Doctoral theses at NTNU, 2023:407

Ali Esmaeily

Network Slicing in Beyond 5G: Implementation, Isolation, and Coexistence

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

Ali Esmaeily

Network Slicing in Beyond 5G: Implementation, Isolation, and Coexistence

Thesis for the Degree of Philosophiae Doctor

Trondheim, November 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

ISBN 978-82-326-7512-8 (printed ver.)
ISBN 978-82-326-7511-1 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:407

Printed by NTNU Grafisk senter

Abstract

The contemporary landscape of telecommunications is defined by the advent of Fifth Generation (5G) and the tantalizing promise of Beyond 5G (B5G) networks. 5G and B5G networks are the keys to fulfilling a massive number of heterogeneous services, applications, and use cases that have come into existence in recent years. Each use case imposes diverse performance requirements regarding latency, availability, reliability, and data rate. Within this landscape, Network Slicing in 5G emerges as a cornerstone technology, addressing the need for tailored and efficient allocation of infrastructure resources to such diverse use cases. However, defining, creating, establishing, managing, and monitoring the operation of network slices turn out to be challenging tasks. Other than this, managing the running network slices brings more challenges. One of these challenges is to operate network slices in an isolated fashion from each other to provide the demanded services via the network slices according to the Quality of Service (QoS) requirements of each service. Besides, resource allocation mechanisms in different networking domains become more complex than traditional networks to simultaneously provide diverse use cases on physical infrastructure. The research contributions of this Ph.D. concentrate on three main parts.

First, this research endeavor provides a rational and practical roadmap for the realization of network slices, with a particular emphasis on small-scale frameworks and controlled testbed environments. The process of defining, creating, and establishing these slices requires accurate Management and Orchestration (MANO) procedures. By delving into the practical sophistication of this work, the study aims to deliver actionable insights to navigate the complexities of network slicing.

Second, the imperative of isolation is essential to address. The research focuses on the isolation concept in terms of performance and security within the Core Network (CN) domain for the coexistence of Enhanced Mobile Broadband (eMBB) and Ultra-Reliable Low-Latency Communication (URLLC) use cases. In this context, the thesis proposes a practical approach that ensures preventing interference and preserving the integrity of data traversing. The approach involves the integration of state-of-the-art Virtual Private Network (VPN) solutions with Virtualized Network Functions (VNFs)/Cloud-Native/Containerized

Network Functions (CNFs) via the MANO entity in the network to establish isolated network slices formed by the VNFs/CNFs.

Third, the research follows up on the intersection of network slicing and the Radio Access Network (RAN) domain. The coexistence of eMBB and URLLC use cases within the RAN domain requires innovative solutions. Through a precise analysis of numerologies within the 5G-New Radio (5G-NR) architecture and the innovative integration of the puncturing technique, this research presents effective solutions that facilitate the harmonious coexistence of these use cases. By doing so, this thesis strengthens the potential of network slices to cater to an array of demanding requirements.

In summary, this thesis reflects the evolving landscape of network architecture. The thesis exploration of network slicing in the context of 5G and beyond represents the importance of adaptability and customization in the realm of telecommunications. Through practical insights and theoretical support, the study reveals the process of network slice creation, highlights the necessity of isolation, and unlocks the potential of network slices to seamlessly accommodate diverse use cases. As the telecommunications world continues to extend, this thesis illuminates a path toward resilient, efficient, and harmonious network infrastructure.

Preface

This dissertation is submitted in partial fulfillment of the requirements for the degree Philosophiae Doctor (Ph.D.). The Ph.D. thesis work was carried out in the Department of Information Security and Communication Technology (IIK), at the Faculty of Information Technology and Electrical Engineering (IE), Norwegian University of Science and Technology (NTNU). The research presented here was conducted under the supervision of Associate Professor Katina Kralevska and the co-supervision of Professor Danilo Gligoroski.

The thesis takes the form of a paper collection. The included papers are the contributions of this Ph.D. in scientific conferences or journals. Page numbering in the papers has been reformatted in order to have consistent numbering within the thesis and deviate from the published versions.

Acknowledgements

Firstly, I want to express my gratitude to a higher power for granting me the patience and strength needed to overcome the challenges I've faced. Many people, both directly and indirectly, have contributed to my journey toward this Ph.D. achievement.

I would like to extend my heartfelt thanks to my supervisor, Associate Professor Katina Kralevska. Her decision to trust me with participation in this fascinating research field is a testament to her faith in me. Throughout my Ph.D. program, her wise guidance, unwavering support, and constant encouragement have been pivotal to my progress, and her advice has been precious.

I am equally thankful to my co-supervisor, Professor Danilo Gligoroski, whose guidance, inspiration, and significant contributions have substantially boosted our collaborative work. I'm also grateful to our administrative staff, especially Mona, Maria, and Pål, whose offices have always been open to address any queries.

In the realm of academia, my colleagues have played an essential role in creating a balanced atmosphere that fosters both work and enjoyable interactions. I want to thank Befe, Mayank, Murad, Sonu, Enio, Ergys, Mattia, Charles, and Kalpanie for the wonderful moments we've shared. I'd also like to acknowledge the thought-provoking discussions with Professor Bjarne Emil Helvik and Professor Yuming Jiang. Their insights have significantly influenced the trajectory of my Ph.D. journey.

Lastly, I owe a debt of gratitude to my family, whose unwavering support has carried me through the challenges. You are the reason for my life, and without you, this journey would have lacked meaning and purpose.

Ali Esmaeily
Trondheim, September 2023

This thesis is dedicated to

the memory of my father,
my beautiful mother,
my wonderful sisters,
and my lovely wife.

Contents

Abstract . . . . . **i**
Preface . . . . . **iii**
Acknowledgements . . . . . **iv**
Contents . . . . . **vi**
Figures . . . . . **viii**
Tables . . . . . **ix**
Acronyms . . . . . **x**

I Research Overview **xvi**
1 Introduction **1**
1.1 Motivation 1
1.2 Highlights of Main Contributions 4
1.3 Thesis Structure 6
2 Background **7**
2.1 Network Slicing Fundamentals 7
2.1.1 Pre-Network Slicing 7
2.1.2 Network Slicing Concept, Terminology, and Lifecycle 8
2.2 Network Slicing Features and 5G Use Cases 12
2.2.1 Network Slicing Features 12
2.2.2 5G Use Cases 13
2.3 Network Slicing Enabling Technologies and Providing E2E Network Slices 14
2.3.1 Network Slicing Enabling Technologies 14
2.3.2 Providing E2E Network Slices 17
2.4 Isolation in Network Slicing 17
2.5 Resource Allocation in Network Slicing 21
2.5.1 Key Requirements for Resource Allocation in Network Slicing 21
2.5.2 Resource Allocation Approaches in Network Slicing 22
2.5.3 Comparison of Resource Allocation Approaches in Network Slicing 22
2.5.4 Resource Allocation Realization in Network Slicing 23

2.6 5G New Radio 24
2.6.1 5G-NR terminology 24
2.6.2 Waveform and Scalable Numerologies in 5G-NR 25
2.6.3 Multiple numerologies for diverse service requirements of 5G use cases 25
2.6.4 Orthogonal Multiple Access vs. Non-Orthogonal Multiple Access RAN Slicing in 5G-NR 27
3 Related Works 30
3.1 State-of-the-art Network Slicing Proof-of-Concepts 30
3.1.1 Small-scale CN slicing testbeds 31
3.1.2 Small-scale RAN slicing testbeds 32
3.1.3 Small-scale E2E slicing testbeds 33
3.2 State-of-the-art solutions for Network Slicing Isolation 35
3.2.1 Simulation-based solutions 35
3.2.2 Implementation-based solutions 36
3.3 State-of-the-art solutions for eMBB and URLLC coexistence in the 5G-NR 37
4 Research Contributions 42
4.1 Research Design 42
4.1.1 Motivation 42
4.1.2 Research Questions and Research Objectives 43
4.1.3 Research Methodology 44
4.2 Contributions of the papers 46
4.2.1 Summary of the Results Contributing to the Thesis 48
5 Concluding Remarks 53
5.1 Conclusion 53
5.2 Future directions 55
Bibliography 56

II Included Papers 71
Paper I 72
Paper II 99
Paper III 107
Paper IV 114
Paper V 125
Paper VI 133

Figures

1.1 5G and 4G KPIs comparison [4]. 2
1.2 6G and 5G KPIs comparison [5]. 2
1.3 Different network slices on a shared physical infrastructure [7]. . 3

2.1 MOCN vs. GWCN [16]. 8
2.2 Distinct network slices created on M-InPs providing several services from different tenants to the end users [21]. 11
2.3 The lifecycle of a NSI [8]. 12
2.4 5G three main use cases [25]. 14
2.5 ETSI NFV framework [31]. 16
2.6 E2E network slice traverses over all network domains [34]. 18
2.7 Network slice isolation concept areas. 18
2.8 5G-NR numerology selection criteria according to frequency ranges, latency requirements, and deployment types [58]. 26
2.9 Different multiple access schemes, OMA: (a) FDMA, (b) TDMA, (c) CDMA, (d) OFDMA, and NOMA: (e) Code-/Power-domain [64]. 28
2.10 RAN slicing for three main use cases via (a). OMA and (b). NOMA in the 5G-NR [65]. 28

4.1 Papers' interconnections and their links to research questions presented in 4.1.2. 47

Tables

2.1 5G-NR flexible numerologies [56]. 25

4.1 Included Papers mapped to research methodologies presented in 4.1.3. 47
4.2 Not included Paper mapped to research methodologies presented in 4.1.3. 47

Acronyms

1G First Generation. 27

2G Second Generation. 27

3G Third Generation. 27

3GPP 3rd Generation Partnership Project. 3

4G Fourth Generation. 1

5G Fifth Generation. i, 1

5G-NR 5G-New Radio. ii, 5, 24

6G Sixth Generation. 1

AAA Authentication, Authorization, and Accounting. 20

AI Artificial Intelligence. 34

API Application Programming Interface. 15, 42, 43

B5G Beyond 5G. i, 1

CAPEX Capital Expenditure. 1

CDMA Code Division Multiple Access. 27

CDN Content Delivery Network. 9

CN Core Network. i, 5, 7

CNF Cloud-Native/Containerized Network Function. i, ii, 5, 9

CP Control Plane. 14

CSC Communication Service Customer. 10

CSP Communication Service Provider. 9

CyP Cyclic Prefix. 24

DC Data Center. 31

DDos Distributed Denial-of-Service. 36

DL Downlink. 25

DP Data Plane. 14

E2E End-to-End. 3, 4

EC Edge Computing. 15

eMBB Enhanced Mobile Broadband. i, 5

eNB Evolved Node B. 31

EPC Evolved Packet Core. 31

ETSI European Telecommunications Standards Institute. 3

FDMA Frequency Division Multiple Access. 27

FR Frequency Range. 25

GWCN Gateway Core Network. 7

HNF Hybrid Network Function. 9

IBN Intent-Based Networking. 33

IMS IP Multimedia System. 31

INI Inter-Numerology Interference. 39

InP Infrastructure Provider. 9

ISI Inter-Symbol Interference. 24

KPI Key Performance Indicator. 1

LCM Lifecycle Management. 4, 16

LTE Long Term Evolution. 31

M(V)NO Mobile (Virtual) NO. 9

M-InP Multiple-Infrastructure Provider. 3, 9

MAC Media Access Control. 24

MANO Management and Orchestration. i, ii, 4, 10, 15

MEC Multi-access Edge Computing. 4, 15

MIMO Multi-Input Multi-Output. 25

MIoT Massive Internet of Things. 1

MME Mobility Management Entity. 7, 8

mMTC massive Machine Type Communication. 14

MNO Mobile NO. 9

MOCN Multi-Operator Core Network. 7

MPLS Multi-Protocol Label Switching. 19

MSP Mobile Service Provider. 9

NF Network Function. 8

NFD Network Function Descriptor. 10

NFV Network Function Virtualization. 3, 4, 9

NFVI Network Function Virtualization Infrastructure. 15

NFVO Network Function Virtualization Orchestrator. 16

NGMN Next Generation Mobile Network. 3

NIM Network Infrastructure Manager. 31

NO Network Operator. 9

NOMA Non-Orthogonal Multiple Access. 5, 27

NS Network Service. 5, 10

NSD Network Service Descriptor. 10

NSI Network Slice Instance. 5

NSID Network Slice Instance Descriptor. 10

NSIT Network Slice Instance Template. 10

NSLP Network Slice Provider. 9

NSSI Network Slice Subnet Instance. 10

OAI OpenAirInterface. 48

OFDM Orthogonal Frequency Division Multiplexing. 24

OFDMA Orthogonal FDMA. 27

OMA Orthogonal Multiple Access. 27

OPEX Operational Expenditure. 1

OSM Open Source MANO. 48

OTT Over the Top. 9

PaaS Platform-as-a-Service. 34

PNF Physical Network Function. 8, 9

PNO Public NO. 9

PoC Proof-of-Concept. 30

PRB Physical Resource Block. 19

QoE Quality of Experience. 31

QoS Quality of Service. i, 5, 8

RAN Radio Access Network. ii, 7

RAT Radio Access Technology. 18

RE Resource Element. 24

RIS Reconfigurable Intelligent Surface. 40

RO Research Objective. 6, 43

RQ Research Question. 6, 43

RRC Radio Resource Control. 24

RRM Radio Resource Management. 32

SA Service Assurance. 31

SCS Subcarrier Spacing. 24

SDN Software Defined Networking. 4, 14

SI Service Instance. 10

SIC Successive Interference Cancellation. 29

SLA Service Level Agreement. 8

SlaaS Slice-as-a-Service. 31

SLO Service Level Objective. 10

srsLTE Software Radio Systems LTE. 48

TDD Time-Division Duplex. 24

TDMA Time Division Multiple Access. 27

TN Transport Network. 8

TTI Transmission Time Interval. 24

UE User Equipment. 10

UL Uplink. 25

UMTS Universal Mobile Telecommunications System. 7

UP User Plane. 14

URLLC Ultra-Reliable Low-Latency Communication. i, 5

VIM Virtualized Infrastructure Manager. 16

VM Virtual Machine. 15

VNF Virtualized Network Function. i, ii, 4, 9

VNFM Virtualized Network Function Manager. 16

VNO Virtual NO. 9

VPN Virtual Private Network. i

VPNaaS Virtual Private Network-as-a-Service. 5, 49

WDM Wavelength Division Multiplexing. 19

Part I

Research Overview

Chapter 1

Introduction

1.1 Motivation

The Fifth Generation (5G) of mobile networks and Beyond 5G (B5G), such as the Sixth Generation (6G) and later generations, are expected to provide various services with heterogeneous requirements compared to the Fourth Generation (4G) and previous generations of cellular networks. In addition to providing services such as UltraHD and 360-degree video streaming, there are novel services that involve innovative healthcare delivery, smart transportation systems, and smart grids. Furthermore, Massive Internet of Things (MIoT) devices, which exponentially spread within a small geographical area, are another type of distinct service in the B5G paradigm [1–3]. Key Performance Indicators (KPIs) are various measurable metrics that display how efficiently a network provides its services. In terms of the required KPIs in 5G,

- video streaming services demand a very high peak data rate of up to several Gbps,
- healthcare, smart transportation, and smart grids necessitate an extremely low delay of less than 1 ms and extra high reliability of 99.999%, and
- MIoT devices are distributed up to 1 million devices/km^2.

Figures 1.1 and 1.2 outline the comparison of the main KPIs for different services in 5G vs. 4G and 6G vs. 5G, respectively.

The major challenge with providing such diverse services is that the physical infrastructure resources are scarce. Thus, these resources need to be employed intelligently to deliver such services. Efficient *Network Sharing* [6] is considered a traditional solution. Through network sharing, multiple operators can participate in infrastructure resource sharing according to their agreed resource allocation plans. This strategy can assist an operator in reducing Capital Expenditures (CAPEXs) and Operational Expenditure (OPEX). As a further development of

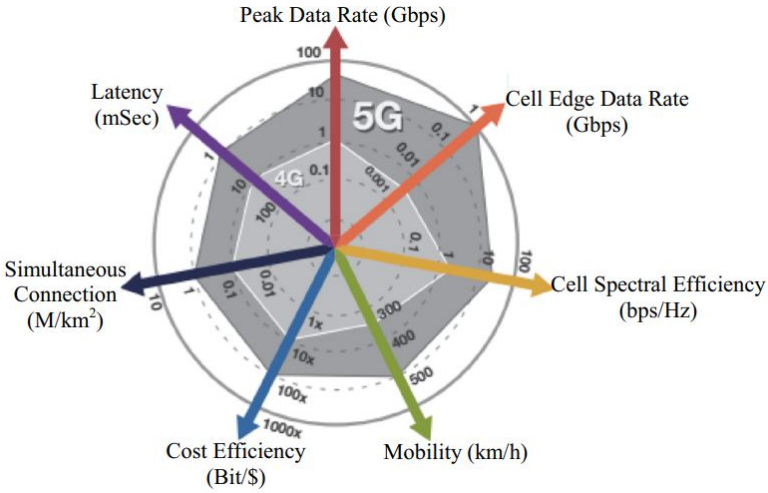


Figure 1.1: 5G and 4G KPIs comparison [4].

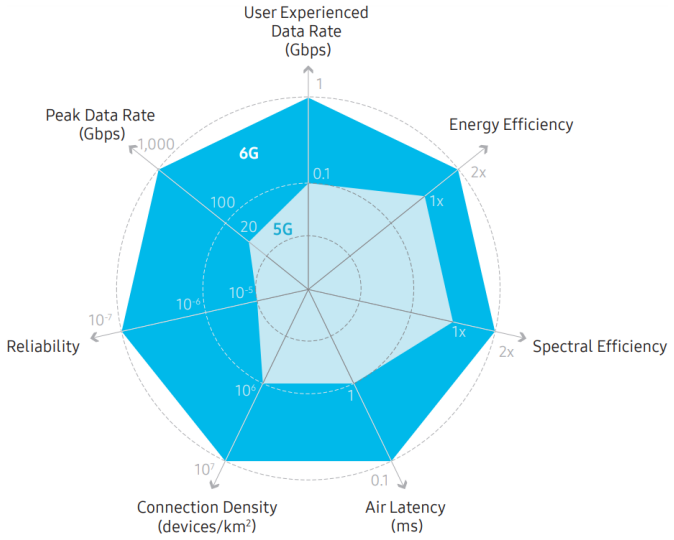


Figure 1.2: 6G and 5G KPIs comparison [5].

network sharing, *Network Slicing* yields the flexibility and dynamicity of allocating the required and appropriate amount of physical resources to the services over the same physical infrastructure simultaneously. In fact, network slicing leverages the running of multiple logical networks on top of physical infrastruc-

ture. Figure 1.3 illustrates three different slices as separate logical networks running on a shared physical infrastructure.

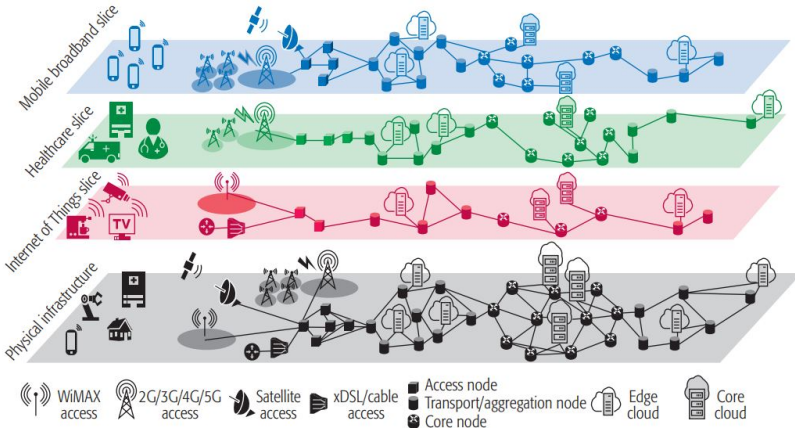


Figure 1.3: Different network slices on a shared physical infrastructure [7].

Although the vision and objectives of 5G are transparent, creating diverse partial or End-to-End (E2E) network slices on the physical infrastructure can be very complex. Moreover, this procedure becomes even more complicated when it comes to providing management and orchestration of (E2E) network slices that may be implemented on Multiple-Infrastructure Providers (M-InPs), efficient and dynamic resource allocation of network slices in different network domains, and separating and isolating the network slices. Consequently, it is evident that network slicing encounters numerous challenges on its way to being fully implemented in 5G and B5G. Due to the complexity and challenges of network slicing implementation, global efforts and initiatives cooperate to propose innovative solutions in order to tackle critical challenges towards network slicing realization. Several standardization organizations such as the 3rd Generation Partnership Project (3GPP), the Next Generation Mobile Network (NGMN), and the European Telecommunications Standards Institute (ETSI) provide specifications for various aspects involved in network slicing establishment. The 3GPP has granted a technical report [8] associated with the 5G network slicing concept, its use cases, its main requirements, and possible solutions for the orchestration and management of network slices. The NGMN has provided the standardization of 5G network slicing [9]. The ETSI has released a document [10] which exposes the relationship between Network Function Virtualization (NFV) technology and 5G network slicing. Besides, network slicing has also gained much attention among the expanding community of researchers in academia and industry to develop novel and unconventional approaches to overcome the chal-

lenges for network slicing realization in B5G.

1.2 Highlights of Main Contributions

This section briefly introduces the main contributions included in this Ph.D. thesis as follows:

- **Paper 1**:

  A. Esmaeily, K. Kralevska, "Small-Scale 5G Testbeds for Network Slicing Deployment: A Systematic Review," Wireless Communications and Mobile Computing, 2021.

  This paper comprehensively looks at small-scale network slice testbeds. It starts with an overview of network slicing technologies, Software Defined Networking (SDN), Network Function Virtualization (NFV), Cloud computing, Multi-access Edge Computing (MEC), and their roles in the Lifecycle Management (LCM). The paper introduces major open-source software packages for network slicing, aligning them with the Network Function Virtualization (NFV) Management and Orchestration (MANO) framework, and outlines design criteria for small-scale testbed deployment. The paper reviews leading small-scale testbeds using these criteria and addresses common deployment challenges.

- **Paper 2**:

  A. Esmaeily, K. Kralevska, D. Gligoroski, "A Cloud-based SDN/NFV Testbed for End-to-End Network Slicing in 4G/5G," 6th IEEE Conference on Network Softwarization (NetSoft), 2020.

  Extending the concepts from Paper 1, Paper 2 details the 5GIIK testbed. It employs open-source software packages to deploy an End-to-End (E2E) network. Notably, the 5GIIK testbed offers dynamic slice provisioning and real-time Virtualized Network Function (VNF) monitoring. The paper's main contribution lies in its comprehensive architecture, which holds promise for future contributions.

- **Paper 3**:

  S. Kielland, A. Esmaeily, K. Kralevska, and D. Gligoroski, "Secure Service Implementation with Slice Isolation and WireGuard," IEEE International

Mediterranean Conference on Communications and Networking (MeditCom), 2022.

Paper 3 aims to enhance traffic isolation solutions in emerging use cases. It introduces an innovative Virtual Private Network-as-a-Service (VPNaaS) integration with the NFV MANO framework, ensuring data confidentiality in Core Network (CN) slices. The integration offers isolation for VNF lifecycle management. The results show potential for performance, security, and simplified configuration.

- **Paper 4**:

  A. Esmaeily and K. Kralevska, "Orchestrating Isolated Network Slices in 5G Networks," under review in EURASIP Journal on Wireless Communications and Networking, 2023.

  Paper 4 extends the idea of Paper 3 and addresses the need for comprehensive VPNaaS solution integration in 5G. Paper 4 then evaluates the performance of VPNaas integration with the NFV MANO framework in Cloud-Native/Containerized Network Function (CNF) environment. The paper defines VNF, Network Service (NS), and Network Slice Instance (NSI) levels' descriptors and implements a cloud-native 5G standalone architecture. Results highlight the performance of some VPN solutions integrated with the NFV MANO framework, which depends on the 5G use case in which such a VPN is going to be utilized.

- **Paper 5**:

  A. Esmaeily, K. Kralevska, T. Mahmoodi, "Slicing Scheduling for Supporting Critical Traffic in Beyond 5G," 19th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2022.

  This paper explores Enhanced Mobile Broadband (eMBB) and Ultra-Reliable Low-Latency Communication (URLLC) slice coexistence over 5G-New Radio (5G-NR), addressing their distinct Quality of Service (QoS) requirements. It proposes an efficient scheduling strategy using Non-Orthogonal Multiple Access (NOMA) resource allocation and puncturing techniques, focusing on enhancing the individual average date rate for each eMBB user while fulfilling URLLC requirements.

- **Paper 6**:

  A. Esmaeily, H. V. K. Mendis, T. Mahmoodi, K. Kralevska, "Beyond 5G Resource Slicing with Mixed-Numerologies for Mission Critical URLLC and eMBB Coexistence," IEEE Open Journal of the Communications Society, 2023.

  Building on Paper 5, Paper 6 combines the puncturing technique with various 5G-NR numerologies. It categorizes URLLC traffic into classes, with the objective of maximizing the sum rate of the eMBB users and meeting URLLC latency and reliability demands. The paper divides the resource allocation problem into sub-problems and introduces an optimization algorithm, promising efficient resource allocation for coexisting eMBB and URLLC slices.

1.3 Thesis Structure

The thesis is structured in two parts. Part I provides an introduction to the topic of the thesis, and it gives an insight into the outcomes and contributions that have been accomplished. Part II presents the papers included in the thesis. In addition to the introductory chapter, the following chapters are included in Part I:

- **Chapter 2** covers the necessary background of the research domains comprised in this thesis.
- **Chapter 3** includes the related works performed in the thesis domain.
- **Chapter 4** comprises two major sections. The first section presents the research design, which includes the Research Questions (RQs), Research Objectives (ROs). The expressed RQs and ROs are specified within the scope of the thesis topic according to the open research questions, knowledge gaps, and challenges highlighted in Chapter 3. Then the employed research methodology is presented. The second section exposes the main contributions of the thesis, a summary of the content of each paper, how the papers are interconnected, and how they are linked to the *RQs*.
- **Chapter 5** concludes the thesis and highlights the potential future research directions.

Chapter 2

Background

This chapter provides an overview of the network slicing fundamentals with main use cases and their requirements, network slicing enabling technologies along with management and orchestration, implementing slicing, resource allocation, and isolation of slices.

2.1 Network Slicing Fundamentals

2.1.1 Pre-Network Slicing

The virtualization concept, which originated in the 60s, was first utilized to develop IBM operating systems [11]. Basically, virtualization refers to a technique to build a virtualized form of a physical entity with the required computing, networking, and storage resources via a set of software-based procedures, which results in a virtual entity with corresponding computing, networking, and storage resources. Virtualization was the first option to utilize in data centers in the 70s [12] and overlay networks in the 80s [13]. In fact, overlay networks that connect network nodes via logical connections over a physical infrastructure can be considered as the starting point of network sharing and, thus, network slicing.

3GPP introduced network sharing in Release 99 of the Universal Mobile Telecommunications System (UMTS) networks [14]. Network sharing started with sharing on-site Radio Access Network (RAN) pieces of equipment, which is referred to as passive network sharing. The second approach focused on sharing base stations, antennas, core network, mobile backhaul equipment, and also contractual-based spectrum, which is referred to as active network sharing. Active network sharing [15], as illustrated in Figure 2.1, is split up into 1) Multi-Operator Core Network (MOCN) shares the RAN and the frequency spectrum in which each involving operator owns a separate Core Network (CN) in 4G networks and 2) Gateway Core Network (GWCN) shares the RAN and also the Mo-

bility Management Entity (MME) of the CN. This latter approach is affordable since more network entities can be shared.

Network slicing is a considerable move forward in network sharing. Accommodating diverse services in B5G requires a high level of flexibility, dynamicity, and programmability in allocating the appropriate amount of physical infrastructure resources to services, and network slicing grants such capabilities to B5G.

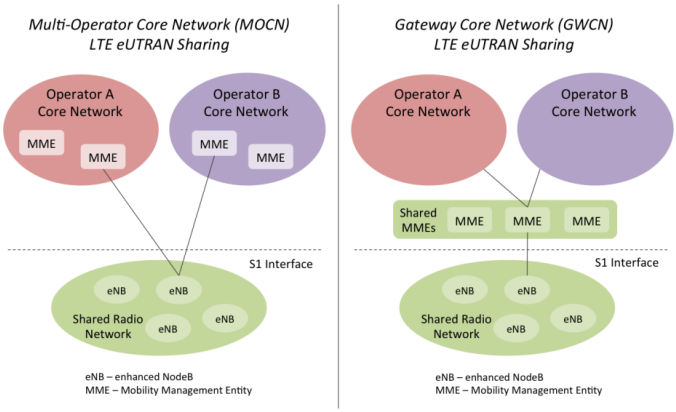


Figure 2.1: MOCN vs. GWCN [16].

2.1.2 Network Slicing Concept, Terminology, and Lifecycle

Network slicing in the 5G network area is a newly defined concept. Network Functions (NFs) [17], are considered as the building blocks of network slices. Network slicing leverages the operation of multiple self-contained NFs on top of common physical infrastructure. In fact, an E2E network slice chains specific NFs in a particular order to form a virtualized, isolated, software-oriented, and programmable environment. Such E2E network slice delivers a distinct network service according to Quality of Service (QoS) specifications expressed in the Service Level Agreement (SLA). These specifications are then translated into slice templates in order to allocate the required resources in the RAN, Transport Network (TN), and CN domains [18]. Hence, network slicing facilitates technical and business innovation by creating customized and distinct network services.

Some of the terminologies in the network slicing context are defined as follows:

- **NFs and Physical Network Functions (PNFs)**: NFs are constructive operational components (networking appliances) such as routers, switches,

firewalls, and load balancers with particular functionalities in the network infrastructure, and they have well-defined exterior interfaces for communicating with each other. If the NFs are implemented on dedicated vendor-specific hardware, they are referred to as PNFs.

- **Softwarized NFs**: With the help of Network Function Virtualization (NFV) as a network architecture framework, PNFs are implemented in software mainly via two prominent approaches: 1) Virtualized NFs (VNF) deployed on virtual machines, and 2) Cloud-Native/Containerized NFs (CNFs) deployed on containers. These VNFs and CNFs are later implemented in data centers or cloud-based environments that run on top of general-purpose (vendor-neutral) hardware.
- **Hybrid Network Functions (HNFs)**: HNFs are composed of PNFs, VNFs, and CNFs.
- **Infrastructure resources**: They are heterogeneous pieces of hardware and required software to host and connect NFs for creating a network slice. In particular, infrastructure resources include hardware and software computing, storage, and networking resources, along with physical assets of the RAN domain. By applying virtualization technology, such resources have to be logically abstracted to be employed in the same fashion as physical resources. Most of the time, infrastructure resources are referred to as *resources*.
- **Infrastructure Provider (InP)**: It is the central entity that owns a specific physical network inside an administrative domain that offers the infrastructure resources via programming interfaces. For creating an E2E slice, there may be more than one involving InP, which is referred to as Multiple-Infrastructure Providers (M-InPs).
- **Network Operator (NO)**: It grants wired and wireless communications services, and it owns the InP or controls its resources to deliver services to Public NOs (PNOs)), Mobile NOs (MNOs), and Virtual NOs (VNOs).
- **Network Slice Provider (NSLP) [19]/Mobile Service Provider (MSP))**: It takes the responsibilities of monitoring, controlling, managing, and orchestrating the corresponding resources from the InP that the network slice demands [20]. Most of the time, NSLP and InP are the same entity. We also consider them as the same entity in this thesis.
- **Business customer/Communication Service Provider (CSP)/Tenant**: It leases virtual resources from (M-)InP(s) in the form of distinct network slice(s) via which the tenant can realize, manage, and provide specific services to its end users. Enterprise or specialized industries like Mobile (Virtual) NOs (M(V)NOs), Verticals (e.g. eHealth, automotive services), Over the Top (OTT)(e.g. Content Delivery Networks (CDNs)) are considered as typical tenants. One of the main objectives of B5G is first to facilitate the

co-existence of multiple tenants and second to administrate the association and intercommunication between them. This capability describes the so-called multi-tenancy environment, which indicates that a single instance of the software and its hosting InP serve multiple tenants.

- **End users/User Equipment (UE)/Communication Service Customers (CSCs)**
  /User Terminals: They consume the services provided by the tenants and are considered as the last link in the service provisioning chain.
- **(E2E) network slice SLA**: In the context of network slicing, the (E2E) network slice SLA refers to a formal agreement between the tenant and the NO with the scope of evaluating and verifying service characteristics and also responsibilities of both parties. An SLA includes information regarding service delivery, billing, legal issues, etc.
- **(E2E) network slice Service Level Objective (SLO)**: It is a section within an SLA that defines specific metrics according to the QoS requirements, such as precise performance requirements for the service, that have to be delivered to the tenants from the (M-)InP(s).
- **Service Instance (SI)**: They represent communication services that are established from tenants to end users. Each service is realized by an SI.
- **Network Slice Instance Template (NSIT)/Catalogue**: It holds the required attributes and SLO specifications that can characterize a particular network slice use case (see Section 2.2.2).
- **Network Slice Instance Descriptor (NSID)**: It interprets the NSIT content into all the essential technical networking information that is needed for network slice deployment. Each NSID includes 1) the required information of Network Service (NS) that create a specific (E2E) NSI, 2) the form of interconnection between them, 3) the demanded resources in order for the slice to operate according to the (E2E) network slice SLO obligations and fulfill the QoS necessities. Each NSID, in turn, is split down into one or several Network Service Descriptors (NSDs). Each NSD contains the required technical information, such as virtual links and connection points between the (P/V/C/H) NFs, which form that specific NS. Finally, each NSD is further broken into the (P/V/C/H) Network Function Descriptors (NFDs) in which more detailed information is included, such as a precise amount of network, computing, and storage resources required for each particular (P/V/C/H) NF to operate appropriately. Such resources in different network domains are allocated by the Management and Orchestration (MANO) entity of the network (see Section 2.3.1).
- **Network Slice Subnet Instance (NSSI)**: It is a collection of (P/V/C/H) NFs with the required resources for regular operation in a particular subnet. The combination of several NSSIs creates an NSI. Each involving NSSI

must participate in the overall QoS of an NSI. A Subnet can be considered a networking domain. Thus, a combination of RAN NSSI, TN NSSI, and CN NSSI paired together can form an E2E NSI.

- **Resource Federation and Service Federation**: Resource federation refers to enabling the interconnection functionality between multiple administrative domains of M-InPs to effectively utilize their resources due to the resource limitation of a single InP. Such functionality grants creating NSIs that span across M-InPs, and convey services that are established and federated over M-InPs.

Figure 2.2 demonstrates some of the terminologies.

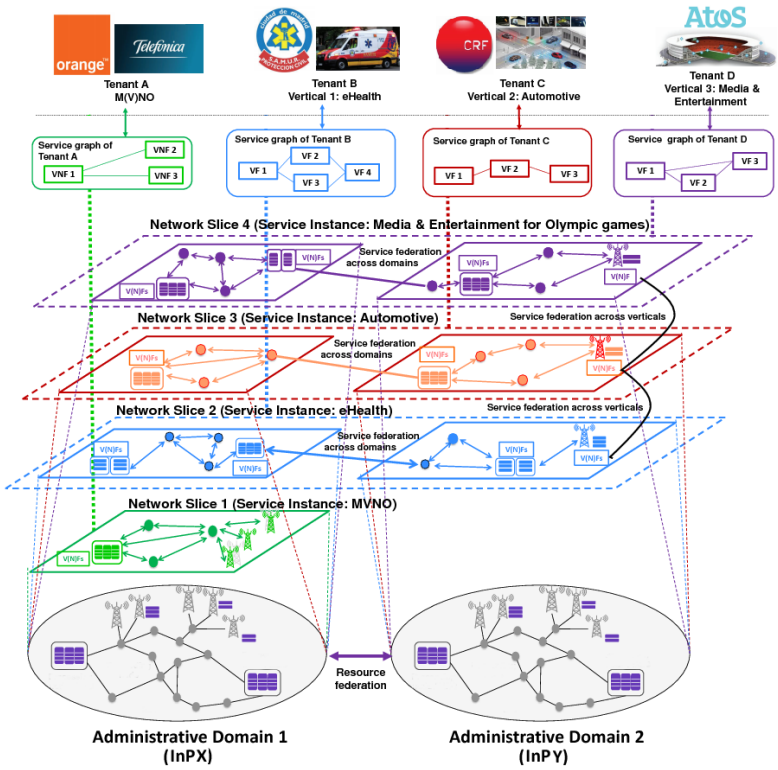


Figure 2.2: Distinct network slices created on M-InPs providing several services from different tenants to the end users [21].

Each NSI experiences a period of existence, called the lifecycle of an NSI, consisting of several phases from the time a slice is established until the moment the slice is deleted from the network. The preparation phase and the lifecycle of

an NSI are illustrated in Figure 2.3 and summarized as follows:

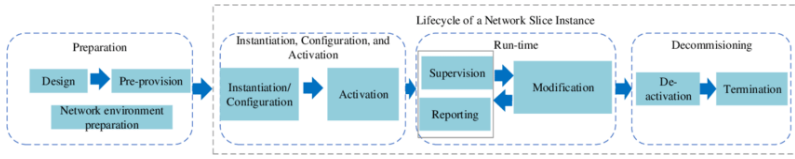


Figure 2.3: The lifecycle of a NSI [8].

- **Preparation**: This phase incorporates the required arrangements in the network prior to the NSI creation, such as defining and verifying NSID, analyzing and assessing the network slice requirements, and capacity planning.
- **Instantiation, Configuration, and Activation**: In this phase, an NSI according to the NSID with all the demanded resources is established. Now, the created NSI holds a communication service.
- **Run-time**: This phase comprises continuous performance monitoring and comparing the results with the KPI values in order to perform possible modification and reconfiguration in the allocated resources to the NSI.
- **Decommissioning**: In this phase, the network slice is de-activated, and by releasing out the allocated resources and returning them to the shared pool of resources, the network slice is terminated.

2.2 Network Slicing Features and 5G Use Cases

2.2.1 Network Slicing Features

There are some fundamental principles for network slicing technology and its operations, which are listed in the following:

- **Automation**: This capability yields on-demand configuration in network slicing. In this case, apart from conventional desired KPIs such as data rate, latency, jitter, etc., new information regarding the operational duration and periodicity of a network slice is also included in the SLA.
- **Customization**: This feature presents efficient network resource utilization according to the service requirements for diverse slices.
- **Elasticity**: It is another crucial feature that is related to network slice operation and the number of allocated resources to the network slice in order to guarantee the SLA while changing network conditions, the number of users, and also user mobility.

- **E2E**: It is an essential characteristic of network slicing to provide various network services from the service providers' locations toward the end users' premises. Providing E2E network slices can be quite challenging since 1) an E2E network slice spans across RAN, TN, and CN domains with heterogeneous involving technologies, and 2) an E2E network slice may develop on M-InPs.
- **Isolation**: It is one of the major properties of network slicing that refers to eliminating any negative impact of other slices or even InP from a particular slice in order for the slice to continue its regular operation. Although isolation guarantees for network slices to operate appropriately as defined in the SLA, providing isolation may lead to inefficient utilization of network resources due to the multiplexing gain reduction.
- **Programmability**: This attribute brings flexibility, dynamicity, and scalability to network slices by controlling and managing the number of allocated resources in different network domains to a particular slice via open APIs, which in fact, results in expediting on-demand service customization and network resource elasticity.
- **Hierarchical resource abstraction**: This attribute in network slicing leads to a recursive virtualization scheme to allow a higher level and more extensive range of resource abstraction in a hierarchical fashion in order to enhance service provisioning.

2.2.2 5G Use Cases

As discussed before, B5G provides services that have diverse performance requirements in terms of data rate, latency, reliability, distribution in a small geographical area, etc. Each of these services belongs to a specific category which is mainly referred to as *use cases* that demand different types of network slices. These use cases with their required KPIs are presented in Figure 2.4 and introduced in the following:

- **enhanced Mobile Broadband (eMBB) [22]**: Such a use case concentrates on providing high data rates in downlink and uplink transmissions, accommodating large data traffic volumes and UEs' connectivity per geographical area, granting wide area coverage and connectivity, and considering high UE mobility. Hence, eMBB needs a high capacity, typically low delay, and also high availability in the network.
- **Ultra-Reliable Low Latency Communication (URLLC) [23]**: As the term explains, this use case focuses on supporting services demanding ultra-reliability along with low latency communications with 99.999% reliability, packet loss of almost one packet out of every 10,000 packets, and maximum 1 ms delay. URLLC facilitates implementing critical services, con-

sisting of a high level of automation, management, control, public safety, disaster recovery, etc. In addition, the URLLC use case requires a very high level of slice isolation and prioritization in data transmission.

- **massive Machine Type Communication (mMTC) [24]**: This use case facilitates data transmission for a very high density of UEs spread in a small geographical region with usually fixed and non-time critical service demands and easiness in their operations, which leads to long battery life. Furthermore, the mMTC use case helps to accommodate diverse connectivity among an enormous number of UEs for supporting high scalability.

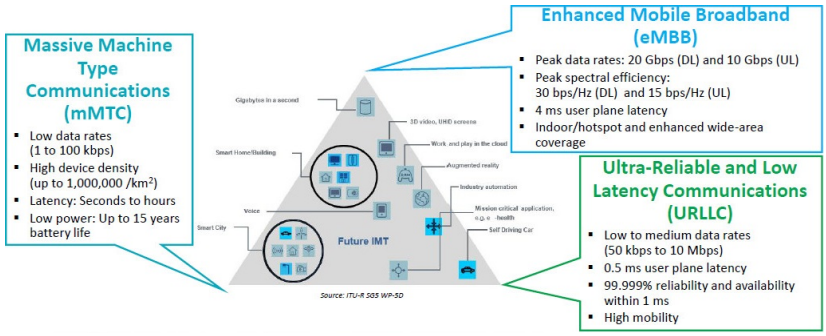


Figure 2.4: 5G three main use cases [25].

2.3 Network Slicing Enabling Technologies and Providing E2E Network Slices

2.3.1 Network Slicing Enabling Technologies

The approach towards softwarization, virtualization, and cloudification as enabling technologies of network slicing has brought tremendous progress and benefits in terms of programmability, flexibility, and innovative ideas to service provisioning. These key enablers for network slicing are introduced in the following:

- **Software Defined Networking (SDN) [26]**: SDN facilitates network management by decoupling the Control Plane (CP) from the Data Plane (DP) (also referred to as User Plane (UP)). In the core of the CP, a logically centralized intelligent entity, called the *SDN controller*, grants on-demand programmability and dynamic configuration for the dummy forwarding

devices in the DP. Utilizing SDN leads to providing flexibility and service-oriented policies, which are crucial for network slicing realization.

- **Cloud computing [27]**: Cloud computing provides storage, computational, and networking resources in order to enable network slicing on single or multiple platforms. Cloud computing offers access to remote resources in shared pools that are administered over the Internet. Cloud computing is based on two major orientations: 1) Cloud-based applications that lead to migrating legacy applications, which were installed on end users' devices or on the organizations' IT infrastructure, towards cloud-based servers in order to present the applications over web browsers, and 2) Cloud-native applications, which introduce those applications that are essentially designed and developed to utilize the benefits of the cloud environment such as modularity, scalability, and Application Programming Interfaces (APIs) integration.
- **Edge Computing (EC) [28]**: By pushing the processing and computing tasks to the edges of the network, EC empowers computing applications and data management close to end users to facilitate delay-sensitive service implementation. The ETSI Multi-access Edge Computing (MEC) and Fog computing are two of the standard EC implementations. MEC [29] concentrates on RAN and fixed AN, and it authorizes third parties to instantiate, control, and manage different services from the edge of the network. Fog computing [30] was presented by Cisco to accommodate data transmission between wirelessly connected devices in a MIoT system.
- **Network Function Virtualization (NFV) [17]**: NFV enables the deployment of initially hardware-implemented PNFs on virtual environments, called VNFs, leveraging the cost-efficiency advantage of Cloud computing. VNFs are deployed on Virtual Machines (VMs) and/or Containers that can be chained together in a particular order on a cloud environment granting specific network services. As shown in Figure 2.5, apart from VNFs, the ETSI NFV architectural framework [31] includes:

    - **Network Function Virtualization Infrastructure (NFVI)**: It presents the logical and physical environment for deploying the VNFs, which includes virtualized storage, computing, and networking resources, with their corresponding supporting hardware components.
    - **Management and Orchestration (MANO)**: It is capable of controlling, managing, and orchestrating the VNFs running on the NFVI. This framework facilitates an on-demand network services establishment considering a distinct collection of NFs, which can be PNFs, VNFs, and/or CNFs, depending on the network service requirements. In order to provide service instantiation process and delivery, the MANO

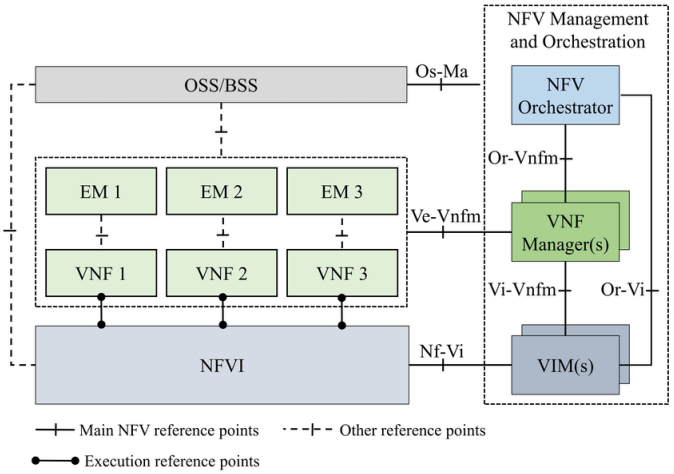


Figure 2.5: ETSI NFV framework [31].

is split up into three main building blocks:

- **Virtualized Infrastructure Manager (VIM)**: It enables policies for controlling the NFVI resources within an InP. VIM is also responsible for receiving and measuring fault and performance information of NFVI resources. Consequently, VIM can control and oversee NFVI resource allocation to the available VNFs/CNFs. In the context of network slicing, the VIM allocates the NFVI resources and controls their association in the service chaining and traffic steering.
- **Virtualized Network Function Manager (VNFM)**: It is in charge of the Lifecycle Management (LCM) of one or multiple VNFs of an NSI. The LCM process comprises configuring, instantiating, maintaining, and terminating the VNFs according to the corresponding descriptors.
- **Network Function Virtualization Orchestrator (NFVO)**: It grants resource and service orchestration in the network. NFVO has two internal sections: 1) resource orchestrator and 2) network service orchestrator. Firstly, the resource orchestrator receives the current information concerning potential physical and virtual resources of the NFVI via the VIM. Secondly, the network service orchestrator implements complete LCM of multiple network services. As a result, NFVO continuously updates the information regarding the available VNFs/CNFs running on top of the NFVI.

Hence, NFVO can initiate various network services. This is done by NSIDs, NSDs, and ultimately (P/V/C/H) NFDs.

2.3.2 Providing E2E Network Slices

Network slicing can be implemented in various network domains [32]. In fact, the physical infrastructure resources can be sliced at distinct network domains, as illustrated in Figure 2.6, including UE, RAN, edge/TN, and CN slicing. Slicing at the UE level allows device-to-device communication. Besides, caching at the UE level facilitates instantaneous content access. Accordingly, effective slicing policies have to be performed at the UE level.

Slicing in the RAN can be performed according to various communication specifications such as reliability, latency, radio access coverage, etc. A RAN slice takes a fraction of the available number of radio resources and determines one type of the 5G-NR numerologies (see Section 2.6.2). RAN slicing considers both orthogonal-based frequency and non orthogonal-based multiple access schemes. RAN slicing mainly concentrates on providing: 1) RAN programmability by utilizing SDN technology in this domain in order to share and manage radio resources dynamically and efficiently among various tenants, 2) slice independence and isolation, and 3) CP and DP functional split requirements to guarantee optimal performance in the RAN domain.

Performing logical partitioning to create CN slices is also essential. Each of the CN slices should produce a complete CN functionality over the InP. Employing SDN technology in the 5G CN results in the separation of DP and CP. Hence, CP functionalities such as access authentication, session, and policy management duties are individually conducted of the DP functionalities such as packet encapsulating/decapsulating tasks and packet forwarding. Thus, such CP and DP separation benefits in the CN can grant 1) independent DP functionalities per network slice, 2) independent CP tasks per network slice, and 3) general CP operations for several network slices.

Apart from the UE level, RAN, and CN resources, performing efficient slicing of the edge/TN resources between multiple tenants is also needed [33]. Therefore, all of the mentioned procedures are required while fulfilling the design specifications of a network slice that traverses over the whole network domains to create an active E2E NSI.

2.4 Isolation in Network Slicing

As mentioned before, one of the critical features of network slicing that has attracted much attention in the research world and industry is implementing E2E network slices that are isolated from each other. Network slice isolation refers

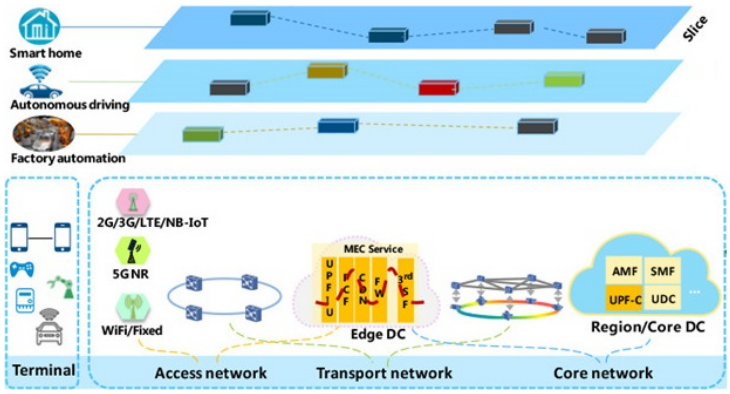


Figure 2.6: E2E network slice traverses over all network domains [34].

to discarding any negative influence of 1) other slices as well as InP from a slice (inter-slice isolation) and 2) building block entities within the same slice (intra-slice isolation); in order for the slice to continue with its regular operation [35]. In each network slicing use case, particular properties, and parameters can be specified in the SLA between the tenants and (M-)InP(s) to define proper isolation requirements. Due to the importance of isolation in network slicing, it is usually analyzed in four main perspectives as summarized in Figure 2.7:

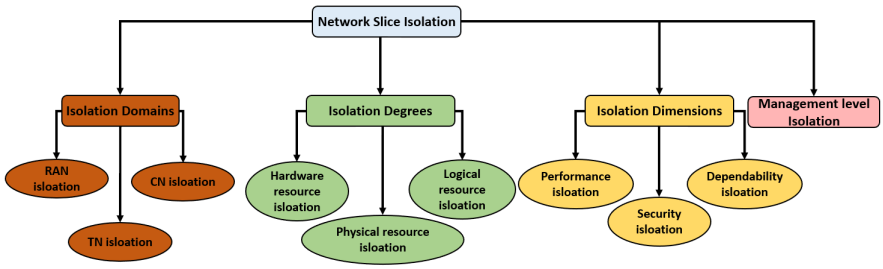


Figure 2.7: Network slice isolation concept areas.

- **Isolation Domains [36, 37]**: E2E slice isolation in different network domains has to be solved in the RAN, TN, and CN domains at the (M-)InP(s) level. Isolation domains in network slicing refer to:
  - **RAN isolation**: In the RAN domain, a specific number/amount of available resources are allocated to diverse network slices, and each network slice must not employ any resources allocated to other slices. Some of RAN resources are different Radio Access Technologys (RATs),

antennas, available frequency band(s), subcarriers, and Physical Resource Blocks (PRBs).
    - **TN isolation**: In the TN domain, traffic flow from a distinct network slice must not be transmitted to another slice. Some of the TN resources include optical fibers, cables, routers, etc.
    - **CN isolation**: In the CN domain, all the slices may utilize the same resources; nevertheless, appropriate strategies are needed to perform independent data packet processing in different slices.

- **Isolation Degrees [38]**: Isolation can be reached with various levels of strength, spanning from no isolation to complete isolation. In general, it holds that the more resources are shared among slices, the lower the isolation degree is obtained. Therefore, since complete isolation is granted only by allocating dedicated physical resources to each slice, it is essential to include different degrees of isolation. The resources of the RAN, TN, and CN domains can be isolated in several degrees at the (M-)InP(s) level. Isolation degrees in network slicing refer to:
    - **Hardware resource isolation**: In this case, perfect isolation is achieved by assigning different hardware entities to each network slice in each domain. Assigning dedicated antennas in the RAN, separate cables/-optical fibers in the TN, and individual servers in the CN domain are typical examples for this case.
    - **Physical resource isolation**: In this case, an intermediate level of isolation is achieved by splitting the available physical resources. Allocating a specific chunk of the available frequency band to different subcarriers in order to transmit data of each slice in the RAN, applying Wavelength Division Multiplexing (WDM) for optical fibers in the TN, and dividing the available physical compute and storage resources of a server between multiple slices in the CN domain are common examples for this case.
    - **Logical resource isolation**: In this last case, isolation is supported at the lowest level of the available resources. Dividing subcarriers into orthogonal-based frequency-time resources in the RAN, Multi-Protocol Label Switching (MPLS) of the data traffic in the TN, and utilizing VMs and containers on shared compute and storage resources of a server in the CN domain are considered standard examples for this case.

- **Isolation Dimensions [38]**: Finally, associations between the tenant demands and network slices have to be taken into account via isolation dimensions. Isolation dimensions in network slicing refer to:

- **Performance isolation**: As discussed, by pairing NSSIs from all network domains (NSSI-RAN, NSSI-TN, and NSSI-CN), an E2E NSI is created. In the case of providing performance isolation for various E2E network slices, it is required to individually evaluate each specific network domain in terms of throughput, latency, jitter, etc., to eliminate performance degradation from each of the involving NSSIs. Due to introducing extra overhead for processing the newly added information resulting from isolation, achieving the expected KPIs of an NSI is usually challenging. Besides, from the MANO perspective, it is essential to have mechanisms that appropriately plan and support the integrity of the resource allocation process.
- **Security isolation**: In the case of providing security isolation, a network slice must be protected from any adversary efforts that result in the malfunctioning of the network slice. It is worth mentioning that not only protecting a particular slice is crucial, but also, in the case of attacking a slice, it must not be spread to other slices. In addition to protection, data privacy in network slices is absolutely essential. In this case, no data of the involving (physical/logical) entities in providing network slices has to be accessed from any (internal/external) unauthorized entity. Hence, 1) providing secure interaction between the tenants and the (M-)InP(s) is vital, and 2) implementing suitable Authentication, Authorization, and Accounting (AAA) methods at the management level of each tenant are necessitated.
- **Dependability isolation**: In the case of providing dependability isolation, any physical/logical failure from the hardware/software levels must not be transferred to other operational network slices. Besides, appropriate redundancy planning for isolated resources needs to be taken into consideration. In addition, extra attention is expected for integrating pairing points of the involving NSSIs from the network domains' borders that may result in a point of failure. Furthermore, 1) performing qualified management tasks at the tenants' level to prevent interfering in the operations between tenants and 2) appropriate configuration at the management level of tenants and the (M-)InP(s) to avoid network slice failure are fundamental.

- **Management level isolation [38]**: Management level isolation aims to supervise and control each slice independently by considering it as an individual and solitary network. Isolation characteristics must be executed among slices at the virtualization level, creation phase, and orchestration phase. Consequently, to achieve each slice's QoS requirements, a number of policies and regulations, including the isolation essentials, need to be determined and appropriately enforced in order to maintain end-user ser-

vices' requirements.

2.5 Resource Allocation in Network Slicing

Due to the dynamicity and heterogeneity of traffic loads coming from the UEs in mobile networks, B5G systems need to bring novel approaches to enable slice-based allocation and management of resources across various network domains [39]. Thus, it is crucial to concentrate on this realm with more details.

2.5.1 Key Requirements for Resource Allocation in Network Slicing

- **Customizability**: Tenants need to customize the allocation of resources and performance of their network slices to satisfy their end users' demands. Hence, due to the typical temporal traffic load fluctuations over diverse network nodes, well-defined interfaces are expected for the tenants to adjust their slices' resource allocation dynamically to fit their end users' requests.
- **Complexity**: Resource allocation strategies' overheads necessitate to be held as low as possible. Such signaling overheads are related to the on-demand setup of network slices and the computational tasks that are required to perform such (re)configuration procedures of slices. It should be noted that there is a trade-off between the level of complexity and customizability of resource allocation in network slicing.
- **Efficiency**: The (M-)InP(s) will desire to achieve a high level of utilization of all types of its resources. This results in reducing CAPEX and OPEX by providing flexible infrastructure resource-sharing policies among multiple slices.
- **Isolation and Privacy**: As explained before, each tenant wants to have a specific level of isolation in its slice(s) in order to ensure that its SLA with the (M-)InP(s) will not be compromised by other slices of other tenants. Besides, due to the resource-sharing nature in the (M-)InP(s) among network slices, it is vital to decrease the leakage probability of sensitive information between tenants in order to protect their privacy. It is worth noting that there is a tradeoff between efficiency and isolation in network slices since resource efficiency is enhanced by relaxing isolation.
- **Cost predictability**: Tenants often tend to have long-term SLA with the (M-)InP(s) for resource allocation that usually results in predictable costs.

2.5.2 Resource Allocation Approaches in Network Slicing

Resource allocation in network slicing is split up into two main approaches, and for each of them, there are many proposed schemes for implementation. The two main approaches are:

- **Share-based [6, 40–43]**: This approach refers to distributing the (M-)InP(s) resources according to pre-agreed fixed shares of resources among multiple tenants. In other words, infrastructure resources in different network domains are shared and allocated to different slices provided for multiple tenants according to the proportion to their shares at each network domain and for each specific network node, such as base stations in the RAN, fiber optics in the TN, and servers in data centers in the CN.
- **Reservation-based [44–50]**: This approach refers to tenants' reservation requests to the (M-)InP(s) in order to grant resources from corresponding network nodes in the network domains to establish the demanded network slices. Such reservation requests may be accepted or declined by the (M-)InP(s) depending on resource availability or any other considerations.

It is worth stating that a significant feature accompanying what is mentioned above is the timing scale at which resource allocation approaches are set up and performed. It includes a long timing scale (months), a short timing scale (days), and a concise timing scale (hours, minutes, seconds) for allocating resources to tenants.

2.5.3 Comparison of Resource Allocation Approaches in Network Slicing

In the following, the share-based and the reservation-based approaches are compared with each other according to the resource allocation key requirements mentioned in subsection 2.5.1.

1. The Reservation-based approach grants a solid guarantee with stable resource allocation to the accepted tenants' requests and their corresponding established slices; however, this privilege is gained with negative impacts, including increasing complexity and signaling overheads, and reducing infrastructure resource efficiency.
2. The majority of the complexity load in the reservation-based approach runs on the (M-)InP(s) side. This is in contrast with the share-based approach that pushes part of the complexity load in performance management to the tenants' sides resulting in relatively simple algorithms on the (M-)InP(s) side.
3. The share-based approach affords some levels of isolation and privacy to

keep the performance at least at the range of static slicing; instead, the reservation-based approach can grant isolation, privacy, and protection by on-demand designing in slicing.

4. In terms of cost predictability, the share-based approach is highly predictable since the resource allocation, in this case, only relies on the proportion of shares; nevertheless, cost predictability in the reservation-based approach depends on the pricing rules applied by the (M-)InP(s).

2.5.4 Resource Allocation Realization in Network Slicing

It is necessary to emphasize that, apart from the approaches mentioned in subsection 2.5.2, extra considerations are needed to be combined with these approaches in order to realize resource allocations for multiple tenants that own various slices with diverse requirements. Such extra considerations include:

- **Forecasting methods**: that prognosticate future requests based on the former traffic loads in order to reallocate infrastructure resources at concise timing scales.
- **Admission control policies for establishing network slices**: that refer to admission or rejection of requests for creating different network slices via the reservation-based approach.
- **Admission control policies for end users of a tenant**: that guarantee service delivery to the admitted end users according to their QoS requirements.
- **End user mobility**: that points to checking the current distribution of the end users of a tenant and also mobility of them.
- **VNFs' placement**: that relates to the location environment for launching the involved VNFs of a slice according to the slice needs and also the availability of resources within each network domain.
- **Computational resource allocation**: that determines different parameters and timelines that are required for setting up, configuring, and (re)allocating other types of infrastructure resources.
- **Radio resource allocation and scheduling**: that define precise schemes to map high-level resource allocation plans into scheduled radio frames. These scheduling schemes need to fulfill the general resource allocation strategy and also particular QoS specifications, such as slices with extremely high reliability and extremely low delay requirements, i.e., URLLC slices.

2.6 5G New Radio

3GPP has developed the 5G new wireless access technology, so-called 5G-New Radio (5G-NR) [51, 52]. The essential technology characteristics comprise reliable and low latency transmissions, advanced antenna technologies, frequency spectrum flexibility to operate in high-frequency bands, interworking in high and low-frequency bands, and dynamic Time-Division Duplex (TDD). The radio interface of the 5G-NR comprises the physical layer [53, 54], and higher layers in the radio protocol stack consist of Media Access Control (MAC) and Radio Resource Control (RRC) layers [55]. 5G-NR is a flexible air interface able to grant a firm base for the future progression of wireless services. Some of the main features or topics related to 5G-NR are briefly discussed in the following.

2.6.1 5G-NR terminology

Some terms are introduced in the following. Some of them are previously used in 4G systems with the same meaning, so they are reused in the 5G-NR. Some others have been updated in their concept, and some new terms have been defined for the 5G-NR.

- **Resource Element (RE)**: That is the smallest unit in the 5G-NR resource grid consisting of 1 subcarrier in the frequency domain and 1 Orthogonal Frequency Division Multiplexing (OFDM) symbol in the time domain.
- **Radio frame**: 5G-NR, similar to the 4G systems, holds 10 subframes, each lasting for 1 ms.
- **Transmission Time Interval (TTI)/(eMBB) time slot**: Corresponds to 1 subframe duration (1 ms) that is required to encapsulate non-delay-sensitive data (transport blocks) from higher radio protocol stack layers and deliver to the physical layer in order to transmit it via the radio interface.
- **Short TTI/(URLLC) mini-slot**: Every TTI is divided into several short TTIs to immediately transmit URLLC traffic.
- **Cyclic Prefix (CyP)**: It is a parameter that is required to remove Inter-Symbol Interference (ISI) due to multipath transmitted signals. 5G-NR supports both normal CyP and extended CyP.
- **Numerology**: It determines a specific pattern for a collection of parameters, including Subcarrier Spacing (SCS) labeled as Δf, CyP, and OFDM symbol length.
- **Physical Resource Block (PRB)**: That is the smallest number of radio resources, in terms of REs, that can be assigned to a UE. A PRB is determined as 12 consecutive subcarriers in the frequency domain and 14 OFDM symbols in the time domain. By considering time slots/mini-slots, different Δf

duration in various numerologies, the size of PRB changes.

2.6.2 Waveform and Scalable Numerologies in 5G-NR

3GPP proposed to deploy the OFDM with CyP in Downlink (DL) and Uplink (UL) transmissions. This decision results in keeping the implementation complexity and cost low for broadband operations and Multi-Input Multi-Output (MIMO) technologies. 5G-NR supports transmission in Frequency Range (FR):

- **FR1 (410 MHz–7.125 GHz)**: known as sub-6 GHz;
- **FR2 (4.25 GHz–52.6 GHz)**: known as millimeter-waves.

Besides, scalable numerologies are also essential to support 5G-NR operations in the FR1 and FR2 for diverse use cases. 5G-NR employs flexible SCS with duration of $\Delta f = 2^{\mu} \times 15$ KHz ($\mu = 0, 1, 2, 3, 4$), which is scaled up from the fundamental 15 KHz SCS for $\mu = 0$ in the 4G systems. Consequently, the CyP in 5G-NR is scaled down by the factor of $2^{-\mu}$ from the CyP in 4G systems which has a length of 4.7 ms. At lower frequency bands, cells can be large, and the numerologies $\mu = 0, 1$ perform well. However, at higher frequency bands, cells are usually small, and the CyP lengths provided by the numerologies $\mu = 2, 3$ will suffice. Table 2.1 presents the parameters of each numerology setup in the 5G-NR.

The equal frame duration in the 5G-NR and the 4G systems facilitates their radio coexistence. Besides, transmissions of delay-sensitive data from the URLLC slices within mini-slots fulfill the extra low latency requirement for critical data communications.

Table 2.1: 5G-NR flexible numerologies [56].

μ	$\Delta f = 2^{\mu} \times 15$ (KHz)	CyP type	#symbols per slot	#slots per frame	#slots per subframe	FR1	FR2	BW of a PRB (KHz)	slot length (ms)	CyP length (μs)	symbol length (μs)
0	15	N *	14	10	1	✓	✗	180	1	4.69	66.67
1	30	N	14	20	2	✓	✗	360	0.5	2.34	33.33
2	60	N,E **	14	40	4	✓	✓	720	0.25	1.17	16.67
3	120	N	14	80	8	✗	✓	1440	0.125	0.57	8.33
4	240	N	14	160	16	✗	✓	2880	0.0625	0.29	4.17

*N: Normal, **E: Extended.

2.6.3 Multiple numerologies for diverse service requirements of 5G use cases

Employing multiple numerologies in the 5G-NR enhances the flexibility of scheduling use cases with diverse service requirements via performing slicing in the RAN. By operating within mini slots, both the DL and UL URLLC transmissions can be scheduled significantly faster than traditional scheduling within the slots

in the 4G systems. Besides, eMBB and mMTC DL and UL transmissions, which are not delay-sensitive compared to the URLLC use cases, can still be transmitted over the slots. Selection criteria of the 5G-NR numerologies for a specific use case depend on several factors [57], such as 1) the type of deployment environment for the 5G-NR operation (rural locations, dense urban, etc.), 2) use case requirements (extremely high reliability and extremely low latency for the URLLC transmissions, very high throughput for the eMBB transmissions, etc.), 3) end user mobility, 4) implementation complexity, or other conditions.

As indicated in Table 2.1, not all of the numerology options are applicable for both the *FR1* and *FR2*. The lower numerology types ($\mu = 0, 1, 2$) support operations in the *FR1* while the higher numerology types ($\mu = 2, 3, 4$) are employed for the operations in the *FR2*. The higher numerologies enhance delay-sensitive transmissions for URLLC use cases. Higher numerologies with wider SCS are suitable for deployments of small cell sizes at higher frequency bands that result in URLLC transmissions within mini slots. The lower numerologies with narrower SCS improve transmissions for the mMTC use cases. For the eMBB use cases, the middle range of numerologies ($\mu = 1, 2$) is the right choice. Figure 2.8 outlines the association between the cell size, frequency, and latency for the numerology selection.

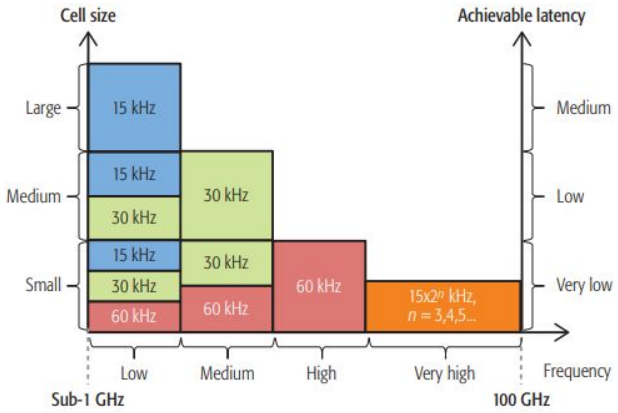


Figure 2.8: 5G-NR numerology selection criteria according to frequency ranges, latency requirements, and deployment types [58].

2.6.4 Orthogonal Multiple Access vs. Non-Orthogonal Multiple Access RAN Slicing in 5G-NR

Past generations of cellular networks have experienced significant changes in terms of employing multiple access techniques. In particular, Frequency Division Multiple Access (FDMA) in 1G, Time Division Multiple Access (TDMA) in 2G, Code Division Multiple Access (CDMA) in 3G, and Orthogonal FDMA (OFDMA) in 4G have been employed as the corresponding key multiple access technologies [59, 60]. Considering the design principles for each of the mentioned techniques, they all belong to the Orthogonal Multiple Access (OMA) [61] scheme in which radio resources are orthogonally allocated to the end users in the time, frequency, code domain, or their combinations. This scheme yields in separately carrying information between end users and base stations with relatively low complexity and cost-efficient receivers. Nevertheless, the number of supported end users is restricted by the number of available orthogonal radio resources in the OMA scheme. Furthermore, wireless channel impairments constantly sabotage their orthogonality, resulting in the necessity of employing highly complex restoring measures for the transmitted information. Hence, it leads to a challenge for the OMA to fulfill the spectral efficiency (for eMBB use case), low latency (for URLLC use case), and massive connectivity (for mMTC use case) requirements in 5G. Thus, advanced solutions have to be considered.

The innovative idea of utilizing the Non-Orthogonal Multiple Access (NOMA) approach has been introduced to support more end users with diverse service requirements than the total number of available orthogonal radio resources in RAN slicing [62, 63]. In the NOMA scheme, the same resources are non-orthogonally allocated to the end users with the cost of receiver complexity increment, which is inevitable for non-orthogonal signals' separation. Figure 2.9 illustrates different multiple access schemes.

As discussed before, due to the latency requirement of the URLLC end users, their communications have to be restricted in time, referring to transmission within mini slots. In contrast, a massive number of passive and active mMTC end users makes it impractical to allocate resources to each of them. Thus, it is essential to provide resources to the active subset of the mMTC devices shared via a random access process. Furthermore, since there is no delay requirement for the mMTC transmissions, the active subset of the mMTC devices can span over multiple time slots similar to the eMBB end users. Figure 2.10 illustrates RAN slicing via OMA and NOMA in the 5G-NR. Each subfigure represents eMBB and mMTC transmissions over a slot and URLLC transmissions over several mini slots.

The NOMA technique is divided into two main categories:

- **Code-domain NOMA [66, 67]**: It is motivated by the classic CDMA scheme,

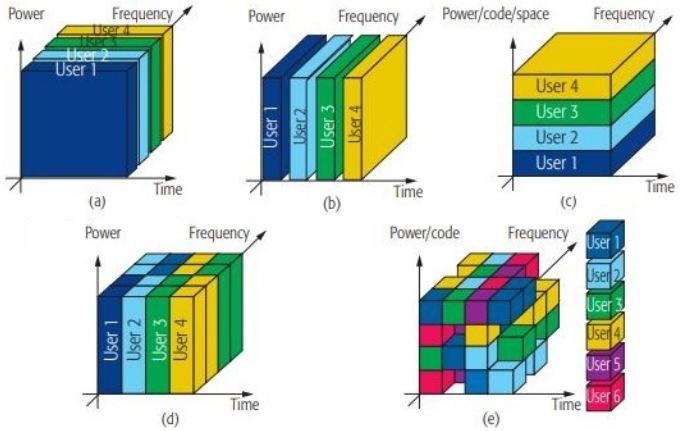


Figure 2.9: Different multiple access schemes, OMA: (a) FDMA, (b) TDMA, (c) CDMA, (d) OFDMA, and NOMA: (e) Code-/Power-domain [64].

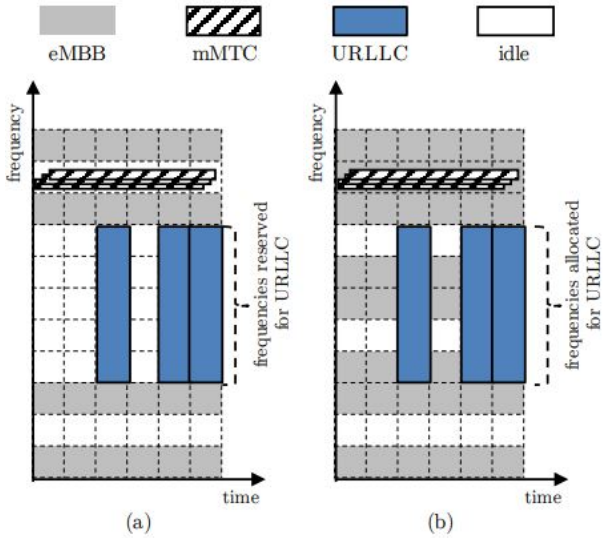


Figure 2.10: RAN slicing for three main use cases via (a). OMA and (b). NOMA in the 5G-NR [65].

in which multiple end users receive the identical time-frequency resources but apply individual user specific spreading sequences. Though, the major difference compared to the CDMA is that the spreading sequences are limited to sparse sequences or non-orthogonal low cross-correlation se-

quences in the NOMA technique.

- **Power-domain NOMA [68, 69]**: It is realized by allocating different power levels to the (eMBB, URLLC, mMTC) signals at the transmitter side (base station/end users) and then superimpose the signals on each other after channel coding and modulation. Power-domain NOMA itself is split up into:
    - **Power-domain NOMA with Superposition [65]**: That refers to allocating *non-zero* power levels to different signals. Such non-zero power signals are superpositioned on the transmitter side, scheduled, multiplexed, and transmitted over the same radio resources, and finally, they are recovered via techniques such as Successive Interference Cancellation (SIC) in the receiver side. The superposition technique grants spectral efficiency enhancement but with the price of receiver complexity increment compared to the OMA scheme to distinguish and recover eMBB, URLLC, and mMTC signals from each other.
    - **Power-domain NOMA with Puncturing [70]**: Due to the complexity of the SIC receivers to recover different signal types and the requirement of URLLC signals to be transmitted, received, and recovered with extremely low delay, the puncturing technique has been proposed as the alternative to the superposition. In the puncturing technique, which is mainly used for the coexistence of the eMBB and URLLC transmissions, the transmitter allocates *zero* power to the eMBB signals whenever the URLLC signals utilize the same radio resources and overlap with the eMBB signals. The puncturing technique eliminates the complexity of SIC but with the price of discarding the eMBB transmissions caused by the overlapped URLLC signal, which may result in a remarkable decrease in the data rate of the eMBB end users.

Chapter 3

Related Works

This chapter presents the most relevant state-of-the-art within the scope of the thesis and for the presented background in Chapter 2.

3.1 State-of-the-art Network Slicing Proof-of-Concepts

There is a significant number of Proof-of-Concepts (PoCs) for network slicing in the literature. Such testbeds grant the possibility of realizing complex network architectures in the network slicing field to assess and improve network performance. Additionally, while research testbeds keep the expense of network deployment moderate, their functionalities, with a reasonable approximation, are comparable to real networks. With the help of NFV technology, such testbeds can usually be realized on standard PCs or servers with a not very high amount of resources and without purchasing specialized hardware and software. Further, the availability of open-source software packages provides opportunities for designing innovative solutions towards 5G and B5G [71].

Among such platforms, small-scale network slicing testbeds are essential for the research community in several aspects. Small-scale testbeds demand less deployment budget compared to large-scale testbeds. Besides, small-scale testbeds, with a compact softwarized version of the demanded entities, are easier to deploy and launch than large-scale ones. Moreover, due to such testbeds' small scaling, they are more flexible to troubleshoot, and resolving potential issues is faster than large-scale testbeds with various associated entities. Ultimately, although the number of practical use cases that can be studied on small-scale testbeds is lower than on large-scale testbeds and real networks, small-scale testbeds can yield similar results to large-scale solutions. This section focuses on those network slicing realization solutions implemented on *small-scale testbeds* consisting of *open-source* software packages. The aspects mentioned earlier are the main criteria for selecting the following testbeds, which are categorized

according to their ability to provide CN slicing, RAN slicing, and E2E slicing.

3.1.1 Small-scale CN slicing testbeds

Authors in [72, 73] deploy containerized implementation of both Evolved Packet Core (EPC) and Evolved Node B (eNB) on a cloud infrastructure. In the testbed, specific cloud-based templates manage applications in a stack of containers, and various services via network slices can be created for eHealth and IoT. [74] operates RAN on licensed Long Term Evolution (LTE) and 5G bands, and the CN comprises EPC and IP Multimedia System (IMS). The testbed focuses on providing low latency services for the URLLC use case, and only CN slicing has been implemented. There are two CN slices for eMBB and URLLC use cases, and they are evaluated according to throughput and delay requirements.

[75] presents Slice-as-a-Service (SlaaS) regarding Data Centers (DCs). In this case, a slice is formed by a combination of DC slices (compute and storage resources) connected by network slices (networking resources) running on their own VIMs and Network Infrastructure Managers (NIMs), respectively. Slices are formed via transformable resources, represented as physically isolated or virtually shared resources. The testbed allows an evaluation case to discover the required time (loading, booting, configuration, and service startup times) to establish different infrastructures. The solution in [76] is a practical implementation of an efficient resource usage model for dynamic and real-time slice (de)allocation in the CN domain. This solution also brings isolation between the involving VNFs of a slice and considers allocation policies for slice requests.

The framework in [77] examines Service Assurance (SA) to satisfy Quality of Experience (QoE) and QoS requirements in the context of network slicing. The framework integrates a novel SA-based architecture with the ETSI MANO platform to carry out monitoring, analytics, management, and reporting tasks in order to guarantee the performance of adaptive video streaming services. The testbed in [78, 79] exhibits the deployment of EPC as a VNF on a cloud environment, and it shows the EPC service instantiation process via a service orchestrator. In this testbed, according to the defined descriptors at the VNF and network service levels, the internal components of the EPC are separately implemented on several VMs, and then they are configured via specific configuration files. This work aims to produce MEC-based services and integrate the EPC with a C-RAN architecture, including functional split capability.

The work in [80, 81] is a solid design for onboarding and scheduling phases in VNF LCM, and it offers a programmable and flexible MEC-enabled framework for IoT traffic. The work enhances the joint orchestration capability for VNFs in the MEC and cloud resources. This platform is deployed on several physical servers for the functionalities of the CN and comparatively lower computa-

tional resources for the MEC. Some autoscaling and VNF placement scenarios are presented to evaluate the testbed functionalities. The study in [82] presents an innovative strategy for orchestrating CN operations in the form of CNFs, optimizing proximity to the edge network. Simultaneously, computational tasks needed for on-demand processing are offloaded to a centralized cloud infrastructure. This approach enables service isolation and network segmentation irrespective of geographical regions, ensuring high scalability. The primary aim is to automate CNFs provisioning, facilitating application deployment across multiple clusters. The authors evaluate latency performance while retaining edge-based network control.

3.1.2 Small-scale RAN slicing testbeds

The platform in [83, 84] provides slicing of RAN resources and also applies isolation between them via employing SDN to grant an independent CP for the RAN domain. Such isolation capability has been examined for multi-service provisioning. The platform in [85] implements RAN slicing via RESTful API automatically. The testbed applies the slice-aware policy in Radio Resource Management (RRM) functionalities for admission control and scheduling processes. Some practical scenarios are arranged to implement RRM functionalities for the scalability of the network, slice admission control, and slice isolation. The work in [86] demonstrates slicing the RAN resources between eMBB and IoT services. Real-time SDN functionality creates isolated RAN slices for IoT and eMBB devices according to their service requirements. The SDN controller performs the scheduling process and admission control decisions. The testbed has been assessed for average DL throughput in IoT and eMBB slices.

This testbed [87] proposes a cost-efficient containerized and orchestrated 5G mobile network equipped with distinct fronthaul and backhaul topologies. The testbed essentially focuses on integrating ML into the management duties in the 5G RAN and TN domains to optimize network performance. The testbed facilitates the deployment of various network topologies on the fronthaul and backhaul by creating an emulated TN. The platform is assessed by monitoring the number of allocated PRBs to different slice requests and VNFs' placement in a cluster of containers. [88] introduces a cloud-native framework that generates VNFs and employs Kubernetes for the distribution and management of these VNFs across worker nodes. Besides, the framework simulates IP transport network effects in both fronthaul and backhaul sections of actual cellular networks to underscore the significance of EC and the potential of RAN slicing.

3.1.3 Small-scale E2E slicing testbeds

Reference [89] concentrates on providing two VNFs for intelligent monitoring and actuating applications. The testbed considers dynamic sharing of the available bandwidth between the two VNFs for establishing an E2E network slice. The work in [90] presents a comprehensive platform, and apart from service orchestration, it integrates SDN functionality in the RAN domain as well. The testbed investigates providing network slices for eHealth and smart cities. [91] provides slice isolation via deploying two containerized CNs. The testbed has been evaluated for connection establishment for both normal LTE UEs and customized UEs with slice selection capability. The solution presented in [92, 93] affords efficient resource utilization for building independent and customizable E2E slices. Created slices for different tenants are multiplexed over shared resources. The UEs can possibly connect to more than one slice simultaneously. The testbed has been assessed in terms of independency between slices and their throughputs.

Testbed in [94, 95] concentrates on deploying two CN slices connecting to the C-RAN architecture via the TN to slice and manage the TN domain. Different phases of a slice lifecycle, from provisioning, allocating a slice to a UE, and managing the slice, are addressed in this solution. TN slicing is accomplished by driving a slicing policy via SDN to establish a connection flow between CN and RAN domains. The platform in [96] proposes QoS-aware programmable policies in DP to create network slices for URLLC use cases with various reliability requirements. The work is evaluated for providing services such as eHealth and real-time video streaming communication via a machine learning-enabled approach for the patient's examination.

The study in [97] investigates E2E network slicing LCM by integrating a service orchestrator with a network slice manager entity. This integration grants a platform for monitoring, collecting, and evaluating the KPI reports belonging to the chained VNFs that create an E2E network slice. In this way, multiple slices are inquired to discover whether the SLA is fulfilled or not. The testbed operates a scenario by forming a private mobile network that provides services with best-effort and broadband QoS types via E2E network slices. The paper [98] offers an Intent-Based Networking (IBN) framework equipped with deep learning that efficiently automates the configuration, service provisioning, service update, and service assurance. In addition, it performs MANO tasks for the lifecycle of multi-domain slice resources and handles run-time resource scalability.

The work in [99] introduces a MANO framework designed for E2E network slicing automation. The framework integrates novel bandwidth management methods and leverages open-source methodologies alongside cutting-edge cloud-native technologies. The study assesses the resource overhead of this framework and evaluates service throughput under different bandwidth policies. The article [100] analyzes a cloud-native wireless architecture utilizing container-based

virtualization to facilitate flexible service deployment. The authors optimize resource allocation within network slicing by employing deep reinforcement learning. Two model-free algorithms are introduced for real-time network state monitoring and adaptive allocation policy training. Their results lead to enhancements in network efficiency.

Paper [101] offers essential platform provisioning and service LCM capabilities for a specific user-generated content multimedia scenario. This is done by utilizing cloud-native models, Platform-as-a-Service (PaaS), and virtual testbed instances. The research showcases the utilization of service-level telemetry from a cloud-native user-generated content application to adjust system resources across the NFVI dynamically. The authors in [102] present a practical framework employing over-the-air transmissions, addressing two key aspects to enhance LCM in 5G and B5G networks: cloud-native deployment of 5G core NFs and comprehensive end-to-end monitoring. The initial step involves deploying a monitoring framework as CNFs across a multi-tier network with a MEC host. Subsequently, the authors demonstrate an end-to-end monitoring system encompassing infrastructure resources and radio metrics.

The majority of the mentioned works in the Sections 3.1.1, 3.1.2, and 3.1.3, focus only on addressing and solving a particular problem in network slicing, and consequently, their proposed platforms are not exhaustive, being able to deal with more open challenges in the network slicing field. In other words, just a few platforms already have included ([98]) or have the potential ([90, 96]) to simultaneously provide:

- slicing in various networking domains to create E2E slices;
- LCM of a slice via NSIDs and (P/V/C/H) NFDs;
- MANO tasks for M-InPs;
- SDN controller integration in their architecture;
- multi-tenancy and multi-RATs communication capabilities;
- run-time monitoring of resources;
- Artificial Intelligence (AI)-enabled functionalities.

The lack of extensive study in the field of small-scale network slicing testbeds is the motivation behind the work in Paper 1 [103]. Paper 2 [104] offers a comprehensive platform that addresses the above-mentioned missing items. The outcome of Paper 2 is the foundation for further practical approaches in network slicing during this Ph.D.

3.2 State-of-the-art solutions for Network Slicing Isolation

The isolation concept is one of the fundamental features of network slicing. Nevertheless, due to its complexity and challenges, relatively few works have focused on it and analyzed the technical details of network slice isolation. We consider the most relevant works in the network slicing isolation field that refer to technical methods and algorithms for network slice isolation. Works that focus on standardization, abstract network architectures, and surveys in the network slice isolation field are not included in this analysis. We review the relevant works from simulation-based and implementation-based perspectives.

3.2.1 Simulation-based solutions

The work in [105] proposes a novel mutual authentication and key establishment protocol utilizing proxy re-encryption. The protocol grants specific authentication between components of a network slice to enable secure connection and protected key establishment among component pairs for slice security isolation. Paper [106] offers a secure keying scheme, by adopting a multi-party computation strategy, which is appropriate for network slicing architecture in the case that the slices are accessed by third-party applications. This mechanism ensures the satisfaction of use cases or devices in which the data is collected. Authors in [107] introduce the network slice trust degree concept and develop a trust value calculation model that includes three parts. As a result, the MANO can efficiently calculate network slice trust value via dedicated weighting parameters to the three parts: 1) cloud model algorithm, 2) user evaluation, and 3) reward/punishment while the slice is in the run-time phase. Such weighting parameters differ according to the diverse security requirements of the network slices.

The work in [108] grants a secure service-oriented authentication structure for MIoT fog computing. As a result, end users can connect to the CN and anonymously obtain IoT services via proper and different network slices chosen by fog nodes. A privacy-preserving slice selection method is offered to protect slice types and access services. Besides, session keys are assigned among end users, local fogs, and IoT servers to ensure secure access in the fog cache and remote servers with low delay. Reference [109] suggests an SDN-enabled cross-authentication model that connects cryptographic and non-cryptographic schemes to solve latency and security challenges. In particular, the strength vector of the received signal at UEs is employed as a fingerprinting reference to create an unpredictable secret key. Afterward, according to the key agreement protocol, a cryptographic scheme is offered by applying the created secret key in order to

enhance the confidentiality and integrity of the authentication for the handover process between multiple base stations. The authors also present a radio trusted zone database to reduce several authentications of UEs within the network.

The work [110] formulates an integer linear programming design for providing secured CN slices. The work considers nodes and links to model CN slices. The authors introduce a heuristic CN slicing method, so-called VIKOR, which 1) classifies the quality of a node via the corresponding resources of that node and topology attributes and 2) performs the k shortest path algorithm [111] to determine the appropriate physical path for the link of a slice. The intention is to select the minimum value of the product of the maximum bandwidth utilization of a link and its hop counts. This approach results in improving the slice acceptance rate.

The study [112] offers a two-layer and less complicated scheduler for efficient and dynamic RAN slicing. The work presents that there can be different trade-offs between performance, isolation, and priority in dynamic radio resource allocation between RAN slices. The work suggests a strategy based on specific UEs' QoS requirements and the number of demanded PRBs for each UE to establish a RAN slice. Nevertheless, the scheme's complexity grows linearly with the number of created slices pointing to inter-slice scheduling and with the number of UEs within a particular slice referring to intra-slice scheduling.

The study [113] contributes to dynamic virtualized RAN slicing policies with mixed traffic. According to the latency and throughput requirements, the authors present a resource allocation method, a heuristic scheme for resource customization of a UE, and a deep reinforcement learning slicing strategy to enhance resource utilization and QoS fulfillment. The proposed algorithm also increases the slices' performances in mixed traffic compared to some of the state-of-the-art benchmarks.

3.2.2 Implementation-based solutions

The study [114] suggests and evaluates a security-aware slice embedding implementation that allows tenants to report security-oriented requirements while restricting InP information revelation. The implementation is performed via the Ubuntu 16.04 operating system (on a machine with i7-6700HQ @ 2.60GHz, 16 GB RAM, and 300 GB of storage), which includes domains operating in separate processes. The work provides a multi-level method to create all the necessary multi-domain slice embeddings for the InP and M-InPs cases. The work [115] presents a solution to proactively reduce Distributed Denial-of-Service (DDos) attacks in the CN domain by applying slice isolation. The authors propose implementing a mathematical model that can afford on-demand slice isolation and ensure delay for CN slices. They implement their idea over a testbed by employ-

ing six physical servers (Intel(R) Xeon(R) CPU E5420 @ 2.50GHz (4 cores), RAM 8GB, and network bandwidth 100Mb/s). Three of these servers are utilized to allocate CN slices, two to act as DDoS nodes and one to be a client. The results confirm a reduction in DDoS attacks and also an increment in the availability of CN slices.

As evident from Sections 3.2.1 and 3.2.2, most of the mentioned papers analyze isolation from the simulation-based perspective, and only a few contributions offer implementation-based approaches. Besides, from the implementation-based perspective, no contribution particularly investigates slice isolation via slice instantiation process by MANO utilizing NSIDs and involving (P/V/C/H) NFDs. Hence, Paper 3 [116] and Paper 4 [117] focus on implementing security and performance isolation among CN slices via the slice instantiation process performed by MANO tasks.

3.3 State-of-the-art solutions for eMBB and URLLC coexistence in the 5G-NR

We review the related works corresponding to the eMBB and URLLC coexistence in the 5G-NR.

The study in [118] focuses on joint support of visual (via eMBB slice) and haptic (via URLLC slice) perceptions across cellular networks. This joint support takes place via OMA and NOMA for sharing the DL resources. The results of the work confirm that the NOMA slicing is the superior approach compared to OMA in order to support an excellent perceptual resolution and also a high rate for haptic perceptions. Paper [119] provides a joint user association and resource allocation in the DL of a fog network. This is done by 1) employing an analytic hierarchy process to manage the QoS prioritization of diversified IoT applications and, consequently, 2) representing a two-sided matching game to establish a solid association between the IoT devices and the fog network. Their proposal results in resource allocation efficiency in the DL transmissions of eMBB and URLLC traffic. Paper [120] grants a risk-sensitive policy according to the conditional value at risk approach for eMBB reliability and a chance constraint for URLLC reliability.

The authors in [121] study the coexistence problem of eMBB and URLLC users in 5G networks. They formulate a joint resource allocation problem that can satisfy both the eMBB user rate and URLLC interrupt probability requirements. They assign mini-slots for URLLC users and calculate the transmission power of URLLC users, ensuring the reliability constraint. A similar study is performed in [122], which also studies the resource slicing problem for 5G eMBB and URLLC services. The resource slicing problem is formulated as an optimiza-

tion problem aiming to maximize the eMBB data rate. The problem is subject to a URLLC reliability constraint while considering the variance of the eMBB data rate to reduce the impact of immediately scheduled URLLC traffic on the eMBB reliability. An optimization-aided deep reinforcement learning-based framework is proposed to solve the formulated problem.

The authors in [123] evaluate the coexistence technique for eMBB and URLLC based upon a punctured scheme. They extend the study to formulate an optimization problem aiming to maximize the minimum expected achievable rate of eMBB UEs while fulfilling the provisions of the URLLC traffic. In study [124], the radio resources are scheduled among the eMBB UEs on a time slot basis, whereas they are handled for URLLC UEs on a mini-slot basis. They use a penalty successive upper bound minimization-based algorithm for eMBB UEs, while the optimal transportation model is adopted to solve the same URLLC UEs problem. They also present a heuristic algorithm for efficiently scheduling PRBs among eMBB UEs.

Authors of [70] model the impact of the URLLC transmission over the scheduled eMBB traffic via loss functions caused by the URLLC traffic. The work in [125] analyzes the multiplexing of the eMBB and URLLC traffic in the C-RAN environment. The work in [126] investigates the performance trade-offs between eMBB and URLLC traffic types in a multi-cell C-RAN architecture under NOMA and OMA access strategies. The work outcome reveals the advantage of employing the orthogonal-based solution for degrading the mutual interference of the eMBB and URLLC traffic. The authors also demonstrate the potential benefits of puncturing in improving the efficiency of fronthaul usage by discarding received mini-slots affected by URLLC interference. The authors in [127] present a puncturing scheme for transmitting low latency communication traffic, multiplexed on a downlink shared channel with eMBB. They also propose recovery mechanisms for the impacted eMBB users to minimize the capacity loss for eMBB users due to low latency communication traffic.

A group of authors considers an optimal resource assignment under different channel conditions within a mixed numerology approach in [128, 129]. The work presented in [130] focuses on the scheduling problem for heterogeneous services within a mixed numerology approach to maximize the number of satisfied users while meeting latency demand and data transmission requirements. Mini-slots enable transmissions that can be performed in a shorter time than the regular slot duration. In higher numerologies, the use of wider SCSs provides shorter slot durations. Consequently, low-latency communications can be enabled by combining mixed numerology and mini-slot approaches. The work in [131] offers a model to optimize the numerology and resource allocation for mixed numerology systems, which employ the mini-slot approach.

The work in [132] aims to maximize the minimum expected achieved rate

of eMBB users and fairness between them by employing a one-to-one matching game to compute appropriate eMBB and URLLC pairs for URLLC resource allocation. The authors of [133] and [134] aim at maximizing the aggregated throughput of the eMBB and URLLC users while mitigating the Inter-Numerology Interference (INI). They consider satisfying the minimum acceptable throughput of the eMBB and maximum allowed delay of the URLLC users according to their corresponding service requirements. The authors propose a deep reinforcement learning INI-aware agent to overcome the computation complexity of the optimization problem. Their method offers a spectrum allocation fulfilling the eMBB and URLLC service requirements while reducing the INI. Finally, they analyze their results delivered by the INI-aware agent when the URLLC traffic statistic is modeled based on mobile and industrial networks.

Reference [135] formulates the RAN slicing problem between eMBB and URLLC users as a multi-timescale problem and proposes a hierarchical deep neural network algorithm to assign radio resources to their corresponding users. The authors model the selection of slice parameters within a time slot as a partially observable Markov decision process and present an algorithm to define configuration parameters for the eMBB and URLLC slices efficiently. In the work in [136], the authors compute the achievable latency for the industrial network scenario based on an accurate system-level simulation. Their primary focus is determining 5G NR configurations that are more relevant for Industry 4.0 applications to analyze the effect of reserving bandwidth for URLLC services. Reference [137] defines a context of the network based on combined statistical characteristics from the wireless channel and UEs' service requirements to train a Mondrian forest to predict an optimal mixed-numerology profile.

The authors of [138] work on solving the challenges of radio resource allocation in the mmWave band of 5G-NR by proposing a deep reinforcement learning-based scheduler. The scheduler allocates resources for a list of UEs to satisfy their different slice's SLA requirements according to the channel quality of each UE. Paper [139] presents a resource allocation strategy that combines latency, control channel, hybrid automatic repeat request, and radio channel quality in determining the transmission resources for different users. The approach minimizes the latency and bypasses unwarranted costly segmentation of URLLC payloads over several transmissions. Reference [140] addresses the problem of joint admission control and resource scheduling for URLLC by utilizing a standard continuous SNR model, where all allocated resource blocks contribute to the success probability, and a binary SNR model, where each resource block is classified as active or inactive according to a SNR threshold. In congestion cases, the work focuses on discovering a subset of users that can be scheduled at the same time.

The authors in [141] develop a joint optimization problem for power and bandwidth allocation with long-term conditions of queues backlog for the eMBB

users. They utilize the Lyapunov drift-plus-penalty technique to create the relationship between the long-term constraints and the short-term optimization problem. Furthermore, they employ a one-to-one matching procedure to solve the slicing puncture problem. The work in [142] designs a coordinated multi-point multi-numerology network to improve the throughput of eMBB and latency of URLLC users. The authors solve a subcarrier and power allocation problem with the objective of maximizing the system sum rate. They show that their designed network has a higher sum data rate, lower delay, and throughput outage compared to the traditional non-coordinated multi-point single numerology scenarios.

Reference [143] concentrates on minimizing the rate loss of the eMBB users and packet segmentation loss of URLLC users while fulfilling the QoS requirements of eMBB and URLLC use cases. They consider the case of one-to-one pairing in which one URLLC packet can be paired with only one eMBB. They employ a bi-level optimization problem that includes one inner and one outer problem. The inner problem seeks to discover the optimal power and frequency resources for each URLLC and eMBB pair, and the outer problem desires to search for the optimal eMBB-URLLC pairing policy. They also generalize the problem for many-to-many pairing while undervaluing the overhead due to URLLC packet segmentation.

The authors in [144] aim at minimizing the decoding error rate of URLLC users while ensuring the demand for the throughput of eMBB users. They propose a block coordinate descent optimization algorithm to obtain the optimal bandwidth allocation, puncture weight, and transmit power. Paper [145] focuses on studying eMBB and URLLC use cases in networks that are assisted by a Reconfigurable Intelligent Surface (RIS). The authors jointly optimize the power and frequency allocation problem and the RIS phase shift matrix to enhance the eMBB sum rate and URLLC reliability. The work in [146] concentrates on eMBB and URLLC use cases in a massive MIMO system by providing a unified information-theoretic framework incorporating an infinite-blocklength analysis of the eMBB spectral efficiency with a finite-blocklength analysis of the URLLC error probability. The work relies on the use of mismatched decoding and saddle-point approximation.

The main objective of most of the mentioned papers in this section is to maximize the sum rate of the eMBB end users without necessarily considering each individual eMBB end user. Nevertheless, one of the fundamental goals of 5G and B5G is to afford a more uniform user experience than 4G systems in terms of higher individual data rates for a larger number of eMBB end users [147]. This motivates the works in Paper 5 [148] and Paper 6 [149] to enhance the data rate for each eMBB end user while still aiming to maximize the sum rate of the eMBB end users. This is done by offering optimized schemes for reducing

the eMBB data rate loss and enhancing the data rate of each eMBB end user. Besides, the URLLC use case can be subcategorized into different groups or classes in which each class can hold diverse latency and/or reliability. In other words, each class can hold stricter/relaxer latency and/or reliability compared to other classes. This is a missing concept in all the mentioned papers in this section and thus motivates the work in Paper 6.

Chapter 4

Research Contributions

This chapter is structured into two major sections. The first section focuses on the research design, encompassing the formulation of Research Questions (RQs). These RQs and ROs are tailored to the thesis topic, addressing the gaps and challenges highlighted in Chapter 3. The chosen research methodology is also presented in this section. The second section highlights the thesis's main contributions, abstracting each paper's content, describing their interconnections, and demonstrating their relevance to the defined RQs.

4.1 Research Design

4.1.1 Motivation

Network slicing expedites multiple flexible logical networks on top of a shared physical infrastructure that allows technical and business innovations. Network slicing integrates physical infrastructure and logical networks into a programmable and open softwarized multi-service network environment. Each of these logical networks needs to be defined clearly in terms of demanded resources in each specific network domain. Then such allocated resources necessitate to be controlled and managed dynamically via an external entity that monitors the performance of the network. Besides, any negative operational impact on a slice, that may lead to a state in which the slice does not operate as intended, must not affect other slices' operations. Furthermore, efficient resource allocation to a particular slice or multiple slices is crucial in order to provide services that operate according to the SLA. Motivated by the mentioned aspects, several principles are listed that shape the corresponding technical operations in network slicing, which are the focus of this thesis. Hence, this work has been directed to:

- Network slice automation and programmability that enable on-demand (partial/E2E) network slice configurations via open Application Program-

ming Interfaces (APIs).
- Network slice isolation that assures performance and security guarantees for slices.
- Network slice customization that guarantees efficient resource allocation for coexisting of the eMBB and URLLC slices in order to fulfill their service requirements.

4.1.2 Research Questions and Research Objectives

This thesis focuses on three main research domains: Network Slice Automation, Network Slice Isolation, and Network Slice Customization. Thus, it is absolutely important to classify the main research goals while studying and investigating network slicing. Consequently, the following Research Questions (RQs) were posed and answered by the work performed towards the Research Objectives (ROs) in this thesis.

RQ1: *How to study and implement network slices on a physical infrastructure?*

RO1: Designing and launching solution is applied in the form of a (relatively small-scale) research testbed equipped with open-source packages to operate as VNFs, which construct different network slices. This is done by creating different descriptors for VNFs, network services, and network slices employed by a service orchestrator to provide lifecycle administration of network slices.

RQ2: *How to provide security and performance isolation between eMBB and URLLC network slices?*

RO2: Data transmission encryption protocols are established to provide isolation between network resources of the operational VNFs of network slices. This mechanism grants inter-slice traffic flow separation, which results in isolation in terms of security and performance. Security isolation eliminates the negative impact of an intentional attack in one slice upon other slices. Performance isolation ensures that the KPIs of service are fulfilled. This goal can be addressed by specific VPN-based solutions that can achieve not only traffic isolation (security isolation) but also high throughput for the eMBB slice and low latency for the URLLC slice (performance isolation). Furthermore, optimal resource allocation between diverse slice types can be accomplished by considering both QoS and security requirements for multiple slice types.

RQ3: *How to optimize the coexistence of eMBB and URLLC traffic on radio resources?*

RO3: The NOMA method is employed that simultaneously serves multiple eMBB and URLLC end users on the same frequency-time resources over short TTIs, also known as mini-slots. NOMA utilizes distinct techniques, such as SIC, which yields higher spectral efficiency and lower latency compared to the conventional OMA. However, the main challenges of SIC are complexity and sometimes a high number of errors in signal detection. Consequently, to overcome such difficulties, the puncturing technique is considered as the main alternative for SIC in NOMA. The puncturing technique interprets any eMBB traffic over a mini-slot overlapping URLLC traffic as eradicated or punctured traffic.

It is worth noting that the RQ1 and RO1 mainly consider implementing network slicing in general, and they do not focus on any specific 5G use case. The RQ2 and RO2 progress one step forward and concentrate on enforcing slice isolation specifically for the eMBB and URLLC in the CN domain. Finally, RQ3 and RO3 consider the coexistence of the eMBB and URLLC in the RAN domain. As it can be observed, on the one hand, RQ1 and RQ2 are linked via implementing network slices and also isolation between them, and on the other hand, RQ2 and RQ3 are connected as they both focus on eMBB and URLLC use cases, RQ2 in the CN domain and RQ3 on the RAN domain.

4.1.3 Research Methodology

Investigating network slicing principles for its implementation and modeling is a research and development process and involves multiple steps. The research methodology in this work comprises the following phases:

1. **Literature review:** The foundation of this study is laid through an extensive literature review, which concentrates on existing research, theories, and practices concerning the coexistence of eMBB and URLLC within the context of 5G and B5G networks. This comprehensive exploration provides the contextual framework necessary for devising novel solutions. This phase is presented in the Chapters 2 and 3. Moreover, the first paper included in this thesis, Paper 1, is the straight outcome of reviewing related works, which is a systematic review corresponding to the *RQ1*.
2. **Determining research questions:** Specific open RQs are meticulously formulated based on the literature's insights as presented in Section 4.1.2. These questions are designed to address the complexities of enabling eMBB and URLLC coexistence through network slicing implementation and isolation. Each question is carefully tailored to extract targeted knowledge, setting the path for the subsequent phases. Each paper in this work an-

swers partially or entirely a specific RQ.

3. **Theoretical solution designs and analysis:** The research methodology then is shifted towards theoretical exploration, where innovative solutions are conceptualized and their potential is analyzed. Theoretical models and frameworks are developed to outline the sophistication of implementing network slicing to enable eMBB and URLLC services while maintaining isolation requirements. This phase supplies the theoretical groundwork for subsequent practical endeavors.
4. **Available open-source packages/software tools for solution creation:** Available software tools and suitable frameworks for crafting network slicing solutions are assessed for their suitability and efficacy. Such tools are required to materialize the theoretical designs effectively and must be aligned with the eMBB and URLLC coexistence.
5. **Modeling and Simulation/Simulation evaluation:** Regarding the *RQ3* that incorporates modeling approaches, we describe the desired properties for the models. We apply methods from various mathematical disciplines, among them probability and algebra. After designing the eligible model, simulations in Matlab or Python are performed. These quantitative simulations enable the assessment of network behavior, resource utilization, and QoS when eMBB and URLLC services share the same infrastructure through network slicing. Paper 5 and Paper 6 are the outcomes of this phase.
6. **Practical approach and Implementation/Experimental evaluation:** In this case, Real-world experimentations are conducted to validate the theory in a controlled environment (testbed) in order to complete simulation findings. This phase concerns the deployment of actual network slicing scenarios, closely emulating the conditions experienced in 5G and B5G networks. Since *RQ1, RQ2* involve practical implementation, we define which specific platforms and software packages are required in order to fulfill the objectives of the related RQs. In this phase, we set up, configure, and program the network under the study. For instance, average knowledge of networking, virtualization techniques for launching the service orchestrator and infrastructure manager, and some programming skills are required in this phase. Paper 2, Paper 3, and Paper 4 are the result of this phase.
7. **Performing test scenarios and Measurements:** We operate several tests to capture and collect the required data regarding the practical approach by launching and executing specific scenarios. Regarding the modeling approach, by executing the simulation multiple times and including required randomness, we capture the required data to be evaluated in the next phase.

8. **Final results and alignment with theory:** First, the obtained results from both simulation and experimental evaluations are meticulously analyzed to determine their alignment with the theoretical predictions. Then, the results are checked to see if they respond to the RQs outlined in the second phase. Next, the results are analyzed in order to discard potential errors in the practical and modeling methods. Finally, the results are assessed to validate the network performance enhancement under the proposed algorithms/techniques(for the modeling method) and the proposed implementation solutions (for the practical method). Such crucial steps establish the efficacy of the proposed solutions in enabling eMBB and URLLC coexistence while ensuring network slicing isolation.
9. **Improvement of Simulation and Experimental evaluations:** The feedback from the initial simulation and experimental phases triggers iterative refinement. This involves further optimization of simulation parameters and enhancing experimental setups, ensuring accuracy and reliability in subsequent evaluations. Therefore, both Simulation and Experimental strategies are regularly improved to have final results that answer the RQs well.
10. **Conclusions:** This phase leads to conclusive insights that address the defined RQs. The findings collectively declare the conclusion of this study, reinforcing the significance of network slicing in facilitating the coexistence of eMBB and URLLC services within the dynamic landscape of B5G networks.

4.2 Contributions of the papers

This section presents the main contributions of the thesis with a relevant overview of each paper mapped to the RQs. This Ph.D. contributed to 7 publications in total. Table 4.1 presents a complete list of the 6 included publications in the thesis and how they are mapped to the research methodologies used in this thesis. The order in which the papers are presented is not necessarily chronological but rather related to the research goals so that it is more straightforward to follow the exposition in the thesis. Table 4.2 shows the contribution that is not included in the thesis.

Figure 4.1 illustrates the interconnections between the papers and how they are linked to the research questions. The contribution of each paper is stated in the following.

Table 4.1: Included Papers mapped to research methodologies presented in 4.1.3.

Research methodology / **Contribution**	Systematic Review	Practical approach and Implementation	Modeling approach and Simulation
Paper 1: Small-Scale 5G Testbeds for Network Slicing Deployment: A Systematic Review [103]	✓	-	-
Paper 2: A Cloud-based SDN/NFV Testbed for End-to-End Network Slicing in 4G/5G [104]	-	✓	-
Paper 3: Secure Service Implementation with Slice Isolation and WireGuard [116]	-	✓	-
Paper 4: Orchestrating Isolated Network Slices in 5G Networks [117]	-	✓	-
Paper 5: Slicing Scheduling for Supporting Critical Traffic in Beyond 5G [148]	-	-	✓
Paper 6: Beyond 5G Resource Slicing with Mixed-Numerologies for Mission Critical URLLC and eMBB Coexistence [149]	-	-	✓

Table 4.2: Not included Paper mapped to research methodologies presented in 4.1.3.

Research methodology / **Contribution**	Systematic Review	Practical approach and Implementation	Modeling approach and Simulation
Paper 7: 5G Network Slice Isolation with WireGuard and Open Source MANO: A VPNaaS Proof-of-Concept [150]	-	✓	-

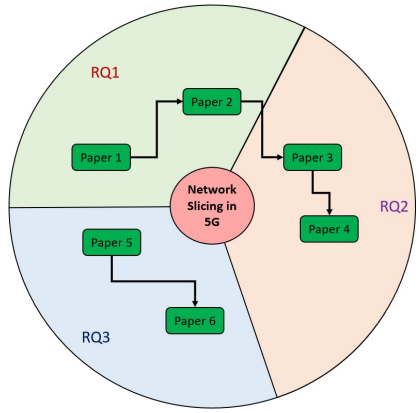


Figure 4.1: Papers' interconnections and their links to research questions presented in 4.1.2.

4.2.1 Summary of the Results Contributing to the Thesis

Paper 1

A. Esmaeily, K. Kralevska, "Small-Scale 5G Testbeds for Network Slicing Deployment: A Systematic Review," Wireless Communications and Mobile Computing, 2021.

The main contribution of the paper is to provide a comprehensive overview of the current state-of-the-art small-scale testbeds for network slice implementation. At first, this work provides a short overview of the network slicing enabling technologies (SDN, NFV, Cloud computing, and MEC), the role of each involved technology in the network slice lifecycle, the fundamental works that focus on large-scale network slicing testbeds, and the lack of similar papers in the case of small-scale testbeds for network slicing. The latter motivates the work in the paper. This paper presents the major open-source software packages to implement network slicing and map them to the ETSI NFV MANO framework. Additionally, primary and secondary design criteria for deploying small-scale testbeds for network slicing are defined. Next, the leading current state-of-the-art small-scale testbeds for network slicing are presented. Subsequently, the predefined design criteria are used as an assessment for them. Finally, common challenges while deploying such testbeds are discussed.

Paper 2

A. Esmaeily, K. Kralevska, D. Gligoroski, "A Cloud-based SDN/NFV Testbed for End-to-End Network Slicing in 4G/5G," 6th IEEE Conference on Network Softwarization (NetSoft), 2020.

While Paper 1 investigates presenting and organizing small-scale testbeds designed with open-source packages for network slicing deployment, Paper 2 further extends Paper 1 by presenting in detail the 5GIIK testbed. Paper 2 introduces a novel testbed called the 5GIIK testbed for network slicing implementation. The testbed comprises several open-source software components. Specifically, the testbed employs 1) the Software Radio Systems LTE (srsLTE) package for deploying the RAN, 2) OpenAirInterface (OAI) for the CN implementation, 3) Open Source MANO (OSM) as the MANO entity, and 4) OpenStack as the VIM in order to perform partial or E2E network slicing. The testbed also integrates SDN-based controllers in its architecture in order to include machine learning-based algorithms in the RAN (5G-EmPOWER controller) and to perform slicing in the TN (M-CORD controller). Some of the main features in the 5G network,

such as utilizing Multi-Radio Access Technologies (M-RATs) and multi-tenancy support, are also addressed in the testbed architecture. The main contribution of the 5GIIK testbed comprises slice provision dynamicity, real-time monitoring of VNFs, and VNF instantiation on different VIMs. Such tasks are accomplished via a hierarchical procedure of creating and onboarding descriptor files for VNF, NS, and NSI levels. In this paper, the 5GIIK has been examined via an initial testing scenario; however, its comprehensive architecture exposes various opportunities for future contributions.

Paper 3

S. Kielland, A. Esmaeily, K. Kralevska, and D. Gligoroski, "Secure Service Implementation with Slice Isolation and WireGuard," IEEE International Mediterranean Conference on Communications and Networking (MeditCom), 2022.

This paper is the first outcome of the work in Paper 2 by utilizing the 5GIIK testbed. It is essential to provide more efficient traffic isolation solutions than the current proposals regarding latency and throughput for the emerging new use cases. This motivates us to present an integration of Virtual Private Network-as-a-Service (VPNaaS) with a service orchestrator component in this work. The framework forms an overlay network that guarantees the confidentiality of data traffic passing between several VNFs in CN slices deployed on a 5G non-standalone network architecture. This is achieved by the integration of the OSM and WireGuard as a relatively simple yet fast and efficient VPN technology that utilizes state-of-the-art cryptography to instantiate OAI EPC VNFs. Specifically, Juju, as the VNFM component of the OSM, interacts with so-called charm files. Charm files handle all the essential procedures to manage the VNFs' lifecycle, such as deployment, configuration, and modification that form WireGuard-enabled VNFs. The performance of the OSM-WireGuard integration in the 5GIIK testbed is examined by measuring the throughput and latency, as interested KPIs, over different interfaces of the OAI EPC for the eMBB and URLLC use cases, respectively. The results exhibit that the integrated OSM-WireGuard framework is a promising solution for 1) providing performance and security isolations in slices with strict latency and throughput requirements and 2) facilitating the configuration and key distribution.

Paper 4

A. Esmaeily and K. Kralevska, "Orchestrating Isolated Network Slices in 5G Networks," under review in EURASIP Journal on Wireless Communications and Networking, 2023.

This paper is an extension of Paper 3. The work presented in Paper 3 provides WireGuard over the OAI EPC VNFs as VPNaaS. Nevertheless, Paper 3 concentrates on delivering VPNaaS with only WireGuard solution on a non-standalone open-source 5G solution. Hence, it is also necessary to have a comparison with the other state-of-the-art VPN solutions in the 5G network in order to make an explicit conclusion. This motivates the work in Paper 4. The experimental tasks in this work are the second outcome of the 5GIIK testbed. The paper focuses on delivering VPNaaS by utilizing WireGuard, IPSec, and OpenVPN solutions. The integration of the VPN solutions with the OSM is referred to as XVPN-OSM architecture. This paper defines specific descriptors for the VNF, NS, and NSI levels according to each VPN solution. In addition, the work implements a cloud-native 5G standalone architecture by using free5GC open-source solution with all the networking components running as CNFs on two separate Kubernetes clusters. The employed approach involves an automated service instantiation that grants end-to-end confidentiality of the traffic passing the CNFs. The OSM Northbound Interface (OSM-NBI) and juju proxy charms easily manage the different VPN solutions with API. The XVPN-OSM architecture provides 1) security isolation that results in secure communication between the involved CNFs of NSs and NSIs, and 2) performance isolation between eMBB with high throughput demand and URLLC with low latency demand. Besides, this architecture demonstrates the advantages/disadvantages of one VPN solution compared to the others. The results verify that the integration of the WireGuard with OSM is still promising, even for complex CNFs in 5G standalone architecture, which is missing in Paper 3.

Paper 5

A. Esmaeily, K. Kralevska, T. Mahmoodi, "Slicing Scheduling for Supporting Critical Traffic in Beyond 5G," 19th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2022.

This paper studies the coexistence of eMBB and URLLC slices over 5G-NR. Due to the distinct QoS requirements of the eMBB slices (with high throughput requirements) and URLLC slices (with extremely high reliability and extremely low latency demands), providing such slices on the same physical infrastructure with a limited amount of resources is challenging. Besides, this process becomes even more challenging when it comes to the RAN because of the stochastic nature of the wireless channels. Hence an efficient scheduling strategy is required for the coexistence of the eMBB and URLLC traffic over the 5G-NR. This paper employs the NOMA resource allocation and puncturing technique to formulate the

achievable rate of the eMBB users by introducing the incoming URLLC traffic as throughput loss of the eMBB users. This paper concentrates on improving the individual average date rate for each eMBB user while fulfilling the URLLC traffic requirements. The paper solves the eMBB/URLLC resource allocation problem by defining a puncturing rate threshold in an optimization algorithm. The proposed optimization algorithm is analyzed via scenarios in which gNB operates with single or multiple antennas toward the eMBB and URLLC users in downlink transmission. The obtained results confirm the efficiency of the proposed algorithm in the resource allocation problem for the coexistence of eMBB and URLLC slices.

Paper 6

A. Esmaeily, H. V. K. Mendis, T. Mahmoodi, K. Kralevska, "Beyond 5G Resource Slicing with Mixed-Numerologies for Mission Critical URLLC and eMBB Coexistence," IEEE Open Journal of the Communications Society, 2023.

This paper is an extension of the work included in Paper 5. While Paper 5 contribution concentrates on allocating radio resources efficiently for eMBB and URLLC coexistence using the puncturing technique; it does not incorporate the various numerologies of the 5G-NR, except for numerology zero. Employing different numerologies grants flexibility in radio resource allocation to different users. Besides, there are remarkably few state-of-the-art studies that focus on radio resource allocation via employing the puncturing technique over various 5G-NR numerologies. These are the motivations behind Paper 6. In addition, Paper 6 reveals its uniqueness by categorizing the URLLC traffic into different classes. In this case, each class represents a portion of the traffic generated by the URLLC users belonging to a particular URLLC use case. In fact, there is no equivalent work that explores together the puncturing method and 5G-NR mixed-numerology to fulfill distinct URLLC classes' requirements (extra-low latency and ultra-high reliability) on the one hand and to maximize the sum rate of the eMBB users, on the other hand. In this way, apart from eMBB users, different URLLC traffic classes can be prioritized as they belong to various URLLC use cases and consequently hold different QoS requirements. The main objective of Paper 6 is to maximize the sum rate of the eMBB users while fulfilling the minimum acceptable data rate of each eMBB user to provide fairness in allocating radio resources. Concurrently, the resource allocation problem has to satisfy the extra low latency and ultra-high reliability requirements of the URLLC users of different URLLC traffic classes. The paper firstly divides the eMBB/URLLC resource allocation problem into three sub-problems and secondly introduces an optimiz-

ation algorithm to obtain efficient outcome. The simulation results demonstrate the applicability and efficiency of the presented algorithm in solving the resource allocation problem for the coexistence of eMBB and URLLC slices.

Chapter 5

Concluding Remarks

5.1 Conclusion

The 5G era and beyond networks involve supporting a rapidly increasing number of use cases and innovative smart services with a wide diversity among them in terms of QoS requirements. Network slicing, which benefits from softwarization, virtualization, and cloudification of the underlying network infrastructure, grants the capability to design and create such smart services in 5G and B5G. As a result, network slicing will transform the perspective of the telecommunication industry.

Firstly, this thesis focuses on practical approaches and modeling strategies for creating network slices with performance and security isolation capabilities. The results of two papers, Paper 1 and Paper 2, indicate the leading practical small-scale solutions towards network slicing implementation and service automation. Moreover, regarding isolation, most of the state-of-the-art works remain in the concept stage. There are only a few works that follow modeling methods for providing isolation between diverse network slices. To the best of our knowledge, there is no significant contribution to pursuing practical solutions for providing isolation between slices. Hence, there is a gap in the state-of-the-art to propose feasible answers to the isolation problem in network slicing. Additionally, the work in this research goes one step further to contribute to service automation by involving slice isolation. We provide descriptors for network slice, network service, and network function levels to practically create, monitor, manage, and control each specific slice during the whole lifecycle of that slice. We also include isolation capability in the service provisioning process. The outcomes of Paper 3 and Paper 4 confirm that the appropriate integration of a service orchestrator and a high-performance VPN solution can efficiently provide traffic isolation between slices. Such isolation guarantees the confidentiality of traffic crossing various VNFs/CNFs in a 5G network slice while satisfying the through-

put and delay requirements for the eMBB and URLLC use cases, respectively.

It is worth mentioning that although Paper 2, Paper 3, and Paper 4 show promising solutions in the 5G and B5G, small-scale testbeds for network slicing implementation may encounter several scalability issues including:

- **Limited resource capacity**: Small-scale testbeds need more computing, storage, and networking resources to emulate larger-scale network environments effectively.
- **Inadequate traffic generation**: Generating realistic traffic at a smaller scale may not sufficiently reflect the complexities of traffic patterns in larger networks.
- **Performance and management overheads**: As testbeds expand, they may encounter increased performance and management overheads, resulting in delays and reduced efficiency.
- **Interference**: As the number of network slices increases, interference between them may become a concern, impacting the QoS requirements of each of them.
- **Scalability of MANO**: Managing and orchestrating multiple network slices can be complex, and small-scale testbeds may not fully address these challenges.
- Network component scalability: The scalability of individual network elements, such as VNFs/CNFs in a small-scale testbed, may not accurately reflect real-world scenarios.
- **Diversity in slice types**: Scalability issues appear when supporting diverse slice types with distinct QoS demands on the same infrastructure.
- **Dynamic scalability**: Network slicing should ideally support dynamic scaling of resources based on user demand, which can be challenging in small-scale testbeds due to limited resources.
- **Security concerns**: Expanding the scale can introduce security vulnerabilities that may not be quite noticeable in small-scale environments.

Consequently, addressing these scalability challenges in small-scale testbeds is essential to effectively validate network slicing solutions and qualify them for large-scale real-world implementations.

Secondly, this thesis discusses, analyzes, and addresses the resource allocation problem for the eMBB and URLLC traffic types in the 5G-NR. By following power-domain NOMA with a puncturing approach, Paper 5 formulates the coexistence of the eMBB and URLLC slices and proposes an algorithm to efficiently solve the eMBB/URLLC resource allocation problem. In this paper, we evaluate the performance under different network setups in order to prove the necessity of applying our algorithm to the eMBB/URLLC resource allocation problem. As a complementary work to the Paper 5, Paper 6 concentrates on the coexistence problem of eMBB and URLLC use cases over different 5G-NR numerologies while

applying the puncturing technique in the numerologies. Besides, by introducing the URLLC traffic classes in Paper 6, we go one step forward to deliver a realistic landscape of a 5G network, on the one hand, with eMBB users, and on the other hand, with URLLC users with different QoS requirements. Simulation results confirm the applicability and efficiency of our proposed resource allocation method.

5.2 Future directions

In the contemporary landscape and even more so in the coming years, vertical industries are increasingly claiming their need for networks that can efficiently and dynamically accommodate their specific use cases. The solution to this growing demand lies in network slicing, which offers the flexibility required to cater to a diverse collection of industry requirements. As the initial phases of 5G deployment have already been initiated in various regions worldwide, the significance of network slicing, situated at the heart of both 5G and B5G networks, cannot be overstated. Although the initial launching phase of 5G has already started in some regions worldwide, network slicing, as an evolving technology in the center of 5G and B5G networks, deserves more consideration from academia and industry.

Since the trend of applying AI-based approaches is exponentially growing, one possible and expected opportunity for future directions could be integrating AI into network slicing. Some primary attempts have been started [151, 152], yet AI-Network Slicing integration necessitates further attention. The application of AI could potentially revolutionize the way network slices are managed, optimized, and dynamically adjusted in response to varying demands.

Furthermore, as the horizon of network slicing extends into B5G networks, routes for enhanced innovation emerge. The increase of IoT devices, the promise of ubiquitous connectivity, and the overflow in data-intensive applications show a landscape of possibilities. In this context, orchestrating complex network slices to accommodate the demands of diverse industries and emerging technologies becomes an exciting realm for future exploration. The potential of leveraging network slicing for cross-domain collaboration and ensuring seamless integration of industries and services is a promising direction.

In summary, the future direction for this thesis stands as a compelling point in which the convergence of network slicing, AI, and the ever-evolving landscape of 5G and B5G networks offer productive grounds for exploration.

Bibliography

[1] F. Qi, X. Zhu, G. Mang, M. Kadoch and W. Li, 'Uav network and iot in the sky for future smart cities,' *IEEE Network*, vol. 33, no. 2, pp. 96–101, 2019. DOI: 10.1109/MNET.2019.1800250.

[2] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran and S. Guizani, 'Internet-of-things-based smart cities: Recent advances and challenges,' *IEEE Communications Magazine*, vol. 55, no. 9, pp. 16–24, 2017. DOI: 10.1109/MCOM.2017.1600514.

[3] M. Sookhak, H. Tang, Y. He and F. R. Yu, 'Security and privacy of smart cities: A survey, research issues and challenges,' *IEEE Communications Surveys Tutorials*, vol. 21, no. 2, pp. 1718–1743, 2019. DOI: 10.1109/COMST.2018.2867288.

[4] R. W, '5g mobile communications for 2020 and beyond vision and key enabling technologies,' 2014. [Online]. Available: https://eucnc.eu/files/keynotes/Roh.pdf.

[5] S. Research, *6g: The next hyper-connected experience for all*, 2020.

[6] A. Antonopoulos, 'Bankruptcy problem in network sharing: Fundamentals, applications and challenges,' *IEEE Wireless Communications*, vol. 27, no. 4, pp. 81–87, 2020. DOI: 10.1109/MWC.001.1900414.

[7] 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*, vol. 55, no. 5, pp. 80–87, 2017. DOI: 10.1109/MCOM.2017.1600935.

[8] 3GPP TR 28.801 (V15.1.0), 'Study on management and orchestration of network slicing for next generation network (release 15),' Jan. 2018.

[9] NGMN. document TR. v1.0.8., 'Description of network slicing concept,' Sep. 2016.

[10] ETSI. document TR. v3.1.1, 'Gr nfv-eve 012; network functions virtualisation (nfv) release 3; evolution and ecosystem,' Dec. 2017.

[11] R. P. Goldberg, 'Survey of virtual machine research,' *Computer*, vol. 7, no. 6, pp. 34–45, 1974. DOI: 10.1109/MC.1974.6323581.

[12] R. P. Goldberg, 'Survey of virtual machine research,' *Computer*, vol. 7, no. 6, pp. 34–45, 1974.

[13] D. Clark, W. Lehr, S. Bauer, P. Faratin, R. Sami and J. Wroclawski, 'The growth of internet overlay networks: Implications for architecture, industry structure and policy,' in *33rd Research Conference on Communication, Information and Internet Policy, Arlington, Virginia*, Citeseer, 2005.

[14] GSMA, 'Mobile infrastructure sharing,' Sep. 2012.

[15] 3GPP Standard TS 23.251, 'Network sharing; architecture and functional description, release 12,' Mar. 2015.

[16] A. Brydon, '3gpp network sharing enhancements for lte,' *WIRELESS BLOG*, 2013. [Online]. Available: https://www.unwiredinsight.com/2013/3gpp-lte-ran-sharing-enhancements.

[17] R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten, F. De Turck and R. Boutaba, 'Network function virtualization: State-of-the-art and research challenges,' *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.

[18] Int. Telecommun. Union, ITU-T Recommendation Y.3011, 'Framework of network virtualization for future networks, next generation network—future networks,' Jan. 2012.

[19] Kelvin Qin, 'Network slicing, use case requirements,' Apr. 2018. [Online]. Available: https://www.gsma.com/futurenetworks/wp-content/uploads/2018/06/Network-Slicing-Use-Case-Requirements-_-FInal-.pdf.

[20] O. Queseth, Ö. Bulakci, P. Spapis, P. Bisson, P. Marsch, P. Arnold, P. Rost, Q. Wang, R. Blom, S. Salsano *et al.*, '5g ppp architecture working group: View on 5g architecture (version 2.0, december 2017),' 2017. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2018/01/5G-PPP-5G-Architecture-White-Paper-Jan-2018-v2.0.pdf.

[21] A. de la Oliva, X. Li, X. Costa-Perez, C. J. Bernardos, P. Bertin, P. Iovanna, T. Deiss, J. Mangues, A. Mourad, C. Casetti, J. E. Gonzalez and A. Azcorra, '5g-transformer: Slicing and orchestrating transport networks for industry verticals,' *IEEE Communications Magazine*, vol. 56, no. 8, pp. 78–84, 2018. DOI: 10.1109/MCOM.2018.1700990.

[22] 3GPP TR 22.863, 'Feasibility study on new services and markets technology enablers for enhanced mobile broadband, release 14,' Jun. 2016.

[23] 3GPP TR 22.862, 'Feasibility study on new services and markets technology enablers for critical communications, release 14,' Jun. 2016.

[24] 3GPP TR 22.861, 'Feasibility study on new services and markets technology enablers for massive internet of things, release 14,' Jun. 2016.

[25] S. Bicheno, '5g reaches an anticlimax at mwc 2019,' 2019. [Online]. Available: https://telecoms.com/495842/5g-reaches-an-anticlimax-at-mwc-2019/.

[26] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, 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.

[27] P. M. Mell and T. Grance, 'Sp 800-145. the nist definition of cloud computing,' Gaithersburg, MD, USA, Tech. Rep., 2011.

[28] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, 'A survey on mobile edge computing: The communication perspective,' *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[29] W. P. ETSI NFV, 'Mobile-edge computing—introductory technical,' Sep. 2014.

[30] M. Abdelshkour, 'Iot, from cloud to fog computing,' *Cisco Blogs*, 2017.

[31] ETSI, 'Network functions virtualisation (nfv): Architectural framework,' *ETSI Gs NFV*, vol. 2, no. 2, 2013.

[32] Q. Li, G. Wu, A. Papathanassiou and M. Udayan, 'An end-to-end network slicing framework for 5g wireless communication systems,' *ArXiv*, vol. abs/1608.00572, 2016.

[33] C. I. Wireless World Research Forum White Paper 3, 'End to end network slicing,' Nov. 2017. [Online]. Available: https://www.wwrf.ch/files/content%5C%20wwrf/publications/outlook/Outlook21.pdf.

[34] H. Yan, 'End-to-end 5g network slicing: Key to digital transformation,' *ZTE*, 2018. [Online]. Available: https://www.zte.com.cn/global/about/magazine/zte-technologies/2018/1/Special-Topic/5G-network-Slicing.html.

[35] R. F. Olimid and G. Nencioni, '5g network slicing: A security overview,' *IEEE Access*, vol. 8, pp. 99 999–100 009, 2020. DOI: 10.1109/ACCESS.2020.2997702.

[36] Z. Kotulski, T. W. Nowak, M. Sepczuk, M. A. Tunia, R. Artych, K. Bocianiak, T. Osko and J.-P. Wary, 'Towards constructive approach to end-to-end slice isolation in 5g networks,' *EURASIP J. Inf. Secur.*, vol. 2018, p. 2, 2018. DOI: 10.1186/s13635-018-0072-0. [Online]. Available: https://doi.org/10.1186/s13635-018-0072-0.

[37] Z. Kotulski, T. W. Nowak, M. Sepczuk and M. A. Tunia, '5g networks: Types of isolation and their parameters in ran and cn slices,' *Computer Networks*, vol. 171, p. 107 135, 2020, ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2020.107135. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128619304797.

[38] A. J. Gonzalez, J. Ordonez-Lucena, B. E. Helvik, G. Nencioni, M. Xie, D. R. Lopez and P. Grønsund, 'The isolation concept in the 5g network slicing,' in *2020 European Conference on Networks and Communications (EuCNC)*, 2020, pp. 12–16. DOI: 10.1109/EuCNC48522.2020.9200939.

[39] A. Banchs, G. de Veciana, V. Sciancalepore and X. Costa-Perez, 'Resource allocation for network slicing in mobile networks,' *IEEE Access*, vol. 8, pp. 214 696–214 706, 2020. DOI: 10.1109/ACCESS.2020.3040949.

[40] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez and A. Azcorra, 'Network slicing for guaranteed rate services: Admission control and resource allocation games,' *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6419–6432, 2018. DOI: 10.1109/TWC.2018.2859918.

[41] P. Caballero, A. Banchs, G. de Veciana and X. Costa-Pérez, 'Network slicing games: Enabling customization in multi-tenant networks,' in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 2017, pp. 1–9. DOI: 10.1109/INFOCOM.2017.8057046.

[42] P. Caballero, A. Banchs, G. de Veciana and X. Costa-Pérez, 'Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads,' *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 3044–3058, 2017. DOI: 10.1109/TNET.2017.2720668.

[43] J. Zheng, P. Caballero, G. de Veciana, S. J. Baek and A. Banchs, 'Statistical multiplexing and traffic shaping games for network slicing,' in *2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2017, pp. 1–8. DOI: 10.23919/WIOPT.2017.7959883.

[44] M. Leconte, G. S. Paschos, P. Mertikopoulos and U. C. Kozat, 'A resource allocation framework for network slicing,' in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 2177–2185. DOI: 10.1109/INFOCOM.2018.8486303.

[45] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso and C. Verikoukis, 'Multi-tenant slicing for spectrum management on the road to 5g,' *IEEE Wireless Communications*, vol. 24, no. 5, pp. 118–125, 2017. DOI: 10.1109/MWC.2017.1700138.

[46] H.-T. Chien, Y.-D. Lin, C.-L. Lai and C.-T. Wang, 'End-to-end slicing as a service with computing and communication resource allocation for multi-tenant 5g systems,' *IEEE Wireless Communications*, vol. 26, no. 5, pp. 104–112, 2019. DOI: 10.1109/MWC.2019.1800466.

[47] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia and A. Banchs, 'Mobile traffic forecasting for maximizing 5g network slicing resource utilization,' in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 2017, pp. 1–9. DOI: 10.1109/INFOCOM.2017.8057230.

[48] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore and X. Costa-Pérez, 'A machine learning approach to 5g infrastructure market optimization,' *IEEE Transactions on Mobile Computing*, vol. 19, no. 3, pp. 498–512, 2020. DOI: 10.1109/TMC.2019.2896950.

[49] D. Bega, M. Gramaglia, M. Fiore, A. Banchs and X. Costa-Perez, 'Deepcog: Cognitive network management in sliced 5g networks with deep learning,' in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019, pp. 280–288. DOI: 10.1109/INFOCOM.2019.8737488.

[50] F. Wei, G. Feng, Y. Sun, Y. Wang, S. Qin and Y.-C. Liang, 'Network slice reconfiguration by exploiting deep reinforcement learning with large action space,' *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2197–2211, 2020. DOI: 10.1109/TNSM.2020.3019248.

[51] 3GPP TR 38.912, 'Study on new radio access technology,' Aug. 2017.

[52] 3GPP TR 38.913, 'Study on scenarios and requirements for next generation access technologies,' Aug. 2017.

[53] 3GPP TS 38.201, 'Nr; physical layer; general description,' Jun. 2018.

[54] 3GPP TS 38.202, 'Nr; services provided by the physical layer,' Jun. 2018.

[55] 3GPP TS 38.331, 'Nr; radio resource control (rrc); protocol specification,' Jun. 2018.

[56] 3GPP TS 38.101-4, 'User equipment (ue) radio transmission and reception; part 4: Performance requirements.,' Mar. 2020.

[57] P. J, '5g techniques for ultra reliable low latency communication,' 2017. [Online]. Available: https://cscn2017.ieee-cscn.org/files/2017/08/Janne_Peisa_Ericsson_CSCN2017.pdf.

[58] A. A. Zaidi, R. Baldemair, V. Moles-Cases, N. He, K. Werner and A. Cedergren, 'Ofdm numerology design for 5g new radio to support iot, embb, and mbsfn,' *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 78–83, 2018. DOI: 10.1109/MCOMSTD.2018.1700021.

[59] A. W. Scott and R. Frobenius, 'Multiple access techniques: Fdma, tdma, and cdma,' in *RF Measurements for Cellular Phones and Wireless Data Systems*. 2008, pp. 413–429. DOI: 10.1002/9780470378014.ch30.

[60] H. Li, G. Ru, Y. Kim and H. Liu, 'Ofdma capacity analysis in mimo channels,' *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4438–4446, 2010. DOI: 10.1109/TIT.2010.2054710.

[61] 3GPP, 'Technical specification group services and system aspects; release 15 description,' 3GPP, Technical Specification (TS), Mar. 2019, Version 1.1.0.

[62] Z. Wu, K. Lu, C. Jiang and X. Shao, 'Comprehensive study and comparison on 5g noma schemes,' *IEEE Access*, vol. 6, pp. 18 511–18 519, 2018. DOI: 10.1109/ACCESS.2018.2817221.

[63] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen and L. Hanzo, 'A survey of non-orthogonal multiple access for 5g,' *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018. DOI: 10.1109/COMST.2018.2835558.

[64] Y. Chen, A. Bayesteh, Y. Wu, B. Ren, S. Kang, S. Sun, Q. Xiong, C. Qian, B. Yu, Z. Ding, S. Wang, S. Han, X. Hou, H. Lin, R. Visoz and R. Razavi, 'Toward the standardization of non-orthogonal multiple access for next generation wireless networks,' *IEEE Communications Magazine*, vol. 56, no. 3, pp. 19–27, 2018. DOI: 10.1109/MCOM.2018.1700845.

[65] P. Popovski, K. F. Trillingsgaard, O. Simeone and G. Durisi, '5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view,' *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018. DOI: 10.1109/ACCESS.2018.2872781.

[66] R. Hoshyar, F. P. Wathan and R. Tafazolli, 'Novel low-density signature for synchronous cdma systems over awgn channel,' *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1616–1626, 2008. DOI: 10.1109/TSP.2007.909320.

[67] D. Guo and C.-c. Wang, 'Multiuser detection of sparsely spread cdma,' *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 3, pp. 421–431, 2008. DOI: 10.1109/JSAC.2008.080402.

[68] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada and T. Nakamura, 'Concept and practical considerations of non-orthogonal multiple access (noma) for future radio access,' in *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, 2013, pp. 770–774. DOI: 10.1109/ISPACS.2013.6704653.

[69] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li and K. Higuchi, 'Non-orthogonal multiple access (noma) for cellular future radio access,' in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, 2013, pp. 1–5. DOI: 10.1109/VTCSpring.2013.6692652.

[70] A. Anand, G. De Veciana and S. Shakkottai, 'Joint scheduling of urllc and embb traffic in 5g wireless networks,' in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 1970–1978. DOI: 10.1109/INFOCOM.2018.8486430.

[71] F. Kaltenberger, A. P. Silva, A. Gosain, L. Wang and T.-T. Nguyen, 'Openairinterface: Democratizing innovation in the 5g era,' *Computer Networks*, vol. 176, p. 107 284, 2020, ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2020.107284. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128619314410.

[72] B. Dzogovic, V. T. Do, B. Feng and T. van Do, 'Building virtualized 5g networks using open source software,' in *2018 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, 2018, pp. 360–366.

[73] B. Dzogovic, B. Santos, V. T. Do, B. Feng, N. Jacot and T. Van Do, 'Connecting remote enodeb with containerized 5g c-rans in openstack cloud,' in *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/ 2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, 2019, pp. 14–19.

[74] M. Arif, O. Liinamaa, I. Ahmad, A. Pouttu and M. Ylianttila, 'On the demonstration and evaluation of service-based slices in 5g test network using nfv,' in *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*, 2019, pp. 1–6.

[75] L. A. Freitas, V. G. Braga, S. L. Corrêa, L. Mamatas, C. E. Rothenberg, S. Clayman and K. V. Cardoso, 'Slicing and allocation of transformable resources for the deployment of multiple virtualized infrastructure managers (vims),' in *4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, 2018, pp. 424–432.

[76] D. Sattar and A. Matrawy, 'Dsaf: Dynamic slice allocation framework for 5g core network,' *ArXiv*, vol. abs/1905.03873, 2019.

[77] J. Kim and M. Xie, 'A study of slice-aware service assurance for network function virtualization,' in *2019 IEEE Conference on Network Softwarization (NetSoft)*, Jun. 2019, pp. 489–497.

[78] T. Dreibholz, 'Flexible 4g/5g testbed setup for mobile edge computing using openairinterface and open source mano,' in *Web, Artificial Intelligence and Network Applications*, Springer International Publishing, 2020, pp. 1143–1153.

[79] A. F. Ocampo, T. Dreibholz, M.-r. Fida, A. Elmokashfi and H. Bryhni, 'Integrating cloud-ran with packet core as vnf using open source mano and openairinterface,' Sydney, New South Wales/Australia: IEEE Computer Society, Nov. 2020.

[80] I. Sarrigiannis, E. Kartsakli, K. Ramantas, A. Antonopoulos and C. Verikoukis, 'Application and network vnf migration in a mec-enabled 5g architecture,' in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2018, pp. 1–6. DOI: 10.1109/CAMAD.2018.8514943.

[81] I. Sarrigiannis, K. Ramantas, E. Kartsakli, P. Mekikis, A. Antonopoulos and C. Verikoukis, 'Online vnf lifecycle management in an mec-enabled 5g iot architecture,' *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4183–4194, 2020. DOI: 10.1109/JIOT.2019.2944695.

[82] O.-M. Ungureanu, C. Vlădeanu and R. Kooij, 'Collaborative cloud-edge: A declarative api orchestration model for the nextgen 5g core,' in *2021 IEEE international conference on service-oriented system engineering (SOSE)*, IEEE, 2021, pp. 124–133.

[83] X. Foukas, M. K. Marina and K. P. Kontovasilis, 'Orion: Ran slicing for a flexible and cost-effective multi-service mobile network architecture,' *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017.

[84] X. Foukas, F. Sardis, F. Foster, M. K. Marina, M. A. Lema and M. Dohler, 'Experience building a prototype 5g testbed,' in *Proceedings of the Workshop on Experimentation and Measurements in 5G*, 2018, pp. 13–18.

[85] K. Koutlia, R. Ferrus, E. Coronado Calero, R. Riggio, F. Palacio, A. Umbert and J. Pérez-Romero, 'Design and experimental validation of a software-defined radio access network testbed with slicing support,' *Wireless Communications and Mobile Computing*, vol. 2019, pp. 1–17, Jun. 2019. DOI: 10.1155/2019/2361352.

[86] S. Costanzo, I. Fajjari, N. Aitsaadi and R. Langar, 'Dynamic network slicing for 5g iot and embb services: A new design with prototype and implementation results,' in *2018 3rd Cloudification of the Internet of Things (CIoT)*, 2018, pp. 1–7.

[87] C. V. Nahum, L. De Nóvoa Martins Pinto, V. B. Tavares, P. Batista, S. Lins, N. Linder and A. Klautau, 'Testbed for 5g connected artificial intelligence on virtualized networks,' *IEEE Access*, vol. 8, pp. 223 202–223 213, 2020. DOI: 10.1109/ACCESS.2020.3043876.

[88] S. Ganji, S. Behnaminia, A. Ahangarpour, E. Mazaheri, S. Baradaran, Z. Zali, M. R. Heidarpour, A. Rakhshan and M. F. Shoyari, *Cn2f: A cloud-native cellular network framework*, 2023. arXiv: 2305.18778 [cs.NI].

[89] P. Mekikis, K. Ramantas, A. Antonopoulos, E. Kartsakli, L. Sanabria-Russo, J. Serra, D. Pubill and C. Verikoukis, 'Nfv-enabled experimental platform for 5g tactile internet support in industrial environments,' *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 1895–1903, 2020.

[90] N. Nikaein, C.-Y. Chang and K. Alexandris, 'Mosaic5g: Agile and flexible service platforms for 5g research,' *SIGCOMM Comput. Commun. Rev.*, vol. 48, no. 3, pp. 29–34, Sep. 2018, ISSN: 0146-4833. DOI: 10.1145/3276799.3276803. [Online]. Available: https://doi.org/10.1145/3276799.3276803.

[91] A. Shorov, '5g testbed development for network slicing evaluation,' in *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2019, pp. 39–44.

[92] G. Garcia-Aviles, M. Gramaglia, P. Serrano and A. Banchs, 'Posens: A practical open source solution for end-to-end network slicing,' *IEEE Wireless Communications*, vol. 25, no. 5, pp. 30–37, 2018.

[93] G. Garcia-Aviles, M. Gramaglia, P. Serrano, F. Gringoli, S. Fuente-Pascual and I. L. Pavon, 'Experimenting with open source tools to deploy a multi-service and multi-slice mobile network,' *Computer Communications*, vol. 150, pp. 1–12, 2020.

[94] C. Huang, C. Ho, N. Nikaein and R. Cheng, 'Design and prototype of a virtualized 5g infrastructure supporting network slicing,' in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, 2018, pp. 1–5.

[95] M. T. Abbas, T. A. Khan, A. Mahmood, J. J. D. Rivera and W. Song, 'Introducing network slice management inside m-cord-based-5g framework,' in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, 2018, pp. 1–2.

[96] Q. Wang, J. Alcaraz-Calero, R. Ricart-Sanchez, M. B. Weiss, A. Gavras, N. Nikaein, X. Vasilakos, B. Giacomo, G. Pietro, M. Roddy, M. Healy, P. Walsh, T. Truong, Z. Bozakov, K. Koutsopoulos, P. Neves, C. Patachia-Sultanoiu, M. Iordache, E. Oproiu, I. G. B. Yahia, C. Angelo, C. Zotti, G. Celozzi, D. Morris, R. Figueiredo, D. Lorenz, S. Spadaro, G. Agapiou, A. Aleixo and C. Lomba, 'Enable advanced qos-aware network slicing in 5g networks for slice-based media use cases,' *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 444–453, 2019.

[97] V. Q. Rodriguez, F. Guillemin and A. Boubendir, '5g e2e network slicing management with onap,' in *2020 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, 2020, pp. 87–94.

[98] K. Abbas, T. A. Khan, M. Afaq and W.-C. Song, 'Network slice lifecycle management for 5g mobile networks: An intent-based networking approach,' *IEEE Access*, vol. 9, pp. 80 128–80 146, 2021. DOI: 10.1109/ACCESS.2021.3084834.

[99] Y.-S. Chiu, L.-H. Yen, T.-H. Wang and C.-C. Tseng, 'A cloud native management and orchestration framework for 5g end-to-end network slicing,' in *2022 IEEE International Conference on Service-Oriented System Engineering (SOSE)*, IEEE, 2022, pp. 69–76.

[100] L. Wang, J. Wu, Y. Gao and J. Zhang, 'Deep reinforcement learning based resource allocation for cloud native wireless network,' *arXiv:2305.06249*, 2023.

[101] S. Kahvazadeh, H. Khalili, R. N. Silab, B. Bakhshi and J. Mangues-Bafalluy, 'Vertical-oriented 5g platform-as-a-service: User-generated content case study,' in *2022 IEEE Future Networks World Forum (FNWF)*, IEEE, 2022, pp. 706–711.

[102] S. Barrachina-Muñoz, M. Payaró and J. Mangues-Bafalluy, 'Cloud-native 5g experimental platform with over-the-air transmissions and end-to-end monitoring,' in *2022 13th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, IEEE, 2022, pp. 692–697.

[103] A. Esmaeily and K. Kralevska, 'Small-scale 5g testbeds for network slicing deployment: A systematic review,' *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–26, 2021.

[104] A. Esmaeily, K. Kralevska and D. Gligoroski, 'A cloud-based sdn/nfv testbed for end-to-end network slicing in 4g/5g,' in *2020 6th IEEE Conference on Network Softwarization (NetSoft)*, IEEE, 2020, pp. 29–35.

[105] V. Sathi, M. Srinivasan, P. Kaliyammal Thiruvasagam and S. Chebiyyam, 'A novel protocol for securing network slice component association and slice isolation in 5g networks,' Oct. 2018, pp. 249–253. DOI: 10.1145/3242102.3242135.

[106] P. Porambage, Y. Miche, A. Kalliola, M. Liyanage and M. Ylianttila, 'Secure keying scheme for network slicing in 5g architecture,' in *2019 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2019, pp. 1–6. DOI: 10.1109/CSCN.2019.8931330.

[107] B. Niu, W. You, H. Tang and X. Wang, '5g network slice security trust degree calculation model,' in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017, pp. 1150–1157. DOI: 10.1109/CompComm.2017.8322724.

[108] J. Ni, X. Lin and X. S. Shen, 'Efficient and secure service-oriented authentication supporting network slicing for 5g-enabled iot,' *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 644–657, 2018. DOI: 10.1109/JSAC.2018.2815418.

[109] C. M. Moreira, G. Kaddoum and E. Bou-Harb, 'Cross-layer authentication protocol design for ultra-dense 5g hetnets,' in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–7. DOI: 10.1109/ICC.2018.8422404.

[110] X. Li, C. Guo, L. Gupta and R. Jain, 'Efficient and secure 5g core network slice provisioning based on vikor approach,' *IEEE Access*, vol. 7, pp. 150 517–150 529, 2019. DOI: 10.1109/ACCESS.2019.2947454.

[111] J. Y. Yen and J. Y. YENt, 'Finding the k shortest loopless paths in a network,' 2007.

[112] D. Marabissi and R. Fantacci, 'Highly flexible ran slicing approach to manage isolation, priority, efficiency,' *IEEE Access*, vol. 7, pp. 97 130–97 142, 2019. DOI: 10.1109/ACCESS.2019.2929732.

[113] G. Sun, K. Xiong, G. O. Boateng, D. Ayepah-Mensah, G. Liu and W. Jiang, 'Autonomous resource provisioning and resource customization for mixed traffics in virtualized radio access network,' *IEEE Systems Journal*, vol. 13, no. 3, pp. 2454–2465, 2019. DOI: 10.1109/JSYST.2019.2918005.

[114] F. Boutigny, S. Betgé-Brezetz, G. Blanc, A. Lavignotte, H. Debar and H. Jmila, 'Solving security constraints for 5g slice embedding: A proof-of-concept,' *Comput. Secur.*, vol. 89, 2020. DOI: 10.1016/j.cose.2019.101662. [Online]. Available: https://doi.org/10.1016/j.cose.2019.101662.

[115] D. Sattar and A. Matrawy, 'Towards secure slicing: Using slice isolation to mitigate ddos attacks on 5g core network slices,' in *2019 IEEE Conference on Communications and Network Security (CNS)*, 2019, pp. 82–90. DOI: 10.1109/CNS.2019.8802852.

[116] S. Kielland, A. Esmaeily, K. Kralevska and D. Gligoroski, 'Secure service implementation with slice isolation and wireguard,' in *2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, IEEE, 2022, pp. 148–153.

[117] A. Esmaeily and K. Kralevska, 'Orchestrating isolated network slices in 5g networks,' *EURASIP Journal on Wireless Communications and Networking*, vol. 2023, 2023.

[118] J. Park and M. Bennis, 'Urllc-embb slicing to support vr multimodal perceptions over wireless cellular systems,' Dec. 2018, pp. 1–7. DOI: 10.1109/GLOCOM.2018.8647208.

[119] S. F. Abedin, M. G. R. Alam, S. M. A. Kazmi, N. H. Tran, D. Niyato and C. S. Hong, 'Resource allocation for ultra-reliable and enhanced mobile broadband iot applications in fog network,' *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 489–502, 2019. DOI: 10.1109/TCOMM.2018.2870888.

[120] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi and C. S. Hong, 'Embb-urllc resource slicing: A risk-sensitive approach,' *IEEE Communications Letters*, vol. 23, no. 4, pp. 740–743, 2019. DOI: 10.1109/LCOMM.2019.2900044.

[121] X. Zhang, X. Guo and H. Zhang, 'RB Allocation Scheme for eMBB and URLLC Coexistence in 5G and Beyond,' *Wireless Communications and Mobile Computing*, vol. 2021, 2021.

[122] M. Alsenwi, N. Tran, M. Bennis, S. Pandey, A. Bairagi and C. S. Hong, 'Intelligent resource slicing for embb and urllc coexistence in 5g and beyond: A deep reinforcement learning based approach,' *IEEE Transactions on Wireless Communications*, vol. PP, pp. 1–1, Feb. 2021. DOI: 10.1109/TWC.2021.3060514.

[123] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran and C. S. Hong, 'A matching based coexistence mechanism between embb and urllc in 5g wireless networks,' in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '19, Limassol, Cyprus: Association for Computing Machinery, 2019, pp. 2377–2384, ISBN: 9781450359337.

[124] A. Bairagi, M. Munir, M. Alsenwi, N. Tran, S. Alshamrani, M. Masud, Z. Han and C. S. Hong. (Mar. 2020). 'Coexistence mechanism between embb and urllc in 5g wireless networks.'

[125] J. Tang, B. Shim and T. Q. S. Quek, 'Service multiplexing and revenue maximization in sliced c-ran incorporated with urllc and multicast embb,' *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 881–895, 2019. DOI: 10.1109/JSAC.2019.2898745.

[126] R. Kassab, O. Simeone and P. Popovski, 'Coexistence of urllc and embb services in the c-ran uplink: An information-theoretic study,' in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6. DOI: 10.1109/GLOCOM.2018.8647460.

[127] K. I. Pedersen, G. Pocovi, J. Steiner and S. R. Khosravirad, 'Punctured scheduling for critical low latency data on a shared channel with mobile broadband,' in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, IEEE, 2017, pp. 1–6.

[128] L. Marijanović, S. Schwarz and M. Rupp, 'Optimal resource allocation with flexible numerology,' in *2018 IEEE International Conference on Communication Systems (ICCS)*, IEEE, 2018, pp. 136–141.

[129] L. Marijanovic, S. Schwarz and M. Rupp, 'A novel optimization method for resource allocation based on mixed numerology,' in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, IEEE, 2019, pp. 1–6.

[130] T. T. Nguyen, V. N. Ha and L. B. Le, 'Wireless scheduling for heterogeneous services with mixed numerology in 5G wireless networks,' *IEEE Communications Letters*, vol. 24, no. 2, pp. 410–413, 2019.

[131] L. Marijanović, S. Schwarz and M. Rupp, 'Multiplexing Services in 5G and Beyond: Optimal Resource Allocation Based on Mixed Numerology and Mini-Slots,' *IEEE Access*, vol. 8, pp. 209 537–209 555, 2020.

[132] Y. Prathyusha and T.-L. Sheu, 'Coordinated resource allocations for embb and urllc in 5g communication networks,' *IEEE Transactions on Vehicular Technology*, 2022.

[133] M. Zambianco and G. Verticale, 'Mixed-numerology interference-aware spectrum allocation for embb and urllc network slices,' in *2021 19th Mediterranean Communication and Computer Networking Conference (MedComNet)*, 2021, pp. 1–8. DOI: 10.1109/MedComNet52149.2021.9501277.

[134] M. Zambianco, 'A reinforcement learning agent for mixed-numerology interference-aware slice spectrum allocation with non-deterministic and deterministic traffic,' *Computer Communications*, vol. 189, pp. 100–109, 2022, ISSN: 0140-3664. DOI: https://doi.org/10.1016/j.comcom.2022.03.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366422000858.

[135] M. Setayesh, S. Bahrami and V. W. Wong, 'Resource slicing for embb and urllc services in radio access network using hierarchical deep learning,' *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 8950–8966, 2022. DOI: 10.1109/TWC.2022.3171264.

[136] M. Mhedhbi, M. Morcos, A. Galindo-Serrano and S. E. Elayoubi, 'Performance evaluation of 5g radio configurations for industry 4.0,' in *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2019, pp. 1–6. DOI: 10.1109/WiMOB.2019.8923609.

[137] D. Kotagiri, A. Sawabe, E. Takahashi, T. Iwai, T. Onishi and Y. Nishikawa, 'Context-based mixed-numerology profile selection for 5g and beyond,' in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, IEEE, 2022, pp. 611–616.

[138] K. Boutiba, M. Bagaa and A. Ksentini, 'Radio resource management in multi-numerology 5g new radio featuring network slicing,' in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 359–364. DOI: 10.1109/ICC45855.2022.9838462.

[139] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi and P. Mogensen, 'Efficient low complexity packet scheduling algorithm for mixed urllc and embb traffic in 5g,' in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, 2019, pp. 1–6. DOI: 10.1109/VTCSpring.2019.8746407.

[140] A. Destounis and G. S. Paschos, 'Complexity of urllc scheduling and efficient approximation schemes,' in *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2019, pp. 1–8. DOI: 10.23919/WiOPT47501.2019.9144114.

[141] Y. Zhao, X. Chi, L. Qian, Y. Zhu and F. Hou, 'Resource allocation and slicing puncture in cellular networks with embb and urllc terminals coexistence,' *IEEE Internet of Things Journal*, 2022.

[142] L.-H. Shen, C.-Y. Su and K.-T. Feng, *Comp enhanced subcarrier and power allocation for multi-numerology based 5g-nr networks*, 2021. arXiv: 2112.04070 [cs.NI].

[143] M. Almekhlafi, M. A. Arfaoui, C. Assi and A. Ghrayeb, 'Superposition-based urllc traffic scheduling in 5g and beyond wireless networks,' *IEEE Transactions on Communications*, vol. 70, no. 9, pp. 6295–6309, 2022. DOI: 10.1109/TCOMM.2022.3194018.

[144] W. Ning, Y. Wang, M. Liu, Y. Chen and X. Wang, 'Mission-critical resource allocation with puncturing in industrial wireless networks under mixed services,' *IEEE Access*, vol. 9, pp. 21 870–21 880, 2021. DOI: 10.1109/ACCESS.2021.3056202.

[145] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi and A. Ghrayeb, 'Joint resource allocation and phase shift optimization for ris-aided embb/urllc traffic multiplexing,' *IEEE Transactions on Communications*, vol. 70, no. 2, pp. 1304–1319, 2022. DOI: 10.1109/TCOMM.2021.3127265.

[146] G. Interdonato, S. Buzzi, C. D'Andrea, L. Venturino, C. D'Elia and P. Vendittelli, *On the coexistence of embb and urllc in multi-cell massive mimo*, 2023. DOI: 10.48550/ARXIV.2301.03575. [Online]. Available: https://arxiv.org/abs/2301.03575.

[147] Qualcomm, *Everything you need to know about 5g*. [Online]. Available: https://www.qualcomm.com/5g/what-is-5g.

[148] A. Esmaeily, K. Kralevska and T. Mahmoodi, 'Slicing scheduling for supporting critical traffic in beyond 5g,' in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, IEEE, 2022, pp. 637–643.

[149] A. Esmaeily, H. K. Mendis, T. Mahmoodi and K. Kralevska, 'Beyond 5g resource slicing with mixed-numerologies for mission critical urllc and embb coexistence,' *IEEE Open Journal of the Communications Society*, vol. 4, pp. 727–747, 2023.

[150] S. Haga, A. Esmaeily, K. Kralevska and D. Gligoroski, '5g network slice isolation with wireguard and open source mano: A vpnaas proof-of-concept,' in *2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2020, pp. 181–187. DOI: 10.1109/NFV-SDN50289.2020.9289900.

[151] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li and J. Rao, 'Ai-assisted network-slicing based next-generation wireless networks,' *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 45–66, 2020. DOI: 10.1109/OJVT.2020.2965100.

[152] N. Salhab, R. Langar and R. Rahim, '5g network slices resource orchestration using machine learning techniques,' *Computer Networks*, vol. 188, p. 107 829, 2021, ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2021.107829. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128621000165.

Part II

Included Papers

Paper I

A. Esmaeily and K. Kralevska, “Small-scale 5g testbeds for network slicing deployment: A systematic review,” Wireless Communications and Mobile Computing, vol. 2021, pp.1-26.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6655216, 26 pages
https://doi.org/10.1155/2021/6655216

Review Article

Small-Scale 5G Testbeds for Network Slicing Deployment: A Systematic Review

Ali Esmaeily and Katina Kralevska

Department of Information Security and Communication Technology, NTNU-Norwegian University of Science and Technology, Trondheim 7491, Norway

Correspondence should be addressed to Ali Esmaeily; ali.esmaeily@ntnu.no

Received 4 November 2020; Revised 31 March 2021; Accepted 15 April 2021; Published 11 May 2021

Academic Editor: Luis Castedo

Developing specialized cloud-based and open-source testbeds is a practical approach to investigate network slicing functionalities in the fifth-generation (5G) mobile networks. This paper provides a comprehensive review of most of the existing cost-efficient and small-scale testbeds that partially or fully deploy network slicing. First, we present relevant software packages for the three main functional blocks of the ETSI NFV MANO framework and for emulating the access and core network domains. Second, we define primary and secondary design criteria for deploying network slicing testbeds. These design criteria are later used for comparison between the testbeds. Third, we present the state-of-the-art testbeds, including their design objectives, key technologies, network slicing deployment, and experiments. Next, we evaluate the testbeds according to the defined design criteria and present an in-depth summary table. This assessment concludes with the superiority of some of them over the rest and the most dominant software packages satisfying the ETSI NFV MANO framework. Finally, challenges, potential solutions, and future works of network slicing testbeds are discussed.

1. Introduction

The fifth-generation (5G) and beyond networks are expected to provide various services compared to the 4G and previous generations of networks. The Quality of Service (QoS) requirements can be quite different in terms of low latency (or even extralow latency), bandwidth, reliability, and availability. Remote surgery, autonomous driving, a massive number of sensors communicating with the network, and video streaming with extrahigh quality are just some of the numerous 5G services. The main concern here is that the physical infrastructure resources are limited and valuable, especially when data traffic demands from different operators increase. Therefore, efficient network sharing [1, 2] is considered as a conventional solution. Through network sharing, multiple operators can share infrastructure resources according to their agreed allocation plans. This approach can help an operator to reduce a significant amount of Capital Expenditure (CAPEX) and Operational Expenditure (OPEX).

As an evolution of network sharing, network slicing brings the flexibility and dynamicity of allocating the required and appropriate amount of physical resources to all service types mentioned above over the same physical infrastructure simultaneously. In fact, network slicing leverages the running of multiple logical networks on top of physical infrastructure. Network Functions (NFs) [3] are constructive operational components (physical networking devices) such as routers, firewalls, and load balancers that have specific functionalities in network infrastructure and hold distinct exterior interfaces for establishing communication between each other. An End-to-End (E2E) network slice [4] is a logical separated (isolated) network, created by chaining NFs, which delivers a particular network service according to QoS requirements via the underlying shared infrastructure in the (Radio) Access Network ((R)AN), Transport Network (TN), and Core Network (CN).

Network Function Virtualization (NFV), Software Defined Networking (SDN), and Cloud computing are considered as

the three enabling technologies for implementing network slicing in 5G.

(i) *NFV* [3] is a network architecture framework where NFs that traditionally used dedicated vendor-specific hardware, so-called Physical NFs (PNFs), are now implemented in software. There are two leading solutions towards softwarized PNFs: (1) Virtualized NFs (VNFs) deployed on virtual machines and (2) Containerized NFs (CNFs) deployed on containers. These VNFs and CNFs, in turn, are then implemented in data centers or on cloud environments that run on top of general-purpose (vendor-neutral) hardware.

(ii) *SDN* [5] enables programmable and dynamic network configuration by separating the Control Plane (CP) and the Data Plane (DP), where a centralized entity (controller) in the CP configures the forwarding devices in the DP.

(iii) *Cloud computing* [6] deploys remote network resources in shared pools that can be administered over the Internet. Cloud computing is based on two principal orientations: (1) Cloud-based applications that point to relocating legacy applications, which were established on end-users' devices or on the companies' IT infrastructure, to cloud-based servers in order to deliver the applications over web browsers, and (2) Cloud-native applications, which refer to those applications that are particularly created and developed to employ the advantages of the cloud environment such as constant development, modularity, Application Programming Interface (API) integration, and scalability.

As mentioned, one of the 5G objectives is to implement ultralow latency services and to serve many devices with different amounts of computing resources. Multi-access Edge Computing (MEC) [7] is an enhancement of cloud computing that reduces the latency in a mobile network by pushing the processing and computing tasks to the edges of the network (such as base stations) to be closer to the devices with a limited amount of resources. This yields in facilitating the operation of delay-sensitive applications in such devices. These enabling technologies bring flexibility, programmability, and efficiency, but at the cost of higher complexity in operating and managing the 5G networks. The necessity for the Management and Network Orchestration (MANO) framework [8], which performs efficient resource management and orchestration between all network elements in the whole architecture, is undeniable.

Figure 1 illustrates a multi-layered architecture of network slice provisioning in 5G.

(i) In the first layer, there is a shared infrastructure layer, which includes heterogeneous hardware and software resources (base station, compute, storage, and networking) spanning over the RAN, TN, and CN domains to host multiple NFs in the second layer. In fact, these resources are sliced according to various service requirements and then will be allocated to different service types.

(ii) In the second layer, there are various NFs (PNFs, VNFs, and CNFs) with certain capabilities, belonging to different network domains. This layer encapsulates the essential configuration and managing operations of the NFs to provide different service types in the third layer.

(iii) In the third layer, according to service specifications, particular PNFs, VNFs, and/or CNFs (from the second layer) are chained in an explicit order with the appropriate amount of resources (from the first layer) to grant a distinctive service instance. The uniqueness of a service instance in this layer has a straightforward association with the business model, which indicates the reason for creating such a service that will be presented via a slice.

(iv) In the fourth layer, the launched service instances from the previous layer constitute E2E network slices. Hence, controlling and management policies on each of the network slices can be achieved independently via the NFV MANO framework.

(v) The NFV MANO framework is in charge of orchestrating all of the mentioned layers. Basically, the NFV MANO delivers all the monitoring, coordinating, controlling, and managing tasks of the available physical and virtual resources in order to maintain an efficient resource utilization between all types of NFs (PNF, VNF, and CNF) in the whole architecture. This results in producing network services that meet the specific service requirements over distinctive network slices.

Since the introduction of the network slicing concepts and specification by the 3rd Generation Partnership Project (3GPP) [9], network slicing has attracted a lot of attention in the past years. Apart from the theoretical aspects of different ways of achieving the 5G objectives, research communities in academia and industry have followed practical approaches to examine different features of 5G and to evaluate the network performance under various use cases. In this regard, practical research works in the 5G area have developed prototype system implementations of individual parts of the mobile network architecture, which are known as research testbeds. Recently, even more complex network architectures have been deployed on such testbeds that support network slicing. Research testbeds grant the possibility to evaluate and enhance network performance. Besides, while research testbeds keep the cost of network deployment low, their functionalities, with a fair approximation, are comparable to real networks. Such testbeds can usually be implemented on standard PCs or servers with a not very high amount of resources and without the need of purchasing specialized hardware and software. Moreover, the availability of open-source software packages provides opportunities for creating innovative solutions towards 5G [10].

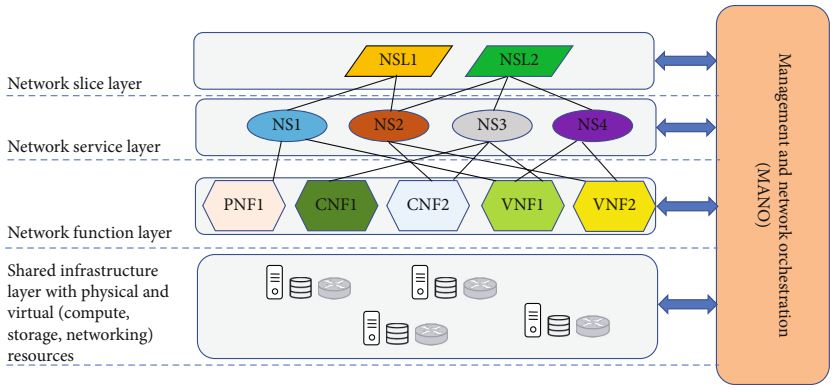


Figure 1: Multi-layered architecture of network slice provisioning in 5G via the composition of VNFs, CNF, and/or PNFs into network services to form network slices.

Deploying testbeds with network slicing capabilities is a challenging and error-prone task as it involves development of a network equipped with fundamental enabling technologies and the ability of programming and configuring the physical infrastructure. Depending on the specific service requirements, the physical and virtual components of a network slicing testbed must satisfy performance requests such as the amount of hardware and software resources (CPU, memory), reliability, and failure rates (dependability analysis) [11]. Nevertheless, the complexity of the testbed deployment process sometimes impacts the utilization of open-source solutions and standard PCs.

Although some excellent surveys have been done on different aspects of network slicing such as [4, 12, 13], just a few works focus on network slicing implementations, in particular, [14–17] elaborate collaborative 5G network slicing research projects and the proposed large-scale testbeds as outcomes of these projects. Reference [14] presents a broad study of five main large-scale SDN testbeds by explaining their design purposes, essential technologies, slicing capability, and use cases. Reference [15] investigates the necessity of network slicing for facilitating the implementation of Internet-of-Things (IoT) intelligent applications and smart services. Bonati et al. in [16] describe open source utilities, frameworks, and hardware components that can be used to instantiate softwarized 5G networks. Barakabitze et al. [17] provide a comprehensive review of 5G networks, a tutorial of the 5G network slicing technology enablers including SDN, NFV, MEC, Cloud/Fog computing, network hypervisors, virtual machines, and containers, as well as an overview of collaborative large 5G network slicing implementations. Nonetheless, there is a lack of a comprehensive survey that presents and evaluates small-scale state-of-the-art 5G network slicing implementations. Small-scale network slicing testbeds are important for the research community in several aspects. Small-scale testbeds require a lower deployment budget compared to large-scale testbeds. Besides, small-scale testbeds, with a compact softwarized version of the required entities, are more effortless to deploy and launch than large-scale ones. Further, due to such testbeds' small scaling, they are more manageable to troubleshoot, and resolving possible issues is faster than large-scale testbeds with multiple involved entities. Eventually, although the number the practical use cases that can be investigated on small-scale testbeds is lower than large-scale testbeds and real networks, small-scale testbeds can afford similar analogous results to large-scale solutions. The aforementioned aspects motivate the work in this paper. We summarize our contributions as follows:

(i) We present the software packages and platforms that fit in the ETSI NFV MANO framework functional blocks for emulating RAN, CN domains, and MANO.

(ii) We define primary and secondary design criteria for network slicing testbeds.

(iii) We provide a detailed study of small-scale state-of-the-art testbeds for deploying network slicing. These testbeds are relatively easy to deploy and usually without requiring a huge financial investment, thus, suitable for university labs.

(iv) We further evaluate the testbeds according to the defined primary and secondary design criteria.

(v) We highlight the typical challenges while deploying such testbeds, and present possible solutions and directions for future work.

The rest of the paper is organized as follows. Section 2 explains the research methodology for this paper. Section 3, firstly, presents the ETSI NFV MANO framework along with possible open-source software solutions for each specific block in this framework, and secondly, outlines the desired criteria for designing network slicing testbeds in 5G. In Section 4, small-scale and cost-efficient state-of-the-art network slicing testbeds are detailed with their specific features. In Section 5, first, we compare the testbeds presented in the previous section, and then, we explain some of the main challenges while deploying such testbeds. Section 6 concludes the paper. Table 1 presents a list of the acronyms used in this paper.

Table 1: List of the used acronyms in this paper.

Abb.	Definition	Abb.	Definition	Abb.	Definition
5G	Fifth generation	4G	Forth generation	3GPP	3rd Generation Partnership Project
AI	Artificial Intelligence	AMF	Access and Mobility Management function	API	Application Programming Interface
BBU	Baseband Unit	CAI	Connected AI	CN	Core Network
CNF	Containerized NF	CP	Control Plane	CAPEX	Capital Expenditure
C-RAN	Cloud-RAN	DC	Data Center	DCAE	Data Collection Analytics & Events
DP	Data Plane	DSAF	Dynamic Slice Allocation Framework	E2E	End-to-End
eMBB	enhanced Mobile Broadband	EPC	Evolved Packet Core	ETSI	European Telecommunications Standards Institute
FCFSFA	First Come First Serve First Available	GUI	Graphical UI	HP LCVNF	High Priority LCVNF
IaaS	Infrastructure-as-a-Service	IIoT	Industrial IoT	IMS	IP Multimedia System
IoT	Internet-of-Things	KPI	Key Performance Indicator	KQI	Key Quality Indicators
L2TP	Layer-2 Tunneling Protocol	LCVNF	Latency Critical VNF	LP LCVNF	Low Piority LCVNF
LTE	Long-Term Evolution	LT VNF	Latency Tolerant VNF	MAC	Medium Access Control
MANO	Management and Network Orchestration	M-CORD	Mobile-Central Office Rearchitected as Datacenter	MEC	Multiaccess Edge Computing
ML	Machine Learning	MME	Mobility Management Entity	MTC	Machine Type Communication
NAS	Network Attached Storage	NBI	Northbound Interface	NF	Network Function
NR	New Radio	NFV	Network Function Virtualization	NFVI	NFV Infrastructure
NFVO	NFV Orchestrator	NIM	Network Infrastructure Manager	NSO	Network Service Orchestrator
OAI	OpenAirInterface	ODL	OpenDayLight	ODTN	Open and Disaggregated Transport Network
OMEC	Open Mobile Evolved Core	ONAP	Open Networking Automation Platform	ONOS	Open Network Operating System
OPEX	Operational Expenditure	OSM	Open Source MANO	OTG	OAI Traffic Generator
OvS	Open virtualSwitch	PaaS	Platform-as-a-Service	PNF	Physical NF
QoE	Quality of Experience	QoS	Quality of Service	RAN	Radio Access Network
RAT	Radio Access Technology	RLC	Radio Link Control	RO	Resource Orchestrator
RRC	Radio Resource Control	RRH	Remote Radio Head	RRM	Radio Resource Management
SA	Service Assurance	SaaS	Software-as-a-Service	SBI	Southbound Interface
SDN	Software Defined Networking	SD-RAN	Software Defined RAN	SEMIoTICS	Smart End-to-end Massive IoT Interoperability, Connectivity, and Security
SLA	Service Level Agreement	SlaaS	Slice-as-a-Service	SRS LTE	Software Radio Systems LTE
TN	Transport Network	UE	User Equipment	UI	User Interface
VDU	Virtual Deployment Unit	VES	Virtual Event Streaming	VIM	Virtualized Infrastructure Manager
VNF	Virtualized NF	VNFFG	VNF Forwarding Graph	VNFM	VNF Manager

2. Research Methodology

Network slicing has become a very hot topic both in academia and industry. This trend has resulted in research on various aspects of network slicing in 5G and a fast-growing number of publications. It is evident that only a portion of these publications introduces implementation solutions for network slicing, i.e., network slicing testbeds. In order to review such publications, we followed a research methodology and defined the procedure to search for related publications, the inclusion and exclusion criteria, and finally, the data collection method to extract pertinent publications. Inclusion and exclusion criteria are used to filter out nonrelevant collected papers. There is also an extra step for quality assessment regarding those publications that pass the inclusion criteria in the final systematization.

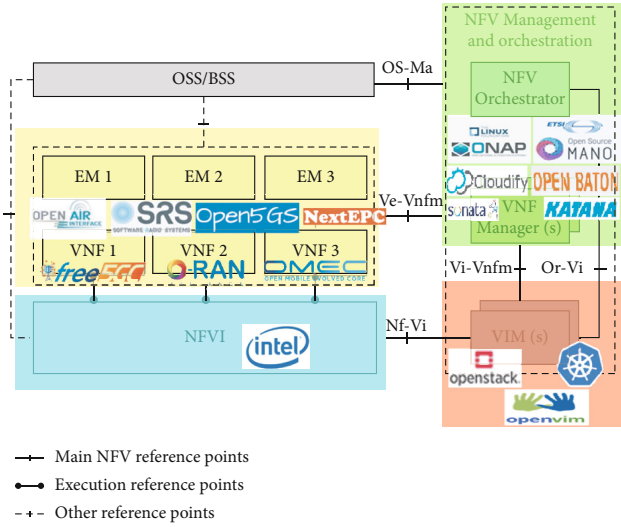


Figure 2: Different open-source software solutions mapped to the ETSI NFV MANO framework [8].

In the first step, we identified the databases for searching for potential relevant publications such as (1) ACM Digital Library, (2) IEEE Xplore, (3) Springer Link, (4) ScienceDirect, and (5) arXiv. Next, we started our searching process with relevant keywords to narrow down the selection area of the scientific publications into the network slicing field and, in particular, the deployment of network slicing. We employed some keywords such as <*5G testbed*>, <*network slicing testbed*>, <*network slicing platform*>, and < *network slicing framework*>.

In the second step, we defined the inclusion criteria, for the publications resulted from the first step, as follows:

(i) Does the publication present a solution for network slicing deployment?

(ii) How is the solution provided? Which software and hardware components are used?

(iii) Is the presented testbed cost-efficient in terms of equipment and also human resources needed for the tested deployment?

We also defined the exclusion criteria as:

(i) A publication that introduces a large-scale testbed for network slicing, which is not possible to be implemented with a small budget.

(ii) A testbed, which is a result of national or international research projects, and those projects have been finished or are no longer active.

In the third step, the publications that meet the inclusion criteria are assessed for their quality. Following questions are applied for quality assessment:

(i) Can the presented testbed be used to investigate different typical use cases in the 5G network slicing, or is the solution just an initial implementation of network slicing with limited capacity for providing few realistic scenarios?

(ii) Does the publication include comprehensive information for the testbed architecture and deployment? Are there any extra and complementary sources included in the publication, that could help other researchers to deploy a similar testbed or a possible future extension?

In the end, we categorize the testbeds following the primary and secondary criteria defined in Section 3.

3. ETSI NFV MANO Framework and Design Criteria for Network Slicing Testbeds

3.1. ETSI NFV MANO Framework and Different Open-Source Software Solutions. ETSI introduces the NFV MANO architecture [8], which is comprised of three main functional blocks. Figure 2 illustrates these blocks with the reference points that connect them. This figure also summarizes some of the preeminent software solutions for each specific block. We focus on combining these solutions into the presented testbeds in Section 4 instead of explaining each one of these software modules individually.

(i) *Virtualized Infrastructure Manager (VIM)* performs controlling mechanisms for the NFV Infrastructure (NFVI) resources within an infrastructure provider. VIM is also responsible for receiving fault measurement and performance information

of NFVI resources. Consequently, VIM can supervise NFVI resources allocation to the available VNFs. OpenStack [18] and OpenVIM [19] (for VNFs) and Kubernetes [20] (for CNFs) are possible solutions for the VIM section.

(ii) *VNF Manager (VNFM)* conducts one or several VNFs and does the lifecycle management of VNFs. VNF lifecycle management involves establishing/-configuring, preserving, and terminating VNFs.

(iii) *NFV Orchestrator (NFVO)* implements resource and service orchestration in the network. NFVO is split up into Resource Orchestrator (RO) and Network Service Orchestrator (NSO). First, RO collects the current information regarding possible physical and virtual resources of NFVI through the VIM. Second, NSO applies a complete lifecycle management of multiple network services. In this way, NFVO keeps updating the information about the available VNFs running on top of NFVI. As a result, NFVO can initiate multiple network services. As part of the lifecycle management, NFVO can also terminate a network service whenever no longer a service request is received for that specific service. In several solutions, NFVO and VNFM are integrated into the MANO section. Open Source MANO (OSM) [21], Open Networking Automation Platform (ONAP) [22], OpenBaton [23], Cloudify [24], SONATA [25], and Katana Slice Manager [26] are considered as the leading integrated solutions for the MANO section. Note that OSM can also perform management and orchestration tasks on PNFs.

Regarding VNFs/CNFs, several open-source software solutions can emulate RAN and CN domains:

(i) *RAN* domain is emulated with Software Radio Systems LTE (srsLTE) [27], OpenAirInterface (OAI) [10, 28], or O-RAN in its *Bronze* release [29, 30]

(ii) *CN* domain is realized with OAI, Open5GS (previously known as NextEPC) [31], Open Mobile Evolved Core (OMEC) [32], or free5GC [33]

Then, via chaining these VNFs/CNFs in the RAN and CN by the NFVO, distinguished service instances, so-called network slice subinstances, are formed. Some solutions for the TN domain, such as Open and Disaggregated Transport Network (ODTN) [34], utilize disaggregated optical equipment and open-source software to create a TN slice subinstance. An E2E network slice instance is created by pairing the definite RAN and CN slice subinstances via the TN slice subinstance [35].

3.2. Design Criteria for Network Slicing Testbeds. Multiple features should be taken into consideration when designing a comprehensive testbed of 5G and beyond networks. We identify the key design criteria for creating a 5G testbed that can emulate a real network's major features and allow us to develop and test new algorithms. They are divided into two groups.

3.2.1. Primary Criteria. These attributes are fundamental for creating a network slicing testbed.

(i) *Support of the main enabling technologies.* The proposed testbed should be based on SDN, NFV, and cloud computing. Therefore, flexibility and dynamicity in the network are granted. SDN and NFV are complementary, hence, combined with cloud computing pave the way for the paradigms Software-as-a-Service (SaaS), Platform-as-a-Service (Paas), and Infrastructure-as-a-Service (IaaS) [3].

(ii) *MANO equipped with dynamic monitoring capability.* The testbed should support management, orchestration, programmability, and dynamic monitoring of different network functions, network services, and network slices. Therefore, the role of the MANO entity is essential that is the result of SDN/NFV utilization in the network architecture [36].

(iii) *Multi-network domain with partial slicing support.* A 5G testbed needs to provide connectivity across all network domains (air interface, (R)AN, TN, and CN) in order to show a practical ability that emulates the main functionalities of the 5G network. Multi-network domain support allows achieving E2E network slicing; however, it is worth noting that network slicing is a capability that can be implemented partially, and testbeds can deploy slicing only in one specific network domain.

(iv) *Multi-tenancy support.* 5G network is expected first to enable the coexistence of multiple tenants that demand the same network functionalities, and second to administrate the cooperation and interaction between them. This capability represents the so-called multi-tenancy environment, which means that a single instance of the software and its supporting infrastructure serves multiple tenants. Multi-tenancy is one of the main aspects of the 5G networks and should be supported in the testbed implementation.

3.2.2. Secondary Criteria. These attributes add extra features to a network slicing testbed apart from those in the *primary* group. Testbeds with these extra features broaden the research scope in the network slicing field.

(i) *Multi-radio access technologies support.* Different Radio Access Technologies (RATs) such as Long-Term Evolution (LTE), WiFi, and 5G New Radio (5G NR) should be deployed on the same platform [37]. Furthermore, Cloud-RAN (C-RAN), as a cloud computing-based architecture, brings cloudification benefits into the RAN domain. C-RAN consists of a cloud-Baseband Unit (BBU) pool and several Remote Radio Heads (RRHs). Since the 5G-RAN domain integrates the mentioned RATs with the corresponding frequency bands and provides them

via the cloud, a solid platform should implement these capabilities.

(ii) *End-to-End network slicing.* The slicing capability should be expanded upon all network domains. An E2E network slice consists of several network slice subnet instances, each belonging to a particular network domain. Therefore, all network slice subnet instances should be provided and chained together to form an E2E network slice.

(iii) *Cross-location support.* One possible solution for experimenting with more realistic scenarios is deploying testbeds located in two geographical areas. In this case, RAN and CN domains are implemented and launched on two geographically separated infrastructures, and a backbone TN interconnects them. The cross-location capability becomes even more essential when evaluating network performance for providing delay-sensitive services in the 5G network. In real-world use cases, the RAN and CN domains are not necessarily located in the same geographical location, and, as mentioned, MEC is the technology answer to expedite the communication between the RAN and CN domains. Hence, cross-location testbeds facilitate measuring service delay and proposing possible solutions for those services that require low latency.

(iv) *Machine Learning(ML)-enabled.* 5G testbeds equipped with ML toolkits enable users to design, verify, and operate machine learning models via a supervised user interface. One possible outcome of using ML techniques in network slicing is to predict wireless channel behavior in the RAN domain. As a result, the available radio resources can be scheduled in an optimized way to maximize the resource usage per end-user or slice in the next transmissions.

(v) *Open-source.* Providing open-source 5G platforms with well-defined interfaces is considered as a huge advantage in deploying 5G testbeds because an open-source testbed can be deployed by other researchers to help foster research and innovation. It helps to reduce the hassle of setting up a working mobile network that on itself is a complicated and error-prone process

These design criteria explained above and outlined in Figure 3 are later used as an assessment for the state-of-the-art testbeds.

4. An Overview of the State-of-the-Art Network Slicing Testbeds

We describe most of the state-of-the-art testbeds designed for implementing network slicing. We exclude large-scale testbeds that are costly in terms of equipment and human resources as some of them are already explained in [16, 17].

4.1. 5G4IoT [38, 39]. This testbed (Figure 4) deploys OAI in containers to virtualize both Evolved Packet Core (EPC)

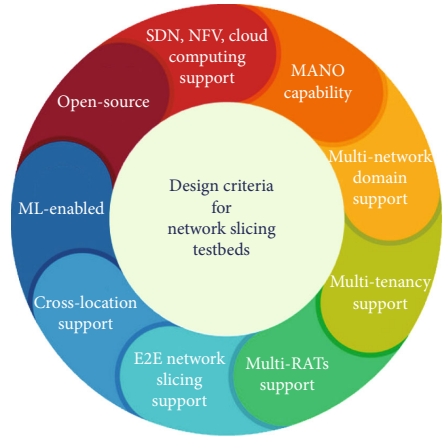


Figure 3: Design criteria for network slicing testbeds.

and eNB. For scalability purposes, the testbed has been implemented in several containers. The testbed is created on a cloud infrastructure based on OpenStack, which is located at Oslo Metropolitan University. There are also a cisco switch and a cisco router in this testbed, separated into two VLANs, which establish the connection between EPC and eNB in two ways. The first method uses SDN Calico (https://projectcalico.org) for layer-3 packet exchange, which provides scalability and dynamic security on the cloud infrastructure for layer-3 routing for IoT. The second approach is Layer-2 Tunneling Protocol (L2TP) to encapsulate the traffic for those IoT applications that need a lower security level. The latter approach is granted by Open virtualSwitch (OvS) without using IPSec. In the testbed, OpenStack Heat templates, as an underlay networking policy, are used to integrate OAI EPC in the OpenStack Neutron. These templates manage cloud-based applications in a stack of containers, and various services via network slices can be created. The testbed is evaluated by producing two isolated network slices for eHealth and light Internet on the same infrastructure.

4.2. 5G Test Network (5GTN) [40]. In this testbed (Figure 5), located at Oulu University, the RAN operates on licensed LTE and 5G bands. The CN comprises EPC and IP Multimedia System (IMS). The CN components are implemented on cloud infrastructure, OpenStack and VMWare. The testbed is aimed at serving different use cases; thus, it includes MEC in the edge to provide low latency services. However, there is no RAN slicing, so only CN slicing is currently implemented. The testbed includes multiple CN domains, which result in sharing radio resources for different services. In this case, each base station in the RAN utilizes a single gateway to access a slice. The authors showcased two slices, slice A and slice B, provided in the CN domain. Slice A from the EPC domain (deployed on OpenStack and orchestrated by CloudBand which is the Nokia platform for NFV orchestration) provides enhanced Mobile Broadband (eMBB) services for IoT and content delivery applications, and slice B in the IMS domain (deployed on VMWare and orchestrated by

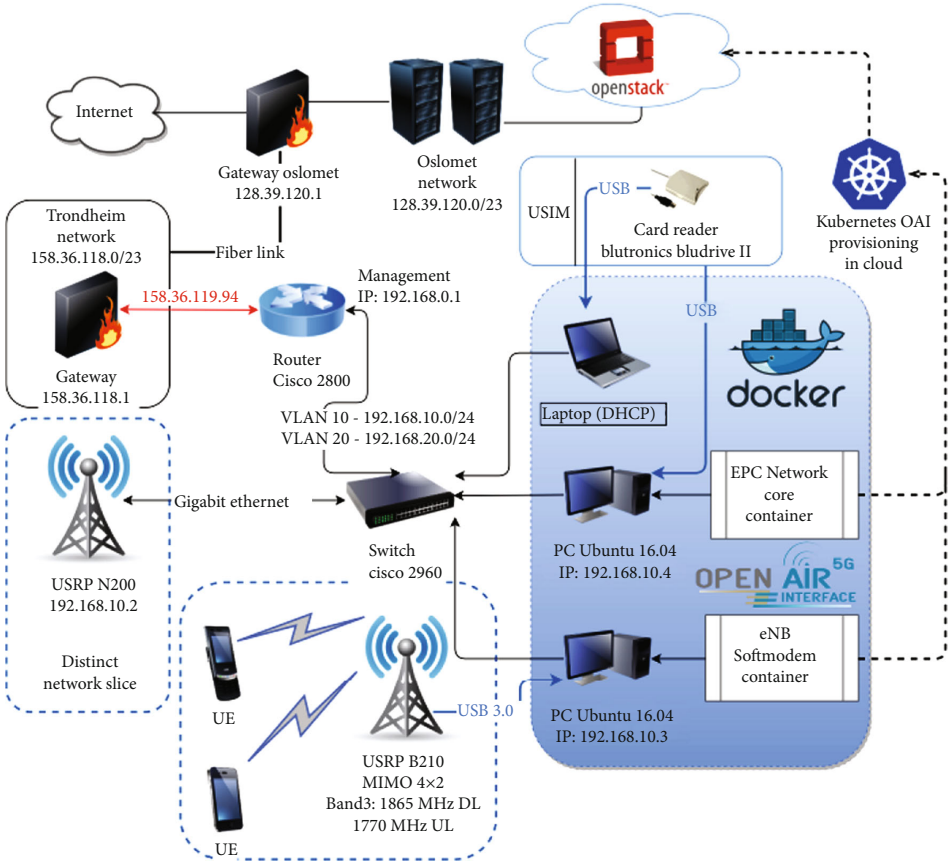


Figure 4: 5G4IoT testbed architecture [38].

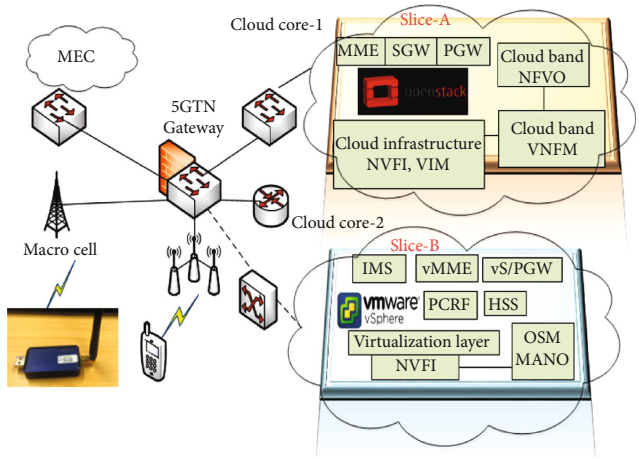


Figure 5: 5GTN testbed architecture [40].

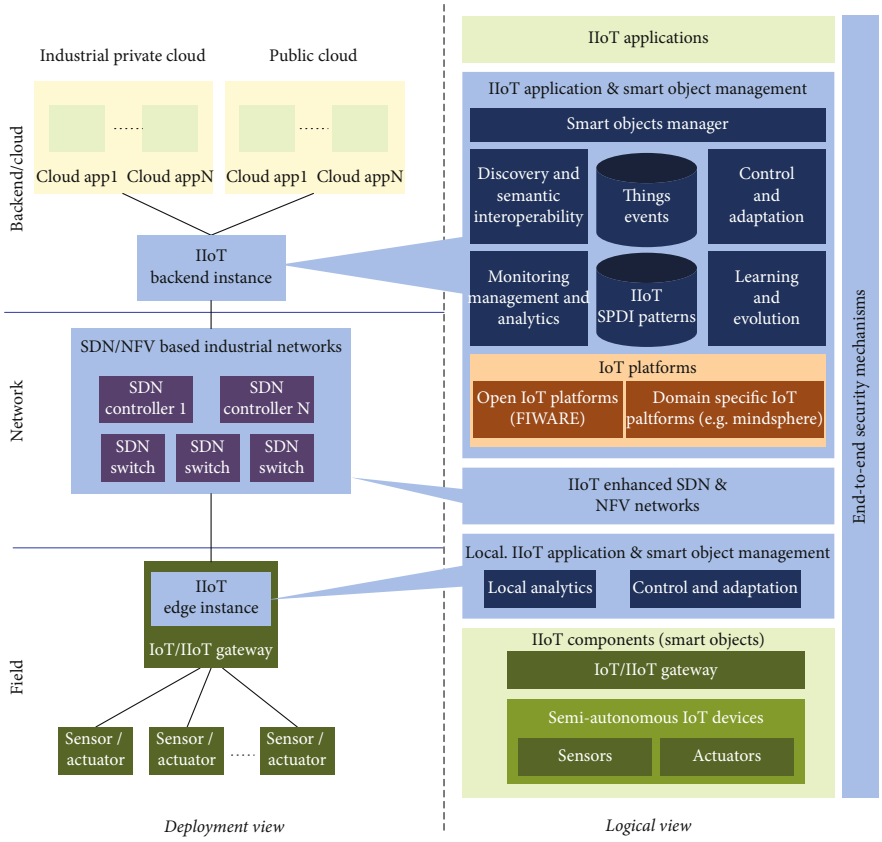


Figure 6: Architecture of SEMIoTICS platform [42].

OSM) provides critical communication and Voice over LTE services. By changing the Access Point Name between EPC (deployed on OpenStack) and IMS (deployed on VMWare), User Equipment (UE) switching between the two slices is possible. The testbed has been examined for CPU utilization, throughput, and delay for the two specific slices.

4.3. 5G Tactile Internet Platform [41]. This testbed (Figure 6) follows the Smart End-to-end Massive IoT Interoperability, Connectivity, and Security (SEMIoTICS) [42] architecture to create a 5G platform based on SDN, NFV, and MEC. The principal objective of SEMIoTICS is to build a framework to provide secure and reliable E2E services with submillisecond latency in actuation operations for IoT/Industrial IoT (IIoT) applications. The SEMIoTICS architecture consists of 3 layers: Backend/Cloud, Networking, and Field layers. The Backend layer is a cloud-based OpenStack platform. It creates several VMs and performs their lifecycle management. The services are provided in several containers and managed by OpenStack Tacker. Currently, there are two deployed VNFs: one for smart monitoring and one for actuating. The Networking layer manages the virtual domain on the testbed and creates tenants to chain VNFs by utilizing

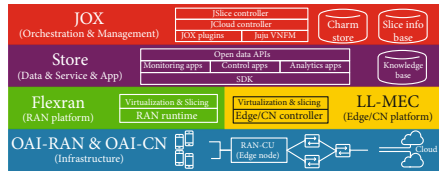


Figure 7: Mosaic5G platform schematic architecture [43].

the SDN controller Neutron. The communication between separated tenants and also with external networks takes place by performing layer 3 routing. The Field layer is responsible for establishing a connection between smart sensors and actuators with the upper layers. This process is done by exchanging messages through IoT/IIoT gateways in the Field layer and virtual SDN switches in the Networking layer. The testbed performance has been assessed for performing E2E slicing and dynamically sharing the available bandwidth between the two VNFs.

4.4. Mosaic5G [43]. Mosaic5G platform (Figure 7) brings flexibility and scalability to service provision. The testbed architecture consists of five software modules along with

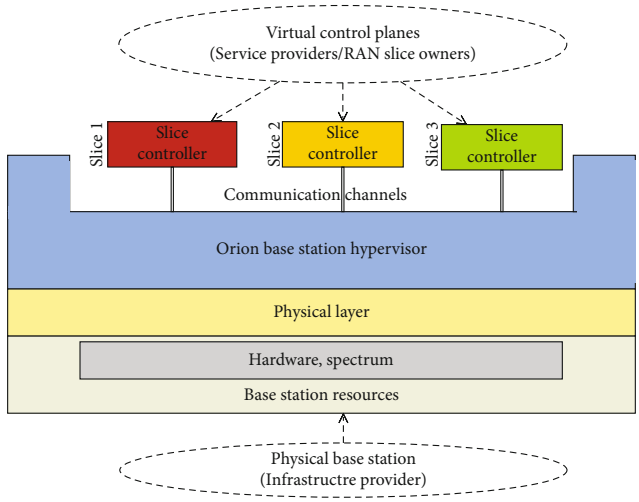


Figure 8: High level of Orion testbed architecture [45].

hardware components: OAI, FlexRAN, LL-MEC, Store, and JOX. OAI emulates both the RAN and the CN domains of LTE networks. FlexRAN [44] is an open-source Software Defined RAN (SD-RAN) entity. FlexRAN delivers one of these two tasks, deployment of controlling mechanisms for several base stations in a centralized way or performing distributed controlling policies. These two actions are done as reconfigurable control functionalities in the RAN domain. LL-MEC separates the control plane and data plane traffic at the edge and the CN domain. In this way, the MEC functionality is achieved. Basically, FlexRAN and LL-MEC perform SDN functionality in the RAN, and in the edge and core domains, respectively. Store includes a set of modules, monitoring, and control applications for developing network applications for a specific use case. JOX plays the role of orchestration in the network to provide several E2E network slices according to NFV MANO platform. Therefore, network slices can be deployed and then optimized based on various service specifications. The Mosaic5G platform has been used for a few use cases such as critical e-Health, V2X communication for intelligent transportation systems, and multi-service management/orchestration for smart cities.

4.5. Orion [45, 46]. The proposed architecture of the Orion testbed (Figure 8) enables dynamic RAN slicing. Orion provides the sharing of RAN resources in addition to applying isolation between slices, and so, operation in one slice cannot degrade the performance of another slice. This is achieved by having an independent control plane in the RAN domain for each slice. As a result, Orion offers the opportunity to deploy different service characteristics in the RAN domain, and it is a concrete step towards realizing RAN-as-a-Service. The testbed consists of two main components: Base Station Hypervisor and virtual control plane. The Base Station Hypervisor performs slice isolation in the RAN domain, while the Hypervisor capability prepares an abstraction layer of available radio resources to the slices in the RAN. In this way, service providers build virtual base stations on top of the Hypervisor in order to create their RAN slices. A separated virtual control plane for each slice interconnects to the Hypervisor to exchange the required signaling messages. This independent deployment enables slice isolation in the RAN domain. Furthermore, the Orion architecture enables a virtual control plane of a slice to connect to multiple base stations via their Hypervisor layer. Several case studies regarding Orion's performance evaluation have been done, such as testing slice isolation and possibilities for E2E- and multi-service provisions.

4.6. 5G Testbed for Network Slicing Evaluation [47]. This testbed (Figure 9) utilizes OAI for both RAN and CN domains. There are two CNs which share radio resources of a single eNB in the RAN. OAI RAN consists of OAISIM that allows simulation of UE and eNB. OAISIM acts like a real RAN domain and simulates the LTE protocol stack. A UE with a Network Slice Selection Assistance Information (NSSAI) capability has been implemented in the testbed. Deploying two CNs in containers provides CN slice isolation. In both CNs, the Access and Mobility Management Function (AMF), which is one of the entities in 5G architecture, has been integrated with the Mobility Management Entity (MME) of the LTE platform. The eNB in the RAN selects a CN according to the NSSAI information, which is provided via the S1-AP interface between the eNB and each CN. The testbed has been appraised for connection establishment for both normal LTE UEs and UEs with an implemented NSSAI capability. In the case of normal LTE UE, it includes required encoding messages during the attach process to the network. In the case of the modified UEs, NSSAI is implemented as an optional field in them. The related CN decodes this NSSAI via the S1-AP interface provided by the eNB in the attach process to the network.

4.7. POSENS [48, 49]. POSENS platform (Figure 10) provides efficient resource utilization for creating independent and

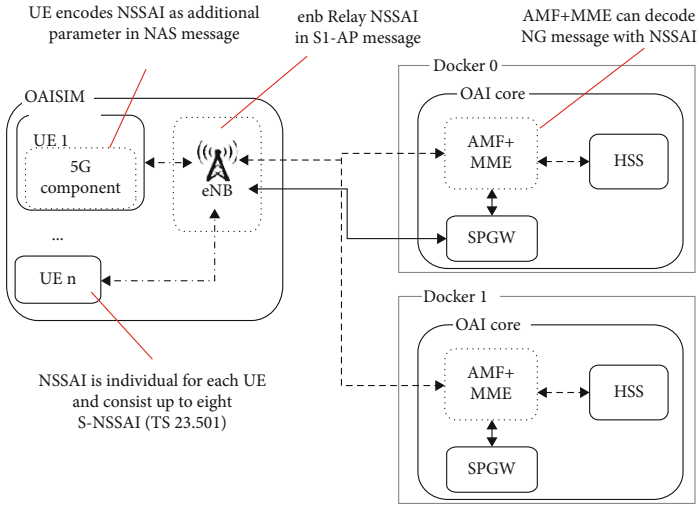


Figure 9: 5G testbed with network slicing support [47].

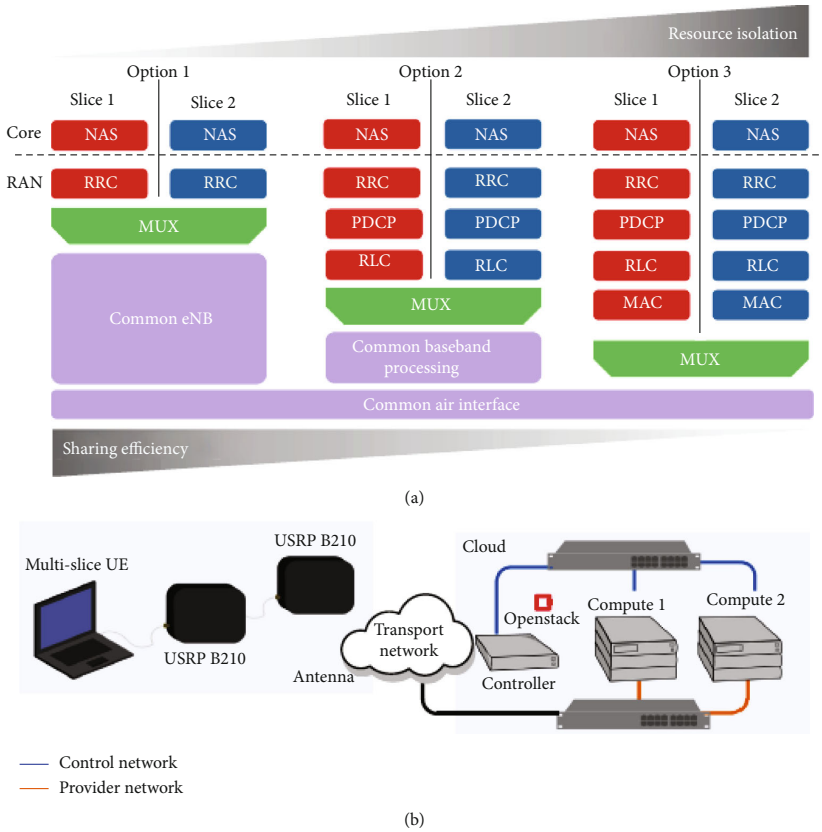


Figure 10: POSENS testbed. (a) RAN slicing options. (b) Architecture [48, 49].

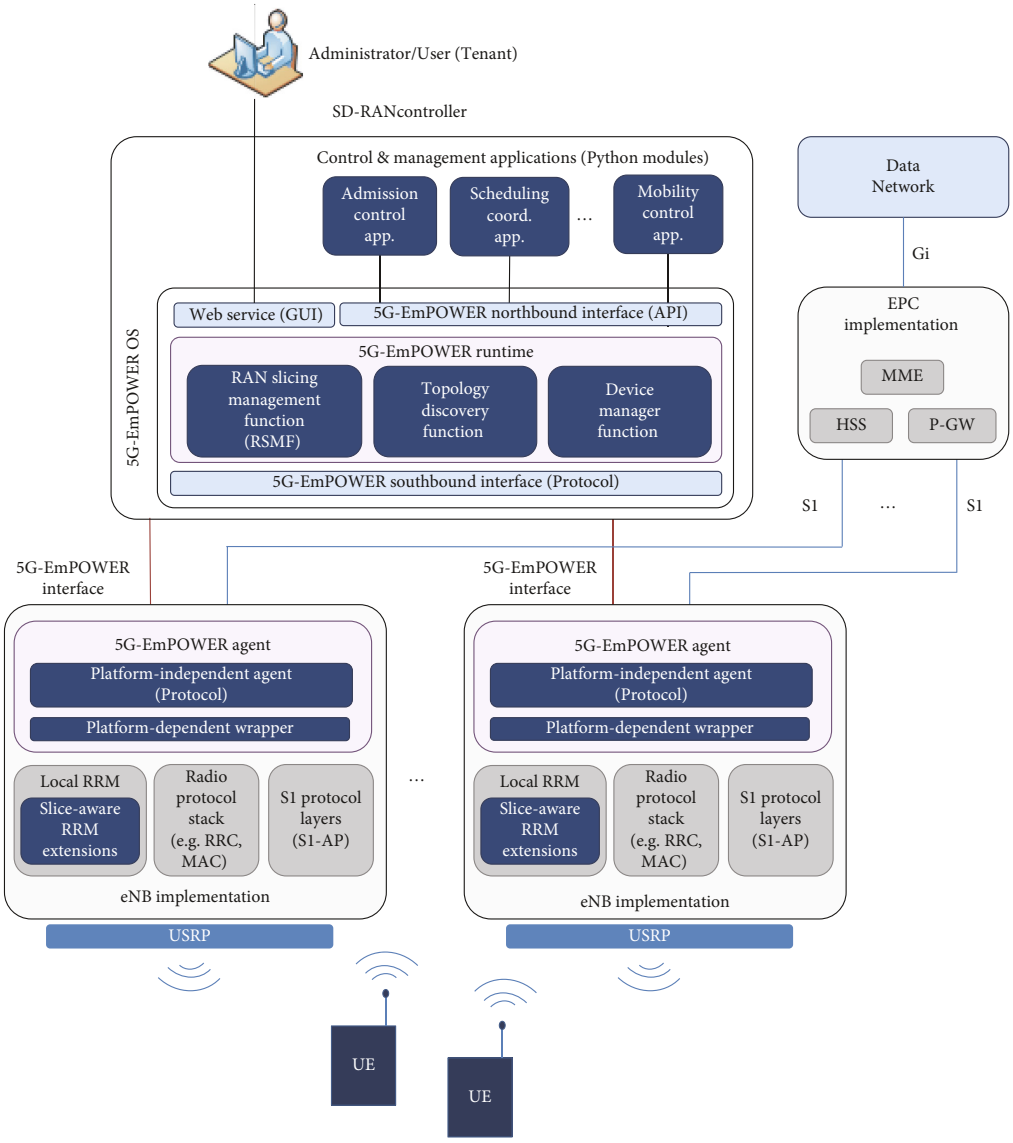


Figure 11: UPC testbed integrated with 5G-EmPOWER SD-RAN controller [50].

customizable E2E slices. Different NFs in the network layer are chained by MANO to create network slices. Then, the slices for different tenants need to be multiplexed/demultiplexed over shared resources. This procedure requires enabling the capability of multiple CN connections to a single RAN domain and the possibility for a UE to benefit from more than one slice simultaneously. In POSENS, the possibility of implementing RAN slicing is discussed via three options: (1) slice-aware shared RAN (slicing protocol stack down to Radio Resource Control (RRC)), where the whole radio domain is shared but CNs are distinguished by the specific services they provide. and a UE can utilize different slices provided by the CNs; (2) slice-specific radio bearer (slicing protocol stack down to Radio Link Control (RLC)), where only cell-specific functionality is shared; and (3) slice-specific RAN (slicing protocol stack down to Medium Access Control (MAC)), which apart from the air interface, slices of different tenants are isolated in other protocol stack layers. The latter case needs specific synchronization policies among slices to be deployed efficiently. Each option holds its own level of performance and implementation complexity and POSENS implements the first option for RAN slicing in its

first release. For CN, POSENS utilizes OAI CN with no modifications. In the case of MANO, the testbed provides per-slice management, and an orchestration mechanism deployed in customized version of OSM. The testbed has been evaluated in terms of independency between slices, throughput, and independent service function chaining.

4.8. UPC University Testbed [50]. UPC platform (Figure 11) implements RAN slicing via RESTful API automatically. The testbed applies the slice-aware policy in Radio Resource Management (RRM) functionalities for admission control and scheduling processes. 5G-EmPOWER [51], acting as a SD-RAN, allows RAN slicing management, and it also shares the available radio resources among the created RAN slices according to RRM descriptors. The interconnection between 5G-EmPOWER and eNB in the RAN is provided via an *EmPOWER Agent* for performing management policies in the data plane. The testbed utilizes OAI or Next EPC for the CN domain. The srsLTE emulates LTE eNB. The *EmPOWER Agent*, in turn, splits up into two sections: (1) *Agent*, which includes protocol parser for EmPOWER exchanged messages and manager entities for different message types. The message type is changed depending on the requested message originated from either the 5G-EMPOWER or the *Agent*, and activation/deactivation of a specific capability on the *Agent* side; (2) *Wrapper*, which converts EmPOWER messages to LTE protocol stack information. Several practical scenarios have been carried out for implementing RRM functionalities for admission control, scalability of the network, isolation among the existent slices.

4.9. Mobile-Central Office Rearchitected as Datacenter-(M-CORD-) Based 5G Framework [52, 53]. This work focuses mainly on OAI integration with the M-CORD framework (Figure 12) and different implementation procedures to deploy LTE network on top of M-CORD. The testbed in [53] extends the previous work further by deploying two CN instances connecting to the C-RAN architecture via the TN in order to slice and manage the TN domain. Notably, different phases of a slice lifecycle from provisioning, allocating a slice to a UE, and managing the slice are provided by this framework. Several entities are integrated into M-CORD which emulate a complete network. XOS performs service orchestration while OpenStack provides the infrastructure for deploying the services via chaining VNFs. Open Network Operating System (ONOS) [54] acts as an SDN controller and separates CP and DP functionalities. Available resources are modified and configured/reconfigured via Graphical User Interface (GUI). TN slicing is performed by running slicing policy via ONOS SDN to establish a connection flow between CN and RAN domains. In fact, ONOS inquires the OpenStack via REST API to receive the necessary information regarding the underlying platform to create TN slice between the CN and RAN domains. ONOS performs management mechanisms on TN slices via its Southbound Interface (SBI).

4.10. Dynamic Network Slicing for 5G IoT and eMBB Services [55]. This testbed (Figure 13) demonstrates sharing of the same RAN resources among eMBB and IoT services. A

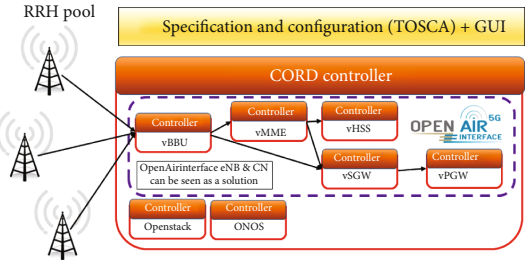


Figure 12: M-CORD framework schematic architecture [52].

Northbound SDN application is designed in this testbed to create isolated RAN slices for IoT and eMBB devices according to their service requirements. IoT devices connect to the C-RAN via a gateway. The real-time slicing decision in C-RAN is performed by an SDN controller (FlexRAN) that connects via its Northbound Interface (NBI) to a Slicing app entity, which includes IoT and eMBB modules. With the help of the scheduling process conducted by the SDN controller, the slicing app determines the number of allocated radio resources to each specific slice. In the testbed architecture, the LTE scheduling mechanism is operated by the SDN controller, where CP of the MAC layer is administered as a Northbound SDN application on the cloud. An agent module is responsible for connection establishment between the slicing scheduler entity in DP and the SDN controller in CP via the SBI. Other actions, such as admission control decision, duration of allocating radio resources to a slice, are also performed by the SDN controller and the slicing app. The testbed has been evaluated in some scenarios, such as measuring average downlink throughput in IoT and eMBB slices.

4.11. Transformable Resources Slicing Testbed for Deployment of Multiple VIMs [56]. This testbed (Figure 14) concentrates on providing Slice-as-a-Service (SlaaS) considering Data Centers (DCs). In this case, a slice is composed of a combination of DC slices (compute and storage resources) attached by Network slices (networking resources) operating on their own VIMs and Network Infrastructure Managers (NIMs), respectively. This work is considered as a supplement for Clayman's model [57]. Clayman's model consists of three layers: (1) orchestration layer, which manages the slice lifecycle, optimizes resource allocation, and coordinates DC slices and Network slices of a particular slice; (2) DC slice and Network slice controllers layer; former creates DC slices and deploys upon request VIM and later creates Network slices between DC slices and deploys upon request NIM; (3) infrastructure layer, which includes all physical resources.

As an extension to the Clayman's model, here, slices are created via so-called transformable resources, which are interpreted as physically isolated (bare metal) or virtually shared resources. The VIMs are responsible for controlling and managing the number of allocated resources to each slice. The testbed utilizes the DC slice controller to deploy VIMs according to general templates for each slice dynamically. As a result, selecting a specific VIM converts to be a choice for a tenant and not a monopolized feature assigned

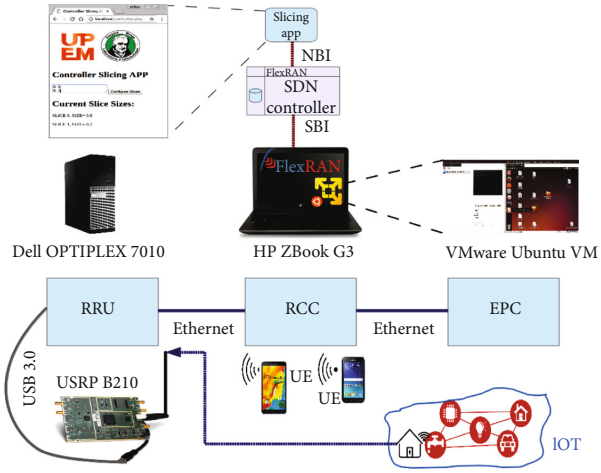


Figure 13: 5G network slicing testbed supporting IoT and eMBB services [55].

by the network provider anymore, and it can be set based on the service specification. Hence, each tenant can now operate its own VIM. The testbed comprises a server running as the DC slice controller and four nodes with the same hardware configuration and equipped with an Arduino to trigger on/off action for each node and inspect each node's status. The testbed offers an evaluation scenario to determine the required time (loading, booting, configuration, and service startup times) to establish different infrastructures (VLSP, Kubernetes, and OpenStack).

4.12. Dynamic Slice Allocation Framework (DSAF) [58]. This testbed (Figure 15) is a practical implementation to evaluate the DSAF paradigm. Basically, DSAF is an efficient resource usage model for dynamic and real-time slice (de)allocation in the CN domain, which is based on minimum CPU utilization and finding links with the lowest delay. DSAF considers allocation policies for slice requests. DSAF also brings isolation between chained NFs of a slice. It is composed of five entities: (1) Orchestrator, which manages slice (de)allocation mechanism and all of the framework elements; (2) Optimization module, which monitors the available CPU and link delays while receiving slice requests; (3) database, which maintains slice request information, slice allocation policies, and the available resources; (4) Optimization Agent, which acts as a mediator entity between the Orchestrator and the Optimization module to exchange information regarding slice allocation approaches; and (5) Hypervisor Agent, which interacts with the Orchestrator by presenting slice state information and performs slice (de)allocation. DSAF performance has been compared with First Come First Serve First Available (FCFSFA) method for different number of VNFs of a slice hosted by a Hypervisor. In these scenarios, the total number of slice requests allocated in DSAF is greater or equal than the FCFSFA scheme.

4.13. SliceNet Platform [59]. This testbed (Figure 16) proposes a QoS-aware network slicing for multiple services with distinct QoS requirements. The testbed focuses on studying use cases for providing critical services with various reliability requirements. It introduces a novel SliceNet platform strategy to provide eHealth services via 5G network slicing. SliceNet offers a realistic network slicing with guaranteed QoS requirements by QoS-programmable policies in the data plane. This is done by implementing traffic engineering functions in both hardware and software levels. Moreover, SliceNet presents a plug and play control layer to let users demand customizable network slices in the network. SliceNet suggests E2E network slicing in both single and multi-domain providers. It also contributes to a cognitive network slice management functionality to enhance the QoS requirement for the services granted by the network slices. In addition to that, SliceNet also operates an ML-enabled method for the patient's examination in critical use cases via real-time video streaming communication from an ambulance to the medical center. This testbed is an extended version of the Mosaic5G platform [43].

4.14. Iquadrat Informatica (IqInf) Testbed [60]. This framework (Figure 17) utilizes OAI for deploying RAN and CN domains, and SDN switches (composed of Open vSwitch and VLAN switch) for building TN. The separation between CP and DP in the RAN domain is achieved by implementing FlexRAN Agent API, which provides a centralized load balancing and handover mechanism while having more than one eNB in this network. The OpenDayLight (ODL) realizes SDN policy in the TN domain. With the help of PHY abstraction mode of oaisim in OAI RAN, emulating practical network scenarios with numerous UEs and eNBs is conceivable. In particular, the OAI Traffic Generator (OTG) delivers network traffic of multiple applications like Voice over IP and Machine Type Communication (MTC) in this testbed. Deploying an orchestration scheme between SDN controllers and within the entire network domains has been considered as a future enhancement for the testbed architecture.

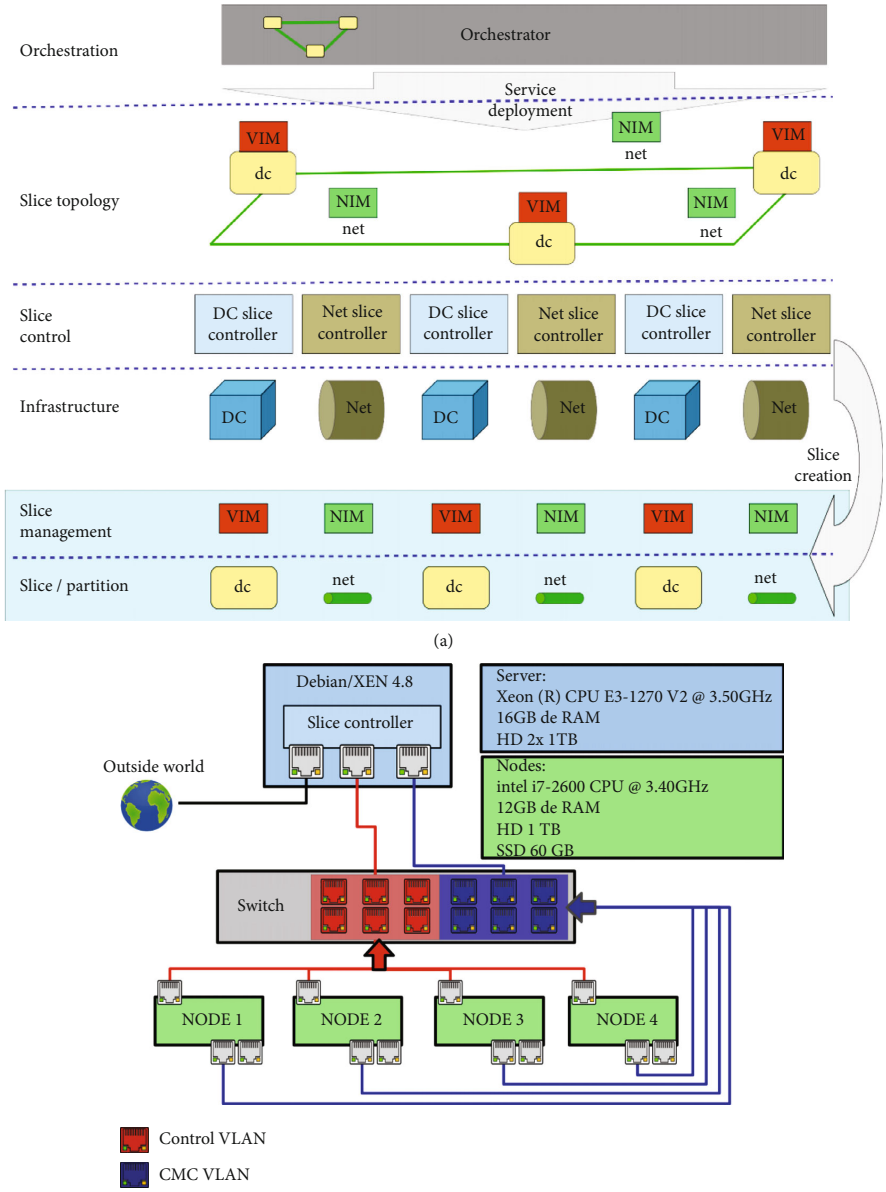


Figure 14: Transformable resources slicing testbed. (a) SlaaS. (b) Architecture [56].

4.15. Slice-Aware Service Assurance (SA) Framework [61]. This framework (Figure 18) examines Service Assurance (SA) in order to satisfy Quality of Experience (QoE) and QoS requirements in the context of network slicing. The framework integrates a novel SA-based architecture to the ETSI MANO platform to assure the services provided by different network slices in a network. Each component in the NFV MANO architecture has a counterpart in the SA-based NFV MANO platform: Slice Assurance, NS Assurance, NFV Assurance, and Infrastructure Assurance. These components operate four actions, including monitoring, analytics, management, and reporting, to guarantee the performance of the corresponding layer. This extended platform supports reporting information from all involved layers in service provisioning in the network slicing context. The platform also facilitates management and orchestration of

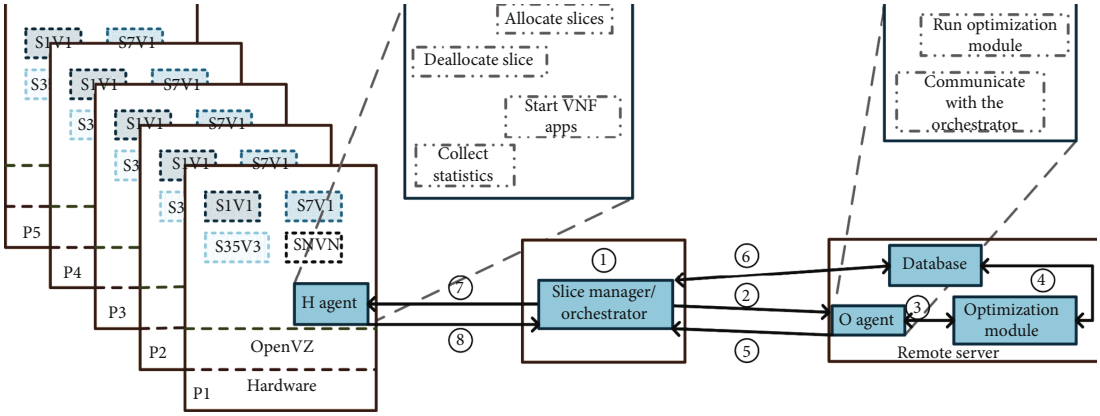


FIGURE 15: Implementing a topology with physical servers (brown blocks) and DSAF components (blue blocks with solid lines representing logical communication paths) [58].

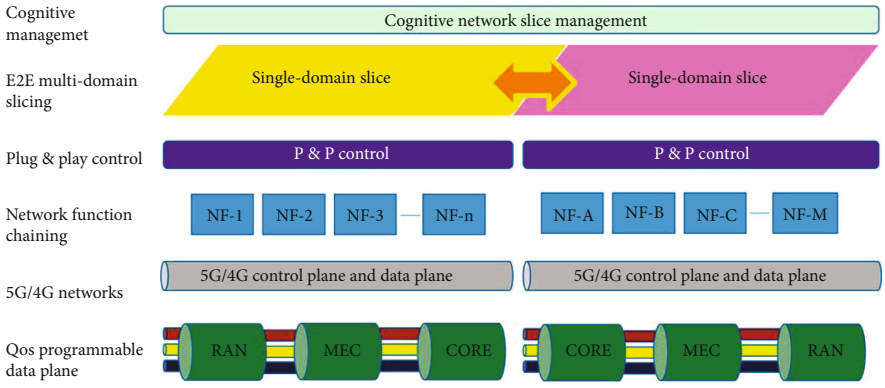


FIGURE 16: E2E network slicing approach in SliceNet platform [59].

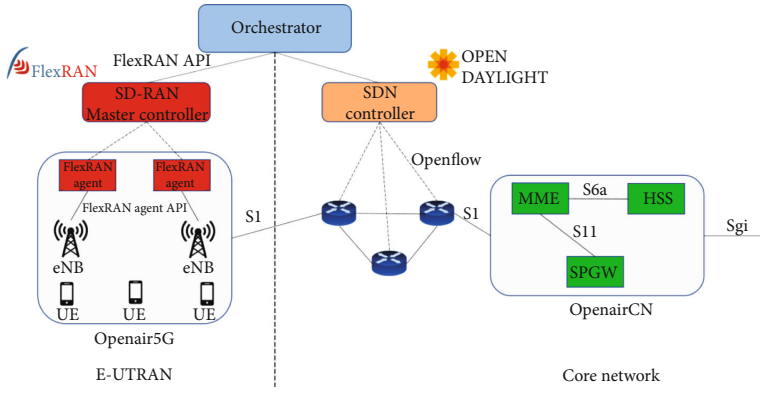


FIGURE 17: IqInf testbed architecture [60].

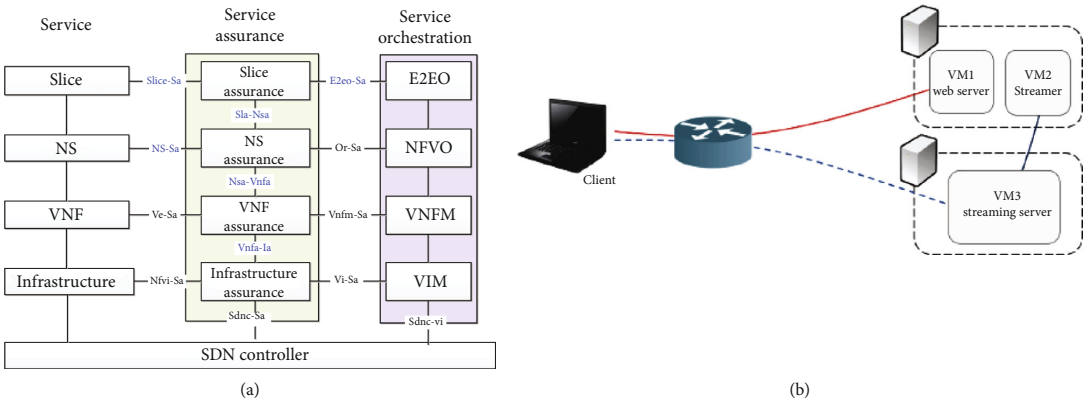


Figure 18: Slice-Aware SA framework. (a) Architecture. (b) Measurement setup. [61].

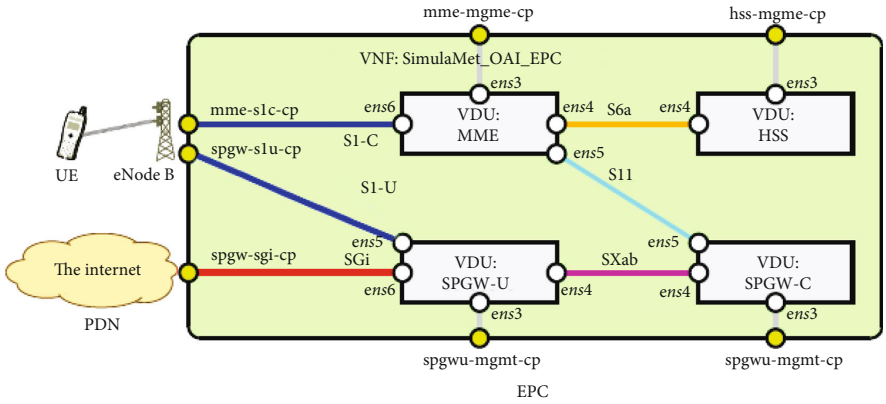


Figure 19: Simula testbed architecture [62].

various NFVs to assure that QoS and QoE requirements are fulfilled. The testbed evaluates the QoE of a service according to multiple service dependability Key Quality Indicators (KQIs). To this end, the testbed implements web content browsing and adaptive video streaming services to appraise infrastructure performance and the variation of KQIs for the service.

4.16. Simula Metropolitan Centre Testbed [62, 63]. This testbed (Figure 19) demonstrates the deployment of OAI-EPC as a VNF on a cloud environment (OpenStack), and it presents the LTE CN service instantiation via OSM. In this testbed, according to the defined descriptors at the VNF and network service levels, the internal components of OAI-EPC are firstly cloned from related repositories. Secondly, they are implemented and configured via special configuration files, Juju Charms, on four separate virtual machines (Virtual Deployment Units (VDUs) in descriptors). The goals of this implementation are to produce MEC services to EPC as well as to integrate EPC with the extended eNB software. Finally, the testbed functionality is evaluated for establishing TCP and SCTP connections in three scenarios: downloading from server to UE, uploading from UE to the server, and bidirectional communication between UE and server. In its recent release, Simula testbed implements a mobile network based on OAI-EPC deployed as a VNF using OSM, which is now integrated with C-RAN architecture with functional split capability for BBU processing functions.

4.17. 5GIIK [64, 65]. 5GIIK is a cross-location testbed (Figure 20), which deploys OAI-EPC and srsLTE eNB as cloud-based VNFs via OSM on two OpenStack platforms launched at two geographically separated areas. In this testbed, the VNF-onboarding process takes place in three phases of the VNF lifecycle. In the first phase, management policies for establishing the VNFs are performed. In the second phase, configured VNFs grant the requested services. In the third phase, reconfiguration of VNFs and monitoring of their Key Performance Indicators (KPIs) in runtime operation are provided. This testbed performs E2E network slicing via a hierarchical process by defining specific descriptors at

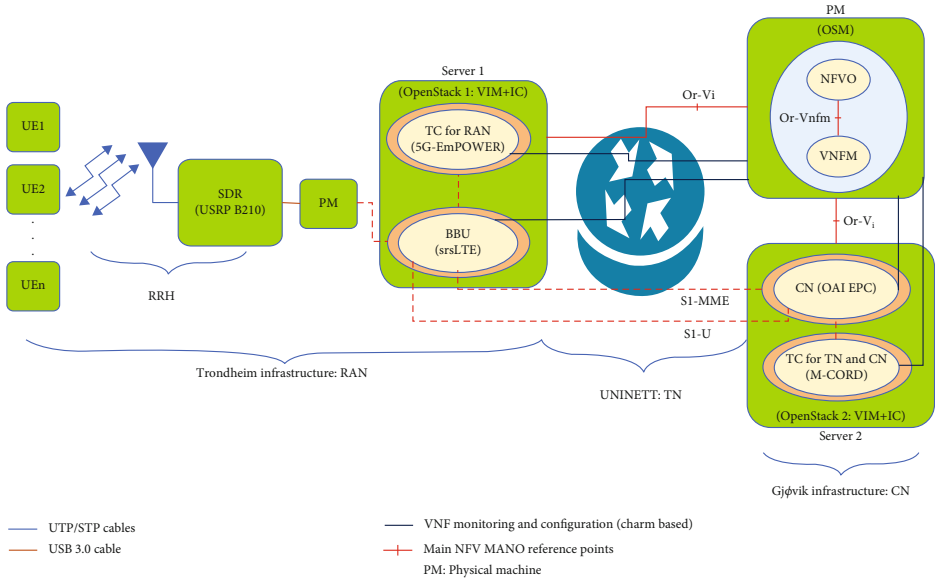


Figure 20: 5GIIK testbed architecture [64].

the VNF, network service, and network slice levels on CN and RAN domains. The testbed also integrates 5G-EmPOWER for RAN and M-CORD for TN and CN as SDN controllers. This results in supporting multi-tenancy in the RAN and also implementing slicing in the TN domains. It is worth noting that the 5G-EmPOWER assists common ML toolkits to facilitate realization and administration of machine learning models in this testbed. 5GIIK extends its capabilities by introducing Wireguard [66] to its architecture as a Virtual Private Network- (VPN-) based solution for providing slice isolation. In this solution, WireGuard-enabled VNFs operate on the NFVI via actions performed by OSM-NBI and Juju proxy charms. As a result, traffic isolation and security isolation, which are two essential features in network slicing, are granted via the integrated OSM-WireGuard framework.

4.18. Integrated Slice Management with ONAP Framework [67]. This testbed (Figure 21) investigates E2E network slicing lifecycle management (modeling, onboarding, instantiation, and operation) by integrating ONAP service orchestrator with a network slice manager entity. This integration grants a platform for (1) monitoring and collecting KPI reports that belong to the chained VNFs that create an E2E network slice and (2) evaluating the provided logs of information. In this way, multiple slices are inquired to trace whether the Service Level Agreement (SLA) between the service provider and service user is met or not. The testbed performs a use case by creating a private mobile network that affords services with best-effort and broadband QoS types via E2E network slices. Firstly, a slice is modeled as an ONAP-Network-Service composed of three VNFs (CP and DP for the CN domain, and a RAN emulator); besides, some policies for guaranteeing the SLA are defined. Secondly, the slice is deployed according to corresponding templates for each VNF. The testbed then utilizes the defined policies to perform slice management by modifying the allocated cloud resources to the two QoS types. Consequently, a dedicated channel grants a higher priority to broadband service compared to the best-effort service.

4.19. BlueArch [68]. BlueArch platform (Figure 22) includes a customized structure of some open-source tools. The testbed serves in three operational modes: simulation, emulation, and access to a physical network to communicate with other platforms. The testbed architecture comprises two main sections. Section one consists of (1) a Network Attached Storage (NAS) server representing shared storage that presents a private network; (2) a gateway router attaches a wireless access point operating on the same private network, an external OpenStack infrastructure, and the Internet; (3) Raspberry Pi accessories for MEC nodes implementation outcomes in producing IoT infrastructure in order to migrate VNFs. Section two consists of six VMs each encompassing a specific functionality: (1) an open-source PfSence (https://www.pfsense.org/) firewall for conducting regular firewall actions and also for traffic shaping, network monitoring, and load balancing; (2) employing ODL, Ryu, and HP-VAN SDN controllers hosted by a Citrix XEN server for yielding a cross-platform controlling of OvS devices in DP; (3) open MANO and RIFT.io hosted by another XEN server that operate as orchestrators and support VNF-onboarding process for network slicing; (4) an application server works as the SDN application layer hosting open-source operating systems clients, which in turn driving GNS3 UI, a hypervisor, and XEN center; (5) a network emulation server involves two

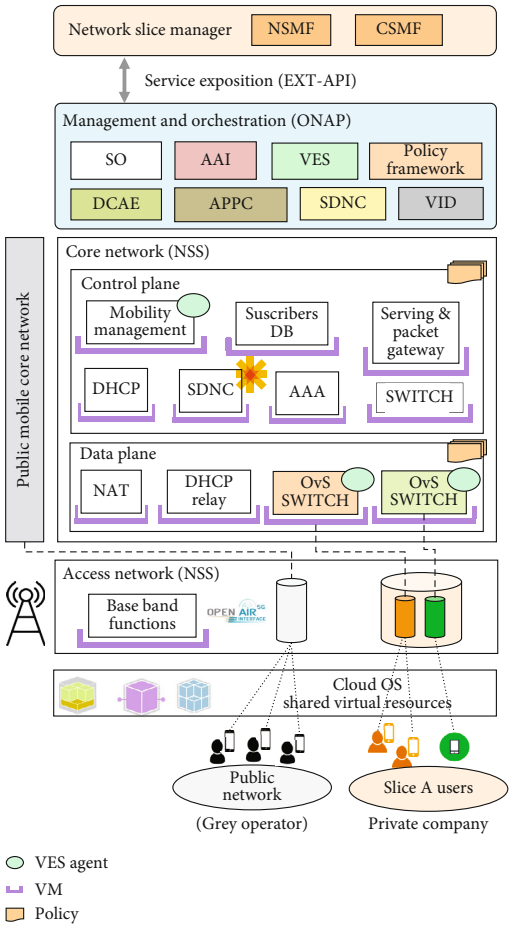


Figure 21: Architecture of the ONAP testbed integrated with slice management [67].

types of Mininet for wired and wireless SDN, and GNS3 Compute for offloaded computation resulting from GNS3 UI; (6) a MySQL-based database server interfacing the testbed with an external platform.

The testbed is evaluated in three use cases: (1) real-time monitoring of resource utilization in disaster recovery by installing ShellMon client on IoT gateways; (2) hosting VNF as a docker container when a MEC node becomes overloaded by taking a self-triggered action to relocate to another MEC node (known as VNF migration); (3) modeling wireless channel and scheduling radio resources in RAN domain employing Matlab and using the testbed to perform SDN functionality.

4.20. MEC-Enabled 5G IoT Platform [69, 70]. This work (Figure 23) is a solid proposal for onboarding and scheduling aspects in VNF lifecycle management, and it presents a programmable and flexible MEC-enabled platform for IoT traffic. In this work, VNFs are categorized into Latency Critical VNFs (LCVNFs) and Latency Tolerant VNFs (LTVNFs). As a result, the applications are also divided into (1) real-time, provided by High Priority LCVNFs (HP LCVNFs), with resources in the MEC, (2) near-real-time, provided by Low Priority LCVNFs (LP LCVNFs), and (3) non-real-time, provided by LTVNFs. The LP LCVNFs and LTVNFs are deployed on the cloud instead of MEC since they do not provide real-time applications. The work improves the joint orchestration capability in the NFVO for the MEC and cloud resources for the mentioned VNF types via two methods: (1) an online placement scheme to deliver the required management tasks at the VNF level according to the data traffic and (2) a latency embedding structure that enables VNF migration and scalability to fulfill service requirements in real time. These two methods are accomplished by introducing (1) an algorithm for VNF Forwarding Graph (VNFFG) in chained VNFs for prioritizing delay-sensitive services and (2) a second algorithm for the real-time allocation of the MEC and cloud resources to the VNFs that takes into account scale-in/out features for diverse service requirements. The testbed is deployed on several physical servers for the functionalities of the core (cloud infrastructure and NFVO) and network edge (MEC) with lower computational resources compared to the core. OpenStack, as the VIM with its telemetry feature, conducts data collection, data monitoring for future resource utilization, and placement policy through its compute schedulers. Furthermore, the OSM provides the NFVO functionality in this testbed. There are some hypervisors located at the core and the edge that afford the computing tasks. The testbed is assessed by some autoscaling, VNF placement, and online VNF scheduling scenarios.

4.21. CAI Testbed [71]. This testbed (Figure 24) offers a cost-efficient virtualized and orchestrated 5G mobile network equipped with containers and distinct fronthaul and backhaul topologies. The testbed mainly concentrates on integrating Artificial Intelligence (AI) using Kubeflow tool [72] to the management tasks in the 5G RAN and TN domains in order to optimize network performance. The testbed, called Connected AI (CAI), with the help of Kubernetes as a container-orchestrator, presents a mobile network composed of OvS devices, Ryu as SDN controller, and the OAI FlexRAN controller. CAI expedites the deployment of various network topologies on the fronthaul and backhaul by creating an emulated TN using Mininet. An AI agent takes various actions in the network by employing the information granted via Ryu and OAI FlexRAN controllers to feed ML models in order to implement several slice configurations. CAI builds a containerized implementation of OAI for C-RAN and Free5GC for the CN using Docker. The CAI testbed is evaluated via two use cases: (1) monitoring the amount of allocated radio resource blocks to different slice requests and (2) VNF placement in a cluster of containers by means of the AI agent.

5. Discussion

5.1. Comparison between Different State-of-the-Art Network Slicing Testbeds. In this section, we compare the testbeds

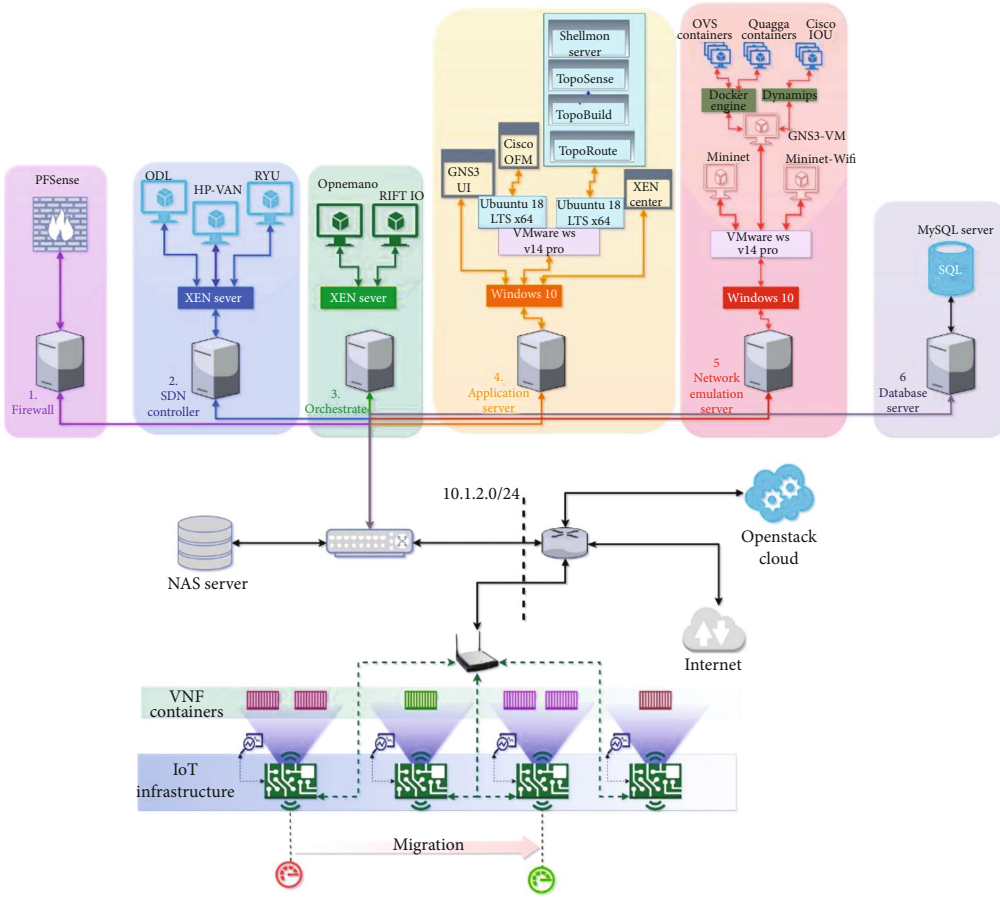


Figure 22: BlueArch testbed architecture [68].

according to the design criteria for network slicing testbeds presented in Section 3. Table 2 summarizes the major characteristics of each testbed. The testbeds in Table 2 can be arranged into two categories.

(i) The first category comprises those testbeds that partially achieve some of the *primary* or *secondary* attributes of the design criteria for network slicing testbeds. In this regard, the testbeds in [38, 40, 45, 46, 50, 55, 56, 58, 60–62, 68–71] present network slicing in a particular network domain, and they do not realize a complete E2E network slicing. Reference [56] applies network slicing within multiple-VIMs (DCs); however, this implementation is limited to one network domain, and it does not present E2E network slicing, which crosses all network domains (RAN, TN, and CN). Other testbeds such as [47] implements E2E network slicing; however, it does not offer MANO capability, multi-RATs, and multi-tenancy facilities in the architecture. The platform in [41] applies light employment of MANO entity and E2E network slicing in its design.

(ii) The second category encompasses the implementations which satisfy all of the *primary* and the majority of *secondary* attributes from the design criteria explained in Section 3. The testbeds such as those in [43, 48, 49, 52, 53, 59, 64, 67] deliver E2E network slicing with MANO privilege in their architectures along with multi-tenancy and multi-RAT support. The testbeds in [59, 64] also incorporate ML-enabled capability in their architectures, and the testbed in [64] is open-source.

5.2. Implementation Challenges for Deploying Network Slicing Testbeds. This section presents some of the current challenges for deploying small-scale network slicing testbeds and summarizes proposed solutions that can slightly mitigate these challenges.

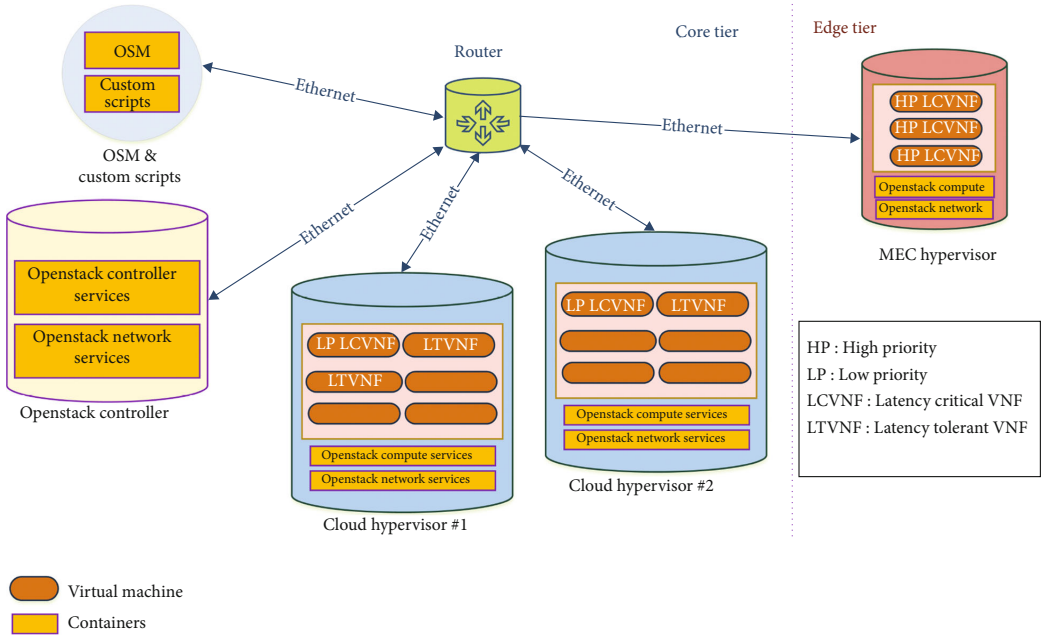


Figure 23: MEC-Enabled 5G IoT architecture [70].

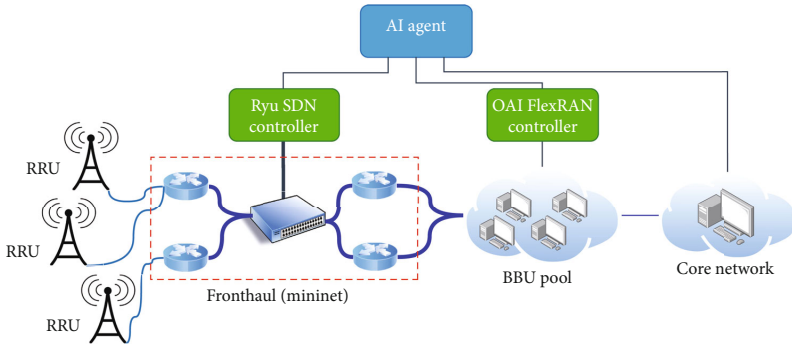


Figure 24: CAI testbed [71].

(i) *Monitoring frameworks for testbeds.* 5G is expected to provide heterogeneous services with distinct QoS requirements via utilizing network slicing. In this regard, the dynamic monitoring of the launched services is essential. This becomes challenging when recognizing the issues of possible performance degradation of the services. In fact, the multi-layered architecture of the 5G network, as shown in Figure 1, causes such challenges. Intelligently identifying such issues requires analyzing multiple possible sources of the problem via particular frameworks to effectively monitor the deployment and performance of services. To partially address this problem, different types of monitoring capabilities are integrated in some of the elaborated testbeds. The testbeds in which OSM acts as an orchestrator in their architectures, such as [40, 45, 46, 59, 62, 64, 70], usually employ the interaction of the system monitoring module (MON) with a monitoring toolkit such as Prometheus [73] for collection of VNFs' metrics and then utilize Grafana [74] to visualize the collected data. The testbeds with an ONAP orchestrator, such as [67], focus on SLA monitoring by exploiting Data Collection Analytics & Events (DCAE) and Virtual Event Streaming (VES) components. Reference [43] benefits from a monitoring application in the store component. The architecture in [61] offers monitoring functions in each layer of SA and also implements virtual monitoring agents or virtual probes at each point of presence to actively observe network services.

Table 2: Comparison of small-scale testbeds for network slicing in 5G, which ✓ denotes supported feature and ✗ denotes unsupported feature.

Testbed	SDN	NFV	Cloud comp.	Multi-domain	Multi-tenancy	MANO	Multi-RATs	E2E slicing	Cross-location	ML-enabled	Open-source	MANO type
1. 5G4IoT [38, 39]	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
2. 5GTN [40]	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✓ (https://5gtn.fi/)	OSM, CloudBand
3. SEMIoTICS [41]	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓ (https://www.semiotics-project.eu/)	OpenStack tacker
4. Mosaic5G [43]	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓ (http://mosaic5g.io/)	JOX
5. Orion [45, 46]	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	OSM and a customized orchestrator
6. 5G Testbed for NS [47]	✗	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓ (https://github.com/ashxz47)	✗
7. POSENS [48, 49]	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓ (https://github.com/wnlUc3m)	Customized OSM
8. UPC testbed [50]	✓	✓	✗	✓	✓	✗	✓	✗	✗	✓	✗	✗
9. M-CORD based testbed [52, 53]	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓ (https://nick133371.github.io/)	XOS
10. NS for 5G IoT and eMBB [55]	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗
11. Transformable resources slicing testbed [56]	✓	✓	✓	✗(E2E slice traverses over RAN, TN and CN.)	✓	✓	✗	✗	✗	✗	✗	VLSP, Kubernetes, and OpenStack
12. DSAF [58]	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	Customized python-based orchestrator
13. SliceNet [59]	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	OSM, OpenBaton
14. IqInf testbed [60]	✓	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗
15. Slice-aware SA testbed [61]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	Service assurance integrated with MANO
16. Simula [62, 63]	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓ (https://github.com/simula/5gvinni-oai-ns)	OSM
17. 5GIIK [64, 65]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓ (https://bit.ly/3rgOgd6)	OSM
18. ONAP based testbed [67]	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	ONAP
19. BlueArch [68]	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	Open MANO, RIFT.io
20. MEC IoT platform [69, 70]	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	OSM
21. CAI [71]	✓	✓	✓	✓	✗	✗	✓	✗	✗	✓	✓ (https://bit.ly/3tXErSX)	✗

(ii) *Cross-location testbeds.* Launching testbeds over separate areas impacts the service performance because of delay, jitter, and packet loss. This issue becomes even more challenging when providing delay-sensitive services. Consequently, discovering techniques to enhance service performance in cross-location deployment is exceptionally important. As mentioned in Table 2, the testbeds in [46, 64] deploy a cross-location architecture for C-RAN (RAN and MEC) and CN on two separate cloud-based infrastructures. In these two testbeds, the MANO entity (OSM), with the help of an SDN-assist feature, partially considers this issue by implementing application-aware traffic flow strategies to mitigate the generated latency because of the cross-location architecture, which results in enhancing connection reliability [46, 75].

(iii) *C-RAN deployment on testbeds.* Implementing C-RAN architecture on a testbed using open-source software packages can be challenging since the interaction between BBU and RRHs entails extremely low latency. Some attempts, such as in [43, 46, 63] resolve this problem by deploying the BBU section with a combination of PNF and VNF. They split the protocol stack of BBU into two sections in their solution instead of launching the BBU completely in a cloud-based environment. In particular, the functionality of the PHY layer of the BBU is split into a lower-PHY as PNF (to run on a physical machine along with RRHs) and higher-PHY as VNF (to run on a cloud infrastructure). In this way, the communication between (lower-PHY layer of) BBU and RRHs fulfills the ultralow delay requirement while keeping the benefit of the cloud-based implementation of (higher-PHY layer of) BBU.

(iv) *Resource management on testbeds with limited infrastructure capacity.* Resource management is considered as another possible challenge while deploying testbeds on infrastructures with limited physical and/or virtual resources. Since diverse services demand various amounts of networking, computing, and storage resources, it is essential to identify optimized methods to allocate available resources to service instances. To deal with this issue, testbeds that adopt OpenStack as VIM in their infrastructures, such as references [38–41, 45, 46, 62, 64], can enable Telemetry Data Collection to gather event and data for utilization statistics of the infrastructure resources.

(v) *Slice isolation on testbeds.* The (intra/inter) slice isolation concept is a common concern while implementing network slicing, and it is not limited to research testbeds. It is worth stating that there are some endeavors to tackle the isolation issue. Testbeds, such as those in [43, 45, 46, 55, 59], which utilize FlexRAN in their architectures, present partial slice isolation in the RAN domain. The testbeds in [48, 49] perform isolation in the RAN domain by slicing the protocol stack down to RRC, RLC, and MAC layers. Nevertheless, introducing and realizing efficient and practical techniques to guarantee isolation in network slicing, especially in the RAN domain, is subject of future work. The work presented in [65] is one step towards providing traffic isolation and security isolation in network slicing.

6. Conclusion

Network slicing testbeds with dedicated management and orchestration entities endeavor to outline and emulate trial and real use cases to achieve network slicing. On this basis and according to pioneer technologies, this paper addresses the principal design criteria for creating and deploying experimental environments for network slicing in 5G. After that, the paper explains the most common small-scale state-of-the-art testbeds for network slicing with their characteristics. The presented testbeds are then reviewed and compared via the design criteria, followed by possible challenges while creating such experimental platforms. Although many efforts have been performed to create testbeds for examining and evaluating network performance under various use cases in network slicing, there are still open research questions in this field.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] A. Antonopoulos, "Bankruptcy problem in network sharing: fundamentals, applications and challenges," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 81–87, 2020.

[2] P. Rost, A. Banchs, I. Berberana et al., "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 84–91, 2016.

[3] R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: state-of-the-art and research challenges," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.

[4] 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*, vol. 55, no. 5, pp. 80–87, 2017.

[5] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, 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.

[6] P. M. Mell and T. Grance, *The NIST Definition of Cloud Computing*, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, USA, 2011.

[7] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: the communication

perspective," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[8] ETSI, *Network Functions Virtualisation (NFV): Architectural Framework*, vol. 2, no. 2, 2013ETSI Gs NFV, 2013.

[9] 3GPP, *Study on Management and Orchestration of Network Slicing for Next Generation Network*, 2017.

[10] F. Kaltenberger, A. P. Silva, A. Gosain, L. Wang, and T.-T. Nguyen, "Openairinterface: democratizing innovation in the 5G era," *Computer Networks*, vol. 176, article 107284, 2020.

[11] Stackify, "The Ultimate Guide to Performance Testing and Software Testing: Testing Types, Performance Testing Steps, Best Practices, and More," April 2021, https://stackify.com/ultimate-guide-performance-testing-and-softwaretesting/.

[12] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: a survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[13] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[14] T. Huang, F. R. Yu, C. Zhang, J. Liu, J. Zhang, and Y. Liu, "A survey on large-scale software defined networking (SDN) testbeds: approaches and challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 891–917, 2017.

[15] L. U. Khan, I. Yaqoob, N. H. Tran, Z. Han, and C. S. Hong, "Network slicing: recent advances, taxonomy, requirements, and open research challenges," *IEEE Access*, vol. 8, pp. 36009–36028, 2020.

[16] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5G networks: state-of-the-art and the road ahead," 2020, https://arxiv.org/abs/2005.10027.

[17] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: a survey of taxonomy, architectures and future challenges," *Computer Networks*, vol. 167, article 106984, 202010.1016/j.comnet.2019.106984.

[18] OpenStack, "OpenStack The Most Widely Deployed Open Source Cloud Software in the World," April 2021, https://www.openstack.org/.

[19] OpenVIM, "Telefónica NFV reference lab," April 2021, https://github.com/nfvlabs/openvim.

[20] Kubernetes, "Kubernetes Production-Grade Container Orchestration," April 2021, https://kubernetes.io.

[21] Open Source MANO (OSM), "OSM Open Source NFV Management and Orchestration (MANO) software stack aligned with ETSI NFV," April 2021, https://osm.etsi.org.

[22] ONAP, "ONAP Open Networking Automation Platform," April 2021, https://www.onap.org/.

[23] OpenBaton, "OpenBaton An extensible and customizable NFV MANO-compliant framework," April 2021, http://openbaton.org.

[24] Cloudify, "Cloudify Multi Cloud Orchestration," April 2021, https://cloudify.co/.

[25] SONATA, "Sonata agile development. testing and orchestration of services in 5g virtualized networks," April 2021, https://www.sonata-nfv.eu.

[26] Katana Wiki Home, "MediaNetworks Laboratory," April 2021, https://github.com/medianetlab/katana-slice_manager/wiki.

[27] I. Gomez-Miguelez, A. Garcia-Saavedra, P. Sutton, P. Serrano, C. Cano, and D. Leith, *Srslte: An Opensource Platform for Lte Evolution and Experimentation*, 2016.

[28] N. Nikaein, M. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface," *English, Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.

[29] O-RAN, "O-RAN Operator Defined Open and Intelligent Radio Access Networks," April 2021, https://www.o-ran.org.

[30] J. Murray and J. Huang, *Blog*, O-RAN, 2020, April 2021, https://www.o-ran.org/blog/2020/6/28/the-2nd-release-of-o-ran-software-bronze-addssupport-for-new-key-elements-of-the-o-ran-architecture-and-updates-to-align-with-the-latest-o-ranspecifications.

[31] Open5GS, "Open5GS Open source project of 5GC and EPC," April 2021, https://open5gs.org/.

[32] TECHNICAL STEERING TEAM (TST), *OMEC Open Mobile Evolved Core*, Open Networking Foundation (ONF), 2020, April 2021, https://www.opennetworking.org/omec/.

[33] free5GC, *free5GC Link the World!*April 2021, https://www.free5gc.org/.

[34] Open Networking Foundation (ONF), "ODTN Open and Disaggregated Transport Network," April 2021, https://www.opennetworking.org/odtn/.

[35] D. Gligoroski and K. Kralevska, "Expanded combinatorial designs as tool to model network slicing in 5G," *IEEE Access*, vol. 7, pp. 54879–54887, 2019.

[36] B. Sonkoly, J. Czentye, R. Szabo et al., "Multi-domain service orchestration over networks and clouds," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 377-378, 2015.

[37] C. Kilinc, M. Ericson, P. Rugeland et al., "5G multi-rat integration evaluations using a common PDCP layer," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, Sydney, NSW, Australia, 2017.

[38] B. Dzogovic, V. T. Do, B. Feng, and T. van Do, "Building virtualized 5G networks using open source software," in *2018 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, pp. 360–366, Penang, Malaysia, 2018.

[39] B. Dzogovic, B. Santos, V. T. Do, B. Feng, N. Jacot, and T. Van Do, "Connecting remote eNodeB with containerized 5G C-RANs in OpenStack cloud," in *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/ 2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pp. 14–19, Paris, France, 2019.

[40] M. Arif, O. Liinamaa, I. Ahmad, A. Pouttu, and M. Ylianttila, "On the demonstration and evaluation of service-based slices in 5G test network using NFV," in *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*, pp. 1–6, Marrakech, Morocco, 2019.

[41] P. Mekikis, K. Ramantas, A. Antonopoulos et al., "NFV-enabled experimental platform for 5G tactile internet support in industrial environments," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 1895–1903, 2020.

[42] SEMIoTICS, "SEMIoTICS Smart End-to-end Massive IoT Interoperability, Connectivity and Security," April 2021, https://www.semiotics-project.eu/.

[43] N. Nikaein, C.-Y. Chang, and K. Alexandris, "Mosaic5G," *ACM SIGCOMM Computer Communication Review*, vol. 48, no. 3, pp. 29–34, 2018.

[44] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "Flexran: a flexible and programmable platform for software-defined radio access networks," in *Proceedings of the 12th International on Conference on Emerging*

Networking EXperiments and Technologies, ser. CoNEXT '16, pp. 427–441, Irvine, California, USA, 2016.

[45] X. Foukas, M. K. Marina, and K. P. Kontovasilis, "Orion: Ran slicing for a flexible and cost-effective multiservice mobile network architecture," in *In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (MobiCom '17). Association for Computing Machinery*, pp. 127–140, New York, NY, USA, 2017.

[46] X. Foukas, F. Sardis, F. Foster, M. K. Marina, M. A. Lema, and M. Dohler, "Experience building a prototype 5G testbed," in *Proceedings of the Workshop on Experimentation and Measurements in 5G (EM-5G'18). Association for Computing Machinery*, pp. 13–18, New York, NY, USA, 2018.

[47] A. Shorov, "5G testbed development for network slicing evaluation," in *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pp. 39–44, Saint Petersburg and Moscow, Russia, 2019.

[48] G. Garcia-Aviles, M. Gramaglia, P. Serrano, and A. Banchs, "POSENS: a practical open source solution for end-to-end network slicing," *IEEE Wireless Communications*, vol. 25, no. 5, pp. 30–37, 2018.

[49] G. Garcia-Aviles, M. Gramaglia, P. Serrano, F. Gringoli, S. Fuente-Pascual, and I. L. Pavon, "Experimenting with open source tools to deploy a multi-service and multi-slice mobile network," *Computer Communications*, vol. 150, pp. 1–12, 2020.

[50] K. Koutlia, R. Ferrus, E. Coronado Calero et al., "Design and experimental validation of a software-defined radio access network testbed with slicing support," *Wireless Communications and Mobile Computing*, vol. 2019, 17 pages, 2019.

[51] E. Coronado, S. N. Khan, and R. Riggio, "5G-EmPOWER: a software-defined networking platform for 5G radio access networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 715–728, 2019.

[52] C. Huang, C. Ho, N. Nikaein, and R. Cheng, "Design and prototype of a virtualized 5G infrastructure supporting network slicing," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pp. 1–5, Shanghai, China, 2018.

[53] M. T. Abbas, T. A. Khan, A. Mahmood, J. J. D. Rivera, and W. Song, "Introducing network slice management inside m-cord-based-5G framework," in *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*, pp. 1-2, Taipei, Taiwan, 2018.

[54] P. Berde, M. Gerola, J. Hart et al., "Onos: towards an open, distributed SDN OS," in *Proceedings of the Third Workshop on Hot Topics in Software Defined Networking, ser. HotSDN'14*, pp. 1–6, Chicago, IL, USA, 2014.

[55] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar, "Dynamic network slicing for 5G IoT and eMBB services: a new design with prototype and implementation results," in *2018 3rd Cloudification of the Internet of Things (CIoT)*, pp. 1–7, Paris, France, 2018.

[56] L. A. Freitas, V. G. Braga, S. L. Corrêa et al., "Slicing and allocation of transformable resources for the deployment of multiple virtualized infrastructure managers (vims)," in *4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, pp. 424–432, Montreal, QC, Canada, 2018.

[57] S. Clayman, "Network slicing supported by dynamic vim instantatiation," in *IETF 100*, Singapore, 2017.

[58] D. Sattar and A. Matrawy, "Dsaf: dynamic slice allocation framework for 5G core network," 2019, https://arxiv.org/abs/1905.03873.

[59] Q. Wang, J. Alcaraz-Calero, R. Ricart-Sanchez et al., "Enable advanced QoS-aware network slicing in 5G networks for slice-based media use cases," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 444–453, 2019.

[60] K. Ramantas, E. Kartsakli, M. Irazabal, A. Antonopoulos, and C. Verikoukis, "Implementation of an SDN-enabled 5G experimental platform for core and radio access network support," in *Interactive Mobile Communication, Technologies and Learning*pp. 791–796, Springer, ChamISBN: 978-3-319-75174-0.

[61] J. Kim and M. Xie, "A study of slice-aware service assurance for network function virtualization," in *2019 IEEE Conference on Network Softwarization (NetSoft)*, pp. 489–497, Paris, France, 2019.

[62] T. Dreibholz, "Flexible 4G/5G testbed setup for mobile edge computing using OpenAirInterface and open source mano," Springer International Publishing, Web, Artificial Intelligence and Network Applications, 2020.

[63] A. F. Ocampo, T. Dreibholz, M.-r. Fida, A. Elmokashfi, and H. Bryhni, *Integrating Cloud-RAN with Packet Core as VNF Using Open Source MANO and OpenAirInterface*, IEEE Computer Society, Sydney, New SouthWales/Australia, 2020.

[64] A. Esmaeily, K. Kralevska, and D. Gligoroski, "A cloud-based SDN/NFV testbed for end-to-end network slicing in 4G/5G," in *2020 6th IEEE Conference on Network Softwarization (NetSoft)*, pp. 29–35, Ghent, Belgium, 2020.

[65] S. Haga, A. Esmaeily, K. Kralevska, and D. Gligoroski, "5G network slice isolation with WireGuard and open source MANO: a VPNaaS Proof-of-Concept," in *2020 IEEE Conference on Network Function Virtualization and Software Defined Networks, NFV-SDN 2020*, pp. 181–187, Leganes-Madrid, Spain, 2020.

[66] J. A. Donenfeld, "WireGuard: Next Generation Kernel Network Tunnel," in *24th Annual Network and Distributed System Security Symposium, NDSS*, The Internet Society, San Diego, CA, USA, 2017, https://www.ndss-symposium.org/ndss2017/ndss-2017-programme/wireguard-next-generation-kernel-network-tunnel/.

[67] V. Q. Rodriguez, F. Guillemin, and A. Boubendir, "5G e2e network slicing management with onap," in *2020 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, pp. 87–94, Paris, France, 2020.

[68] S. Ghosh, "Bluearch - an implementation of 5G testbed," *Journal of Communications*, vol. 14, pp. 1110–1118, 2019.

[69] I. Sarrigiannis, E. Kartsakli, K. Ramantas, A. Antonopoulos, and C. Verikoukis, "Application and network VNF migration in a MEC-enabled 5G architecture," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 1–6, Barcelona, Spain, 2018.

[70] I. Sarrigiannis, K. Ramantas, E. Kartsakli, P. Mekikis, A. Antonopoulos, and C. Verikoukis, "Online VNF lifecycle management in an MEC-enabled 5G IoT architecture," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4183–4194, 2020.

[71] C. V. Nahum, L. D. N. M. Pinto, V. B. Tavares et al., "Testbed for 5G connected artificial intelligence on virtualized networks," *IEEE Access*, vol. 8, pp. 223202–223213, 2020.

[72] E. Bisong, "Kubeflow and kubeflow pipelines," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pp. 671–685, Apress, Berkeley, CA, USA, 2019, ISBN: 978-1-4842-4469-2.

[73] Prometheus, "Prometheus From metrics to insight," April 2021, https://prometheus.io/.

[74] Grafana, "Grafana Labs Your observability wherever you need it," April 2021, https://grafana.com/.

[75] Technical Steering Committee, *Osm Release Eight Notes*, Open Source MANO (OSM), 2020, April 2021, https://osm.etsi.org/wikipub/images/5/56/OSM_Release_EIGHT_-_Release_Notes.pdf.

Paper II

A. Esmaeily, K. Kralevska, and D. Gligoroski, "A Cloud-based SDN/NFV Testbed for End-to-End Network Slicing in 4G/5G," 2020 6th IEEE Conference on Network Softwarization (NetSoft), Ghent, Belgium, 2020, pp. 29-35, doi: 10.1109/NetSoft48620.2020.9165419.[1]

[1]

A Cloud-based SDN/NFV Testbed for End-to-End Network Slicing in 4G/5G

Ali Esmaeily, Katina Kralevska, and Danilo Gligoroski
Dep. of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU)
Email: {ali.esmaeily, katinak, danilog}@ntnu.no

***Abstract*—Network slicing aims to shape 5G as a flexible, scalable, and demand-oriented network. Research communities deploy small-scale and cost-efficient testbeds in order to evaluate network slicing functionalities. We introduce a novel testbed, called 5GIIK, that provides implementation, management, and orchestration of network slices across all network domains and different access technologies. Our methodology identifies design criteria that are a superset of the features present in other state-of-the-art testbeds and determines appropriate open-source tools for implementing them. 5GIIK is one of the most comprehensive testbeds because it provides additional features and capabilities such as slice provision dynamicity, real-time monitoring of VMs and VNF-onboarding to different VIMs. We illustrate the potentials of the proposed testbed and present initial results.**

***Index Terms*—Open-source testbed, 4G, 5G, SDN, NFV, Cloud, OSM, Orchestrator, E2E network slicing.**

I. Introduction

Network slicing is considered as an enabling technology to achieve the ambitious expectations of the fifth generation of mobile networks (5G), such as providing various services with different requirements over the same network. An end-to-end (E2E) network slice [1] is an isolated logical network that provides a specific network service based on an accurately defined service demand upon a shared physical infrastructure. Each slice can then be controlled and managed independently.

Network Function Virtualization (NFV) [2], Software Defined Networking (SDN) [3] and Cloud computing [2] are the three key technologies for implementing network slicing in 4G/5G. Since 5G aims to provide ultra-low latency services, Multi-access Edge Computing (MEC) is a complementary technology to cloud computing. In addition to these enabling technologies, the existence of an entity which performs efficient resource management and orchestration is inevitable. The Management and Network Orchestration (MANO) framework coordinates between available physical and virtual networking, storage and compute resources. These resources are required for creating, managing and delivering services through different slices. ETSI has developed a NFV MANO framework [4] that is composed of three functional blocks connected via reference points presented in Figure I:

- Virtualized Infrastructure Manager (VIM) controls the NFV Infrastructure (NFVI) resources within an operator's infrastructure domain. Thus, VIM is able to gather performance and fault measurement information of these resources. VIM also oversees the allocation of the NFVI resources to the available Virtual Network Functions (VNFs).
- VNF Manager (VNFM) supervises a VNF or multiple VNFs and performs the life cycle management of VNF instances. Life cycle management includes setting up, maintaining and taking down VNFs.
- NFV Orchestrator (NFVO) manages resource and service orchestration and is responsible for the entire life cycle management of various network services. Firstly, NFVO collects information about physical and virtual resources located in NFVI via the VIM. Secondly, NFVO updates its information about the available VNFs in NFVI continuously. In this way, NFVO initializes several network services by chaining particular PNFs and/or VNFs. NFVO can maintain and terminate a network service whenever there is no call for that specific service.

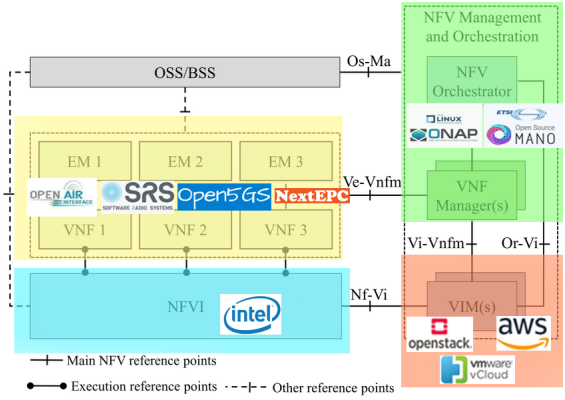


Fig. 1. Different open-source software solutions mapped to the ETSI-NFV MANO framework [4].

One cost-efficient way of testing new solutions in 5G is by developing testbeds with functionalities close to real networks. By cost-efficient, we mean that setting up the testbed does not require the purchase of specialized purpose hardware and software, i.e. the software solutions are easily deployable and maintainable on standards PCs. Figure I shows some of the leading open-source solutions for the different modules in the NFV MANO framework. A network slice can span over all network domains: 1) Radio Access Network (RAN) emulated with Software Radio Systems LTE (srsLTE), OpenAirInterface (OAI) or Open5GS; 2) Transport Network (TN); 3) Core Network (CN) implemented with OAI or NextEPC. The MANO

978-1-7281-5684-2/20/$31.00 ©2020 IEEE

entity, represented with Open Source MANO (OSM) or Open Networking Automation Platform (ONAP), coordinates the available resources. OpenStack, AWS or WMware vCloud suite can act as the VIM.

Our contribution: The contribution of this paper is threefold. First, we combine the open-source solutions: srsLTE for the RAN, OAI for the CN, OSM for the MANO entity and OpenStack for the VIM, in order to build an E2E network slicing testbed called 5GIIK. The value of the presented testbed is in its impact as a tool for experimentation around the community. Second, we identify the desired features for novel 4G/5G testbeds, including multi-tenancy support, orchestrating capabilities, support of multi-radio access technologies, deployment of E2E network slicing and open-source. Last but not least, we give an overview of common small-scale, non-federated and cost-efficient testbeds that are easily deployable without requiring a substantial financial investment. These testbeds are compared based on the identified features, and the comparison serves as a selection criterion for the proposed testbed. We show that our testbed is one of the most comprehensive testbeds because it enables additional capabilities such as slice provision dynamicity, real-time monitoring of VMs and VNF instantiation to different VIMs.

II. Small-scale testbeds for network slicing

A. Design criteria for network slicing testbeds

We first identify the key attributes for creating a 5G testbed that can emulate the principal features of a real network. These attributes are later used as assessment criteria for state-of-the-art testbeds.

- **Support of the main enabling technologies:** The proposed testbed should be based on SDN, NFV and cloud computing. Therefore, flexibility, programmability and dynamicity in the network are granted, in addition to providing an orchestration on different network levels.
- **Multi-domain support:** A 5G testbed should span across all network domains (air interface, RAN, TN, CN) in order to provide realization and management of E2E network slicing.
- **Multi-radio access technologies support:** 5G integrates different radio access technologies (RATs), thus, Long Term Evolution (LTE), WiFi and New Radio (NR) should be deployed on the same platform.
- **Multi-tenancy support:** Several mobile network operators or service providers (over-the-top players) to be able to share infrastructure by each of them acquiring one or several network slices.
- **End-to-end network slicing:** The slicing should be deployed across all network domains, i.e. a network slice instance consists of network slice subnet instances from different domains [5].
- **Open-source:** The testbed is open-source with well-defined interfaces.

B. An overview of the state-of-the-art network slicing testbeds

We give here a short description of the existing testbeds and compare them in Table I.

1) Secure 5G4IoT Lab [6]: This testbed deploys OAI in containers to virtualize both Evolved Packet Core (EPC) and eNB. For scalability purposes, the testbed has been implemented in several containers. It has been evaluated by producing two isolated network slices (eHealth and Internet light) on the same infrastructure.

2) 5G Test Network (5GTN) [7]: In this testbed, located at Oulu University, the RAN operates on licensed LTE and 5G bands. By changing the Access Point Name between EPC (deployed on OpenStack) and IP Multimedia System (deployed on VMWare), UE switching between two slices is possible. The testbed has been tested for CPU utilization, throughput and delay for the two specific slices.

3) 5G Tactile Internet platform [8]: This testbed follows the SEMIoTICS architecture, consisting of backend/cloud, networking and field layers, to create a 5G platform for providing E2E services for industrial IoT applications with sub-millisecond latency. The testbed performance has been assessed for performing E2E slicing and dynamically sharing the available bandwidth between two VNFs, one for smart monitoring and one for actuating.

4) Mosaic5G [9]: This testbed brings flexibility and scalability to service provision. The testbed architecture consists of five software modules along with hardware components: OAI, FlexRAN, LL-MEC, Store and JOX. The Mosaic5G platform has been used for a few use cases such as critical e-Health, V2X communication for intelligent transportation systems and multi-service management/orchestration for smart cities.

5) Orion [10]: The architecture of Orion provides the sharing of RAN resources in addition to providing isolation between slices, and so, operation in one slice cannot degrade the performance of another slice. This is achieved by having an independent control plane in the RAN domain for each slice. As a result, Orion offers the opportunity to deploy different service characteristics in the RAN domain and it is a concrete step towards realizing RAN-as-a-Service.

6) 5G Testbed for Network Slicing Evaluation [11]: The testbed utilizes OAI for both RAN and CN domains. There are two CNs which share radio resources of a single eNB in the RAN. The testbed has been appraised for connection establishment for both normal LTE UEs and UEs with an implemented Network Slice Selection Assistance Information.

7) POSENS [12]: POSENS provides efficient resource utilization for creating independent and customizable E2E slices. RAN slicing can be realized via three possibilities: 1) Slice-aware shared RAN where the whole radio domain is shared, but CNs are distinguished by the specific services they provide and a UE can utilize different slices provided by the CNs; 2) Slice-specific radio bearer where only cell-specific functionality is shared; and 3) Slice-specific RAN where apart from the air interface, slices of different tenants are isolated in other protocol stack layers.

8) UPC University testbed [13]: This testbed implements automatically RAN slicing via RESTful API. The testbed applies the slice-aware policy in Radio Resource Management (RRM) for admission control and scheduling processes. 5G-

TABLE I
COMPARISON OF SMALL SCALE TESTBEDS FOR NETWORK SLICING IN 5G.

Testbed	SDN	NFV	Cloud comp.	Multi-domain	Multi-tenancy	MANO	Multi-RATs	E2E slicing	Open-source
1. Secure 5G4IoT Lab [6]	✓	✓	✓	✓	✓	-	-	-	-
2. 5G Test Network (5GTN) [7]	✓	✓	✓	✓	-	✓	✓	-	✓
3. 5G Tactile Internet platform [8]	✓	✓	✓	✓	✓	✓	-	✓	✓
4. Mosaic5G [9]	✓	✓	✓	✓	✓	✓	✓	✓	✓
5. Orion [10]	✓	✓	✓	✓	✓	✓	-	-	-
6. 5G Testbed for Network Slicing [11]	-	✓	✓	✓	-	-	-	✓	✓
7. POSENS [12]	✓	✓	✓	✓	✓	✓	✓	✓	✓
8. UPC University testbed [13]	✓	✓	-	✓	✓	-	✓	-	-
9. M-CORD based 5G Frameworks [14]	✓	✓	✓	✓	✓	✓	✓	✓	✓
10. NS for 5G IoT and eMBB [15]	✓	✓	✓	✓	✓	-	✓	-	-
11. CHARISMA [16]	✓	✓	✓	✓	✓	✓	-	✓	✓
12. Slice-Aware Service Assurance [17]	✓	✓	-	-	-	✓	-	-	-
13. Simula Metropolitan Centre [18]	✓	✓	✓	✓	✓	✓	✓	-	✓
14. 5GIIK (our proposal)	✓	✓	✓	✓	✓	✓	✓	✓	✓

EmPOWER [19], as the central entity in the testbed, allows RAN slicing management and it also shares the available radio resources among the created RAN slices according to RRM descriptors.

9) Mobile-Central Office Re-Architected as Datacenter (M-CORD) based 5G framework [14]: The work in [14] focuses mainly on OAI integration with the M-CORD platform and different implementation procedures to deploy LTE network on top of M-CORD.

10) Dynamic Network Slicing for 5G IoT and eMBB services [15]: This testbed demonstrates the sharing of the same RAN resources among enhanced Mobile Broadband (eMBB) and IoT services. The real-time slicing decision in C-RAN is performed by a SDN controller (FlexRAN) that connects via its Northbound Interface to an entity called Slicing app, which includes IoT and eMBB modules.

11) CHARISMA testbed [16]: This testbed has been designed for practical analysis in the CHARISMA project, and the goal is to bring the network processing close to the users. By employing Ethernet Virtual Connections and Virtual LAN-ID concepts in the testbed, multi-tenancy and slice isolation are achieved. WiFi technology and cloud-based servers present the RAN and CN domains, respectively.

12) Slice-Aware Service Assurance Framework [17]: This testbed measures Quality of Experience (QoE) of a specific service according to the several service dependability Key Quality Indicators (KQIs). The testbed provides web content browsing and adaptive video streaming services to assess infrastructure performance and the KQIs alteration for each service.

13) Simula Metropolitan testbed [18]: This testbed demonstrates the deployment of OAI-EPC as a VNF on a cloud environment, and it presents the LTE CN service instantiation via OSM. The goals of this implementation are to produce MEC services to EPC and to integrate EPC with the extended eNB software. The functionality of the testbed is evaluated for establishing TCP and SCTP connections for downloading from server to UE, uploading from UE to a server, and bidirectional communication between UE and server.

As a summary of Table I, we can classify the presented testbeds into two main categories. The first category includes those testbeds which provide some capabilities in network slicing; however, they do not attain to all of the design criteria for realization of E2E network slicing. While [6], [7], [10], [13], [15], [17], [18] focus on one specific network domain and provide slicing just for that particular domain, [11] provides E2E slicing but does not have a separate entity for management and orchestration. Papers [8], [16] offer only a light MANO implementation. The solutions in [8], [11], [16] deploy E2E network slicing with MANO capability but without multi-RATs or multi-tenancy support. The testbeds in [9], [12], [14] belong to the second category, which complies with all of the designing principles. Nevertheless, 5GIIK offers more features and capabilities such as slice provision dynamicity, real-time monitoring of VMs and VNF-onboarding to different VIMs, which differentiate it from other testbeds in the second category. Section IV describes the 5GIIK features.

III. 5GIIK - OUR PROPOSED TESTBED

We consider all features elaborated in the previous section and propose 5GIIK - a testbed architecture that grants an E2E network slicing with MANO capability, that supports multi-tenancy and multi-RATs and at the same time it is a cost-efficient design. In this section, we first present the network architecture with all associated components, and then we explain the network slice creation and instantiation in 5GIIK.

A. 5GIIK testbed architecture

A high-level description of the 5GIIK testbed is given in Figure 2. It is composed of several entities that emulate real 4G/5G networks. The RAN and CN parts of the testbed are implemented in Trondheim and Gjøvik campuses of NTNU, respectively. The IP backbone network, which is provided by Norway's National Research and Education Network (UNINETT), is used as TN in our platform. Our testbed virtualizes not only the CN but also the Base Band Unit (BBU) of RAN into the cloud to build a Cloud-RAN (C-RAN) architecture. The remote radio head section of the C-RAN is managed by a Software Defined Radio (SDR) and connected antennas to the SDR. The architecture, illustrated in Figure 3, is partially motivated by the work in [1], [6] and [14]. To simplify, the connections of NFV MANO reference

points with external entities are not depicted. Next, we evaluate the different open-source solutions for modules in the NFV MANO framework in Figure I and explain our reasoning behind the selected solution for the proposed architecture in Figure 3.

- *VNFs:* OAI is a flexible solution for emulating LTE systems and implements the full protocol stack of 3GPP standard in Evolved-UMTS Terrestrial Radio Access Network (E-UTRAN) and EPC. OAI can be used to build a complete LTE network (eNB, EPC and UE) on a PC or Virtual Machine (VM). NextEPC[1], as its name indicates, implements just the CN of a 4G/5G system. Open5GS[1] provides a complete implementation of a 4G/5G. However, the lack of detailed documentation about its specifications is a repelling point for choosing it. The srsLTE solution emulates the whole system and has a well-structured code for future improvements. The srsLTE library is modular and utilizes single instruction multiple data operations for increasing its performance in the system. It has a light implementation regarding the CN. From a hardware perspective, the srsLTE library can operate with various front end RFs and it is able to provide interfaces for different types of Ettus USRP pieces of equipment.
- *NFV orchestrator:* There are various solutions, but the main competition is between ONAP[1] and OSM. Considering the compatibility of these orchestrators with different VIMs, OSM supports several VIMs and can manage them at the same time. Regarding resource usage (CPU and memory), again, OSM bests ONAP by utilizing fewer resources compared to ONAP [20].
- *VIM:* There are solutions such as OpenStack, VMware vCloud Director[1] and Amazon Web Services[1]. OpenStack is more potent than others for infrastructure orchestration, scalability and resource utilization. Besides, OpenStack can be deployed on standard machines with the appropriate amount of resources.

Following this elaboration, 5GIIK testbed uses OAI as CN and srsLTE as RAN. The MANO entity in 5GIIK is represented by OSM, which is developed in Python and operates on Linux. OSM combines NFVO, VNFM of the NFV MANO architecture. As a result, configuration and abstraction of VNFs, orchestration, and the management of the network services are feasible. Since release 4, OSM uses Docker container technology and cloud-based solutions. 5GIIK also integrates SDN controllers to its architecture. SDN-based Tenant Controller (TC) is needed to provide L2 VLANs to manage tenant VNFs located in different VIMs while implementing network slicing in TN. We integrate two TCs for the whole network domains to make our design more generic.

1) 5G-EmPOWER as TC for RAN domain: 5G-EmPOWER controller [19], also known as EmPOWER, is an open multi-access network operating system. It is created upon a single platform that consists of general-purpose hardware (x86) and Linux. 5G-EmPOWER is RAT-agnostic and by using a data plane programmability policy, it manages the virtualized network resources of multiple radio nodes (WiFi Access Points, LTE eNBs and 5G NR). It communicates with the eNB in the C-RAN to manage and control the radio resource allocation to the end-users. Consequently, it supports multiple virtual networks (tenants) on top of the same physical infrastructure. The latest version of 5G-EmPOWER is compatible with srsLTE eNB, and it fits well in our testbed.

[1]https://nextepc.org/, https://open5gs.org/, https://www.onap.org/, https://www.vmware.com, https://aws.amazon.com/, https://opencord.org/, https://www.openstack.org/, https://www.softwareradiosystems.com/, https://osm.etsi.org/, https://www.opennetworking.org/onos/

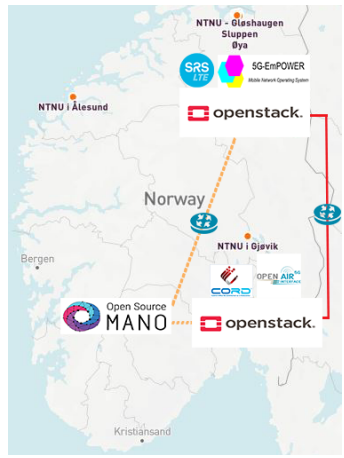


Fig. 2. The used open-source solutions mapped to testbed premises in Trondheim and Gjøvik.

2) M-CORD as TC for TN and CN domains: M-CORD[1] is a cloud-based solution built on SDN, NFV technologies. It encompasses both virtualization of RAN functions (vRAN) and a virtualized CN (vEPC) to allow mobile edge applications and services using a micro-service architecture. M-CORD disaggregates and virtualizes network functions and operator services. By integrating the ONOS[1] controller in its architecture, M-CORD implements network slicing in TN in order to form an E2E network slice.

For deploying E2E network service orchestration, OSM interacts with VIMs via Or-Vi interfaces. OSM performs lifecycle management of NF configuration, operation and monitoring by interacting with Physical/Virtual NFs (EPC, BBU) via charm configuration files. The infrastructure management section of NFV MANO is divided into VIM and Infrastructure SDN based Controller (IC), represented in both OpenStack 1 and 2 in Figure 3. VIM, in cooperation with IC, manages and controls the infrastructure layer, both physical and virtual resources, via Nf-Vi.

B. Network slice instantiation in 5GIIK testbed

Now we describe how a network slice is instantiated by the OSM in the 5GIIK testbed, and the procedure is illustrated button-up in Figure 4.

VNF Descriptors (VNFDs) are located in the first level of creating network slices. VNFD is a file that retains the

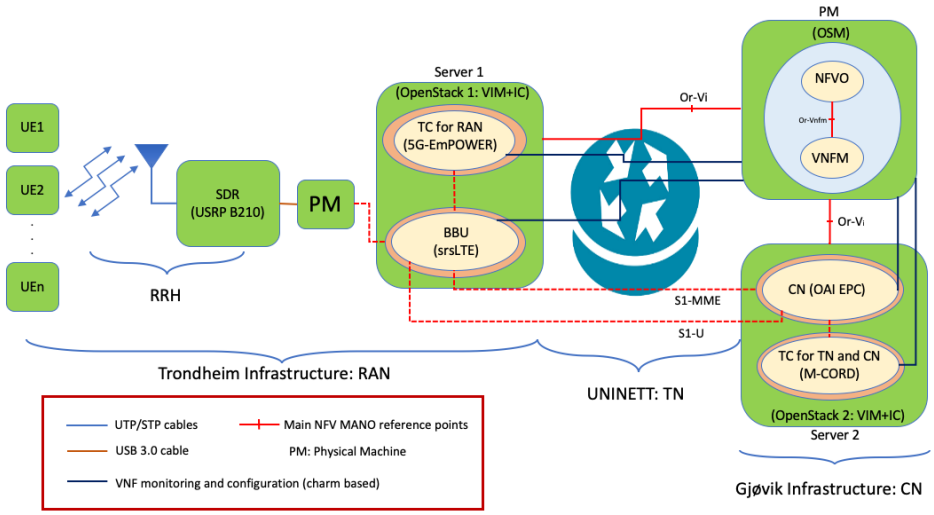


Fig. 3. 5GIIK testbed architecture.

information such as the software image that the VNF needs to be built on as well as CPU, memory and storage that the VNF needs for high-performance operation, internal virtual links (vls) between Virtualization Deployment Units (VDUs) inside a VNF and a lifecycle event of a network slice. A management network (mgmt) is needed to assign IP addresses to the launched VDUs. In this level, two VNFDs are required; OAI EPC-VNFD and srsLTE eNB-VNFD. OAI EPC contains four entities: Home Subscriber Server (HSS), Mobility Management Entity (MME), Control Plane of the Service Packet Gateway (SPGW-C) and User Plane of the Service Packet Gateway (SPGW-U). Thus, OAI EPC-VNF includes four VDUs, one for each entity, while srsLTE eNB-VNF includes one VDU in the context of OSM.

Network Service Descriptors (NSDs) are positioned in the second level of the slice formation. NSD comprises different launched VDUs that a network service needs to operate. NSD also includes external connection points (cps) between the demanded VDUs. The same as the VNFD level, two NSDs are essential here; OAI-EPC-NSD and srsLTE eNB-NSD.

Finally, in the third level of a network slice creation, a Network Slice Instance Descriptor (NSID) chains the established service instances and forms a network slice. In this level, one NSID is needed to chain the launched services in the CN and RAN domains. To summarize, these are the steps to instantiate an E2E network slice:

1) At the VNFD level, OSM instantiates the required VDUs (VMs) via its resource orchestrator block to the VIM. In this level, the necessary resources are allocated, and specific interfaces are configured on each VDU.
2) At the NSD level, according to the launched VMs, particular service instance(s) is (are) created.
3) At the NSID, the service instances are chained and create an E2E network slice.

Apart from the crucial information that has to be defined on each descriptor level, some customized information can be set as well. For instance, it is possible to define some metric parameters for running VNFs to be collected via VIM. In this way, performing periodic network monitoring at the infrastructure level is achieved.

IV. 5GIIK Features and Initial Experiments

The proposed testbed provides a whole range of features that can be exploited for developing various new solutions for network slicing in wireless and mobile networks.

- 5GIIK is cross-domain and it spans over the whole network in order to support E2E slicing.
- The SDN functionality in 5GIIK enables studies of numerous use cases and scenarios for new resource allocation techniques. In particular, 5G-EmPOWER supports common machine learning toolkits, which is a missing capability for most of the testbeds mentioned in Table I. Hence, slice-aware traffic marking strategies can provide dynamicity in slice provision for different use cases, and it can assign the available radio resources to the end-users in an optimized fashion.
- 5GIIK allows multi-RAT implementation. The recent containerized-based 5G-EmPOWER release (under an APACHE 2.0 License) is compatible with WiFi access points and srsLTE (release 19.09) to perform RAN slicing for both of these RATs. As a result, RAN slicing for both LTE and WLAN is applicable.
- Multi-tenancy that is fundamental in the 5G era, especially in the RAN domain, is supported in our architecture. In particular, 5G-EmPOWER grants 5GIIK to create two different tenants via its web interface. By defining a Mobile Network Operator (MNO) such as PLMN-ID=A, radio resources can be shared among two Mobile Virtual Network Operators

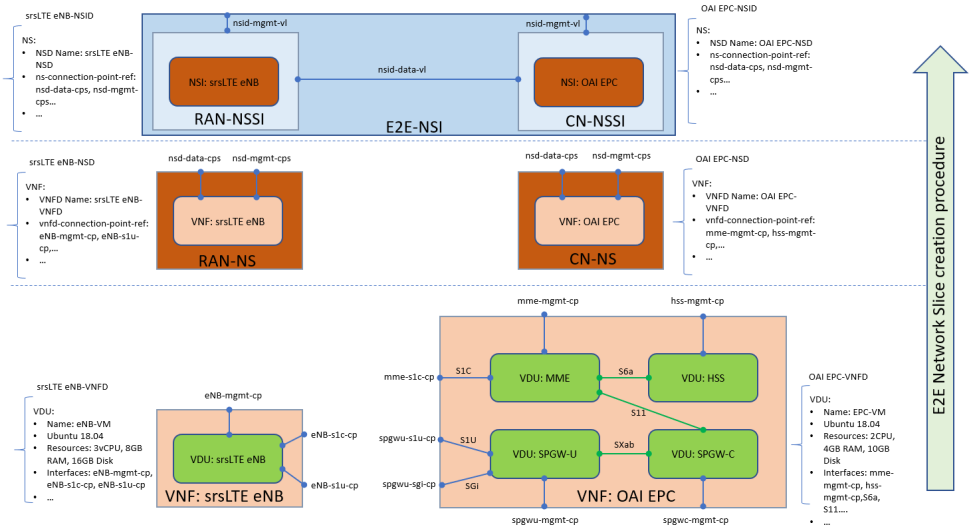


Fig. 4. Procedure of E2E network slice creation.

(MVNOs) so-called foo and bar. Each of the foo and the bar has its own unique MVNO-identifier. Then, on each of these MVNOs, one or several network slices can be created. Consequently, one user equipment can be configured in such a way to connect to the desired slice on a particular MVNO. Figure 5 demonstrates two scenarios in which there are one and two MVNOs with their slices, respectively.

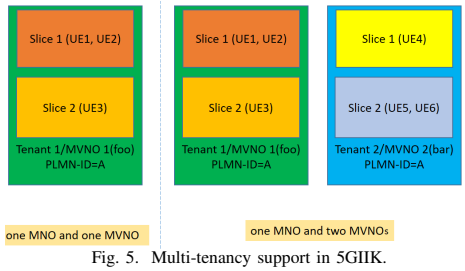


Fig. 5. Multi-tenancy support in 5GIIK.

- OSM and OpenStack perform management and orchestration in 5GIIK. Besides, OSM can manage multiple-VIMs, since our testbed is implemented on two OpenStack infrastructures (RAN in Trondheim and CN in Gjøvik).
- OSM module in 5GIIK offers E2E network slice provisioning. Firstly, VNFDs specify the desired images with their demanded resources (CPU, memory, and storage) via OpenStack. Secondly, based on the NSDs, specific service instances are created. Finally, the NSID determines how to chain these service instances to create an E2E network slice that traverses the whole network domains.
- OSM provides the possibility to perform the VNF-onboarding process. In VNF-onboarding, the VNF lifecycle has three phases – so-called days. In day-0, management policies for VNFs' instantiation are established. In day-1, VNFs are configured, and they can provide the demanded services. In day-2, the possibilities of reconfiguring VNFs and monitoring their Key Performance Indicators (KPIs) in runtime operation are granted. Hence, the onboarding process can be done for a variety of VNFs to build favorable VNF packages on OpenStack.
- Prometheus is a system monitoring toolkit that can be integrated with OSM in our testbed. Specific metrics such as CPU utilization and average memory utilization can be defined at VNF descriptors. Consequently, Prometheus retrieves the collected metrics and performs real-time monitoring of all active/detective VMs.

Implementation challenges: Deploying C-RAN is a challenge since the communication between BBU and RRH demands very low latency, and it is essential to implement BBU close to RRH. It is even more challenging when the network delivers delay-sensitive services with ultra-low latency requirements. Resource management in VIM(s) is considered as another challenge, mainly when a VIM is not capable of launching instances (VMs that are running VNFs) with a high amount of assigned resources in terms of CPU and memory. It is crucial to know how to allocate available resources to multiple instances in a VIM.

A. Initial testing

5GIIK is installed on three similar Intel machines (i7-4790 CPU @ 3.60GHz, 32GB RAM), which are running two OpenStack platforms and OSM. In order to evaluate the testbed performance regarding CPU and memory usage, we carried out one initial test. The test involves transferring and downloading files with different sizes (500 MB and 1 GB) from one VM in one OpenStack to another VM in the second OpenStack. First, one VNF descriptor is created on the OSM to define the required image with its demanded resources (CPU, memory and storage). Furthermore, the management

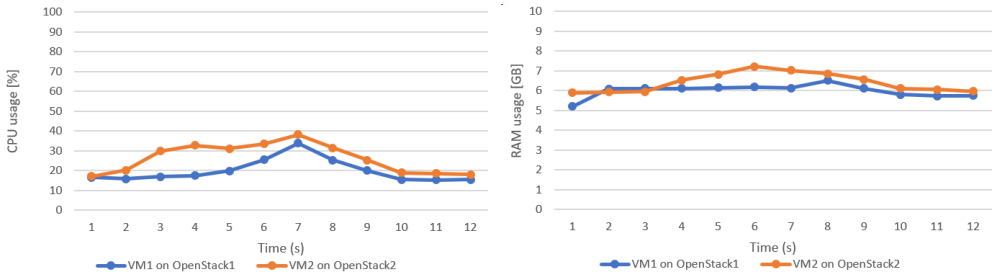


Fig. 6. CPU and RAM usage during a transfer of two files from one VM to the other in 5GIIK.

network has to be set to assign an IP address to the service instance. Subsequently, one service descriptor is created to determine how the base image can launch the service instance. Considering the VNF and service descriptors, two similar service instances (two VMs running Ubuntu 16.04 with one virtual CPU, 16 GB of RAM and 20 GB of storage) are launched on the two VIMs. Figure 6 illustrates the CPU and RAM measurement. VM1 downloads the smaller file and then starts sending it to the VM2. A similar approach for VM2 takes place but with the larger file. As expected, VM1 utilizes less amount of resources (up to 33.9% of CPU and 6,5 GB of RAM) compared to the VM2 (up to 36.2% of CPU and 7,2 GB of RAM).

V. Conclusion

Following the modern development of tools and technologies that enable network slicing, we composed a list of several design criteria for constructing testbeds. Then we summarized some small-scale testbeds with their main features analyzed through the criteria list that we composed. We also proposed 5GIIK - a testbed that performs E2E network slicing with the capability of management and orchestration of network resources. 5GIIK is an open-source-based architecture and its flexibility provides the opportunity to create innovative algorithms, patterns and solutions in the network slicing realm.

References

[1] 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 Comm. Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

[2] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Comm Surv Tut*, vol. 20, pp. 2429–2453, 2018.

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

[4] ETSI, "Network functions virtualisation (nfv): Architectural framework," *ETSI Gs NFV*, vol. 2, no. 2, 2013.

[5] D. Gligoroski and K. Kralevska, "Expanded combinatorial designs as tool to model network slicing in 5g," *IEEE Access*, vol. 7, pp. 54 879–54 887, 2019.

[6] B. Dzogovic, V. T. Do, B. Feng, and T. van Do, "Building virtualized 5G networks using open source software," in *IEEE Sym. on Comp. App. Ind. Elec. (ISCAIE)*, 2018, pp. 360–366.

[7] M. Arif, A. Pouttu, O. Liinamaa, M. Ylianttila, and I. Ahmad, "On the Demonstration and Evaluation of Service- Based Slices in 5G Test Network using NFV," Apr. 2019.

[8] P. Mekikis *et al.*, "NFV-enabled Experimental Platform for 5G Tactile Internet Support in Industrial Environments," *IEEE Transactions on Industrial Informatics*, p. 1, 2019.

[9] N. Nikaein, C.-Y. Chang, and K. Alexandris, "Mosaic5G: Agile and Flexible Service Platforms for 5G Research," *SIGCOMM Comput. Commun. Rev.*, vol. 48, pp. 29–34, 2019.

[10] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN Slicing for a Flexible and Cost-Effective Multi-Service Mobile Network Architecture," in *Proc. of the 23rd Annual Int. Conf. on Mobile Computing and Networking*, 2017, pp. 127–140.

[11] A. Shorov, "5G Testbed Development for Network Slicing Evaluation," in *IEEE Conf. of Russian Young Researchers in Electrical and Electronic Engineering*, 2019, pp. 39–44.

[12] G. Garcia-Aviles, M. Gramaglia, P. Serrano, and A. Banchs, "POSENS: A Practical Open Source Solution for End-to-End Network Slicing," *IEEE Wireless Comm.*, vol. 25, no. 5, pp. 30–37, 2018.

[13] K. Koutlia *et al.*, "Design and Experimental Validation of a Software-Defined Radio Access Network Testbed with Slicing Support," *Wireless Comm. and Mob. Comp.*, pp. 1–17, Jun. 2019.

[14] C. Huang, C. Ho, N. Nikaein, and R. Cheng, "Design and Prototype of A Virtualized 5G Infrastructure Supporting Network Slicing," in *IEEE 23rd Int. Conf. on Digital Signal Processing (DSP)*, 2018, pp. 1–5.

[15] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar, "Dynamic Network Slicing for 5G IoT and eMBB services: A New Design with Prototype and Implementation Results," in *3rd Cloudification of the Internet of Things (CIoT)*, 2018, pp. 1–7.

[16] P. K. Chartsias *et al.*, "SDN/NFV-based end to end network slicing for 5G multi-tenant networks," in *2017 European Conf. on Netw. and Comm. (EuCNC)*, 2017, pp. 1–5.

[17] J. Kim and M. Xie, "A study of slice-aware service assurance for network function virtualization," in *2019 IEEE Conference on Network Softwarization (NetSoft)*, Jun. 2019, pp. 489–497.

[18] T. Dreibholz, "Flexible 4g/5g testbed setup for mobile edge computing using openairinterface and open source mano," in *Web, Artificial Intelligence and Network Applications*, Springer International Publishing, 2020, pp. 1143–1153.

[19] E. Coronado, S. N. Khan, and R. Riggio, "5G-EmPOWER: A Software-Defined Networking Platform for 5G Radio Access Networks," *IEEE Trans. Network and Service Mngm.*, vol. 16, no. 2, pp. 715–728, 2019.

[20] G. M. Yilma, F. Z. Yousaf, V. Sciancalepore, and X. P. Costa, "On the challenges and kpis for benchmarking open-source NFV MANO systems: OSM vs ONAP," *ArXiv*, vol. abs/1904.10697, 2019.

Paper III

S. Kielland, A. Esmaeily, K. Kralevska, and D. Gligoroski, "Secure Service Implementation with Slice Isolation and WireGuard," 2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), Athens, Greece, 2022, pp. 148-153, doi: 10.1109/MeditCom55741.2022.9928730.[2]

Secure Service Implementation with Slice Isolation and WireGuard

Sondre Kielland, Ali Esmaeily, Katina Kralevska, and Danilo Gligoroski
Department of Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU)
Email: {sondrki, ali.esmaeily, katinak, danilo.gligoroski}@ntnu.no

***Abstract*—Network slicing enables the provision of services for different verticals over a shared infrastructure. Nevertheless, security is still one of the main challenges when sharing resources. In this paper, we study how WireGuard can provide an encrypted Virtual Private Network (VPN) tunnel as a service between network functions in 5G setting. The open source management and orchestration entity deploys and orchestrates the network functions into network services and slices. We create multiple scenarios emulating a real-life cellular network deploying VPN-as-a-Service between the different network functions to secure and isolate network slices. The performance measurements demonstrate from 0.8 Gbps to 2.5 Gbps throughput and below 1ms delay between network functions using WireGuard. The performance evaluation results are aligned with 5G key performance indicators, making WireGuard suited to provide security in slice isolation in future generations of cellular networks.**

***Index Terms*—OSM, WireGuard, VPN, NFV, 5G, Network slice, URLLC, eMBB.**

I. Introduction

The enrollment of 5G non-standalone cellular networks is already in operation by mobile network operators. In developing Beyond 5G (B5G) networks, several planned functionalities will enable verticals to establish their services with diverse Quality-of-Service (QoS) requirements on shared physical infrastructure. Providing End-to-End (E2E) services over isolated network slices is a key factor to empower multiple services on a shared infrastructure. To develop agile B5G networks for supporting applications with different QoS requirements, Network Function Virtualisation (NFV), Software-Defined Networking (SDN) and Multi-Access Edge Computing (MEC) are the main enabling technologies [1], [2].

An NFV Management and Orchestration (MANO) entity connected to one or several Virtual Infrastructure Managers (VIMs) controls and monitors the deployment of Network Services (NSs) by deploying necessary infrastructure resources. For an agile network deployment, the NFV MANO also administrates connections between Virtual Network Functions (VNFs), including creation of virtual networks with the help of SDN. Therefore, instead of manually creating and connecting the NSs together, the NFV MANO helps operators to deploy and control Network Functions (NFs) automatically. With its automatic and reusable functionality, a large number of NFs and NSs can be rapidly deployed on a single or multiple VIMs.

Cloud infrastructures that can be rented or shared are necessary to utilize resources efficiently for financial and load distribution purposes. Introducing shared infrastructure raises further security challenges. Securing application data transfer over shared networks is one example of such a security challenge. A countermeasure that can be initiated against such security concerns is operating Virtual Private Network (VPN) between NFs. However, establishing VPN tunnels introduces additional overhead. For services dependent on low latency or high throughput, the additional overhead may affect their service performance.

NFV MANO can provide traffic isolation for NFs in NSs by deploying VPN tunneling between NFs and interconnecting them [3]. In this way, the secure tunneling isolates Network Slice Instances (NSIs) and the provided NSs via the NSIs. Nevertheless, this approach is only feasible if the VPN does not introduce significant overhead violating QoS requirements. The deployment of VPN between VNFs in an automatic mode in order to provide security isolation between slices and the effect of the introduced overhead on the performance isolation among slices in a shared environment are still open research questions.

In this paper, we implement and analyze the performance of WireGuard for providing slice isolation in 5G environment. WireGuard [4] is a straightforward yet immediate VPN solution that functions via the Linux kernel and employs state-of-the-art cryptography approaches. Open Source MANO (OSM) orchestrates NSs and NSIs, and establishes VPN tunnels between the VNFs. The integrated WireGuard-OSM architecture provides: 1) secure communication between the involved VNFs of NSs and NSIs - slice isolation; 2) performance isolation between the slices. The performance analysis shows that the integrated WireGuard-OSM architecture meets the required Key Performance Indicator (KPI) values in terms of high throughput for enhanced Mobile Broadband (eMBB) slices and low latency for Ultra Reliable Low Latency Communication (URLLC) slices. We make the code publicly available[1] to the research community.

The remainder of this paper is organized as follows. Section II provides a literature overview of practical approaches for secure isolation between slices. Section III presents the system architecture. The implementation steps are explained in Section IV. The performance evaluation results are presented in Section V. Finally, Section VI concludes the paper.

[1]https://github.com/sondrki/TTM4905/

978-1-6654-9825-8/22/$31.00 ©2022 IEEE

II. Related Work

The isolation concept between network slices can be studied from security, performance, and dependability aspects [5]. In addition, the Confidentiality, Integrity and Availability (CIA) triad is a widely used way of looking at different security aspects. A shared infrastructure introduces security challenges in all dimensions of the CIA triad. The key feature of shared infrastructures is that an attack on or from another party sharing the infrastructure should not affect the other sharing parties. This definition of CIA is harmonic with the isolation concept in network slicing. Other parties should also be unaffected when it comes to performance and dependability, extending the availability dimension. The workload, the number of resources, and hardware or software failure of another NS should not reduce the performance of an NF in a separate NS or NSI.

While 5G intends to fix some security issues present in the previous generations of cellular networks, it also introduces several new security threats. Some of them are raised by providing services via network slices. Paper [6] explores and classifies different security challenges of 5G networks. Proper isolation of logical resources is essential to avoid introducing several new risks. Eavesdropping and tampering with data, for instance, are two vectors an attacker could use to interfere with security if the application data is not properly encrypted. Hantouti et al. suggest that operators should deploy encrypted tunnels as a way to establish trust between Service Functions (SFs) to provide packet integrity and prevent bypassing of policies [7].

The work in [8] proposes a novel mutual authentication and key establishment protocol utilizing proxy re-encryption. The protocol grants specific authentication between components of a network slice to enable secure connection for protected key establishment among component pairs for slice security isolation. Paper [9] offers a secure keying scheme by adopting a multi-party computation strategy, which is appropriate for network slicing architecture in the case that third-party applications access the slices. This mechanism ensures the satisfaction of use cases or devices in which the data is collected.

Both Haga et al. in [10] and Vidal et al. in [11] focus on how a VPN can be deployed using OSM. Reference [10] demonstrates how WireGuard can be added in VNFs and compares the performance of WireGuard and OpenVPN. This proof-of-concept is carried out using two VNFs in a single NS with manual configuration of peer connectivity in WireGuard. For the peer setup, keys and other necessary information are obtained manually. Vidal et al. in [11] uses IPsec as VPN solution to provide link-layer connectivity for multi-site deployments. In this work, OSM deploys multiple NSs connected through one VNF at each NFV Infrastructure (NFVI). These VNFs handle the link layer abstraction for the other VNFs. IPsec is used to secure the connection between the link layer providing VNFs. Keys and connection parameters are supplied by the operator when instantiating the NSI.

To the best of our knowledge, none of the state-of-the-art works presents a secure service automation provisioning utilizing complex and real-life NFs. This motivates us to integrate WireGuard tunneling with OSM, which grants secure communication between NFs in order to establish automated and realistic network services. As a result, this system architecture guarantees security and performance isolation between NSIs.

III. System Architecture

Day-0, Day-1, and Day-2 operations are terminologies used in the OSM community referring to the stages of Life-Cycle Management (LCM) of NFs. The steps in Figure 1 are used to handle LCM of NFs via the NF onboarding process and they link closely to Day-0 to Day-2 operations. In Figure 1,

- Day-0 phase focuses on necessary instantiation, including charms and descriptor creation/editing, validation, packaging, and emulation;
- Day-1 phase concentrates on service initialization containing test, release, and deploy;
- Day-2 phase covers runtime actions comprising operate and monitor steps.

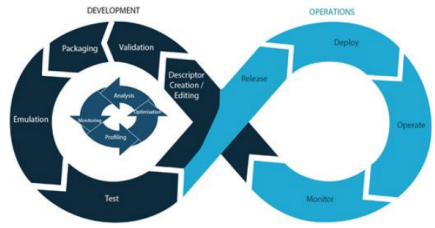


Fig. 1. Steps for service lifecycle [12].

OSM has three inbuilt supporting applications for LCM [13]. Cloud-init is responsible for the initial Day-0 operations like setting username and password. For Day-1 operations, Helm charts or Juju charms can be used, while Day-2 operations are also possible with Juju. The difference between Helm and Juju is that Helm is used solely for Kubernetes-based Network Functions (KNFs), while Juju is also usable at NS level and for VNFs that are not Kubernetes (K8s) based [14], [15]. We have used cloud-init and Juju charms for OSM onboarding in our implementation.

Further, Juju has two operation modes: native and proxy. Native charms run operations directly inside a VNF. On the other hand, proxy charms use a centrally placed controller, VNF Configuration and Abstraction (VCA), to manage the Day-1 and Day-2 actions. The VCA connects to the VNFs through their management interface and instructs the VNFs. The VCA-VNF connection uses the Secure Shell (SSH) protocol by default. In the paper, we have used proxy charms with a VCA installed co-located and integrated with OSM. Both the VCA and OSM are, therefore, able to access the VNFs management interface to execute their actions.

To build user-defined actions, Juju uses Python scripts. The connection to the OSM instance is made through the description files of the VNFs, NSs, Juju config files describing metadata, and the available Day-1 and Day-2 actions. For the OSM integration of proxy charms, the *charms.osm.sshproxy* library is provided by OSM to take care of, among other tasks, the basic Juju proxy peer setup.

In addition to running actions in VNFs, Juju can be used to create relations between Juju units for management, scaling, and handling dependencies across VNFs. We use Juju relations to transfer WireGuard peer information between VNFs.

Figure 2 illustrates how we use proxy charms and relations in Juju to create a bridge for transferring information between VNFs. The figure shows the architecture for the multi-site demonstration. Note that we used a single-VIM, moving the Home Subscriber Server (HSS) into *VIM 1*, for the performance evaluation results presented in Section IV. The architecture for the single-VIM setup is as illustrated in the rightmost half of the figure showing *VIM 1*.

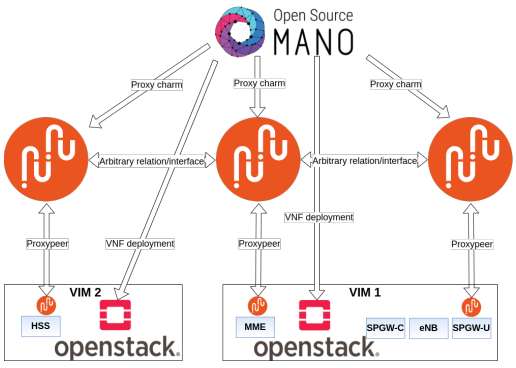


Fig. 2. Interactions between elements in our Juju proxy implementation.

Key distribution is a task that often requires manual steps when establishing a VPN tunnel. Manual setup can be time-consuming for dynamic environments or environments with many interfaces that need to be secured. If the tenant manager needs to do configuration, the NS is only usable after initializing the VPN tunnels. However, if we apply the approach presented by Vidal et al. [11] and input the necessary information, including keys, the application can start sending data immediately after Day-1 actions have finished. Using a Key Management System (KMS) is a similar approach. However, OSM does not provide such functionality. To use the KMS approach additional functionality outside of the OSM framework must be added.

To perform key management, we use a non-standard approach using Juju relations with the requirement of using proxy charms for our VNFs. By using Juju relations, we create new individual keys for every new deployment of an interface and make the application of the NS usable directly after the Day-1 tasks finish. Furthermore, with our approach, the private keys are only stored inside the VNFs. The public key and other necessary information for the peer setup get automatically transferred to the peer.

IV. Implementation

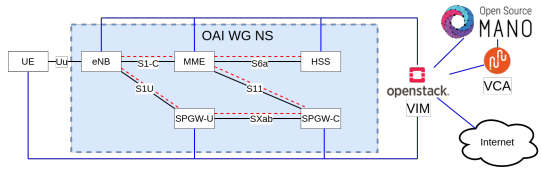


Fig. 3. Architecture of our implementation.

To implement WireGuard in a realistic 5G environment we created a NS with Evolved Packet System (EPS) components from OpenAirInterface (OAI) [16]. We then added WireGuard connectivity on the different interfaces. Figure 3 shows the deployed architecture. OSM is used to communicate with MicroStack VIM [17]. The VIM hosts different VNFs, creates virtual networks and performs routing of outgoing traffic from the VNFs, represented by solid blue lines. WireGuard tunnel is created automatically between the VNFs on the interfaces in the NS, represented by the red dotted lines. In addition to the primary VIM, we utilized a second VIM in order to explore the EPS NS deployment in multiple sites.

A. Development

We followed these steps to prepare the deployments: 1) compose a virtualized EPS, 2) set up a mechanism for automatic WireGuard peering, 3) structure NSs into Network Slice Template (NST), and lastly, 4) test the WireGuard connectivity in a multi-site deployment. The code for the descriptors and charms is publicly available on GitHub. In the following paragraphs, we further describe the development steps for creating the descriptors and the scripts.

1. Composing a Virtualized EPS: In [18], Dreibholz implements an Evolved Packet Core (EPC) with HSS, Mobility Management Entity (MME), and a combined Serving Gateway (SGW) and Packet Data Network Gateway (PGW) separated in two components, Service Packet Gateway-User plane (SPGW-U) and Service Packet Gateway-Control plane (SPGW-C), for user- and control-plane tasks, respectively. To extend this NS with real-life traffic, we add a virtualized eNodeB (eNB). Further, we create an User Equipment (UE) in a Virtual Machine (VM) kept outside the NS. The UE is still able to connect to the eNB after instantiating the NS with manual network setup in MicroStack. To establish the air interface, Uu, we have compiled and used OAI simulation option. When connecting the UE to the eNB, we verify that the different EPS components function as expected and provide service to the UE. The UE connects to an outer network through the SPGW-U via the eNB. At this first step of implementation, we still have not included WireGuard between the components.

We chose to build the NS by spreading the EPS components into separate VNFs. This approach allows to split the VNFs

in the VIMs. Extending it to a multi-site environment gives us the opportunity to emulate a scenario where other components, for instance, MEC, are deployed closer to the end-users. The VNFs distributed to remote sites are able to communicate with the core securely with the help of WireGuard.

2. Automatic WireGuard Peering: Manually setting up VPN tunnels between several interfaces can be time-consuming. Thus, we use Juju relations for automatic peering with no extra information given to the other end of the peer at the time of instantiating the NS. The first step in the automatic peering is the establishment of relationships between VNFs on both sides. Then the paired VNFs retrieve information like *public key*, *endpoint*, and *listening port* to communicate with each other. Wireguard usually employs the following cryptographic primitives: elliptic Curve25519 for key exchange, then HKDF for the key derivation, and finally, the bulk encryption work is performed by the symmetric primitive ChaCha20Poly1305 for Authenticated Encryption with Associated Data (AEAD) [4]. All of these primitives have excellent performance in software supporting the objective of NFV. Moreover, due to the lack of considerable overhead and latency, and remarkable efficiency, ChaCha20Poly1305 AEAD performs significantly in terms of ping time and throughput for the URLLC and eMBB slices, respectively.

To establish WireGuard connectivity on all interfaces given in Figure 3, we changed the IP address configuration in the components. Changing the interface addresses is necessary to route application data over the VPN tunnel and, at the same time to ensure that applications inside the VNF have been installed and started correctly even when waiting for the tunnel establishment. Besides, to verify that the NS runs WireGuard, we connect the UE and observe that it connects and gets Packet Data Network (PDN) service.

Further, in order to observe how resources affect the WireGuard performance, we have prepared a copy of the EPS NS with WireGuard connectivity with doubled resources.

3. NST creation: After having a working NS with WireGuard connectivity between the interfaces, we include it in two NSTs to observe if and how the performance is affected by providing security with WireGuard. The two NSTs have different values of quality indicators corresponding to different 5G QoS Identifiers (5QIs) [19]. The QoS parameters correspond to eMBB and URLLC use-cases, respectively. Further, the NST is prepared with only the management interfaces of the VNFs. The management interfaces are attached to the external connection points in the NSTs.

4. Multi-site deployment: To verify that the automatic peering setup also works in a multi-site environment, we have separated the HSS VNF to a second VIM. When using OpenStack/MicroStack, the external floating IP address is by default not known inside a VM. However, the VCA can retrieve the management IP address to perform its actions. To find the floating IP addresses of the VNFs, we use the same function that Juju employs for its *proxypeer* connection between a Juju unit at the VCA and the Virtual Deployment Unit (VDU) in the VNF. After the endpoint IP address is found, the MME and HSS connect automatically with WireGuard connectivity. A prerequisite for multi-site WireGuard connectivity is to use a port opened in the firewalls.

B. Proof-of-Concept for VPN-as-a-Service

With the automatic peering, we presented a few steps to add WireGuard as a VPN-as-a-Service (VPNaaS). Here we summarize all steps to build the proof-of-concept.

1) Append installation of WireGuard in cloud-init.
2) Add name and parameters for Day-1 and Day-2 actions in the actions.yaml file.
3) Add relations between VNFs in the metadata.yaml file.
4) Include the Python code to append the charm script. The name of the relationship must correspond between the name used in metadata.yaml and the listener in the *__init__* function of the Python script.
5) Add the actions from actions.yaml into Day-1, Day-2 operations in the VNF Descriptor (VNFD). To create the WireGuard tunnel as a Day-1 operation, the relevant actions should be included in the *initial-config-primitive* section in the VNFDs. Day-2 actions are placed in the *config-primitive* section.
6) While the default implementation sets up the VPN, Day-2 actions can be used for further configuration and maintenance, for instance, if a new connection should be added towards a NF.

V. Performance Evaluation

To assess the performance of WireGuard in the 5G network, we conducted measurement tests in both the control and user plane, with and without WireGuard capability. We utilized both arbitrary data and the UE to generate realistic traffic in the network. We observe the impact of integrating secure communication with Wireguard on the performance metrics that should be aligned with the 5G KPI [20].

While producing arbitrary data for high network load, we measure the latency and Service Response Time (SRT) in the control plane, combining multiple EPS components. In general, the following tasks are done to test the performance of NSs and NSIs:

- Observe SRT on the MME when the UE connects;
- Observe throughput and latency in the user plane with the UE over S1-U interface;
- Measure throughput and latency between components in the EPS in the control plane over S1-C and S6a interfaces.

A. Lab Environment

The primary VIM is a server running MicroStack with resources of 56 vCPUs, 126 GB RAM, and 915 GB storage. The second VIM, used for multi-site deployment, also runs MicroStack but has fewer resources with a total of 9 vCPUs, 32 GB RAM, and 150 GB storage. For the EPS NS a total of 14 vCPU, 27 GB RAM and 110 GB storage are utilized. According to the limiting ISP, the bandwidth between the two NFVIs is specified to be 200 Mbps. For the VNFs to communicate securely across the VIMs, WireGuard tunnel is established

between the NFVIs. Our measurement shows a throughput of approximately 180 Mbps between the MicroStack instances. A nested WireGuard tunnel is used when adding WireGuard on the S6a interface for the multi-site deployment. The internal throughput of the NFVI where the primary VIM runs is 20 Gbps. Table I gives a summary of the resources used for the VNFs.

TABLE I
VNF INFORMATION OF THE OAI EPS NS.

VNF name	Operating System	Number of virtual CPUs	Amount of RAM (GB)	Storage (GB)
HSS	Ubuntu18.04	4	8.0	20
MME	Ubuntu18.04	2	4.0	20
SPGWU	Ubuntu18.04	1	3.0	20
SPGWC	Ubuntu18.04	3	4.0	30
eNB	Ubuntu18.04	4	8.0	20
UE	Ubuntu18.04	2	4.0	20

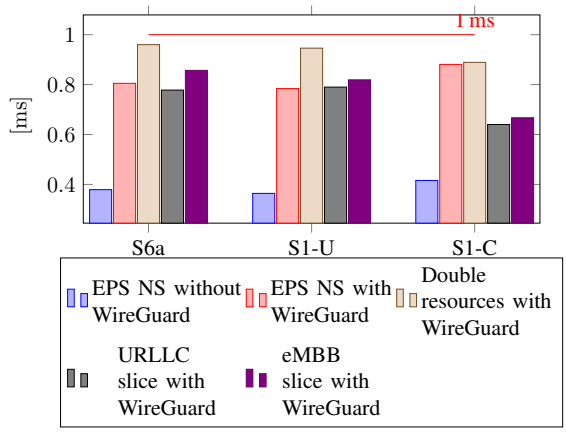


Fig. 4. Latency comparison for different interfaces with and without WireGuard functionality.

B. Observations

Before adding the VPN tunnels, we are able to capture connection information such as the International Mobile Subscriber Identity (IMSI), network realms, and hostnames at the VIM. However, after we introduce WireGuard, the only information observable at the VIM is the use of the WireGuard protocol and link-layer discovery messages.

For the control plane, we observe the SRT for the HSS application to a connecting UE. When monitoring SRT of the HSS application including networking from the MME, the NS with WireGuard has the lowest average SRT. In particular, with ten successful connections for the UE, the average SRT of the Diameter protocol drops from 6.156 ms for the EPS without WireGuard capability to 5.377 ms when WireGuard is added. When doubling the resources on the EPS NS with WireGuard, SRT of 5.607 ms is measured. Based on the other measurements, it is likely that the HSS application itself is the delaying part. With a reduced number of connections, we have not observed a negative effect on the SRT when using WireGuard.

A comparison of the latency measurements for different instances and interfaces is shown in Figure 4. The red line in the figure indicates 1 ms, representing one of the E2E KPI for URLLC applications in 5G. All single-site instances achieve lower latency than the 1 ms. However, adding WireGuard introduces a visible overhead when comparing the NS without WireGuard to the other instances in Figure 4. On the other hand, we observe that the average latencies for the S1-C interface in the eMBB and URLLC NSIs (illustrated in grey and purple) are lower than the other two counterpart measurements. It is worth noting that doubling the resources does not necessarily reduce the latency, confirming that the latency depends on multiple factors such as 5QI parameters and the workload of components in the NS.

Figure 5 compares the throughput between components with WireGuard enabled on different interfaces across instances. The red line represents the 100 Mbps downlink user data rate KPI. We highlight three main results from observing the throughput. The first one is that, unlike the latency, the throughput changes according to the available resources. When comparing the NS with double resources to the others, the throughput is higher for the NS with the double resources. The second observation is that the throughput over the Uu interface is significantly lower than the other measurements. The throughput over the Uu is around 1.7 Mbps, while the average throughput for the S1-U is over 1 Gbps making the Uu the bottleneck of the EPS. The last observation is about the maximum throughput when averaging over 10 minutes. For the NS with double resources, we observe throughput of 2.2 Gbps for the S1-U. For the other instances, a range from 770 Mbps to 1.48 Gbps is measured.

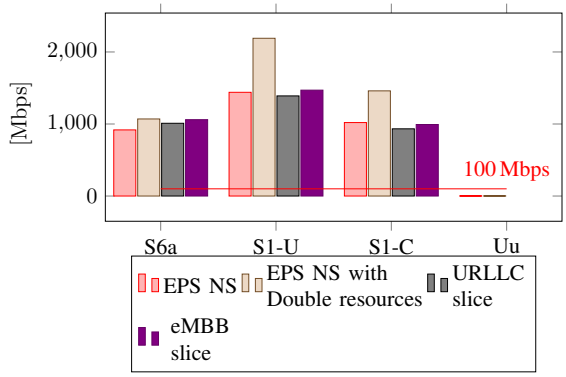


Fig. 5. Throughput comparison for different interfaces with WireGuard.

Figure 6 compares the throughput in the two NSIs. We observe that the performance over diverse interfaces differs when running each NSI alone and when the two NSIs are running simultaneously. For instance, the throughput at the S6a

interface reaches up to 1.1 Gbps for the URLLC slice when it is operating alone and simultaneously with the eMBB slice. However, the throughput at the S1-U interface is 1.43 Gbps for the URLLC slice separately and it reduces a bit to 1.41 Gbps when it is running simultaneously with the eMBB slice. Regarding S1-C interface, the throughput reaches 1.12 Gbps for the separate URLLC and it decreases to 0.97 Gbps when the eMBB slice is also working. In general, the differences between the NSIs are minor, meaning that WireGuard is a promising solution for slice isolation of eMBB and URLLC slices.

It should be noted that we observe a total throughput of approximately 3 Gbps, which is lower than the internal networking throughput of around 20 Gbps when testing with a workload on the same logical interface for the two NSIs simultaneously. As we approach the internal networking limit for the throughput, we detect more considerable differences between the NSIs based on their QoS parameters and the allocated resources.

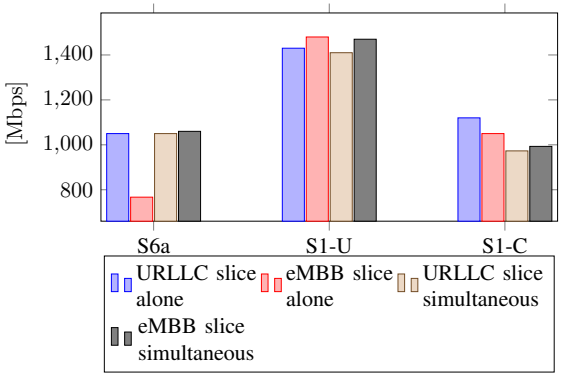


Fig. 6. Throughput comparison with WireGuard for NSIs - measured separately and simultaneously.

In the multi-site deployment, we take measurements over the S6a interface, which is the only one that differs from the other NSs and NSIs. As expected, the throughput is lower, and the latency is higher than in the other instances. The performance is lower even without WireGuard between the VNFs. However, we observe that WireGuard adds overhead in this scenario as well. In the multi-site NS, the average latency over 1000 ICMP packets increases from 18.355 ms to 19.769 ms when using WireGuard. For the average throughput, we observe a reduction from 179 Mbps to 156 Mbps, which is expected based on the given 200 Mbps bandwidth.

VI. Conclusions

By using Juju relations and providing a proof of concept for using WireGuard as VPNaaS, we showed that WireGuard can be implemented with automatic peer setup after instantiating. The performance measurements demonstrate that WireGuard is suitable for applications with requirements corresponding to several of the 5G KPI values. We show that WireGuard can be used as VPNaaS in the context of 5G networks and beyond in order to provide secure communication and slice isolation.

Replacing the arbitrary Juju relations with a KMS, using a 5G Core network instead of EPC components, adding multiple UEs, and evaluating scenarios in which fulfilling service requirements (especially throughput) are beyond WireGuard capability are potential directions for future investigation.

References

[1] B. Blanco, J. O. Fajardo, I. Giannoulakis, E. Kafetzakis, S. Peng, J. Pérez-Romero, I. Trajkovska, P. Sayyad Khodashenas, L. Goratti, M. Paolino, and E. Sfakianakis, "Technology pillars in the architecture of future 5g mobile networks: Nfv, mec and sdn," *Computer Standards and Interfaces*, vol. 54, 01 2017.

[2] A. Esmaeily and K. Kralevska, "Small-scale 5g testbeds for network slicing deployment: A systematic review," *Wireless Communications and Mobile Computing*, vol. 2021, 2021.

[3] Z. Kotulski, T. Nowak, M. Sepczuk, M. Tunia, R. Artych, K. Bocianiak, T. Osko, and J.-P. Wary, "On end-to-end approach for slice isolation in 5g networks. fundamental challenges," in *Federated Conf. on Computer Science and Information Systems (FedCSIS)*, 2017, pp. 783–792.

[4] J. A. Donenfeld, "Wireguard: next generation kernel network tunnel." in *NDSS*, 2017, pp. 1–12.

[5] A. J. Gonzalez, J. Ordonez-Lucena, B. E. Helvik, G. Nencioni, M. Xie, D. R. Lopez, and P. Grønsund, "The isolation concept in the 5g network slicing," in *2020 European Conference on Networks and Communications (EuCNC)*. IEEE, 2020, pp. 12–16.

[6] H. Kim, "5g core network security issues and attack classification from network protocol perspective." *J. Internet Serv. Inf. Secur.*, vol. 10, no. 2, pp. 1–15, 2020.

[7] H. Hantouti, N. Benamar, and T. Taleb, "Service function chaining in 5g amp; beyond networks: Challenges and open research issues," *IEEE Network*, vol. 34, no. 4, pp. 320–327, 2020.

[8] V. N. Sathi, M. Srinivasan, P. K. Thiruvasagam, and S. R. M. Chebiyyam, "A novel protocol for securing network slice component association and slice isolation in 5g networks," in *Proceedings of the 21st ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWIM '18, 2018, p. 249–253.

[9] P. Porambage, Y. Miche, A. Kalliola, M. Liyanage, and M. Ylianttila, "Secure keying scheme for network slicing in 5g architecture," in *2019 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2019, pp. 1–6.

[10] S. Haga, A. Esmaeily, K. Kralevska, and D. Gligoroski, "5g network slice isolation with wireguard and open source mano: A vpnaas proof-of-concept," in *2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2020, pp. 181–187.

[11] I. Vidal, B. Nogales, D. Lopez, J. Rodríguez, F. Valera, and A. Azcorra, "A secure link-layer connectivity platform for multi-site nfv services," *Electronics*, vol. 10, no. 15, 2021.

[12] 5G-PPP Architecture Working Group, "View on 5g architecture," https://tinyurl.com/2p9dxph4, accessed: 07.01.2022.

[13] ETSI OSM, "Etsi-nfv-vnfd," https://osm.etsi.org/docs/user-guide/05-osm-usage.html, accessed: 19.07.2022.

[14] ——, "Etsi-nfv-nsd," https://tinyurl.com/26dt45xv, accessed: 19.07.2022.

[15] ——, "Etsi-nfv-vnfd," https://tinyurl.com/2p9yp7cr, accessed: 19.07.2022.

[16] O. S. Alliance, "Openairinterface," accessed: 04.01.2022. [Online]. Available: https://openairinterface.org/

[17] A. Esmaeily, K. Kralevska, and D. Gligoroski, "A cloud-based sdn/nfv testbed for end-to-end network slicing in 4g/5g," in *2020 6th IEEE Conference on Network Softwarization (NetSoft)*, 2020, pp. 29–35.

[18] T. Dreibholz, "Flexible 4g/5g testbed setup for mobile edge computing using openairinterface and open source mano," in *Workshops of the international conference on advanced information networking and applications*. Springer, 2020, pp. 1143–1153.

[19] ETSI, "System architecture for the 5g system (5gs)," ETSI, Tech. Rep. TS 123 501 V16.6.0, October 2020. [Online]. Available: https://tinyurl.com/2p8392jt

[20] ——, "Why do we need 5g?" https://www.etsi.org/technologies/mobile/5g, accessed: 19.07.2022.

Paper IV

A. Esmaeily and K. Kralevska, "Orchestrating Isolated Network Slices in 5G Networks," under review in EURASIP Journal on Wireless Communications and Networking, 2023.

This paper is under review for publication and is therefore not included.

Paper V

A. Esmaeily, K. Kralevska, and T. Mahmoodi, "Slicing Scheduling for Supporting Critical Traffic in Beyond 5G," 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2022, pp. 637-643, doi: 10.1109/CCNC49033.2022.9700671.[3]

Slicing Scheduling for Supporting Critical Traffic in Beyond 5G

Ali Esmaeily*, Katina Kralevska*, and Toktam Mahmoodi†
*Dept. of Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU), Norway
Email: {ali.esmaeily, katinak}@ntnu.no
†Dept. of Engineering, King's College London, London, UK
Email: toktam.mahmoodi@kcl.ac.uk

***Abstract*—One of the most challenging services fifth-generation (5G) mobile network is designed to support, is the critical services in-need of very low latency, and/or high reliability. It is now clear that such critical services will also be at the core of beyond 5G (B5G) networks. While 5G radio design accommodates such supports by introducing more flexibility in timing, how efficiently those services could be scheduled over a shared network with other broadband services remains as a challenge. In this paper, we use network slicing as an enabler for network sharing and propose an optimization framework to schedule resources to critical services via puncturing technique with minimal impact on the regular broadband services. We then thoroughly examine the performance of the framework in terms of throughput and reliability through simulation.**

Keywords: B5G, eMBB, URLLC, coexistence, resource allocation, puncturing, critical traffic.

I. Introduction

The 5G mobile network came with the promise of ten times better performance in all directions [1]. However, the main paradigm shift has been in supporting services from industry which otherwise have had a dedicated network. Supporting critical services was enabled with two main enablers: the possibility to offer ultra-high reliability and low-latency and the capability to share one network between services with different needs, known as network slicing. It is now clear that the industry support and critical services will be one of the main targets for *beyond 5G* (B5G) networks.

The 5G standard has introduced classes of services in order to encapsulate different requirements. The enhanced Mobile Broadband (eMBB) was introduced as an advanced version of 4G mobile broadband with higher throughput, while the Ultra-Reliable Low-Latency Communication (URLLC) was introduced to capture the needs of critical industry data traffic. The goal of the eMBB service is to attain a high data rate while delivering an acceptable reliability level. In contrast, URLLC services require stringent latency and reliability constraints to support critical industries such as smart factories, autonomous driving, or remote surgery [2]. Such supports are addressed in the 5G New Radio (5G-NR) standard, i.e., the 3rd Generation Partnership Project (3GPP) RAN1/RAN2 [3], [4]. These specifications entail methods for eMBB service to obtain a high data rate and at the same time introduce flexible numerology allowing shorter transmission time, which then could be used for an immediate transmission of a smaller amount of latency-sensitive data through grant-free access [5]. Scheduling URLLC traffic over the same resources that were provisioned for eMBB introduces challenges for the eMBB traffic; hence the rich literature on the coexistence of eMBB and URLLC is reviewed in Section II.

Network slicing is seen as one of the leading enabling technologies for sharing network resources among multiple tenants of a network (including vertical industries), which can provide services with diverse requirements in B5G. Accordingly, radio resource scheduling is crucial to efficiently utilize shared resources between slices in order to meet various services' requirements of the tenants [6]. Hence, 3GPP RAN1/RAN2 specifications facilitate the realization of slicing over 5G-NR via 1) RAN awareness feature to perform traffic administration for slices belonging to different tenants, and 2) policy enforcement and radio resource management for the RAN slices [7]. Such incorporation between 3GPP RAN1/RAN2 and network slicing drives efficient, flexible, and controllable radio resource sharing among slices.

In this paper, we use network slicing as a concept for sharing the network resources between URLLC and eMBB traffic. We consider using regular Transmission Time Intervals (TTIs) for eMBB traffic and short TTIs for URLLC traffic. While the eMBB traffic is scheduled and will be transmitted as planned, the URLLC traffic will be transmitted immediately by *puncturing* the eMBB transmission slot. We extend an existing loss model in the literature [8] and accurately express it to capture the impact of this *puncturing* on the eMBB throughput. To this end, the main contributions of this paper are as follows:

- Characterizing the resource allocation problem for the coexistence of eMBB/URLLC traffic scheduling using the puncturing technique with the main objective of maximizing the minimum data rate of each eMBB user.
- Precisely formulating the loss function definition to capture the impact of puncturing, resulting from overlapped URLLC traffic, on each eMBB user's throughput and per TTI and for every particular allocated radio resource to each eMBB user.
- Presenting an optimization framework ensuring the loss in eMBB throughput due to scheduling URLLC traffic is minimal; hence achievable data rate for the eMBB

978-1-6654-3161-3/22/$31.00 ©2022 IEEE

users is not affected significantly. We define a *puncturing rate threshold* to limit such impact. Moreover, We benchmark our proposed solution with the state-of-the-art approaches. Simulation results confirm that the proposed method can 1) fulfill URLLC reliability requirements and 2) at the same time maintain the minimum achievable rate of the worst-case eMBB user, close to the minimum acceptable data rate for the eMBB users even for a high amount of incoming URLLC load. Worst-case eMBB user refers to the user located at the cell edge (with low allocated power or poor channel gain) or the most punctured user with the overlapped URLLC load.

The remainder of this paper is organized as follows. Section II provides an overview of the literature on eMBB and URLLC coexistence. Section III presents the system model we use in this paper and problem formulation of the optimization framework. The simulation results are presented in Section IV. Section V concludes the paper.

II. Related work

The conventional orthogonal-based radio resource allocation mechanism is not suited for the coexistence of URLLC and eMBB traffic [9]. One of the proposals from the 3rd Generation Partnership Project (3GPP) to efficiently multiplex eMBB and URLLC data transmissions via the 5G-NR is the superposition/puncturing scheme. Superposition/puncturing [10] is performed by applying non orthogonal-based scheduling [11] of both eMBB and URLLC traffic on the same radio channel simultaneously. Superposition/puncturing scheme is recognized as a promising technique to enable the coexistence of the eMBB and URLLC transmissions over the 5G-NR and thus has attracted much attention in academia and industry. Reference [8] models the impact of the URLLC transmission over the scheduled eMBB traffic via loss functions caused by the URLLC traffic. Reference [12] investigates the multiplexing of the eMBB and URLLC traffic in the Cloud RAN (C-RAN) environment. eMBB and URLLC traffic are transmitted via multicast and unicast transmission mode, respectively. The authors also provide a framework in order to maximize the revenue stream of the C-RAN provider. The study in [13] investigates mutual support of visual (over eMBB slice) and haptic (over URLLC slice) perceptions over cellular networks. Paper [14] suggests a two-sided matching game for a joint user association and resource allocation problem, using an analytic hierarchy process, which yields in enhancing resource allocation in the downlink eMBB and URLLC transmissions for a fog network. The authors in [15] utilize a decomposition technique for the integrated eMBB and URLLC resource scheduling problem into two separate problems. For the case of eMBB, the authors employ the penalty successive upper bound minimization method over TTIs, and for the URLLC case, a transportation rule is applied over short TTIs. Paper [16] considers the efficiency of adopting the orthogonal-based and non orthogonal-based scheduling for the eMBB and URLLC traffic in a multi-cell C-RAN system. The work outcome reveals the advantage of using the orthogonal-based solution for degrading the mutual interference of the eMBB and URLLC traffic. The authors in [10] suggest a communication-theoretic basis for eMBB, mMTC, and URLLC services. The results showcase the performance of both orthogonal-based and non orthogonal-based slicing for different service types. The study in [17] utilizes a matching game for the joint eMBB and URLLC traffic. The authors denote an optimization approach to maximize the minimum demanded eMBB data rate and, at the same time, analyze URLLC reliability constraints. Reference [18] presents a risk-sensitive strategy according to the conditional value at risk method for eMBB reliability and a chance constraint for URLLC reliability. The work in [19] provides an optimization problem obtained from an intelligent resource allocation scheme based on the puncturing approach by considering reliability for eMBB and URLLC services. The authors apply a deep reinforcement learning policy to discover the total number of punctured mini-slots of the whole eMBB users.

Unlike those works, which mainly focus on maximizing the sum rate of the eMBB users, this paper concentrates on maximizing individual minimum achievable data rate for the eMBB users. We describe the resource allocation problem for each eMBB user that experiences a negative impact on its data rate due to the incoming URLLC traffic. Such traffic punctures some or even all of the allocated resources to the eMBB user in a time slot.

III. System model and problem formulation

A. System Model

In this network, we consider downlink eMBB and URLLC traffic, i.e., transmission from the network to the pieces of User Equipment (UEs) over a single gNB that can operate with single or multiple antennas $j \in \mathcal{J} = \{1, 2, ..., J\}$. The gNB schedules the eMBB and URLLC traffic and transmits the corresponding data for each service type via its antennas towards eMBB and URLLC users over flat i.i.d Rayleigh fading channels. The gNB serves $k \in \mathcal{K} = \{1, 2, ..., K\}$ number of eMBB and $n \in \mathcal{N} = \{1, 2, ..., N\}$ number of URLLC UEs. The time domain is split into equally spaced time slots (TTIs) for the eMBB UEs' transmissions. Each time slot is subdivided into a fixed number of M equally spaced mini-slots (short TTIs) where $m \in \mathcal{M} = \{1, 2, ..., M\}$ denotes a mini-slot. In the frequency domain, the radio resources are divided into $b \in \mathcal{B} = \{1, 2, ..., B\}$ Resource Blocks (RBs). Each RB b contains 12 sub-carriers in the frequency domain and 14 OFDM symbols in the time domain. Since there is no strict latency requirement for serving the eMBB users, the RBs are allocated to them at the beginning of each time slot. However, the sporadic URLLC requests can arrive at any time within a time slot, and due to the extreme latency requirement of such requests, the gNB needs to serve them immediately in a mini-slot instead of waiting for the next time slot. The gNB punctures previously scheduled eMBB transmissions in mini-slots by applying zero power to these transmissions to serve the URLLC requests promptly.

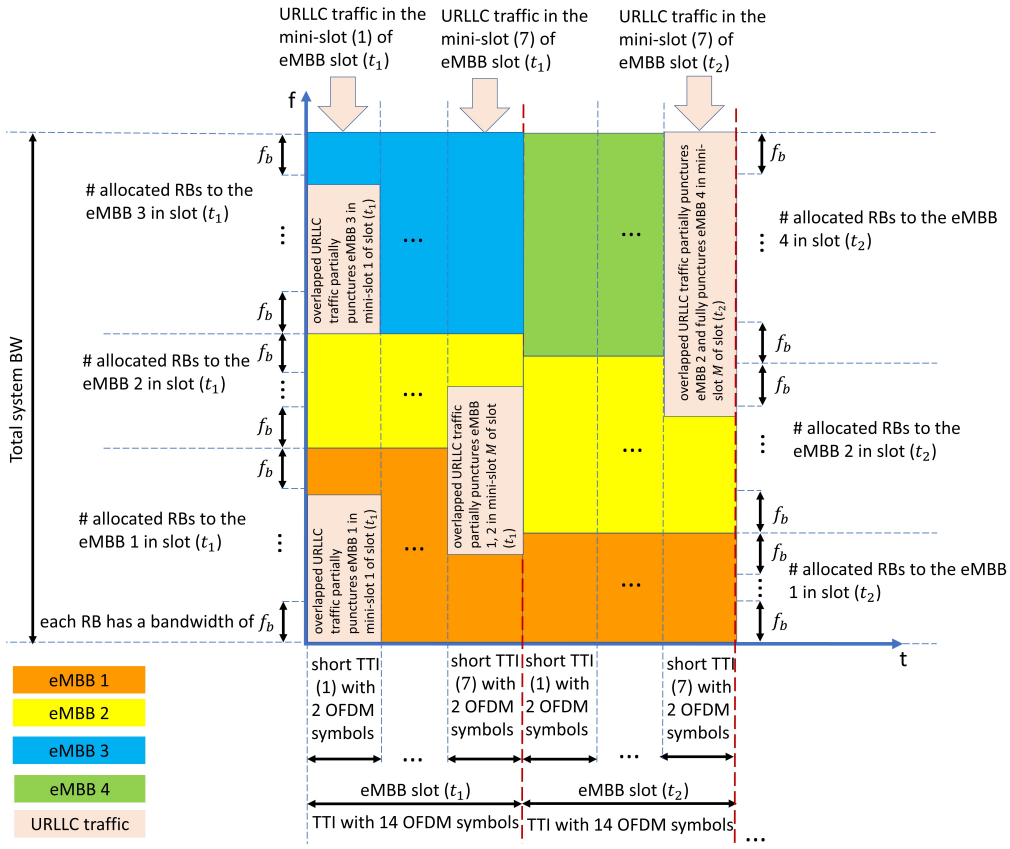


Fig. 1: eMBB/URLLC coexistence enabled by the puncturing mechanism for the numerology $\mu = 0$.

In 5G-NR, unlike 4G, the bandwidth of a RB, f_b, and time slot duration, TTI, are not fixed, and they are set according to specific values of sub-carrier spacing, Δf. Hence, there are several so-called numerologies in 5G-NR according to the values of Δf. Figure 1 illustrates the puncturing mechanism for the coexistence of eMBB/URLLC traffic for the numerology zero-labeled as $\mu = 0$ with Δf= 15 KHz, f_b= 180 KHz, TTI = 1 msec which contains 14 OFDM symbols, and each short TTI $\approx$ 142 μsec that occupies 2 OFDM symbols. Let consider the first mini-slot of the first time slot. The sporadic incoming URLLC traffic in the first mini-slot $m = 1$ impacts the previously scheduled eMBB users with the allocated radio resources in this mini-slot. gNB decides to map the URLLC traffic to some of the eMBB UEs in this mini-slot. Hence, some of the resources of the eMBB UEs 1 and 3, $k = 1, 3$ are punctured by the overlapped URLLC traffic.

Accordingly, the maximum achievable rate for an eMBB user k at the time slot t over the whole allocated RBs can be formulated as follows:

$$r_k^{eMBB}(t) = [\phi_k^{eMBB}(t) - \gamma_k^{eMBB}(t)] \times r_{k,peak}^{eMBB}(t) \quad (1)$$

where the $\phi_k^{eMBB}(t)$ is the total amount of radio resources allocated to the eMBB user k at time slot t, $\gamma_k^{eMBB}(t)$ is called the total loss function which indicates the fraction of punctured resources allocated to eMBB user k at time slot t due to the incoming URLLC requests, and $r_{k,peak}^{eMBB}(t)$ is the total achievable data rate of the eMBB user k at time slot t. This formulation is general, and by following the Shannon channel capacity, it can be further extended to:

$$r_k^{eMBB}(t) = \sum_{b=1}^{B} \Big[\Big(x_{kb}(t) f_b - \gamma_{kb}^{eMBB}(t) \Big) \times log_2 \Big(1 + \frac{\sum_{j=1}^{J} p_{kb}^j(t) h_{kb}^j(t)}{\sigma^2} \Big) \Big] \quad (2)$$

where $x_{kb}(t)$ is the resource allocation coefficient, $x_{kb}(t)$ = 1 denotes that the RB b is allocated to the eMBB user k at time slot t and $x_{kb}(t)$ = 0 shows no allocation; f_b is the bandwidth of the RB b; $p_{kb}^j(t)$ is the transmission power from the antenna j of the gNB over the RB b to the eMBB user k at time slot t; $h_{kb}^j(t)$ is the Rayleigh fading channel gain of

the transmission from the antenna j of the gNB over the RB b to the eMBB user k at time slot t; σ^2 is the noise power; and finally $\gamma_{kb}^{eMBB}(t)$ indicates the fraction of punctured RB b that is allocated to eMBB user k at time slot t. Now, let $D_m(t)$ be a random variable indicating the number of incoming URLLC packets in the mini-slot m of time slot t. Hence, the total incoming URLLC packets in the time slot t is equal to $D(t) = \sum_{m=1}^{M} D_m(t)$. As a result, the $\gamma_k^{eMBB}(t)$ can be formulated as follows:

$$\begin{aligned}\gamma_k^{eMBB}(t) &= \sum_{b=1}^{B} \gamma_{kb}^{eMBB}(t) \\ &= \Big[\sum_{b=1}^{B} x_{kb}(t) f_b \times \rho_{kb}(t) \frac{D(t)}{|B| \times |M|}\Big]\end{aligned} \tag{3}$$

where $\rho_{kb}(t) \in [0,1]$ indicates the weight of puncturing; and $|B| \times |M|$ presents the total system capacity in terms of frequency-time resources. The URLLC traffic is upper bounded by total system capacity, i.e., $D(t) \leq |B| \times |M|$. The $\rho_{kb}(t)$ identifies the pattern of overlapping total URLLC traffic in the time slot t on the eMBB user k resources in order to utilize (puncture) them for the URLLC transmission. According to the pattern of puncturing the eMBB resources, the $\gamma_k^{eMBB}(t)$ function can be approximated as a regular algebraic function. In this paper, we define $\gamma_k^{eMBB}(t)$ as first and second-degree non-decreasing polynomial known as linear and convex quadratic functions, respectively, where $\gamma_k^{eMBB}(t) \in \Big[0, \sum_{b=1}^{B} x_{kb}(t) f_b\Big]$. Hence, for each eMBB user k in time slot t if:

- $\gamma_k^{eMBB}(t) = 0$, no puncturing occurs;
- $0 < \gamma_k^{eMBB}(t) < \sum_{b=1}^{B} x_{kb}(t) f_b$, partial puncturing happens;
- $\gamma_k^{eMBB}(t) = \sum_{b=1}^{B} x_{kb}(t) f_b$, full puncturing appears.

It should be noted that the individual achievable data rate for the eMBB user k in time slot t holds a higher value if this user suffers from a resource deduction scheme following the convex function rather than the linear function.

Until now, we have only considered the latency requirement for the incoming URLLC requests by scheduling them on top of eMBB transmissions. Regarding the reliability requirement of URLLC traffic, let θ_{max} be the outage probability threshold and η be the URLLC packet size, then the reliability of URLLC UEs can be represented as [18]:

$$\begin{aligned}Pr(error) = Pr\Big\{ \sum_{n=1}^{N} \sum_{k=1}^{K} \Big[\frac{\gamma_k^{eMBB}(t)}{f_b N} \times \\ log_2\Big(1 + \frac{\sum_{j=1}^{J} p_{nb}^{j}(t) h_{nb}^{j}(t)}{\sigma^2}\Big)\Big] \leq \eta D(t)\Big\} \leq \theta_{max}.\end{aligned} \tag{4}$$

Under the joint eMBB/URLLC resource allocation problem, the objective is to maximize the data rate for each of the eMBB UEs and at the same time fulfill the URLLC UEs' requirements in terms of extra low delay and high reliability:

$$\max_{p,\gamma} \min_{k \in \mathcal{K}} \mathbb{E}\{\sum_{t=0}^{T} r_k^{eMBB}(t)\} \tag{5a}$$

$$subject\ to \quad Pr(error) \leq \theta_{max} \tag{5b}$$

$$\sum_{k=1}^{K} \sum_{b=1}^{B} \sum_{j=1}^{J} p_{kb}^{j}(t) \leq P_{max} \tag{5c}$$

where the P_{max} is the maximum transmission power from the gNB towards all types of the UEs.

B. Solving the coexistence optimization problem

Here we present the proposed algorithm to find an optimal solution for Eq. (5a). In this algorithm, first, we set the minimum acceptable data rate R_{min} for the eMBB users. Subsequently, in each time slot t we define a *puncturing rate threshold* $th^{eMBB}(t)$ according to the loss functions for all eMBB users. The selection criteria for calculating $th^{eMBB}(t)$ is as follows:

$$th^{eMBB}(t) = \begin{cases} \max_{\forall k \in \mathcal{K}}\{\ \gamma_k^{eMBB}(t)\}, \\ \qquad 0 \leq \gamma_k^{eMBB}(t) < \sum_{b=1}^{B} x_{kb}(t) f_b; \\ \max_{\forall k \in \mathcal{K}}\{\ \gamma_k^{eMBB}(t)\} - \textit{offset}, \\ \qquad \gamma_k^{eMBB}(t) = \sum_{b=1}^{B} x_{kb}(t) f_b; \end{cases} \tag{6}$$

where *offset* indicates a constant value to tune $th^{eMBB}(t)$ if the second condition holds. It is worth noting that the first condition for defining $th^{eMBB}(t)$ is much more likely to happen than the second one. After setting a value for $th^{eMBB}(t)$, we proceed to calculate the achievable rate for each eMBB user k in the time slot t. Next, we check whether the achievable rate is less than R_{min}. If $r_k^{eMBB}(t) < R_{min}$, then we map the incoming URLLC load to another possible eMBB user k' with the allocated RB b' if at least one of the following conditions is fulfilled:

- higher power value, i.e. $p_{k'b'}^{j'}(t) > p_{kb}^{j}(t)$;
- larger channel gain value, i.e. $h_{k'b'}^{j'}(t) > h_{kb}^{j}(t)$;
- lower loss function, i.e. $\gamma_{k'}^{eMBB}(t) < \gamma_k^{eMBB}(t)$;

otherwise we hold with the current eMBB user k. In other words, the challenge corresponds to the minimum rate belonging to the most punctured eMBB users, which negatively impacts the performance of the system if the minimum rate would be less than the R_{min}. Hence, tracking each eMBB user rate is crucial in each time slot within a frame and for the whole transmission period. As a result, the optimization algorithm protects those eMBB users with low power allocation, bad channel quality, and less allocated RBs to avoid worsening their date rate by over-puncturing. Algorithm 1 summaries the steps.

Algorithm 1 Algorithm for eMBB/URLLC coexistence

Input: $t \in T, b \in \mathcal{B}, k \in \mathcal{K}, j \in \mathcal{J}, p^{j}_{kb}(t), h^{j}_{kb}(t), \gamma^{eMBB}_{k}(t)$
Output: Solution to Eq. (5a) for eMBB/URLLC coexistence
Set R_{min}
Define $th^{eMBB}(t)$ according to Eq. (6)
for $t \in T$ **do**
  for $k \in K$ **do**
    for $j \in J$ **do**
      Calculate $r^{eMBB}_{k}(t)$ based on $th^{eMBB}(t)$
      if $r^{eMBB}_{k}(t) < R_{min}$ **then**
        Map the URLLC load to eMBB user k' in case $p^{j'}_{k'b'}(t) > p^{j}_{kb}(t)$, $h^{j'}_{k'b'}(t) > h^{j}_{kb}(t)$, or $\gamma^{eMBB}_{k'}(t) < \gamma^{eMBB}_{k}(t)$
      else
        Puncture the current eMBB user k
      end if
    end for
  end for
end for

TABLE I: Simulation parameters.

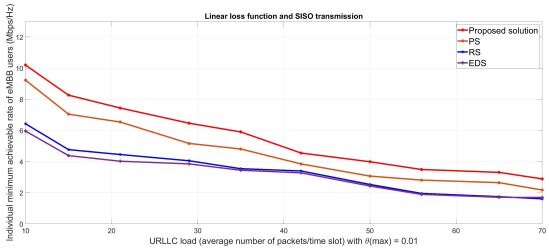
Simulation parameter	Value
Cell radius(m)	500
Number of mini-slots	7
Number of OFDM symbols per mini-slot	2
Number of eMBB users	5
URLLC traffic model	Poisson process
f_b (KHz)	180
Total BW (MHz)	20
Min guard band for numerology $\mu = 0$ (KHz)	692.5
Number of resource blocks	103
R_{min} (Mbps)	5
P_{max} (dBm)	40
Time slot length (msec)	1
Mini-slot length (μsec)	142
Time frame length (msec)	10
URLLC packet size (Bytes)	50

IV. Performance evaluation

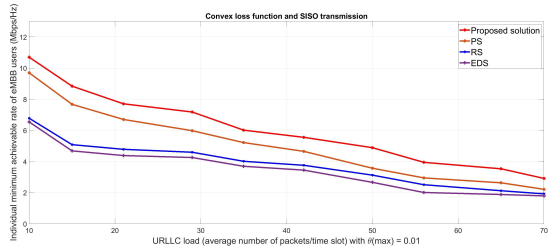
In this section, we verify the efficiency of our proposed algorithm through simulations and evaluate the performance. Our objective is to show the increase of the individual minimum achievable data rate for each eMBB user in the following analysis. We analyze and simulate the RAN domain using MATLAB R2019b with the CVX toolbox. In our simulated RAN, we consider one gNB located at the center of the cell coverage zone with a 500 m radius. The gNB operates on 20 MHz in the downlink mode, which serves several eMBB and URLLC UEs that are randomly distributed within the coverage zone. Besides, the gNB schedules eMBB and URLLC traffic in the downlink transmission over flat i.i.d Rayleigh fading channels. Table I summarizes the main simulation parameters. We benchmark the performance of our proposed solution with the well-known state-of-the-art approaches, including: 1) Punctured Scheduling (PS) [20]: PS selects the RBs with the highest MCS allocated to eMBB users and punctures them in order to serve URLLC traffic; 2) Random Scheduler (RS) [8]: RS serves the incoming URLLC traffic by randomly selecting pre-allocated RBs to the eMBB users; and 3) Equally Distributed Scheduler (EDS) [17]: EDS serves the incoming URLLC traffic by equally choosing pre-allocated RBs to each of the eMBB users.

A. eMBB data rate influenced by puncturing with URLLC load

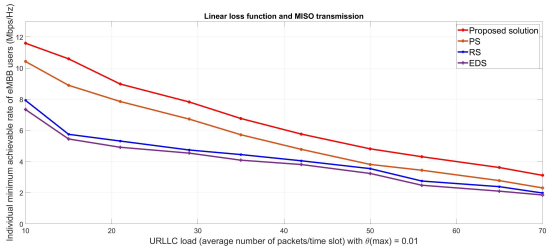
We first investigate the performance of the proposed algorithm in terms of resource allocation for the individual minimum achievable rate of the eMBB users. The deduction of the allocated resources to the eMBB users is represented by the $\gamma^{eMBB}_{k}(t)$ function in each time slot. We assume that the type of this function for the simulation environment is either a second-degree non-decreasing polynomial linear or convex quadratic function. Moreover, gNB can operate either with single or multiple numbers of antennas towards eMBB and URLLC users. We also consider that both eMBB and URLLC users are equipped with only a single antenna for data transmission. Hence the operation in downlink between the gNB and the users happens either in Single- or Multiple-Input Single Output configurations known as SISO and MISO, respectively. Figure 2 illustrates four different *regimes* that may happen via transmission of the data in the downlink in the form of *(type of $\gamma^{eMBB}_{k}(t)$, type of the transmission configuration)*. For each *regime*, we study the individual minimum achievable eMBB data rate per user via the proposed optimization algorithm, PS, RS, and EDS solutions. By increasing the URLLC load per time slot, depending on the scheduling strategy, some or all of the eMBB users may be influenced by puncturing. Particularly, in the *(linear, SISO) regime* illustrated in Figure 2a, for a number of 40 URLLC packets per time slot (considered as a mid-range number of URLLC packets per time slot), the minimum achievable rate for each eMBB user can reach up to 3.3, 3.5, 4.2, and 5.1 Mbps for the EDS, RS, PS, and our proposed solution respectively. By applying the optimization algorithm, gNB searches for at least one possible pre-scheduled eMBB candidate with higher allocated power, higher channel gain, or lower loss function to map full or partial URLLC load to that eMBB user while at the same time satisfying the minimum acceptable data rate for the eMBB users. The optimization process enhances even the worst-case eMBB user data rate to achieve up to 5.1 Mbps which is still greater than the R_{min}. The performance of the proposed algorithm is also prominent by increasing the minimum data rate up to 10.1 Mbps for the low amount of URLLC packets per time slot (10 packets). The same logic follows for the other *regimes* as well. The most reliable case is *(convex, MISO) regime*, presented in Figure 2d. This *regime* holds the convex loss function with less puncturing impact on the eMBB users than the linear loss function, and gNB operates with



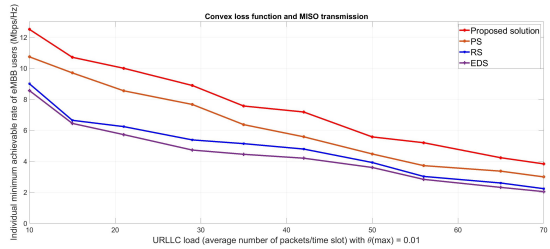
(a) *(linear, SISO) regime*.



(b) *(convex, SISO) regime*.



(c) *(linear, MISO) regime*.



(d) *(convex, MISO) regime*.

Fig. 2: Individual minimum achievable eMBB data rate for different number of URLLC packets per time slot.

multiple antennas towards all users in the downlink. In the *(convex, MISO) regime*, the proposed algorithm can improve the minimum data rate for the worst-case eMBB user up to 12.5 Mbps for 10 URLLC packets per time slot while PS, RS, and EDS can ultimately achieve up to 10.8, 9, and 8.8 Mbps respectively for the same user. It is worth considering that the efficiency of the proposed algorithm is noticeable even for a high amount of URLLC load with the rate of up to 57 packets per time slot, where the individual minimum achievable data rate is equal to the R_{min}. Due to the sporadic nature of such packets, the probability of having a very high number of URLLC packets per time slot is low, and thus the proposed algorithm is close to real scenarios. Besides, we assume the size of a URLLC packet is 50 bytes; however, smaller packet sizes are also expected, which results in less puncturing of the eMBB users. By keeping the individual minimum data rate close to the R_{min}, the network guarantees that each eMBB user receives at least minimum resources, which are required for normal web browsing and light video streaming. However, full HD video streaming with very high resolution demands some buffer time. In fact, with this strategy, the network does not allow to fully puncture eMBB users, and it keeps the data rate at a minimum level to avoid reducing the individual eMBB data rate significantly. The proposed algorithm outperforms PS, RS, and EDS solutions in different regimes under the same amount of URLLC load per time slot.

B. eMBB reliability region for different URLLC load

Here we analyze the reliability of the eMBB users. We set R_{min} equal to 5 Mbps and consider the most reliable transmission *(convex, MISO) regime*. As Figure 3 illustrates, applying the proposed algorithm during the transmission towards the eMBB users delivers a more reliable communication compared to the other scheduling policies in the downlink. Specifically, the eMBB users experience 91% reliable transmission for 10 incoming URLLC packets per slot while PS, RS, and EDS can provide reliable transmission up to 86%, 82%, and 80%, respectively. By increasing the intensity of the URLLC packets per slot, the eMBB reliability decreases to 71% for a very high number of URLLC packets (70 packets per time slot) which, in fact, is less likely. The proposed algorithm surpasses the other solutions even for a high number of URLLC packets per slot, and the gap between our strategy and its closest competitor, PS, is significant. The proposed algorithm is 10% more reliable than the PS case for 70 URLLC packets per time slot.

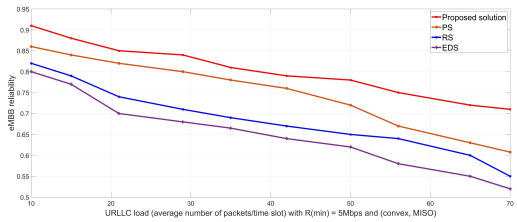


Fig. 3: eMBB reliability for different number of URLLC packets per slot with R_{min} = 5 Mbps and *(convex, MISO) regime*.

V. Conclusion

We investigated the coexistence problem of eMBB and URLLC in 5G-NR. We formulated the puncturing data rate

problem for each eMBB user in order to study the impact of the incoming URLLC traffic, which must be scheduled immediately within mini-slots due to its extra low latency requirement. We proposed an optimization algorithm to enhance the minimum eMBB data rate per user and evaluated its performance with various loss functions, gNB transmission configuration regimes, and some state-of-the-art solutions. As a result, the proposed algorithm improves the data rate per eMBB user, even for the worst-case eMBB user. Besides, by applying the proposed optimization algorithm, the eMBB users experience a more reliable transmission than the other approaches.

REFERENCES

[1] Ericsson, "5G: what is it?" ERICSSON, white paper, October 2014.

[2] H. Mendis, P. E. Heegaard, and K. Kralevska, "5g network slicing as an enabler for smart distribution grid operations," 2019.

[3] 3GPP, "RAN1 - Radio Layer 1 (Physical layer)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.331, 03 2020, version 16.0.0.

[4] ——, "RAN2 - Radio layer 2 and Radio layer 3 Radio Resource Control," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.824, 03 2019, version 16.0.0.

[5] 3GPP, "TSG RAN WG1 88," 3GPP, Tech. Rep., February 2017.

[6] A. Esmaeily and K. Kralevska, "Small-scale 5g testbeds for network slicing deployment: A systematic review," *Wireless Communications and Mobile Computing*, vol. 2021, 2021.

[7] GSM Association, "E2E Network Slicing Architecture," GSMA, White Paper, June 2021, version 1.0.

[8] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5g wireless networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 1970–1978.

[9] 3GPP, "Technical specification group services and system aspects; release 15 description," 3GPP, Technical Specification (TS), March 2019, version 1.1.0.

[10] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[11] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[12] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced c-ran incorporated with urllc and multicast embb," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 881–895, 2019.

[13] J. Park and M. Bennis, "Urllc-embb slicing to support vr multimodal perceptions over wireless cellular systems," pp. 1–7, 12 2018.

[14] S. F. Abedin, M. G. R. Alam, S. M. A. Kazmi, N. H. Tran, D. Niyato, and C. S. Hong, "Resource allocation for ultra-reliable and enhanced mobile broadband iot applications in fog network," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 489–502, 2019.

[15] A. Bairagi, M. Munir, M. Alsenwi, N. Tran, S. Alshamrani, M. Masud, Z. Han, and C. S. Hong. (2020, 03) Coexistence mechanism between embb and urllc in 5g wireless networks.

[16] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of urllc and embb services in the c-ran uplink: An information-theoretic study," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[17] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, and C. S. Hong, "A matching based coexistence mechanism between embb and urllc in 5g wireless networks," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '19. Association for Computing Machinery, 2019, p. 2377–2384.

[18] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, and C. S. Hong, "embb-urllc resource slicing: A risk-sensitive approach," *IEEE Communications Letters*, vol. 23, no. 4, pp. 740–743, 2019.

[19] M. Alsenwi, N. Tran, M. Bennis, S. Pandey, A. Bairagi, and C. S. Hong, "Intelligent resource slicing for embb and urllc coexistence in 5g and beyond: A deep reinforcement learning based approach," *IEEE Transactions on Wireless Communications*, vol. PP, pp. 1–1, 02 2021.

[20] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, 2017, pp. 1–6.

Paper VI

A. Esmaeily, H. V. K. Mendis, T. Mahmoodi, and K. Kralevska, "Beyond 5G Resource Slicing With Mixed-Numerologies for Mission Critical URLLC and eMBB Coexistence," in IEEE Open Journal of the Communications Society, vol. 4, pp. 727-747, 2023, doi: 10.1109/OJCOMS.2023.3254816.

Received 20 January 2023; revised 16 February 2023; accepted 2 March 2023. Date of publication 9 March 2023; date of current version 22 March 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3254816

Beyond 5G Resource Slicing With Mixed-Numerologies for Mission Critical URLLC and eMBB Coexistence

ALI ESMAEILY [1] (Student Member, IEEE), H. V. KALPANIE MENDIS [1], TOKTAM MAHMOODI [2] (Senior Member, IEEE), AND KATINA KRALEVSKA [1] (Member, IEEE)

[1]Department of Information Security and Communication Technology, Norwegian University of Science and Technology, 7491 Trondheim, Norway

[2]Department of Engineering, King's College London, WC2R 2LS London, U.K.

CORRESPONDING AUTHOR: A. ESMAEILY (e-mail: ali.esmaeily@ntnu.no)

ABSTRACT Network slicing has been a significant technological advance in the 5G mobile network allowing delivery of diverse and demanding requirements. The slicing grants the ability to create customized virtual networks from the underlying physical network, while each virtual network can serve a different purpose. One of the main challenges yet is the allocation of resources to different slices, both to best serve different services and to use the resources in the most optimal way. In this paper, we study the radio resource slicing problem for Ultra-Reliable Low Latency Communications (URLLC) and enhanced Mobile Broadband (eMBB) as two prominent use cases. The URLLC and eMBB traffic is multiplexed over multiple numerologies in 5G New Radio, depending on their distinct service requirements. Therein, we present our optimization algorithm, Mixed-numerology Mini-slot based Resource Allocation (MiMRA), to minimize the impact on eMBB data rate due to puncturing by different URLLC traffic classes. Our strategy controls such impact by introducing a puncturing rate threshold. Further, we propose a scheduling mechanism that maximizes the sum rate of all eMBB users while maintaining the minimum data rate requirement of each eMBB user. The results obtained by simulation confirm the applicability of our proposed resource allocation algorithm.

INDEX TERMS B5G, eMBB, numerology, puncturing, resource allocation, URLLC.

I. INTRODUCTION

THE NEXT-GENERATION mobile networks are intended to support the diverse requirements of the vertical industries, thus, to support a wide range of devices and applications. The fifth generation (5G) and beyond 5G (B5G) networks expand not only mobile broadband services compared to the fourth generation (4G) but also address new service-oriented use cases that involve innovative healthcare delivery, smart transportation systems, factory automation, and smart grids.

To address the diversity of applications and services served by 5G, the community categorized these services into three classes. These include massive Machine-Type Communications (mMTC), enhanced Mobile Broadband (eMBB), and Ultra-Reliable Low-Latency Communications (URLLC). mMTC is designed to provide connectivity for thousands of devices spread over a wide coverage. mMTC requires a low data rate and low power connectivity for enormous amounts of sensor/actuator devices (i.e., the Internet of Things (IoT)). eMBB deals with high data rates, high spectral efficiency, and low latency and can be considered as a direct extension of the 4G broadband services. URLLC necessitates significantly low End-to-End (E2E) latency and ultra-high reliability, and it is associated with tactile Internet [1]. URLLC is also referred to as mission-critical communications as it enables real-time control and automation of dynamic processes in various fields, such as industrial process automation and manufacturing, power distribution, or traffic management and safety.

The major challenge in providing such diverse and abundant services is that the physical infrastructure resources are scarce to meet all requirements. Thus, these resources need

to be deployed intelligently to deliver such services. In order to fulfill the above-specified diverse performance requirements imposed by B5G use cases, the next-generation mobile network has to be redesigned. Network slicing is a promising paradigm, which allows heterogeneous services to coexist within the same network architecture. The approach towards softwarization, virtualization, and cloudification as enabling technologies of network slicing has brought tremendous progress and benefits in terms of programmability, flexibility, and innovative ideas to service provisioning. Hence, network slicing leverages the benefits of a virtualized resource sharing environment enabled by Software-Defined Networking (SDN) and Network Function Virtualization (NFV) [2], [3], [4]. Based on softwarization and virtualization, it is capable of enabling Network-as-a-Service (NaaS) [5] and allows the coexistence of multiple networks on the same physical infrastructure. An E2E network slice is composed of the Radio Access Network (RAN), transport and Core Network (CN) sub-network slices in between the end (user) devices [6]. In this work, we consider slicing in the RAN, which is a constituting part of an E2E network slice.

5G New Radio (5G NR) follows the same principles of Orthogonal Frequency Division Multiple Access (OFDMA) technology which was adopted in Long Term Evolution (LTE) and LTE-Advanced (LTE-A). 5G NR supports multiple waveform configurations, which results in scalable numerologies. The resulting flexible frequency-time lattice is designed to support diverse requirements imposed by different traffic classes. URLLC users are typically mission-critical; therefore, they need to be prioritized over the eMBB users, which are typically considered best effort users. The coexistence of eMBB and URLLC users in the same mobile network is, hence, demanding given the trade-off between simultaneously achieving high data rates for the eMBB users and the ultra-reliability and low latency for the URLLC users.

A. RELATED WORK

Resource allocation and orchestration are vital aspects of network slicing as the logical E2E slices are realized upon a shared resource pool. To this end, numerous research works have considered radio resource allocation and proposed various scheduling algorithms. The 3rd Generation Partnership Project (3GPP) [7] has proposed a superposition/puncturing method for multiplexing URLLC and eMBB traffic in 5G cellular systems. The authors in [8] study the coexistence problem of eMBB and URLLC users in 5G networks. They formulate a joint resource allocation problem that can satisfy both eMBB user rate and URLLC interrupt probability requirements. They assign mini-slots for URLLC users and calculate the transmission power of URLLC users, ensuring the reliability constraint. A similar study is performed in [9], which also studies the resource slicing problem for 5G eMBB and URLLC services. The resource slicing problem is formulated as an optimization problem that aims at maximizing the eMBB data rate. The problem is subject to a URLLC reliability constraint while considering the variance of the eMBB data rate to reduce the impact of immediately scheduled URLLC traffic on the eMBB reliability. An optimization-aided deep reinforcement learning-based framework is proposed to solve the formulated problem.

The dynamic multiplexing scheme [7] is recognized as a promising technique to enable the coexistence of the eMBB and URLLC transmissions over the 5G NR and thus has attracted much attention in academia and industry. The authors in [10] evaluate the coexistence technique for eMBB and URLLC based upon a punctured scheme. They extend the study to formulate an optimization problem aiming to maximize the minimum expected achievable rate of eMBB User Equipment (UEs) while fulfilling the provisions of the URLLC traffic. In study [11], the radio resources are scheduled among the eMBB UEs on a time slot basis, whereas they are handled for URLLC UEs on a mini-slot basis. They use a penalty successive upper bound minimization-based algorithm for eMBB UEs, while the optimal transportation model is adopted to solve the same URLLC UEs problem. They also present a heuristic algorithm for efficient scheduling of PRBs among eMBB UEs.

Authors of [12] model the impact of the URLLC transmission over the scheduled eMBB traffic via loss functions caused by the URLLC traffic. The work in [13] analyzes the multiplexing of the eMBB and URLLC traffic in the Cloud-RAN (C-RAN) environment. The work in [14] investigates the performance trade-offs between eMBB and URLLC traffic types in a multi-cell C-RAN architecture under Non-Orthogonal Multiple Access (NOMA) and OMA access strategies. The work outcome reveals the advantage of employing the orthogonal-based solution for degrading the mutual interference of the eMBB and URLLC traffic. The authors also demonstrate the potential benefits of puncturing in improving the efficiency of fronthaul usage by discarding received mini-slots affected by URLLC interference. The authors in [15] present a puncturing scheme for transmitting low latency communication traffic, multiplexed on a downlink shared channel with eMBB. They also propose recovery mechanisms for the impacted eMBB users to minimize the capacity loss for eMBB users due to low latency communication traffic. A group of authors considers an optimal resource assignment under different channel conditions within a mixed numerology approach in [16], [17]. The work presented in [18] focuses on the scheduling problem for heterogeneous services within a mixed numerology approach aiming to maximize the number of satisfied users while meeting latency demand and data transmission requirements. Mini-slots enable transmissions that can be performed in a shorter time than the regular slot duration. In higher numerologies, the use of wider Sub-Carrier Spacings (SCSs) provides shorter slot durations. Consequently, low-latency communications can be enabled by combining mixed numerology and mini-slot approaches. 3GPP proposed mixed numerology with mini-slots that use single numerology with shorter slot durations than a regular slot for that predefined numerology

in order to support multiple services on the same carrier [19]. The work in [20] offers a model to optimize the numerology and resource allocation for mixed numerology systems, which employ the mini-slot approach.

The work in [21] aims to maximize the minimum expected achieved rate of eMBB users and fairness between them by employing a one-to-one matching game to compute appropriate eMBB and URLLC pairs for URLLC resource allocation. The authors of [22] and [23] aim at maximizing the aggregated throughput of the eMBB and URLLC users while mitigating the Inter-Numerology Interference (INI). They consider satisfying the minimum acceptable throughput of the eMBB and maximum allowed delay of the URLLC users according to their corresponding service requirements. The authors propose a deep reinforcement learning INI-aware agent to overcome the computation complexity of the optimization problem. Their method offers a spectrum allocation fulfilling the eMBB and URLLC service requirements while reducing the INI. Finally, they analyze their results delivered by the INI-aware agent when the URLLC traffic statistic is modeled based on mobile and industrial networks. Reference [24] formulates the RAN slicing problem between eMBB and URLLC users as a multi-timescale problem and proposes a hierarchical deep neural network algorithm to assign radio resources to their corresponding users. The authors model the selection of slice parameters within a time slot as a partially observable Markov decision process and present an algorithm to define configuration parameters for the eMBB and URLLC slices efficiently.

The work in [25], the authors compute the achievable latency for the industrial network scenario based on an accurate system-level simulation. Their primary focus is determining 5G NR configurations that are more relevant for Industry 4.0 applications to analyze the effect of reserving bandwidth for URLLC services. Reference [26] defines a context of the network based on combined statistical characteristics from the wireless channel and UEs' service requirements to train a Mondrian forest to predict an optimal mixed-numerology profile. The authors of [27] work on solving the challenges of radio resource allocation in the mmWave band of 5G NR by proposing a deep reinforcement learning-based scheduler. The scheduler allocates resources for a list of UEs to satisfy their different slice's SLA requirements according to the channel quality of each UE. Paper [28] presents a resource allocation strategy that combines latency, control channel, hybrid automatic repeat request, and radio channel quality in determining the transmission resources for different users. The approach minimizes the latency and bypasses unwarranted costly segmentation of URLLC payloads over several transmissions. Reference [29] addresses the problem of joint admission control and resource scheduling for URLLC by utilizing a standard continuous SNR model, where all allocated resource blocks contribute to the success probability, and a binary SNR model, where each resource block is classified as active or inactive according to a SNR threshold. In congestion cases, the work focuses on discovering a subset of users that can be scheduled at the same time.

The authors in [30] develop a joint optimization problem for power and bandwidth allocation with long-term conditions of queues backlog for the eMBB users. They utilize the Lyapunov drift-plus-penalty technique to create the relationship between the long-term constraints and the short-term optimization problem. Furthermore, they employ a one-to-one matching procedure to solve the slicing puncture problem. The work in [31] designs a coordinated multi-point multi-numerology network to improve the throughput of eMBB and latency of URLLC users. The authors solve a subcarrier and power allocation problem with the objective of maximizing the system sum rate. They show that their designed network has a higher sum data rate, lower delay, and throughput outage compared to the traditional non-coordinated multi-point single numerology scenarios. Reference [32] concentrates on minimizing the rate loss of the eMBB users and packet segmentation loss of URLLC users while fulfilling the QoS requirements of eMBB and URLLC use cases. They consider the case of one-to-one pairing in which one URLLC packet can be paired with only one eMBB. They employ a bi-level optimization problem that includes one inner and one outer problem. The inner problem seeks to discover the optimal power and frequency resources for each URLLC and eMBB pair, and the outer problem desires to search for the optimal eMBB-URLLC pairing policy. They also generalize the problem for many-to-many pairing while undervaluing the overhead due to URLLC packet segmentation.

The authors in [33] aim at minimizing the decoding error rate of URLLC users while ensuring the demand for the throughput of eMBB users. They propose a block coordinate descent optimization algorithm to obtain the optimal bandwidth allocation, puncture weight, and transmit power. Paper [34] focuses on studying eMBB and URLLC use cases in networks that are assisted by a Reconfigurable Intelligent Surface (RIS). The authors jointly optimize the power and frequency allocation problem and the RIS phase shift matrix to enhance the eMBB sum rate and URLLC reliability. The work in [35] concentrates on eMBB and URLLC use cases in a massive MIMO system by providing a unified information-theoretic framework incorporating an infinite-blocklength analysis of the eMBB spectral efficiency with a finite-blocklength analysis of the URLLC error probability. The work relies on the use of mismatched decoding and saddlepoint approximation.

Compared to the works presented above, in our previous work [36], we maximize the data rate for each of the eMBB users while guaranteeing a minimum acceptable data rate requirement per eMBB user. We develop the resource allocation problem by formulating a loss function for each eMBB user that experiences an adverse impact on its data rate due to the puncturing by the incoming URLLC traffic. We aim to minimize such negative impact of URLLC traffic upon eMBB users by introducing a puncturing rate threshold. In

TABLE 1. Comparison of related work and the proposed work in this paper for eMBB and URLLC coexistence, where ✓ denotes that the corresponding work covers the topic and ✗ denotes that the corresponding work does not cover the topic.

Related Work	Puncturing method	Mixed Numerology	URLLC latency	URLLC reliability	URLLC traffic classification	eMBB Loss function definition	Power allocation	eMBB data rate	Resource Block allocation	Channel state
[8]	✓	✗	✗	✓	✗	✗	✓	✓	✓	✗
[9]	✓	✗	✗	✓	✗	✗	✓	✓	✓	✓
[10]	✓	✗	✓	✓	✗	✓	✗	✓	✓	✗
[11]	✓	✗	✓	✓	✗	✓	✗	✓	✓	✗
[12]	✓	✗	✗	✗	✗	✓	✗	✓	✓	✓
[13]	✗	✗	✓	✗	✗	✗	✓	✓	✗	✓
[14]	✓	✗	✓	✗	✗	✗	✓	✓	✗	✓
[15]	✓	✗	✓	✗	✗	✗	✗	✓	✗	✓
[16]	✗	✓	✗	✗	✗	✗	✗	✓	✗	✓
[17]	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓
[18]	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗
[20]	✗	✓	✓	✗	✗	✗	✗	✓	✗	✓
[21]	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓
[22], [23]	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓
[24]	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
[25]	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗
[26]	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗
[27]	✗	✓	✓	✗	✗	✗	✗	✓	✓	✗
[28]	✗	✗	✓	✓	✗	✗	✗	✓	✓	✗
[29]	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗
[30]	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓
[31]	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓
[32]	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓
[33]	✓	✗	✗	✓	✗	✓	✓	✓	✓	✗
[34]	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓
[35]	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓
[36]	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓
Our work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

this paper, we extend our previous work by incorporating mixed numerologies for different URLLC traffic classes. We also introduce the MiMRA algorithm for the resource-slicing problem between URLLC and eMBB traffic. Some of the recent related works, such as [22], [23], [30], [32], and [34], address the main technical challenges in the eMBB and URLLC coexistence problem. Nevertheless, no work considers URLLC traffic classification. Although all of the URLLC use cases share low latency and ultra-high reliability characteristics, each specific use case holds its own distinct and exclusive value, such as the Motion control use case compared to the Closed-loop control use case as discussed further in Section II-B. Such URLLC use cases also need prioritization in serving by the network. Thus, classifying URLLC traffic is crucial. Besides, each specific URLLC class holds a different packet size, and this feature becomes extra-critical when transmitting such packets promptly. Moreover, considering accurate power allocation to the eMBB and URLLC users is also vital in order to fulfill their service requirements while overcoming noise and interference. Consequently, there is a need for simultaneously addressing puncturing over mixed numerologies while differentiating the URLLC traffic. This motivates our contributions in this paper, outlined in the next section. Table 1

summarizes the comparison of the related work and our proposed work.

B. CONTRIBUTIONS

The main contributions of this paper are as follows:

- We describe the radio resource allocation problem for the coexistence of eMBB and URLLC traffic scheduling by employing the puncturing method over different numerologies. We formulate the resource allocation problem according to the eMBB resource block allocation, eMBB power allocation, and scheduling of different URLLC traffic classes. Our main objective is to maximize the *sum rate* of the eMBB users while fulfilling the minimum acceptable data rate of each eMBB user in order to deliver fairness in allocating radio resources. Concurrently, the resource allocation problem has to satisfy the extra low latency and ultra-high reliability requirements of the URLLC users.
- We categorize the URLLC traffic into different classes. Each class represents a portion of the traffic that has been generated by the URLLC users belonging to a particular URLLC use case. To the best of our knowledge and following Table 1, there is no similar work that investigates together the puncturing method and mini-slots with 5G NR mixed-numerology to fulfill distinct URLLC classes' requirements (extra low latency and ultra-high reliability) on the one hand and to maximize the sum rate of the eMBB users on the other hand. In this way, apart from eMBB users, we can also differentiate and prioritize URLLC traffic classes as they belong to various URLLC use cases and thus hold different QoS requirements.
- We define precisely a loss function of the eMBB user's data rate to capture the impact of puncturing by the overlapped traffic of each URLLC class according to the number and size of the URLLC packets within each class. The loss function is expressed per TTI in each specific numerology and for every particular radio resource allocated to each eMBB user.
- We propose an optimization strategy called Mixed-numerology Mini-slot based Resource Allocation (MiMRA) that guarantees the loss in eMBB data rate due to the overlapped URLLC traffic is minimal. Consequently, the achievable data rate for the eMBB users is not impacted immensely. Furthermore, we represent a *puncturing rate threshold* to limit the such impact.

C. ORGANIZATION

The remainder of this paper is organized as follows. In Section II we present a few fundamental concepts that are related to this work. Section III explains the system model of our network. In Section IV, we describe the proposed optimization strategy for eMBB/URLLC coexistence. In Section V, we illustrate the numerical results of the analysis. Finally, Section VI concludes the paper.

II. PRELIMINARIES

In this section, we discuss a few fundamental elements upon which the work of this paper is built.

A. 5G NEW RADIO

5G New Radio (NR) is designed to support deployment across a wide range of frequencies. Two different frequency ranges are designated for 5G NR named: Frequency Range 1 (FR-1) and Frequency Range 2 (FR-2) [37]. The bands in FR-1 are envisaged to carry much of the traditional cellular mobile communications traffic. The higher frequency bands in the range FR-2 aim to provide short range very high data rate capability for the 5G radio. Thus, 5G NR can operate in both the sub-6 GHz bands, some of which are traditionally used by previous standards, as well as millimeter wave (mmWave) bands with a shorter range but higher available channel bandwidths.

1) 5G SCALABLE NUMEROLOGIES

Distinct from LTE-A, 5G NR supports multiple waveform configurations referred to as numerologies. A numerology represents a set of parameters such as SCS, OFDM symbol length, and Cyclic Prefix (CP). LTE supports carrier bandwidths of up to 20 MHz with a mainly fixed OFDM numerology (15 KHz SCS). Nevertheless, NR is designed to offer scalable OFDM numerologies to support diverse spectrum bands and deployment models. This is achieved by creating multiple numerologies formed by scaling the basic LTE SCS with 2^{μ}, where μ is an integer between 0 and 4. The numerology is selected independently of the frequency band, with possible SCS of 15 KHz to 240 KHz. Regardless of the numerology, the length of a radio frame and a subframe are always 10 ms and 1 ms, respectively. The difference is the number of time slots within a subframe and the number of symbols within a time slot.

Table 2 presents the main features of each of the five numerologies defined in 5G NR [38]. The following is the terminology used in this paper.

- *Numerology:* A numerology represents a set of parameters such as SCS, OFDM symbol length, and CP.
- *Frame:* Similar to LTE, 10 subframes, each lasting for $1ms$ construct one frame.
- *Slot:* A slot consists of 14 OFDM symbols and is transmitted within a transmission time interval (TTI).
- *Transmission time interval (TTI)/(eMBB) time slot:* Corresponds to 1 subframe duration ($1ms$) that is required to encapsulate non delay-sensitive data (transport blocks) from higher radio protocol stack layers and deliver it to the physical layer in order to transmit it via the radio interface.
- *Resource Block (RB):* In this paper, a RB in 5G NR is defined as 12 consecutive subcarriers in the frequency domain and 14 symbols in the time domain. With different sizes of slots and subcarriers of different numerologies, the size of the RB may change, as illustrated in Figure 1.

TABLE 2. 5G new radio numerologies [38].

Numerology, μ	SCS [KHz]	#symbols per slot	#slots per subframe	Cyclic prefix (CP)	Symbol duration [μs]	CP duration [μs]
0	15	14	1	Normal	71.43	4.69
1	30	14	2	Normal	35.71	2.34
2	60	14, 12	4	Normal, extended	17.86	1.17
3	120	14	8	Normal	8.92	0.57
4	240	14	16	Normal	4.46	0.29

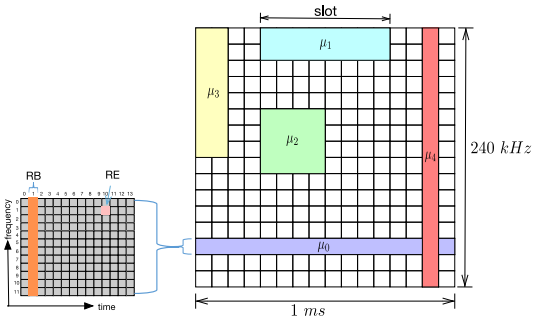


FIGURE 1. 5G flexible numerology structures.

- *Resource Element (RE):* RE is the smallest unit within the 5G NR resource grid, consisting of one subcarrier in the frequency domain and one OFDM symbol in the time domain.
- *Cyclic Prefix (CP):* CP is required to eliminate Inter-Symbol Interference (ISI) due to multipath signals. 5G NR supports both normal CP and extended CP. With a normal CP, each slot is formed by 14 OFDM symbols, while 12 are only available when using an extended CP.

5G NR can support a wide range of services, devices, and deployments. Another new concept in 5G NR is called Bandwidth Part (BWP). A BWP is a set of contiguous RBs configured inside a channel bandwidth; thus, the width of a BWP may be smaller than or equal to the channel bandwidth. The motivation behind introducing BWP is such that it could be challenging to use the larger 5G bandwidths for the UEs and to support UE bandwidth adaptation to help reduce device power consumption [39]. Besides, BWPs can be of various numerologies, which can be employed to decrease the latency for particular services, such as in URLLC use cases.

Employing multiple numerologies in the 5G NR enhances the flexibility of scheduling use cases with diverse service requirements via performing slicing in the RAN. With a shorter duration of slots, transmissions can be scheduled much faster than a traditional LTE-based network. Furthermore, NR enables both uplink and downlink transmissions within a slot, making it possible to support low latency traffic. In addition, different numerologies support multiple deployment scenarios from sub-1 GHz range to mmWave applications. The higher numerologies $\mu = 3$ and $\mu = 4$ support high frequencies in the mmWave range defined in the range of FR-2. Furthermore, since the symbol length and SCS are inversely proportional to each other, wider SCSs reduce the CP length, which is an overhead to a system. This is especially useful for smaller cells where delay spread is low. For applications that tolerate longer delay spread, narrower SCSs are preferable. In the work [40], the authors present a criterion for selecting the most preferable numerology for a set of services under given network conditions.

However, the flexibility provided by the multi-numerology structure of 5G NR introduces a non-orthogonality into the system. This causes interference between the multiplexed numerologies, known as Inter-Numerology Interference (INI). Several papers analyze INI and present various INI management techniques [22], [23], [41], [42], [43].

2) 5G MINI-SLOT

A slot is a unit for transmission used by the gNB scheduling mechanism. Typically, a slot occupies either 14 (normal CP) or 12 (extended CP) OFDM symbols (see Table 2). 5G NR allows transmissions with a flexible start position and a duration shorter than a regular slot duration, which is referred to as a mini-slot. A mini-slot is the minimum scheduling unit used in 5G NR. Though, in principle, a mini-slot can be as short as one OFDM symbol in Release 15, mini-slots are limited to 2, 4, and 7 OFDM symbols [44]. In low-latency scenarios such as URLLC, a transmission needs to begin immediately without waiting for the start of a slot. Mini-slot transmission facilitates achieving lower latency in 5G NR architecture and minimizes interference to other RF links. Unlike slots, mini-slots are not tied to the frame structure. This helps in puncturing the existing frame without waiting to be scheduled.

B. URLLC AND EMBB USE CASES

A vertical domain is an industry or group of enterprises in which similar products or services are developed, produced, and provided. The operation of a vertical industry is hugely interdependent on an Information and Communications Technology (ICT) infrastructure. Depending on the products or the services they deliver, the underlying ICT infrastructure must be designed in a way that it can accommodate all

TABLE 3. Communication service performance requirements for different URLLC use cases [45].

Use case	availability (%)	reliability (MTBF)	maximum E2E latency	data size (Bytes)
Motion control in factory automation	99, 9999	10 years	$1ms$ – $500ns$	-
Closed-loop control in process automation	99, 9999	>1 year	$<10ms$	-
Fault location, isolation, and restoration (FLISR)	99, 9999	-	$<5ms$	<1500
Wind power plant network	99, 9999999	10 years	$16ms$	-

service requirements demanded by such vertical industries. The vertical domains addressed in this paper are:

- *Power distribution:* Modern society is highly dependent on the reliability and resiliency of the power grid. The energy sector is currently subject to a fundamental change caused by the evolution toward renewable energy, i.e., an increasing number of power plants based on solar and wind power. These changes lead to bi-directional electricity flows and increased dynamics of the power system. New sensors and actuators are being deployed in the power system to efficiently monitor and control the volatile conditions of the grid, requiring real-time information exchange. The emerging power distribution grid is also referred to as the smart distribution grid. The smartness enhances insight into both the grid as a power network and the grid as a system of systems.
- *Power generation:* This domain comprises all aspects of centralized power generation, i.e., the centralized conversion of chemical energy and other forms of energy into electrical energy. Examples of pertinent systems are large gas turbines, steam turbines, combined-cycle power plants, and wind farms. In addition, the planning and installation of respective equipment and plants, as well as the operation, monitoring, and maintenance of these plants are encompassed by this vertical domain.
- *Manufacturing:* The manufacturing industry is currently subject to a fundamental change due to the fourth industrial revolution (Industry 4.0). It requires improvements in flexibility, versatility, resource efficiency, cost efficiency, worker support, and quality of industrial production and logistics in order to address the needs of increasingly volatile and globalized markets.

In the following, we present several use cases within selected 5G vertical industries requiring URLLC or eMBB communication services, which some of them will be considered later in a scenario for our simulations.

1) URLLC USE CASES

- *Factory automation in manufacturing:* Factory automation serves the automated control, monitoring, and optimization of processes in a factory. It deals with applications such as closed-loop control, motion controllers, robotics, and computer-integrated manufacturing. Factory automation is a key enabler for industrial mass production with high quality and cost-efficiency. Thus, related applications are characterized by strict requirements on the underlying communication infrastructure regarding availability and latency.
- *Process automation in manufacturing:* In the closed-loop control use case for process automation, several sensors are installed in a plant, and each sensor performs continuous measurements. The measurement data is transported to a controller, which takes a decision to set actuators. The latency and determinism in this use case are crucial. Therefore, this use case has very stringent requirements in terms of latency and service availability.
- *Fault Location, Isolation and Service Restoration (FLISR) in power distribution:* The FLISR is an essential operation to support the self-healing of power distribution grids. Typically, in power distribution grids, each feeder section has a controller device. Using Peer-to-Peer (P2P) communication among the Intelligent Electronic Devices (IEDs), the system operates autonomously without the intervention of the control center. In P2P communication via IEC 61850 GOOSE (Generic Object-Oriented Substation Event) messages are sent periodically (in steady-state) by each IED to neighboring IEDs of the same feeder and are not acknowledged. The data rate per IED is low in steady-state, but GOOSE bursts with high data rates occur, especially during fault situations, and require low E2E latency and high reliability.

Table 3 presents the different communication service performance requirements for different URLLC use cases mentioned above.

2) EMBB USE CASES

- *Remote grid surveillance in power distribution:* Critical infrastructures such as power distribution grids must be continuously monitored and controlled. Such critical infrastructures are heavily exposed to threats posed by malicious actors as well as potentially catastrophic natural disasters. As a result, there is a trend for smart distribution grids to incorporate video, photography, Unmanned Aerial Vehicles, and drones for visual surveillance for the supervision and observation of grid equipment.
- *Augmented (AR) or Virtual Reality (VR):* Use cases and applications also exist that require very high data rates

as offered by eMBB, such as augmented or virtual reality. Cloud-based AR/VR is the key technology enabling games, education, training.
- *Video streaming from event venues:* One potential application is large spectator events such as sports games or concerts, where spectators are located far away from the physical event location but are able to experience it, for instance, via live video streaming through social media. Also, the spectators are able to experience a front-row view of the action despite their physical location as a benefit of VR.

C. URLLC AND EMBB COEXISTENCE STRATEGIES

The incoming URLLC packets to a gNB have to be immediately sent through the scheduled eMBB transmissions and cannot be queued due to the strict latency requirements of URLLC traffic. The conventional orthogonal-based radio resource allocation mechanism is not suited for the coexistence of URLLC and eMBB traffic [46]. 3GPP defines two approaches for the coexistence of these heterogeneous services with distinct requirements.

1) DYNAMIC MULTIPLEXING

The superposition or puncturing scheme is one of the proposals from 3GPP to efficiently multiplex eMBB and URLLC data transmissions via the 5G NR [7]. eMBB traffic is scheduled at the beginning of slots. URLLC packets may arrive during an ongoing eMBB transmission, and URLLC traffic can be immediately overlapped at any mini-slot. If eMBB transmissions are allocated zero power when URLLC traffic is overlapped, then it is referred to as the puncturing of eMBB transmissions. If the gNB chooses non-zero transmission powers for both eMBB and overlapping URLLC traffic, that is referred to as the superposition of URLLC traffic over eMBB traffic. It is worth mentioning that there is a tradeoff between employing superposition instead of puncturing. Utilizing superposition will enhance the performance in terms of the eMBB sum rate. Nevertheless, this advantage comes with the cost of 1) eMBB user's interference over URLLC user resulting in increasing the risk of violating the URLLC reliability requirement, and 2) computational complexity in URLLC users due to performing the Successive Interference Cancellation (SIC) technique [47]. Besides, there is no guarantee of delivering fairness in allocating resources between eMBB users since the objective of superposition is to improve the eMBB sum rate and not necessarily to fulfill the minimum acceptable data rate of each eMBB user. There has been a solution by allocating more power to the URLLC user, compared to the eMBB user, in order to reduce bit-error-rate and therefore higher reliability, and eliminate using SIC in the URLLC user [48]. However, in this solution, it is assumed that the gNB allocates more power to the URLLC user. This method is against one of our objectives, as operating such a method results in disregarding accurate and optimum power allocation between eMBB and URLLC users. Another solution is employing the superposition or puncturing technique according to the gNB decision. This approach may not be feasible either, as URLLC traffic needs to be transmitted immediately. Due to decision time, switching time, and processing time between conducting superposition or puncturing technique by the gNB, employing such an approach can violate the low latency requirement of URLLC packets.

2) ORTHOGONAL SCHEDULING

The gNB pre-reserves a number of frequency channels for URLLC traffic. Two reservation mechanisms fall under the orthogonal scheduling; semi-static reservation and dynamic reservation [9]. In the semi-static scheme, the gNB intermittently broadcasts the frame structure configurations. However, in the dynamic reservation, the frame structure information is updated frequently and dynamically using the control channel of a scheduled user. The downside of this approach is that resources reserved for URLLC will be wasted in case there is no URLLC transmission. Furthermore, the dynamic scheme needs additional control overhead compared to the semi-static scheme.

D. ASSUMPTIONS

Several assumptions are considered in the problem formulation.

- We assume eMBB and URLLC downlink transmissions with different service requirements in terms of data rate, latency, and packet size. The URLLC traffic is coming from several URLLC priority classes. Each URLLC priority class contains data flow of a certain number of URLLC UEs that generate packets with a specific incoming rate (high and medium compared to the eMBB users), and they have a particular delay requirement.
- We focus on the dynamic puncturing of allocated resources to eMBB users by overlapping the URLLC traffic on the same radio resources.

III. SYSTEM MODEL

In this section, we explain the system model, we formulate the problem and we also present the proposed algorithm for eMBB/URLLC coexistence. Table 4 summarizes the notation used in this paper.

A. SYSTEM MODEL AND PROBLEM FORMULATION

We analyze and study downlink eMBB and URLLC traffic, i.e., transmitting traffic from a single gNB, that can operate with single or multiple antennas $j \in \mathcal{J} = \{1, 2, \ldots, J\}$, to User Equipment (UEs). For the sake of simplicity, we consider single antenna eMBB and URLLC UEs to envision Massive MIMO scenarios, as assumed in [49].

The gNB schedules the eMBB and URLLC traffic and transmits the corresponding data for each service type via its antennas towards eMBB and URLLC users over flat independent and identically distributed (i.i.d.) Rayleigh

TABLE 4. List of parameters used in the paper.

Notation	Meaning
$\mathcal{I}$	Set of URLLC traffic classes
μ	Set of numerologies
$\mathcal{J}$	Set of antennas
$\mathcal{K}$	Set of eMBB UEs
$\mathcal{N}$	Set of URLLC UEs
L	Maximum number of iterations
B_μ	Set of RBs in numerology μ
$\Delta f_{(\mu=\chi)}$	SCS of numerology $\mu = \chi$
m_μ	Mini-slot within the numerology μ
b_μ	RB b within the numerology μ
t_μ	Time slot t of numerology μ
$R^{eMBB}_{k_\mu}(t_\mu)$	Achievable rate in numerology μ for eMBB user k_μ at the time slot t_μ
$\phi^{eMBB}_{k_\mu}(t_\mu)$	Total amount of radio resources allocated to the eMBB user k_μ at time slot t_μ
$\gamma^{eMBB}_{k_\mu}(t_\mu)$	Total loss function (fraction of punctured RBs b_μ allocated to eMBB user k_μ at time slot t_μ)
$x_{k_\mu b_\mu}(t_\mu)$	Resource allocation coefficient for eMBB user k_μ
f_{b_μ}	Bandwidth of the RB b in numerology μ
$p^j_{k_\mu b_\mu}(t_\mu)$	Transmission power from the antenna j of the gNB over the RB b_μ to the eMBB user k_μ at time slot t_μ
$p^j_{n_{i\mu} b_\mu}(t_\mu)$	Transmission power from the antenna j of the gNB over the RB b_μ to the URLLC user n_i in time slot t_μ
$h^j_{k_\mu b_\mu}(t_\mu)$	Rayleigh fading channel gain of the transmission from the antenna j of the gNB over the RB b_μ to the eMBB user k_μ at time slot t_μ
$h^j_{n_{i\mu} b_\mu}(t_\mu)$	Rayleigh fading channel gain of the transmission from the antenna j of the gNB over the RB b_μ to the URLLC user n_i in time slot t_μ
$\sigma^2_{Total_{k_\mu}}$	Total interference and noise power impacts eMBB user k_μ
$\sigma^2_{Total_{n_{i\mu}}}$	Total interference and noise power impacts URLLC user $n_{i\mu}$
$D^i_{m_\mu, n_i}(t_\mu)$	Random variable indicating the number of incoming URLLC packets generated by n_i user in mini-slot m_μ
$\eta^i_{m_\mu, n_i}(t_\mu)$	Instantaneous packet size of URLLC UE $n_i \in \mathcal{N}_i = \{1, 2, ..., N_i\}$ belonging to the class i in the mini-slot m_μ of time slot t_μ
$D_{total}(t_\mu)$	Total incoming URLLC traffic at time slot t_μ
π_{k_μ}	Number of punctured mini-slots of eMBB user k in numerology μ
$th^{eMBB}(t_\mu)$	Puncturing rate threshold
θ^i_{max}	Outage probability threshold of the URLLC class i
R_{min}	Minimum acceptable eMBB data rate

fading channels. The gNB serves $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ total number of eMBB and $n \in \mathcal{N} = \{1, 2, \ldots, N\}$ total number of URLLC UEs within a set of numerologies $\mu \in \boldsymbol{\mu} = \{0, \ldots, 4\}$. Figure 2 illustrates a symbolic puncturing mechanism for the coexistence of eMBB and URLLC traffic classes for the numerologies $\mu = 0$, $\mu = 1$, and $\mu = 2$ with $\Delta f_{(\mu=0)} = 15$ KHz, $\Delta f_{(\mu=1)} = 30$ KHz, and $\Delta f_{(\mu=2)} = 60$ KHz, $f_{b_{\mu=0}} = 180$ KHz, $f_{b_{\mu=1}} = 360$ KHz, and $f_{b_{\mu=2}} = 720$ KHz, $\text{TTI}_{(\mu=0)} = 1$ ms, $\text{TTI}_{(\mu=1)} = 0.5$ ms, and $\text{TTI}_{(\mu=2)} = 0.25$ ms, respectively. Within each specific numerology μ,

- The time domain is split into equally spaced time slots (TTIs) for the eMBB UEs' transmissions. The time slot is then subdivided into a fixed number of M_μ equally spaced mini-slots (short TTIs) where $m_\mu \in \mathcal{M}_\mu = \{1, 2, \ldots, M_\mu\}$ denotes a mini-slot within the numerology μ.
- The radio resources in the frequency domain are divided into $b_\mu \in \mathcal{B}_\mu = \{1, 2, \ldots, B_\mu\}$ RBs. Each RB b_μ contains 12 sub-carriers in the frequency domain and 14 OFDM symbols in the time domain.
- We refer to each eMBB user as k_μ, since depending on the gNB configuration, each eMBB user can be served via different numerologies in the various corresponding time slots.

According to the incoming arrival rates, the latency, and reliability requirements of different URLLC use cases mentioned in Table 3, the URLLC UEs are sub-categorized into different traffic classes $i \in \mathcal{I} = \{1, 2, \ldots, I\}$. In each class i, a subset number of URLLC UEs N_i generate a traffic volume within mini-slot m_μ. Since there is no strict latency requirement for serving the eMBB users, the RBs are allocated to them at the beginning of each time slot. However, the sporadic URLLC requests can arrive at any time within a time slot, and due to the extreme latency requirement of such requests, the gNB needs to serve them immediately in a mini-slot instead of waiting for the next time slot. Therefore, the gNB punctures previously scheduled eMBB transmissions in mini-slots by applying zero power to these transmissions to serve the URLLC requests promptly.

The sporadic URLLC traffic impacts the previously scheduled eMBB users with the allocated radio resources in some mini-slots. Suppose there are two different URLLC traffic classes, i and $i+1$. Let us assume that these URLLC traffic classes arrive at the first mini-slot of the first time slot for the three numerologies $\mu = 0$, $\mu = 1$, and $\mu = 2$, as it is shown in Figure 2. gNB determines to map the URLLC traffic to the eMBB UEs 1, 3, and 5, i.e., $k = 1$, $k = 3$, and $k = 5$. In particular, the URLLC traffic of class i and $i+1$ punctures 1) the $m_0 = 1$ of the first slot of $k = 1$ with 2 OFDM symbols per mini-slot, 2) the $m_1 = 1$ of the first slot of $k = 3$ with 7 OFDM symbols per mini-slot, and 3) the $m_2 = 1$ of the first slot of $k = 5$ with 7 OFDM symbols per mini-slot. The same idea is repeated for the last mini-slots of the eMBB UEs $k = 2$, $k = 4$, and $k = 8$ in which URLLC traffic classes i and $i+1$ arrive randomly, and gNB determines to puncture them. The URLLC traffic of class i and $i+1$ punctures 1) the $m_0 = 7$ of the first slot of $k = 2$ with 2 OFDM symbols per mini-slot, 2) the $m_1 = 2$ of the first slot of $k = 4$ with 7 OFDM symbols per mini-slot, and 3) the $m_2 = 2$ of the first slot of $k = 8$ with 7 OFDM symbols per mini-slot. The idea is that the generated URLLC packets belonging to different URLLC

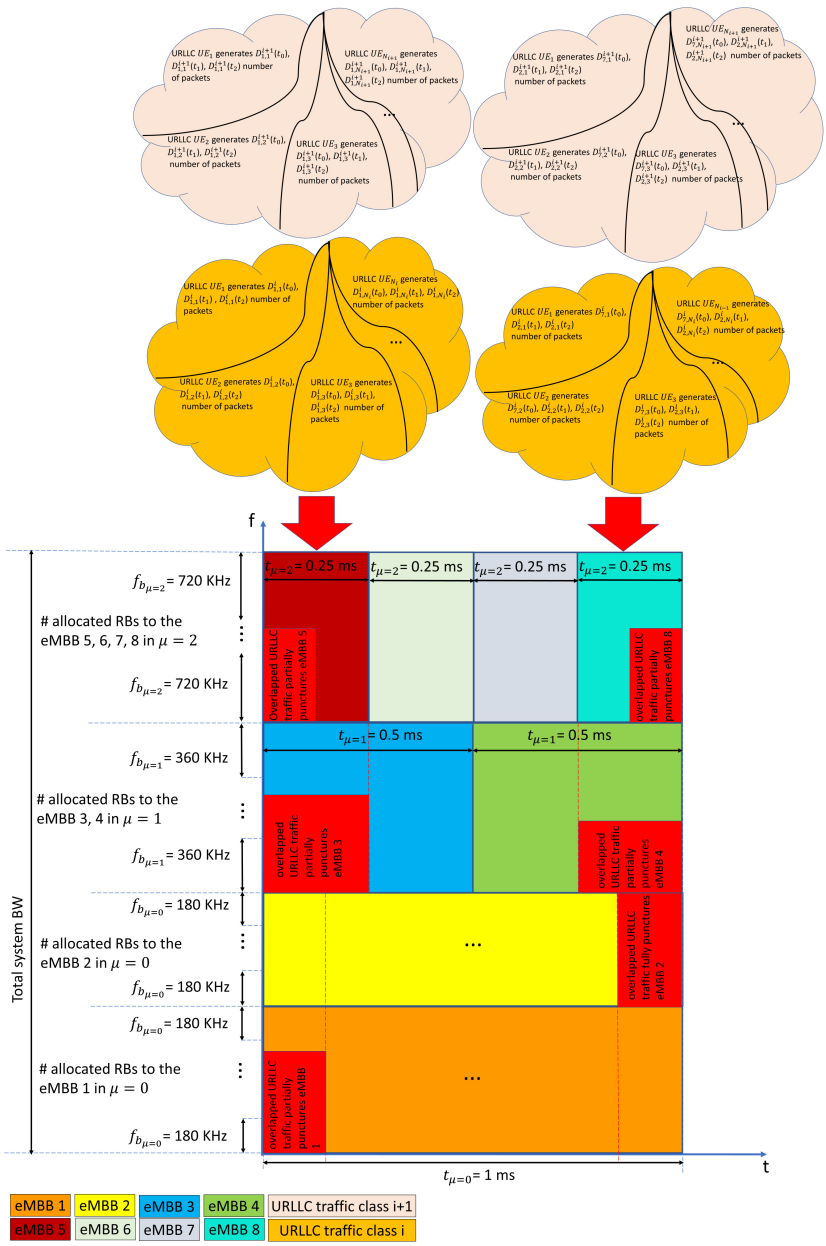


FIGURE 2. **Coexistence of eMBB and URLLC traffic classes in downlink via the puncturing mechanism for $\mu = 0$, $\mu = 1$, and $\mu = 2$ numerologies.**

classes are served. Hence, some of the allocated resources to the $k = 1, 2, 3, 4, 5, 8$ are punctured by the overlapped URLLC traffic.

Accordingly, the maximum achievable rate in the particular numerology μ for an eMBB user k_μ at the time slot t_μ over the whole allocated RBs can be formulated as follows:

$$R_{k_\mu}^{eMBB}(t_\mu) = \left[\phi_{k_\mu}^{eMBB}(t_\mu) - \gamma_{k_\mu}^{eMBB}(t_\mu)\right] \times R_{k_\mu,peak}^{eMBB}(t_\mu) \quad (1)$$

where the $\phi_{k_\mu}^{eMBB}(t_\mu)$ is the total amount of radio resources allocated to the eMBB user k_μ at time slot t_μ, $\gamma_{k_\mu}^{eMBB}(t_\mu)$ is the total loss function which indicates the fraction of punctured resources allocated to eMBB user k_μ at time slot

t_μ due to the incoming URLLC requests, and $R^{eMBB}_{k_\mu,peak}(t_\mu)$ is the peak achievable data rate of the eMBB user k_μ at time slot t_μ. This formulation is general, and by following the Shannon channel capacity, it can be further extended to:

$$R^{eMBB}_{k_\mu}(t_\mu) = \sum_{b_\mu=1}^{B_\mu}\Bigg[\Big(x_{k_\mu b_\mu}(t_\mu)f_{b_\mu} - \gamma^{eMBB}_{k_\mu b_\mu}(t_\mu)\Big) \times log_2\left(1+\frac{\sum_{j=1}^{J} p^j_{k_\mu b_\mu}(t_\mu)h^j_{k_\mu b_\mu}(t_\mu)}{\sigma^2_{Total_{k_\mu}}}\right)\Bigg] \tag{2}$$

where $x_{k_\mu b_\mu}(t_\mu)$ is the resource allocation coefficient, $x_{k_\mu b_\mu}(t_\mu) = 1$ denotes that the RB b_μ is allocated to the eMBB user k_μ at time slot t_μ and $x_{k_\mu b_\mu}(t_\mu) = 0$ shows no allocation; f_{b_μ} is the bandwidth of the RB b_μ; $p^j_{k_\mu b_\mu}(t_\mu)$ is the transmission power from the antenna j of the gNB over the RB b_μ to the eMBB user k_μ at time slot t_μ; $h^j_{k_\mu b_\mu}(t_\mu)$ is the Rayleigh fading channel gain of the transmission from the antenna j of the gNB over the RB b_μ to the eMBB user k_μ at time slot t_μ; $\sigma^2_{Total_{k_\mu}} = \sigma^2_{ICI_\mu} + \sigma^2_{ISI_\mu} + \sigma^2_{Cher_{k_\mu}} + \sigma^2_{INI_{k_\mu}} + \sigma^2_{Noise}$ indicates the summation of Inter-carrier Interference (ICI), Inter-symbol Interference (ISI), Channel estimation error (Cher), INI, and noise power, respectively [20], [50]; and finally, $\gamma^{eMBB}_{k_\mu b_\mu}(t_\mu)$ indicates the fraction of punctured RB b_μ that is allocated to eMBB user k_μ at time slot t_μ.

Now, in each specific numerology μ, we consider $D^i_{m_\mu,n_i}(t_\mu)$ as a random variable indicating the number of incoming packets per mini-slot duration and $\eta^i_{m_\mu,n_i}(t_\mu)$ as the instantaneous packet size of URLLC UE $n_i \in \mathcal{N}_i = \{1,2,\ldots,N_i\}$ belonging to the class i in the mini-slot m_μ of time slot t_μ. Hence, the total incoming URLLC traffic in the time slot t_μ is equal to $D_{total}(t_\mu) = \sum_{m_\mu=1}^{M_\mu}\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\eta^i_{m_\mu,n_i}(t_\mu)D^i_{m_\mu,n_i}(t_\mu)$. As a result, the $\gamma^{eMBB}_{k_\mu}(t_\mu)$ can be formulated as follows:

$$\begin{aligned}\gamma^{eMBB}_{k_\mu}(t_\mu) &= \sum_{b_\mu=1}^{B_\mu}\gamma^{eMBB}_{k_\mu b_\mu}(t_\mu)\\ &= \left[\sum_{b_\mu=1}^{B_\mu}x_{k_\mu b_\mu}(t_\mu)f_{b_\mu} \times \rho_{k_\mu b_\mu}(t_\mu)\frac{D_{total}(t_\mu)}{|B_\mu|\times|M_\mu|}\right]\end{aligned} \tag{3}$$

where $\rho_{k_\mu b_\mu}(t_\mu) \in [0,1]$ indicates the weight of puncturing, and $|B_\mu|\times|M_\mu|$ presents the total system capacity in terms of frequency-time resources. The URLLC traffic is upper bounded by total system capacity, i.e., $D_{total}(t_\mu) \le |B_\mu|\times|M_\mu|$. The $\rho_{k_\mu b_\mu}(t_\mu)$ identifies the pattern of overlapping total URLLC traffic in the time slot t_μ on the eMBB user k_μ resources in order to utilize (puncture) them for the URLLC transmission. The loss function is bounded $\gamma^{eMBB}_{k_\mu}(t_\mu) \in [0, \sum_{b_\mu=1}^{B_\mu}x_{k_\mu b_\mu}(t_\mu)f_{b_\mu}]$. Hence, for each eMBB user k_μ in time slot t_μ if:

- $\gamma^{eMBB}_{k_\mu}(t_\mu) = 0$, no puncturing;
- $0 < \gamma^{eMBB}_{k_\mu}(t_\mu) < \sum_{b_\mu=1}^{B_\mu}x_{k_\mu b_\mu}(t_\mu)f_{b_\mu}$, partial puncturing;
- $\gamma^{eMBB}_{k_\mu}(t_\mu) = \sum_{b_\mu=1}^{B_\mu}x_{k_\mu b_\mu}(t_\mu)f_{b_\mu}$, full puncturing happens.

Regarding URLLC traffic, the data rate of URLLC UE $n_i \in \mathcal{N}_i = \{1,2,\ldots,N_i\}$ belonging to the class i can be approximated as [51], [52]:

$$r^{URLLC}_{n_{i\mu}}(t_\mu) = \sum_{k_\mu=1}^{K_\mu}\Bigg[\left(\frac{\gamma^{eMBB}_{k_\mu}(t_\mu)}{\sum_{i=1}^{I}N_i} \times \log_2\left(1+\frac{\sum_{j=1}^{J}p^j_{n_{i\mu}b_\mu}(t_\mu)h^j_{n_{i\mu}b_\mu}(t_\mu)}{\sigma^2_{Total_{n_{i\mu}}}}\right)\right) - \sum_{b_\mu=1}^{B_\mu}\Psi^{URLLC}_{n_{i\mu}b_\mu}(t_\mu)\Bigg] \tag{4}$$

where $p^j_{n_{i\mu}b_\mu}(t_\mu)$ is the transmission power from the antenna j of the gNB over the RB b_μ to the URLLC user n_i in time slot t_μ; $h^j_{n_{i\mu}b_\mu}(t_\mu)$ is the Rayleigh fading channel gain of the transmission from the antenna j of the gNB over the RB b_μ to the URLLC user n_i in time slot t_μ; and $\sigma^2_{Total_{n_{i\mu}}}$ specifies total interference and noise power which negatively affects the URLLC user n_i in numerology μ. The $\Psi^{URLLC}_{n_{i\mu}b_\mu}(t_\mu)$ indicates the finite block-length channel coding regime in order to calculate the achievable rate of URLLC users which is given as:

$$\Psi^{URLLC}_{n_{i\mu}b_\mu}(t_\mu) = \sqrt{\frac{1-\left(1+\frac{p^j_{n_{i\mu}b_\mu}(t_\mu)h^j_{n_{i\mu}b_\mu}(t_\mu)}{\sigma^2_{Total_{n_{i\mu}}}}\right)^{-2}}{C^{URLLC}_{n_{i\mu}b_\mu}(t_\mu)}} \times \frac{Q^{-1}(\varepsilon^d_{n_i})}{ln(2)} \tag{5}$$

where $C^{URLLC}_{n_{i\mu}b_\mu}$ is the number of symbols in the mini-slot m_μ of time slot t_μ for the URLLC user n_i over the RB b_μ; and $Q^{-1}(.)$ is the inverse of the Gaussian Q-function of the decoding error probability for the URLLC user n_i.

The latency requirement of the URLLC user n_i over a particular numerology μ needs to satisfy the following [10], [21]:

$$\sum_{m_\mu=1}^{M_\mu}\eta^i_{m_\mu,n_i}(t_\mu).D^i_{m_\mu,n_i}(t_\mu) \le r^{URLLC}_{n_{i\mu}}(t_\mu) \tag{6}$$

which indicates that the achieved data rate of URLLC user n_i has to be greater than the total incoming URLLC traffic of this particular URLLC user n_i in the time slot t_μ in order to satisfy its latency requirement in the numerology μ.

Regarding the reliability condition, we should know that the requests from all the URLLC users n_i of all the classes I within time slot t_μ have to be served in order to ensure that the reliability is satisfied. Thus, if θ^i_{max} ($\theta^i_{max} \ll 1$) represents

the outage probability threshold of the URLLC class i, we can define the reliability for each class as follows [9]:

$$Pr\left[\sum_{n_i=1}^{N_i} r_{n_{i\mu}}^{URLLC}(t_\mu) \le \sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu,n_i}^{i}(t_\mu).D_{m_\mu,n_i}^{i}(t_\mu)\right] \le \theta_{max}^{i} \tag{7}$$

which shows that the probability, in which the total number of served URLLC users (N_i) is less than the incoming URLLC traffic of all the users within URLLC class i, has to be less than θ_{max}^{i} in order to satisfy the reliability requirement.

As a result, the objective of the optimal resource allocation problem is to maximize the sum data rate of eMBB users over all utilized numerologies while ensuring that the individual data rate of each eMBB user is lower bounded by the minimum acceptable eMBB data rate, i.e., R_{min} to guarantee the fairness between eMBB users. Besides, at the same time, the resource allocation problem is required to fulfill the URLLC UEs' requirements in terms of extra low latency and ultra-high reliability. Consequently, the sum data rate maximization problem is formulated as follows:

$$\boldsymbol{P_0}: \max_{x,p,\rho} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} \tag{8a}$$

subject to :

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu}\sum_{j=1}^{J} p_{k_\mu b_\mu}^{j}(t_\mu) + \sum_{\mu=0}^{4}\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\sum_{b_\mu=1}^{B_\mu}\sum_{j=1}^{J} p_{n_{i\mu}b_\mu}^{j}(t_\mu) \le P_{max},$$

$$p_{k_\mu b_\mu}^{j}(t_\mu) \ge 0,\ p_{n_{i\mu}b_\mu}^{j}(t_\mu) \ge 0 \tag{8b}$$

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu} x_{k_\mu b_\mu}(t_\mu) \le 1, \qquad x_{k_\mu b_\mu}(t_\mu) \in \{0,1\},$$

$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{8c}$$

$$\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) \le |B_\mu|, \forall \mu \in \boldsymbol{\mu} \tag{8d}$$

$$\rho_{k_\mu b_\mu}(t_\mu) \in [0,1]$$

$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{8e}$$

$$\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu,n_i}^{i}(t_\mu).D_{m_\mu,n_i}^{i}(t_\mu) \le r_{n_{i\mu}}^{URLLC}(t_\mu);$$

$$\forall n_i \in \mathcal{N}_i, \quad \forall i \in \mathcal{I}, \quad \forall \mu \in \boldsymbol{\mu} \tag{8f}$$

$$Pr\left[\sum_{n_i=1}^{N_i} r_{n_{i\mu}}^{URLLC}(t_\mu) \le \sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu,n_i}^{i}(t_\mu).D_{m_\mu,n_i}^{i}(t_\mu)\right] \le \theta_{max}^{i},$$

$$\forall i \in \mathcal{I}, \quad \forall \mu \in \boldsymbol{\mu} \tag{8g}$$

In this resource allocation problem, constraint (8b) defines the maximum transmission power budget via the gNB antennas in the downlink, P_{max}, towards all the eMBB and URLLC users. RBs' allocation among eMBB users is presented via constraint (8c). RBs' restriction in each numerology is presented by constraint (8d). Constraint (8e) indicates the weight of eMBB puncturing by the overlapped URLLC traffic. Finally, the latency and reliability requirements of the URLLC users are presented via (8f) and (8g), sequentially.

IV. OPTIMIZATION METHOD

In order to discover an optimal solution to the problem $\boldsymbol{P_0}$, it is necessary to find suitable mini-slots to position URLLC traffic on them while considering all potential RBs and power budget combinations of the eMBB users within different numerologies. Such a solution requires to satisfy eMBB users in terms of high data rate and, at the same time, URLLC users in terms of ultra-high reliability and extra low latency. Nonetheless, this procedure makes the solving approach complex as $\boldsymbol{P_0}$ is a non-convex problem. Since the optimization problem is mixed-integer nonlinear programming, we need to simplify this problem in order to reduce its complexity and make it solvable in a reasonable time. Hence, to find an appropriate solution to the $\boldsymbol{P_0}$ problem, we employ the decomposition and relaxation-based strategy for the eMBB and URLLC resource allocation problem. This results in converting $\boldsymbol{P_0}$ to a convex optimization problem. In this method, first, we decompose $\boldsymbol{P_0}$ into three sub-problems: $\boldsymbol{P_1}$ refers to the eMBB RBs allocation, $\boldsymbol{P_2}$ leads to the power allocation, and $\boldsymbol{P_3}$ considers URLLC traffic scheduling. Then, we relax the binary variable $x_{k_\mu b_\mu}(t_\mu)$ to a continuous variable in problem $\tilde{\boldsymbol{P}}_{\boldsymbol{1}}$. Then, the fractional solution is rounded to get a solution to the original integer problem, $\boldsymbol{P_1}$. Subsequently, we also utilize Markov's inequality expression in order to linearly estimate (8g) requirement. Finally, we prove the convexity of $\tilde{\boldsymbol{P}}_{\boldsymbol{1}}$, $\boldsymbol{P_2}$, and $\boldsymbol{P_3}$ sub-problems. Then each problem is solved individually based on its structure in order to achieve a practical solution with low computation complexity. The CVX toolbox [53], [54] is then used when solving each sub-problem.

A. EMBB RESOURCE BLOCK ALLOCATION PROBLEM

By decomposing $\boldsymbol{P_0}$ problem, while assuming p and ρ are constant values, the resource allocation problem, $\boldsymbol{P_1}$, is represented as follows:

$$\boldsymbol{P_1}: \max_{x} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} \tag{9a}$$

subject to :

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu} x_{k_\mu b_\mu}(t_\mu) \le 1, \quad x_{k_\mu b_\mu}(t_\mu) \in \{0,1\},$$

$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{9b}$$

$$\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) \le |B_\mu|, \forall \mu \in \boldsymbol{\mu} \tag{9c}$$

The existence of an integer variable in problem (9a) leads us to relax the $x_{k_\mu b_\mu}(t_\mu)$ to a continuous variable, $\bar{x}_{k_\mu b_\mu}(t_\mu)$, in order to avoid complexity in solving this problem. Hence, we can convert the original problem to $\bar{\boldsymbol{P}}_{\mathbf{1}}$ as follows:

$$\bar{\boldsymbol{P}}_{\mathbf{1}}: \max_{\bar{x}} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} + \upsilon\omega \tag{10a}$$

subject to :

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\bar{x}_{k_\mu b_\mu}(t_\mu) \le 1+\omega \qquad 0 \le \bar{x}_{k_\mu b_\mu}(t_\mu) \le 1,$$
$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{10b}$$

$$\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu}\bar{x}_{k_\mu b_\mu}(t_\mu) \le |B_\mu| + \omega, \forall \mu \in \boldsymbol{\mu} \tag{10c}$$

where $\omega = \max\{0, \sum_{k_\mu=1}^{K_\mu} x_{k_\mu b_\mu}(t_\mu) - 1\}$ is the rounding error value introduced by relaxing the integer variable, and υ is the weighting factor of ω with a negative value. The feasible solution to $\bar{\boldsymbol{P}}_{\mathbf{1}}$ is obtained with the minimum rounding error constraint, i.e., $\omega \to 0$.

Lemma 1: For constant values of p and ρ, $\bar{\boldsymbol{P}}_{\mathbf{1}}$ is a convex optimization problem.

Proof: It is worth noting that $R_{k_\mu}^{eMBB}(t_\mu)$ and its constraints are linear functions with respect to $\bar{x}_{k_\mu b_\mu}(t_\mu)$. The same applies to (10a) and its constraints, (10b) and (10c) concerning $\bar{x}_{k_\mu b_\mu}(t_\mu)$; thus, $\bar{\boldsymbol{P}}_{\mathbf{1}}$ is a convex optimization problem.

Finally, we need to convert the relaxed $\bar{x}_{k_\mu b_\mu}(t_\mu)$ variable back to the original binary variable $x_{k_\mu b_\mu}(t_\mu)$ after solving problem (9a). By determining $\alpha \in [0, 1]$ defined in [55], the conversion can be represented as:

$$x_{k_\mu b_\mu}(t_\mu) = \begin{cases} 1, & \text{if } \bar{x}_{k_\mu b_\mu}(t_\mu) \ge \alpha; \\ 0, & O.W. \end{cases} \tag{11}$$

■

B. EMBB POWER ALLOCATION PROBLEM

By decomposing $\boldsymbol{P_0}$ problem and presuming $\bar{x}$ and ρ as fixed values, the power allocation problem $\boldsymbol{P_2}$ is considered as follows:

$$\boldsymbol{P}_{\mathbf{2}}: \max_{p} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} \tag{12a}$$

subject to :

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu}\sum_{j=1}^{J} p_{k_\mu b_\mu}^{j}(t_\mu) + \sum_{\mu=0}^{4}\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\sum_{b_\mu=1}^{B_\mu}\sum_{j=1}^{J} p_{n_{i\mu} b_\mu}^{j}(t_\mu) \le P_{max},$$
$$p_{k_\mu b_\mu}^{j}(t_\mu) \ge 0, \ p_{n_{i\mu} b_\mu}^{j}(t_\mu) \ge 0 \tag{12b}$$

Lemma 2: For fixed values of $\bar{x}$ and ρ, $\boldsymbol{P_2}$ is a convex optimization problem.

Proof: We calculate the Hessian matrix of $R_{k_\mu}^{eMBB}$ in order to investigate whether it is a convex or a concave function. According to the definition of a semi-definite matrix, we need to calculate the result of $z^T \times \boldsymbol{H_R} \times z$, which is a real number. In this expression, z is a real column vector, z^T is the transpose of z, and $\boldsymbol{H_R}$ is the Hessian matrix of $R_{k_\mu}^{eMBB}$ which is defined as follows:

$$\boldsymbol{H_R} = \begin{bmatrix} \frac{\partial^2 R}{\partial \bar{x}^2} & \frac{\partial^2 R}{\partial \bar{x}\partial \rho} & \frac{\partial^2 R}{\partial \bar{x}\partial p} \\ \frac{\partial^2 R}{\partial \rho \partial \bar{x}} & \frac{\partial^2 R}{\partial \rho^2} & \frac{\partial^2 R}{\partial \rho \partial p} \\ \frac{\partial^2 R}{\partial p \partial \bar{x}} & \frac{\partial^2 R}{\partial p \partial \rho} & \frac{\partial^2 R}{\partial p^2} \end{bmatrix} \tag{13}$$

Since in $\boldsymbol{P_2}$ we consider the $\bar{x}$ and ρ as fixed values in $R_{k_\mu}^{eMBB}$, thus, all of the $\boldsymbol{H_R}$ elements except the $(\boldsymbol{H_R})_{3,3} = \frac{\partial^2 R}{\partial p^2}$ are zero. By taking the second-order derivative of the $R_{k_\mu}^{eMBB}$ with respect to $p_{k_\mu b_\mu}^{j}(t_\mu)$ we obtain:

$$(\boldsymbol{H_R})_{3,3} = \frac{\partial^2 R}{\partial p^2} = -\frac{\sum_{b_\mu=1}^{B_\mu}\left[\left(x_{k_\mu b_\mu}(t_\mu) f_{b_\mu} - \gamma_{k_\mu b_\mu}^{eMBB}(t_\mu)\right) \times \left(\frac{\sum_{j=1}^{J} h_{k_\mu b_\mu}^{j}(t_\mu)}{\sigma^2_{Total_{k_\mu}}}\right)^2\right]}{ln(2) \times \left(1 + \frac{\sum_{j=1}^{J} p_{k_\mu b_\mu}^{j}(t_\mu) h_{k_\mu b_\mu}^{j}(t_\mu)}{\sigma^2_{Total_{k_\mu}}}\right)^2} \tag{14}$$

which obviously is always negative for any $p_{k_\mu b_\mu}^{j}(t_\mu)$ value. Now we calculate the result of $z^T \boldsymbol{H_R} z$ as follows:

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \quad \forall z \in \mathbb{R}^3 \tag{15}$$

$$z^T \times \boldsymbol{H_R} \times z = z_3 (\boldsymbol{H_R})_{3,3} z_3 = (\boldsymbol{H_R})_{3,3} {z_3}^2 \le 0 \tag{16}$$

The result indicates that $\boldsymbol{H_R}$ is a negative semi-definite matrix, and consequently, $R_{k_\mu}^{eMBB}$ is a concave function. Since we want to maximize $\boldsymbol{P_2}$ and due to the linearity constraint of (12b) with respect to $p_{k_\mu b_\mu}^{j}(t_\mu)$, $\boldsymbol{P_2}$ is a convex optimization problem. ■

C. URLLC TRAFFIC SCHEDULING

In this section, first, we employ the Markov inequality expression [56] in order to simplify the constraint (8g) to a linear condition as follows:

$$Pr\left[\sum_{n_i=1}^{N_i} r_{n_{i\mu}}^{URLLC}(t_\mu) \le \sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu, n_i}^{i}(t_\mu).D_{m_\mu, n_i}^{i}(t_\mu)\right] \le \frac{\mathbb{E}\left[\sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu, n_i}^{i}(t_\mu).D_{m_\mu, n_i}^{i}(t_\mu)\right]}{\sum_{n_i=1}^{N_i} r_{n_{i\mu}}^{URLLC}(t_\mu)} \le \theta_{max}^{i} \tag{17}$$

Finally, decomposing $\boldsymbol{P_0}$ problem, while supposing $\bar{x}$ and p as invariant, yields in URLLC scheduling problem, $\boldsymbol{P_3}$, as follows:

$$\boldsymbol{P}_{\mathbf{3}}: \max_{\rho} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} \tag{18a}$$

subject to :

$$\rho_{k_\mu b_\mu}(t_\mu) \in [0,1]$$
$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{18b}$$

$$\sum_{m_\mu=1}^{M_\mu} \eta^i_{m_\mu,n_i}(t_\mu).D^i_{m_\mu,n_i}(t_\mu) \le r^{URLLC}_{n_{i\mu}}(t_\mu);$$
$$\forall n_i \in \mathcal{N}_i, \quad \forall i \in \mathcal{I} \tag{18c}$$

$$\frac{\mathbb{E}\left[\sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta^i_{m_\mu,n_i}(t_\mu).D^i_{m_\mu,n_i}(t_\mu)\right]}{\theta^i_{max}} \le \sum_{n_i=1}^{N_i} r^{URLLC}_{n_{i\mu}}(t_\mu),$$
$$\forall n_i \in \mathcal{N}_i, \quad \forall i \in \mathcal{I} \tag{18d}$$

Lemma 3: For invariant values of $\bar{x}$ and p, $\boldsymbol{P_3}$ is a convex optimization problem.

Proof: By considering the equivalent value of the loss function expressed in (3) and substituting it in (2), it is evident that $R^{eMBB}_{k_\mu}(t_\mu)$ is a linear function with respect to $\rho_{k_\mu b_\mu}(t_\mu)$. Since (18c) and (18d) are also linear constraints with respect to $\rho_{k_\mu b_\mu}(t_\mu)$; thus, $\boldsymbol{P_3}$ is a convex optimization problem. ■

D. MIMRA AS A SOLUTION OF PROBLEM (8a)

In this section, we present our proposed *Mixed-numerology Mini-slot based Resource Allocation (MiMRA)* algorithm to find an optimal solution for Eq. (8a). First, MiMRA algorithm converts $\boldsymbol{P_0}$ to $\bar{\boldsymbol{P_1}}$, $\boldsymbol{P_2}$, and $\boldsymbol{P_3}$ sub-problems by relaxation and decomposition technique. Next, the algorithm sets the minimum acceptable eMBB data rate, i.e., R_{min}, in order to guarantee fairness between eMBB users. In this way, the algorithm not only maximizes the sum rate of the eMBB users but also ensures that each individual eMBB user will achieve at least the R_{min} value. In each iteration l the algorithm searches for $\tilde{\boldsymbol{x}}^{(l)}, \boldsymbol{p}^{(l)}, \boldsymbol{\rho}^{(l)}$ as a solution from a feasible convex set. Then within each numerology μ, the algorithm specifies the number of punctured mini-slots for a particular eMBB user in that numerology, k_μ, as follows:

$$\pi_{k_\mu} = \left\lceil \frac{\rho_{k_\mu b_\mu}(t_\mu) D_{total}(t_\mu)}{|B_\mu|} \right\rceil \tag{19}$$

where $\pi_{k_\mu} \in \{0, 1, 2, \ldots, M_\mu\}$. The MiMRA algorithm also defines a *puncturing rate threshold*, i.e., $th^{eMBB}(t_\mu)$, according to the loss functions for all eMBB users. The selection criteria for calculating $th^{eMBB}(t_\mu)$ is as follows:

$$th^{eMBB}(t_\mu) = \begin{cases} \max_{\forall k_\mu \in \mathcal{K}_\mu}\left\{\gamma^{eMBB}_{k_\mu}(t_\mu)\right\}, \\ 0 \le \gamma^{eMBB}_{k_\mu}(t_\mu) < \sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) f_{b_\mu}; \\ \max_{\forall k_\mu \in \mathcal{K}_\mu}\left\{\gamma^{eMBB}_{k_\mu}(t_\mu)\right\} - offset_\mu, \\ \gamma^{eMBB}_{k_\mu}(t_\mu) = \sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) f_{b_\mu}; \end{cases} \tag{20}$$

where $offset_\mu$ indicates a constant value to tune $th^{eMBB}(t_\mu)$ if the second condition in (20) holds. After setting a value for $th^{eMBB}(t_\mu)$, the algorithm proceeds to calculate the number of punctured mini-slots for each eMBB user and the achievable data rate to verify whether each eMBB user can at least attain R_{min} or not. Thereafter, if $\sum_{\mu=0}^{4}\sum_{t_\mu=0}^{T_\mu} R^{eMBB}_{k_\mu}(t_\mu) < R_{min}$, then, depending on the $\rho_{k_\mu b_\mu}(t_\mu) \in [0,1]$, the algorithm maps part or the whole of incoming URLLC load, $D_{total}(t_\mu)$, to another possible eMBB user k'.

gNB holds the channel state information of the users. It continuously tracks the eMBB users within its coverage area by keeping a log of their channel conditions and the distance they are located. gNB searches to find the potential eMBB user k' with the allocated RB in the same numerology, b'_μ, or with RB in another numerology, $b'_{\mu'}$, if at least one of the following six conditions in Eq. (21) is fulfilled. Otherwise the algorithm runs for another round of iteration, $l+1$, to find another possible set of solutions, i.e., $\tilde{\boldsymbol{x}}^{(l+1)}, \boldsymbol{p}^{(l+1)}, \boldsymbol{\rho}^{(l+1)}$ until it converges.

$$\begin{cases} \text{a) if } k' \text{ is allocated a higher power than } k \text{ in } \mu \text{ or } \mu': \\ \begin{cases} p^{j'}_{k_\mu b'_\mu}(t_\mu) > p^{j}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu \\ p^{j'}_{k_{\mu'} b'_{\mu'}}(t_{\mu'}) > p^{j}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu' \end{cases} \\ \text{b) if } k' \text{ has a larger channel gain than } k \text{ in } \mu \text{ or } \mu': \\ \begin{cases} h^{j'}_{k_\mu b'_\mu}(t_\mu) > h^{j}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu \\ h^{j'}_{k_{\mu'} b'_{\mu'}}(t_{\mu'}) > h^{j}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu' \end{cases} \quad \mathbf{k} \xrightarrow{D_{total}(t_\mu)} \mathbf{k}' \\ \text{c) if } k' \text{ has a lower loss function than } k \text{ in } \mu \text{ or } \mu': \\ \begin{cases} \gamma^{eMBB}_{k_\mu b'_\mu}(t_\mu) < \gamma^{eMBB}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu \\ \gamma^{eMBB}_{k_{\mu'} b'_{\mu'}}(t_{\mu'}) < \gamma^{eMBB}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu' \end{cases} \end{cases} \tag{21}$$

As a result, the proposed algorithm tries to maximize the sum rate of all eMBB users while also considering each individual eMBB user to achieve the minimum acceptable data rate to fulfill the QoS requirement. Algorithm 1 summarizes the above steps.

E. COMPLEXITY ANALYSIS OF THE PROPOSED ALGORITHM

This subsection represents the computational complexity of the proposed MiMRA algorithm. In order to calculate the complexity of the algorithm, we notice that it is composed of some nested loops. The computational complexity of MiMRA is $\mathcal{O}(|L||\boldsymbol{\mu}||I||T_\mu||K_\mu||J|)$. Nevertheless, the $|\boldsymbol{\mu}|$, $|I|$, and $|J|$ have finite values and cannot get very high arbitrary values. As we know, the largest value of numerologies, $|\boldsymbol{\mu}|$, is 4. In addition, according to [57], the URLLC traffic classes, $|I|$, are mainly categorized into up to 8 different classes. Finally, according to [58], the maximum number of antennas that so far have been practically implemented in Massive MIMO base stations is 64. Consequently

Algorithm 1 MiMRA Algorithm for eMBB/URLLC Coexistence

Input: $\mu \in \boldsymbol{\mu}, i \in I, t \in T, b \in \mathcal{B}, k \in \mathcal{K}$, $j \in \mathcal{J}, h^{j}_{kb}(t), P_{max}$
Output: Solution to Eq. (8a) and providing fairness between eMBB users.
Relax $\boldsymbol{x}$ to $\bar{\boldsymbol{x}}$, and decompose $\boldsymbol{P_0}$ to $\bar{\boldsymbol{P}}_1$, $\boldsymbol{P_2}$, and $\boldsymbol{P_3}$
Set R_{min}
for $l \leftarrow 0$ to L **do**
  Find $\bar{\boldsymbol{x}}^{(l)}, \boldsymbol{p}^{(l)}, \boldsymbol{\rho}^{(l)}$ from feasible convex set as a solution of $\bar{\boldsymbol{P}}_1$, $\boldsymbol{P_2}$, and $\boldsymbol{P_3}$ respectively.
  Find $\boldsymbol{x}^{(l)}$ via Eq. (11).
  Define $th^{eMBB}(t_\mu)$ according to Eq. (20).
  for $\mu \in \boldsymbol{\mu}$ **do**
    for $i \in I$ **do**
      for $t_\mu \in T_\mu$ **do**
        for $k_\mu \in K_\mu$ **do**
          for $j \in J$ **do**
            Calculate π_{k_μ}, $R^{eMBB}_{k_\mu}(t_\mu)$ based on $th^{eMBB}(t_\mu)$
            if $\sum_{\mu=0}^{4} \sum_{t_\mu=0}^{T_\mu} R^{eMBB}_{k_\mu}(t_\mu) < R_{min}$ **then**
              According to $\rho_{k_\mu b_\mu}(t_\mu)$ map part or the whole $D_{total}(t_\mu)$ to k' in case of Eq. (21).
            else
              Go back to step: (7).
            end if
          end for
        end for
      end for
    end for
  end for
end for

the actual total computational complexity of MiMRA is $\mathcal{O}(|L||T_\mu||K_\mu|)$.

V. PERFORMANCE EVALUATION

In this section, we demonstrate the efficiency of our proposed algorithm through simulations and evaluate the performance of the algorithm.

A. NETWORK SCENARIO

We consider a shared 5G NR infrastructure with several URLLC and eMMB users in coexistence as illustrated in Figure 2. Notice that in this scenario, we would like to serve the generated URLLC packets belonging to different URLLC classes according to their priority, defined in Table 5. The common RAN physical resources are logically shared to transmit URLLC and eMBB traffic towards corresponding users in the downlink. The URLLC traffic is generated by the power distribution and manufacturing verticals belonging to two distinct URLLC classes. The eMBB traffic is produced from video streaming of a popular sport tournament. The different types of traffic are characterized by the following scenario which is used to determine the simulation parameters.

- URLLC traffic class 1: Generated by the IEDs placed in power distribution grids which broadcast GOOSE messages when an event (e.g., alarm, failure, or any mission-critical event) occurs [59]. We imagine that a failure occurs in the observed geographical area (which falls under the coverage of the gNB) and investigate the impact of the injected GOOSE messages into the network.
- URLLC traffic class 2: We assume a large manufacturing factory continuously operating in the same geographical area. Few sensors are installed inside the processing section to obtain measurements and perform process automation.
- eMBB traffic: Meantime, a largely popular sport event is assumed to be happening thus, several residents in the area are video streaming the live broadcast of the event with the HD quality up to 4K resolution.

TABLE 5. Simulation parameter configurations.

Type of traffic	URLLC class $i = 1$	URLLC class $i + 1 = 2$	eMBB
No. of users	10	10	20
Air latency [1] [ms]	1 [2]	5	-
Minimum data rate [Mbps]	1.5 [3]	1	-
Priority	high	medium	low
Traffic model	Poisson	Poisson	Full-buffered

[1] It is considered that standard air latency requirements are assigned to 20% of the corresponding E2E latency requirements [57], [60].
[2] Fault case.
[3] Fault case.

B. SIMULATION SETUP

We study and simulate the 5G RAN domain in a dense urban microcell scenario. In our simulated 5G NR, we consider one gNB operating in the FR-1 with 8 antennas towards the downlink, located at the center of the cell coverage zone with a 500 m radius. The operating center frequency is set to 3.5 GHz and $P_{max} = 40$ dBm. Several single antenna eMBB and URLLC users are randomly distributed within the coverage zone. Besides, the gNB schedules eMBB and URLLC transmissions over flat i.i.d Rayleigh fading channels. The remaining system parameters are listed in Table 5. In order to provide practical results comparable to realistic scenarios, the target KPI values of eMBB and URLLC services are extracted from specification documents [60].

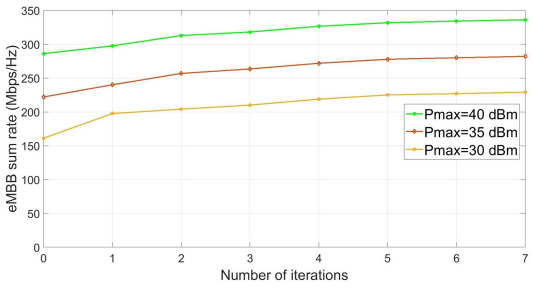


FIGURE 3. Convergence of MiMRA algorithm with various gNB maximum transmit power, P_{max} = {30, 35, 40} dBm, and R_{min} = 8Mbps.

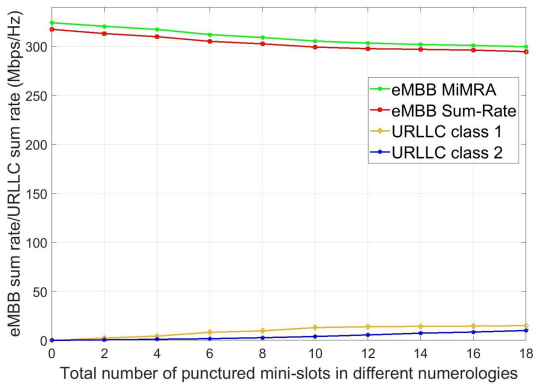


FIGURE 4. Achievable eMBB/URLLC sum rates over the total number of punctured mini-slots in different numerologies.

Extensive simulations are carried out under the following situations. It is considered that the gNB punctures the pre-scheduled eMBB traffic towards the corresponding users in the downlink by transmitting the overlapped URLLC traffic classes modeled as Poisson processes. In addition, air latency of the URLLC traffic classes is also considered in the simulation while assuming the eMBB traffic is not delay-sensitive compared to the URLLC traffic. The gNB utilizes numerologies $\mu = 0, 1, 2$ to transmit eMBB traffic over all of the available RBs in each numerology [40], [61]. The corresponding time slots for each numerology, $t_{\mu=0} = 1$ ms, $t_{\mu=1} = 0.5$ ms, and $t_{\mu=2} = 0.25$ ms, are subdivided into number of M_0, M_1, and M_2 mini-slots, respectively. The gNB punctures the eMBB traffic with URLLC traffic class $i = 1, i+1 = 2$ over all the utilized numerologies. We first evaluate the performance of the network in terms of the achievable sum rate of the eMBB and different classes of URLLC users under the total number of punctured mini-slots in different numerologies. Then we investigate the sum rate of the eMBB users for two various minimum acceptable rates per eMBB user under the number of URLLC packets generated from several URLLC users within class $i = 1, i+1 = 2$. After that, we repeat the previous evaluation, but this time under different gNB transmit power values. Next, we analyze the obtainable sum rate of the eMBB users for two diverse maximum allowed delay requirements of the URLLC users under different gNB transmit power values.

C. PERFORMANCE RESULTS

First, we evaluate the convergence speed of the MiMRA algorithm. As illustrated in Figure 3, we investigate how fast MiMRA converges according to the different values for the maximum transmit power, P_{max}, of gNB towards different users. We can observe that the eMBB sum rate converges fast and evolves to saturated status after around 5 iterations. Besides, it can be noticed that we have a higher eMBB sum rate employing higher transmit power. In particular, the eMBB sum rate can reach up to 336Mbps for P_{max} = 40 dBm. In contrast, the eMBB sum rate can obtain less value for smaller transmit power from the gNB. It is obvious that the reason is because of having a higher SNR value for larger P_{max}, which results in holding a larger eMBB data rate.

In Figure 4, we illustrate the achievable sum rates of URLLC classes $i = 1, i+1 = 2$ and eMBB users over the total number of punctured mini-slots in different numerologies. In this part, we set R_{min} = 15Mbps. The figure represents a gradual decrease/increase of the achievable eMBB/URLLC sum rates, respectively. In this figure, we compare the performance of the MiMRA algorithm with the baseline approach, *Sum-Rate* [62] scheduler, whose objective is to maximize the average sum rate of eMBB users via involving the puncturing strategy. As the figure reveals, the eMBB users achieve their maximum sum rate with no punctured mini-slots. By receiving URLLC traffic from either class, $i = 1$ or $i+1 = 2$ or both, gNB starts puncturing the eMBB users over various numerologies depending on the number of RBs required to fulfill the URLLC users, the priority of the URLLC users, and their latency requirements. As a result, the gNB assigns the demanded RBs to the URLLC class $i = 1$ and then URLLC class $i+1 = 2$ due to their QoS requirements. As the gNB tries to satisfy the URLLC users, the sum rate of the eMBB users decreases. Specifically, by puncturing up to 18 mini-slots in various numerologies and assigning the necessary RBs, the gNB serves the URLLC users of class $i = 1$ and $i + 1 = 2$ to reach their sum rate up to 15Mbps and 10Mbps, respectively. These sum rate values are appropriate to transmit the URLLC packets towards the corresponding URLLC users in different classes in the downlink. Regarding the eMBB users, as it can be seen from the figure, the MiMRA algorithm outperforms the *Sum-Rate* since even with a high number of punctured mini-slots (18 mini-slots), the MiMRA algorithm is still able to deliver the minimum acceptable data rate for each eMBB user R_{min} = 15Mbps to provide of up to 299.57Mbps as the sum rate of the eMBB users.

We next evaluate the sum rates of the eMBB users according to the allocated power from the gNB. We consider two

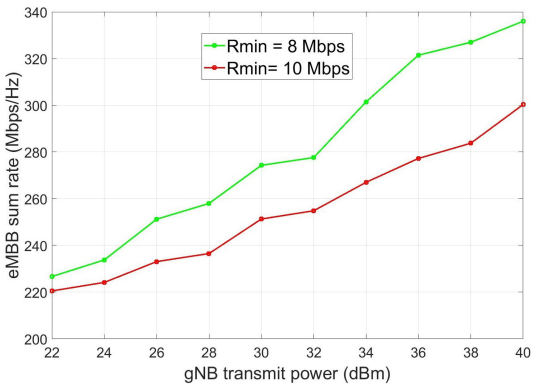


FIGURE 5. Achievable sum rates of eMBB users with two minimum acceptable data rate values versus different gNB transmit power.

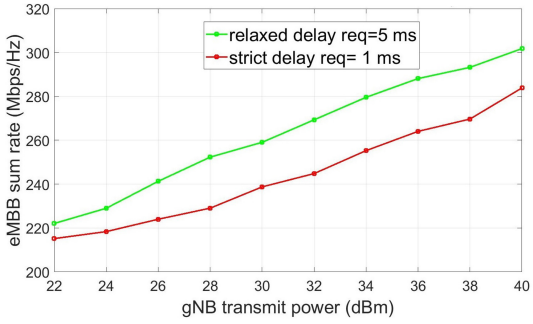


FIGURE 6. Achievable sum rates of eMBB users affected by traffic from two URLLC classes with two maximum allowed delay requirements versus different gNB transmit power.

scenarios. We set $R_{min} =$ 8Mbps and 10Mbps as the minimum acceptable data rates for the QoS requirements of the eMBB users in these scenarios. As Figure 5 illustrates, we can observe that higher QoS requirement for R_{min} results in lower sum rate performance due to stringent service requirements. The gNB potentially requires more resources to serve those highly-demanding eMBB users. Nevertheless, delivering the $R_{min} =$ 8Mbps for each eMBB user is more feasible and attainable rather than the $R_{min} =$ 10Mbps in case of increasing the number of eMBB and URLLC users.

Figure 6 demonstrates a follow-up graph to present the achievable sum rates of eMBB users impacted by traffic from two URLLC classes with two maximum delay requirements versus different gNB transmit power. As it can be observed, the sum rates of the eMBB users degrade with the stringent delay requirement. It means that in the case of strict latency, the system performance will confine more than the relaxed delay requirement case. The gNB requires more RBs with shorter time slots to satisfy the delay requirement of the URLLC class with stricter latency, 1ms, compared to the URLLC class with a relaxed latency, 5ms. Consequently, in

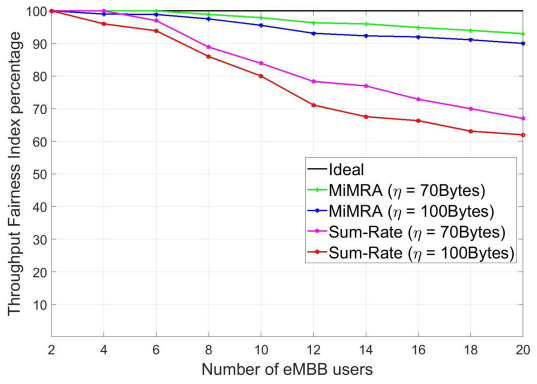


FIGURE 7. Fairness between eMBB users.

the case of incoming URLLC traffic with an even stricter value than 1 ms, the sum rates of the eMBB users are further reduced.

Figure 7 illustrates fairness in allocating the demanded resources between the eMBB users. We compare the performance of the MiMRA algorithm with the *Sum-Rate* scheduler under different packet sizes of the URLLC traffic classes. We want to investigate how much the fairness values provided by the MiMRA and *Sum-Rate* differ from the ideal (desired) case in which there is perfect fairness in allocating the required resources between the eMBB users. The fairness among the eMBB users is calculated based on Jain's Fairness index [63]. As it can be seen from the Figure, for a smaller packet size of $\eta =$ 70Bytes, the MiMRA algorithm performs well compared to the *Sum-Rate* outcome as MiMRA fairness is close to the ideal value (desired fairness) even for a large number of eMBB users (20 users). We observe the same performance for a larger packet size of $\eta =$ 100Bytes, as MiMRA again outperforms *Sum-Rate*. In both cases, there is a large difference between the fairness resulting from the MiMRA algorithm and the *Sum-Rate*. In particular, for $\eta =$ 70Bytes, MiMRA and *Sum-Rate* grant up to 93% and 67%, and for $\eta =$ 100Bytes, provide up to 90% and 62% fairness, respectively. This performance is due to Eq. (21) and considering $R_{min} =$ 8Mbps in order to find the perfect candidate for puncturing.

In Figure 8, we study the sum rates of the eMBB users versus the URLLC traffic load. In particular, we set $R_{min}=$ 6Mbps and 8Mbps as the two minimum acceptable data rates. Then, we evaluate the performance of the network for handling the incoming URLLC load with two maximum outage probability threshold values, $\theta^1_{max} = 0.001$ and $\theta^2_{max} = 0.01$, belonging to URLLC class $i = 1$ and $i+1 = 2$, respectively. As observed from the Figure, the eMBB users can reach high values for their sum rates when the number of URLLC packets is zero. As the incoming URLLC traffic classes arrive with different outage probabilities, the gNB punctures eMBB users to serve the URLLC traffic types. In

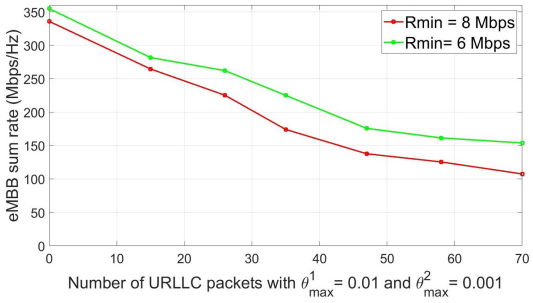


FIGURE 8. Achievable sum rates of eMBB users with two minimum acceptable data rate values URLLC load.

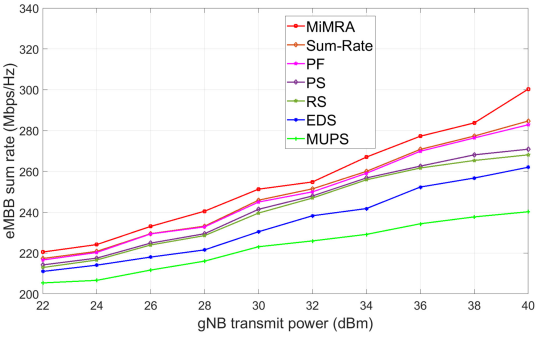


FIGURE 9. Comparison of eMBB sum rates with MiMRA and baselines for different gNB transmit power when R_{min} = 8Mbps.

such a situation, the gNB both tries to fulfill the requirements for two URLLC traffic classes, and to puncture eMBB users to the extent that each eMBB user can still achieve the minimum acceptable data rate. By considering the fairness provided by the MiMRA algorithm, for up to 38 URLLC packets, including both traffic classes, we can achieve the promised minimum data rate of up to 10.5Mbps per user if the R_{min} = 6Mbps. For the same number of URLLC packets, we reach up to 8Mbps per user if the R_{min} = 8Mbps. As the number of URLLC packets increases up to 70 packets, the gNB can still provide 7.65Mbps if the R_{min} = 6Mbps, but it ultimately can deliver up to utmost 5.37Mbps if the R_{min} = 8Mbps. It is worth mentioning that the MiMRA algorithm performs well since the sporadic behavior of URLLC traffic makes it rare to have such a high value of URLLC packets (70) in a concise period of a time slot.

Figure 9 illustrates the eMBB sum rates versus different gNB transmit power values. In particular, we compare the performance of MiMRA with 1) *Sum-Rate* that adopts a puncturing strategy to maximize the sum-rate of all eMBB users; 2) *Random Scheduler (RS)* [12] that transmits the incoming URLLC traffic by randomly picking pre-allocated RBs to the eMBB users; 3) *Proportional Fair (PF)* [64] which attempts to use the variations of channel conditions by

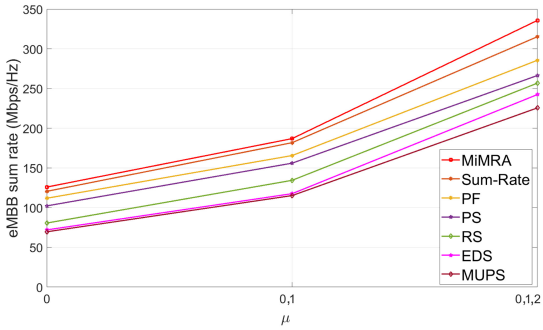


FIGURE 10. Comparison of MiMRA and baselines for eMBB sum rates versus different numerologies, μ, for R_{min} = 8Mbps, P_{max} = 40 dBm, and 10 URLLC packets.

assigning resources to users with the most suitable conditions for the upcoming period; 4) *Punctured Scheduling (PS)* [15] (also known as user-based puncturing) that chooses the RBs with the highest MCS assigned to eMBB users, and it punctures them to serve URLLC traffic; 5) *Equally Distributed Scheduler (EDS)* [10] which equally selects pre-allocated RBs to each of the eMBB users to serve the URLLC traffic; and 6) *Multi-User Preemptive Scheduling (MUPS)* [65]. We consider the incoming URLLC load θ^1_{max} = 0.001 and θ^2_{max} = 0.01, belonging to URLLC class $i = 1$ and $i + 1 = 2$, respectively and in total 28 packets. It is observed from the Figure that for lower power values, P_{max} = 22 dBm, MiMRA provides up to 220.5Mbps. Under the same condition and URLLC load, *Sum-Rate*, *PF*, *PS*, *RS*, *EDS*, and *MUPS* grant up to 217.1Mbps, 216.4Mbps, 214.2Mbps, 212.2Mbps, 210.8Mbps, and 205Mbps respectively. This exposes that there is a 3.4Mbps gap between MiMRA and the second best algorithm, which is *Sum-Rate*. Besides, since MiMRA exhibits high fairness, it can be inferred that each eMBB user can achieve up to at least 11.2Mbps, which is still higher than the R_{min} = 8Mbps. MiMRA utilizes a higher power value, P_{max} = 40 dBm, to deliver up to 300Mbps in order to perform even better than before with the price of consuming higher power. In this case, *Sum-Rate*, *PF PS*, *RS*, *EDS*, and *MUPS* provide 282.1Mbps, 281Mbps, 269.8Mbps, 267.9Mbps, 261.3Mbps, and 240.1Mbps, respectively. The difference between MiMRA performance and *Sum-Rate* is almost 17.9Mbps. Considering the high fairness of MiMRA, it can deliver up to 15Mbps. As a result, by keeping the data rate per eMBB user higher than or close to the R_{min}, the network guarantees that each eMBB user receives at least minimum resources, which are required for full HD video streaming with very high resolution with almost zero buffer time. In fact, with the MiMRA algorithm, the gNB does not permit to puncture any of the eMBB users completely, and it maintains the data rate at a level to avoid decreasing per eMBB data rate per user remarkably.

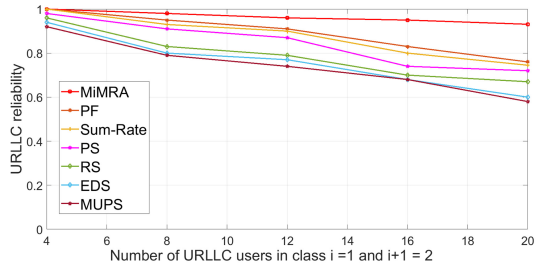
Figure 10 represents the performance of MiMRA in achieving a higher eMBB sum rate compared to baselines

with respect to the numerology values, μ. For $\mu = 0$, MiMRA shows a relatively similar value to what *Sum-Rate* delivers for the eMBB sum rate. However, MiMRA performance is slightly better than the *Sum-Rate*. As the value of μ increases, the eMBB sum rates of different approaches also grow. This is due to the increase in the SCS of RB with their respective numerologies. As the value of μ evolves, MiMRA outperforms the other solutions. In particular, MiMRA attains almost 336Mbps while *Sum-Rate* obtains 315Mbps over the employed numerologies.

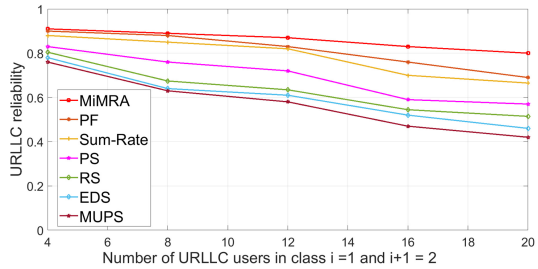
Usually in order to enhance *sum rate*, eMBB users with high channel gains need to be assigned more RBs than users with low channel gains. Nevertheless, this worsens the performance of eMBB users with poor channel conditions, especially if they are punctured by the incoming URLLC traffic as well. This results in acquiring significantly low fairness among the eMBB users. Hence, to maximize the *sum rate*, the number of RBs allocated to users with poor channel conditions has to be high. This creates a crucial dilemma between having high *sum rate* and fairness among the eMBB users [21]. As can be comprehended from the simulation results, the MiMRA algorithm resolves this challenge by setting R_{min} and defining $th^{eMBB}(t_\mu)$ to maximize the *sum rate* of the eMBB users and deliver fairness among them. Additionally, since the B5G era incorporates dealing with various URLLC use cases with distinct QoS requirements, MiMRA simultaneously ensures to fulfill diverse URLLC classes' demands for extra low latency and ultra-high reliability. Thus, URLLC traffic classes belonging to critical use cases are served with the highest priority, as discussed in the following.

Figure 11(a) and 11(b) illustrate the URLLC reliability of two classes (class 1 and class 2) with two packet sizes. As can be observed in 11(a), reliability drops as the number of URLLC users increases. In particular, with $\eta = 70$Bytes, for a few URLLC users (up to 12 users), MiMRA grants reliability of up to 96% while *PF* provides up to 90%. As the number of URLLC users grows (20 users), MiMRA still guarantees the URLLC reliability of up to 94% while *PF* can deliver maximum reliability of up to 76%. In the case of 11(b), we can see that the reliability reduces even further when the URLLC packet size increases. When the number of URLLC users is 20, MiMRA provides up to 80% reliability while the second best, *PF*, offers up to 69.8% reliability.

Figure 12 compares the URLLC delay Cumulative Distribution Function (CDF) of different baselines for 10 eMBB and 10 URLLC users of the mission-critical case with a delay requirement of 1ms. In particular, *MUPS* delivers the largest delay as it cannot counteract the strict URLLC delay requirement with an appropriate resource allocation that satisfies eMBB and URLLC users simultaneously. *EDS* provides the second largest delay, which still cannot satisfy the delay constraint of 1ms. Nevertheless, MiMRA outperforms other baseline solutions by meeting the delay requirement of URLLC packets.



(a) URLLC reliability for $\eta = 70$Bytes



(b) URLLC reliability for $\eta = 100$Bytes

FIGURE 11. Comparison of URLLC reliability with MiMRA and baselines versus different numbers of URLLC users with two URLLC packet sizes.

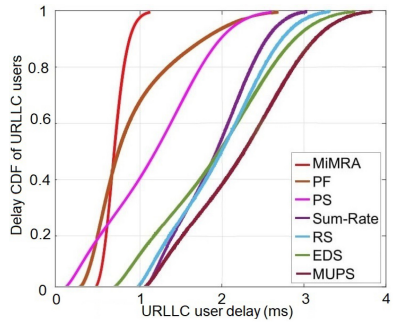


FIGURE 12. Delay CDF of the URLLC users in class $i = 1$ with a delay requirement of 1ms.

VI. CONCLUSION

In this paper, we propose an optimization framework to solve the resource allocation problem of coexisting URLLC and eMBB users in 5G NR over various numerologies. Furthermore, we study the impact of the incoming URLLC traffic, which is scheduled immediately into the mini-slots, instead of eMBB users, due to the stringent latency requirement. Our main goal is to maximize the sum rate for the eMBB users and to achieve a minimum acceptable data rate for the individual eMBB users ensuring fairness. The simulation results show that the proposed algorithm MiMRA enhances the sum rate of eMBB users while, at the same

time, each eMBB user can still achieve a minimum acceptable data rate. Thus, the eMBB users experience a more reliable transmission than the other studied approaches.

REFERENCES

[1] Q. Zhang and F. H. Fitzek, "Mission critical IoT communication in 5G," in *Proc. 1st Int. Conf. FABULOUS*, Ohrid, Republic of Macedonia, 2015, pp. 35–41.

[2] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead," *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107516.

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

[4] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN—Key technology enablers for 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2468–2478, Nov. 2017.

[5] P. Costa, M. Migliavacca, P. Pietzuch, and A. L. Wolf, "NaaS: Network-as-a-service in the cloud," in *Proc. 2nd USENIX Workshop Hot Topics Manage. Internet, Cloud, Enterprise Netw. Services (Hot-ICE)*, 2012, pp. 1–6.

[6] D. Gligoroski and K. Kralevska, "Expanded combinatorial designs as tool to model network slicing in 5G," *IEEE Access*, vol. 7, pp. 54879–54887, 2019.

[7] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Access Capabilities, V 16.3.0*, 3GPP Standard TS 36.306, 2020.

[8] X. Zhang, X. Guo, and H. Zhang, "RB allocation scheme for eMBB and URLLC coexistence in 5G and beyond," *Wireless Commun. Mobile Comput.*, vol. 2021, Oct. 2021, Art. no. 6644323.

[9] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.

[10] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, and C. S. Hong, "A matching based coexistence mechanism between eMBB and uRLLC in 5G wireless networks," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, 2019, pp. 2377–2384.

[11] A. K. Bairagi et al., "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, Mar. 2021.

[12] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 1970–1978.

[13] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue Maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019.

[14] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6.

[15] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, 2017, pp. 1–6.

[16] L. Marijanović, S. Schwarz, and M. Rupp, "Optimal resource allocation with flexible numerology," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, 2018, pp. 136–141.

[17] L. Marijanovic, S. Schwarz, and M. Rupp, "A novel optimization method for resource allocation based on mixed numerology," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[18] T. T. Nguyen, V. N. Ha, and L. B. Le, "Wireless scheduling for heterogeneous services with mixed numerology in 5G wireless networks," *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 410–413, Feb. 2020.

[19] "5G; study on new radio (NR) access technology," 3GPP, Sophia Antipolis, France, Rep. 38.912, 2018.

[20] L. Marijanović, S. Schwarz, and M. Rupp, "Multiplexing services in 5G and beyond: Optimal resource allocation based on mixed numerology and mini-slots," *IEEE Access*, vol. 8, pp. 209537–209555, 2020.

[21] Y. Prathyusha and T.-L. Sheu, "Coordinated resource allocations for eMBB and URLLC in 5G communication networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 8, pp. 8717–8728, Aug. 2022.

[22] M. Zambianco and G. Verticale, "Mixed-numerology interference-aware spectrum allocation for eMBB and URLLC network slices," in *Proc. 19th Mediterr. Commun. Comput. Netw. Conf. (MedComNet)*, 2021, pp. 1–8.

[23] M. Zambianco and G. Verticale, "A reinforcement learning agent for mixed-numerology interference-aware slice spectrum allocation with non-deterministic and deterministic traffic," *Comput. Commun.*, vol. 189, pp. 100–109, May 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366422000858

[24] M. Setayesh, S. Bahrami, and V. W. Wong, "Resource slicing for eMBB and URLLC services in radio access network using hierarchical deep learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 8950–8966, Nov. 2022.

[25] M. Mhedhbi, M. Morcos, A. Galindo-Serrano, and S. E. Elayoubi, "Performance evaluation of 5G radio configurations for industry 4.0," in *Proc. Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, 2019, pp. 1–6.

[26] D. Kotagiri, A. Sawabe, E. Takahashi, T. Iwai, T. Onishi, and Y. Nishikawa, "Context-based mixed-numerology profile selection for 5G and beyond," in *Proc. IEEE 19th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2022, pp. 611–616.

[27] K. Boutiba, M. Bagaa, and A. Ksentini, "Radio resource management in multi-numerology 5G new radio featuring network slicing," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 359–364.

[28] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient low complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G," in *Proc. IEEE 89th Veh. Technol. Conf. (VTC-Spring)*, 2019, pp. 1–6.

[29] A. Destounis and G. S. Paschos, "Complexity of URLLC scheduling and efficient approximation schemes," in *Proc. Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOPT)*, 2019, pp. 1–8.

[30] Y. Zhao, X. Chi, L. Qian, Y. Zhu, and F. Hou, "Resource allocation and slicing puncture in cellular networks with eMBB and URLLC terminals co-existence," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18431–18444, Oct. 2022.

[31] L.-H. Shen, C.-Y. Su, and K.-T. Feng, "CoMP enhanced subcarrier and power allocation for multi-numerology based 5G-NR networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 5460–5476, May 2022.

[32] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Superposition-based URLLC traffic scheduling in 5G and beyond wireless networks," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6295–6309, Sep. 2022.

[33] W. Ning, Y. Wang, M. Liu, Y. Chen, and X. Wang, "Mission-critical resource allocation with puncturing in industrial wireless networks under mixed services," *IEEE Access*, vol. 9, pp. 21870–21880, 2021.

[34] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, "Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1304–1319, Feb. 2022.

[35] G. Interdonato, S. Buzzi, C. D'Andrea, L. Venturino, C. D'Elia, and P. Vendittelli, "On the coexistence of eMBB and URLLC in multi-cell massive MIMO," 2023, *arXiv:2301.03575*.

[36] A. Esmaeily, K. Kralevska, and T. Mahmoodi, "Slicing scheduling for supporting critical traffic in beyond 5G," in *Proc. IEEE 19th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2022, pp. 637–643.

[37] *User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone*, 3GPP TS 36.101-1, R16, V16.0.0, Sep. 2020.

[38] "User equipment (UE) radio transmission and reception; part 4: Performance requirements," 3GPP, Sophia Antipolis, France, Rep. TR 38.101-4, V16.0.0, Mar. 2020.

[39] X. Lin, D. Yu, and H. Wiemann, "A primer on bandwidth parts in 5G new radio," 2020, *arXiv:2004.00761*.

[40] H. V. K. Mendis, P. E. Heegaard, V. Casares-Giner, F. Y. Li, and K. Kralevska, "Transient performance modelling of 5G slicing with mixed numerologies for smart grid traffic," in *Proc. IEEE 26th Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, 2021, pp. 1–7.

[41] A. B. Kihero, M. S. J. Solaija, and H. Arslan, "Inter-numerology interference for beyond 5G," *IEEE Access*, vol. 7, pp. 146512–146523, 2019.

[42] X. Zhang, L. Zhang, P. Xiao, D. Ma, J. Wei, and Y. Xin, "Mixed numerologies interference analysis and inter-numerology interference cancellation for windowed OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7047–7061, Aug. 2018.
[43] A. F. Demir and H. Arslan, "Inter-numerology interference management with adaptive guards: A cross-layer approach," *IEEE Access*, vol. 8, pp. 30378–30386, 2020.
[44] *5G NR; Base Station (BS) Radio Transmission and Reception*, 3GPP Standard TS 138 104, V15.2.0, Release 15, 2018.
[45] *Technical Specification Group Services and System Aspects; Service Requirements for Cyber-Physical Control Applications in Vertical Domains*, 3GPP Standard TS 22.104, V16.5.0, Sep. 2020.
[46] "Technical specification group services and system aspects," 3GPP, Sophia Antipolis, France, Rep. TR 21.915, V1.1.0, Release 15, Mar. 2019.
[47] K. Ying, J. M. Kowalski, T. Nogami, Z. Yin, and J. Sheng, "Coexistence of enhanced mobile broadband communications and ultra-reliable low-latency communications in mobile front-haul," in *Proc. Broadband Access Commun. Technol. XII*, Jan. 2018, Art. no. 105590C.
[48] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Joint resource and power allocation for URLLC-eMBB traffics multiplexing in 6G wireless networks," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.
[49] T. L. Marzetta, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
[50] R. Nissel and M. Rupp, "OFDM and FBMC-OQAM in doubly-selective channels: Calculating the bit error probability," *IEEE Commun. Lett.*, vol. 21, no. 6, pp. 1297–1300, Jun. 2017.
[51] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[52] J. Scarlett, V. Y. F. Tan, and G. Durisi, "The dispersion of nearest-neighbor decoding for additive non-Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 81–92, Jan. 2017.
[53] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[54] M. Grant and S. Boyd. "CVX: MATLAB software for disciplined convex programming." Accessed: Dec. 2022. [Online]. Available: http://cvxr.com/cvx/
[55] U. Feige, M. Feldman, and I. Talgam-Cohen, "Oblivious rounding and the integrality gap," in *Proc. Approx. Randomization, Combinatorial Optim. Algorithms Techn. (APPROX/RANDOM)*, 2016, pp. 1–28.
[56] O. Ibe, *Markov Processes for Stochastic Modeling*. Oxford, U.K.: Newnes, 2013.
[57] K. S. Kim et al., "Ultrareliable and low-latency communication techniques for tactile Internet services," *Proc. IEEE*, vol. 107, no. 2, pp. 376–393, Feb. 2019.
[58] "Massive MIMO systems for 5G and beyond networks." Accessed: Dec. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7284607/#:~:text=Theoretically%2C%20Massive%20MIMO%20systems%20can, in%20massive%20MIMO%20base%20station
[59] T. Bhattacharjee and M. Jamil, "GOOSE publishing and receiving operations of IEC 61850 enabled IEDs," in *Proc. IEEE 1st Int. Conf. Energy, Syst. Inf. Process. (ICESIP)*, 2019, pp. 1–6.
[60] "New services and applications with 5G ultra-reliable low latency communications." 5Gamericas. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2019/07/5G_Americas_URLLLC_White_Paper_Final__updateJW.pdf
[61] O. Aydin et al., "D4.1 draft air interface harmonization and user plane design," EU Project, Brussels, Belgium, document METIS-II/D4.1, May 2016.
[62] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "eMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr. 2019.
[63] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination," Eastern Res. Lab., Digit. Equip. Corp., Hudson, MA, USA, Rep. DEC-TR-301, 1984.
[64] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC traffic within eMBB services in 5G NR: Fair scheduling," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1080–1093, Feb. 2021.
[65] A. A. Esswie and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, 2018, pp. 136–141.

ALI ESMAEILY (Student Member, IEEE) received the master's degree in telecommunications engineering from the Polytechnic University of Catalonia in 2018. He is currently pursuing the Ph.D. degree with the Department of Information Security and Communication Technology, Norwegian University of Science and Technology. His research interests include cloud computing, service orchestration, 5G NR, and network slicing.

H. V. KALPANIE MENDIS received the bachelor's degree (Hons.) in electrical and information engineering from the University of Ruhuna, Sri Lanka, in 2015, and the master's degree in information and communication technology from the University of Agder, Norway, with a focus on 5G ultra-reliable communications. She is currently pursuing the Ph.D. degree with the Norwegian University of Science and Technology. Her research interests lie in the areas of 5G end-to-end network slicing, intent-based networking, multi-RAT architectures, software-defined networking, network function virtualization, management and orchestration of networks, and dependability.

TOKTAM MAHMOODI (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Iran, in 2002, and the Ph.D. degree in telecommunications from King's College London, U.K., in 2009, where she is currently the Head of the Centre for Telecommunications Research with the Department of Engineering. Her research interests include mobile communications, network intelligence, and low-latency networking.

KATINA KRALEVSKA (Member, IEEE) received the M.Sc. degree in mobile and wireless communications from Ss. Cyril and Methodius University, Skopje, Macedonia, in 2012, and the Ph.D. degree in telematics from the Norwegian University of Science and Technology, Trondheim, Norway, in 2016, where she has been an Associate Professor with the Department of Information Security and Communication Technology since 2018. She was the Deputy Head of the Department for two years from 2019 to 2020. Her research interests and activities lie in the areas of next-generation networks, coding theory, and blockchain.

ISBN 978-82-326-7512-8 (printed ver.)
ISBN 978-82-326-7511-1 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology