Communications, doi: 10.1109/JSAC.2020.3018837.

[55] H. Hashida, Y. Kawamoto and N. Kato, "Intelligent Reflecting Surface Placement Optimization in Air-Ground Communication Networks Toward 6G," in IEEE Wireless Communications, doi: 10.1109/MWC.001.2000142.

[56] R. K. Saha, "Licensed Countrywide Full-Spectrum Allocation: A New Paradigm for Millimeter-Wave Mobile Systems in 5G/6G Era," in IEEE Access, vol. 8, pp. 166612-166629, 2020, doi: 10.1109/ACCESS.2020.3023342.

[57] J. A. Hodge, K. V. Mishra and A. I. Zaghloul, "Intelligent Time-Varying Metasurface Transceiver for Index Modulation in 6G Wireless Networks," in IEEE Antennas and Wireless Propagation Letters, doi: 10.1109/LAWP.2020.3025333.

[58] A. Celik, A. Chaaban, B. Shihada and M. -S. Alouini, "Topology Optimization for 6G Networks: A Network Information-Theoretic Approach," in IEEE Vehicular Technology Magazine, doi: 10.1109/MVT.2020.3017152.

[59] M. Kishk, A. Bader and M. -S. Alouini, "Aerial Base Station Deployment in 6G Cellular Networks Using Tethered Drones: The Mobility and Endurance Tradeoff," in IEEE Vehicular Technology Magazine, doi: 10.1109/MVT.2020.3017885.

[60] W. Lu et al., "SWIPT Cooperative Spectrum Sharing for 6G-Enabled Cognitive IoT Network," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.3026730.

[61] N. Chi, Y. Zhou, Y. Wei and F. Hu, "Visible light communication in 6G: Advances, challenges, and prospects," in IEEE Vehicular Technology Magazine, doi: 10.1109/MVT.2020.3017153.

[62] J. Liu, W. Liu, X. Hou, Y. Kishiyama, L. Chen and T. Asai, "Non-Orthogonal Waveform (NOW) for 5G Evolution and 6G," 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, London, United Kingdom, 2020, pp. 1-6, doi: 10.1109/PIMRC48278.2020.9217361.

[63] A. Vanelli-Coralli, A. Guidotti, T. Foggi, G. Colavolpe and G. Montorsi, "5G and Beyond 5G Non-Terrestrial Networks: trends and research challenges," 2020 IEEE 3rd 5G World Forum (5GWF), Bangalore, India, 2020, pp. 163-169, doi: 10.1109/5GWF49715.2020.9221119.

[64] L. Zhen, A. K. Bashir, K. Yu, Y. D. Al-Otaibi, C. H. Foh and P. Xiao, "Energy-Efficient Random Access for LEO Satellite-Assisted 6G Internet of Remote Things," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.3030856.

[65] S. Liao, J. Wu, J. Li and K. Konstantin, "Information-Centric Massive IoT based Ubiquitous Connected VR/AR in 6G: A Proposed Caching Consensus Approach," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.3030718.

[66] R. Alghamdi et al., "Intelligent Surfaces for 6G Wireless Networks: A Survey of Optimization and Performance Analysis Techniques," in IEEE Access, doi: 10.1109/ACCESS.2020.3031959.

[67] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao and K. Wu, "Artificial Intelligence-Enabled Intelligent 6G Networks," in IEEE Network, doi: 10.1109/MNET.011.2000195.

[68] H. Huang, S. Hu, T. Yang and C. W. Yuan, "Full Duplex Non-orthogonal Multiple Access with Layers-based Optimized Mobile Relays Subsets Algorithm in B5G/6G Ubiquitous Networks," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.3033553.

[69] Y. Liu, X. Wang, G. Boudreau, A. B. Sediq and H. Abouzeid, "A Multi-Dimensional Intelligent Multiple Access Technique for 5G Beyond and 6G Wireless Networks," in IEEE Transactions on Wireless Communications, doi: 10.1109/TWC.2020.3032631.

[70] A. Mukherjee, P. Goswami, M. A. Khan, L. Manman, L. Yang and P. Pillai, "Energy Efficient Resource Allocation strategy in Massive IoT for Industrial 6G Applications," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.3035608.

[71] I. Yildirim, A. Uyrus and E. Basar, "Modeling and Analysis of Reconfigurable Intelligent Surfaces for Indoor and Outdoor Applications in Future Wireless Networks," in IEEE Transactions on Communications, doi: 10.1109/TCOMM.2020.3035391.

[72] S. Wan, J. Hu, C. Chen, A. Jolfaei, S. Mumtaz and Q. Pei, "Fair-Hierarchical Scheduling for Diversified Services in Space, Air and Ground for 6G-Dense Internet of Things," in IEEE Transactions on Network Science and Engineering, doi: 10.1109/TNSE.2020.3035616.

[73] S. Han et al., "Artificial-Intelligence-Enabled Air Interface for 6G: Solutions, Challenges, and Standardization Impacts," in IEEE Communications Magazine, vol. 58, no. 10, pp. 73-79, Oct. 2020, doi: 10.1109/MCOM.001.2000218.

[74] M. Matinmikko-Blue, S. Yrjölä and P. Ahokangas, "Spectrum Management in the 6G Era: The Role of Regulation and Spectrum Sharing," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083851.

[75] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed and J. C. Zhang, "Artificial Intelligence-Enabled Cellular Networks: A Critical Path to Beyond-5G and 6G," in IEEE Wireless Communications, vol. 27, no. 2, pp. 212-217, Apr. 2020, doi: 10.1109/MWC.001.1900323.

[76] J. Zhu, M. Zhao, S. Zhang and W. Zhou, "Exploring the road to 6G: ABC — foundation for intelligent mobile networks," in China Communications, vol. 17, no. 6, pp. 51-67, June 2020, doi: 10.23919/JCC.2020.06.005.

[77] V. Ziegler, H. Viswanathan, H. Flinck, M. Hoffmann, V. Räisänen and K. Hätönen, "6G Architecture to Connect the Worlds," in IEEE Access, vol. 8, pp. 173508-173520, 2020, doi: 10.1109/ACCESS.2020.3025032.

[78] H. Li, F. Fang, Z. Ding, "Joint resource allocation for hybrid NOMA-assisted MEC in 6G networks," in Digital Communications and Networks, vol. 6, no. 3, pp. 241-252, 2020.

[79] H. Xu, P. V. Klaine, O. Onireti, B. Cao, M. Imran, L. Zhang, "Blockchain-enabled resource management and sharing for 6G communications," in Digital Communications and Networks, vol. 6, no. 3, pp. 261-269, 2020.

[80] Y. Wei, M. Peng, Y. Liu, "Intent-based networks for 6G: Insights and challenges," in Digital Communications and Networks, vol. 6, no. 3, pp. 270-280, 2020.

[81] M. Wang, T. Zhu, T. Zhang, J. Zhang, S. Yu, W. Zhou, "Security and privacy in 6G networks: New areas and new challenges," in Digital Communications and Networks, vol. 6, no. 3, pp. 281-291, 2020.

[82] Y. Fu, K. N. Doan, T. Q. S. Quek, "On recommendation-aware content caching for 6G: An artificial intelligence and optimization empowered paradigm," in Digital Communications and Networks, vol. 6, no. 3, pp. 304-311, 2020.

[83] Y. Chen, W. Liu, Z. Niu, Z. Feng, Q. Hu, T. Jiang, "Pervasive intelligent endogenous 6G wireless systems: Prospects, theories and key technologies," in Digital Communications and Networks, vol. 6, no. 3, pp. 312-320, 2020.

[84] International Telecommunication Union (ITU), "IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," ITU Publications, M.2083-0, 2015.

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

[86] A. Ghosh, A. Maeder, M. Baker and D. Chandramouli, "5G Evolution: A View on 5G Cellular Technology Beyond 3GPP Release 15," in IEEE Access, vol. 7, pp. 127639-127651, 2019, doi: 10.1109/ACCESS.2019.2939938.

[87] P. Kumari, S. A. Vorobyov and R. W. Heath, "Adaptive Virtual Waveform Design for Millimeter-Wave Joint Communication–Radar," in IEEE Transactions on Signal Processing, vol. 68, pp. 715-730, 2020, doi: 10.1109/TSP.2019.2956689.

[88] A. Yazar and H. Arslan, "Flexible Multi-Numerology Systems for 5G New Radio," in River Publishers Journal of Mobile Multimedia, vol. 14 no.4, pp. 367-394, 2018.

[89] A. B. Kihero, M. S. J. Solaija and H. Arslan, "Inter-Numerology Interference for Beyond 5G," in IEEE Access, vol. 7, pp. 146512-146523, 2019, doi: 10.1109/ACCESS.2019.2946084.

[90] Z. E. Ankarali, B. Peköz and H. Arslan, "Flexible Radio Access Beyond 5G: A Future Projection on Waveform, Numerology, and Frame Design Principles," in IEEE Access, vol. 5, pp. 18295-18309, 2017, doi: 10.1109/ACCESS.2017.2684783.

[91] M. M. Sahin and H. Arslan, "Multi-functional Coexistence of Radar-Sensing and Communication Waveforms," in IEEE Vehicular Technology Conference (VTC-Fall), Victoria, Canada, 2020.

[92] A. A. Sabah and H. Arslan, "NOMA for Multi-Numerology OFDM Systems," in Hindawi Wireless Communications and Mobile Computing, vol. 2018, pp. 1-9, 2018, doi:10.1155/2018/8514314.

[93] A. Tusha, S. Doğan and H. Arslan, "A Hybrid Downlink NOMA With OFDM and OFDM-IM for Beyond 5G Wireless Networks," in IEEE Signal Processing Letters, vol. 27, pp. 491-495, 2020, doi: 10.1109/LSP.2020.2979059.

[94] M. M. Şahin and H. Arslan, "Waveform-Domain NOMA: The Future of Multiple Access," 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 2020, pp. 1-6, doi: 10.1109/ICCWorkshops49005.2020.9145077.

[95] A. Maatouk, E. Çalışkan, M. Koca, M. Assaad, G. Gui and H. Sari, "Frequency-Domain NOMA With Two Sets of Orthogonal Signal Waveforms," in IEEE Communications Letters, vol. 22, no. 5, pp. 906-909, May 2018, doi: 10.1109/LCOMM.2018.2810118.

[96] M. B. Çelebi and H. Arslan, "Theoretical Analysis of the Co-Existence of LTE-A Signals and Design of an ML-SIC Receiver," in IEEE Transactions on Wireless Communications, vol. 14, no. 8, pp. 4626-4639, Aug. 2015, doi: 10.1109/TWC.2015.2424244.

[97] A. M. Jaradat, J. M. Hamamreh and H. Arslan, "Modulation Options for OFDM-Based Waveforms: Classification, Comparison, and Future Directions," in IEEE Access, vol. 7, pp. 17263-17278, 2019, doi: 10.1109/ACCESS.2019.2895958.

[98] P. Yang, Y. Xiao, Y. L. Guan, M. Di Renzo, S. Li and L. Hanzo, "Multidomain Index Modulation for Vehicular and Railway Communications: A Survey of Novel Techniques," in IEEE Vehicular Technology Magazine, vol. 13, no. 3, pp. 124-134, Sept. 2018, doi: 10.1109/MVT.2018.2814023.

[99] S. Doğan, A. Tusha and H. Arslan, "NOMA With Index Modulation for Uplink URLLC Through Grant-Free Access," in IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 6, pp. 1249-1257, Oct. 2019, doi: 10.1109/JSTSP.2019.2913981.

[100] A. Yazar and H. Arslan, "A Flexibility Metric and Optimization Methods for Mixed Numerologies in 5G and Beyond," in IEEE Access, vol. 6, pp. 3755-3764, 2018, doi: 10.1109/ACCESS.2018.2795752.

[101] A. F. Demir and H. Arslan, "Inter-Numerology Interference Management With Adaptive Guards: A Cross-Layer Approach," in IEEE Access, vol. 8, pp. 30378-30386, 2020, doi: 10.1109/ACCESS.2020.2972287.

[102] E. Memisoglu, A. B. Kihero, E. Basar and H. Arslan, "Guard Band Reduction for 5G and Beyond Multiple Numerologies," in IEEE Communications Letters, vol. 24, no. 3, pp. 644-647, Mar. 2020, doi: 10.1109/LCOMM.2019.2963311.

[103] A. Yazar and H. Arslan, "Reliability Enhancement in Multi-Numerology Based 5G New Radio Using INI-Aware Scheduling," in EURASIP Journal on Wireless Communications and Networking, vol. 2019 no. 110, pp. 1-14, 2019.

[104] M. J. Colella, J. N. Martin and F. Akyildiz, "The HALO network™," in IEEE Communications Magazine, vol. 38, no. 6, pp. 142-148, June 2000, doi: 10.1109/35.846086.

[105] T. Han, X. Ge, L. Wang, K. S. Kwak, Y. Han and X. Liu, "5G Converged Cell-Less Communications in Smart Cities," in IEEE Communications Magazine, vol. 55, no. 3, pp. 44-50, Mar. 2017, doi: 10.1109/MCOM.2017.1600256CM.

[106] L. Wang, T. Han, Q. Li, J. Yan, X. Liu and D. Deng, "Cell-Less Communications in 5G Vehicular Networks Based on Vehicle-Installed Access Points," in IEEE Wireless Communications, vol. 24, no. 6, pp. 64-71, Dec. 2017, doi: 10.1109/MWC.2017.1600401.

[107] M. S. J. Solaija, H. Salman, A. B. Kihero, M. I. Saglam, H. Arslan, "Generalized Coordinated Multipoint Framework for 5G and Beyond," arXiv:2008.06343 [eess.SP], Aug. 2020.

[108] M. S. Ali, E. Hossain, A. Al-Dweik and D. I. Kim, "Downlink Power Allocation for CoMP-NOMA in Multi-Cell Networks," in IEEE Transactions on Communications, vol. 66, no. 9, pp. 3982-3998, Sept. 2018, doi: 10.1109/TCOMM.2018.2831206.

[109] J. Ding and J. Cai, "Two-Side Coalitional Matching Approach for Joint MIMO-NOMA Clustering and BS Selection in Multi-Cell MIMO-NOMA Systems," in IEEE Transactions on Wireless Communications, vol. 19, no. 3, pp. 2006-2021, Mar. 2020, doi: 10.1109/TWC.2019.2961654.

[110] B. Han, J. Lianghai and H. D. Schotten, "Slice as an Evolutionary Service: Genetic Optimization for Inter-Slice Resource Management in 5G Networks," in IEEE Access, vol. 6, pp. 33137-33147, 2018, doi: 10.1109/ACCESS.2018.2846543.

[111] D. Sattar and A. Matrawy, "Optimal Slice Allocation in 5G Core Networks," in IEEE Networking Letters, vol. 1, no. 2, pp. 48-51, June 2019, doi: 10.1109/LNET.2019.2908351.

[112] D. A. Chekired, M. A. Togou, L. Khoukhi and A. Ksentini, "5G-Slicing-Enabled Scalable SDN Core Network: Toward an Ultra-Low Latency of Autonomous Driving Service," in IEEE Journal on Selected Areas in Communications, vol. 37, no. 8, pp. 1769-1782, Aug. 2019, doi: 10.1109/JSAC.2019.2927065.

[113] Q. Zhang, H. Sun, Z. Wei and Z. Feng, "Sensing and Communication Integrated System for Autonomous Driving Vehicles," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 2020, pp. 1278-1279, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162963.

[114] H. Turkmen, M. S. J. Solaija, H. M. Furqan, H. Arslan, "Generalized Radio Environment Monitoring for Next Generation Wireless Networks," arXiv:2008.06203 [eess.SP], Aug. 2020.

[115] W. Lee and I. F. Akyildiz, "Optimal spectrum sensing framework for cognitive radio networks," in IEEE Transactions on Wireless Communications, vol. 7, no. 10, pp. 3845-3857, October 2008, doi: 10.1109/T-WC.2008.070391.

[116] S. Yarkan and H. Arslan, "Exploiting location awareness toward improved wireless system design in cognitive radio," in IEEE Communications Magazine, vol. 46, no. 1, pp. 128-136, Jan. 2008, doi: 10.1109/MCOM.2008.4427241.

[117] H. F. Ates, S. M. Hashir, T. Baykas and B. K. Gunturk, "Path Loss Exponent and Shadowing Factor Prediction From Satellite Images Using Deep Learning," in IEEE Access, vol. 7, pp. 101366-101375, 2019, doi: 10.1109/ACCESS.2019.2931072.

[118] D. Sabella et al., "A flexible and reconfigurable 5G networking architecture based on context and content information," 2017 European Conference on Networks and Communications (EuCNC), Oulu, 2017, pp. 1-6, doi: 10.1109/EuCNC.2017.7980669.

[119] Z. Feng, Z. Fang, Z. Wei, X. Chen, Z. Quan and D. Ji, "Joint radar and communication: A survey," in China Communications, vol. 17, no. 1, pp. 1-27, Jan. 2020, doi: 10.23919/JCC.2020.01.001.

[120] G. Naik, J. Park, J. Ashdown, W. Lehr, "Next Generation Wi-Fi and 5G NR-U in the 6 GHz Bands: Opportunities & Challenges," arXiv:2006.16534, 2020.

[121] H. B. Yilmaz, T. Tugcu, F. Alagöz and S. Bayhan, "Radio environment map as enabler for practical cognitive radio networks," in IEEE Communications Magazine, vol. 51, no. 12, pp. 162-169, Dec. 2013, doi: 10.1109/MCOM.2013.6685772.

[122] A. Imran, A. Zoha and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," in IEEE Network, vol. 28, no. 6, pp. 27-33, Nov.-Dec. 2014, doi: 10.1109/MNET.2014.6963801.

[123] Q. Mao, F. Hu and Q. Hao, "Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 2595-2621, Fourthquarter 2018, doi: 10.1109/COMST.2018.2846401.

[124] C. Zhang, P. Patras and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2224-2287, thirdquarter 2019, doi: 10.1109/COMST.2019.2904897.

[125] N. C. Luong et al., "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3133-3174, Fourthquarter 2019, doi: 10.1109/COMST.2019.2916583.

[126] Y. Sun, M. Peng, Y. Zhou, Y. Huang and S. Mao, "Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3072-3108, Fourthquarter 2019, doi: 10.1109/COMST.2019.2924243.

[127] M. Chen, U. Challita, W. Saad, C. Yin and M. Debbah, "Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3039-3071, Fourthquarter 2019, doi: 10.1109/COMST.2019.2926625.

[128] J. Wang, C. Jiang, H. Zhang, Y. Ren, K. Chen and L. Hanzo, "Thirty Years of Machine Learning: The Road to Pareto-Optimal Wireless Networks," in IEEE Communications Surveys & Tutorials, doi: 10.1109/COMST.2020.2965856.

[129] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan and X. Chen, "Convergence of Edge Computing and Deep Learning: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 869-904, Secondquarter 2020, doi: 10.1109/COMST.2020.2970550.

[130] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. Chen and L. Hanzo, "Machine Learning Paradigms for Next-Generation Wireless Networks," in IEEE Wireless Communications, vol. 24, no. 2, pp. 98-105, Apr. 2017, doi: 10.1109/MWC.2016.1500356WC.

[131] R. Li et al., "Intelligent 5G: When Cellular Networks Meet Artificial Intelligence," in IEEE Wireless Communications, vol. 24, no. 5, pp. 175-183, Oct. 2017, doi: 10.1109/MWC.2017.1600304WC.

[132] Z. Chang, L. Lei, Z. Zhou, S. Mao and T. Ristaniemi, "Learn to Cache: Machine Learning for Network Edge Caching in the Big Data Era," in IEEE Wireless Communications, vol. 25, no. 3, pp. 28-35, June 2018, doi: 10.1109/MWC.2018.1700317.

[133] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. Alouini and R. Zhang, "Wireless Communications Through Reconfigurable Intelligent Surfaces," in IEEE Access, vol. 7, pp. 116753-116773, 2019, doi: 10.1109/ACCESS.2019.2935192.

[134] M. D. Renzo et al., "Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How it Works, State of Research, and Road Ahead," in IEEE Journal on Selected Areas in Communications, doi: 10.1109/JSAC.2020.3007211.

[135] C. Xu, L. Yang and P. Zhang, "Practical Backscatter Communication Systems for Battery-Free Internet of Things: A Tutorial and Survey of Recent Research," in IEEE Signal Processing Magazine, vol. 35, no. 5, pp. 16-27, Sept. 2018, doi: 10.1109/MSP.2018.2848361.

[136] N. Van Huynh, D. T. Hoang, X. Lu, D. Niyato, P. Wang and D. I. Kim, "Ambient Backscatter Communications: A Contemporary Survey," in IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 2889-2922, Fourthquarter 2018, doi: 10.1109/COMST.2018.2841964.

[137] R. Long, Y. Liang, H. Guo, G. Yang and R. Zhang, "Symbiotic Radio: A New Communication Paradigm for Passive Internet of Things," in IEEE Internet of Things Journal, vol. 7, no. 2, pp. 1350-1363, Feb. 2020, doi: 10.1109/JIOT.2019.2954678.

[138] Q. Zhang, Y. Liang and H. V. Poor, "Symbiotic Radio: A New Application of Large Intelligent Surface/Antennas (LISA)," 2020 IEEE Wireless Communications and Networking Conference (WCNC), Seoul, Korea (South), 2020, pp. 1-6, doi: 10.1109/WCNC45663.2020.9120455.

[139] S. Buzzi, C. I, T. E. Klein, H. V. Poor, C. Yang and A. Zappone, "A Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 4, pp. 697-709, Apr. 2016, doi: 10.1109/JSAC.2016.2550338.

[140] J. Huang, C. Xing and C. Wang, "Simultaneous Wireless Information and Power Transfer: Technologies, Applications, and Research Challenges," in IEEE Communications Magazine, vol. 55, no. 11, pp. 26-32, Nov. 2017, doi: 10.1109/MCOM.2017.1600806.

[141] T. D. Ponnimbaduge Perera, D. N. K. Jayakody, S. K. Sharma, S. Chatzinotas and J. Li, "Simultaneous Wireless Information and Power Transfer (SWIPT): Recent Advances and Future Challenges,"

in IEEE Communications Surveys & Tutorials, vol. 20, no. 1, pp. 264-302, Firstquarter 2018, doi: 10.1109/COMST.2017.2783901.

[142] J. M. Hamamreh, H. M. Furqan and H. Arslan, "Classifications and Applications of Physical Layer Security Techniques for Confidentiality: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1773-1828, Secondquarter 2019, doi: 10.1109/COMST.2018.2878035.

[143] J. M. Hamamreh, M. Yusuf, T. Baykas and H. Arslan, "Cross MAC/PHY layer security design using ARQ with MRC and adaptive modulation," 2016 IEEE Wireless Communications and Networking Conference, Doha, 2016, pp. 1-7, doi: 10.1109/WCNC.2016.7564987.

[144] S. Dwivedi, M. Zoli, A. N. Barreto, P. Sen and G. Fettweis, "Secure Joint Communications and Sensing using Chirp Modulation," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, 2020, pp. 1-5, doi: 10.1109/6GSUMMIT49458.2020.9083884.

[145] Y. Hu and R. Zhang, "A Spatiotemporal Approach for Secure Crowdsourced Radio Environment Map Construction," in IEEE/ACM Transactions on Networking, doi: 10.1109/TNET.2020.2992939.

AUTHORS

A. Yazar received his B.Sc. degree in electrical engineering from Eskisehir Osmangazi University, Eskisehir, Turkey in 2011, M.Sc. degree in electrical engineering from Bilkent University, Ankara, Turkey in 2013, and Ph.D. degree in electrical engineering from Istanbul Medipol University, Istanbul, Turkey in 2020. He is currently general coordinator as a member of the Communications, Signal Processing, and Networking Center (CoSiNC) at Istanbul Medipol University. His current research interests are flexible waveform design, radio resource management techniques, and the role of machine learning in wireless communications systems.

S. Doğan Tusha received the B.Sc. degree in electronics and telecommunication engineering from Kocaeli University, Kocaeli, Turkey, in 2015, and the Ph.D. degree in electrical and electronics engineering from Istanbul Medipol University, Istanbul, Turkey, in 2020. She is currently a post-doctoral researcher in the Communications, Signal Processing, and Networking Center (CoSiNC) at Istanbul Medipol University, Istanbul, Turkey. Her research interests include index modulation, millimeter-wave frequency bands, nonorthogonal multiple accessing (NOMA), and random access techniques for next-generation wireless networks.

H. Arslan (IEEE Fellow, IEEE Distinguished Lecturer) received his BS degree from the Middle East Technical University (METU), Ankara, Turkey in 1992; his MS and Ph.D. degrees were received respectively in 1994 and 1998 from Southern Methodist University (SMU), Dallas, TX. From January 1998 to August 2002, he was with the research group of Ericsson, where he was involved with several projects related to 2G and 3G wireless communication systems. Since August 2002, he has been with the Electrical Engineering Department, at the University of South Florida, where he is a Professor. In December 2013, he joined Istanbul Medipol University to found the Engineering College, where he has worked as the Dean of the School of Engineering and Natural Sciences. In addition, he has worked as a part-time consultant for various companies and institutions including Anritsu Company and The Scientific and Technological Research Council of Turkey.

Dr. Arslan conducts research in wireless systems, with emphasis on the physical and medium access layers of communications. His current research interests are on 5G and beyond radio access technologies, physical layer security, interference management (avoidance, awareness, and cancellation), cognitive radio, multi-carrier wireless technologies (beyond OFDM), dynamic spectrum access, coexistence issues, non-terrestial communications (High Altitude Platforms), joint radar (sensing) and communication designs. Dr. Arslan has been collaborating extensively with key national and international industrial partners and his research has generated significant interest in companies such as InterDigital, Anritsu, NTT DoCoMo, Raytheon, Honeywell, Keysight technologies. Collaborations and feedback from industry partners has significantly influenced his research. In addition to his research activities, Dr. Arslan has also contributed to wireless communication education. He has integrated the outcomes of his research into education which lead him to develop a number of courses at the University of South Florida. He has developed a unique "Wireless Systems Laboratory" course (funded by the National Science Foundation and Keysight technologies) where he was able to teach not only the theory but also the practical aspects of wireless communication system with the most contemporary test and measurement equipment.

Dr. Arslan has served as general chair, technical program committee chair, session and symposium organizer, workshop chair, and technical program committee member in several IEEE conferences. He is currently a member of the editorial board for the IEEE Surveys and Tutorials and the Sensors Journal. He has also served as a member of the editorial board for the IEEE Transactions on Communications, the IEEE Transactions on Cognitive Communications and Networking (TCCN), and several other scholarly journals by Elsevier, Hindawi, and Wiley Publishing.

ON THE EVOLUTION OF INFRASTRUCTURE SHARING IN MOBILE NETWORKS: A SURVEY

Lorela Cano[1], Antonio Capone[2], Brunilde Sansò[3]
[1,2]Politecnico di Milano, Milan, Piazza Leonardo da Vinci, 32, 20133, Milano MI, Italy, [3]Polytechnique Montréal, 2500 Chemin de Polytechnique, Montréal, QC H3T 1J4, Canada,

NOTE: Corresponding author: Lorela Cano (lorela.cano@polimi.it)

Abstract – *Infrastructure sharing for mobile networks has been a prolific research topic for more than three decades now. The key driver for Mobile Network Operators to share their network infrastructure is cost reduction. Spectrum sharing is often studied alongside infrastructure sharing although on its own it is a vast research topic outside the scope of this survey. Instead, in this survey we aim to provide a complete picture of infrastructure sharing both over time and in terms of research branches that have stemmed from it such as performance evaluation, resource management etc. We also put an emphasis on the relation between infrastructure sharing and the decoupling of infrastructure from services, wireless network virtualization and multi-tenancy in 5G networks. Such a relation reflects the evolution of infrastructure sharing over time and how it has become a commercial reality in the context of 5G.*

Keywords – 5G, infrastructure sharing, mobile networks, multi-tenancy, spectrum sharing, wireless network virtualization

1. INTRODUCTION

Infrastructure sharing in mobile networks is a multi-faceted problem involving not only academic and industrial research entities but also national and international regulatory entities [51, 52, 65, 66], standardization bodies [1–4] and vendors [45,107]. In essence, infrastructure sharing in mobile networks is the shared use of existing or jointly deployed network infrastructure among multiple Mobile Network Operators (MNOs).

Based on which network elements (nodes) MNOs agree to/can share, there are two main types of sharing: *passive* and *active*, the latter comprising the former. Passive sharing (also referred to as site sharing or co-location [49]) implies the sharing of the site physical space and of the non-active elements on the site (such as shelter, cabinet, mast, etc. [49,104]). Instead, active sharing extends to active elements of the Radio Access Network (RAN) (such as antennas, Base Transceiver Stations/Base Station Controller for 2G, Node B/Radio Network Controller for 3G, eNode B for 4G, and gNodeB for 5G) and part of the core nodes (in fact, core node elements related to user billing and accounting are not shared).

The phenomenon of infrastructure sharing has disrupted the business model of a *conventional* MNO, that is, an MNO which is by itself responsible for (i) purchasing a spectrum license, (ii) deploying and managing the network infrastructure, (iii) tailoring services for their subscribers (e.g., voice, data, etc.) and (iv) handling their billing and accounting. The main reason for MNOs to share infrastructure is to divide the infrastructure cost among them and hence make their business more profitable. In these lines, infrastructure sharing has accompanied the technology migrations from 2G to 3G and from 3G to 4G due to the high upfront cost met by MNOs during these migrations. In turn, in 5G networks, infrastructure sharing, besides from being a means for cost-reduction, it is also an important pillar of the 5G architecture. Another paradigm strongly linked to infrastructure sharing is spectrum sharing. The need for spectrum sharing comes from spectrum being an intrinsically scarce resource, even more so in the context of 5G, given its target throughputs. However, spectrum sharing alone is a really vast research topic and will be outside the scope of this survey unless combined with infrastructure sharing.

What's more, in this paper we will also address some literature on Wireless Network Virtualization (WNV) [91] and network slicing (enabling multi-tenancy) in the context of 5G [6], since both are based on infrastructure and spectrum sharing. Conversely, WNV and network slicing can be seen as enablers for infrastructure and spectrum sharing. Besides, another concept closely related to infrastructure and spectrum sharing is that of the *decoupling of infrastructure from services*, which was envisioned by some of the early literature on infrastructure sharing (see Section 2). The concept has been further carried out in the context of WNV and then in the context of network slicing. In fact, the different research efforts on introducing Software-Defined Networking (SDN), virtualization in general and Network Functions Virtualization (NFV) in particular into mobile networks seem to have converged into the 5G

architecture as enablers for network slicing.

Infrastructure sharing in this broader sense has been a very prolific research topic over the last three decades. Samdanis *et al.* in [125] provide a compelling analysis of the path from infrastructure sharing to multi-tenancy. However, to the best of our knowledge, our survey is the first[1] comprehensive study on how the infrastructure sharing topic in mobile networks has evolved over time, i.e., with the advent of the different mobile network generations, and which research branches have spurred from this topic. Reviewing this evolution is particularly important now that networking slicing is being introduced in 5G (from release 16 onwards) and operators are looking for models for sharing infrastructure costs and to invest more in new services and applications, collaborating with different players of vertical industrial sectors. Moreover, it is becoming clear to the telecommunications industry sector that some form of infrastructure sharing will be the common basis on which networks will be deployed in different countries and services will evolve and diversify, going beyond 5G and preparing the ground for the next generation.

This survey is organized in the following fashion. Due to the change in the nature of problems studied over time, we first make a broad chronological classification of the literature into *early* works and *recent and up-to-date* works. For the latter, we further identify several research branches/categories. The overall picture of our classification is depicted in Fig. 1. An overview of the early works on the topic is provided in Section 2. Further, in Section 3, we focus on the more recent and up-to-date works. Then in Section 4 we make a critical discussion of the research area related to infrastructure sharing and provide an outlook of future research directions. Finally, conclusions are drawn in Section 5.
For readers' ease, in Table 1 we provide the definitions of the acronyms and abbreviations used in the paper.

3GPP	3rd Generation Partnership Project
BS	Base Station
C-RAN	Cloud Radio Access Network
DCS	Digital Cellular System
EDGE	Enhanced Data rates for GSM Evolution
GERAN	GSM EDGE RAN
GSM	Global System for Mobile communications
IaaS	Infrastructure as a Service
InP	Infrastructure Provider
IoT	Internet of Things
IP	Integer Programming
JV	Joint Venture
MLFG	Multi-Leader-Follower Game
mmWave	millimiter Wave
MNO	Mobile Network Operator
MVNO	Mobile Virtual Network Operator
NaaS	Network as a Service
NFV	Network Functions Virtualization
NSP	Network Service Provider
OTT	Over The Top
PRB	Physical Resource Block
QoS	Quality of Service
RAN	Radio Access Network
RRH	Remote Radio Head
SaaS	Software as a Service
SDN	Software-Defined Networking
SINR	Signal-to-Interference-plus-Noise Ratio
SLA	Service Level Agreement
SP	Service Provider
UMTS	Universal Mobile Telecommunications System
VNO	Virtual Network Operator
VO	Virtual Operator
W-CDMA	Wideband Code Division Multiple Access
WNV	Wireless Network Virtualization
xG	xth mobile network Generation

Table 1 – Definitions of acronyms and abbreviations

2. EARLY WORKS

[16,49,55,70,115,118,142] are among the earliest articles on infrastructure sharing (combined at times also with spectrum sharing). With the exception of [70][2], these articles have tended to:

- address technical issues of different sharing alternatives,
- assess the financial profitability through techno-economic approaches,
- state regulatory standpoints and provide guidelines for the latter and
- conceive new paradigms for the mobile market.

In [118], which dates back to 1994, Ramsdale states that national roaming[3] is part of the specifications of the Digital Cellular System at 1800 MHz (DCS 1800), unlike the Global System for Mobile Communications at 900 MHz (GSM 900), which supported international roaming only. National roaming was introduced in the DCS 1800 to improve coverage due to smaller cell sizes at 1800 MHz (as opposed to 900 MHz).

Instead, the work in [55] shows the positive impact of infrastructure sharing in financial terms for the Universal Mobile Telecommunications Systems (UMTS), especially for lowly populated areas in which network deployment is dictated by coverage instead of capacity.

[1] This survey is based on the PhD thesis of Lorela Cano [24].

[2] The study in [70] is an early work on the problem of scheduling users of multiple operators arising from the case when a 3G, facility-based MNO hosts several Mobile Virtual Network Operators (MVNOs): the authors propose a non-pre-emptive priority queuing model for circuit-switched traffic applied through an admission control scheme.

[3] National roaming is an infrastructure sharing alternative that allows users of an operator which does not provide coverage in certain areas of a country to be served by the network of another operator of that country covering such areas.

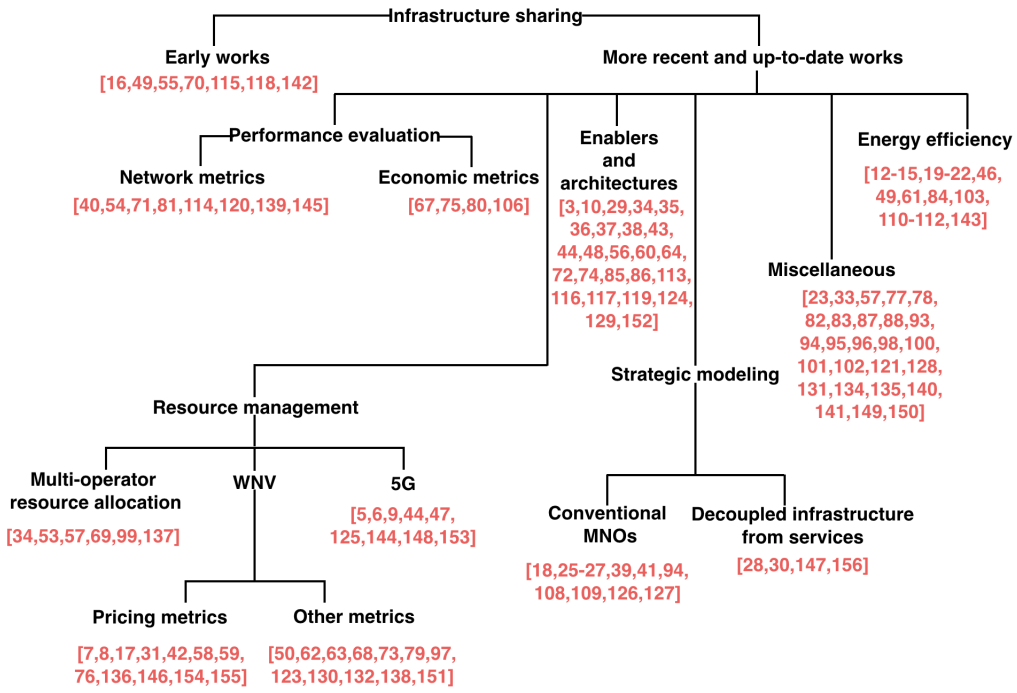


Fig. 1 – Literature classification map

In turn, MVNOs are suggested as a means to monetize spare resources of an MNO.

Park *et al.* in [115] discuss issues faced by MNOs worldwide when deploying Wideband Code Division Multiple Access (W-CDMA) and propose spectrum trading and infrastructure sharing as means to accelerate the deployment of W-CDMA. However, they emphasize that such means should be cautiously treated by regulators.

The study in [49] proposes a spreadsheet-based financial model to estimate the economic profitability of multiple sharing alternatives and shows that cost can be further reduced if the network operations are outsourced or a joint venture is created.

The authors in [142] discuss technical aspects concerning the infrastructure sharing alternatives at the time; they also anticipate two crucial paradigms: (i) dynamic spectrum trading and (ii) the decoupling of the network infrastructure from services, enabled by infrastructure sharing. It is worth noticing that both these paradigms are ongoing research topics even nowadays. Similarly, according to [16], the advantages of network sharing go beyond cost reduction: based on the product life cycle model, the authors suggest that, under an appropriate regulatory framework, network sharing can steer the monolithic mobile networks industry toward the decoupling of the network infrastructure from services for end users. In other words, based on [142] and [16] infrastructure sharing would lead to new stakeholders such as network/infrastructure providers (InPs) and service providers (SPs) which were expected to emerge in the mobile market, the former being responsible for network planning, deployment and management while the latter for dealing only with the development of novel services (possibly specialized and targeting specific market segments [16]).

When analyzing these early works on infrastructure sharing in mobile networks, we have to consider the specific technical limitations that have constrained the approaches for 2G, 3G and partially 4G network, to some aspects of the problem only. In particular, being the spectrum one of the most important assets of a mobile network and being it easily shared among physically separated networks, it has been widely studied considering the locality of interference generated and the re-usability in different geographical areas.

As far as the physical infrastructure is concerned, the first works on sharing have focused on economic aspects and market regulation policies associated to the introduction of MVNOs. However, the main limitation of these approaches was due to the mobile technology that prevented a significant service and performance differentiation among users of MVNOs and MNOs. Therefore, sharing policies had to be based on other objectives such as cooperative coverage of low population areas and cost sharing of radio towers.

As mentioned above, the key aspect that we take from these works is the decoupling of network services from the infrastructure that provides them. Only recently, however, this concept has become fully exploitable thanks to the network virtualization technologies that allow a fine grain differentiation of the network behavior with respect to different applications and groups of

users. This radical change of the technology scenario, mainly due to the new architectural solutions and service definition of 5G, did not cancel the main issues analyzed by the early works on infrastructure sharing, such as the economic aspects of cost sharing and their relation with resource allocation or partitioning.

3. MORE RECENT AND UP-TO-DATE WORKS

In the more recent and up-to-date literature, there is a tendency to address *specific problems*, e.g., the problem of resource management, for *specific sharing scenarios*, e.g., infrastructure and spectrum sharing at the RAN. There are at least two ways to go about the classification of this literature, one being problem-centric and the other being methodology-centric. We have opted for the first one in order to highlight the fact that there are many aspects to infrastructure sharing and hence provide the reader with the bigger picture on the topic. Methodology details are discussed only when deemed necessary.

Under the problem-centric classification, we have identified the following research branches/categories for the revised articles: (i) *performance evaluation*, (ii) *resource management*, (iii) *enablers* and *architectures*, (iv) *energy efficiency*, (v) *strategic modeling* and (vi) *miscellaneous*.

It is worth pointing out that some of the articles may fit in more than one category, but for each such article, we have opted for a single category, the one we believe is the most salient.

3.1 Performance evaluation

Several authors have addressed the gains of particular infrastructure and/or spectrum sharing scenarios in terms of *network performance metrics*, such as throughput, coverage probability etc. (see e.g., [71, 114, 139, 145]) and/or *economic* ones such as CAPEX/OPEX reduction (see e.g., [67, 75, 80, 106]). The common approach is to benchmark such scenarios against the baseline case when no sharing takes place and the involved MNOs build individual networks instead. Methodology-wise, both theoretical, mainly stochastic geometry analysis (see e.g., [54,71,81,145]), and simulation approaches (see e.g., [40,114,120]) have been adopted. For instance, the work in [114] proposes a virtualized architecture to enable two types of spectrum sharing other than the classical one and capacity sharing (national roaming) and compares the different sharing alternatives with no sharing case. The performance metrics considered in [114] are the sector load and packet drop probability.

The authors in [40] analyse how the time and space correlation of the MNO individual traffic loads impacts the gains of infrastructure sharing in the case when MNOs decide to pool together their respective networks. Kibilda *et al.* [81] resort to stochastic geometry to calculate the gains of sharing for the cases of infrastructure and/or spectrum pooling. Their key finding is that the infrastructure and spectrum sharing gains do not sum up when combined since full sharing (infrastructure+spectrum) introduces a trade-off between the data rate and coverage.

As 5G is expected to make use of the millimeter wave (mmWave) frequencies [11], the gains of infrastructure and/or spectrum in these frequencies have become the object of several recent articles. For instance, Gupta *et al.* in [54] provide a stochastic geometry-based theoretical analysis on the gains of spectrum sharing using a simplified antenna and channel model for the mmWave frequency range. In particular, in [54] it is shown how narrow beams are key for spectrum sharing in the mmWaves. A very similar investigation to [81] is carried by Rebato *et al.* in [120] for mmWaves; the authors highlight the impact of the channel model accuracy when carrying out a quantitative analysis of the sharing gains. The recent work in [71] also addresses infrastructure and spectrum sharing at mmWaves and it resorts to stochastic geometry to derive the probability of Signal-to-Interference-plus-Noise Ratio (SINR) coverage as a performance metric.

In Table 2 we provide a visual overview of the classification of the different articles that were included in the performance evaluation category. As can be seen from the table, methodologically-wise, the authors use mainly stochastic geometry, simulation and optimization. The other method found was empirical analysis. With respect to the type of measures used to evaluate performance, we can see that network measures are fairly diverse: even though most work in this category deals with physical layer measures such as SINR, networking measures such as traffic load, sector overload or packet drop probability are also considered. Not surprisingly, less diversity can be found in the economic measures' category.

3.2 Resource management

Problems of resource management arise whenever infrastructure sharing is combined with spectrum sharing, as users of multiple MNOs/MVNOs have to be assigned resources from a shared pool.

Several studies ([34, 53, 99, 137]) have proposed algorithms for a multi-operator scheduler, namely when users of multiple MNOs have to be scheduled in the finite resources available in a shared Base Station (BS). Assuming MNOs agree *a priori* on the resource shares, i.e., how to split the available BS resources among them, the work in [137] adopts the concept of Generalized Processor Sharing for a multi-operator scheduler. For the same setting, Malanchini *et al.* [99] explore the trade-

Classification \ Methodology		Stochastic Geometry	Simulation	Optimization	Other
Network Measures	traffic load		[40]	[67]	
	SINR	[54], [71]	[120]	[80]	
	SINR coverage probability	[71], [81]			
	Throughput			[67]	
	User rate	[81], [145]			
	Sector overload		[114]		
	Packet drop probability		[114]		
Economic Measures	Revenue		[75]		
	CAPEX		[75]		[106]
	OPEX		[75]		
	Miscellaneous		[75]	[67], [80]	

Table 2 – Performance evaluation classification

off between satisfying the resource shares and improving the overall (system) spectral efficiency when the agreed resource shares are violated in a controlled fashion. The work in [53] considers a global scheduler taking decisions for clusters of BSs and therefore scheduling users of multiple MNOs over a 3D time-frequency-space resource grid. In [53] scheduling is performed with the objective of maximizing the overall system utility. The authors in [34] propose a BS virtualization scheme which performs scheduling in two levels, namely, among MNOs, and for each MNO, among its user flows. Hew *et al.* in [57] consider a network shared by multiple MNOs, each of them serving both a set of end users and a set of MVNOs. In this context, the problem of resource allocation is tackled in two steps: first, the resource sharing among MNOs, and then the resource sharing among the users and the MVNOs of each MNO, where the resource sharing at each step is modeled as a bargaining problem. The study in [69] suggests an algorithm that fairly allocates the shared radio resources among MNOs. In [105] the authors propose Remote Radio Head (RRH) assignment algorithms for an SDN-based Cloud Radio Access Network (C-RAN) shared by multiple MNOs.

Concerning WNV, the problem of resource management is crucial in the interaction between an InP and its SPs. In the context of this paper, an InP is an entity which is responsible for the infrastructure deployment, management and operation and does not serve end users directly whereas an SP is an entity which does not have any resources of its own but purchases or rents resources from an InP to provision services for its end users. It is worth noticing that the terminology concerning the SP varies across different articles: such an entity is also referred to as a Virtual Operator (VO), a Virtual Network Operator (VNO) or a Mobile Virtual Network Operator (MVNO). Also notice that the conventional MVNO obtains resources from an MNO which serves end users of its own, unlike the InP. The key difference lies in the fact that a conventional MVNO competes with its MNO, while there is no such competition between an InP and its SPs/VOs/VNOs/MVNOs. In these lines, some articles tend to "misuse" the term InP when they consider the InP to provide services also to end users. Additionally an InP is also referred to as a Network Service Provider (NSP). Moreover, the work in [91] envisions three different types of stakeholders in line with the ones in the cloud computing domain, i.e., the InP providing Infrastructure as a Service (IaaS), the MVNO providing Network as a Service (NaaS) and the SP providing Software as a Service (SaaS). For instance, in [89] the authors address a scenario in which there are multiple InPs, a single MVNO and multiple SPs where the MVNO acts as a reseller of resources from InPs to SPs. It should also be noted that the terms *slicing* and *slice* are also misused in some articles in non-5G contexts, in the sense that, such articles do not consider problem instances that account for 5G service requirements.

There is a large body of literature on resource managements concerning InPs and SPs in the context of WNV. The vast majority of articles in this literature considers a single InP and multiple SPs (see e.g., [7, 8, 17, 31, 42, 50, 58, 59, 62, 63, 68, 73, 76, 79, 97, 123, 130, 132, 133, 136, 138, 146, 151, 154, 155]). However, there are exceptions: e.g., the work in [32] considers a single InP and a single VNO which serves multiple users through an SDN-based virtualized network provided by the InP. The VNO faces the problem of scheduling its users, each characterized by a maximum delay over a finite time period, through resources rented by the InP with the objective of minimizing the payments made to the InP for the rented resources. There are also articles which consider both multiple InPs and multiple SPs (and few other variations with multiple InPs) which however are more pertinent to Section 3.5 hence discussed therein.

As for the literature on a single InP and multiple SPs,

it can be broadly classified into two groups based on whether the resource management is driven by pricing ([7, 8, 17, 31, 42, 58, 59, 76, 136, 146, 154, 155]) or not ([50, 62, 63, 68, 73, 79, 97, 123, 130, 132, 133, 138, 151]). For instance, Ho *et al.* in [58] consider the case when there is a single InP serving multiple MVNOs, each characterized by a fixed number of users and a Service Level Agreement (SLA) given in terms of a minimum resource requirement and a maximum aggregate rate (over all its users). The InP has to decide how to price and allocate its available BS resources among all users of all MVNOs so as to maximize its profit while guaranteeing the SLA of each MVNO. In this work MVNOs are also self-interested as the goal of each MVNO is to maximize its own profit given by the difference between the total rate obtained from resources allocated by the InP and their cost. The problem is then modeled as a one-leader multi-follower variant of the Stackelberg game with the InP being the leader and each MVNO being a follower. Instead, Kamel *et al.* in [73] address a scheduling problem over one time frame which is modeled through mathematical programming. In details, there is a single InP and a set of VOs, each having a fixed number of users and a minimum resource requirement (total Physical Resource Blocks (PRBs) over the time frame). The InP has to decide to which user to assign each PRB and the amount of power to allocate to each PRB so as to maximize the total rate over the time frame while satisfying the maximum power constraint, the minimum resource requirement of each VO and a VO-specific proportional fairness constraint for cell-center and cell-edge users.

In 5G, the problem of resource management reemerges in the context of multi-tenancy and its enabler, network slicing ([5, 6]). Tenants (such as MVNOs, Over The Top (OTT) providers and vertical industries) have distinct requirements to support their services which have to be translated into appropriate network resources. It is worth noting that network slicing does not involve only the RAN segment but it can be end-to-end. However, the problem of resource management at the RAN segment has brought about a significant amount of attention from the research community due to the intrinsically complex nature of the radio (wireless) access. For instance, the authors in [125] propose the "5G Network Slice Broker", a centralized scheduler based on the 3rd Generation Partnership Project (3GPP) specifications for network sharing. The proposed scheduler has a global view of the shared network and applies admission control and resource allocation, translating the tenants' request, with given SLAs, into available network resources. Other examples on resource management at the RAN in the context of multi-tenancy/network slicing are given in [9, 44, 47, 122, 144, 148, 153].

Summarizing, the Resource Management category is a very rich part of the infrastructure sharing literature. Within this category, we have identified three subcategories:

1. partitioning and allocation of resources shared by multiple operators,
2. the literature on Wireless Network Virtualization that mostly deals with the sharing of resources between a single InP and multiple SPs; this subcategory, can be further subdivided into:
   (a) articles that base their modelling on pricing issues and
   (b) articles that base their modelling on other issues, such as performance metrics, and
3. a large body of 5G literature that deals with resource management and network slicing.

3.3 Enablers and architectures

Although the different alternatives for infrastructure and spectrum sharing can be financially attractive for MNOs, they where not always supported by the 3GPP specifications; in fact, while a basic type of network sharing was supported as of Release 5, there was no support for more involved network sharing scenarios for the 3GPP GSM EDGE[4] RAN (GERAN) prior to Release 10 ([3]).

Standardization apart, the research community has largely contributed on the topics of enabling network sharing, e.g., through novel architectures. While passive sharing (i.e., site/tower sharing) is the simplest network sharing alternative to implement, the different types of active sharing demand architectural changes in mobile networks e.g., to guarantee the isolation of the involved MNOs in terms of their private information in order to avoid harming competition, or they demand changes at the protocol stack level to implement the novel resource management algorithms etc. According to [64], radio resource management should be delegated to a third party provider to ensure isolation and therefore not to interfere with competition. In [56] the authors introduce AppRAN which relies on a centralized scheduler to perform application-level resource allocation for a shared RAN.

In particular, different flavors of virtualization have been widely considered by the research community as candidate enablers for network sharing. For instance, the virtualized network architecture proposed in [60] can support network sharing. Other papers that resort to virtualization are e.g., [10, 34, 72, 117, 152]. In particular, the authors of [38] and of [43] propose the "Network without Borders", namely the virtualized pool of (heterogeneous) wireless resources for which infrastructure and spectrum pooling are essential. Costanzo *et al.* in [37] suggest an architecture for 4G RAN sharing based on SDN and NFV.

[4]Enhanced Data rates for GSM Evolution

In the context of enabling network slicing in 5G networks, there is a myriad of papers that propose architectures or test prototypes based on (i) NFV and/or SDN (see e.g, [35, 36, 85, 86, 113, 119]), (ii) changes to the RAN protocol stack (see e.g., [48, 116, 124]), or (iii) using features of the new 5G radio ([44]) etc. In particular, the work in [29] proposes an architecture to support network slicing in ultra-dense networks, the one in [74] presents an architecture that supports Internet Of Things (IoT) slices whereas the one in [129] dwells on combining 3GPP specifications for 5G with NFV.

3.4 Energy efficiency

Infrastructure and spectrum sharing allow to reduce the energy-consumption OPEX cost particularly in cases when the aggregated network resources (infrastructure and/or spectrum) are redundant. For instance, in rural areas where capacity is not an issue, MNOs can decommission a subset of the aggregated BSs and/or operate at a subset of the aggregated frequency carriers [49], which reduces the energy consumption and (indirectly) the environmental impact. In these lines, since MNOs dimension their networks based on the peak-load traffic predictions, there is intrinsically resource redundancy during the off-peak periods in their individual networks. Consequently, MNOs can agree to roam users of each other during the off-peak periods, e.g., overnight, and switch off a subset of their BSs (see e.g., [13, 21]). While the vast majority of infrastructure (and spectrum) sharing problems revolve around economic and technical aspects, some papers (see e.g., [12–15,19–22,46,61,84,103,110–112,143]) have taken an energy-efficiency/green networking perspective.

3.5 Strategic modeling

This branch consists of articles that deal with decision-making problems such as MNOs deciding whether to enter a sharing agreement or not, SPs selecting InPs from which to obtain resources etc. In these lines we can further split this category into two subcategories: (i) infrastructure sharing among conventional MNOs and (ii) infrastructure sharing for decoupled infrastructure from services (involving InPs and SPs etc.). Such articles naturally resort to mathematical programming and to game theory in particular when the involved actors are assumed rational, self-interested and payoff-maximizing entities.

3.5.1 Infrastructure sharing among conventional MNOs

The following articles concern either greenfield deployment of shared networks [18, 25–27, 108, 109, 127] or the case when shared networks are created by pooling together the existing network infrastructure of at least two MNOs [39, 41, 94, 126].

Blogowski *et al.* in [18] deal with the particular scenario when two MNOs have to deploy BSs over a given set of candidate sites. For each site, each MNO has to decide whether to install a BS or not; in the former case, if both MNOs decide to install a BS, it is assumed that it is profitable for both to install a single shared BS. The problem is formulated as a non-cooperative game where the payoff of each player (MNO) is given by its total profit (revenues - cost), calculated over all BSs. It is assumed that each site can serve a given (arbitrary) number of users, e.g., those under its coverage area, which means there are no capacity constraints associated with the sites. Instead, coverage constraints are present and they are expressed as a minimum percentage of users to be served by each MNO (a common constraint associated for spectrum licensees). When the coverage constraint is absent, MNOs can decide independently for each site. Otherwise, the game is no longer separable. The authors describe the propriety of the Nash equilibria of the game for different relationships of the payoff matrix (i.e., by establishing relations between the payoffs obtained under different strategy profiles) and also suggest a centralized solution which Pareto dominates all Nash Equilibria.

[108, 109, 127] address the problem of infrastructure and spectrum sharing arising when a set of MNOs, each with a given number of users (market share) and own spectrum license, plan a greenfield Long-Term Evolution (LTE) deployment. The strategic problem of coalition formation, namely, which subsets of MNOs voluntarily sign long-term infrastructure and spectrum sharing agreements, is modeled by means of non-cooperative game theory. We address a very similar problem to [108, 109, 127] in [25, 26] resorting to cooperative game theory in [26] and non-cooperative game theory in [25]. Unlike in [108,109,127], in [25,26] we (i) account for both the technical and economic aspects of sharing reflected in the payoff function definition and (ii) do not split the shared infrastructure cost among MNOs *a priori*; how these cost are split is an outcome of the model (game). In turn in [27], we address a similar scenario to [25, 26] but without spectrum pooling. Moreover, in [27] we consider two different cases deriving from two different perspectives, the one of a regulatory entity favoring the users and the MNOs' perspective as profit-maximizers. We model the former case through Mixed Integer Linear Programming and the latter through cooperative game theory.

The authors in [41] consider the case when a set of MNOs agrees to pool together their current individual RAN networks but make joint decisions for future decommissions, network expansion and upgrades of their shared network; a greedy procedure is proposed to solve the multi-period network planning.

Similarly to the "sale-leaseback" approach of Tower Companies (see e.g., [90]), the work in [39] assumes a set

of self-interested MNOs decide to pool together their respective network infrastructures and create a Joint Venture (JV), responsible for managing their shared network. In turn, MNOs will leaseback network capacity from the JV. The authors propose a Stackelberg game to determine the shares MNOs obtain from the JV and the prices set by the JV to the MNOs and by the MNOs to their respective users.

Notably, the user perspective is considered in [94], which investigates the problem of user-to-BS association when multiple MNOs decide to pool together their respective network infrastructures. The authors propose a non-cooperative game to model the problem of each user selecting its serving BS from the shared pool, independently, so that its individual data rate is maximized.

The work in [126] represents a fresh take on infrastructure sharing. Its authors consider a set of MNOs with individual but overlapping infrastructures (BSs) and individual spectrum licenses; in this setting one of the MNOs (the buyer) can purchase the use of BSs of the other MNOs (the sellers) for serving its own users at its own licensed spectrum. The buyer MNO evaluates whether it can provide a given (Quality of Service) QoS to its own users through its own infrastructure by increasing the transmission power of its BSs or by purchasing BSs from the seller MNOs. In the latter case, the buyer MNO has to decide from which seller MNOs to buy from and what fraction of their BSs to purchase so as to minimize its expenditures while satisfying the QoS of its users. In turn, the seller MNOs have to decide the fraction of their own BSs to sell so as to maximize their profit (payment from the buyer MNO minus cost of sold BSs) where the competition in quantity among the seller MNOs is modeled as a Cournot market.

3.5.2 Infrastructure sharing for decoupled infrastructure from services

We remind the reader that we have discussed the varying terminology used across different articles related to the infrastructure sharing for decoupled infrastructure from services in Section 3.2 and that we have maintained the authors' terminology for the considered stakeholders when describing their articles and, when necessary, we provide clarifications on how they compare to our definitions of InPs and SPs.

It is worth pointing out that, across the different articles very distinct mathematical approaches have been used to study the interaction among InPs and SPs.

Rather exceptionally, the study in [30] tackles the interaction among InPs and MVNOs (analogous to SPs) from the MVNO perspective. In fact, the authors in [30] consider multiple InPs but a single MVNO and propose a model based on contract theory in which the MVNO acts as the employer whereas the InPs as employees.

Instead, Wei *et al.* in [147] take a centralized approach. Specifically, the work in [147] considers multiple InPs and multiple VNOs (analogous to SPs) in the context of WNV. Here, each InP has a given set of users of its own; resources allocated to its own users are referred to as local slices and the total rate across the local slices should be above a given minimum for each InP. Instead, resources allocated to users of an MVNO are referred to as foreign slices. Each InP is characterized by a given bandwidth (number of subchannels) and power budget for the downlink of a BS. The problem consists in determining the number of subchannels and amount of power to allocate to each slice by each InP. The objective is to maximize the total rate across all slices while satisfying the bandwidth and power constraints and the minimum rate requirement for the local slices of each InP. Consequently, the problem is formulated by means of an Integer Programming (IP) model. In this model an MVNO can be simultaneously served by multiple InPs, likewise an InP can simultaneously serve multiple MVNOs.

The authors in [156] propose a hierarchical (two layer) combinatorial auction to model the interactions among multiple InPs, multiple MVNOs (analogous to SPs), and multiple end users concerning the resource allocation at the BS level (the resources here being transmission power, number of channels and number of antennas).

In [28] we propose a novel framework based on a Multi-Leader-Follower Game (MLFG) to study the techno-economic interactions among multiple InPs and multiple SPs in a 5G context.

Table 3 summarizes the main issues that are considered in this subcategory: what are the actors that intervene in the infrastructure sharing scheme and what is the modelling and mathematical approach that is taken in each case.

Article	Actors	Approach
[30]	many InPs - one SP	contract theory
[147]	many InPs - many SPs	IP model
[156]	many InPs - many SPs	auction theory
[28]	many InPs - many SPs	MLFG

Table 3 – Infrastructure sharing for decoupled infrastructure from services

3.6 Miscellaneous

3.6.1 Infrastructure sharing for mobile network segments other than the access

Infrastructure sharing and multi-tenancy can also be applied to specific segments of a mobile network other than the access. For instance, the studies in [23, 98, 128, 140, 141] address sharing of the backhaul network whereas

the one in [83] deals with the sharing of the core network.

3.6.2 Infrastructure sharing among different types of networks

In the following paragraph we provide some examples of heterogeneous infrastructure sharing. The work in [78] studies sharing among different Radio Access Technologies (RATs), the one in [102] addresses sharing between LTE femtocells and Wi–Fi hotspots whereas the one in [100] investigates 3G offloading over Wi–Fi. Kibilda *et al.* [82] deal with sharing among MNOs and OTTs. In [101] the authors propose a RAN architecture for both infrastructure and spectrum sharing between the MNOs and safety services. Instead the study in [95] concerns infrastructure sharing between mobile services and smart grid utilities or intelligent transportation services. Lin *et al.* in [93] address backhaul sharing among mobile networks and fixed networks whereas Simo-Reigadas *et al.* in [131] suggest exploiting the community infrastructure as backhaul for 3G.

3.6.3 Infrastructure sharing for networks other than mobile

The concept of infrastructure sharing is not exclusive to mobile networks. In fact, it has been applied to fixed access networks and problems related to the latter have been recently addressed in the literature (see e.g., [33] and [77]). Apart from fixed access networks, infrastructure sharing has also been proposed for Wi–Fi networks, e.g., in [121].

3.6.4 Spectrum sharing

As previously stated, the overall literature on the different types of spectrum sharing alone (i.e., not combined with infrastructure sharing) is *per se* very vast. Unsurprisingly, as spectrum is a scarce resource for the MNOs, many papers within this literature resort to different game theory models (see e.g., [94, 135, 149, 150]).

3.6.5 MVNO business model

The relation among the MNO, its MVNO(s) and the end users has been largely addressed through game theory as well (see e.g., [57, 87, 88, 96, 134]).

4. DISCUSSION AND OUTLOOK

The large body of literature analyzed in this survey and the impact it has had over the years on the standardization of mobile technologies and the practices adopted around the world, testify that mobile networks are important infrastructures with high costs which can be shared in some scenarios in order to offer better and more convenient services to end users. Also, in terms of regulatory strategies adopted by national authorities in different countries, that over the years have favoured the introduction of MVNOs and roaming mechanisms, we can observe that the wide service availability at reasonable prices has been considered particularly important also with respect to natural market competition.

With the arrival of 5G, we are observing a renewed interest in sharing strategies due to the specific virtualization technology available and the standardization of network slicing. For the first time, the dynamic allocation of network resources and the service specialization on different slices, allow serving different groups of users according to different quality levels. This will likely generate the interest of new players specialized in vertical application domains, in order to take the role of slice tenants and sharing the resources of the network infrastructures with others. Even extreme scenarios where communications resources are traded in real time on automated markets are now possible [92], like it already happens in other sectors like energy networks. With this regard, the use of the large toolbox created by research over several years will certainly be an important asset to be used to shape sharing and trading instruments.

There are however, other important evolution trends of the technology that will probably influence the sharing methods beyond 5G. We want to mention here two of them that we believe are particularly relevant:

- the extreme distribution of access infrastructure with the so-called smart radio environments, and
- the full virtualization of connectivity in open and cloud-based architectures.

As for the smart radio environments, they consider the introduction of new equipment at the radio interface of mobile networks, which includes smart repeaters with large antenna arrays and controllable reflective surfaces. This kind of evolution trend is making clear that in the future the deployment of multiple physical infrastructures by different operators will become more and more difficult and the focus of MNO attention will shift from optimizing their own network to that of efficiently controlling the resources leased from the common infrastructure.

While the full virtualization of connectivity will be complete relatively soon, the evolution started years ago with the separation of the network logic from the pure transmission technology. There are here however, important novelties associated with open source approaches like Open RAN that are becoming popular and that are fostering the transition to cloud-based solutions where the value for service providers will be in designing and implementing advanced services based on the effective use of transmission resources.

5. CONCLUSION

Infrastructure sharing in mobile networks has been a pervasive research topic over the last three decades and has produced a significant body of work.

One interesting takeaway from this survey is that while researchers sought enabling technologies to materialize infrastructure sharing in 3G and 4G networks, in 5G networks instead, infrastructure sharing became an important pillar of the 5G architecture which means that in turn 5G enables infrastructure sharing from a business point of view. In these lines one can easily argue the presence of infrastructure sharing also in future networks.

As for the mobile ecosystem, it is worth noting that the concepts of decoupling infrastructure from services and dynamic spectrum trading have been anticipated in the literature over two decades ago but they have come into being only recently (mainly in 5G networks).

Of the several research branches within the infrastructure and spectrum sharing topics, resource management in the context of resource sharing has been and will be one of the most active research branches given the current and future need for dynamic resource sharing.

REFERENCES

[1] 3GPP. TR 22.852, v13.1.0, Study on Radio Access Network (RAN) sharing enhancements (Release 13), September 2014.

[2] 3GPP. TR 22.951, v15.0.0, Service aspects and requirements for network sharing (Release 15), July 2018.

[3] 3GPP. TS 23.251, v15.1.0, Network sharing; Architecture and functional descriptions (Release 15), September 2018.

[4] 3GPP. TS 32.130, v15.0.0, Telecommunication management; Network sharing; Concepts and requirements (Release 15), June 2018.

[5] 5G NORMA. Deliverable D3.3, 5G NORMA network architecture - final report, October 2017.

[6] 5GPPP Architecture Working Group. View on 5G architecture, December 2017.

[7] H. Ahmadi, I. Macaluso, I. Gomez, L. DaSilva, and L. Doyle. Virtualization of spatial streams for enhanced spectrum sharing. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2016.

[8] Ö. U. Akgül, I. Malanchini, V. Suryaprakash, and A. Capone. Dynamic resource allocation and pricing for shared radio access infrastructure. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2017.

[9] O. Al-Khatib, W. Hardjawana, and B. Vucetic. Spectrum sharing in multi-tenant 5G cellular networks: Modeling and planning. *IEEE Access*, 7:1602–1616, 2019.

[10] L. Anchora, M. Mezzavilla, L. Badia, and M. Zorzi. A performance evaluation tool for spectrum sharing in multi-operator LTE networks. *Computer Communications*, 35(18):2218–2226, November 2012.

[11] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang. What will 5g be? *IEEE Journal on selected areas in communications*, 32(6):1065–1082, 2014.

[12] C. Anglano, M. Guazzone, and M. Sereno. Maximizing profit in green cellular networks through collaborative games. *Computer Networks*, 75:260–275, 2014.

[13] A. Antonopoulos, E. Kartsakli, A. Bousia, L. Alonso, and C. Verikoukis. Energy-efficient infrastructure sharing in multi-operator mobile networks. *IEEE Communications Magazine*, 53(5):242–249, 2015.

[14] O. Aydin, E. A. Jorswieck, D. Aziz, and A. Zappone. Energy-spectral efficiency tradeoffs in 5G multi-operator networks with heterogeneous constraints. *IEEE Transactions on Wireless Communications*, 16(9):5869–5881, 2017.

[15] Y. Bao, J. Wu, S. Zhou, and Z. Niu. Bayesian mechanism based inter-operator base station sharing for energy saving. In *Communications (ICC), 2015 IEEE International Conference on*, pages 49–54. IEEE, 2015.

[16] C. Beckman and G. Smith. Shared networks: Making wireless communication affordable. *IEEE Wireless Communications*, 12(2):78–85, April 2005.

[17] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez. Optimising 5G infrastructure markets: The business of network slicing. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pages 1–9. IEEE, 2017.

[18] A. Blogowski, P. Chrétienne, and F. Pascual. Network sharing by two mobile operators: beyond competition, cooperation. *RAIRO-Operations Research*, 49(3):635–650, 2015.

[19] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis. Game theoretic approach for switching off base stations in multi-operator environments. In *2013 IEEE International Conference on Communications (ICC)*, pages 4420–4424. IEEE, 2013.

[20] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis. Auction-based offloading for base station switching off in heterogeneous networks. In *2016 European Conference on Networks and Communications (EuCNC)*, pages 335–339. IEEE, 2016.

[21] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis. Game-theoretic infrastructure sharing in multioperator cellular networks. *IEEE Transactions on Vehicular Technology*, 65(5):3326–3341, 2016.

[22] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis. Multiobjective auction-based switching-off scheme in heterogeneous networks: To bid or not to bid? *IEEE Transactions on Vehicular Technology*, 65(11):9168–9180, 2016.

[23] C. Caillouet, D. Coudert, and A. Kodjo. Robust optimization in multi-operators microwave backhaul networks. In *Global Information Infrastructure Symposium-GIIS 2013*, pages 1–6. IEEE, 2013.

[24] L. Cano. *Game-theoretic frameworks for the techno-economic aspects of infrastructure sharing in current and future mobile networks*. PhD thesis, Polytechnique Montréal and Politecnico di Milano, 2020.

[25] L. Cano, A. Capone, G. Carello, M. Cesana, and M. Passacantando. Cooperative infrastructure and spectrum sharing in heterogeneous mobile networks. *IEEE Journal on Selected Areas in Communications*, 34(10):2617–2629, oct 2016.

[26] L. Cano, A. Capone, G. Carello, M. Cesana, and M. Passacantando. A non-cooperative game approach for RAN and spectrum sharing in mobile radio networks. In *22th European Wireless Conference*, pages 1–6, Oulu, Finland, May 18-20, 2016.

[27] L. Cano, A. Capone, G. Carello, M. Cesana, and M. Passacantando. On optimal infrastructure sharing strategies in mobile radio networks. *IEEE Transactions on Wireless Communications*, 16(5):3003–3016, may 2017.

[28] L. Cano, G. Carello, M. Cesana, M. Passacantando, and B. Sansò. Modeling the techno-economic interactions of infrastructure and service providers in 5G networks with a multi-leader-follower game. *IEEE Access*, 7:162913–162940, dec 2019.

[29] C.-Y. Chang, N. Nikaein, O. Arouk, K. Katsalis, A. Ksentini, T. Turletti, and K. Samdanis. Slice orchestration for multi-service disaggregated ultra-dense RANs. *IEEE Communications Magazine*, 56(8):70–77, 2018.

[30] Z. Chang, D. Zhang, T. Hämäläinen, Z. Han, and T. Ristaniemi. Incentive mechanism for resource allocation in wireless virtualized networks with multiple infrastructure providers. *IEEE Transactions on Mobile Computing*, 19(1):103–115, 2018.

[31] Z. Chang, K. Zhu, Z. Zhou, and T. Ristaniemi. Service provisioning with multiple service providers in 5g ultra-dense small cell networks. In *IEEE PIMRC 2015*, pages 1895–1900, 2015.

[32] X. Chen, H. Zhang, and Z. Han. Delay-tolerant resource scheduling in large-scale virtualized radio access networks. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

[33] B. Cornaglia, G. Young, and A. Marchetta. Fixed access network sharing. *Optical Fiber Technology*, 26:2–11, 2015.

[34] X. Costa-Pérez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan. Radio access network virtualization for future mobile carrier networks. *IEEE Communications Magazine*, 51(7):27–35, July 2013.

[35] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar. DEMO: SDN-based network slicing in C-RAN. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–2. IEEE, 2018.

[36] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar. A network slicing prototype for a flexible cloud radio access network. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–4. IEEE, 2018.

[37] S. Costanzo, D. Xenakis, N. Passas, and L. Merakos. Augmented RAN with SDN Orchestration of Multi-tenant Base Stations. *Wireless Personal Communications*, 96(2):2009–2037, 2017.

[38] L. A. DaSilva, J. Kibilda, P. DiFrancesco, T. K. Forde, and L. E. Doyle. Customized services over virtual wireless networks: The path towards networks without borders. In *Future Network and Mobile Summit (FutureNetworkSummit), 2013*, pages 1–10. IEEE, 2013.

[39] X. Deng, J. Wang, and J. Wang. How to Design a Common Telecom Infrastructure for Competitors to be Individually Rational and Collectively Optimal. *IEEE Journal on Selected Areas in Communications*, 35(3):736–750, 2017.

[40] P. Di Francesco, F. Malandrino, and L. A. DaSilva. Mobile network sharing between operators: a demand trace-driven study. In *Proceedings of the 2014 ACM SIGCOMM workshop on Capacity sharing workshop*, pages 39–44. ACM, 2014.

[41] P. Di Francesco, F. Malandrino, T. K. Forde, and L. A. DaSilva. A sharing-and competition-aware framework for cellular network evolution planning. *IEEE Transactions on Cognitive Communications and Networking*, 1(2):230–243, June 2015.

[42] S. D'Oro, F. Restuccia, T. Melodia, and S. Palazzo. Low-complexity distributed radio access network slicing: Algorithms and experimental results. *IEEE/ACM Transactions on Networking*, 26(6):2815 – 2828, 2018.

[43] L. Doyle, J. Kibilda, T. K. Forde, and L. DaSilva. Spectrum without bounds, networks without borders. *Proceedings of the IEEE*, 102(3):351–365, 2014.

[44] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano. 5G RAN slicing for verticals: Enablers and challenges. *IEEE Communications Magazine*, 57(1):28–34, 2019.

[45] Ericsson. Network sharing. https://www.ericsson.com/us/ourportfolio/networks-services/network-sharing?nav=marketcategory004 [Online; Accessed: 2017-05-15].

[46] M. J. Farooq, H. Ghazzai, E. Yaacoub, A. Kadri, and M.-S. Alouini. Green virtualization for multiple collaborative cellular operators. *IEEE Transactions on Cognitive Communications and Networking*, 3(3):420–434, 2017.

[47] A. Fendt, S. Lohmuller, L. C. Schmelz, and B. Bauer. A network slice resource allocation and optimization model for end-to-end mobile networks. In *2018 IEEE 5G World Forum (5GWF)*, pages 262–267. IEEE, 2018.

[48] R. Ferrus, O. Sallent, J. Pérez-Romero, and R. Agusti. On 5G radio access network slicing: Radio interface protocol features and configuration. *IEEE Communications Magazine*, 56(5):184–192, 2018.

[49] T. Frisanco, P. Tafertshofer, P. Lurin, and R. Ang. Infrastructure sharing for mobile network operators; From a deployment and operations view. In *IEEE International Conference on Information Networking (ICOIN)*, pages 1–5, January 2008.

[50] L. Gao, P. Li, Z. Pan, N. Liu, and X. You. Virtualization framework and VCG based resource block allocation scheme for LTE virtualization. In *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pages 1–6. IEEE, 2016.

[51] F. Grijpink, S. Newman, S. Sandoval, M. Strandell-Jansson, and W. Torfs. A "New Deal": Driving investment in Europe's telecoms infrastructure. https://tmt.mckinsey.com/content/industry/Telecommunications/page/19, 2012. [Online; Accessed: 2016-03-30].

[52] GSMA. Mobile infrastructure sharing. https://www.gsma.com/mobilefordevelopment/programme/connected-society/mobile-infrastructure-sharing-report/, November 2008. Accessed: 11-01-2019.

[53] A. Gudipati, L. E. Li, and S. Katti. RadioVisor: A slicing plane for Radio Access Networks. In *Proceedings of the Third Workshop on Hot Topics in Software Defined Networking*, HotSDN '14, pages 237–238, 2014.

[54] A. K. Gupta, J. G. Andrews, and R. W. Heath. On the feasibility of sharing spectrum licenses in mmwave cellular systems. *IEEE Transactions on Communications*, 64(9):3981–3995, 2016.

[55] J. Harno. 3G business case successfulness within the constraints set by competition, regulation and alternative technologies. *JOURNAL-COMMUNICATIONS NETWORK*, 1(2):159–165, 2002.

[56] J. He and W. Song. Appran: Application-oriented radio access network sharing in mobile networks. In *Communications (ICC), 2015 IEEE International Conference on*, pages 3788–3794. IEEE, 2015.

[57] S. L. Hew and L. B. White. Cooperative resource allocation games in shared networks: Symmetric and asymmetric fair bargaining models. *IEEE Transactions on Wireless Communications*, 7(11):4166–4175, November 2008.

[58] T. M. Ho, N. H. Tran, S. A. Kazmi, and C. S. Hong. Dynamic pricing for resource allocation in wireless network virtualization: A stackelberg game approach. In *IEEE ICOIN 2017*, pages 429–434, 2017.

[59] T. M. Ho, N. H. Tran, L. B. Le, Z. Han, S. A. Kazmi, and C. S. Hong. Network virtualization with energy efficiency optimization for wireless heterogeneous networks. *IEEE Transactions on Mobile Computing*, 18(10):2386–2400, 2018.

[60] M. Hoffmann and M. Staufer. Network virtualization for future mobile networks: General architecture and applications. In *2011 IEEE international conference on communications workshops (ICC)*, pages 1–5. IEEE, 2011.

[61] M. F. Hossain, K. S. Munasinghe, and A. Jamalipour. Energy-efficient inter-RAN cooperation for non-collocated cell sites with base station selection and user association policies. *Wireless Networks*, 25(1):269–285, 2019.

[62] F.-T. Hsu and C.-H. Gan. Resource allocation with spectrum aggregation for wireless virtual network embedding. In *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, pages 1–5. IEEE, 2015.

[63] M. Hu, Y. Chang, Y. Sun, and H. Li. Dynamic slicing and scheduling for wireless network virtualization in downlink LTE system. In *2016 19th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 153–158. IEEE, 2016.

[64] J. Hultel, K. Johansson, and J. Markendahl. Business models and resource management for shared wireless networks. In *IEEE 60th Vehicular Technology Conference (VTC2004-Fall)*, volume 5, pages 3393–3397, September 2004.

[65] Industry Canada. Framework for mandatory roaming and antenna tower and site sharing. http://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/h$_s$f10290.*html*, 2013.[*Online*; *Accessed* : $2016 - 03 - 30$].

[66] ITU. Mobile and wireless network regulation. http://www.ictregulationtoolkit.org/2.6. [Online; Accessed: 2016-03-28].

[67] T. Janssen, R. Litjens, and K. W. Sowerby. On the expiration date of spectrum sharing in mobile cellular networks. In *2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 490–496. IEEE, 2014.

[68] Y. Jia, H. Tian, S. Fan, P. Zhao, and K. Zhao. Bankruptcy game based resource allocation algorithm for 5G Cloud-RAN slicing. In *IEEE WCNC 2018*, pages 1–6, 2018.

[69] K. Johansson. *Cost effective deployment strategies for heterogeneous wireless networks*. PhD thesis, KTH Information and Communication Technology, 2007.

[70] K. Johansson, M. Kristensson, and U. Schwarz. Radio resource management in roaming based multi-operator wcdma networks. In *Vehicular Technology Conference, 2004. VTC 2004-Spring. 2004 IEEE 59th*, volume 4, pages 2062–2066. IEEE, 2004.

[71] R. Jurdi, A. K. Gupta, J. G. Andrews, and R. W. Heath. Modeling infrastructure sharing in mmWave networks with shared spectrum licenses. *IEEE Transactions on Cognitive Communications and Networking*, 4(2):328–343, 2018.

[72] M. Kalil, M. Youssef, A. Shami, A. Al-Dweik, and S. Ali. Wireless resource virtualization: opportunities, challenges, and solutions. *Wireless Communications and Mobile Computing*, 16(16):2690–2699, 2016.

[73] M. I. Kamel, L. B. Le, and A. Girard. LTE wireless network virtualization: Dynamic slicing via flexible scheduling. In *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pages 1–5. IEEE, 2014.

[74] E. Kapassa, M. Touloupou, P. Stavrianos, and D. Kyriazis. Dynamic 5G Slices for IoT applications with diverse requirements. In *2018 Fifth International Conference on Internet of Things: Systems, Management and Security*, pages 195–199. IEEE, 2018.

[75] M. Katsigiannis, T. Smura, T. Casey, and A. Sorri. Techno-economic modeling of value network configurations for public wireless local area access. *NETNOMICS: Economic Research and Electronic Networking*, 14(1):27–46, November 2013.

[76] S. M. A. Kazmi and C. S. Hong. A matching game approach for resource allocation in wireless network virtualization. In *Proceedings of the 11th International Conference on ubiquitous information management and communication*, pages 1–6, 2017.

[77] K. J. Kerpez, J. M. Cioffi, P. J. Silverman, B. Cornaglia, and G. Young. Fixed access network sharing. *IEEE Communications Standards Magazine*, 1(1):82–89, 2017.

[78] M. A. Khan, A. C. Toker, C. Troung, F. Sivrikaya, and S. Albayrak. Cooperative game theoretic approach to integrated bandwidth sharing and allocation. In *IEEE International Conference on Game Theory for Networks (GameNets '09)*, pages 1–9, May 2009.

[79] S. Khatibi and L. M. Correia. Modelling of virtual radio resource management for cellular heterogeneous access networks. In *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, pages 1152–1156. IEEE, 2014.

[80] J. Kibilda and L. A.DaSilva. Efficient coverage through inter-operator infrastructure sharing in mobile networks. In *IEEE IFIP Wireless Days (WD)*, pages 1–6, November 2013.

[81] J. Kibiłda, N. J. Kaminski, and L. A. DaSilva. Radio access network and spectrum sharing in mobile networks: A stochastic geometry perspective. *IEEE Transactions on Wireless Communications*, 16(4):2562–2575, 2017.

[82] J. Kibilda, F. Malandrino, and L. A. DaSilva. Incentives for infrastructure deployment by over-the-top service providers in a mobile network: A cooperative game theory model. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.

[83] W. Kiess, M. R. Sama, J. Varga, J. Prade, H.-J. Morper, and K. Hoffmann. 5G via evolved packet core slices: Costs and technology of early deployments. In *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–7. IEEE, 2017.

[84] G. Koutitas, G. Iosifidis, B. Lannoo, M. Tahon, S. Verbrugge, P. Ziridis, Ł. Budzisz, M. Meo, M. A. Marsan, and L. Tassiulas. Greening the airwaves with collaborating mobile network operators. *IEEE Transactions on Wireless Communications*, 15(1):794–806, 2016.

[85] F. Kurtz, C. Bektas, N. Dorsch, and C. Wietfeld. Network slicing for critical communications in shared 5G infrastructures-an empirical evaluation. In *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, pages 393–399. IEEE, 2018.

[86] L.-V. Le, B.-S. P. Lin, L.-P. Tung, and D. Sinh. SDN/NFV, Machine Learning, and Big Data Driven Network Slicing for 5G. In *2018 IEEE 5G World Forum (5GWF)*, pages 20–25. IEEE, 2018.

[87] H. Le Cadre and M. Bouhtou. An interconnection game between mobile network operators: Hidden information forecasting using expert advice fusion. *Computer networks*, 54(17):2913–2942, 2010.

[88] H. Le Cadre and M. Bouhtou. Modelling MNO and MVNO's dynamic interconnection relations: is cooperative content investment profitable for both providers? *Telecommunication Systems*, 51(2-3):193–217, 2012.

[89] T. LeAnh, N. H. Tran, D. T. Ngo, and C. S. Hong. Resource allocation for virtualized wireless networks with backhaul constraints. *IEEE Communications Letters*, 21(1):148–151, 2017.

[90] T. Levine, P. Eijsvoogel, and M. Reede. Passive infrastructure sharing. http://www.allenovery.com/SiteCollectionDocuments/Passive2012. [Online; Accessed: 2016-03-28].

[91] C. Liang and F. R. Yu. Wireless virtualization for next generation mobile cellular networks. *IEEE wireless communications*, 22(1):61–69, 2015.

[92] A. Lieto, I. Malanchini, S. Mandelli, E. Moro, and A. Capone. Strategic network slicing management in radio access networks. *IEEE Transactions on Mobile Computing*, 2020.

[93] P. Lin, J. Zhang, Q. Zhang, and M. Hamdi. Enabling the femtocells: A cooperation framework for mobile and fixed-line operators. *IEEE Transactions on Wireless Communications*, 12(1):158–167, 2013.

[94] Y.-T. Lin, H. Tembine, and K.-C. Chen. Interoperator spectrum sharing in future cellular systems. In *Global Communications Conference (GLOBECOM), 2012 IEEE*, pages 2597–2602. IEEE, 2012.

[95] Y. Lostanlen. From heterogeneous wireless networks to sustainable efficient ICT infrastructures. In *2013 7th European Conference on Antennas and Propagation (EuCAP)*, pages 1360–1363. IEEE, 2013.

[96] M. H. Lotfi and S. Sarkar. The economics of competition and cooperation between mnos and mvnos. In *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2017.

[97] X. Lu, K. Yang, Y. Liu, D. Zhou, and S. Liu. An elastic resource allocation algorithm enabling wireless network virtualization. *Wireless Communications and Mobile Computing*, 15(2):295–308, 2015.

[98] J. Lun and D. Grace. Software defined network for multi-tenancy resource sharing in backhaul networks. In *2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pages 1–5. IEEE, 2015.

[99] I. Malanchini, S. Valentin, and O. Aydin. Wireless resource sharing for multiple operators: Generalization, fairness, and the value of prediction. *Computer Networks*, 100:110–123, 2016.

[100] L. Mamatas, I. Psaras, and G. Pavlou. Incentives and algorithms for broadband access sharing. In *Proceedings of the 2010 ACM SIGCOMM workshop on Home networks*, pages 19–24, 2010.

[101] D. Marabissi and R. Fantacci. Heterogeneous public safety network architecture based on RAN slicing. *IEEE Access*, 5:24668–24677, 2017.

[102] J. Markendahl and M. Nilson. Business models for deployment and operation of femtocell networks; – Are new operation strategies needed for mobile operators? In *21st European Regional ITS Conference, Copenhagen*, September 2010.

[103] M. A. Marsan and M. Meo. Network sharing and its energy benefits: A study of European mobile

network operators. In *2013 IEEE Global Communications Conference (GLOBECOM)*, pages 2561–2567. IEEE, 2013.

[104] D.-E. Meddour, T. Rasheed, and Y. Gourhant. On the role of infrastructure sharing for mobile network operators in emerging markets. *Computer Networks*, 55(7):1576–1591, May 2011.

[105] O. Narmanlioglu and E. Zeydan. Efficient RRH assignments for mobile network operators in shared cellular network architecture. In *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pages 1103–1108. IEEE, 2017.

[106] B. Naudts, M. Kind, F.-J. Westphal, S. Verbrugge, D. Colle, and M. Pickavet. Techno-economic analysis of software defined networking as architecture for the virtualization of a mobile network. In *2012 European workshop on software defined networking*, pages 67–72. IEEE, 2012.

[107] Nokia. White paper - network sharing: Delivering mobile broadband more efficiently and at lower cost. nokia.com. [Online; Accessed: 2017-05-15].

[108] F. Offergelt, F. Berkers, and G. Hendrix. If you can't beat 'em, join 'em; Cooperative and non-cooperative games in network sharing. In *IEEE 15th International Conference on Intelligence in Next Generation Networks (ICIN)*, pages 196–201, October 2011.

[109] F. H. Offergelt. Saphyre: Cooperation among competitors – analysing sharing scenarios for mobile network operators using game theory. Master's thesis, Leiden University, The Netherlands, 2011.

[110] M. Oikonomakou, A. Antonopoulos, L. Alonso, and C. Verikoukis. Cooperative base station switching off in multi-operator shared heterogeneous network. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.

[111] M. Oikonomakou, A. Antonopoulos, L. Alonso, and C. Verikoukis. Evaluating cost allocation imposed by cooperative switching off in multioperator shared hetnets. *IEEE Transactions on Vehicular Technology*, 66(12):11352–11365, 2017.

[112] M. Oikonomakou, A. Antonopoulos, L. Alonso, and C. Verikoukis. Fairness in multi-operator energy sharing. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

[113] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira. Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges. *IEEE Communications Magazine*, 55(5):80–87, 2017.

[114] J. S. Panchal, R. D. Yates, and M. M. Buddhikot. Mobile network resource sharing options: Performance comparisons. *IEEE Transactions on Wireless Communications*, 12(9):4470–4482, 2013.

[115] J. S. Park, M. Kim, and H. J. Lee. Analysis of European 3G markets and advanced strategies for 3G development. In *The 7th International Conference on Advanced Communication Technology, 2005, ICACT 2005.*, volume 1, pages 428–431. IEEE, 2005.

[116] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí. On the configuration of radio resource management in a sliced RAN. In *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*, pages 1–6. IEEE, 2018.

[117] M. Rahman, C. Despins, and S. Affes. Analysis of CAPEX and OPEX benefits of wireless access virtualization. In *IEEE International Conference on Communications (ICC) Workshops*, pages 436–440, June 2013.

[118] P. Ramsdale. Personal communications in the UK—Implementation of PCN using DCS 1800. *International Journal of Wireless Information Networks*, 1(1):29–36, 1994.

[119] R. Ravindran, A. Chakraborti, S. O. Amin, A. Azgin, and G. Wang. 5G-ICN: Delivering ICN services over 5G using network slicing. *IEEE Communications Magazine*, 55(5):101–107, 2017.

[120] M. Rebato, M. Mezzavilla, S. Rangan, and M. Zorzi. Resource sharing in 5g mmwave cellular networks. In *Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on*, pages 271–276. IEEE, 2016.

[121] M. Richart, J. Baliosian, J. Serrati, J.-L. Gorricho, R. Agüero, and N. Agoulmine. Resource allocation for network slicing in WiFi access points. In *2017 13th International conference on network and service management (CNSM)*, pages 1–4. IEEE, 2017.

[122] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed. Scheduling wireless virtual networks functions. *IEEE Transactions on network and service management*, 13(2):240–252, 2016.

[123] B. Rouzbehani, L. M. Correia, and L. Caeiro. Radio resource and service orchestration for virtualised multi-tenant mobile Het-Nets. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–5. IEEE, 2018.

[124] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti. On radio access network slicing from a radio resource management perspective. *IEEE Wireless Communications*, 24(5):166–174, 2017.

[125] K. Samdanis, X. Costa-Perez, and V. Sciancalepore. From network sharing to multi-tenancy: The 5G network slice broker. *IEEE Communications Magazine*, 54(7):32–39, 2016.

[126] T. Sanguanpuak, S. Guruacharya, E. Hossain, N. Rajatheva, and M. Latva-aho. Infrastructure sharing for mobile network operators: analysis of trade-offs and market. *IEEE Transactions on Mobile Computing*, 17(12):2804–2817, 2018.

[127] SAPHYRE. Deliverable D5.5, Business models, cost analysis and advices for spectrum policy and regulation for scenario III (full sharing), 2013.

[128] O. Semiari, W. Saad, M. Bennis, and Z. Dawy. Inter-operator resource management for millimeter wave multi-hop backhaul networks. *IEEE Transactions on Wireless Communications*, 16(8):5258–5272, 2017.

[129] M.-K. Shin, S. Lee, S. Lee, and D. Kim. A way forward for accommodating NFV in 3GPP 5G systems. In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 114–116. IEEE, 2017.

[130] F. Shirzad and M. Ghaderi. Cloud-based spectrum sharing in virtual wireless networks. In *2016 IEEE 24th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 196–204. IEEE, 2016.

[131] J. Simo-Reigadas, E. Municio, E. Morgado, E. M. Castro, A. Martinez, L. F. Solorzano, and I. Prieto-Egido. Sharing low-cost wireless infrastructures with telecommunications operators to bring 3g services to rural communities. *Computer Networks*, 93:245–259, 2015.

[132] H. M. Soliman and A. Leon-Garcia. A novel neuro-optimization method for multi-operator scheduling in cloud-RANs. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.

[133] H. M. Soliman and A. Leon-Garcia. QoS-aware frequency-space network slicing and admission control for virtual wireless networks. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2016.

[134] F. Sun, B. Liu, F. Hou, H. Zhou, L. Gui, and J. Chen. Cournot equilibrium in the mobile virtual network operator oriented oligopoly offloading market. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.

[135] F. Teng, D. Guo, and M.-L. Honig. Sharing of unlicensed spectrum by strategic operators. In *IEEE Global Conference for Signal Processing and Communications (GlobalSIP)*, pages 288–292, December 2014.

[136] T. D. Tran and L. B. Le. Resource allocation for efficient bandwidth provisioning in virtualized wireless networks. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2017.

[137] S. Valentin, W. Jamil, and O. Aydin. Extending generalized processor sharing for multi-operator scheduling in cellular networks. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*, pages 485–490. IEEE, 2013.

[138] J. van de Belt, H. Ahmadi, L. E. Doyle, and O. Sallent. A prioritised traffic embedding mechanism enabling a public safety virtual operator. In *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, pages 1–5. IEEE, 2015.

[139] F. Vaz, P. Sebastiao, L. Goncalves, and A. Correia. Femtocell deployment in LTE-A networks: A sustainability, economical and capacity analysis. In *IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pages 3423–3427, September 2013.

[140] D. P. Venmani, Y. Gourhant, and D. Zeghlache. Divide and share: A new approach for optimizing backup resource allocation in LTE mobile networks backhaul. In *2012 8th International Conference on Network and Service management (CNSM) and 2012 Workshop on Systems Virtualization Management (SVM)*, pages 189–193. IEEE, 2012.

[141] D. P. Venmani, Y. Gourhant, and D. Zeghlache. ROFL: Restoration of failures through link-bandwidth sharing. In *2012 IEEE Globecom Workshops*, pages 30–35. IEEE, 2012.

[142] J. Village, K. Worrall, and D. Crawford. 3g shared infrastructure. In *Third International Conference on 3G Mobile Communication Technologies (Conf. Publ. No. 489)*, pages 10–16. IET, 2002.

[143] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso, and C. Verikoukis. Cooperation incentives for multi-operator C-RAN energy efficient sharing. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

[144] P. L. Vo, M. N. Nguyen, T. A. Le, and N. H. Tran. Slicing the edge: Resource allocation for RAN network slicing. *IEEE Wireless Communications Letters*, 7(6):970–973, 2018.

[145] S. Wang, K. Samdanis, X. C. Perez, and M. Di Renzo. On spectrum and infrastructure sharing in multi-operator cellular networks. In *2016 23rd International Conference on Telecommunications (ICT)*, pages 1–4. IEEE, 2016.

[146] J. Wei, K. Yang, G. Zhang, and Z. Hu. Pricing-based power allocation in wireless network virtualization: A game approach. In *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 188–193. IEEE, 2015.

[147] J. Wei, K. Yang, G. Zhang, and X. Lu. A QoS-Aware Joint Power and Subchannel Allocation Algorithm for Mobile Network Virtualization. *Wireless Personal Communications*, 104(2):507–526, 2019.

[148] D. Wu, Z. Zhang, S. Wu, J. Yang, and R. Wang. Biologically inspired resource allocation for network slices in 5G-enabled Internet of Things. *IEEE Internet of Things Journal*, 6(6):9266–9279, 2018.

[149] Y. Xiao, Z. Han, C. Yuen, and L. A. DaSilva. Carrier aggregation between operators in next generation cellular networks: A stable roommate market. *IEEE Transactions on Wireless Communications*, 15(1):633–650, 2016.

[150] Y. Xiao, C. Yuen, P. Di Francesco, and L. A. DaSilva. Dynamic spectrum scheduling for carrier aggregation: A game theoretic approach. In *Communications (ICC), 2013 IEEE International Conference on*, pages 2672–2676. IEEE, 2013.

[151] D. Xu and Q. Li. Resource allocation in wireless virtualized networks with energy harvesting. In *2016 IEEE International Conference on Communication Systems (ICCS)*, pages 1–6. IEEE, 2016.

[152] Y. Zaki, L. Zhao, C. Goerg, and A. Timm-Giel. LTE mobile network virtualization: Exploiting multiplexing and multi-user diversity gain. *Mobile Networks & Applications*, 16(4):424–432, August 2011.

[153] J. Zheng, P. Caballero, G. De Veciana, S. J. Baek, and A. Banchs. Statistical multiplexing and traffic shaping games for network slicing. *IEEE/ACM Transactions on Networking*, 26(6):2528–2541, 2018.

[154] R. Zhou, X. Yin, Z. Li, and C. Wu. Virtualized resource sharing in cloud radio access networks: An auction approach. *Computer Communications*, 114:22–35, 2017.

[155] K. Zhu, Z. Cheng, B. Chen, and R. Wang. Wireless virtualization as a hierarchical combinatorial auction: An illustrative example. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2017.

[156] K. Zhu and E. Hossain. Virtualization of 5g cellular networks as a hierarchical combinatorial auction. *IEEE Transactions on Mobile Computing*, 15(10):2640–2654, 2016.

AUTHORS

L. Cano (BSc. 12, MSc. 14, PhD 20) is currently a researcher at Politecnico di Milano. She recently obtained a double degree PhD at Politecnico di Milano and Polytechnique Montréal. Her main research interests are in the techno-economic characterization of infrastructure sharing in networks based on game theoretical models.

A. Capone (PhD 1998) is full professor at Politecnico di Milano where is he also the Dean of the School of Industrial and Information Engineering, member of the university strategy team POLIMI2040, and director of the Advanced Network Technologies Laboratory (ANTLab). He is associate editor with IEEE Trans. on Mobile Computing and Elsevier Computer Communications, and member of the TPC of major conferences in networking. His research interests include radio resource management and planning of wireless networks, software defined networks and switching architectures. On these topics he has published more than 300 papers. He is a fellow of the IEEE.

B. Sansò (Ph.D.89) is a full professor of telecommunication networks in the department of Electrical Engineering of Polytechnique Montréal and a member of GERAD, a world-renowned applied mathematics research center. Over her long career, she has received many awards and honors, has published and consulted extensively for industry and the mainstream media and has been part of major international committees. She leads the LORLAB, a research group dedicated to developing effective applied mathematics methods to the design and performance of wireless and wireline telecommunication networks. She has a special interest in network robustness and sustainability.

INDEX OF AUTHORS

T

V

X

Y

International
Telecommunication
Union

Telecommunication
Standardization Bureau (TSB)

Place des Nations
CH-1211 Geneva 20
Switzerland

ISSN: 2616-8375
Published in Switzerland
Geneva, December 2020