Doctoral Thesis

Network Slicing Management for 5G Radio Access Networks

Author:
Oscar Adamuz-Hinojosa

Supervisors:
Dr. Juan M. Lopez-Soler
Dr. Pablo Ameigeiras

A thesis submitted in fulfillment of the requirements
to obtain the International Doctor degree as part of the
Doctoral Program in Information and Communication Technologies
in the

Wireless and Multimedia Networking Lab Research Group
Department of Signal Theory, Telematics and Communications
University of Granada

Granada, 9th, March, 2022

Editor: Universidad de Granada. Tesis Doctorales
Autor: Óscar Adamuz Hinojosa
ISBN: 978-84-1117-337-7
URI: http://hdl.handle.net/10481/74957

"There will be obstacles. There will be doubters. There will be mistakes. But with hard work, there are no limits"

Michael Phelps

Acknowledgements

After an intensive period of four years, writing this acknowledgment means to me the finishing touch of my doctoral thesis. Carrying out this doctoral thesis has had a big impact on me, not only on a research level, but also on a personal level. Therefore, I would like to reflect on the people who have supported and helped me so much throughout this period. I would never have been able to finish this work without the guidance of my supervisors, the help from the members of my research group WiMuNet (Wireless and Multimedia Networking Lab), my labmates, and the support from my family and friends.

Firstly, I have been extremely lucky to have Prof. Pablo Ameigeiras Gutiérrez and Prof. Juan Manuel López Soler as supervisors of my doctoral thesis. They have cared so much about my work, and they have provided me useful comments and remarks through the learning process of this doctoral thesis. Prof. Pablo Ameigeiras Gutiérrez has worked side by side with me all along the doctoral journey. He has taught me practically everything I know about scientific research. There are not enough words to express my gratitude to him for his effort and admirable dedication during this journey. Prof. Juan Manuel López Soler has always pursued my interests as if they were his own and he has always shared with me his experience in his research life. He also consistently allowed this doctoral thesis to be my own work, but steered me in the right direction whenever he though I needed it. I could not have imagined having a better supervisors.

I would like to extend my gratitude to the remaining members of my research group WiMuNet for their comradeship. I am deeply indebted to Pablo Muñoz Luengo for supporting and helping me with this doctoral thesis. Although he has not been formally my supervisor, I am grateful to his valuable comments on this work. Without his guidance and support, the realization of this doctoral thesis would not be possible. Special thanks go out to Prof. Juan José Ramos Muñoz and Jorge Navarro Ortiz. Their doors were always open whenever I have run into a trouble spot or have had questions about my research.

I would like to express my sincere gratitude to my labmates. Special thanks go out to Jonathan Prados Garzón, which significantly helps me when I took my first steps on the research world. I will never forget when he willingly shared his precious time during the development of my first simulator. His debugging skills saved me weeks of work. I would like to express my sincere gratitude to Jose Antonio Ordóñez Lucena for his stimulating discussions, patience, and enthusiasm. I will never forget our anecdotes in the lab, especially the ones in the day we submitted our first manuscript. I miss your technical support from the day you signed with Telefónica but at the same time I wish you all the best in your career. I would also like to thank Pilar Andrés Maldonado for her invaluable assistance, responsiveness, and technical help. I miss those days you invited after work all the labmates to your house for dinner, playing video games and watching movies. I wish you all the best in your career in Nokia. Furthermore, I would like to thanks my labmates Lorena Chinchilla Romero and Natalia Chichilla Romero for their time and company. It is a pleasure to have them as colleagues during this long path. All the days are more enjoyable with you, especially the coffee breaks when we try to solve the newspaper's hieroglyphic.

Further, I would like to thank Dr. Vincenzo Sciancalepore and Dr. Xavier Costa Pérez for having me during the research stay and giving me the chance to work with them at NEC Laboratories Europe. Their support, ideas, and guidance made me to improve considerably my research skills. Despite all the issues related to the COVID-19 pandemic, they put all their effort to ease the development of my research stay.

I must express my very profound gratitude to my mother Josefina, my sister Lorena and my stepfather Paco for providing me with unfailing support and continuous encouragement throughout my entire life. Their love and never-ending support always encouraging me to do the best for my studies. I would also like to express my most sincere gratitude to Laura for her love and patience. She inspired me to complete this doctoral thesis. All of them have experienced my ups and downs during the development of this thesis and they have helped me keeping me harmonious. This doctoral thesis would not have been possible without their support.

Finally, I would like to thank to my friends who provided a much needed form of escape from my studies and helped me keeping things in perspective. They

have given me their wise counsel and sympathetic ears. You are always there for me.

Abstract

In the last years, the emergence of Fifth Generation (5G) mobile networks has boosted the society digitalization. Specifically, the research community has materialized its efforts in contributions that will allow the Mobile Network Operators (MNOs) to deploy novel services with diverging requirements in terms of performance and functionalities. The International Mobile Telecommunication (IMT)-2020 standard has grouped these services into three categories: enhanced Mobile Broadband (eMBB), ultra-Reliable Low Latency Communication (uRLLC) and massive Machine Type Communication (mMTC).

These novel services impose diverging and conflicting performance constraints that would difficult their coexistence into a "one-size-fits-all" network architecture. To address that issue, network slicing has emerged as a technological solution. It consists of logically splitting the MNO's physical network infrastructure into a set of independent virtual networks, denominated network slices, each tailored to support the specific requirements of a particular service.

To manage and orchestrate network slices, different Standards Developing Organizations (SDOs) have provided novel contributions in their standards according to their business focus and expertise field.

With respect to the Radio Access Network (RAN), the main standardization body is the Third Generation Partnership Project (3GPP). This SDO considers RAN slicing from a functional viewpoint, i.e., it focuses on the RAN functionalities which a RAN slice requires and how they must be configured to provide a customized service with specific performance and functional requirements. To manage the lifecycle of a RAN slice, the 3GPP has defined a set of management entities which will make use of pre-defined templates to automate the deployment and orchestration of RAN slices.

To achieve the flexibility and modularity that a RAN slice requires, some of its constituent network functions could be implemented by software, i.e., as Vir-

tualized Network Functions (VNFs). However, the management of VNFs and the orchestration of their underlying resources are out of the 3GPP expertise field. The SDO responsible for standardizing the mechanisms to virtualize the network functions is the European Telecommunications Standards Institute (ETSI), specifically the Network Function Virtualization (NFV) group. This SDO has mainly defined the NFV-Management and Orchestration (MANO) framework and the NFV templates. Using these pre-defined templates, the NFV-MANO can automatically instantiate, scale and release VNF instances and their underlying virtual resources.

To deploy and operate the RAN slice's constituents implemented as VNFs, the MNO must build understanding on NFV and how this technology links with the concept of network slicing from the 3GPP perspective. Under this context, one of the main objectives of this dissertation is to design an architectural framework which harmonizes the 3GPP and the ETSI-NFV viewpoints on network slicing to automate the lifecycle management of RAN slices.

To address this objective, this thesis first analyzes how NFV allows the MNO to manage the lifecycle of those RAN slice's network functions which are virtualized. Among all the lifecycle operations, this dissertation focuses on the scaling operation. The reason is the scaling operation comprises the main NFV-MANO management procedures, i.e., adding/removing the virtualized resources of a VNF instance, and deploying/terminating a VNF instance. In this thesis, we specifically analyze the different procedures the NFV-MANO may trigger to automatically scale VNFs. Furthermore, we also propose an ETSI-NFV compliant workflow that clarifies the interactions and information exchanges between the NFV-MANO entities for one representative scaling procedure.

The second step to design a 3GPP/ETSI-NFV-based architectural framework for RAN slicing is to link the 3GPP and NFV templates to automate the lifecycle management of RAN slices. To that end, we propose a description model that harmonizes the 3GPP and ETSI-NFV viewpoints on RAN slicing. Specifically, our model aims to enable the customization and deployment of the virtualized RAN slice's constituents over a multi-cellular environment. The proposed solution benefits from the reusability provided by the NFV descriptors to define the underlying resources of the RAN slice's constituents. To customize the behavior of a RAN slice, the most representative radio parameters to configure its func-

tionalities have been identified. Furthermore, to facilitate the comprehension of the proposal, we also provide an example of the description of three RAN slices for eMBB, mMTC and uRLLC services.

Another key aspect to manage RAN slices is how a 3GPP/ETSI-NFV-based architectural framework can use efficiently the available infrastructure resources to instantiate the virtualized RAN slice's constituents. To that end, this dissertation studies the key aspects for sharing the RAN functionalities implemented as VNFs among RAN slices. Specifically, we propose a description model to define the lifecycle management of shared RAN slice's constituents using the 3GPP and the NFV management templates.

Once the 3GPP/ETSI-NFV-based architectural framework for RAN slicing is designed, the next steps are related to analyze how this framework must determine the amount of MNO's infrastructure resources to instantiate and operate RAN slices. In this sense, the first challenge an MNO faces is how this framework must plan in advance a set of RAN slices requested by the network slice consumers. To address this challenge, the MNO must rely on mechanisms to translate the service requirements of each RAN slice onto the amount of infrastructure resources allocated to these RAN slices. This amount of resources must ensure the performance requirements of these RAN slices are met in the long term. Designing these mechanisms is challenging, especially if the radio interface is considered. The reason is the capacity provided to the RAN slices from this interface is variable since it depends on channel effects such as shadowing, fast-fading and/or interference.

Under this context, other objective of this dissertation is to address the planning of radio resources for multiple RAN slices. Specifically, we visit the problem of computing the radio resource quotas for the requested and already deployed RAN slices during a planning window. A radio resource quota is a bound for the amount of radio resources which the MNO may allocate to a RAN slice in a cell during the planning window. Each quota ensures a RAN slice has enough radio resources in a cell to meet its performance requirements throughout this planning window. In this thesis, we focus on scenarios where the MNO plans in advance RAN slices with performance requirements in terms of either Guaranteed Bit Rate (GBR) or latency and reliability.

In a scenario where the MNO plans RAN slices with GBR requirements, this

thesis assumes the MNO must also guarantee the probability of blocking a User Equipment (UE) session (i.e., the probability of rejecting a data session because there are not resources to satisfy its GBR) is below a certain threshold. Based on that, we propose an analytical model which given (a) the radio resource quota for a RAN slice; (b) the GBR for each UE session; and (c) the distributions for the UE session arrivals and the UE session duration, provides the UE blocking probability in an Orthogonal Frequency-Division Multiple Access (OFDMA) cell. Our model is based on a Multi-dimensional Erlang-B system. It meets the reversibility property which means the proposed model allows the adoption of an arbitrary distribution for the UE session duration. Furthermore, the proposed model considers an arbitrary distribution for the average channel quality within the cell. Another innovation is this model considers the channel gain of a packet scheduler when it dynamically allocates radio resources. The validation results show an estimation error for the UE blocking probability below 1.5%.

Considering the previous model, the next step is to design a planning mechanism to determine the radio resource quotas which ensure (a) the GBR requirements for each RAN slice in the long term, and (b) their UE blocking probabilities are below target upper bounds. To that end, we propose a mathematical framework based on game theory for planning the radio resources of several RAN slices in a multi-cellular environment. In this framework, the planning procedure is formulated as multiple ordinal potential games, one per requested or already deployed RAN slice. Detailed simulations have been performed to demonstrate the effectiveness of the proposed solution in terms of performance, adaptability, and renegotiation capability.

Focusing on a scenario where the MNO plans RAN slices with requirements in terms of latency and reliability, it must ensure the packet transmission delay for each RAN slice via the radio interface does not exceed a target upper bound with a certain probability, i.e., the violation probability. To model the delay bound for a RAN slice in a single cell, this thesis presents a Stochastic Network Calculus (SNC)-based model. It considers as inputs the radio resource quota for this RAN slice, the target violation probability, and its traffic demand. The validation results conducted after simulations show the proposed model provides an upper and conservative estimation of the amount of radio resources required by a RAN slice to achieve a delay bound with the imposed violation probability.

Finally, based on the SNC-based model, we propose a mathematical framework to plan in advance the radio resource quotas for several uRLLC RAN slices in a multi-cellular environment. This framework aims to compute the radio resource quotas of each RAN slice in such a way that the differences between the achieved delay bounds, i.e., those resulting from this assignment, and the target delay bounds are minimized. Detailed simulations have been performed to show the benefits of using the proposed framework in a scenario with radio resource scarcity.

Contents

List of Abbreviations

3GPP	Third Generation Partnership Project.
4G	Fourth Generation.
5G	Fifth Generation.
5G-PPP	5G Infrastructure Public Private Partnership.
5QI	5G Quality of Service Indicator.
AC	Admission Control.
ACTN	Abstraction and Control of Traffic Engineered Networks.
AGV	Automated Guided Vehicle.
AI	Artificial Intelligence.
AMC	Adaptive Modulation and Coding.
API	Application Programming Interface.
AR	Augmented Reality.
BBU	Base Band Unit.
BLER	Block Error Rate.
BSS	Business Support System.
BW	Bandwidth.
BWP	Bandwidth Part.
C-RAN	Cloud RAN.
CAPEX	Capital Expenditure.
CCDF	Complementary Cumulative Distribution Function.
CDF	Cumulative Distribution Function.
CDMA	Code Division Multiple Access.

CN Core Network.
CPRI Common Public Radio Interface.
CQI Channel Quality Indicator.
CTMC Continuos-Time Markov Chain.
CU Centralized Unit.

DC Dual Connectivity.
DCN Dedicated Core Network.
DL Downlink.
DNC Deterministic Network Calculus.
DRB Data Radio Bearer.
DRP Dynamic Resource Provisioning.
DRPA Dynamic Resource Provisioning Algorithm.
DU Distributed Unit.

E2E End-to-End.
EBB Exponentially Bounded Burstiness.
EBF Exponentially Bounded Fluctuation.
eCPRI evolved Common Public Radio Interface.
EDF Earliest Deadline First.
EM Element Manager.
eMBB enhanced Mobile Broadband.
eNB evolved NodeB.
ETSI European Telecommunications Standards Institute.

FCAPS Fault, Configuration, Accounting, Performance, and Security.
FCFS First-come First-served.
FDD Frequency Division Duplex.
FIFO First In First Out.

GBR Guaranteed Bit Rate.
gNB next Generation NodeB.
GSMA Global System for Mobile Communications Alliance.
GST Generic Slice Template.

HARQ Hybrid Automatic Repeat Request.

HDR High Data Rate.

IETF Internet Engineering Task Force.
IL Instantiation Level.
IMT International Mobile Telecommunication.
IoT Internet of Things.
IST Individual Slice Template.
ITU International Telecommunication Union.

KPI Key Performance Indicator.

LA Link Adaptation.
LAA Licensed-Assisted Access.
LOS Line-of-sight.
LTE Long Term Evolution.

MAC Medium Access Control.
MANO Management and Orchestration.
MCS Modulation and Coding Scheme.
MEC Multi-access Edge Computing.
MGF Moment Generating Function.
MIMO Multiple-input Multiple-output.
MM Mobility Management.
mMTC massive Machine Type Communication.
MNO Mobile Network Operator.
MVNO Mobile Virtual Network Operator.

NE Nash Equilibrium.
NEF Network Exposure Function.
NEST Network Slice Type.
NFMF Network Function Management Function.
NFV Network Function Virtualization.
NFVI Network Function Virtualization Infrastructure.
NFVI-PoP Network Function Virtualization Infrastructure Point of Presence.
NFVO Network Function Virtualization Orchestrator.
NG-RAN Next Generation Radio Access Network.

NIC Network Interface Card.
NLOS Non-Line-of-sight.
NMS Network Management System.
NR New Radio.
NRF Network Function Repository Function.
NS Network Service.
NS-IL Network Service Instantiation Level.
NSD Network Service Descriptor.
NSMF Network Slice Management Function.
NSSMF Network Slice Subnet Management Function.
NSST Network Slice Subnet Template.

OFDM Orthogonal Frequency Division Multiplexing.
OFDMA Orthogonal Frequency-Division Multiple Access.
OPEX Operational Expenditure.
OSS Operations Support System.
OTT Over-The-Top.

PDCP Packet Data Convergence Protocol.
PDF Probability Density Function.
PDU Protocol Data Unit.
PF Proportional Fair.
PHY Physical Layer.
PMF Probability Mass Function.
PNF Physical Network Function.
PNFD Physical Network Function Descriptor.
PRB Physical Resource Block.
PS Packet Scheduling.

QoS Quality of Service.

RAC Radio Admission Control.
RAN Radio Access Network.
RAT Radio Access Technology.
RB Resource Block.
RBG Resource Block Group.

RL Reinforcement Learning.
RLC Radio Link Control.
RRC Radio Resource Control.
RRH Radio Remote Header.
RRM Radio Resource Management.
RSRP Received Signal Received Power.
RSRQ Received Signal Received Quality.
RSSI Received Signal Strength Indication.
RU Radio Unit.

S-NSSAI Single Network Slice Selection Assistance Information.
SA Scaling Aspect.
SBA Service-Based Architecture.
SC Small Cell.
SD Slice Differentiator.
SDAP Service Data Adaptation Protocol.
SDN Software-Defined Networks.
SDO Standards Developing Organization.
SINR Signal-to-Interference-plus-Noise Ratio.
SL Scale Level.
SLA Service Level Agreement.
SNC Stochastic Network Calculus.
SON Self-Organizing Network.
SST Slice/Service Type.

TDD Time Division Duplex.
TN Transport Network.
TTI Transmission Time Interval.

UE User Equipment.
UL Uplink.
UM Unackowledge Mode.
uRLLC ultra-Reliable Low Latency Communication.

V2X Vehicle-to-Everything.
VBR Variable Bit Rate.

VCD Virtual Compute Descriptor.
vCPU virtualized CPU.
VDU Virtual Deployment Unit.
VIM Virtual Infrastructure Manager.
VL Virtual Link.
VLD Virtual Link Descriptor.
VM Virtual Machine.
VNF Virtualized Network Function.
VNF-IL Virtualized Network Function Instantiation Level.
VNFC Virtualized Network Function Component.
VNFD Virtualized Network Function Descriptor.
VNFFG Virtual Network Function Forwarding Graph.
VNFFGD Virtual Network Function Forwarding Graph Descriptor.
VNFM Virtual Network Function Manager.
vNIC virtual Network Interface Card.
VPN Virtual Private Network.
VR Virtual Reality.
VSD Virtual Storage Descriptor.

WCDMA Wideband Code Division Multiple Access.
WiMAX Worldwide Interoperability for Microwave Access.
WP Work Package.

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Setting the Scene

1.1 The Digital Transformation of Industry Verticals

Fifth Generation (5G) mobile networks have emerged as a technological solution to boost the society digitalization. Specifically, the industry and academy are materializing their efforts in standards and contributions that will allow the Mobile Network Operators (MNOs) to deploy novel communication services with diverging requirements in terms of performance and functionality. This is opening up innovation opportunities for industry verticals [1]. For example, the authors of [2] have identified more than 200 industry digitization use cases enabled or partially enhanced by 5G mobile networks. These disruptive use cases may be found in sectors such as agriculture, tourism (e.g., museums), transportation, healthcare, education (e.g., convention centers), retail industry (e.g., shopping malls), transport hubs (e.g., ports and airports), sport facilities (e.g., stadiums), energy industry, military bases, or manufacturing. In order to simplify the categorization of these communication services, the International Mobile Telecommunication (IMT)-2020 standard has grouped them into the following three categories [3]:

- **enhanced Mobile Broadband (eMBB)**: This category comprises human-centric use cases for accessing to multimedia content, services and data. The Fourth Generation (4G) mobile networks have been designed and optimized to support broadband services. However, the continued increasing demand for these communication services and the emerging of enhanced

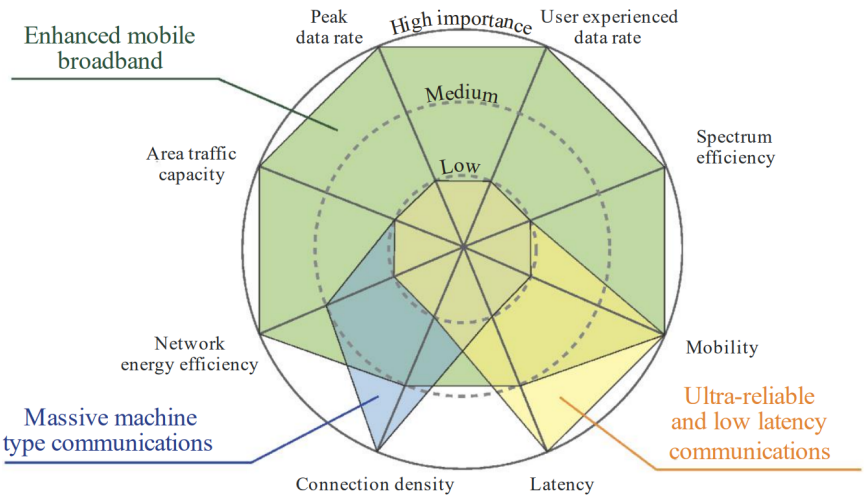


Figure 1.1: 5G communication service categories and their key capabilities [3]

multimedia and entertainment services will soon consume the capacity available in 4G mobile networks. Examples of communication services within this category are ultra-high definition video, augmented reality, or smart offices.

- **massive Machine Type Communication (mMTC)**: This category comprises those communication services in which the devices communicate each others without human intervention. Furthermore, these communication services are characterized by a huge number of connected low-cost devices equipped with long-life batteries, typically transmitting infrequent, small, and non-delay-sensitive data. The operation of the 4G mobile networks is not optimized for these communication services. Additionally, 4G mobile networks offer a poor connection density. Examples of communication services within this category are smart city, smart agriculture or smart metering.

- **ultra-Reliable Low Latency Communication (uRLLC)**: This category comprises communication services with stringent requirements in terms of latency and reliability. Furthermore, these services cover both human and machine-centric communications. The operation of the 4G mobile networks is also not optimized for this service category. Examples of

communication services within this category are: self-driving cars, industrial automation or remote surgery.

Fig. 1.1 summarizes the three main categories for 5G communication services and their main Key Performance Indicators (KPIs).

1.2 Network Slicing as Main Enabler to Provide Industry Vertical Services in 5G Mobile Networks

The novel 5G communication services impose diverging and conflicting performance constraints that would difficult their coexistence into a "one-size-fits-all" network architecture, as the one defined in the 4G mobile networks [4]. To address this issue, network slicing has been proposed as a technological solution. It consists of logically splitting the MNOs' physical network infrastructure into a set of independent virtual networks, denominated network slices, each tailored to support the specific requirements of a particular communication service [5]. In Fig. 1.2, we show an illustrative example of the network slicing concept. Specifically, we depict three network slices providing Internet of Things (IoT), healthcare, and mobile broadband services, respectively, over the MNO's physical network infrastructure.

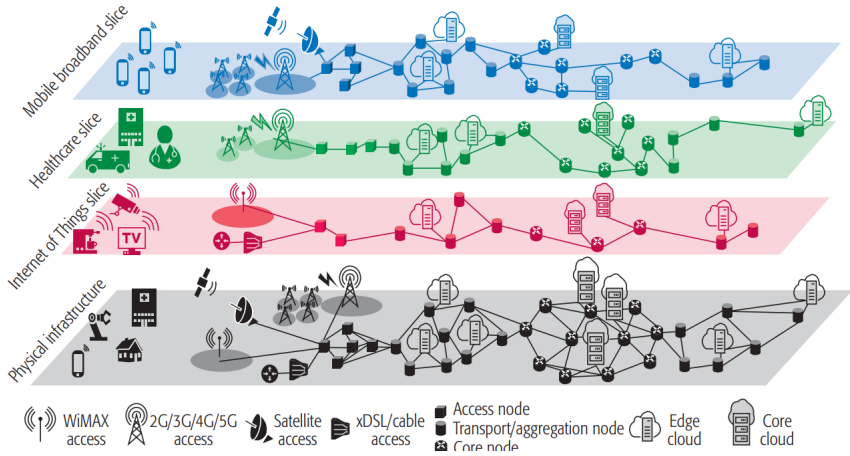


Figure 1.2: Three 5G network slices running over a common physical network infrastructure [5].

The notion of splitting a physical network infrastructure into logical networks is not a new concept in mobile networks [6]. Focusing on the Core Network (CN), the Third Generation Partnership Project (3GPP) has already standardized a network feature known as Dedicated core (DECOR). It allows the MNOs to deploy multiple CNs over the same physical network infrastructure while some levels of flexibility and resource sharing are offered to different service consumers. Network slicing goes beyond DECOR by allowing the MNO full flexibility to manage multiple communication services, each with specific performance requirements. This is possible through the use of "softwarization" technologies such as Network Function Virtualization (NFV) and Software-Defined Networks (SDN). The former allows the MNO, through the virtualization of network functions, to provide a modular logical architecture and a flexible placement of network functions in the MNO's infrastructure. The latter allows the MNO to simplify the forwarding functions and to provide a more advanced separation of control and user plane functionalities. Section 1.3 provides more details on these technologies.

Concerning the Radio Access Network (RAN) segment, there exist RAN sharing techniques [7]. They mainly consist of sharing the spectrum and/or the RAN infrastructure among two or more MNOs. Network slicing goes beyond the notion of RAN sharing [8]. Specifically, network slicing does not only allow the MNO to share its radio resources and physical infrastructure with other MNOs but also to create virtual RAN instances on demand with tailored sets of network functions (e.g., scheduling, mobility management). These network functions will better suit individual service requirements while these RAN instances are isolated among each other.

To bring network slicing to the 5G mobile network, the MNO has to satisfy a set of conditions and principles. The most important are summarized as follow [9]:

- **Management automation**: It enables the MNO to configure, deploy, operate and terminate network slices without human intervention. To that end, these management tasks must rely on signaling-based mechanisms which continuously check the performance and functional requirements of each network slice are met throughout its lifetime.

- **Isolation among network slices**: It is defined in terms of performance,

security and management [5]:

  - Performance: The MNO must ensure that the specific performance requirements of a network slice are always met, regardless of the congestion and performance levels of other network slices.
  - Security: Security attacks in one network slice cannot impact on other network slices. Moreover, each network slice should have independent security functions that prevent unauthorized entities to read or write the specific network slice configuration.
  - Management: From the viewpoint of a network slice consumer, each network slice may be managed as a separate network. To that end, the MNO must define a set of mechanisms which allow each network slice consumer to individually access to certain network slice capabilities.

- **Customization**: It assures the MNO uses efficiently the network functions and infrastructure resources allocated to a network slice while their functional and performance requirements are met throughout its lifetime.

- **Elasticity**: It ensures the MNO satisfies the performance requirements of a network slice under varying: (a) radio and network conditions, (b) amount of attached User Equipments (UEs), (c) traffic demand, and (d) geographical serving area because of UE mobility.

- **Programmability**: It allows the network slice consumers to control the allocated resources for each network slice and its configuration via open Application Programming Interfaces (APIs) that expose network slice capabilities. This facilitates on-demand service customization and resource elasticity.

- **End-to-End (E2E)**: It allows the MNO to deploy a network slice along different administrative, i.e., infrastructure locations managed by different providers, and network domains, i.e., CN, RAN, and Transport Network (TN).

- **Hierarchical abstraction**: It allows the MNO to execute a resource abstraction procedure, which may be repeated in successively higher levels.

The goal of this procedure is the resources of a network slice, allocated to a particular network slice consumer, can be further traded to another third player.

1.3 Technologies Related to Network Slicing

This section provides a brief overview of the main technologies which will enable the MNO to deploy and operate network slices.

1.3.1 Network Function Virtualization (NFV)

In traditional networks there is a strong coupling between the network functions (e.g., load balancing, mobility support, etc) and vendor-dependent proprietary hardware. This approach presents some drawbacks. On the one hand, every time a network operator requires a network upgrade or a new network function it has to (a) acquire new expensive proprietary hardware devices; and (b) find room, power supply and cooling systems for these hardware devices. On the other hand, it requires highly qualified personnel to design, integrate, and operate an increasingly complex network infrastructure [10]. Thus, this approach makes it difficult for operators to include new network features and services with agility and reduced costs.

To overcome the mentioned issues, NFV has emerged as a pivotal technological solution. As Fig. 1.3 depicts, this technology consists of decoupling network functions from proprietary hardware and enabling them to run as software components, denominated as Virtualized Network Functions (VNFs) on commodity servers [10, 12]. This means each network function which a network slice requires could be built with one or more virtualization containers, e.g., Virtual Machines (VMs) and/or LinuX containers (LXC).

The NFV technology introduces some differences with respect to the traditional approach in the way the MNO could provision a communication service. The most significant differences are summarized below [13]:

- **Decoupling software from hardware**: Since network functions are no longer a collection of integrated specific-purpose hardware and software entities, evolution of both are independent of each other. It involves the

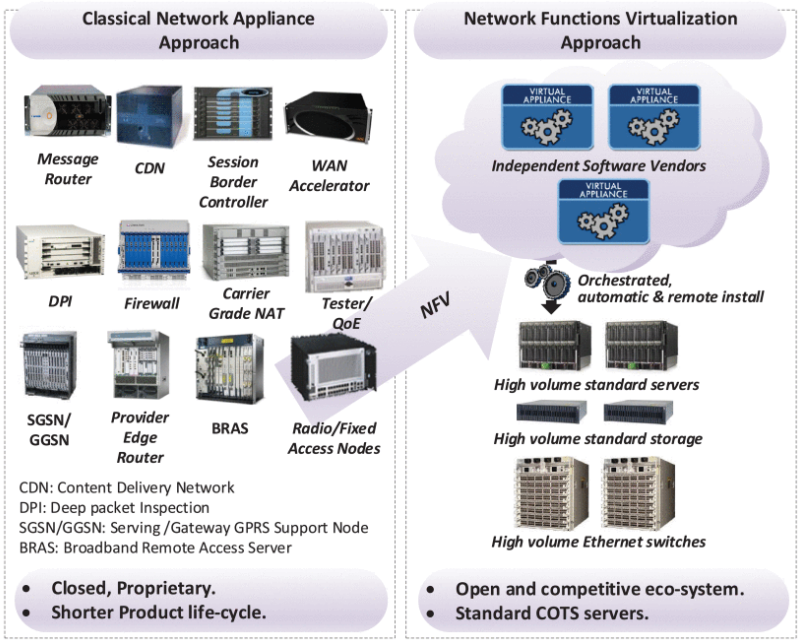


Figure 1.3: Traditional hardware-based network appliances approach versus NFV approach [11].

MNO might make progress on software updates separately from hardware updates, and vice versa.

- **Flexible deployment of network functions**: The software decoupling from hardware helps the MNO to reassign and share the resources of its physical network infrastructure to the different communication services. It might help the MNO to deploy new network services faster over the same network infrastructure.

- **Dynamic scaling**: Decoupling network functionalities into instantiable software components provides the MNO a greater flexibility to scale actual performance of these network functions in more dynamic way and with finer granularity according the incoming traffic demand. This means the MNO could change on the fly the amount of virtualized resources allocated for a network function according its input traffic load.

1.3.2 Software Defined Networking (SDN)

SDN is a technology which is mainly based on four principles: (a) the decoupling of control and data planes, (b) the logical centralization of the control plane, (c) the programmability of the network, and (d) the use of open interfaces [14]. As Fig. 1.4 shows, SDN completely separates the control and user planes in a network infrastructure enabling in this way the network programmability. This means the MNO may program the network devices by an external entity, known as SDN controller [15]. Specifically, the control plane of the network consists of a logically centralized SDN controller implemented in software that controls a set of low-cost and simple network devices. These devices comprise the user plane of the network. In this approach, the SDN controller has a global view of the network, and could make traffic management decisions according to the operational policies imposed by the MNO [16].

SDN will bring substantial benefits to the MNO to implement network slices, especially in the TN segment. The use of a logical centralized SDN controller will facilitate the MNO to automate the management of the TN resources allocated for each network slice. Specifically, this means the forwarding rules of each network

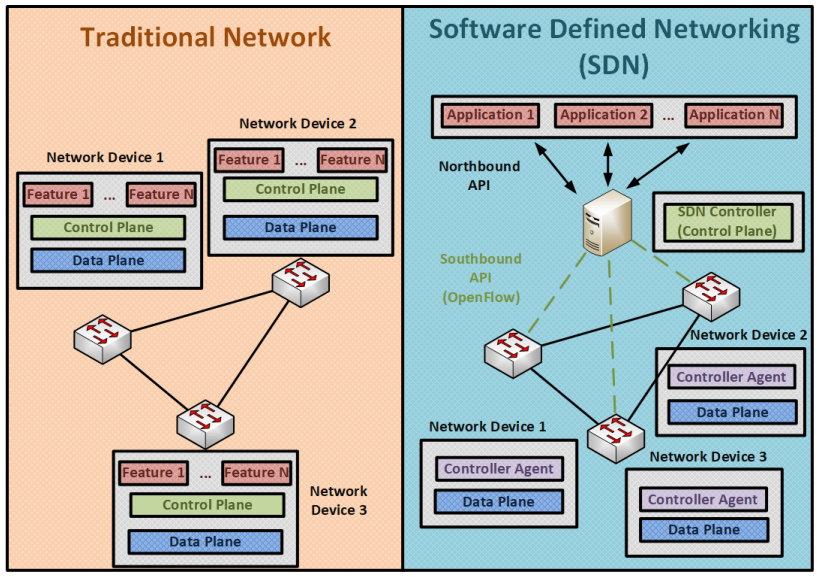


Figure 1.4: Traditional network approach versus SDN approach. Adapted from [17] and [18].

device within the MNO's infrastructure can be dynamically modified according to the traffic demands of each network slice.

1.3.3 Multi-access Edge Computing (MEC)

Multi-access Edge Computing (MEC) is a key emerging technology in 5G mobile networks. As Fig. 1.5 depicts, this technology aims at extending computing, storage, and networking resources from centralized clouds to the edge of the RAN [19]. It allows the MNO to (a) reduce traffic bottlenecks in the CN, (b) assist in the offload of heavy computational tasks from power constrained User Equipment (UE) to edge servers, and (c) bring the CN data plane functionalities to edge servers with the goal of reducing the experienced packet transmission delay.

The use of MEC along with network slicing can provide potential for implementing a wide range of new services. Specifically, MEC is of particular interest for uRLLC and mMTC services. In the case of network slices providing uRLLC services, the MNO can leverage the servers located in the edge to reduce the packet transmission delay experienced by the UEs attached to these network slices. In the case of mMTC, the MNO can use the edge servers to collect infor-

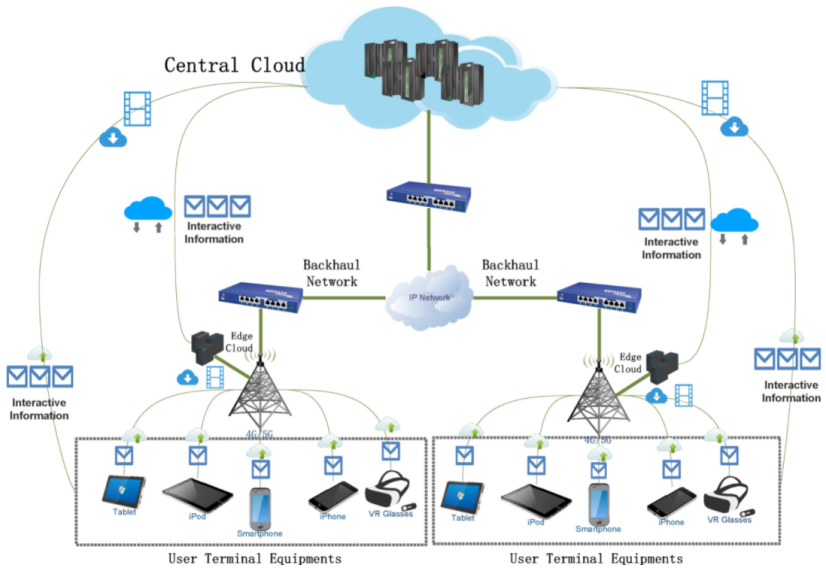


Figure 1.5: Overview of a MEC system [20].

mation from the UEs, process it (e.g., get some statistics) and use the resulting information to introduce novel utilities for the network slice consumers.

1.4 Main Challenges on Network Slicing

To fully realize a network-slicing-based 5G mobile network, several outstanding challenges need to be addressed. The most significant are summarized as follow [8, 21, 22, 23]:

- **RAN virtualization**: This concept goes beyond the notion of RAN sharing. Specifically, RAN virtualization should provide the MNO the capability of creating isolated virtual RAN instances with tailored sets of network functionalities (e.g., scheduling, mobility management) and performance requirements. The virtualization technology has been already applied on the CN. However, using this technology in the RAN is challenging since the capacity of the wireless access nodes mainly depends on time-varying channel effects such as shadowing, fast-fading and/or interference from neighbor access nodes. This means the virtualization methods used for the CN cannot be directly implemented in the RAN. Therefore, new mechanisms to virtualize the wireless access nodes must be designed.

- **Automated management**: To avoid manual efforts and errors, the management of network slices should be implemented in an automated way, i.e., without human intervention. This is challenging since the MNO must consider many dimensions and technologies to (a) instantiate, activate, operate, and terminate network slices; (b) adjust load balance, charging policies, security, and Quality of Service (QoS) for each network slice; (c) abstract and isolate the resources allocated for each network slice; and (d) implement resource sharing mechanisms at inter-slice and intra-slice levels.

- **Resource Sharing**: It can be implemented by static or dynamic partitions. Since the traffic demand of each network slice may change throughout its lifetime, the dynamic resource sharing could make the usage of the MNO's infrastructure resources more efficient with respect to the static resource sharing. Implementing dynamic resource sharing techniques is challenging

because the MNO must also guarantee the performance requirements of each network slice are met, i.e., performance isolation, in the long term.

- **Security**: This is a critical problem to be addressed because of network slices share the same MNO's network infrastructure. Furthermore, each network slice may have specific security policy requirements. All this involves the MNO must design the security mechanisms for a network slice considering the impact of these mechanisms on security aspects of the other network slices.

- **Mobility management**: On the one hand, the mobility support may be optional for some network slices. For example, network slices for mMTC services could not require mobility functions because of fixed position of UEs, e.g., sensors. On the other hand, network slices requiring mobility management could differ in their requirements. For example, the mobility requirements of eMBB services are different from uRLLC services. Thus, the design of a slice-oriented mobility management protocol is imperative to tackle mobility challenges in network slicing.

- **Service composition with fine-grained network functions**: From the MNO viewpoint, it easier composing coarse-grained functions for a network slice because fewer interfaces need to be defined to chain these functions. The drawback of this approach is the reduced flexibility for adapting the network slices to their varying traffic demands. For this reason, fine-grained network functions are more desirable. This means a scalable and interoperable means for network slice composition with fine-grained functions should be considered by the MNO.

- **Translation from service requirements to resource requirements**: A significant challenge for the realization of a network slice is how to go from a high-level description of the service requirements (e.g., UE throughput, number of UEs, delay bound for a packet, etc) to the concrete amount of resources (e.g., radio resources, computing resources, networking resources, etc) required to meet these requirements. To address this problem, the MNO needs specific description languages per network domain to define the network slice requirements and the amount of infrastructure resources to

implement such network slice. Furthermore, the MNO needs a mathematical framework capable of translating these requirements into the amount of required infrastructure resources. This is challenging, specially in the RAN domain where the capacity offered to a network slice depends on the channels effects such as shadowing, fast-fading and/or cell interference.

- **E2E network slice orchestration and management**: A network slice could be deployed in multiple administrative and network domains. Since each domain has specific requirements, it would be desirable defining multiple management entities, each responsible for managing the lifecycle of a network slice in a specific domain. The challenge lies on coordinating the management and orchestration tasks of each domain while the performance requirements of each network slice are met in the long term. For example, scaling up the radio resources allocated for a network slice in the RAN should impact on the amount of resources allocated for such network slice in the CN.

1.5 Scope and Objectives of the Thesis

Network slicing is a technological solution which will play a key role in the 5G mobile networks. Specifically, this solution will allow a MNO to economically provide emerging communication services, each with a specific set of requirements in terms of functionality and performance, over a common physical network infrastructure.

In a first attempt, the research community has put their efforts into the integration of network slicing in the CN segment. Then, considering the benefits provided by this technology, the analysis of network slicing has been extended to other network domains as the TN and the RAN. In this thesis, we focus on the RAN domain.

One of the main challenges in network slicing is the RAN virtualization. To address it, we need first to build understanding on NFV and analyze in detail how this technology enables the MNO to deploy and operate RAN slices. The Standards Developing Organization (SDO) responsible for standardizing NFV is the European Telecommunications Standards Institute (ETSI)-NFV group. This

SDO has defined the NFV-Management and Orchestration (MANO) framework to manage the lifecycle of the network functions which are implemented as software, i.e., the VNFs. Thus, it is key to understand which RAN functionalities could be implemented as VNFs and how the NFV-MANO can instantiate, scale up/down and release the virtual resources allocated for these VNFs.

The ETSI-NFV group has also defined a set of management templates, known as NFV descriptors, which can be used by the NFV-MANO to instantiate, scale and release the virtual resources allocated for each VNF in an automated way. The challenge lies in describing with the NFV templates the virtual resources of the VNFs which will accommodate the fluctuations of the spatial and temporal traffic demands of a RAN slice in a cellular environment.

Another interesting aspect on NFV is how to share the same VNF instances among multiple RAN slices. This approach may involve statistical multiplexing gains on the utilization of virtual resources. However, achieving the customization level required by each RAN slice is challenging. This means the impacts of sharing RAN functionalities among RAN slices must be analyzed in terms of customization and isolation.

Other challenge to be addressed in RAN slicing lies on the translation from the service requirements of a RAN slice onto the amount of resources (i.e., radio resources, computing resources, networking resources, etc) which the MNO requires to deploy this RAN slice and operate it throughout its lifetime. In addition to the ETSI-NFV group, the 3GPP also plays a key role in this context. This SDO is responsible for defining the network functions which a RAN slice requires, including their functionalities and capabilities. Thus, it is key to understand from the management viewpoint how the ETSI-NFV and 3GPP concepts on network slicing can be used by the MNO to proceed from the deployment request of a RAN slice to the provision of such RAN slice.

On the one hand, how the service requirements of a RAN slice can be mapped to the 3GPP understanding on RAN slicing, i.e., to the information models to describe the RAN functionalities, must be analyzed. On the other hand, work must also be done on how the MNO should map the 3GPP information models which define the RAN functionalities to the NFV templates which describe the virtual resources to deploy the RAN slice's VNFs.

In addition to linking the 3GPP and ETSI-NFV information models for net-

work slicing, the MNO must also rely on a mathematical framework to compute the amount of resources required to deploy multiple RAN slices over a common cellular environment. Specifically, this framework must plan the amount of resources which simultaneously satisfy the performance requirements of these RAN slices in the long term.

In this vein, the main objective of this thesis is to study management mechanisms to automate the deployment and orchestration of RAN slices under an architectural framework based on 3GPP and ETSI-NFV specifications. To that end, this thesis addresses the following specific objectives:

- **Objective 1**: Designing an architectural framework based on 3GPP and ETSI-NFV standards for RAN slicing. This objective is decomposed into the following sub-objectives:
  - **Sub-objective 1.1**: Designing management procedures to deploy and operate the virtualized part of a RAN slice with the NFV-MANO.
  - **Sub-objective 1.2**: Proposing a 3GPP/NFV-based architectural framework for managing RAN slices throughout their lifetime.
  - **Sub-objective 1.3**: Defining a mechanism for sharing RAN functionalities and their underlying resources among multiple RAN slices.
- **Objective 2**: Designing, implementing and evaluating management mechanisms for planning RAN slices. This dissertation focuses on RAN slices for services with requirements in terms of either Guaranteed Bit Rate (GBR) or latency and reliability. Under this context, this objective is decomposed into the following sub-objectives:
  - **Sub-objective 2.1**: Designing, implementing and evaluating a solution for planning RAN slices providing services with GBR requirements.
  - **Sub-objective 2.2**: Designing, implementing and evaluating a solution for planning RAN slices providing services with stringent requirements in terms of latency and reliability.

1.6 Research Methodology

This dissertation aims to design a management framework for RAN slicing based on the 3GPP and NFV standards. To address this objective, a classical scientific approach has been followed. First, we have studied in depth the main contributions in standards of these SDOs on network slicing. Then, we have reviewed the main state-of-the-art proposals on management frameworks for RAN slicing based on the 3GPP and ETSI-NFV standards. After reviewing the research community's contributions on this topic, we have identified the main gaps on how the network slicing perspectives from the 3GPP and the ETSI-NFV must be linked to manage the lifecycle of several RAN slices over a multi-cellular environments. To address these gaps, we have proposed a 3GPP/ETSI-NFV-based architectural framework to manage the lifecycle of multiple RAN slices.

Another objective to be addressed in this thesis is the design of management mechanisms for planning in advance the deployment of several RAN slices in a multi-cellular environment. To achieve this goal, the first task has been to design mathematical models which allow the MNO to translate the performance requirements of a single RAN slice onto the amount of radio resources which satisfy them in the long term. To that end, we have reviewed the main state-of-the-art solutions on modeling this translation. In the review process, we have studied the main mathematical tools used in these solutions such as queueing theory, Markov chains, or network calculus. Furthermore, we have also identified the main gaps of the solutions. Then, we have designed, implemented and evaluated two mathematical models to perform the translation from performance requirements onto the amount radio resources for RAN slices offering GBR and uRLLC services, respectively. The proposed models address some of the gaps previously identified. Finally, we have validated the proposed models by means of simulation, demonstrating they are feasible for performing the mentioned translation.

The proposed models are necessary but insufficient for planning RAN slices offering GBR and uRLLC services. Specifically, these models must be used by a mathematical framework to determine the amount of radio resources allocated for each RAN slice and their distribution in the multi-cell environment. This allocation must ensure the performance requirements of these RAN slices are met in the long term. In this thesis, we have focused on scenarios where the MNO must plan

either GBR services or uRLLC services. It means we have designed two mathematical frameworks, one per scenario. Prior the design of these frameworks, we have reviewed the state-of-the-art solutions for provisioning RAN slices. In this review, we have analyzed the main gaps of these solutions and the mathematical tools used for provisioning RAN slices. Based on the previous analysis, we have designed two mathematical frameworks for planning RAN slices offering GBR services and uRLLC services, respectively. Finally, we have implemented and evaluated both frameworks, demonstrating their effectiveness in scenarios with resource scarcity.

1.7 Publications

The study carried out in this dissertation and the proposed solutions have resulted in articles which have been published in renowned international journals, magazines, and conferences. These articles are:

- **Paper A**: **O. Adamuz-Hinojosa**, J. Ordonez-Lucena, P. Ameigeiras, J. J. Ramos-Munoz, D. Lopez and J. Folgueira, "**Automated Network Service Scaling in NFV: Concepts, Mechanisms and Scaling Workflow**," IEEE Communications Magazine, vol. 56, no. 7, pp. 162-169, July 2018. DOI: 10.1109/MCOM.2018.1701336. IF=10.356 (Q1).

- **Paper B**: **O. Adamuz-Hinojosa**, P. Munoz, J. Ordonez-Lucena, J. J. Ramos-Munoz and J. M. Lopez-Soler, "**Harmonizing 3GPP and NFV Description Models: Providing Customized RAN Slices in 5G Networks**," IEEE Vehicular Technology Magazine, vol. 14, no. 4, pp. 64-75, Dec. 2019. DOI: 10.1109/MVT.2019.2936168. IF=7.92 (Q1).

- **Paper C**: **O. Adamuz-Hinojosa**, P. Munoz, P. Ameigeiras and J. M. Lopez- Soler, "**Sharing gNB components in RAN slicing: A perspective from 3GPP/NFV standards**," IEEE Conference on Standards for Communications and Networking (CSCN), Granada, Spain, October 2019. DOI: 10.1109/CSCN.2019.8931318.

- **Paper D**: P. Munoz, **O. Adamuz-Hinojosa**, J. Navarro-Ortiz, O. Sallent, J. Perez-Romero, "**Radio Access Network Slicing Strategies at Spec-**

trum Planning Level in 5G and Beyond," IEEE Access, vol. 8, pp. 79604-79618, May 2020. DOI: 10.1109/ACCESS.2020.2990802. IF=3.74 (Q2).

- **Paper E**: **O. Adamuz-Hinojosa**, P. Ameigeiras, P. Munoz, and J. M. Lopez- Soler, "**Analytical Model for the UE Blocking Probability in an OFDMA Cell providing GBR Slices**," IEEE Wireless Communications and Networking Conference (WCNC), Nanjing, China, March 2021.

In addition to the previous publications, this thesis have also generated articles which are currently under a review process. These articles are the following:

- **Paper F**: **O. Adamuz-Hinojosa**, P. Ameigeiras, P. Munoz, and J. M. Lopez- Soler, "**UE Blocking Probability Model for 5G GBR Slices**," Submitted to IEEE Transactions on Wireless Communications.

- **Paper G**: **O. Adamuz-Hinojosa**, P. Munoz, P. Ameigeiras, and J. M. Lopez- Soler, "**Potential-Game-Based 5G RAN Slice Planning for GBR Services**," Submitted to IEEE Transactions on Mobile Computing.

- **Paper H**: **O. Adamuz-Hinojosa**, V. Sciancalepore, P. Ameigeiras, J. M. Lopez- Soler, and X. Costa-Pérez, "**A Delay-driven RAN Slicing Orchestrator to support B5G uRLLC Services**," Submitted to IEEE Transactions on Wireless Communications.

Lastly and in parallel to the study carried out in this dissertation, the Ph.D. student has also produced the following articles:

- L. Geng, L. Qiang, J. Ordonez-Lucena, **O. Adamuz-Hinojosa**, P. Ameigeiras, D. Lopez and L. Contreras, "**COMS Architecture**," IETF draft-geng-coms-architecture-02, March 2018.

- J. Ordonez-Lucena, **O. Adamuz-Hinojosa**, P. Ameigeiras, P. Munoz, J.J. Ramos-Munoz, J. Folgueira and D. Lopez, "**The Creation Phase in Network Slicing: From a Service Order to an Operative Network Slice**," European Conference on Networks and Communications (EuCNC), Ljubljana, Slovenia,June 2018. DOI: 10.1109/EuCNC.2018.8443255.

- D. Camps-Mur, M. Ghoraishi, J. Gutierrez, J. Ordonez-Lucena, T. Cogalan, H. Haas, A. Garcia, V. Sark, E. Aumayr, S. Meer, S. Yan, A. Mourad, **O. Adamuz-Hinojosa**, J. Perez-Romero, M. Granda, R. Bian, "**5G-CLARITY: Integrating 5GNR, WiFi and LiFi in Private Networks with Slicing Support**," European Conference on Networks and Communications (EuCNC), Dubrovnik, Croatia, June 2020.

- P. Munoz, **O. Adamuz-Hinojosa**, P. Ameigeiras, J. Navarro-Ortiz, J.J. Ramos-Munoz, "**Backhaul-Aware Dimensioning and Planning of Millimeter-Wave Small Cell Networks**," Electronics, vol. 9, no. 1429, Sep 2020. DOI: 10.3390/electronics9091429. IF=2.412 (Q2).

- J. Prados-Garzon, P. Ameigeiras, J. Ordonez-Lucena, P. Munoz, **O. Adamuz-Hinojosa**, D. Camps-Mur, "**5G Non-Public Networks: Standardization, Architectures and Challenges**," IEEE Access, vol. 9, pp. 153893-153908, Nov. 2021. DOI: 10.1109/ACCESS.2021.3127482. IF=3.367 (Q2).

- T. Cogalan, D. Camps-Mur, J. Gutiérrez, S. Videv, V. Sark, J. Prados-Garzon, J. Ordonez-Lucena, H. Khalili, A. Fernández-Fernández, M. Goodarzi, A. Yesilkaya, R. Bian, S. Raju, M. Ghoraishi, H. Haas, **O. Adamuz-Hinojosa**, A. García, C. Colman-Mexnier, A. Mourad, E. Aumayr, "**5G CLARITY: 5G-Advanced Private Networks Integrating 5GNR, WiFi and LiFi**," IEEE Communication Magazine, vol. 60, no. 2, pp. 73-79, Feb. 2022, doi: 10.1109/MCOM.001.2100615. IF=9.619 (Q1).

1.8 Thesis Outline

This thesis is presented as a compendium of articles, most of them already published and the rest under evaluation. This means the contributions and findings of the thesis are collected in these articles, which are presented in Parts II and III. Opening each of these parts we include a chapter which gathers *i*) the background and literature review; *ii*) the description of the problem addressed in these articles; and *iii*) the main contributions of these articles. This chapter will help the readers to understand the content of these articles and how they are related to

each other. In addition, this thesis is bookended by the introductory Part I and Part IV, which provides the main conclusions and recommendations for future investigations.

The outline of this thesis is the following:

- **Part I**: Includes this chapter, where we introduce and motivate the research carried out in this thesis.
- **Part II**: Provides the proposal of a 3GPP/ETSI-NFV-based architectural framework to automate the management of RAN slices throughout their lifetimes. Papers A, B and C compose this part, prefaced by a literature review which summarizes some of the most relevant studies in this topic, and a short overview in which the research questions and main findings from the Papers A, B, and C are compiled.
- **Part III**: Comprises the definition, implementation and evaluation of the mathematical models and frameworks which allow the MNO to plan RAN slices with performance requirements in terms of either GBR or latency and reliability. Papers D, E, F, G, and H form the main body of Part III. Prior to this, we provide a state-of-the-art review which gathers the main contributions on this topic. Furthermore, an overview is included to shed lights on the connection between the Papers D-H and the main contributions of the investigations.
- **Part IV**: Concludes the dissertation, providing recommendations and future paths for research on related topics.

The thesis makes use of numerous abbreviations which are spelled out in their first appearance for each chapter. We recommend that the reader use the List of Abbreviations included before Part I. A reference list is included at the end of each chapter. Note that references that are cited in different chapters may not be represented by the same number in all chapters.

References

[1] A. Aijaz, "Private 5G: The Future of Industrial Wireless," *IEEE Ind. Electron. Mag.*, vol. 14, no. 4, pp. 136–145, 2020.

[2] J. Karlsson, "5G for business: a 2030 market compass," *Ericsson, Business Potential Report, Oct*, 2019.

[3] I. Vision, "Framework and overall objectives of the future development of IMT for 2020 and beyond," *International Telecommunication Union (ITU), Document, Radiocommunication Study Groups*, 2015.

[4] ITU-T Rec. Y.3101, "Requirements of the IMT-2020 network," Jan. 2018.

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

[6] A. Kaloxylos, "A Survey and an Analysis of Network Slicing in 5G Networks," *IEEE Commun. Stand. Mag.*, vol. 2, no. 1, pp. 60–65, 2018.

[7] 3GPP TR 22.852 V.13.1.0, "Study on Radio Access Network (RAN) sharing enhancements (Release 13)," Sep. 2014.

[8] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, 2017.

[9] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 3, pp. 2429–2453, 2018.

[10] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Netw.*, vol. 28, no. 6, pp. 18–26, 2014.

[11] A. U. Rehman, R. L. Aguiar, and J. P. Barraca, "Network Functions Virtualization: The Long Road to Commercial Deployments," *IEEE Access*, vol. 7, pp. 60439–60464, 2019.

[12] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network Function Virtualization: State-of-the-Art and Research Challenges," *IEEE Commun. Surv. Tutor.*, vol. 18, no. 1, pp. 236–262, 2016.

[13] ETSI, "ETSI GS NFV 002 v1.1.1 Network Functions Virtualization (NFV)," *Architectural Framework. sl: ETSI*, 2013.

[14] O. N. Fundation, "SDN Architecture," *ONF TR-521, Issue 1.1*, 2016.

[15] "SDN overview." `https://www.opennetworking.org/sdn-definition/`. Accessed: 2018-07-19.

[16] H. Freeman and R. Boutaba, "Networking industry transformation through softwarization [The President's Page]," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 4–6, 2016.

[17] M. Jammal, T. Singh, A. Shami, R. Asal, and Y. Li, "Software defined networking: State of the art and research challenges," *Comput. Netw.*, vol. 72, pp. 74–98, 2014.

[18] P. Goransson, C. Black, and T. Culver, *Software defined networks: a comprehensive approach*. Morgan Kaufmann, 2016.

[19] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 3, pp. 1657–1681, 2017.

[20] C. Dong and W. Wen, "Joint Optimization for Task Offloading in Edge Computing: An Evolutionary Game Approach," *Sensors*, vol. 19, no. 3, 2019.

[21] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, 2017.

[22] P. Rost *et al.*, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, 2017.

[23] X. Li *et al.*, "Network Slicing for 5G: Challenges and Opportunities," *IEEE Internet Comput.*, vol. 21, no. 5, pp. 20–27, 2017.

Part II

Management Framework for Network Slicing in the Radio Access Network (RAN)

Chapter 2

Background and Problem Description

The first Section provides the network slicing perspectives from the most representative Standards Developing Organizations (SDOs) and the telecommunication industry. Then, the second Section gives an overview of the most representative management frameworks for Radio Access Network (RAN) slicing which have been proposed in the literature. Next, the third Section describes the problem addressed in the Part II of this dissertation. Finally, the fourth Section states the thesis contributions on this topic.

2.1 Network Slicing Perspectives from Standards Developing Organizations (SDOs), Telecommunication Industry and Academia

Given the potential benefits of network slicing, the main SDOs along with the telecommunication industries and academia have made efforts in the last years to integrate network slicing in the Fifth Generation (5G) mobile networks. Below, we summarize how these organizations view of what a network slice is and what should be its purpose according to their business focus and expertise field.

2.1.1 Third Generation Partnership Project (3GPP) Perspective

The Third Generation Partnership Project (3GPP) focuses on network slicing from the application viewpoint. Specifically, this SDO defines a network slice as all the managed network functions, with their supporting resources (e.g., hardware resources, and virtualized compute, storage, and networking resources), which provide a certain set of services to serve a certain business purpose [1]. From the 3GPP perspective, a network slice could be provisioned in the Core Network (CN) and the RAN. To manage a network slice in these network domains, the 3GPP has defined an information model to describe its composition into network functions and its operational features [2]. Furthermore, the 3GPP has also defined a management framework to manage and orchestrate network slices [3]. With this framework and the 3GPP information models, network slicing may be incorporated in the standardized 5G architecture for the CN and the RAN segments.

Due to the increasing complexity to manage and orchestrate a network slice in these network domains, the 3GPP has also introduced the concept of network slice subnet [3]. A network slice subnet represents a group of network functions that are part or complete constituents of a network slice. Thereby, the network slice information model could reference one or more information models, each describing a specific network slice subnet. Focusing on the information model of a single network slice subnet, it describes the subset of constituent network functions. Specifically, this information model (a) defines how these network functions must be deployed and operated; and (b) contains parameters that characterizes these network functions at application level, thus the Mobile Network Operator (MNO) must properly configure these parameters when these network functions are deployed.

Focusing on the RAN segment, each access node, as know as next Generation NodeB (gNB), may be split into a Centralized Unit (CU) and several Distributed Units (DUs). This split enables the MNO to customize the distribution of the radio protocol functions for each network slice in such a way that the specific requirements of the corresponding communication service can be met [4]. Additionally, the 3GPP has defined the 5G air interface, i.e., the New Radio (NR) interface, to incorporate novel features such as broader bandwidths or flexible

Transmission Time Intervals (TTIs). This allows the MNO to adapt this interface for each network slice [5].

In the CN segment, the 3GPP has integrated network slicing along with a Service-Based Architecture (SBA) to enable the incorporation of the classical CN functionalities (e.g. mobility management, session management, etc) with specific vertical functions [6]. Unlike the point-to-point interconnection among the control plane CN functions, in a SBA 5G-CN all the control plane network functions use REST-based services to provide one or more capabilities that can be discovered, requested, or subscribed by other network functions [7]. To aid these purposes, the 5G SBA further defines two new functional entities: the Network Function Repository Function (NRF), which provides network function service discovery capabilities, and the Network Exposure Function (NEF), which exposes services from network functions inside the network to authorized external applications to adjust the network configuration.

2.1.2 European Telecommunications Standards Institute (ETSI)-NFV Perspective

To achieve the flexibility and modularity that a network slice requires, some of their network functions can be implemented by software, i.e., by Virtualized Network Functions (VNFs) [8]. However, the lifecycle management of these VNFs and the orchestration of their resources goes beyond the 3GPP scope. The European Telecommunications Standards Institute (ETSI)-Network Function Virtualization (NFV) group is playing a significant role on these tasks. To that end, the ETSI-NFV group has defined the concept of Network Service (NS) as well as its information model. A NS is a composition of network functions which may be implemented as VNFs and/or Physical Network Functions (PNFs). This means a NS can be regarded as a resource-centric view of a network slice for those cases where the network slice would contain at least one VNF [8].

The concept of NS introduced by the ETSI-NFV group is key for network slicing. Particularly, three scenarios may be considered [9]:

- The network slice consists of a single NS.
- The network slice consists of a composite NS.

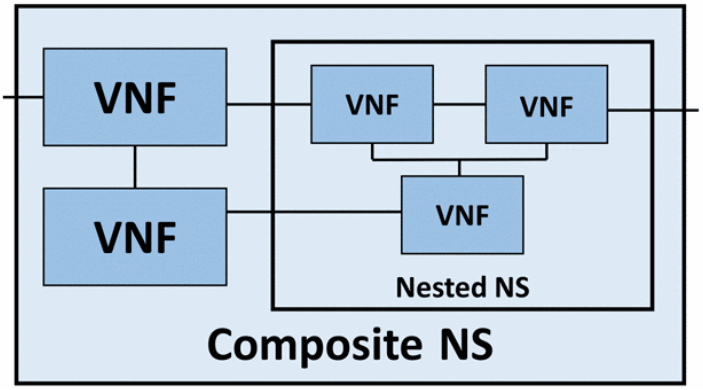


Figure 2.1: An example of a composite NS. This NS consists of two VNFs and one simple NS [9].

- The network slice consists of a concatenation of single and/or composite NSs.

A simple NS includes one or more VNFs, and virtual links providing connectivity between them. In search of modularity and recursiveness, NFV provides the ability to include in the design of a NS one or more nested NSs. The result is a composite NS as the one depicted in Fig. 2.1.

To manage the lifecycle of a NS, the ETSI-NFV group has also defined the NFV architectural framework [10]. This framework is depicted in Fig. 2.2 and comprises the following functional blocks:

- **Network Function Virtualization Infrastructure (NFVI)**: The set of physical resources to build up the environment in which the VNFs are deployed. The physical resources mainly consists of (a) commodity computing and storage hardware; and (b) commodity network devices and links that provide processing, storage, and connectivity to the VNFs. These resources are abstracted through a virtualization layer yielding virtual computing, storage and network resources.

- **VNFs**: The software implementations of the network functions. A VNF comprises one or more Virtualized Network Function Components (VNFCs), each performing a well-defined part of the VNF functionality [11]. In

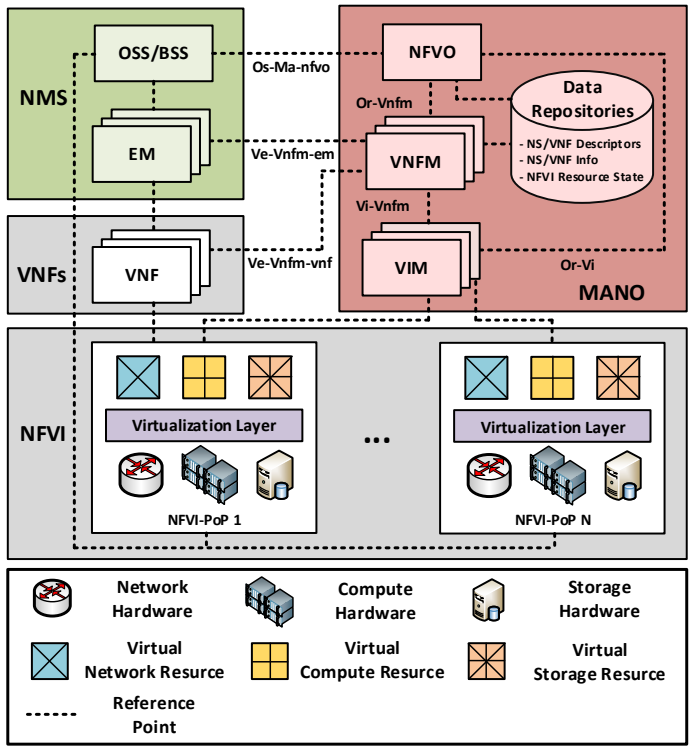


Figure 2.2: ETSI-NFV Architectural Framework.

turn, a VNFC might have several instances. Each VNFC instance is hosted in a single virtualization container, e.g., a Virtual Machine (VM), which is built with a specific amount of virtual computing and storage resources.

- **NFV-Management and Orchestration (MANO)**: The collection of all the entities, interfaces and data repositories to manage the lifecycle of the NSs and orchestrate their underlying resources [12]. The NFV-MANO consists of the following three functional blocks:

  - **Network Function Virtualization Orchestrator (NFVO)**: This functional block is in charge of the lifecycle management of one or more NSs, e.g., their instantiation, update, query, scaling, and termination. Furthermore, this functional block is also responsible for the management coordination of the constituents VNFs of each NS and the underlying NFVI resources.

  - **Virtual Network Function Manager (VNFM)**: This block is responsible for the lifecycle management of one or more VNFs (e.g., instantiation, update, query, scaling, termination).
  - **Virtual Infrastructure Manager (VIM)**: This functional block is responsible for the management of the NFVI resources.

- **Network Management System (NMS)**: The collection of management entities focused on traditional (i.e., non virtualized-related) management tasks, complementary to those tasks performed by the NFV-MANO entities. The NMS comprises the following functional blocks [12]:

  - **Element Manager (EM)**: This block is responsible for Fault, Configuration, Accounting, Performance, and Security (FCAPS) management functionalities for a single VNF.
  - **Operations Support System (OSS)/ Business Support System (BSS)**: The collection of systems and management applications that service providers use to operate their business and network services.

2.1.3 Global System for Mobile Communications Alliance (GSMA) Perspective

Focusing on the provisioning phase of a network slice, when a tenant (i.e, the network slice consumer) requests the MNO to allocate a network slice satisfying a particular set of service requirements, it is convenient for the MNO to unambiguously define the service requirements from different tenants; and represent them in a common language, with the purposes of facilitating their translation into appropriate network slice provisioning actions. In this regard, the Global System for Mobile Communications Alliance (GSMA) has developed a universal network slice blueprint that provides a point of convergence between the MNO and the tenants on network slicing understanding. This blueprint, known as Generic Slice Template (GST) [13] [14], contains a set of attributes that can be used to characterize a network slice. Furthermore, these attributes can be also used by the tenants and the MNO to agree on the Service Level Agreement (SLA).

A Network Slice Type (NEST) describes the characteristics of a network slice by means of filling GST attributes with values based on tenant provided service requirements. Different NESTs allow describing different types of network slices. For network slices based on 5G service categories, e.g., enhanced Mobile Broadband (eMBB) or ultra-Reliable Low Latency Communication (uRLLC), the MNO may have a set of standardized NESTs (S-NESTs). For network slices addressing specific industry use cases, the MNO can define private NESTs (P-NESTs). Both S-NESTs and P-NESTs are registered and published in the MNO's service catalog.

Once the MNO has received the network slice order request from a tenant, the MNO's product order service has to map the S/P-NEST attributes with the network slice information model defined by the 3GPP.

2.1.4 Internet Engineering Task Force (IETF) Perspective

Despite the valuable efforts on standardizing network slicing, the 3GPP and ETSI-NFV group focus on RAN/CN segments and the virtualization environment to deploy a netwrok slice in these domains, respectively. This means Transport Network (TN) is out of scope of these SDOs. The Internet Engineering Task Force (IETF) is the SDO responsible for standardizing network slicing in the TN domain. Most of the IETF work on network slicing has focused on architectural analysis and data modeling. It has proposed some enhancements to the protocols for collecting telemetry and for providing additional control of network resources.

In [15] IETF puts network slicing and network virtualization in the context of the Internet architecture. Specifically, this document compares and contrast different approaches and provides an overview of IETF virtualization technologies. Another key aspect is the relevance of Abstraction and Control of Traffic Engineered Networks (ACTN) on network slicing. It defines a Software-Defined Networks (SDN)-based architecture that relies on the concept of network and service abstraction to detach network and service control from the underlying data plane. In [16], IETF describes how to use the SDN approach, at the core of ACTN, to manage virtual networks so each network slice could be separately operated. In [17] IETF presents how it's data models for service operation can be coordinated with topology management models and with device and protocol

management models. Finally, [18] specifies a framework which can use a combination of existing, modified, and new networking technologies to provide an enhanced Virtual Private Network (VPN) service. Specifically, this framework mainly relies on VPN and Traffic Engineering (TE) technologies to provide connectivity services with advanced characteristics such as low latency guarantees, bounded jitter, or isolation among services.

2.1.5 5G Infrastructure Public Private Partnership (5G-PPP) Research Projects' Perspective

In parallel and constantly being feedbacked with the SDOs specifications, some 5G Infrastructure Public Private Partnership (5G-PPP) research projects are shedding light on the key aspects of network slicing in the 5G architecture.

In a first intent, some of the first phase projects such as 5G-NORMA [19], METIS-II [20], and SESAME [21] have conceptually defined how the control plane functionalities of the network slices might be included in the 5G architecture. However, none of these projects has focused on the management plane of network slicing. Additionally, the 5GEx project is the only one project of the 5G-PPP first phase that covered the multi-domain aspect, providing the first step towards End-to-End (E2E) network slicing provision capabilities [22].

The analysis of the network slicing in the 5G architecture has been also of key importance in the most of 5G-PPP Phase 2 projects. Some of them address the 5G architecture in a specific environment, such as 5G ESSENCE and 5G TRANSFORMER. The 5G ESSENCE project [23] focuses on edge cloud computing and small cells, thus it concentrates on the RAN segment. Their 5G architecture supports network slicing, and it is compliant with 3GPP and ETSI-NFV specifications. The 5G TRANSFORMER aims to transform the current mobile transport network into a SDN/NFV-based mobile transport and computing platform, which brings the network slicing paradigm into transport network by provisioning and managing transport slices throughout a federated virtualized infrastructure [24].

Other projects like MATILDA study the mechanisms required to deploy and orchestrate vertical specific applications over dynamically created network slices. To that end, MATILDA project promises a network-aware applications orches-

trator [25]. Finally, 5G MoNArch and SliceNet projects study the management of network slices across multiple administrative domains. Particularly, 5G MoNArch implements and deploys two testbeds for two vertical use cases with different requirements: Extreme mobile broadband in a touristic city, and a reliable industrial communications in a sea port [26]. In turn, SliceNet implements three primary uses cases to validate their architecture proposal: smart grid self-healing, e-health smart connected ambulance, and an intelligent public lighting system in a smart city [27].

2.2 Representative Management Frameworks for RAN Slicing

Initially, the research community puts their efforts on designing architectural frameworks to manage the lifecycle of network slices in the CN segment. Then, with the purpose of extending the main findings in the CN to an E2E environment, the industry and academia put their focus on the RAN and the TN segments. Concerning the RAN, slicing this segment is challenging due to the following reasons:

- Unlike CN resources (e.g., dedicated hardware and/or virtual resources), the RAN resources (i,e., spectrum) are more limited and thus, the MNO can only consider a fixed amount of radio resources to satisfy the traffic demands of each network slice. Furthermore, the channel effects such as fast-fading, shadowing or inter-cell interference hinder the MNO to provide the capacity which each network slice require and thus, meet their performance requirements.
- The radiofrequency functionalities cannot be virtualized, thus they cannot leverage the benefits of virtualization. Furthermore, most of the baseband functionalities require a real time processing, thus their virtualization is difficult. This involves the MNO puts a further effort to reach the degree of customization required by the RAN functionalities of each network slice.

The first attempts to overcome these challenges focus on designing an architectural framework for managing the lifecycle of several network slices in the

RAN segment. These frameworks are mainly focused on specific aspects such as: (a) the mechanisms for managing the RAN functionalities of each network slice; (b) configuring the RAN functionalities to provide the degree of customization required by each network slice; and (c) the mechanisms for managing the radio resource allocation for each network slice and their attached User Equipments (UEs).

Concerning the mechanisms for managing the RAN functionalities of a network slice, the authors of [28] have defined management interfaces and information models to support the dynamic and automatic deployment of network slices in the RAN. They have also discussed the complexity of such automation and have provided an illustrative description of the applicability of the overall framework and information models in the context of a neutral host provider which offers RAN slices to third party service providers. In [29], the authors have proposed a RAN slicing framework focused on the data management aspects for communication services in the Industry 4.0. Specifically, they have provided hybrid communications management and decentralized data distribution solutions supported by a hierarchical and multi-tier network architecture. In [30], the authors have proposed a RAN runtime slicing system to enable flexible slice customization on top of disaggregated RAN infrastructures. Furthermore, they have introduced a runtime SDK to facilitate the development of agile control applications able to monitor, control, and program the underlying RAN modules. Considering that, they have also provided a prototype based on the OpenAirInterface and MOSAIC5G platform to demonstrate how slicing and programmability can be achieved in two use cases. In [31], the same authors have also extended their framework to support slice-based multi-service chain creation and chain placement, with an auto-scaling mechanism to increase the network performance. The authors of [32] have presented a novel slicing scheme for RAN based on control/user plane separation. Specifically, they have divided the evolved NodeB (eNB) into two entities for transmitting control data and user data, respectively. Thereby, they facilitate the control and user plane separation. The authors of [33] have proposed enhanced network slicing in Fog-RANs (F-RANs). They have comprehensively presented a novel architecture and related key techniques, including radio and cache resource management, for network slicing in F-RANs. In [34] a Not Only (NO) Stack based vRAN has been proposed to be employed in

the 5G mobile communication system. In this proposal, the baseband processing and storage resources are sliced and orchestrated agilely to support multi-Radio Access Technology (RAT). The authors of [35] have provided a comprehensive overview of the 5G RAN design guidelines, key design considerations, and functional innovations. They have also depicted the air interface landscape that is envisioned for 5G, and have elaborated on how this will likely be harmonized and integrated into and overall 5G RAN, in the form of concrete control and user plane design considerations and architectural enablers for RAN slicing. In [36], the authors have considered how network slicing can use Cloud RAN (C-RAN) as an enabler for the required prerequisite network virtualization. Specifically, they have focused on the fronthaul network. They have shown how using a packet-switched fronthaul for network slicing will bring great advantages and enable the use of different functional splits, while the price to pay is minor decrease in fronthaul length due to latency constraints. In [37], the authors have proposed a multitier, Self-Organizing Network (SON) with a middle-tier SON, centralized SON, and distributed SON to support increasing network dynamicity for different RAN slices. They have also analyzed a lower-layer RAN split to optimize fronthaul and so further improve radio efficiency by utilizing the multitier SON among various radio nodes as well as the DUs and the CU of each gNB.

With respect to the mechanisms and techniques to configure the RAN functionalities of each RAN slice, the authors of [38] have focused on how enabling lower-layer flexibility in the RAN affects the development of RAN slicing, particularly in relation to ensuring isolation between RAN slices. Specifically, they have provided an approach that permits the allocation of resources to a service-type to be performed separately to resource allocation for individual services belonging to that type. In [39], the authors have identified the different RAN slice granularity options, that is, how to define RAN slices by combing customer and service needs. They have also presented how the 5G NR features can be used for facilitating RAN slice implementation and provide typical configurations for different RAN slice types from technology and RAN architecture perspectives. The authors of [40] have proposed a framework which aims to provide different Licensed-Assisted Access (LAA) configuration (i.e., each can be seen as a different RAN configuration mode) to be used on demand in a service-oriented manner. They have also provided an illustrative example to show that dynamic radio topology coupled

with LAA as a service is a promising and complementary enhancement for RAN slicing. In [41], the authors have provided a comprehensive analysis of the impact that the realization of RAN slicing has on the different layers of the radio interface protocol architecture. Furthermore, they have also proposed a framework for the support and specification of RAN slices based on the definition of a set of configuration descriptors that characterize the features, policies and resources to be put in place across the radio protocol layers of the gNBs. In [42], the authors have analyzed the RAN slicing problem in a multi-cell network in relation to the Radio Resource Management (RRM) functionalities that can be used as a support for splitting the radio resources among the RAN slices. Specifically, they have compared the granularity in the assignment of radio resources and the degrees of isolation and customization of four different RAN slicing approaches.

Regarding the management mechanisms for radio resource allocation in RAN slicing, the authors of [43] have proposed a two-level Medium Access Control (MAC) scheduling framework that can effectively handle Uplink (UL) and Downlink (DL) transmissions of network slices of different characteristics over a shared RAN. They have applied different characteristics over a shared RAN and different per-slice scheduling policies. In addition, they have focused on reducing latency for uRLLC services. In [44], the same authors have proposed a fully programmable network slicing architecture based on (a) the 3GPP-Dedicated Core Network (DCN); (b) a flexible RAN to enforce network slicing; and (c) a two-level MAC scheduler to abstract and share the physical resources among network slices. Based on that, they have also developed a proof of concept on RAN slicing on top of OpenAirInterface to derive key performance results in terms of flexibility and dynamicity to share the RAN resources among multiple network slices.

2.3 Problem Description

Despite there exists excellent contributions in the literature on architectural frameworks to manage RAN slices, there are still many challenges to be addressed in order to get a management framework fully based on the 3GPP and ETSI-NFV standards. These challenges are mainly related to the RAN virtualization, the RAN slice management in an automated way and the resource sharing.

Focusing on the RAN virtualization, many of the state-of-the-art solutions

on RAN slicing architectural frameworks omit how the MNO must deploy and operate the RAN slice's constituents implemented as VNFs. To address this problem, we must build understanding on NFV and how this technology links with the concept of network slicing from the 3GPP perspective. To that end, we need first to understand how the NFV-MANO can be used to deploy and operate the constituent VNFs of a RAN slice. Among all the NFV-MANO operations, the scaling operation is the most interesting to be analyzed. The reason is this operation comprises the main management procedures, i.e., adding/removing the virtualized resources of a VNF instance, and deploying/terminating a VNF instance.

On one hand, most of works dealing with the scaling operation in virtualized environments focus on mechanisms to estimate and allocate virtual resources to the VNFs. Furthermore, the rules and input data that these proposals use for the scaling operation are not retrieved from NFV templates; instead, they are specified manually for each use case. This may led to less agile and more error-prone scaling solutions, where the automation in NFV-MANO is not fully exploited.

On the other hand, there are not existing works that analyze the effect of modeling the NFV templates in the scaling operation. The existing NFV-MANO platforms use their own data modeling languages for their management templates. This leads to non-compatible workflows for the scaling operation, thus the reusability and portability provided by NFV is wasted.

Concerning the description of a RAN slice from the 3GPP perspective, the management templates defined by this SDO neglects the virtual resource requirements of the virtualized part of a RAN slice. To that end, these management templates could use the NFV descriptors. Notwithstanding, describing the virtual resources to accommodate the fluctuations of spatial and temporal traffic demands of a RAN slice in a cellular environment is challenging. In addition, the management templates defined by the 3GPP considers the gNB functionalities of a RAN slice from an application viewpoint. However, this SDO has not specified how these functionalities must be configured to meet the requirements for a specific communication service, e.g., an eMBB service, an uRLLC service or a massive Machine Type Communication (mMTC) service.

Some solutions for describing RAN slices have been proposed in the literature,

however the 3GPP completed the 5G-NR specifications after these works were published. This means the impact of the NR parameters in the RAN segment has not been analyzed in depth yet. Additionally, although these works have considered partially-virtualized gNBs, they have also neglected the description of the virtual resources required to build up these gNBs. Thereby, describing the spatial and temporal traffic demands of a RAN slice with NFV descriptors is still an open question.

Additionally, the majority of solutions proposed in the literature assume a single VNF instance accommodates a gNB component, i.e., a CU or a DU. This approach guarantees the customization of each RAN slice, however the resource utilization can be inefficient. Sharing VNF instances among RAN slices could involve statistical multiplexing gains on the utilization of virtual resources. However, achieving the customization level required by each RAN slice is challenging.

Some works have pursued a tradeoff solution between customization and resource utilization. Notwithstanding, they have analyzed neither the impacts of sharing gNB components on the customization of each RAN slice, nor the main factors that enable the isolation among RAN slices. Additionally, these works have focused on the lifecycle management from the 3GPP viewpoint, neglecting the ETSI-NFV perspective.

2.4 Thesis Contributions

Papers A-C gather the main contributions and findings of the Part II of dissertation. They are summarized as follows:

1. **Providing an overview of the automation of the NS scaling operation in NFV, addressing the options and boundaries introduced by ETSI normative specification (From Paper A).**

   An overview of the structure of the ETSI Network Service Descriptor (NSD) is provided. We identify the NSD fields most relevant for scaling, with special emphasis on the Instantiation Levels (ILs). These levels specify the different sizes an NS instance can adopt throughout its lifecycle. This limits the scaling of an NS instance to one of the discrete set of levels defined in the NSD. Thus, their correct design is key to ensure appropriate scaling

operations with NFV.

To facilitate their understanding, and to show how they are constructed in a NSD, we propose a simple example of an NSD. We also provide an overview of the automated NS scaling operation, analyzing in depth the options and boundaries introduced by the ETSI NSD information model. We show the different scaling procedures that the NFV framework has available, and how they can be triggered in an automated manner. This includes the proposal of an ETSI-compliant workflow for a representative scaling procedure. Our proposal clarifies how the different MANO blocks interact, specifying the information they exchange in each step.

2. **Proposing a description model that harmonizes 3GPP and ETSI-NFV viewpoints to manage RAN slices (From Paper B).**

   A description model for RAN slicing is provided. By harmonizing 3GPP and ETSI-NFV scopes, the proposed solution allows the management of virtualized gNB functionalities, and their customization by setting predefined radio parameters. Thereby, as a result a MNO could efficiently provide RAN slice subnets to accommodate the services demanded by verticals on a geographical area with specific spatial and temporal traffic demands. To gain insight into this proposal, this thesis provides an example where RAN slices for eMBB, uRLLC, and mMTC are described based on the proposed solution.

3. **Analyzing the key aspects in the 3GPP and NFV standards for sharing gNB components (From Paper C).**

   This thesis sheds light on the key aspects in 3GPP and NFV standards for sharing gNB components between RAN slices. To that end, this thesis (a) identifies the main scenarios for sharing gNB components; (b) analyzes the impact of sharing gNB components on the customization level of each RAN slices (i.e., the specific functional and performance behavior which each slice could reach); (c) determines the main factors that enable the isolation between RAN slices; and (d) proposes a description model to define the lifecycle management of a shared gNB component using the 3GPP and NFV management templates.

References

[1] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 3, pp. 2429–2453, 2018.

[2] 3GPP TS 28.541 V.16.4.0, "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3 (," Mar.

[3] 3GPP TS 28.530 V.16.1.0, "Management and orchestration; Concepts, use cases and requirements (Release 16)," Dec. 2019.

[4] 3GPP TS 38.401 V.15.5.0, "NG-RAN; Architecture description (Release 15)," Mar. 2019.

[5] 3GPP TS 38.300 V.16.6.0, "NR; NR and NG-RAN Overall description; Stage 2 (Release 15)," Dec. 2019.

[6] 3GPP TS 23.501 V.16.0.2, "System Architecture for the 5G System; Stage 2 (Release 16)," Apr. 2019.

[7] H. C. Rudolph, A. Kunz, L. L. Iacono, and H. V. Nguyen, "Security Challenges of the 3GPP 5G Service Based Architecture," *IEEE Commun. Stand. Mag.*, vol. 3, no. 1, pp. 60–65, 2019.

[8] ETSI GS NFV-EVE 012 V3.1.1, "Network Functions Virtualization (NFV); Evolution and Ecosystem; Report on Network Slicing Support with ETSI NFV Architecture Framework," Dec. 2017.

[9] J. Ordonez-Lucena *et al.*, "The Creation Phase in Network Slicing: From a Service Order to an Operative Network Slice," in *EuCNC, Ljubljana, Slovenia*, pp. 1–36, 2018.

[10] ETSI GS NFV 002 V1.1.1, "Network Functions Virtualisation (NFV); Architectural Framework," Oct. 2013.

[11] ETSI GS NFV-SWA 001 V1.1.1, "Network Functions Virtualisation (NFV); Virtual Network Functions Architecture," Dec. 2014.

[12] ETSI GS NFV-MAN 001 V1.1.1, "Network Functions Virtualisation (NFV); Management and Orchestration," Dec. 2014.

[13] GSMA PRD NG.116 v2.0, "Generic network Slice Template," 2019.

[14] 3GPP TS 28.531 V.16.5.0, "Management and orchestration; Provisioning; (Release 16)," Mar. 2020.

[15] J. Arkko *et al.*, "Considerations on Network Virtualization and Slicing," Mar. 2018.

[16] D. King *et al.*, "Applicability of Abstraction and Control of Traffic Engineered Networks (ACTN) to Network Slicing," Oct. 2018.

[17] Q. Wu *et al.*, "A Framework for Automating Service and Network Management with YANG," Oct. 2019.

[18] J. Dong *et al.*, "A Framework for Enhanced Virtual Private Networks (VPN+) Services," Feb. 2020.

[19] P. Rost *et al.*, "Mobile network architecture evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 84–91, 2016.

[20] I. Da Silva *et al.*, "5G RAN architecture and functional design," *METIS II white paper*, 2016.

[21] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective," *IEEE Wirel. Commun.*, vol. 24, no. 5, pp. 166–174, 2017.

[22] C. J. Bernardos, B. P. Gerö, M. Di Girolamo, A. Kern, B. Martini, and I. Vaishnavi, "5GEx: realising a Europe-wide multi-domain framework for software-defined infrastructures," *Trans. Emerg. Telecommun. Technol.*, vol. 27, no. 9, pp. 1271–1280, 2016.

[23] M. R. Spada, J. Pérez-Romero, A. Sanchoyerto, R. Solozabal, M. A. Kourtis, and V. Riccobene, "Management of Mission Critical Public Safety Applications: the 5G ESSENCE Project," in *EuCNC, Valencia, Spain*, pp. 155–160, 2019.

[24] A. de la Oliva *et al.*, "5G-TRANSFORMER: Slicing and Orchestrating Transport Networks for Industry Verticals," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 78–84, 2018.

[25] P. Gouvas *et al.*, "Design, Development and Orchestration of 5G-Ready Applications over Sliced Programmable Infrastructure," in *ITC*, vol. 2, pp. 13–18, 2017.

[26] Ö. Bulakci *et al.*, "Overall 5G-MoNArch architecture and implications for resource elasticity," 2018.

[27] Q. Wang *et al.*, "Slicenet: End-to-End Cognitive Network Slicing and Slice Management Framework in Virtualised Multi-Domain, Multi-Tenant 5G Networks," in *IEEE BMSB*, pp. 1–5, 2018.

[28] R. Ferrús, O. Sallent, J. Pérez-Romero, and R. Agustí, "On the automation of RAN slicing provisioning: solution framework and applicability examples," *EURASIP J. Wirel. Commun. Netw.*, vol. 2019, no. 1, pp. 1–12, 2019.

[29] M. C. Lucas-Estañ, M. Sepulcre, T. P. Raptis, A. Passarella, and M. Conti, "Emerging trends in hybrid wireless communication and data management for the industry 4.0," *Electronics*, vol. 7, no. 12, p. 400, 2018.

[30] C.-Y. Chang and N. Nikaein, "Closing in on 5G Control Apps: Enabling Multiservice Programmability in a Disaggregated Radio Access Network," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 80–93, 2018.

[31] C.-Y. Chang *et al.*, "Slice Orchestration for Multi-Service Disaggregated Ultra-Dense RANs," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 70–77, 2018.

[32] H. Zhao, L. Zhao, K. Liang, and C. Pan, "Radio access network slicing based on C/U plane separation," *China Commun.*, vol. 14, no. 12, pp. 134–141, 2017.

[33] H. Xiang, W. Zhou, M. Daneshmand, and M. Peng, "Network Slicing in Fog Radio Access Networks: Issues and Challenges," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 110–116, 2017.

[34] J. Zeng, X. Su, J. Gong, L. Rong, and J. Wang, "A 5G virtualized RAN based on NO Stack," *China Commun.*, vol. 14, no. 6, pp. 199–208, 2017.

[35] P. Marsch *et al.*, "5G Radio Access Network Architecture: Design Guidelines and Key Considerations," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 24–32, 2016.

[36] L. M. Larsen, M. S. Berger, and H. L. Christiansen, "Fronthaul for Cloud-RAN enabling network slicing in 5G mobile networks," *Wirel. Commun. Mob. Comput.*, vol. 2018, 2018.

[37] J. Yang and Y.-S. Chan, "Toward an Intelligent, Multipurpose 5G Network: Enhancing Mobile Wireless Networks," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 53–60, 2019.

[38] C. Sexton, N. Marchetti, and L. A. DaSilva, "Customization and Trade-offs in 5G RAN Slicing," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 116–122, 2019.

[39] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN Slicing for Verticals: Enablers and Challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, 2019.

[40] E. Pateromichelakis, O. Bulakci, C. Peng, J. Zhang, and Y. Xia, "LAA as a Key Enabler in Slice-Aware 5G RAN: Challenges and Opportunities," *IEEE Commun. Stand. Mag.*, vol. 2, no. 1, pp. 29–35, 2018.

[41] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, 2018.

[42] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective," *IEEE Wirel. Commun.*, vol. 24, no. 5, pp. 166–174, 2017.

[43] A. Ksentini, P. A. Frangoudis, A. PC, and N. Nikaein, "Providing Low Latency Guarantees for Slicing-Ready 5G Systems via Two-Level MAC Scheduling," *IEEE Netw.*, vol. 32, no. 6, pp. 116–123, 2018.

[44] A. Ksentini and N. Nikaein, "Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, 2017.

Chapter 3

Paper A. Automated Network Service Scaling in NFV: Concepts, Mechanisms and Scaling Workflow

Authors:

Oscar Adamuz-Hinojosa, Jose Ordonez-Lucena, Pablo Ameigeiras, Juan Jose Ramos-Munoz, Diego Lopez, and Jesus Folgueira.

The paper has been published in IEEE Communication Magazine, July, 2018.

Abstract

Next-generation systems are anticipated to be digital platforms supporting innovative services with rapidly changing traffic patterns. To cope with this dynamicity in a cost-efficient manner, operators need advanced service management capabilities such as those provided by Network Function Virtualization (NFV). NFV enables operators to scale network services with higher granularity and agility than today. For this end, automation is key. In search of this automation, the European Telecommunications Standards Institute (ETSI) has defined a reference NFV framework that make use of model-driven templates called Network Service Descriptors (NSDs) to operate network services through their lifecycle. For the scaling operation, an NSD defines a discrete set of instantiation levels among which a network service instance can be resized throughout its lifecycle. Thus, the design of these levels is key for ensuring an effective scaling. In this article, we provide an overview of the automation of the network service scaling operation in NFV, addressing the options and boundaries introduced by ETSI normative specifications. We start by providing a description of the NSD structure, focusing on how instantiation levels are constructed. For illustrative purposes, we propose an NSD for a representative Network Service (NS). This NSD includes different instantiation levels that enable different ways to automatically scale this NS. Then, we show the different scaling procedures the NFV framework has available, and how it may automate their triggering. Finally, we propose an ETSI-compliant workflow to describe in detail a representative scaling procedure. This workflow clarifies the interactions and information exchanges between the functional blocks in the NFV framework when performing the scaling operation.

3.1 Introduction

Network softwarization is an unprecedented techno-economic transformation trend that takes advantage of commodity hardware, programmability and reusability of software to provide cost optimizations and service innovation in next-generation networks. Network Function Virtualization (NFV) is a key enabler in this trend. It brings novel practices for flexible and agile Network Service (NS) provisioning and management.

The European Telecommunications Standards Institute (ETSI) has defined a reference NFV architectural framework [1] for the purpose of management and orchestration of NSs in multi-vendor, multi-network environments. This framework consists of three Management and Orchestration (MANO) functional blocks that make use of model-driven templates to deploy and operate multiple instances of different NSs (and their constituents) over a common infrastructure. The Network Service Descriptor (NSD) is the name of the template used for NSs. With the information gathered in a NSD, the MANO blocks are able to manage NS instances throughout their lifecycle with great agility and full automation.

Scaling is a key lifecycle management operation in NFV. Scaling with NFV allows operators to automatically resize NSs at runtime to handle load surges with performance guarantees. This brings dynamicity and cost reduction compared to today's scaling practices, where NS capacity is statically over-dimensioned for the highest predictable traffic peak. To achieve the required automation when scaling, an appropriate model for the NSD is needed.

On one hand, most of works dealing with scaling focus on mechanisms/strategies for virtual resource estimation (e.g., [2]) and allocation (e.g., [3]). The policy-based rules and input data that these proposals use for the scaling operation are not retrieved from an NSD; instead, they are specified manually for every NS. This may led to less agile and more error-prone scaling solutions, where the automation in NFV is not fully exploited.

On the other hand, there are no existing works that analyze the effect the NSD modeling has in the NS scaling operation. The existing MANO platforms based on the NFV framework (e.g., Open Baton, OSM, ONAP or Tacker) use their own data modeling languages (e.g., TOSCA, YANG) for their NSDs. This leads to non-compatible workflows for the NS scaling operation, avoiding reusability and portability of scaling solutions across different NFV platforms [4]. To enable their interoperability, ETSI works on the development of normative specifications for the NFV information model, including interface description, and a platform- and technology-agnostic model for the NSD. Understanding this standardized model, and adapting the existing data models to it, is key for successful scaling operations with the existing MANO platforms. In this line, ETSI NFV has recently started in [5] a work targeted at mapping the TOSCA data model with the NSD information model.

The contribution of this paper is twofold. First, we provide an overview of the structure of the ETSI NSD. We address those NSD fields most relevant for scaling, placing emphasis on the instantiation levels. These levels specify the different sizes an NS instance can adopt throughout its lifecycle. This limits the scaling of an NS instance to one of the discrete set of levels defined in the NSD. Thus, their correct design is key to ensure appropriate scaling operations with NFV. To facilitate their understanding, and show how they are constructed in a NSD, we propose a simple example of an NSD. Secondly, we provide an overview of the automated NS scaling operation, analyzing in depth the options and boundaries introduced by the ETSI NSD information model. We show the different scaling procedures that the NFV framework has available, and how they can be triggered in an automated manner. This includes the proposal of an ETSI-compliant workflow for a representative scaling procedure. This workflow clarifies how the different MANO blocks interact, specifying the information they exchange in each step.

This article is organized as follows. First, we present a background of those NFV concepts relevant for scaling. Then, we provide an insight into the NSD. Next, we detail the most relevant NS scaling procedures, and propose a workflow for one of them. Finally, we remark some conclusions.

3.2 Background on Key NFV Concepts

In this section, we describe the concept of NS and the NFV architectural framework.

3.2.1 The Concept of NS

An NS is a composition of network functions. According to ETSI NFV, network functions may be implemented as Virtualized Network Functions (VNFs) and physical network functions. For simplicity, we only consider the former in this paper.

An NS consists of a set of VNFs, Virtual Links (VLs), and Virtual Network Function Forwarding Graphs (VNFFGs). VLs are abstractions of physical links that logically connect together VNFs. To specify how these connections are made along the entire NS, one or more VNFFGs are used. A VNFFG describes the

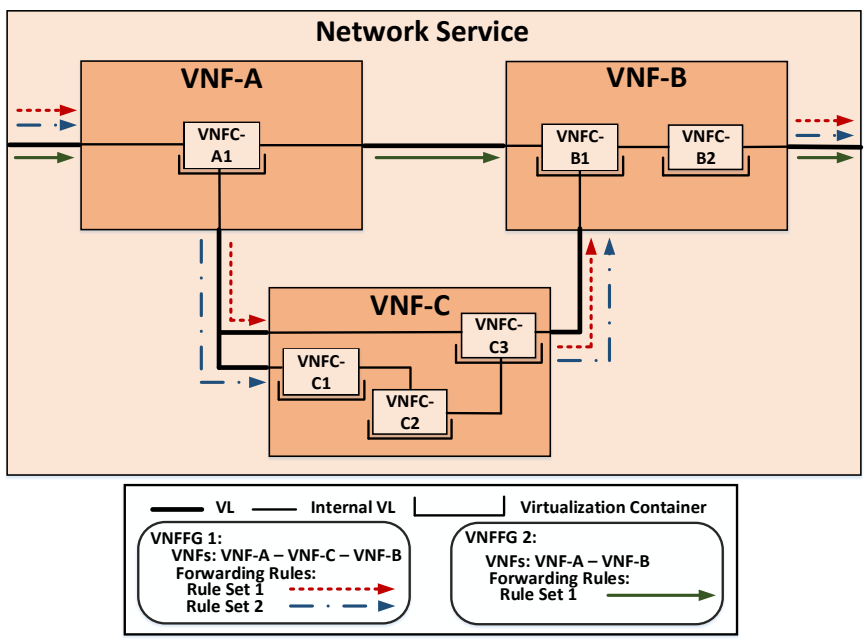


Figure 3.1: NS internal composition. In this example, we have defined two VNFFGs, and we have associated each with a different network plane: VNFFG1 for user plane traffic, and VNFFG2 for management plane traffic. Note that VNFFG1 includes two set of forwarding rules for traffic steering, enabling the definition of two user plane traffic flows, e.g. for distinct processing

topology of (the entire or part of) the NS, and optionally includes forwarding rules to describe how traffic shall flow between the VNFs defined in this topology.

For a fine-grained control of its scalability, performance and reliability, a VNF might be decomposed into one or more Virtualized Network Function Components (VNFCs) [6], each performing a well-defined subset of the entire VNF functionality. Each VNFC is hosted in a single virtualization container, and connected with other VNFCs through internal VLs.

Fig. 3.1 shows an example of an NS with its constituents.

3.2.2 NFV Architectural Framework

ETSI has defined a reference NFV architectural framework [1] for the deployment and operation of NSs. As seen in Fig. 3.2, this framework consists of three main working domains: the Network Function Virtualization Infrastructure (NFVI)

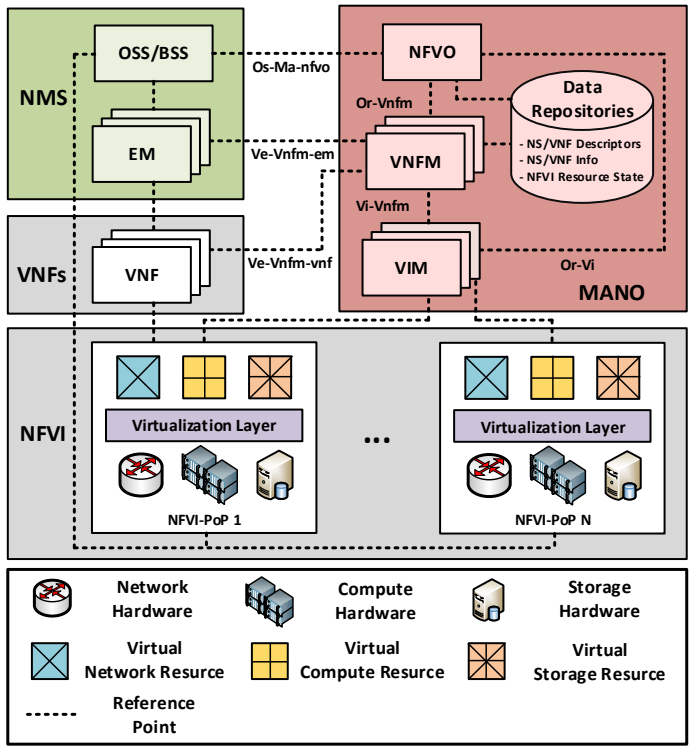


Figure 3.2: ETSI NFV Architectural Framework. The three working domains, and their constituent functional blocks communicate together using a set of reference points.

and VNFs, MANO, and Network Management System (NMS).

The NFVI is the collection of resources that make up the cloud on top of which VNFs run. With the help of a virtualization layer, the underlying physical resources are abstracted and logically partitioned into virtual resources, used for hosting and connecting VNFs. NFVI might span across several geographically remote Network Function Virtualization Infrastructure Point of Presences (NFVI-PoPs), enabling multi-site VNF deployments.

MANO focuses on the virtualization-specific deployment and operation tasks in the NFV framework [7]. MANO consists of three functional blocks, including:

- Virtual Infrastructure Managers (VIMs), each managing the resources of one or more NFVI-PoPs.
- Virtual Network Function Managers (VNFMs), focused on the lifecycle

management of the VNFs, and responsible for their performance and fault management at virtualized resource level.

- Network Function Virtualization Orchestrator (NFVO), that orchestrates NFVI resources across VIMs, and performs NS lifecycle management.

The MANO also includes data repositories to assist these blocks with their tasks. These repositories include: (a) NS and VNF Descriptors, (b) information about all the NS/VNF instances during their lifecycle (NS/VNF Info), and (c) updated information about the state (allocated/reserved/available) of NFVI resources.

Finally, the NMS focuses on traditional (non virtualized-related) management tasks, orthogonal to those defined in MANO. NMS comprise:

- Element Managers (EMs), responsible for the fault, performance, configuration, accounting, and security management of the VNFs at application level.

- Operations Support System (OSS)/Business Support System (BSS), comprising traditional systems and management applications that help operators to provision and operate their NSs.

3.3 NS Description

In this section, we first study the NSD structure. Then, we show an example of an NSD. For illustrative purposes, we propose an NSD for the NS shown in Fig. 3.1.

3.3.1 NSD Overview

An NSD is a deployment template that contains machine-processable information used by MANO blocks to create instances of an NS, and operate them throughout their lifetime. An NSD is constructed from a set of attributes and other descriptors (see Fig. 3.3).

The attributes that an NSD includes enable specifying how to deploy and operate instances of an NS. In this work, we consider those that are most relevant for NS scaling (see Fig. 3.3):

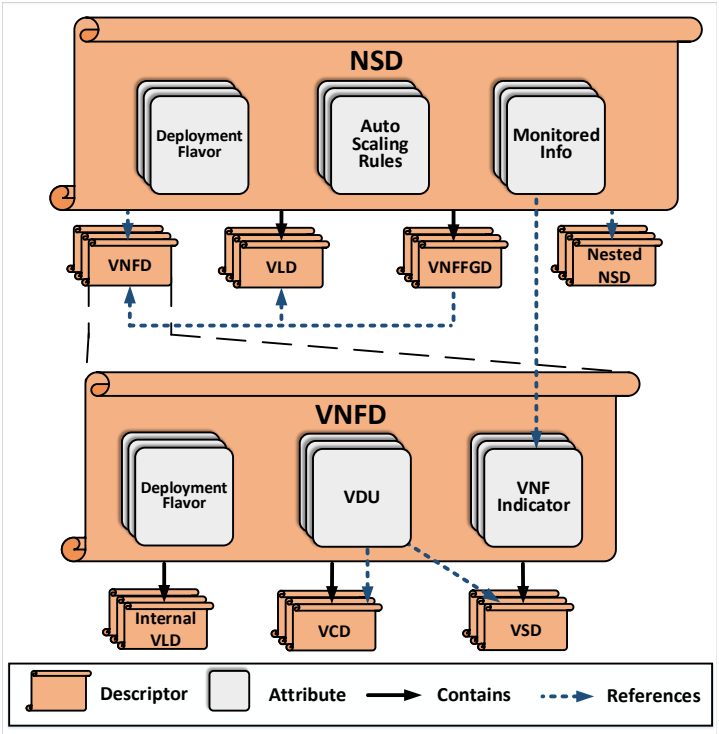


Figure 3.3: NSD structure. Only the descriptors and attributes that are most relevant for NS scaling are shown.

- *Monitored Info*: specifies the information to be tracked for NS performance and fault management. This information includes resource-related performance metrics (at NS/VNF level), and VNF indicators from NS's constituent VNFs.

- *Auto Scaling Rules*: contain rules that enables triggering scaling actions on an NS instance when a condition involving Monitored Info is not satisfied. The NFV information model allows expressing these rules as customized scripts provided at instantiation time. The language used for these scripts shall support conditions involving not only logical/comparison operators with scalar values, but also complex analytical functions able to process statistical data correlated from different sources.

- *Deployment Flavors*: describe specific deployment configurations for the NS. For a more detailed description, see section 3.3.2.

To describe the deployment and operational behavior of NS constituents, NSD contains and references a set of descriptors, including Virtualized Network Function Descriptors (VNFDs), Virtual Link Descriptors (VLDs), and Virtual Network Function Forwarding Graph Descriptors (VNFFGDs) [8]. A VNFD contains information required to deploy and operate instances of a VNF. A VLD provides information of a VL, including the deployment configurations available for VL instantiation. These configurations are specified through Deployment Flavors. Different configurations results in different performance and reliability levels for VNF connectivity. Finally, a VNFFGD references the VNFDs and VLDs for topology description.

From the above descriptors, we concentrate on the VNFDs. The NSD references information of VNFDs that is essential for NS scaling. Similar to an NSD, a VNFD also includes descriptors and attributes.

The descriptors that VNFD contains provide a detailed view on the VNF internal composition. Particularly, a VNFD includes Virtual Compute Descriptors (VCDs), Virtual Storage Descriptors (VSDs), and internal VLDs. The first two specify the virtual compute and storage resources that are needed for VNFCs hosting, while the latter specifies the performance requirements for VNFCs connectivity.

In terms of attributes, a VNFD includes one or more:

- *VNF Indicators*: represent performance/fault-related events that provide information of the VNF at application level.
- *Virtual Deployment Units (VDUs)*: describe how to create and operate instances of VNFCs; hence, an VDU can be seen as a VNFC descriptor. A VDU specifies the compute resources (and optionally storage resources) that a virtualization container needs to host a VNFC. To that end, it references one VCD (and optionally one or more VSDs).
- *Deployment Flavors*: similar to those defined in the NSD, but applied to VNFs.

As seen, the models for the NSD and VNFD are very similar: VNFs/VDUs connected by VLDs, and defining various Deployment Flavors. Thus, some initiatives like the Superfluidity project [4] have suggested the idea of having a

common, reusable information model for both of them, so that they can share the same interfaces and lifecycle management operations. This model shall be recursive and scalable. ETSI NFV already enables this recursivity and scalability at NS level, with the concept of composite NSs (i.e., an NS composed of smaller, nested NSs) [8]. However, ETSI NFV considers the necessity to maintain different models for NSDs and VNFDs, due to some technical issues that can be found in [8, 9].

From the perspective of NS scaling, the Deployment Flavors within an NSD are key attributes, as they contain the instantiation levels permitted for an NS instance. These levels are constructed with information included in the flavors of the VNFDs and VLDs. In section 3.3.2, we describe the different flavors, studying how they enable the definition of different instantiation levels in the NSD.

3.3.2 Deployment Flavors and Instantiation Levels

As seen earlier, there are three types of deployment flavors: VL flavors, VNF flavors, and NS flavors.

Selecting a VL flavor enables selecting specific Quality of Service (QoS) parameters (latency, jitter, etc.) and transport reliability for a VL.

Each VNF flavor within a VNFD can be used to define a different deployment configuration for a VNF. A given deployment configuration specifies the functionality and the performance level(s) allowed for instantiating the VNF. To specify the VNF functionality (i.e. which features need to be activated for the VNF), the VNF flavor indicates which (subset of) VNFCs need to be deployed. Particularly, the flavor references the VDUs to be used for their instantiation. To specify the performance level(s) permitted for VNF instantiation (i.e. which amount of resources are needed for each selected VNFC), the flavor defines one or more VNF-ILs. Each VNF-IL indicates, for each VDU referenced in the flavor, the number of VNFC instances that need to be deployed from this VDU.

Finally, NS flavoring enables adjusting the functionality and the level of performance of an NS. A NS flavor selects the VNFs and VLs to be deployed as part of the NS, and their actual flavors. For each selected VNF flavor, the subset of VNF-ILs to be used is also specified. With this information, an NS flavor defines one or more Network Service Instantiation Levels (NS-ILs). NS-ILs are similar

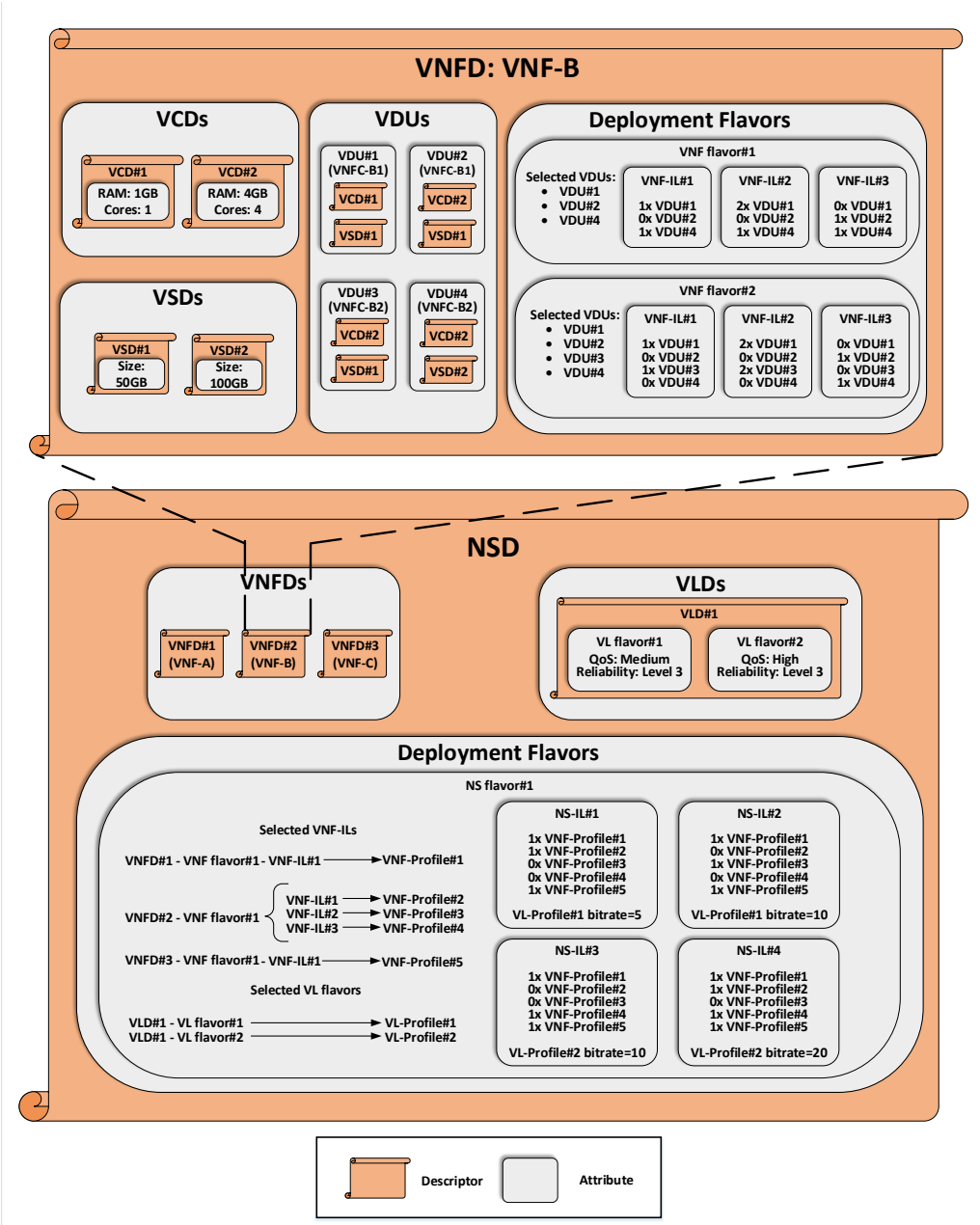


Figure 3.4: A NSD proposal for the NS given in Fig. 3.1. Please note that only the most relevant attributes for scaling are shown. For better understandability, the Virtualized Network Function Instantiation Levels (VNF-ILs) selected for the NS flavor are referred to as VNF-Profiles. Similarly, the VL flavors selected for the NS flavor are referred to as VL-Profiles. The profile term is also used in ETSI NFV. See [8] for more information.

to VNF-ILs, but at NS level. Each NS-IL specifies:

- The number of VNF instances to be deployed from each VNF-IL specified in the NS flavor.
- The VLs required for VNF connectivity, their bitrate, and the VL flavors used for their instantiation.

As seen above, NS-ILs are key for NS scaling. The fact that an NS instance needs to be scaled means that the current NS-IL is no longer valid for that instance, and hence a new NS-IL must be used. In this case, the NFVO must select, among the finite set of NS-IL predefined in the NS flavor, the optimum one for scaling the instance (see section 3.4). Note that the operator's policy adopted for NS flavor design (the number of NS-ILs, and the differences among them in terms of resource requirements) has a great impact on the NS scaling operation at operation time.

To clarify these concepts, in Fig. 3.4 we provide an example of an NSD for the NS shown in Fig. 3.1. For simplicity, we only address the VNFD corresponding to VNF-B: VNFD#2. The rest of VNFDs could be constructed similarly.

VNFD#2 includes four VDUs. From these VDUs, instances of VNFC-B1 and VNFC-B2 with different capacity can be created. For example, VDU#1 and VDU#2 are used to create instances of VNFC-B1 with low/high capacity, respectively. VNFD#2 also includes two flavors: VNF Flavor#1, enabling only VNFC-B1 scaling; and VNF Flavor#2, support the scaling of both VNFC-B1 and VNFC-B2. Each flavor includes three VNF-ILs, differing in their resource requirements.

The NSD references the descriptors of the NS's constituent VNFs: VNFD#1, VNFD#2 and VNFD#3. For simplicity, we assume VNFD#1 and VNFD#3 have each a single flavor with a single VNF-IL. VNF-A, VNF-B, and VNF-C are interconnected through VLs that can be instantiated from the two defined VLs flavors. In this example, we consider a single NS flavor that only allows the scaling of VNF-B, restricting VNF-A and VNF-C to a single instance each. This flavor presents four predefined NS-ILs. These NS-ILs enable two scaling cases: (a) increasing/decreasing the capacity of a VNF-B instance, and (b) adding/removing a VNF-B instance. The first three NS-ILs are used for (a), where there

is one instance of VNF-B. When NS-IL#3 is reached, the VNF-B instance can no longer increment its capacity. At this point, the only way to scale this NS is by adding a new VNF-B instance, as stated in NS-IL#4. With this NS-IL, the NS instance has two instances of VNF-B.

3.4 NS Scaling Automation

In this section, we describe how the NS scaling operation may be automated with MANO, considering the boundaries that ETSI NFV specifications impose. We detail the input information that the NFVO takes to determine if an NS instance needs to be scaled. Assuming the scaling is required, we show the different scaling procedures that NFVO may trigger. Finally, we propose a detailed workflow to describe one of them, illustrating the messages the MANO blocks exchange in that procedure.

3.4.1 Boundaries and Procedures

Although NS scaling can be manually triggered, automation enables operators to fully exploit NFV benefits. To automate the scaling triggering, NFVO has a customizable software module (e.g. supporting NS-specific code) that runs a Dynamic Resource Provisioning Algorithm (DRPA). The DRPA determines when an operative NS instance needs to be scaled, and the NS-IL which optimizes the scaling of that instance according to a set of criteria. This optimum NS-IL will then be used by the NFVO to trigger an appropriate scaling procedure.

The DRPA takes the following input parameters:

- Performance and fault data, as specified in the Monitored Info attribute (see Fig. 3.3). This includes periodical resource-related performance metrics [6, 10, 11, 12, 13], and/or asynchronous alarms (performance metric-based threshold crossing, and VNF indicator value changes) [11, 12, 13].

- Runtime information of the NS instance and each constituent VNF instance, accessible from the NS Info and each VNF Info.

- The entire set of NS-ILs and VNF-ILs available for use in the NSD and VNFDs. These levels, built by NSD/VNFD developers at design time,

cannot be changed at operation time. In case they need to be updated, DevOps strategies like those proposed in [14] could be used.

- Resource capacity information from each accessible VIM. This information can be found in the data repositories (see Fig. 3.2).

The DRPA applies the Auto Scaling Rules to the incoming performance/fault data. If they are not satisfied, NS scaling is required. In that case, the DRPA determines the NS-ILs that are candidate to satisfy the performance/fault criteria specified in the Auto Scaling Rules. Over these candidates, the DRPA applies the pertinent optimization criteria (e.g., minimize resource costs, energy consumption) and a set of constraints (e.g., available resource capacity, placement constraints) to output:

- The optimum NS-IL.
- The NFVI-PoPs that will accommodate the virtual resources associated to this optimum NS-IL. Moving from the current NS-IL towards the optimum NS-IL may entail the allocation and release of resources. For each new resource to be allocated, the DRPA selects the NFVI-PoP where this resource will be accommodated. Next, NFVO determines which VIMs provide access to the selected NFVI-PoPs.

Using these outputs, the NFVO triggers one of the following NS scaling procedures [15]:

- *VNF scaling*: One or more VNF instances in the NS instance modify their capacity by changing their VNF-ILs, and hence by adding and/or removing VNFC instances. This procedure assumes the new VNF-ILs are selected from the VNF flavor currently used.
- *Adding/Removing VNF instances*: Instances of existing VNFs are added/removed in the NS instance. For each VNF instance to be added, it is required to select its VNF-IL from a given VNF flavor.

As seen above, the specific procedure to be triggered is subjected to the differences that exist, in terms of VNF-ILs, between the two NS-ILs: the NS-IL of the NS instance, and the optimum NS-IL that DRPA has chosen. The

possibility of choosing between different candidate NS-ILs (and hence triggering one of the above procedures) in the scaling operation adds flexibility compared to the autoscaling strategies present in the existing MANO solutions. The data models (TOSCA, YANG, or Heat Orchestration Templates [HOT]) used for their descriptors of NSs (and their constituents) have predefined, rigid autoscaling policies that do not allow choosing between different level of resources; instead, they set the level towards the NS (or one of its constituents) shall be scaled to.

3.4.2 Scaling Operation Workflow

As seen earlier, the goals of the VNF Scaling and Adding/Removing VNF instances procedures are different. However, the ways of performing them with MANO are very similar. Indeed, the addition/removal of VNF instances is no more than an extension of a VNF Scaling, with the peculiarity that the former implies (a) adding/removing all the VNFC instances of each VNF instance, and (b) instantiate/modify/remove VLs for VNF connectivity. Due to limited space, we concentrate on VNF Scaling in this paper.

In Fig. 3.5, we show the workflow messages for scaling a single VNF instance of an NS. These messages have been grouped into distinct phases: information collection, scaling triggering, resource allocation and resource release. Although the last two phases may be performed independently, it could happen both are required in the same scaling scenario (e.g., replacing a running VNFC instance by other instance of greater capacity). In that case, allocation goes before release to guarantee service continuity (e.g., when starting the new VNFC instance, the old instance can be deleted).

In the information collection phase, the NFVO gathers performance/fault data from VNFMs and VIMs. Performance metrics at NS/VNF level are reported with Performance Information Available Notifications, and performance metric-based threshold crossed values with Threshold Crossed Notification (steps 1-2). VNF Indicator values are changed by EMs, and notified to NFVO by VNFMs (step 3).

In the scaling triggering phase (step 4) the DRPA uses the above information, along with the information the NFVO has accessible from data repositories (NSD/VNFDs, NS/VNF Info, and NFVI resources state) to decide the optimum

NS-IL. If we assume the optimum NS-IL differs from the existing one in the VNF-IL of a single VNF instance, a VNF scaling procedure will be triggered. This would happen, for example, if the DRPA decides to scale an instance of the NS shown in Fig. 3.1, moving it from NS-IL#1 to NS-IL#3 (see Fig. 3.4). This means scaling the VNF-B instance from VNF-IL#1 to VNF-IL#3. In other words, incrementing the capacity of the VNFC-B1 instance, while leaving the VNFC-B2 instance unmodified. Note that this change involves both resource allocation and resource release phases.

The NFVO requires the VNFM to scale the VNF instance, sending it the new VNF-IL in the Scale VNF to Level Request[1]. Now, VNFM can initiate the scaling operation. To that end, the VNFM provides this lifecycle management operation with an unique ID using the Scale VNF to Level Response (step 5). The VNFM will use this ID to notify the NFVO the start and later the result of this operation[2]. Finally, the VNFM consults the VNF Info and the VNFD to compare the current VNF-IL against the new one. From this comparison, the resources to be allocated and/or released for this scaling operation can be derived.

In the resource allocation phase, we distinguish the following sub-phases:

- *Resource Reservation* (optional [7]): Prior to this sub-phase, VNFM asks NFVO for permission to allocate resources. To that end, VNFM sends NFVO the IDs of VDUs and internal VLs that map to the resources to be allocated (step 6). Although the NFVO already had this information after step 5, ETSI specifications impose this information exchange [12]. Then, the resource reservation sub-phase begins. From the output of the DRPA, the NFVO knows which NFVI-PoPs shall accommodate the resources to be allocated, and the VIMs providing access to those NFVI-PoPs. Now, these resources can be reserved for later allocation. Each selected VIM receives three reservation requests (step 7), one for each resource type (compute, storage, and network) it has to reserve in the NFVI-PoPs under its management. These requests include the placement constraints applicable to

[1]Scale VNF operation message might also be used, but it has some limitations. See [12] for more information.

[2]Some lifecycle management operations require sending NFVO start and result notifications. For simplicity, all these notifications are omitted in the workflow

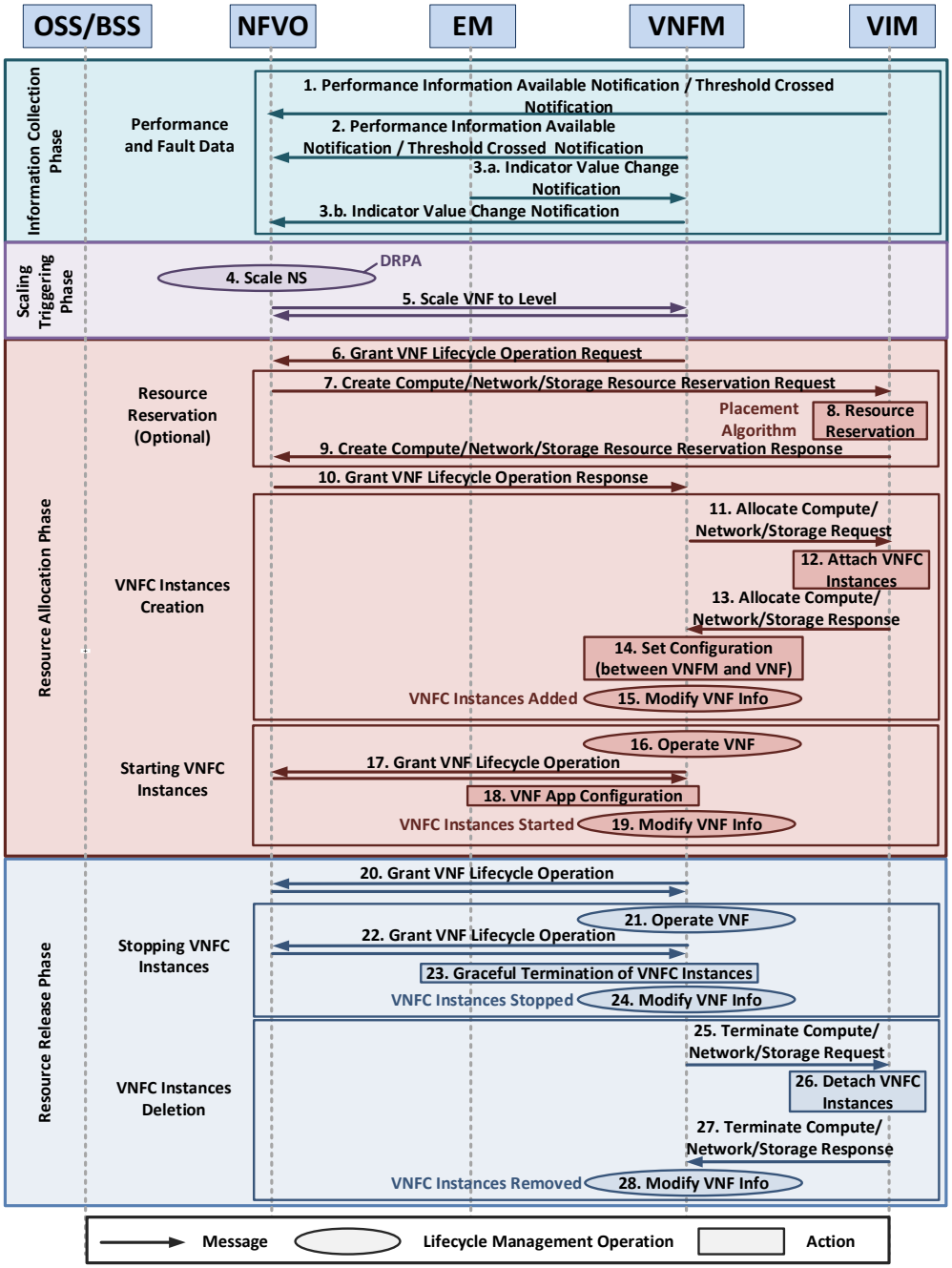


Figure 3.5: Workflow for the VNF Scaling Procedure.

the specified resources. The VIM uses these constraints to perform a placement algorithm (step 8), deciding the appropriate NFVI-PoPs' resource zones [10] where resources are reserved. Then, the VIM sends the NFVO (step 9) the IDs of reserved resources. Finally, the NFVO sends the VNFM those IDs, and connectivity information [12] for each selected VIM. Now, VNFM knows how to access those VIMs, and which one may allocate each resource. If this sub-phase is not performed, two issues need to be considered. First, only steps 6 and 10 are executed. Secondly, the placement algorithm is now performed after resource allocation request (see next sub-phase).

- *VNFC Instances Creation*: The VNFM sends the reservations IDs to the corresponding VIMs (step 11) for resource allocation (step 12). At this point, VNFC instances have been created and their connectivity enabled. The IDs of the allocated resources are then sent to the VNFM (step 13). In step 14, the VNFM triggers the configuration of VNFC instances. Finally, VNFM updates the VNF Info in the data repository (step 15) to reflect the creation of new VNFC instances, and set their state to STOPPED.

- *Starting VNFC Instances*: To start the functionality of the new instances, the VNFM triggers an Operate VNF lifecycle operation (step 16). This operation will force (at the end of this sub-phase) the change of the instances' state from STOPPED to STARTED. Once the NFVO grants this operation (step 17), the new VNFC instances are configured at application level (step 18). To that end, VNFM communicates with the EM. Lastly, VNFM updates the VNF Info (step 19), changing the state of the new instances from STOPPED to STARTED. Note that some of the running VNFC instances could be affected by the creation of the new ones, and hence need to be (re)configured in terms of connectivity (e.g. new interfaces, updated link requirements) and/or application (e.g. sending/receiving packets to/from the new instances). In that case, the VNFM would order the corresponding VIMs and/or EM, respectively to make the necessary changes.

For the Resource Release Phase, we have two sub-phases:

- *Stopping VNFC Instances*: In step 20, VNFM asks NFVO for permission

to release resources. Then, VNFM triggers an Operate VNF lifecycle operation (step 21) to gracefully terminate some VNFC instances (forcing the stopping of VNFC instance at the end of this sub-phase). In step 23, affected instances are (re)configured (following counterpart strategies to those specified in the Starting VNFC instances sub-phase), and instances to be terminated are shut down. Finally, the state of stopped instances is changed from STARTED to STOPPED (step 24).

- *VNFC Instances Deletion*: The VNFM sends to the corresponding VIMs the IDs of the resources that host and connect the stopped VNFC instances (step 25). At this point, these instances are deleted (step 26). Then, the VIMs send back to the VNFM the resource IDs of released resources (step 27). After receiving those IDs, the VNFM updates the VNF Info (step 28) to reflect the instance deletion.

3.5 Conclusions

In this article, we shed light on the NS scaling operation with NFV. The options for automatically scaling a NS with the NFV framework are limited by the way the NSD is constructed. During its lifecyle, an NS instance only can move among the instantiation levels defined in the NSD, so their design is critical to ensure an effective automated scaling. In this work, we have analyzed how these levels are built in a NSD. To facilitate their understanding, we have proposed an NSD example, where different instantiation levels are included for scaling a NS.

We also have shown the different procedures the NFVO may trigger to scale an NS instance according to ETSI specifications, and how NFVO may automate them. To that end, the NFVO runs a DRPA that, taking NSD content and information of the operative NS instance, determines the optimum instantiation level towards the NS instance must be scaled to. This output forces the way the scaling procedures are performed in NFV. For one representative procedure, we have proposed an ETSI-compliant workflow that clarifies the interactions and information exchanges between the functional blocks in the NFV framework.

Acknowledgments

This work is partially supported by the Spanish Ministry of Economy and Competitiveness, the European Regional Development Fund (Project TEC2016-76795-C6-4-R), the Spanish Ministry of Education, Culture and Sport (FPU Grant 16/03354), and the University of Granada, Andalusian Regional Government and European Social Fund under Youth Employment Program.

References

[1] ETSI GS NFV 002 V1.1.1, "Network Functions Virtualisation (NFV); Architectural Framework," Oct. 2013.

[2] R. Mijumbi, S. Hasija, S. Davy, A. Davy, B. Jennings, and R. Boutaba, "Topology-Aware Prediction of Virtual Network Function Resource Requirements," *IEEE Trans. Netw. Serv. Manag.*, vol. 14, no. 1, pp. 106–120, 2017.

[3] J. Gil Herrera and J. F. Botero, "Resource Allocation in NFV: A Comprehensive Survey," *IEEE Trans. Netw. Serv. Manag.*, vol. 13, no. 3, pp. 518–532, 2016.

[4] S. Salsano, F. Lombardo, C. Pisa, P. Greto, and N. Blefari-Melazzi, "RDCL 3D, a model agnostic web framework for the design and composition of NFV services," in *2017 IEEE Conference on NFV-SDN*, pp. 216–222, 2017.

[5] ETSI GS NFV-SOL 001 V0.6.0, "Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; NFV Descriptores based on TOSCA specification," Mar. 2018.

[6] ETSI GS NFV-SWA 001 V1.1.1, "Network Functions Virtualisation (NFV); Virtual Network Functions Architecture," Dec. 2014.

[7] ETSI GS NFV-MAN 001 V1.1.1, "Network Functions Virtualisation (NFV); Management and Orchestration," Dec. 2014.

[8] ETSI GS NFV-IFA 014 V2.4.1, "Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Network Service Templates Specification," Feb. 2018.

[9] ETSI GS NFV-IFA 011 V2.4.1, "Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; VNF Packaging Specification," Feb. 2018.

[10] ETSI GS NFV-IFA 005 V2.4.1, "Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Or-Vi reference point - Interface and Information Model Specification," Feb. 2018.

[11] ETSI GS NFV-IFA 006 V2.4.1, "Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Vi-Vnfm reference point - Interface and Information Model Specification," Feb. 2018.

[12] ETSI GS NFV-IFA 007 V2.4.1, "Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Or-Vnfm reference point - Interface and Information Model Specification," Feb. 2018.

[13] ETSI GS NFV-IFA 008 V2.4.1, "Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Ve-Vnfm reference point - Interface and Information Model Specification," Feb. 2018.

[14] S. Dräxler *et al.*, "SONATA: Service programming and orchestration for virtualized software networks," in *2017 IEEE ICC Workshops*, pp. 973–978, 2017.

[15] ETSI GS NFV-IFA 013 V2.4.1, "Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Os-Ma-Nfvo reference point - Interface and Information Model Specification," Feb. 2018.

Chapter 4

Paper B: Harmonizing 3GPP and NFV Description Models: Providing Customized RAN Slices in 5G Networks

Authors:

Oscar Adamuz-Hinojosa, Pablo Munoz, Jose Ordonez-Lucena, Juan Jose Ramos-Munoz, and Juan Manuel Lopez-Soler.

The paper has been published in the IEEE Vehicular Techonology Magazine, December, 2019.

Abstract

The standardization of Radio Access Network (RAN) in mobile networks has traditionally been led by Third Generation Partnership Project (3GPP). However, the emergence of RAN slicing has introduced new aspects that fall outside 3GPP scope. Among them, network virtualization enables the particularization of multiple RAN behaviors over a common physical infrastructure. Using Virtualized Network Functions (VNFs) that comprise customized radio functionalities, each virtualized RAN, denominated RAN slice, could meet its specific requirements. Although 3GPP specifies the description model to manage RAN slices, it can neither particularize the behavior of a RAN slice nor leverage the Network Function Virtualization (NFV) descriptors to define how its VNFs can accommodate its spatial and temporal traffic demands. In this article, we propose a description model that harmonizes 3GPP and European Telecommunications Standards Institute (ETSI)-NFV viewpoints to manage RAN slices. The proposed model enables the translation of RAN slice requirements into customized virtualized radio functionalities defined through NFV descriptors. To clarify this proposal, we provide an example where three RAN slices with disruptive requirements are described following our solution.

4.1 Introduction

The Fifth Generation (5G) networks aim to boost the digital transformation of industry verticals. These verticals may bring a wide variety of unprecedented services with diverging requirements in terms of functionality and performance. Considering each service separately and building a Radio Access Network (RAN) accordingly would be unfeasible in terms of cost. To economically provide these services, Radio Access Network (RAN) slicing has emerged as a solution [1]. It consists of the provision of multiple RAN slice subnets, each adapted to the requirements of a specific service, over a common wireless network infrastructure.

The leading standardization body on RAN slicing is the Third Generation Partnership Project (3GPP). It defines a RAN slice subnet as a set of next Generation NodeBs (gNBs) that are arranged and configured to provide a particular RAN behavior. To manage its lifecyle, the 3GPP defines the RAN Network Slice

Subnet Management Function (NSSMF) and the Network Function Management Functions (NFMFs) as the management entities; and the RAN Network Slice Subnet Template (NSST) as the deployment template [2].

To achieve the flexibility and modularity that a RAN slice subnet requires, some gNB functionalities can be implemented by software, i.e., by Virtualized Network Functions (VNFs) [3]. However, the lifecycle management of VNFs and the orchestration of their resources goes beyond 3GPP scope. The European Telecommunications Standards Institute (ETSI), specifically the Network Function Virtualization (NFV) group, is playing a significant role on these tasks. To that end, ETSI-NFV has defined the NFV Management and Orchestration (MANO) and NFV descriptors.

Focusing on RAN slicing descriptors, the RAN NSST considers the gNB functionalities of a RAN slice subnet. However, the 3GPP has not specified how these functionalities must be configured to meet the requirements for a specific service, typically enhanced Mobile Broadband (eMBB), ultra-Reliable Low Latency Communication (uRLLC), and massive Machine Type Communication (mMTC). Additionally, the RAN NSST neglects the resource requirements for the virtualized deployment of some gNB functionalities. For this, the RAN NSST could use the NFV descriptors. Notwithstanding, describing the virtual resources to accommodate the fluctuations of spatial and temporal traffic demands of a RAN slice subnet is a challenge.

Recent works have addressed the description of RAN slice subnets. For instance, the authors of [4] propose a set of configuration descriptors to parametrize the features, policies and radio resources within the gNBs of a RAN slice subnet. With these descriptors, this work provides a first attempt to define the customized behavior of a RAN slice subnet. However, 3GPP completed the New Radio (NR) specifications after that work, thus the impact of the NR parameters in RAN have not been analyzed in depth yet. Additionally, although this work considers partially-virtualized gNBs, it neglects the description of the virtual resources required to build up them. Thereby, describing the spatial and temporal traffic demands of a RAN slice subnet with NFV descriptors is still an open question.

In this article, we provide a description model for RAN slicing. By harmonizing 3GPP and ETSI-NFV scopes, the proposed solution allows the management of virtualized gNB functionalities, and their customization by setting predefined

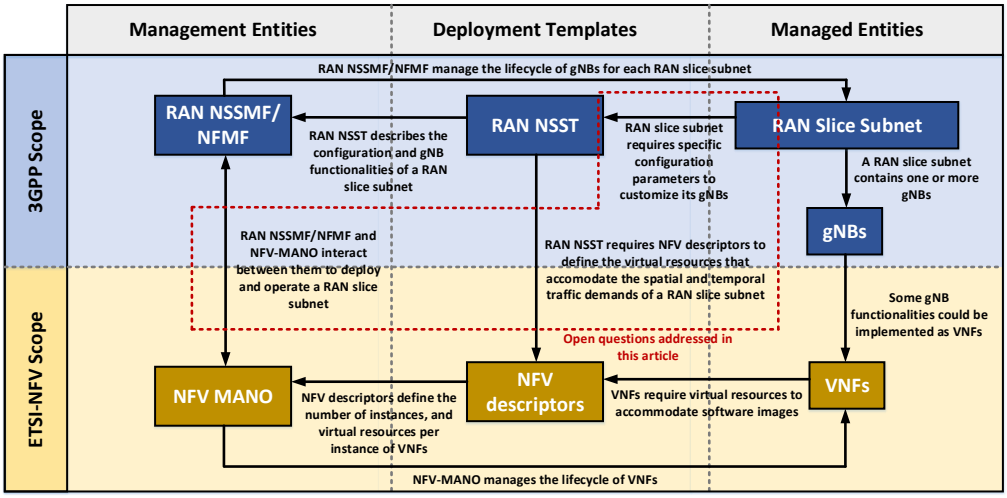


Figure 4.1: Relationship between the 3GPP and ETSI-NFV scopes for the deployment and operation of RAN slices subnets. The aspects within the dotted box are open questions that are addressed in this article.

radio parameters. Thereby, an operator could efficiently provide RAN slice subnets to accommodate the services demanded by verticals on a geographical area with specific spatial and temporal traffic demands. To gain insight into this proposal, we provide an example where RAN slice subnets for eMBB, uRLLC, and mMTC are described based on the proposed solution. For comprehensibility purposes, Fig. 4.1 illustrates the context and the addressed issues of this article.

4.2 RAN Slicing Enablers

4.2.1 NG-RAN Architecture

The 3GPP has defined the Next Generation Radio Access Network (NG-RAN) as the 5G RAN architecture. This architecture comprises gNBs connected to the 5G Core Network. Each gNB provides NR user/control plane protocol terminations towards the User Equipments (UEs). In turn, each gNB comprises one Centralized Unit (CU), multiple Distributed Units (DUs) and multiple Radio Units (RUs) [5].

As depicted in Fig. 4.2, the gNB functionalities are distributed over CU, DUs and RUs in a flexible way. The RUs comprises at least radio-frequency circuitry,

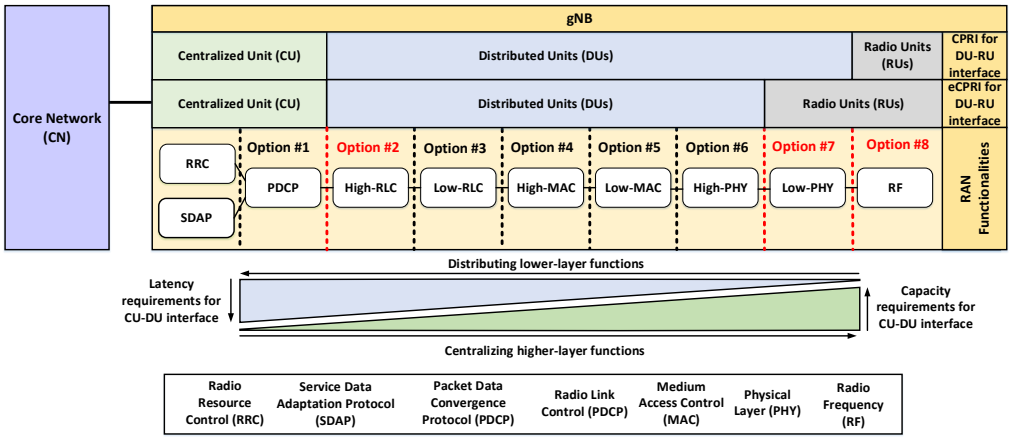


Figure 4.2: 3GPP functional split options for the gNB. Among these split options, the option #2 is the best candidate for CU-DUs splitting and the options #7 and #8 for DUs-RUs splitting in short-term deployments. Note that the latency requirements for CU-DU interface refers to the maximum tolerable latency provided by this transport link. Above this value, the data transmission between CU and DU would be desynchronized.

thus their functionalities are implemented as Physical Network Functions (PNFs), i.e., dedicated hardware appliances. The remaining functionalities, gathered in the DUs and the CU, may be virtualized as VNFs. The DUs contain low-layer functionalities whereas the CU includes high-layers functionalities. According to 3GPP, there exists up to eight options to split radio functionalities between the CU and DUs. The aim of functional split is to leverage the benefits of virtualization (e.g., reducing costs and dynamic scalability) and centralization (e.g., statistical multiplexing gains).

However, the majority of these options present a set of issues and challenges that will difficult their short-term implementation [5]. For this reason, there is a consensus in the industry and academia that the most feasible implementation is the option #2 for splitting CU-DUs. This option could be implemented on the basis of Dual Connectivity (DC) standard.

Regarding the functional split for DUs-RUs, the Common Public Radio Interface (CPRI) has arisen as a standard for implementing option #8. It enables the transmission of baseband signals over transport links. The main drawback of this option is the higher capacity required for these links. To relieve the data rate demands between DUs and RUs, the evolved Common Public Radio Interface

(eCPRI) standards proposed aggregating the low-layer functionalities of Physical Layer (PHY) in the RU, resulting in the split option #7. Furthermore, eCPRI allows an efficient and flexible radio data transmission via a packet-transport network like IP or Ethernet. However, the aggregation of Low-PHY functionalities leads to a significantly higher cost of RUs. In this article, we assume that the implementation of split options #7 or #8 will depend on the features of the transport network in each deployment area.

In short-term deployments, the CU will be executed as a VNF in a Network Function Virtualization Infrastructure Point of Presence (NFVI-PoP), i.e., a cloud site where VNFs can run, while DUs will be likely implemented as PNFs.

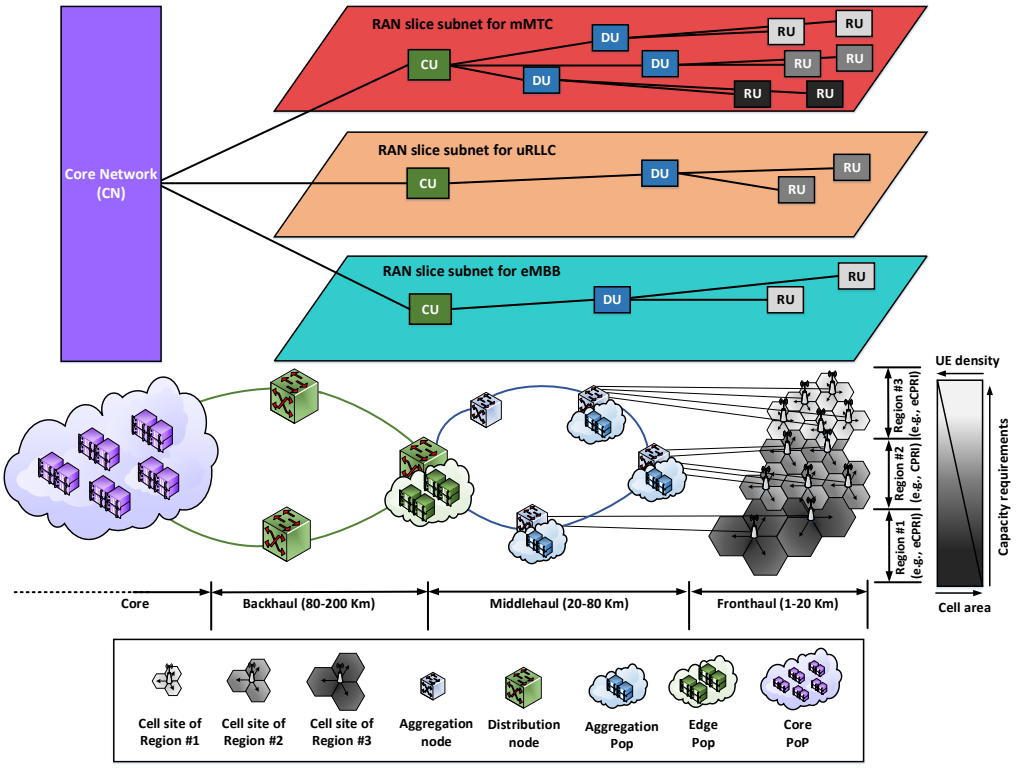


Figure 4.3: Deployment perspective of RAN slice subnets for mMTC, uRLLC, and eMBB, respectively. By way of example, the RAN slice subnet for mMTC is deployed over the three regions. The RAN slice subnet for uRLLC is deployed over the Region #2. The RAN slice subnet for eMBB is deployed over the Region #1. Furthermore, fronthaul links for Regions #1 and #3 use eCPRI whereas for Region #2 use CPRI.

There are two main reasons. First, the software images of DUs must be optimized to execute ms procedures. Secondly, to satisfy the stringent latency requirements, NFVI-PoPs hosting DUs must be installed near users, even closer the NFVI-PoPs hosting CUs.

Despite these issues, researchers are working on the DU virtualization. Some works (e.g., [6]) consider a hierarchical structure of NFVI-PoPs to enable the virtualization of both, the CU and the DU. Furthermore, some gNB software implementations (e.g., OpenAirInterface [6]) are considering the CU-DUs split.

Assuming virtualized CU and DUs in this article, the RAN infrastructure requires a hierarchical structure of NFVI-PoPs in addition to cell sites, as depicted in Fig. 4.3. These NFVI-PoPs might be hosted in the aggregation and distribution nodes that connect the cell sites with the Core Network [7]. Since an aggregation node serves multiple RUs, the hosted NFVI-PoP, could allocate DUs per each RAN slice subnet that requires the coverage area of these RUs. Similarly, the NFVI-PoP hosted in a distribution node could allocate CUs serving the DUs of each RAN slice subnet.

Focusing on an aggregation NFVI-PoP, if the geographical region served by this NFVI-PoP has a high UE density, the allocated DU of a RAN slice subnet will usually serve more cell sites, thus requiring more virtual resources to deal with the aggregated traffic. Similarly, a DU serving a region with low cell sites density, will usually require less virtual resources.

In an edge NFVI-PoP, the amount of virtual resources required by a CU depends on the number of served DUs and the cell sites density supported by each DU.

4.2.2 3GPP RAN Slicing Management Functions and Descriptor

To manage the lifecycle of RAN slice subnets, the 3GPP has defined the RAN NSSMF and the NFMFs [2]. The RAN NSSMF (a) translates the performance and functional requirements of a gNB into the amount of the virtual resources that accommodate the gNBs; and (b) manages the Fault, Configuration, Accounting, Performance, and Security (FCAPS) of the gNBs from the application perspective. Each NFMF is specific for a type of gNB component (i.e., CU, DUs, or RUs), and is controlled by the RAN NSSMF to carry out the activities related

to (b).

To automate the lifecycle management of each RAN slice subnet, the RAN NSSMF uses RAN NSSTs. Each RAN NSST defines the gNB functionalities, and their specific configuration to meet the specific performance requirements of a service type (i.e., eMBB, uRLLC, and mMTC). To identify this service type, the RAN NSST contains the Single Network Slice Selection Assistance Information (S-NSSAI) [8]. This 3GPP parameter consists of two fields: Slice/Service Type (SST) and Slice Differentiator (SD). SST provides a value that identifies the service type of the slice, i.e., SST=1 for eMBB, SST=2 for uRLLC, and SST=3 for mMTC. SD is optional and allows differentiation amongst multiple network slices with the same SST value, e.g., slices for different tenants.

4.2.3 NFV MANO and Descriptors

To manage VNFs, ETSI-NFV has defined the NFV MANO [9]. It comprises:

- Virtual Infrastructure Manager (VIM), which manages the virtual resources from one or NFVI-PoPs.
- Virtual Network Function Manager (VNFM), which manages the VNFs throughout their lifecycle. It is also responsible for their performance and fault management from the virtualization viewpoint.
- Network Function Virtualization Orchestrator (NFVO), which combines PNFs and VNFs to create Network Services (NSs), managing them throughout their lifecycle.

To automate the lifecycle management of NSs and their VNFs and/or PNFs, the NFV-MANO uses the NFV descriptors: Network Service Descriptor (NSD), Virtualized Network Function Descriptor (VNFD) and Physical Network Function Descriptor (PNFD).

Each NSD (and VNFD) defines a set of attributes. Among them, the flavors provide different options to deploy an instance of a NS (and VNF). For example, each flavor might add some extra functionalities to that instance. In turn, each flavor defines one or more Instantiation Levels (ILs), each specifying a different amount of virtual resources for the instance deployed from that flavor. Defining

several ILs enables the adaptation of the required amount of virtual resources to guarantee the performance of an instance of NS (and VNF) supporting traffic fluctuations. For more detailed information about flavors and ILs, see [9].

Finally, since NFV-MANO focuses on virtualization, the PNFDs only contain information required to connect PNFs with VNFs.

4.3 RAN Slice Description Proposal

4.3.1 Harmonizing 3GPP and NFV Descriptors: A Prerequisite for Managing RAN Slices Subnets

To manage the gNBs taking part in each RAN slice subnet, the RAN NSSMF must rely on RAN NSSTs and NFV descriptors.

On the one hand, the RAN NSST focuses on the description of the gNBs of a RAN slice subnet from an application perspective (i.e., information on their functionalities and configuration parameters). The aim of a RAN NSST is to adapt the behavior of the gNBs to meet the requirements of a specific service type (e.g., eMBB). However, the RAN NSST neglects the description of the resources to deploy the virtualized part of these gNBs.

On the other hand, the NFV provides information on the virtual resources that are required to accommodate the spatial and temporal traffic demands of the CU and DUs of a gNBs. This means that NFV descriptors could enable the deployment of the virtualized part of a gNB. However, NFV descriptors are agnostic to the application layer configuration of the CU and DUs.

With the combined use of 3GPP and NFV descriptors, the gNBs of a RAN slice subnet could be deployed and operated. Accordingly, we first analyze the most representative configuration parameters to customize the behavior of a gNB. Then, we propose a description model that harmonizes the scopes of the RAN NSSTs and NFV descriptors to manage the gNBs taking part in different RAN slice subnets. Finally, we explain how the RAN NSSMF and NFMFs interwork with the NFV-MANO to manage RAN slice subnets with the proposed model and configuration parameters.

4.3.2 Configuration Parameters in RAN NSST

According to Table 4.1, the most representative parameters are classified into two groups: 3GPP NR, and network management algorithms.

The 3GPP NR comprises those parameters related to the physical transmission. Among them, the waveform and numerology, the operations bands, the slot format, the 5G Quality of Service Indicators (5QIs), and the Modulation and Coding Schemes (MCSs) are discussed below.

The waveform is based on Orthogonal Frequency Division Multiplexing (OFDM). It consists of several orthogonally-spaced subcarrier with a spacing of $15 \cdot 2^{\mu}$ KHz [10], where μ is the numerology (μ=0, 1, 2, 3 and 4). The higher the numerology is, the shorter the Transmission Time Interval (TTI) is. Decreasing the TTI enables gNBs to transmit UE data faster; and add a margin to increase the number of retransmissions in the hybrid automatic repeat request function. Therefore, shorter TTIs are suitable for RAN slice subnets that require low latency and high reliability. Additionally, high-speed UEs can benefit from shorter TTIs, taking advantage of the time invariant characteristics of the channel.

The NR operation bands includes 450-6000 MHz and 24250-52600 MHz [11]. Each band might accommodate carriers from 5 to 400 MHz. The bandwidth of the selected carrier depends on the required service data rate and the UE density in the geographical regions where the RAN slice subnet is deployed.

The selection of the operation bands also fixes the transmission mode, i.e., Frequency Division Duplex (FDD) or Time Division Duplex (TDD). In case of the TDD mode, there exists predefined slot formats that assign downlink and uplink bits at OFDM symbol level [10]. The selection of the slot format for a given RAN slice subnet depends on the symmetry between its downlink and uplink requirements.

Table 4.1: Relationship between RAN slice subnet requirements and the configuration parameters to customize the behavior of the gNB functionalities for a RAN slice subnet

Configuration parameters to customize the behavior of a RAN slice subnet		**RAN slice subnet requirements**						
		Latency	Maximum Mobility speed	Throughput per UE	UE density	Reliability	Priority Level
3GPP NR	Waveform and numerology	✓	✓				
	Operation bands	✓	✓	✓	✓	✓	
	Slot Format			✓	✓		
	5QIs	✓				✓	✓
	MCS			✓			
Network management algorithms	e.g., Radio Resource Management (RRM); and Self-Organizing Network (SON) techniques	✓		✓	✓	✓	✓

The 5QI specifies the class that ensures a specific Quality of Service (QoS) forwarding behavior in the RAN domain [8]. Each class is mainly characterized by a priority level, a packet delay budget and a packet error rate. Each parameter has a direct impact on the performance of a RAN slice subnet. For example, with the packet delay budget and the packet error rate, the RAN NSSMF can control the latency and reliability level of the RAN slice subnet, respectively. Similarly, with the priority level the RAN NSSMF can weigh the utilization of radio resources shared among the RAN slice subnets, providing in this way multiplexing gains.

The MCS is a modulation scheme and coding rate tuple that provides a given throughput for an UE. Each gNB selects the MCS per each UE based on its current radio conditions. NR defines two set of MCS tuples: one is compatible with the MCSs defined in LTE and the other extends that range to include a higher modulation scheme, thus enabling higher throughput in NR. Each RAN slice subnet should only use the set of MCSs that best meet its throughput requirements.

Network management algorithms are usually proprietary and include vendor-specific parameters. However, some parameters could be configured by the RAN NSSMF, allowing the definition of slice-specific network management algorithms that optimize the operation of each RAN slice subnet. Network management includes traditional RRM functionalities (e.g., packet scheduling) and SON techniques (e.g., mobility robustness optimization).

4.3.3 Description Model to Manage RAN Slice Subnets

Fig. 4.4 shows the proposed description model to define the management of gNBs of several RAN slice subnets. Each RAN NSST references (a) a common NSD that describes the underlying resources of a gNB; and (b) contains the specific configuration parameters for this gNB (i.e., those adapted to a specific SST).

The gNBs of any RAN slice subnet may be stick to the option #2 for CU-DU functional split. However, the option selected for DU-RU split can vary between #7 and #8, depending on the technology used for the underlying fronthaul links, i.e., CPRI or eCPRI. For this reason, the gNB NSD defines three flavors: one supporting only CPRI, other supporting only eCPRI, and the other supporting the joint usage of both technologies in case that they were implemented in a

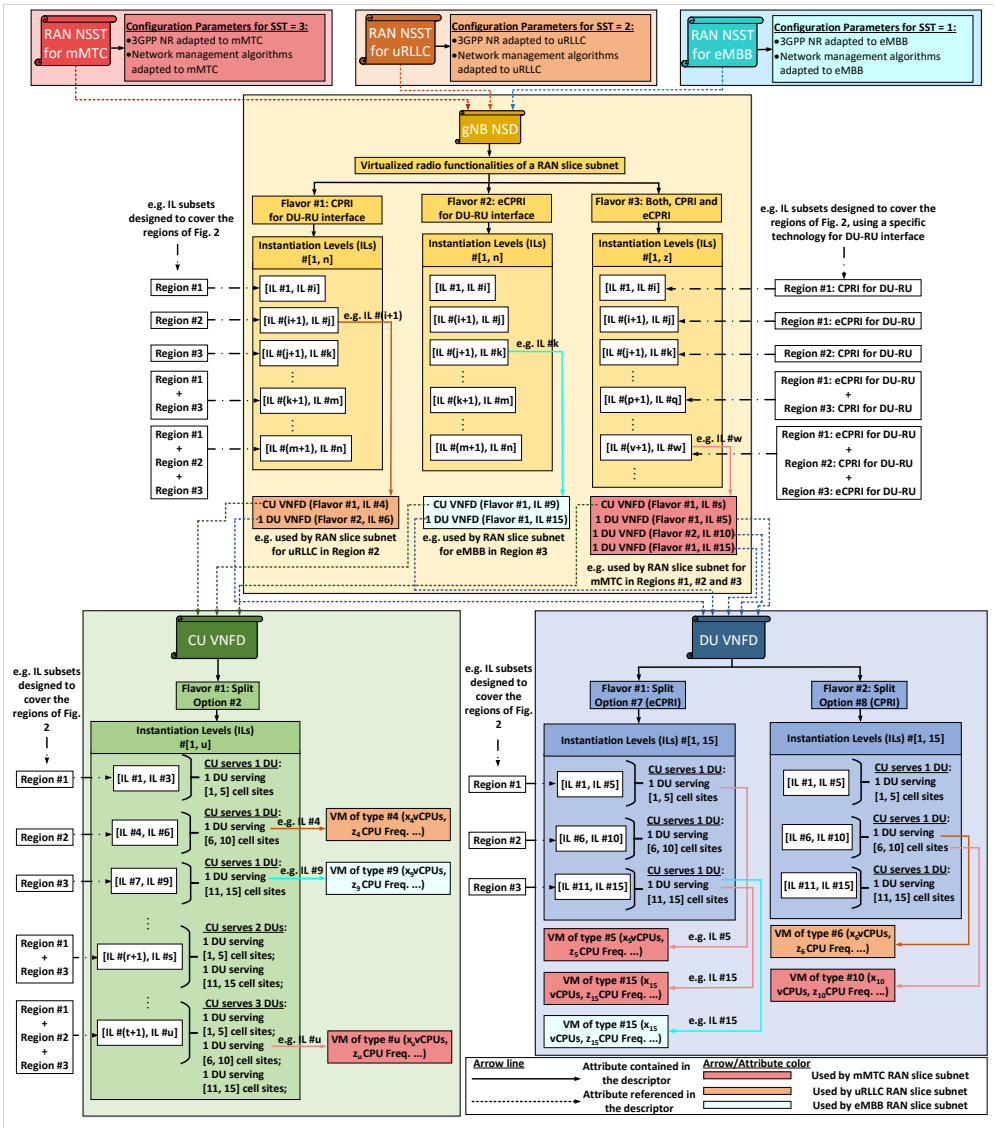


Figure 4.4: Proposed model to define the management of a gNB for each RAN slice subnet. By way of example, the gNBs of the three RAN slice subnets presented in Fig. 4.3 are described. To deploy these gNBs, the RAN NSSMF selects in the gNB NSD the tuples (Flavor #3, IL #w), (Flavor #1, IL #(i+1)) and (Flavor #2, IL #k) for mMTC, uRLLC and eMBB RAN slice subnets, respectively. Note that the mMTC RAN slice subnet requires both, the CPRI and eCPRI for DU-RU interfaces.

specific deployment area.

Each flavor in the gNB NSD defines different subsets of ILs depending on the

type of region(s) (e.g., those represented in Fig. 4.3) covered by a gNB. Since the number of cell sites located in each region is different, each subset gathers those ILs adapted to the possible range of the aggregated traffic demands for a certain number of cell sites located in one or more regions. In the case of using flavor #3, the technology for implementing the DUs-RUs interface also conditions the IL subsets (e.g., two IL subsets for Region #1, one describing CPRI interfaces and the other describing eCPRI interfaces).

For a given subset, each IL defines the number of DUs in the gNB as well as the virtual resources required to deploy these DUs and the CU. The amount of virtual resources is completely different for the CU and DUs. Whereas each DU may serve a number of cell sites in a region, a CU may aggregate the traffic from DUs serving one or more regions, each with different cell sites density. For that reason, the gNB NSD distinguishes between the resources requirements of CU and DU by referencing a different VNFD (with specific flavor and IL) in each case.

Focusing on DU VNFD, it contains two flavors for the specification of the DU-RU functional split. Each flavor enables the split option #7 and #8, respectively. In turn, each flavor defines subsets of ILs, each gathering those ILs adapted to the possible traffic demands from a specific range of number of cell sites served by a DU. Each IL defines the characteristics (i.e., number of cores, CPU frequency, RAM capacities, etc.) of the Virtual Machine (VM) that hosts the DU functionalities. The utilization of this VM mainly depends on aspects such as the amount of radio resources and the MCSs used per each UE [12].

Similarly, the CU VNFD contains one flavor to define the split option #2 for CU-DUs. Each flavor also contains subsets of ILs. However, in this case, these subsets define ILs to support a specific number of DUs since a CU might aggregate DUs from different regions. Depending on the number of DUs served by the CU and their capacities, the characteristic of the VM that host the CU differs between ILs.

Finally, since RUs are fixed in specific locations, the gNB NSD cannot include references to the PNFDs to be reusable in any deployment area. The RAN NSSMF is responsible for selecting the specific PNFDs to define the RUs of a RAN slice subnet.

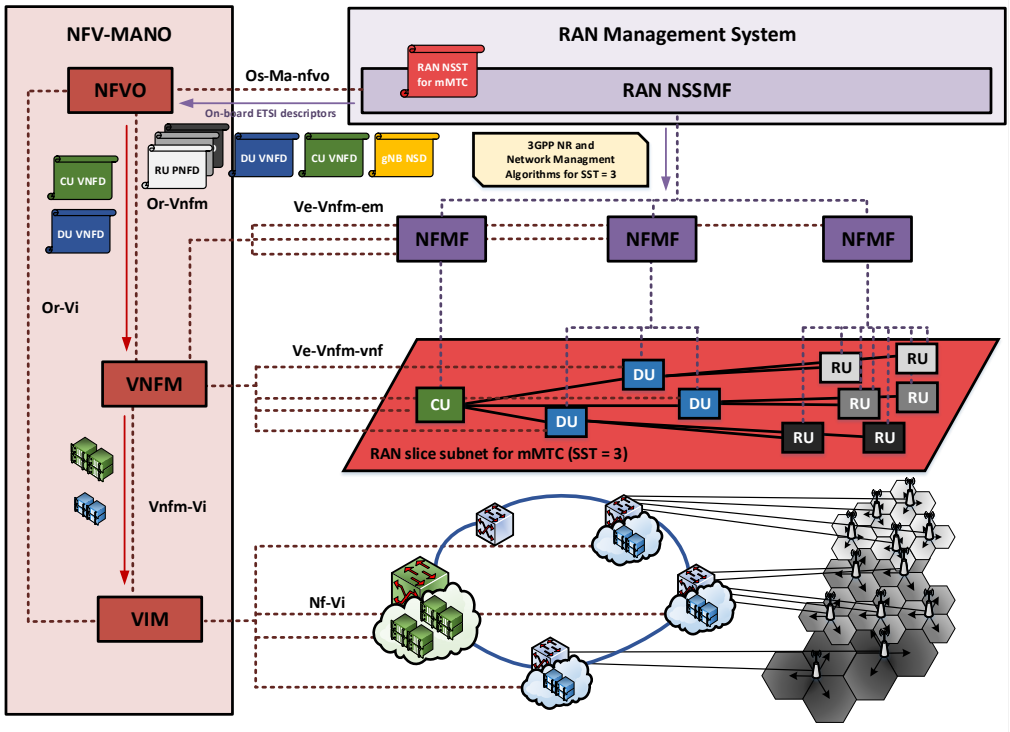


Figure 4.5: RAN slicing management framework. By way of example, this framework manages the deployment and operation of the RAN slice subnet for mMTC (see Figs. 4.3 and 4.4).

4.3.4 RAN NSSMF, NFMFs and NFV-MANO Interworking under a Unified Framework

To manage the gNBs taking part in a RAN slice subnet, there is a necessity to define a unified framework where 3GPP entities (i.e., RAN NSSMF, NFMFs) and ETSI-NFV (NFV-MANO) can work with each other. Examples of tentative integration have been proposed in [2, 3]. In Fig. 4.5, this unified framework is depicted. Each management entity is common for all RAN slice subnets.

When a vertical requests a service for a specific geographical area, the RAN NSSMF first selects a RAN NSST whose SST matches with the requested service. Then, the RAN NSSMF determines which RUs cover the geographical area.

Once the RUs are selected, the RAN NSSMF computes the number of gNBs that will include these RUs. To that end, based on the UE density in each region

of the deployment area, the RAN NSSMF performs the following actions:

- Select the flavor of the gNB NSD according to the technology of fronthaul links (i.e., flavor #1 for CPRI or flavor #2 for eCPRI). If this deployment area comprises fronthaul networks using both technologies, the flavor #3 is selected, since it enables the definition of a DU-RU interface with any of these technologies. For example, the mMTC RAN slice subnet shown in Fig. 4.3 requires the flavor #3 because Regions #1 and #3 usee eCPRI fronthaul links, and Region #2 uses CPRI fronthaul links.

- Compute the number of DUs. To that end, for every region on the deployment area, the RAN NSSMF determines the optimal IL subset for one DU. This IL subset can accommodate the fluctuations of the temporal traffic demands of this area. If this DU cannot serve the entire region, additional DUs with a specific IL subset are included to meet the required capacity in this region. Thereby, the spatial traffic demands of this area are also accommodated.

- Determine the number of CUs that will serve the DUs. Considering the latency constraints due to the physical distance between a CU and the selected DUs, the RAN NSSMF (a) optimally distribute these DUs between a specific number of CUs; and (b) select the optimal IL subset for each CU. The aim of this procedure is minimizing the number of CUs to benefit from the statistical multiplexing gains provided by this centralization approach. The number of CUs is equivalent to the number of gNBs.

- Search across the IL subsets of the gNB NSD, the subset that reference the selected IL subset for each DU and the CU that serves these DUs. Thereby, the RAN NSSMF derives the optimal IL subset for each gNB.

Next, the RAN NSSMF proceeds with the on-boarding of the NFV descriptors along with the selected flavor and the IL subset per each gNB. With this information, the NFVO can instantiate the gNBs and scale them throughout the lifecycle of the RAN slice subnet. The VNFM and VIM also play a key role during the lifecycle through the management of CU/DUs and their underlying virtual resources, respectively. For more information, see [9].

Table 4.2: RAN slice subnet requirements for eMBB [13], mMTC [13] and uRLLC [14]. Note that the geographical regions might be mapped to the ones presented in Fig. 4.3. More precisely, industrial area to Region #1, suburban area to Region #2 and city center to Region #3

RAN Slice Subnet Requirements		**RAN Slice Subnets**		
		eMBB (e.g., UHD streaming)	mMTC (e.g., pollution control)	uRLLC (e.g., remote controlled drones)
Latency (ms)		10	Seconds to hours	5
Maximum Mobility Speed (Km/h)		10	0	250
Throughput per UE	Uplink (Mbps):	50	0.1	25
	Downlink (Mbps):	300	0.1	1
UE density		5.000 UEs/Km^2	500.000 UEs/Km^2	50 UEs/Km^2
Reliability (%)		Not specified	Not specified	99.999
Priority Level		Low	Medium	High
UE type		Pedestrians	Stationary sensors	Remote-controlled vehicles
Geographical regions		City center	City center, industrial area, and suburban area	Suburban area

To customize the behavior of the gNBs, the RAN NSSMF uses the configuration parameters defined in the RAN NSST. With this information, the RAN NSSMF configures the CU and DUs through the specific NFMFs, which apply the parameters provided by the RAN NSSMF.

4.4 Example of RAN Slice Description

To clarify our proposal, this section provides an example where RAN slice subnets for eMBB, uRLLC, and mMTC are deployed over the same city (see details in Table 4.2). These RAN slice subnets can be mapped to the ones presented in Fig. 4.3.

Table 4.3 summarizes the configuration parameters of each RAN NSSTs as well as the information derived by the RAN NSSMF to instantiate the RAN slice subnets. Below, we discuss this information.

4.4.1 eMBB

For eMBB, the RAN NSST defines a numerology of μ=2 to fulfill the latency requirement of 10 ms. Any operation band supports this numerology. However, the adopted carriers should be used with the maximum available bandwidth to support the required high throughput.

Table 4.3: Configuration parameters of each RAN NSST as well as the information derived by the RAN NSSMF. Note 1: Currently NR specifications do not provide any operation band supporting μ=4. Note 2: Flexible OFDM symbols might be used for both, downlink and uplink. Note 3: These levels match with those shown in Fig. 4.4.

Configuration Parameters		**RAN Slice Subnets**		
		RAN NSST for eMBB	**RAN NSST for mMTC**	**RAN NSST for uRLLC**
3GPP NR	Waveform and numerology	μ = 2	μ = 0	μ = 3 Note 1
	Operation bands	450-6000 MHz (max. carrier bandwidth 100 MHz), 24250 to 52600 MHz (max. carrier bandwidth 400 MHz)	450-6000 MHz (carrier bandwidth 5 MHz)	24250 to 52600 MHz
	Slot Format	#28 (12 OFDM symbols for downlink, 1 OFDM symbol for uplink, and 1 flexible OFDM symbol). Note 2	#45 (6 OFDM symbols for downlink, 6 OFDM symbol for uplink, and 2 flexible OFDM symbol). Note 2	#10 (13 OFDM symbol for uplink, and 1 flexible OFDM symbol). Note 2
	5QIs	5QI=80 (default priority level = 66, packet delay budget = 10 ms, packet error rate 10-6)	5QI=4 (default priority level = 50, packet delay budget = 300 ms, packet error rate 10-6)	5QI=81 (default priority level = 11, packet delay budget = 5 ms, packet error rate 10-5)
	MCS	(π/2) BPSK, QPSK, and 16/64/256 QAM	(π/2) BPSK, QPSK and 16/64 QAM	(π/2) BPSK, QPSK, and 16/64 QAM
Network management algorithms	e.g., RRM; and SON techniques	e.g., a dynamic scheduler for guaranteed throughput	e.g., a semi-persistent scheduler	e.g., a dynamic scheduler for guaranteed delay
Information derived by the RAN NSSMF				
RUs covering the deployment area		RUs of the city center	RUs of the entire city	RUs of the suburban area
Number of gNBs		N1	N2	N3
Selected flavor in the gNB NSD		Flavor #2 (eCPRI)	Flavor #3 (CPRI+eCPRI)	Flavor #1 (CPRI)
Subset of ILs from the selected flavor in the gNB NSD		ILs per Region #3, i.e., [IL #(j+1), IL #k]. Note 3	ILs per Region #1 (eCPRI) + Region #2 (CPRI) + Region #3 (eCPRI), i.e., [IL #(v+1), IL #w]. Note 3	ILs per Region #2, i.e., [IL #(i+1), IL #j]. Note 3

Assuming the selection of TDD mode (common for the three use cases), the slot format #28 is set since it allocates the majority of slots for downlink traffic. Concerning QoS classes, the RAN NSST per RAN determines the value #80 because it guarantees a latency lower than 10 ms.

The RAN NSST also specifies the utilization of the extended set of MCSs to provide the highest throughput values (i.e., those obtained from 256 QAM).

Considering the packet scheduling scheme as an example of network management algorithm, the RAN NSST selects a scheme that provides robust and adaptive data transmission. Particularly, the best option is a dynamic scheduler (as opposed to persistent scheduling) which also guarantees the throughput [15].

Finally, the RAN NSSMF selects the flavor #2 and the subset of ILs for Region #3 because they are adapted to the cell site density in a city center, and the fronthaul network of this region implements eCPRI for DU-RU interfaces.

4.4.2 mMTC

For mMTC, the selected numerology is the lowest because the latency is not critical. Additionally, carriers' bandwidth should be the lowest possible, as the

required throughput is low. Due to the small bandwidth, these carriers can only be allocated in the lower operation bands.

With respect to the 5QI, the RAN NSST selects the value #4, because it is the most latency-tolerant while the priority level is not too low.

Regarding the scheduling scheme, the semi-persistent scheduler is the best option since the traffic pattern of the sensors is deterministic, since the information is periodically exchanged with the network [15].

Finally, the RAN NSSMF selects the flavor #3 and the IL subset that considers a CU aggregating DUs serving three different regions over the entire city. Furthermore, this IL subset considers the implementation of eCPRI for the fronthaul networks of Region #1 and Region #3, and CPRI for the fronthaul network of Region #2.

4.4.3 uRLLC

For uRLLC, the RAN NSST selects the highest numerology due to the stringent latency of 5 ms. This numerology forces the utilization of the highest operations bands. The slot format requires a larger amount of slots allocated in the uplink than in the downlink because vehicles continuously collect and send environment information to the remote drivers. Regarding 5QIs, only the #8 guarantees a latency below 5ms.

Finally, the RAN NSSMF selects the flavor #1 and the subset of ILs per Region #2 because they are adapted to the cell site density of the suburban area, and the fronthaul network of this region implements CPRI for DU-RU interfaces.

4.5 Conclusions

RAN slicing enables the provision of different service types over a common wireless network infrastructure. Leveraging the NFV benefits, the CU and DUs of the gNBs for a RAN slice subnet could be customized and adapted to its requirements. Although the RAN NSST considers the gNB functionalities, the 3GPP has not identify which parameters and how they must be customized to provide a RAN slice subnet its expected behavior. Additionally, the RAN NSST neglects the resource requirements for the virtualized deployments of CUs and DUs over a geographical region with fluctuating spatial and temporal traffic demands. With

the aim of enabling the customization and deployment of the gNBs of a RAN slice subnet, we have proposed a description model that harmonizes the 3GPP and ETSI-NFV viewpoints for RAN slicing. The proposed solution benefits from the reusability provided by NFV descriptors to define the underlying resources of the CU and DUs of the gNBs for several RAN slice subnets. To customize the behavior of each RAN slice subnet, we have identified the most representative radio parameters to configure their gNBs. Finally, to facilitate the comprehension of the proposal, an example composed of three RAN slice subnets for eMBB, mMTC and uRLLC scenarios has been provided.

Acknowledgments

This work is partially supported by the Spanish Ministry of Economy and Competitiveness, the European Regional Development Fund (Project TEC2016-76795-C6-4-R), the Spanish Ministry of Education, Culture and Sport (FPU Grant 17/01844), and the University of Granada, Andalusian Regional Government and European Social Fund under Youth Employment Program.

References

[1] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN Slicing for Verticals: Enablers and Challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, 2019.

[2] 3GPP TS 28.533 V.15.1.0, "Management and Orchestration; Architecture Framework (Release 15)," Dec. 2018.

[3] ETSI GS NFV-EVE 012 V3.1.1, "Network Functions Virtualization (NFV); Evolution and Ecosystem; Report on Network Slicing Support with ETSI NFV Architecture Framework," Dec. 2017.

[4] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, 2018.

[5] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 1, pp. 146–172, 2019.

[6] C.-Y. Chang *et al.*, "Slice Orchestration for Multi-Service Disaggregated Ultra-Dense RANs," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 70–77, 2018.

[7] ITU, "GSTR-TN5G - Transport network support of IMT-2020/5G," Feb. 2018.

[8] 3GPP TS 23.501 V.16.0.2, "System Architecture for the 5G System; Stage 2 (Release 16)," Apr. 2019.

[9] O. Adamuz-Hinojosa *et al.*, "Automated Network Service Scaling in NFV: Concepts, Mechanisms and Scaling Workflow," *IEEE Commun. Mag.*, vol. 56, pp. 162–169, July 2018.

[10] 3GPP TS 38.211 V.15.4.0, "NR; Physical channels and modulation (Release 15)," Dec. 2018.

[11] 3GPP TS 38.211 V.15.4.0, "NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (Release 15)," Dec. 2018.

[12] A. Younis, T. X. Tran, and D. Pompili, "Bandwidth and Energy-Aware Resource Allocation for Cloud Radio Access Networks," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 10, pp. 6487–6500, 2018.

[13] Next Generation Mobile Networks (NGMN), "NGMN 5G White Paper." `https://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf`. [Online; Accessed Jan. 2019].

[14] 3GPP TS 22.186 V.16.1.0, "Enhancement of 3GPP support for V2X scenarios; Stage 1 (Release 16)," Dec. 2018.

[15] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 2, pp. 678–700, 2013.

Chapter 5

Paper C: Sharing gNB components in RAN slicing: A perspective from 3GPP/NFV standards

Authors:

Oscar Adamuz-Hinojosa, Pablo Munoz, Pablo Ameigeiras, and Juan Manuel Lopez-Soler.

The paper has been published in the IEEE Conference on Standards for Communications and Networking (CSCN), October, 2019.

Abstract

To implement the next Generation NodeB (gNB) that are present in every Radio Access Network (RAN) slice subnet, Network Function Virtualization (NFV) enables the deployment of some of the gNB components as Virtualized Network Functions (VNFs). Deploying individual VNF instances for these components could guarantee the customization of each RAN slice subnet. However, due to the multiplicity of VNFs, the required amount of virtual resources will be greater compared to the case where a single VNF instance carries the aggregated traffic of all the RAN slice subnets. Sharing gNB components between RAN slice subnets could optimize the trade-off between customization, isolation and resource utilization. In this article, we shed light on the key aspects in the Third Generation Partnership Project (3GPP)/NFV standards for sharing gNB components. First, we identify four possible scenarios for sharing gNB components. Then, we analyze the impact of sharing on the customization level of each RAN slice subnet. Later, we determine the main factors that enable isolation between RAN slice subnets. Finally, we propose a 3GPP/NFV-based description model to define the lifecycle management of shared gNB components.

5.1 Introduction

In the next years vertical industries may bring a wide variety of services with diverging requirements in terms of functionality and performance [1]. To economically provide them over a common wireless network infrastructure, Radio Access Network (RAN) slicing has emerged as a solution [2]. It consists of the provision of multiple RAN slice subnets, each adapted to the requirements of a specific service. To that end, RAN slicing could relies on Network Function Virtualization (NFV). This technology enables the customization of the next Generation NodeBs (gNBs) present in every RAN slice subnet through their implementation as Virtualized Network Functions (VNFs).

The Third Generation Partnership Project (3GPP) is playing a significant role on RAN slicing standardization. It has defined the gNB components, i.e. the Centralized Unit (CU), the Distributed Units (DUs), and the Radio Units (RUs); and the Fifth Generation (5G) radio protocol stack [3]. It has also specified the

management entities and their mechanisms to handle the lifecycle of RAN slice subnets [4]. However, these contributions are not enough to provide RAN slice subnets because the management of those gNB components implemented as VNFs goes beyond the 3GPP scope. The leading standardization body on network virtualization is the European Telecommunications Standards Institute (ETSI), specifically the NFV group, which has defined the management framework and its mechanisms to handle the lifecycle of VNFs [5].

Based on the 3GPP and ETSI-NFV contributions, several research projects are developing specific solutions for RAN slicing [6]. The majority of these solutions assume a single VNF instance to accommodate a gNB component of a specific RAN slice subnet. This approach guarantees the customization of each RAN slice subnet, however the resource utilization can be inefficient. For example, let us assume an individual gNB component for each RAN slice subnet and a fixed resource capacity per VNF instance. Then, if the required capacity of two or more RAN slice subnets fits into one VNF instance, sharing a VNF instance will be a more efficient solution than deploying separate VNF instances.

Sharing VNF instances could involve statistical multiplexing gains on the utilization of virtual resources. However, achieving the customization level required by each RAN slice subnet is a challenge. Some research projects such as 5G-PICTURE [7], SliceNet [8] or 5G-MoNArch [9] has pursued a trade-off solution between customization and resource utilization. Notwithstanding, these projects have analyzed neither the impacts of sharing gNB components on the customization of each RAN slice subnet, nor the main factors that enable the isolation between RAN slice subnets. Additionally, these projects has focused on the lifecycle management from the 3GPP viewpoint, neglecting the NFV perspective.

In this paper, we shed light on the key aspects in 3GPP/NFV standards for sharing gNB components between RAN slice subnets, typically, enhanced Mobile Broadband (eMBB), ultra-Reliable Low Latency Communication (uRLLC), and massive Machine Type Communication (mMTC). To that end, we (a) identify the main scenarios for sharing gNB components; (b) analyze the impact of sharing on the customization level of each RAN slice subnet; (c) determine the main factors that enable the isolation between RAN slice subnets; and (d) propose a description model to define the lifecycle management of a shared gNB component using the 3GPP/NFV management templates.

The remainder of this article is as follow. Section 5.2 overviews the 3GPP/NFV standardization for RAN slicing. Section 5.3 analyzes the key aspects and enablers for sharing gNB components. Section 5.4 provides a 3GPP/NFV-based description model to define the lifecycle management of a shared gNB component. Finally, Section 5.5 draws the main conclusions of this work.

5.2 3GPP/NFV Standardization for RAN Slicing

5.2.1 3GPP Next Generation RAN Architecture

The 3GPP Next Generation Radio Access Network (NG-RAN) is composed of gNBs, which provide wireless connectivity to the User Equipments (UEs) through the New Radio (NR) protocol stack [3]. From a functional viewpoint, a gNB is composed of RUs, DUs and a CU. The functionalities of the NR protocol stack are distributed over these components in a flexible way. The RUs comprise at least the antennas and the radio-frequency circuitry, thus they must be implemented as hardware. The remaining functionalities might be virtualized and they are split into multiple DUs and one CU. The DUs contain the low-layer functionalities whereas the CU includes the high-layer functionalities. The aim of this split is to leverage the benefits of virtualization and centralization. Additionally, the CU could be split into two entities, each gathering the control and data plane functionalities, respectively[1]. For simplicity, this article does not assume control and data plane split.

There exists eight options to split the gNB functionalities as Fig. 5.1 shows. However, there is a consensus in the industry and academia that feasible implementations in the short-term are option #2 for CU-DU, and option #7 for DU-RU [11].

From the infrastructure perspective, the partial virtualization of gNBs requires the existence of Network Function Virtualization Infrastructure Point of Presences (NFVI-PoPs) in addition to cell sites. A NFVI-PoP is a cloud site that hosts the virtual resources to accommodate VNF instances. These NFVI-PoPs, classified as aggregation and edge NFVI-PoPs, connect the cell sites with the core network through a hierarchical approach [12].

[1]The benefits and drawbacks of this approach could be consulted in [10]

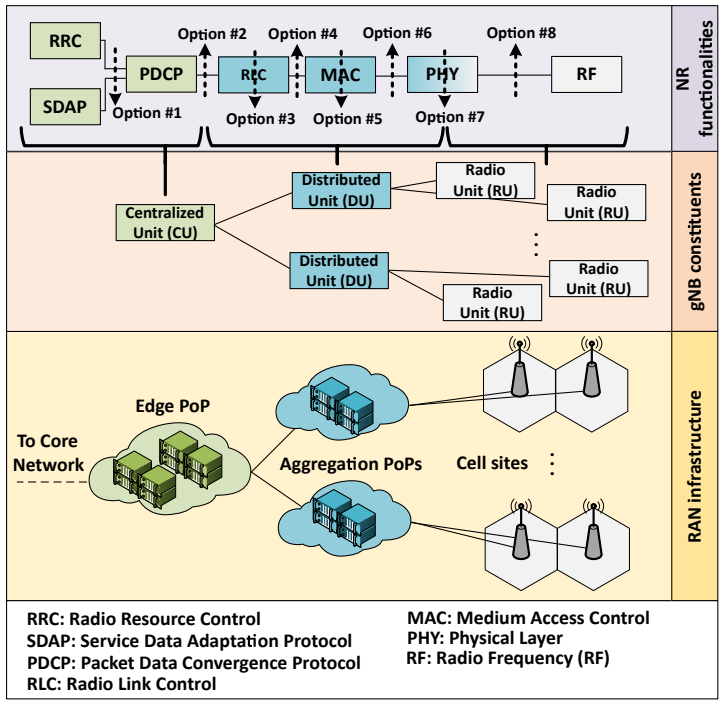


Figure 5.1: glsNG-RAN architecture. For comprehensibility purposes, we assume the CU and the DUs are virtualized.

5.2.2 Enabling RAN Slicing in the NG-RAN Architecture

To slice the NG-RAN architecture into multiple RAN slice subnets, the CU and the DUs should be individually deployed for each RAN slice subnet. Thereby, the gNB functionalities could be customized to meet the specific requirements of each RAN slice subnet.

Regarding its functionalities, a gNB component comprises not only the NR functionalities depicted in Fig. 5.1, but also procedures for Radio Resource Management (RRM) [13]. Those RRM procedures that are time-sensitive, e.g., Packet Scheduling (PS), Link Adaptation (LA), etc, are hosted in the DU. The remaining RRM procedures, e.g.,Mobility Management (MM), Radio Admission Control (RAC), etc, are hosted in the CU to leverage the benefits of centralization.

Each RRM procedure is controlled by a vendor-specific algorithm that guarantees the performance requirements of each UE while the available radio resources are efficiently used. To consider RAN slicing in each RRM procedure, is rea-

sonable to implement a two-level algorithm: inter-slice and intra-slice [13]. At inter-slice level, a RRM algorithm copes with the management of all the RAN slice subnets considering the available radio resources on the whole RAN infrastructure. At intra-slice level, this algorithm is specific for a RAN slice subnet and it is designed to meet its requirements. Furthermore it only considers the allocated radio resources for this RAN slice subnet.

To have a complete picture of the RAN infrastructure, the implementation of RRM algorithms at inter-slice level could be hosted outside the gNB. Additionally, the frequency work of the RRM algorithms (i.e., number of times that they are executed in a period of time) at inter-slice and intra-slice levels cannot be the same. The algorithm at inter-slice level deals with a enormous amount of information. This fact hinders its coordination with the intra-slice implementation of the RRM algorithm, specially if the last is time-sensitive [13].

Focusing on the RRM algorithms at intra-slice level, their decisions are individually applied to each Data Radio Bearer (DRB). For example, the RAC could accept the establishment request for a DRB whose user data require a specific throughput and latency. To apply the RRM decisions, the Radio Resource Control (RRC) layer configures the remaining layers to provide a specific treatment of the user data at each DRB [14]. The configured layers are: Service Data Adaptation Protocol (SDAP), Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), Medium Access Control (MAC) and Physical Layer (PHY).

5.2.2.1 SDAP

This layer is responsible for mapping the traffic flows received from the core network to DRBs with a specific Quality of Service (QoS), thus the configuration of these DRBs is key to provide the UEs of a RAN slice subnet a service with a given QoS.

5.2.2.2 PDCP

This layer applies to each DRB functionalities such as ciphering, robust header compression, or packet duplication. The first two introduce a considerable latency, thus they could be disabled to the RAN slice subnets for uRLLC since they require low latency [15]. On the contrary, packet duplication is recommend-

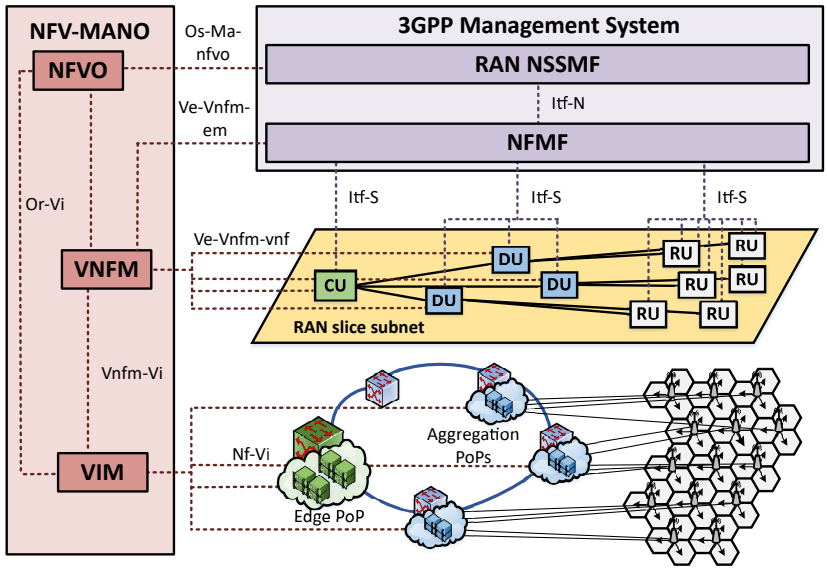


Figure 5.2: 3GPP/NFV-based framework for RAN slicing management.

able for uRLLC RAN slice subnets because they require a high reliability.

5.2.2.3 RLC

This layer comprises functionalities such as segmentation and transfer mode. Concerning segmentation, only eMBB RAN slice subnets will require it to split packets on smaller units because they deal with large payloads. For transmission of user plane data, Acknowledge Mode (AM) is appropriate for uRLLC RAN slice subnets. For those where reliability is not a critical requirement, Unackowledge Mode (UM) is a better option [15].

5.2.2.4 MAC

This layer contains features such as the Hybrid Automatic Repeat Request (HARQ) or the slot format. The HARQ could be specifically configured to optimize the performance of a RAN slice subnet such as the spectral efficiency for eMBB, the coverage for mMTC or the round-trip time for uRLLC.

For Time Division Duplex (TDD) operation, the slot format could be adapted to balance the number of Orthogonal Frequency Division Multiplexing (OFDM)

symbols between the downlink and uplink as a function of the traffic symmetry in each RAN slice subnet (e.g., more downlink bits for eMBB RAN slice subnets).

Additionally, the MAC layer comprises other RRM procedures (e.g., PS, LA, etc). Focusing on PS, the algorithm and the optimization criteria could be could be adapted to optimally distribute the radio resources between the UEs attached to a specific RAN slice subnet [16]. Some examples: semi-persistent planing is better for transmitting periodic information of mMTC services; or optimization criteria such as guaranteeing latency and throughput are appropriate for uRLLC and eMBB services, respectively.

5.2.2.5 PHY

This layer is responsible for aspects such as the numerology or the Modulation and Coding Scheme (MCS). Each RAN slice subnet might require a different numerology. For example, uRLLC RAN slice subnets can benefit from higher numerologies to transmit data with lower latency due to a shorter transmission time interval. To enable a single carrier to support several numerologies, the bandwidth is divided into a set of bandwidth parts, each defining a specific numerology [17]. Thereby, if several RAN slice subnets require different numerologies, they could share the same carrier but using different bandwidth parts. In case of using the same numerology, they could also share the radio resources within a bandwidth part.

5.2.3 Management of RAN Slice Subnets

A RAN slice subnet comprises gNBs that are configured to provide the required behavior. In turn, the components of each gNB could be implemented as VNFs or Physical Network Functions (PNFs), i.e., dedicated hardware.

To manage the lifecycle of RAN slice subnets, the 3GPP and ETSI-NFV have proposed in [4, 5] a RAN slicing management framework as depicted in Fig. 5.2. This management framework requires the interoperation of the NFV-Management and Orchestration (MANO) and the 3GPP management system. The NFV-MANO comprises three functional blocks: Virtual Infrastructure Manager (VIM), Virtual Network Function Manager (VNFM), and Network Function Virtualization Orchestrator (NFVO) [18]. With these functional blocks, the RAN

slicing management framework could only perform tasks related the virtualization of some gNB components, thus NFV-MANO is not enough to manage RAN slice subnets. Specifically NFV-MANO cannot (a) translate the performance and functional requirements of a gNB into the amount of the virtual resources that accommodate the gNB components; and (b) manage the Fault, Configuration, Accounting, Performance, and Security (FCAPS) of the gNB components from the application perspective. These tasks are performed by the network slice subnet management service provider, and the network function service provider, both belonging to the 3GPP management system. For simplicity, we denote these entities as RAN Network Slice Subnet Management Function (NSSMF) and Network Function Management Function (NFMF), respectively (see the example of service management providers in section A.4 of [19]). The RAN NSSMF performs tasks (a) and (b) while the NFMF is controlled by the RAN NSSMF to carry out the activities related to (b) in the gNB components. Since the RAN NSSMF is in charge of configuring the gNB components, it could host the inter-slice implementation of those RRM algorithms that are non-time sensitive. In such case, the RAN NSSMF would transfer the inter-slice decisions to the intra-slice algorithm implementations hosted in each gNB component.

To automate the lifecycle management of RAN slice subnets, the RAN slicing management framework relies on a set of predefined templates. The main template is the RAN Network Slice Subnet Template (NSST), proposed by the 3GPP. It could define the gNB components of a RAN slice subnet; and the parameters for its FCAPS management at application level [12]. Since some of the gNB components might be virtualized, the RAN NSST must reference to the NFV management templates to describe the lifecycle of the VNFs that host them [12]. In Section 5.4, we shed light on the utilization of the 3GPP/NFV management templates.

5.3 Analysis of Key Aspects and Enablers for Sharing gNB Components

5.3.1 Main Scenarios for Sharing gNB Components: Enabling Customization

As depicted in Fig. 5.3, there are four main scenarios for sharing the gNB components between several RAN slice subnets. For all scenarios, we assume the CU and the DUs to be implemented as VNFs, and the RUs as PNFs. Below, we discuss each scenario focusing on the CU and DUs.

- **Scenario #1. The CU/DUs are specific to each RAN slice subnet**: This scenario enables the full customization of each RAN slice subnet because the RRM algorithms at intra-slice level could be specifically implemented to meet their performance requirements (e.g., an intra-slice implementation of PS that guarantees per-UE throughput). This fact involves that, based on the decision of the intra-slice RRM algorithms in a RAN slice subnet, the RRC layer can specifically configure the DRB treatment along the entire NR protocol stack. This scenario is the easiest for earlier implementations since the gNB components are slice-agnostic.

  Despite its benefits in terms of customization, this scenario is the least efficient in terms of resource utilization since it presents specific VNF instances to implement the CU/DUs of each RAN slice subnet. It might involve the underuse of the virtual resources available in the edge/aggregation NFVI-PoPs. For example, let us assume a fixed resource capacity per DU instance, e.g., one virtualized CPU (vCPU), each belonging to a different RAN slice subnet. If the sum of the resource consumption of two DU instances (e.g., 65 % and 15 % of vCPU utilization, respectively) is less than the resource capacity of a single DU instance (e.g., 80 % $<$ 100 %), two vCPUs will be used when only one is required.

  A limitation of this scenario is that the isolation between RAN slice subnets might even not be guaranteed in spite of presenting separate VNF instances for their CU/DUs. For example, implementing the VNF instances through

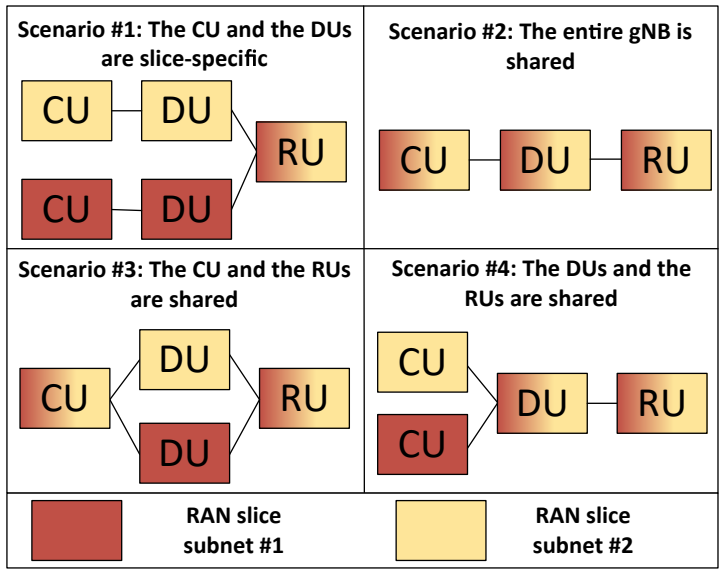


Figure 5.3: Main scenarios for sharing the components of a gNB between several RAN slice subnets. We assume that RUs are shared in each scenario

Virtual Machines (VMs)[2] hinders their isolation due to the virtualization of the Network Interface Cards (NICs) as part of the infrastructure located in an edge/aggregation NFVI-PoP. Used to interconnect VMs, the virtual Network Interface Card (vNIC) of a VM could negatively affect the transmission performance on the vNICs of other VMs [20]. This is due to the fact that increasing the number of VNF instances, and thus the number of vNICs, elevates the interrupt requests and context switching time between the VMs and the hipervisor (i.e., the software, firmware or hardware that creates the VMs). This means that if the number of CU/DUs hosted in an edge/aggregation NFVI-PoP is high, the RAN slice subnets that comprise these CU/DUs could suffer performance degradation even without sharing spectrum.

- **Scenario #2. The entire gNB is shared between RAN slice subnets**: Unlike scenario #1, this scenario is the most efficient in terms of resource utilization since it presents less VNF instances to accommodate

[2]Note that in this article the term VM refers to a virtualization technology in general (i.e., KVM, Linux containers, dockers, etc).

the traffic demands of all the RAN slice subnets. Thereby, the utilization of the resource capacity on the aggregation/edge NFVI-PoPs could be optimized. For instance, assuming the same scenario as the example used in Scenario #1, a shared VNF instance will only require a single vCPU to process the traffic of two RAN slice subnets (i.e., a DU consuming the 80 % of one vCPU).

In this scenario, the vNICs are shared between RAN slice subnets. This means that the user data of each RAN slice subnet transverse the same vNICs, thus the average waiting time of a packet in the vNIC buffer increases. Despite this isolation problem, the main waiting time could be reduced for higher priority packets by controlling some radio parameters in the shared gNB constituents (see Section 5.3.2 for more details). Additionally, this scenario presents a reduced number of VNF instances, and thus the number of vNICs, involving a decrease of the interrupt requests and context switching time between these instances and the hypervisor. Thereby, the transmission performance on the vNICs is not as negatively affected as in scenario #1.

From a functional perspective, the customization level of each RAN slice subnet could be constrained in this scenario. If the RRM algorithms at intra-slice level are shared between RAN slice subnets, the configuration of the DRB treatment in the shared NR protocol stack could not be independently adapted according to the specific requirements of each RAN slice subnet.

The solution to customize the behavior of each RAN slice subnet is making slice-aware (a) the RRM algorithms at intra-slice level and (b) the RRC layer. This means providing them the intelligence to identify the association between a RAN slice subnet and a DRB in order to specifically configure its treatment along the remaining NR protocol layers.

A key element to make slice-aware the RRM algorithms and the RRC layer is the Single Network Slice Selection Assistance Information (S-NSSAI)[21]. Defined by the 3GPP, this parameter classifies a network slice in one of the three main service types (i.e., eMBB, mMTC, and uRLLC). Optionally, the S-NSSAI can define a specific subtype within eMBB, mMTC, or uRLLC

(e.g., for a specific vertical use case). This parameter is used for associating a Protocol Data Unit (PDU) session with a network slice. Since a PDU session comprises the QoS flows mapped to the DRBs of a specific RAN slice subnet, the slice-aware RRM algorithms (and RRC layer) could identify the association between a RAN slice subnet and a DRB through the S-NSSAI. Thereby, at intra-slice level, both the RRM algorithms and the RRC configuration can be adapted for each RAN slice subnet.

Despite the evident utility of the S-NSSAI, making slice-aware the RRM algorithms at intra-slice level (and the RRC layer) involves a higher complexity in their designs. Furthermore, the execution of these algorithms could be more costly in terms of computational requirements. This fact could degrade the performance of the time-sensitive RRM procedures located in the shared DU, thus decreasing the QoS provided by each RAN slice subnet.

Even though the RRM algorithms at intra-slice level (and RRC layer) were not slice-aware, sharing the CU/DUs could be useful for the UE attachment to several RAN slice subnets. In this use case, each RAN slice subnet would comprise two sets of gNBs, each implementing scenario #1 and #2, respectively. Those gNBs implementing scenario #2 would only process signalling messages for attaching the UEs to each RAN slice subnet. After UE attachment, the user data of each RAN slice subnet would be processed by those gNBs implementing scenario #1.

- **Scenario #3. The CU/RUs are shared between RAN slice subnets**: This scenario is not as efficient as scenario #2 in terms of resource utilization because less VNF instances are shared between RAN slice subnets. However, it could present a higher level of customization since the RRM algorithms at intra-slice level located in the DUs can be specifically implemented for each RAN slice subnets.

  The main drawback of this scenario is that the RRM algorithms at intra-slice level in the shared CU and the RRC layer must be slice-aware. However, the complexity of making slice-aware the RRM procedures at intra-slice level in the CU and the RRC layer is not as high as in the scenario #2 because they are not as time-sensitive as the RRM procedures located

in the DU.

Assuming slice-awareness, the shared CU should identify the DU that process the traffic of a specific RAN slice subnet. To that end, the CU could use a matching table that maps the S-NSSAI of each RAN slice subnet with the identifier of the corresponding slice-specific DU (i.e., the DU ID [3]). Thereby, the user data associated to a DRB (specific for a RAN slice subnet) could be delivered to the correspond DU after its processing by the shared CU.

- **Scenario #4. The DUs/RUs are shared between RAN slice subnets**: This scenario is more efficient than the scenario #3 in terms of resource utilization but not as efficient as the scenario #2. The reason of that is that the number of DUs is higher than the number of CUs, thus the number of VNF instances can be considerably reduced in scenario #4.

  Regarding the level of customization, unlike the scenario #3, the intra-slice RRM algorithms in the CU and the RRC layer are specific to each RAN slice subnet, thus their DRB treatment along the entire NR protocol can be adapted to meet their requirements. The main issue of this scenario is that the intra-slice RRM algorithms located in the DU must be slice-aware. Another drawback of this scenario is that the complexity of making slice-aware the RRM procedures is higher than the scenario #3 because these algorithms in the DU are time-sensitive.

  Assuming slice-awareness, the shared DU should identify the source/target CU for the user data associated to each DRB (specific for each RAN slice subnet) to properly process them. To that end, the DU could use a matching table that maps the identifier of each CU (i.e., CU ID) with a S-NSSAI. Thereby, extracting the CU ID of the received user data, the DU could identify the RAN slice subnet that these data belong to, and apply them a specific processing. Note that the CU ID is not currently defined by the 3GPP because assuming non-sharing scenario involves a unique CU per gNB, thus the gNB ID is enough. To enable sharing scenarios, the 3GPP should define the CU ID in future specifications.

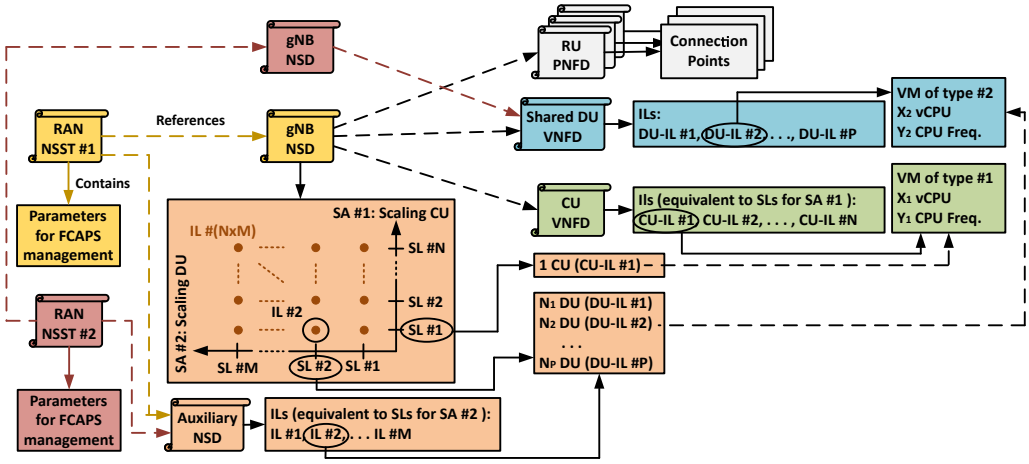


Figure 5.4: Proposed model to describe shared DU instances using 3GPP/NFV management templates. Note that the model for sharing CU instances will be the same except for: (a) the CU Virtualized Network Function Descriptor (VNFD) is shared instead of the DU VNFD; (b) the Instantiation Levels (ILs) of the Auxiliary Network Service Descriptor (NSD) would be equivalent to the Scale Levels (SLs) for Scaling Aspect (SA) #1; and (c) these ILs would reference to the CU-ILs. To avoid redundancy information, the specific CU VNFD and RU Physical Network Function Descriptors (PNFDs) for RAN slice subnet #2 are not shown

5.3.2 Sharing Virtualized gNB Components: Enabling Isolation

To leverage the benefits of sharing gNB components, the isolation between RAN slice subnets must be guaranteed (i.e., if one RAN slice subnet suffers performance degradation, the performance of others RAN slice subnets must remain unaltered). This means the cumulative resource consumption of all RAN slice subnets in the VNF instances that accommodate the shared gNB components cannot exceed their resource capacity. Consequently, the processing layers in NR and operations that significantly impact the resource consumption must be controlled for each RAN slice subnet.

Several works such as [22, 23] have modeled the vCPU consumption by the PHY layer of a virtualized LTE evolved NodeB (eNB) implemented with Open Air Interface [24]. Although they do not consider a 5G gNB, their contributions qualitatively identify the main factors that increase the vCPU consumption of a DU. These factors are: (a) a higher modulation order, which exponentially increases the CPU utilization; (b) a higher number of Physical Resource Blocks

(PRBs), which linearly increases the CPU utilization by an offset; and (c) a higher code rate (i.e., less redundant bits), which linearly increases the CPU utilization.

The average MCSs assigned to the scheduled UEs in each RAN slice subnet must be considered by the RAN NSSMF to estimate the number of PRBs allocated for each RAN slice subnet in coarse time scales[3]. If the RAN NSSMF had a model of the vCPU consumption in function of the MCS and the PRBs, it could control the vCPU consumption in the shared DU instances by each RAN slice subnet. Using this model, the RAN NSSMF could implement mechanisms (e.g., for admission control) to avoid that RAN slice subnets exceed a given percentage of vCPU consumption for each shared DU instance, guaranteeing in this way their isolation at computing resource level.

Although the MCS and the number of PRBs are parameters controlled in the DU, they also impacts the vCPU consumption of a CU because the amount of user data processed by this gNB component directly depends on these parameters [13]. However, at the moment of writing this paper there are no models for the vCPU consumption in the CU because the works in the literature have not considered the CU/DU split and they have focused on the PHY layer (the most vCPU consuming).

In addition to the vCPU consumption, the number of PRBs assigned to each RAN slice also impacts the performance of the vNICs used by the CU/DUs. This is due to more (less) PRBs involves more (less) user data processed by these vNICs. In the case that one of these gNB components is shared, the amount of processed user data by a shared vNIC depends of the PRBs assigned to each RAN slice subnet. By using a model that relates the number of PRBs with the mean waiting time of packets in the vNIC buffer, the RAN NSSMF could also control the PRB assignment for each RAN slice subnet in coarse time scales to avoid excessive buffer delays.

Models for the vCPU consumption and the waiting time on a vNIC buffer should be proposed in future researches to guarantee isolation between RAN slice subnets when a gNB component is shared.

[3]This is distinct from PRB scheduling, which allocates the PRBs assigned to each RAN slice subnet to their UEs

5.4 3GPP/NFV-based Description Model to Manage the Lifecycle of a Shared gNB Component

In [12], we proposed a model to describe the lifecycle management of the gNBs for several RAN slice subnets using the 3GPP/NFV management templates. Despite this model enables the customization of the gNBs and their adaptation to the temporal and spatial traffic demands of each RAN slice subnet, it assumes that the entire gNB is slice-specific (i.e., scenario #1). In this work, we go a step further by proposing a description model that considers gNB sharing.

Hereinafter, we discuss those aspects of the proposed description model that enables the sharing of a gNB component. For more detailed information about other aspects (i.e., regardless sharing), see [12]. For clarity, we focus on scenario #4. Notwithstanding, the proposed model can be easily adapted for scenarios #2 and #3.

Fig. 5.4 shows the proposed model. The 3GPP/NFV management templates are hierarchically structured. On the left are the RAN NSSTs, each defining the FCAPS parameters and the gNBs of a RAN slice subnet. While the FCAPS parameters are included in the RAN NSST, each gNB is described in a gNB NSD that is referenced by the RAN NSST. Note that each RAN NSST references a specific gNB NSD. Managed by the NFVO, a gNB NSD contains a set of attributes to define the lifecycle management of the entire gNB. In this paper, we focus on the SLs, SAs, and ILs since they are key for the instantiation and scaling operations [25]. Each SL defines the number of CU (shared DU) instances and their resource capacity to guarantee the performance of the RAN slice subnet, given a specific traffic demand on a particular geographical area (e.g., a cellular infrastructure with 20 RUs). In turn, the SLs are grouped into two SAs. Each SA defines an independent scaling for the CU (or the shared DU) instances. To ease the management of the SLs of both SAs, the gNB NSD defines ILs. Each IL is the combination of two SLs, one per each SA (e.g, the IL #2 is the combination of the SL #1 for SA #1 and SL #2 for SA #2).

To define the underlying virtual resources of the CU (shared DU) instances in each SL, the gNB NSD must reference a VNFD per each gNB component. Managed by the VNFM, a CU (shared DU) VNFD defines SLs, SAs, and ILs in a similar way as the gNB NSD. The main difference lies in the fact that

these attributes directly define the VMs and their capabilities (i.e., number of vCPUs, CPU freq., etc) to accommodate the CU (shared DU) instances. In this description model, since each CU (shared DU) instance is mapped to a single VM, only one SA is required, thus the SLs and the ILs are used interchangeably in the CU (shared DU) VNFDs. Note that the shared DU VNFD is referenced by the gNB NSD of each RAN slice subnet while the CU VNFD is specific for each one.

Lastly, in addition to CU/DU VNFDs, the gNB NSD also references RU PNFDs. Managed by the NFVO, each RU PNFD defines the physical connectivity points of a single RU.

Deepening on the SAs of the gNB NSD, the SLs of SA #1 define one CU instance whose resource capacity is described in the CU-ILs. Since a gNB has a unique CU, the SLs of SA #1 are equivalent to the CU-ILs. Regarding SA #2, each SL defines the required number of shared DU instances per each DU-IL (i.e., a VM with fixed capabilities).

When the DU instances are shared between several RAN slice subnets, the vCPU consumption in a DU instance might not affect to the majority of slice-specific CU instances. For example, if this increase is due to the user data of a single RAN slice subnet, only the vCPU consumption of one CU instance also increases. In this case, if a single SA was used in the gNB NSD, the design of the SLs would be more complex. Specifically, this design should consider (a) the number of RAN slice subnets that could share each DU instance; (b) their traffic demands; and (c) all the possible combinations for correlating the traffic demands on a shared DU and the slice-specific CUs.

The design complexity of using a single SA can be easily reduced if the shared DUs and the slice-specific CU are scaled independently. For that reason, two SAs have been defined in the gNB NSD, each for scaling independently the slice-specific CU of each RAN slice subnet, and the shared DUs.

Defining two SAs is required but insufficient to scale shared DU instances. Since the gNB instances (i.e., including CU and DU instances) cannot be shared for all the RAN slice subnets, gNB instances per each RAN slice subnet should reference the same shared DU instances. In this case, if a shared DU instance needs to scale, multiple scaling operations should be triggered, one per gNB instance (and per RAN slice subnet). Furthermore, these scaling operations should

be coordinated to select the same SL of SA #2 to scale the DU. To avoid this scaling complexity, ETSI-NFV suggests the definition of an Auxiliary NSD [26]. This management template only defines ILs which coincide with the SLs for SA #2. With this approach, when a shared DU needs to scale, the scaling operation is only executed in the auxiliary Network Services (NSs). After finishing this operation, the IL of each gNB instance must be updated according to the IL of the auxiliary NSs. To that end, the SL of SA #2 is equivalent to the new IL in the auxiliary netowrk service, and the SL of SA #1 is changed according to the specific traffic demand of the RAN slice subnet that the CU belongs.

5.5 Conclusions

In this article, we shed light on the key aspects for sharing gNB components between RAN slice subnets. If the RRM algorithms at intra-slice level and the RRC layer are slice-specific or slice-aware, the gNB components could be shared because the treatment of the NR functionalities for the DRBs of each RAN slice subnet could be specifically configured. We have also identified that controlling the number of PRBs allocated to each RAN slice subnet and the MCSs assigned to their UEs, the isolation between RAN slice subnets can be guaranteed in a gNB component implemented as VNF. Finally, we have proposed a description model to define the lifecycle management of shared gNB components using the 3GPP/NFV management templates.

Ackowledgment

This work is partially supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (Project TEC2016-76795-C6-4-R), the Spanish Ministry of Education, Culture and Sport (FPU Grant 17/01844) and the Andalusian Knowledge Agency (project A-TIC-241-UGR18).

References

[1] J. Ordonez-Lucena *et al.*, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," *IEEE Commun. Mag.*, vol. 55, pp. 80–87, May 2017.

[2] S. E. Elayoubi *et al.*, "5G RAN Slicing for Verticals: Enablers and Challenges," *IEEE Commun. Mag.*, vol. 57, pp. 28–34, Jan. 2019.

[3] 3GPP TS 38.401 V.15.5.0, "NG-RAN; Architecture description (Release 15)," Mar. 2019.

[4] 3GPP TS 28.533 V.15.1.0, "Management and Orchestration; Architecture Framework (Release 15)," Dec. 2018.

[5] ETSI GS NFV-EVE 012 V3.1.1, "Network Functions Virtualization (NFV); Evolution and Ecosystem; Report on Network Slicing Support with ETSI NFV Architecture Framework," Dec. 2017.

[6] 5G-PPP, "Second Wave of Research & Innovation Projects," Nov. 2017.

[7] 5G-PICTURE D2.2, "System architecture and preliminary evaluations," May 2018.

[8] SliceNet D2.2, "Overall Architecture and Interfaces Definition," Jan. 2018.

[9] 5G-MoNArch D4.1, "Architecture and mechanism for resource elasticity provisioning," June 2018.

[10] 3GPP TR 28.531 V.15.0.0, "Study of separation of nr control plane (cp) and user plane (up) for split option 2; (release 15)," Dec. 2017.

[11] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Commun. Surveys Tuts.*, vol. 21, pp. 146–172, Firstquarter 2019.

[12] O. Adamuz-Hinojosa *et al.*, "Harmonizing 3GPP and NFV description models to provide customized RAN slices in 5G networks," *IEEE Veh. Technol. Mag*, 2019. DOI 10.1109/MVT.2019.2936168.

[13] M. Shariat *et al.*, "A Flexible Network Architecture for 5G Systems," *Wireless Commun. and Mobile Computing*, Feb. 2019.

[14] 3GPP TS 38.331 V.15.6.0, "NR; Radio Resource Control (RRC) protocol specification (Release 15)," June 2019.

[15] METIS-II D2.4, "Final Overall 5G RAN Design," June 2017.

[16] C. Chang and N. Nikaein, "RAN Runtime Slicing System for Flexible and Dynamic Service Execution Environment," *IEEE Access*, vol. 6, pp. 34018–34042, 2018.

[17] C. Sexton, N. Marchetti, and L. A. DaSilva, "Customization and Trade-offs in 5G RAN Slicing," *IEEE Commun. Mag.*, vol. 57, pp. 116–122, April 2019.

[18] O. Adamuz-Hinojosa *et al.*, "Automated Network Service Scaling in NFV: Concepts, Mechanisms and Scaling Workflow," *IEEE Commun. Mag.*, vol. 56, pp. 162–169, July 2018.

[19] 3GPP TS 28.531 V.16.0.0, "Management and orchestration; provisioning (release 16)," Dec. 2018.

[20] C. Xu *et al.*, "On Multiple Virtual NICs in Cloud Computing: Performance Bottleneck and Enhancement," *IEEE Syst. J.*, vol. 12, pp. 2417–2427, Sep. 2018.

[21] 3GPP TS 23.501 V.16.0.2, "System Architecture for the 5G System; Stage 2 (Release 16)," Apr. 2019.

[22] A. Younis, T. X. Tran, and D. Pompili, "Bandwidth and Energy-Aware Resource Allocation for Cloud Radio Access Networks," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 6487–6500, Oct 2018.

[23] S. Khatibi, K. Shah, and M. Roshdi, "Modelling of Computational Resources for 5G RAN," *EuCNC, Ljubljana, Slovenia*, pp. 1–5, June 2018.

[24] Open Air Interface, [online] Available: `http://www.openairinterface.org/`.

[25] ETSI GS NFV-IFA 014 V3.2.1, "Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Network Service Templates Specifications," Apr. 2019.

[26] ETSI GS NFV-IFA 013 V3.2.1, "Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Os-Ma-Nfvo reference point - Interface and Information Model Specification," Apr. 2019.

Part III

Planning Solutions for RAN Slices Providing Services with Requirements in Terms of GBR or Latency

Chapter 6

Background and Problem Description

The first Section overviews the main state-of-the-art solutions to provide RAN slices in an automated way. Then, the second Section describes the problem addressed in the Part III of this dissertation. Finally, the last Section summarizes the thesis contributions on this topic.

6.1 Representative Solutions to Provision RAN Slices

Most of the available state-of-the-art solutions to provision Radio Access Network (RAN) slices are focused on (a) admission control; (b) deployment mechanisms and (c) dynamic resource provisioning.

Regarding the admission control, the authors of [1] have proposed a Markov-based model for characterizing the resource sharing in RAN slicing scenarios. They have considered several tenants requesting Guaranteed Bit Rate (GBR) services. This model is able to capture different admission control policies, which determine the transition probabilities between the different states in the model. These authors have also evaluated this model by measuring different performance metrics such as the blocking probability, the degradation probability, the throughput, and the occupation. In [2], the authors have proposed an admission control for RAN slices providing enhanced Mobile Broadband (eMBB) or ultra-Reliable Low Latency Communication (uRLLC) services in a Cloud RAN (C-RAN) envi-

ronment. The goal of the proposed admission control is to maximize the Mobile Network Operator (MNO)'s revenue by properly admitting the network slice requests, subject to the limited physical resource constraints. They have formulated the revenue maximization problem as a mixed-integer nonlinear programming and they have exploited an efficient approach (i.e., successive convex approximation and semidefinite relaxation) to solve it. In [3], the authors have studied the problem of virtual sensor network management from the perspective of a Single shared Sensor Network Infrastructure Provider (SSN-IP) that leases its physical resources to multiple concurrent application providers. Specifically, the authors have introduced a joint optimization framework to solve the problem of Application Admission Control and wireless Sensor Network Slicing (SNS). The proposed framework optimally decides (a) if new applications are admitted in the network; and (b) how to allocate the physical resources to the multiple concurrent applications, while considering constraints at the sensor node level (i.e., processing power and storage) as well as at the network level.

Concerning the deployment mechanisms to instantiate multiple RAN slices, the authors of [4] have proposed a base station agnostic scheme that creates RAN slices taking into account their performance requirements. The proposed framework considers the bottlenecks in the air capacity, backhaul and fronthaul transport networks capacity as well as the delay requirements imposed by the different service types. In [5], the authors have designed near-optimal low-complexity distributed RAN slicing algorithms. First, they have modeled the RAN slicing problem as a congestion game, and they have demonstrated that such game admits a unique Nash Equilibrium (NE). Then, they have evaluated the price of anarchy of the NE, i.e., the efficiency of the NE as compared with the social optimum. Next, they have proposed two fully-distributed algorithms that probably converge to the unique NE without revealing privacy-sensitive parameters from the slice tenants. In [6] the authors have focused on RAN slicing scenarios where Multi-access Edge Computing (MEC) services and traditional services coexist. They have formulated the utility maximization problem as a two-level problem, thereby guaranteeing the inter-slice isolation. To solve the problem they have proposed the Information Prediction and Dynamic Programming based RAN slicing (IP&DP-RS) algorithm, which is based on Support Vector Regression (SVR), Fuzzy Information Granulation (FIG), and Dynamic Programming (DP)

method. This algorithm can realize the intra-slice customization and optimize the network utility with high fairness in polynomial time complexity.

With respect to the dynamic resource provisioning in RAN slicing, the authors of [7] have investigated the problem of how to split the radio resources between multiple RAN slices in a scenario with Vehicle-to-Everything (V2X) and eMBB services involving Uplink (UL), Downlink (DL) and sidelink communications. Specifically, they have proposed a RAN slicing strategy based on an off-line reinforcement learning followed by a low-complexity heuristic approach to determine the split of resources assigned to the eMBB and V2X slices. In [8], the authors have proposed a novel framework to enable RAN slicing with diverse requirements in terms of throughput and latency. Specifically, RAN slices are scheduled based on the Earliest Deadline First (EDF) principle with further enhancements to cope with traffic dynamics. The authors of [9] have studied the potential advantages of allowing a non-orthogonal sharing of RAN resources in UL communications for a set of eMBB, massive Machine Type Communication (mMTC), and uRLLC devices to a common base station. These authors have revealed that their approach can lead, in some regiments, to significant gains in terms of performance tradeoffs among the three generic services as compared to orthogonal slicing. In [10], the authors have analyzed RAN slicing for multiple services using RAN resources such as base station, cache storage, and backhaul capacity. They have formulated the RAN slicing problem as a bi-convex problem that accounts for dependencies between resource allocation for each RAN slice and coordination of sharing the same resources. In [11], the authors have proposed an architecture based on Heterogeneous Network and Spectrum Sharing paradigms. This architecture is managed by means a three-tier scheduler that (a) allows a suitable resource partitioning between dedicated and shared base stations; and (b) introduces RAN-sharing and network virtualization to efficiently manage resource allocation in the shared base stations, while maintain independent allocation policies for commercial and public safety user without resorting to very complex multi-dimensional schedulers. In [12], the authors have considered the criterion for dynamic resource allocation among RAN slices based on a weighted proportionally fair objective. This approach achieves desirable fairness/protection across the RAN slices and their associated users. They have also shown that the objective is NP-hard, making an exact solution impracti-

cal. Furthermore, they have designed a distributed semi-online algorithm which meets performance guarantees in equilibrium and quickly converges to a region around the equilibrium point. In [13], the authors have studied the radio resource allocation problem for RAN slicing in a fronthaul-limited C-RAN. Specifically, they have considered a multi-cell virtualized wireless network supporting different slices with the user to Radio Remote Header (RRH)/Base Band Unit (BBU) allocation constraints. To address this problem, these authors have proposed a two-step iterative algorithm that jointly associates users to RRHs and BBUs, and allocates power and antennas. The goal of this algorithm is to maximize the system sum rate, while maintaining the RAN slice isolation. In [14], the authors have proposed a framework which considers coordination among several base stations to create an abstraction of systems' radio resources so that multiple RAN slices can be served, in a heterogeneous environment. With the use of this framework, the system's resources are dynamically distributed according to users' needs on an isolated and on-demand basis.

6.2 Problem Description

The Third Generation Partnership Project (3GPP) has identified four phases to describe the lifecycle management of a network slice instance: preparation, commissioning, operation and decommissioning [15]. These phases are depicted in Fig. 6.1.

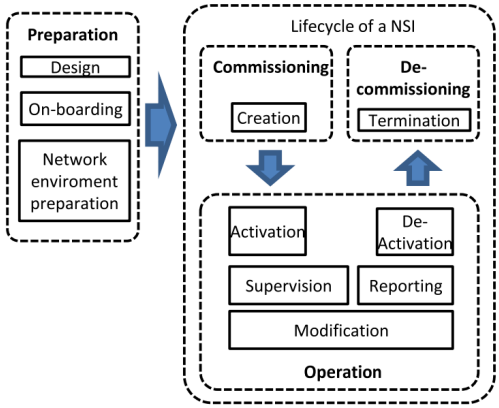


Figure 6.1: Lifecycle phases of a network slice instance [16].

The preparation phase begins when the MNO receives the deployment request for a network slice from a tenant. In this phase, the MNO performs the design and capacity planning of this network slice. It prepares the network environment and also performs the on-boarding and evaluation of network slice's constituents for testing purposes. In the commissioning phase the network slice is deployed. This means all the needed resources are allocated to satisfy the tenant requirements under the current traffic conditions. During the operation phase, the amount of allocated resources for this network slice may change according to its traffic demand. Finally, the lifetime of the network slice ends in the decommissioning phase, in which the MNO releases the assigned resources.

Most of the state-of-the-art solutions for RAN slicing assumes that the preparation phase triggers an admission control and then, once the RAN slice is accepted, its radio resources are determined by a dynamic resource provisioning algorithm executed in the operation phase.

Regarding the admission control, the authors of these proposals assume the tenant first provides the MNO with the performance requirements of the requested service. Then, the MNO verifies the feasibility of deploying it under the current traffic conditions. If it was feasible, the available radio resources would be reallocated for both the RAN slices already deployed and the requested one. If it was not possible, the request would be rejected.

Concerning the dynamic resource provisioning, the authors of these solutions assume the traffic demands of the deployed RAN slices may dynamically change. In such case, an algorithm reallocates the available radio resources among these RAN slices for guaranteeing their performance requirements. To that end, some works assume that upper and lower bounds are somehow, and maybe conservatively, established for such dynamic radio resource assignments.

Despite the state-of-the-art proposals provide excellent contributions for the admission control and the dynamic resource provisioning in RAN slicing, they do not consider the majority of constraints imposed by the Standards Developing Organizations (SDOs) responsible for standardizing the management operations in RAN slicing. This means the MNOs may face difficulties in integrating the state-of-the-art solutions in standard-based management architectures for RAN slicing.

Additionally, the majority of the state-of-the-art solutions assume a MNO

serves requests for deploying RAN slices on demand. This means this MNO executes an admission control immediately the requests are received and then, it deploys those RAN slices which are feasible (e.g., those whose requirements could be met). This approach has the benefit of accelerating the decisions for admitting or rejecting new RAN slice requests. However, these decisions are made on the spot, thus there exists some limitations in the way of renegotiating the Service Level Agreement (SLA) conditions between the MNO and the tenants in case of resource scarcity. Furthermore, the MNO has not the opportunity to reconfigure or update the RAN infrastructure due to the MNO must deploy immediately those RAN slices which were admitted. Finally, since the admission control is a short term and somehow myopic procedure, the admission of a deployment request could involve the rejection of more attractive (e.g., in economic terms) subsequent coming requests because of the resource scarcity.

Unlike the previous approach, this thesis considers the preparation phase of a RAN slice from an alternative perspective. It consists of planning in advance the amount of radio resources required by one or more RAN slices for a given planning window. Specifically, the MNO first translates the performance requirements of each RAN slice into network resources. To that end, the MNO needs to consider the lifetimes of all the RAN slices, which could be partially overlapped over the planning window; and, accordingly estimate the spatio-temporal traffic intensity experienced by each RAN slice in such window. With this information, the MNO can determine the busy hour, i.e., the time period when the RAN infrastructure suffers the worst-case inter-cell interference. Considering the busy hour, the MNO derives the amount of radio resources required by each RAN slice in each cell. The MNO then checks if the performance requirements of all the RAN slices are met with such allocation. If the checking procedure is successful, the MNO uses the derived amount of radio resources to define the upper and lower bounds needed by the dynamic resource provisioning algorithm for allocating radio resources to the deployed RAN slices during the operation phase.

Although our approach is more complex than the approach considered by most of the state-of-the-art solutions for RAN slicing. It provides the MNO more flexibility in case the checking procedure fails. Specifically, it allows the MNO to execute one or more of the following options: (a) adding more radio resources to the RAN infrastructure; (b) renegotiating the SLAs with one or more tenants;

and/or (c) rejecting the least attractive RAN slices. At the moment of writing this thesis, there are not solutions following this approach.

6.3 Thesis Contributions

Papers D-H gather the main contributions and findings of Part III of this dissertation. They are summarized as follows:

1. **Proposing different spectrum planning strategies, each giving a certain degree of flexibility to allocate resources per RAN slice (From Paper D).**

   This thesis proposes different ways of hiring capacity to the MNO and, based on them, we analyze a set of spectrum planning strategies with different degrees of flexibility for allocating spectrum resources for multiple RAN slices. The proposed strategies are evaluated in terms of scalability, spectrum isolation, utilization, and efficiency.

2. **Proposing an analytical model to evaluate the User Equipment (UE) blocking probability in an Orthogonal Frequency-Division Multiple Access (OFDMA) cell (From Papers E and F).**

   This thesis proposes an analytical model for assessing the UE blocking probability in a GBR slice for an OFDMA cell under Poisson session arrivals. Using our model, the MNO can decide the number of radio resources required by a cell to provide a GBR slice while the UE blocking probability is below a given threshold. The main novelty of the proposed model is the use of a Multidimensional Erlang-B system which meets the reversibility property. It means this model allows the adoption of an arbitrary distribution for the UE session duration. Additionally, this property involves the solution for the state probabilities has product form, thus it eases their computation. Furthermore, this model may take as input any distribution for the average channel quality within the cell. Additionally, our model considers the channel gain of a packet scheduler when it dynamically allocates radio resources.

3. **Analyzing the role of the RAN slicing architectural framework in the radio resource planning (From Paper G).**

   This thesis addresses the radio resource allocation problem from the perspective of RAN slice planning. To that end, we provide a step-by-step description of the role of the RAN slicing architectural framework in the radio resource planning for multiple RAN slices.

4. **Proposing a mathematical framework for planning the radio resources of RAN slices offering GBR services (From Paper G).**

   This thesis proposes a mathematical framework capable of translating the GBR requirements of the requested communication services into the minimum radio resource quota assigned for each RAN slice in each cell. Each quota guarantees the UE blocking probability for a RAN slice in a cell is below an upper bound under the inter-cell interference levels presented in the busy hour. More specifically we use game theory to model the radio resource planning in RAN slicing. The problem is formulated as multiple ordinal potential games, one per requested (or already deployed) RAN slice. In each game the players are the cells and their actions are the allocation of radio resources for each considered RAN slice. The goal of each game is to guarantee the GBR requirements of each considered RAN slice, while its UE blocking probability in each cell is below the upper bound. The existence of a NE solution is also demonstrated. Thus, we design novel strategies to solve the formulated problem. These strategies are based on better response dynamics and aim to minimize the UE blocking probability for all the RAN slices.

5. **Proposing an analytical model based on Stochastic Network Calculus (SNC) to evaluate the packet delay bound for a RAN slice, given the probability of exceeding such bound (From Paper H).**

   This dissertation proposes a SNC-based model which provides the packet delay bound of an uRLLC RAN slice in a single cell. To that end, this model considers as inputs: *i*) the amount of dedicated radio resources for this RAN slice, *ii*) the probability the packet delay is above the delay bound, *iii*) the Cumulative Distribution Function (CDF) for the Signal-to-Interference-

plus-Noise Ratio (SINR) experienced by the users served by this RAN slice, and *iv*) the traffic demand of this RAN slice, i.e., the distribution of the packet arrival rate, and the distribution of the packet size. In our model, the packet size distribution could be arbitrary. Additionally, to compute the CDF for the SINR perceived by the user, which is served by a specific uRLLC RAN slice, we use a novel model based on stochastic geometry. This model considers the impact of the interference incurred by multiple RAN slices deployed in neighbor cells on the capacity the serving cell offers to the serving RAN slice.

6. **Proposing a mathematical framework for planning the radio resources of RAN slices offering uRLLC services (From Paper H).**

   This thesis proposes a mathematical framework that—using the proposed SNC-based model—plans in advance the deployment of multiple RAN slices with different requirements in terms of traffic demand, latency and reliability. It aims to derive the dedicated radio resources for each RAN slice which satisfy their requirements throughout its lifetime. Specifically, the proposed framework relies on a novel heuristic to derive the amount of radio resources, which minimize the difference between the delay bounds achieved with such resources and the target delay bounds.

References

[1] I. Vilà, O. Sallent, A. Umbert, and J. Pérez-Romero, "An Analytical Model for Multi-Tenant Radio Access Networks Supporting Guaranteed Bit Rate services," *IEEE Access*, vol. 7, pp. 57651–57662, 2019.

[2] 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, 2019.

[3] C. Delgado *et al.*, "Joint Application Admission Control and Network Slicing in Virtual Sensor Networks," vol. 5, no. 1, pp. 28–43, 2018.

[4] G. Tseliou, F. Adelantado, and C. Verikoukis, "NetSliC: Base Station Agnostic Framework for Network SliCing," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3820–3832, 2019.

[5] S. D'Oro, F. Restuccia, T. Melodia, and S. Palazzo, "Low-Complexity Distributed Radio Access Network Slicing: Algorithms and Experimental Results," *IEEE ACM Trans Netw.*, vol. 26, no. 6, pp. 2815–2828, 2018.

[6] P. Zhao, H. Tian, S. Fan, and A. Paulraj, "Information Prediction and Dynamic Programming-Based RAN Slicing for Mobile Edge Computing," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 614–617, 2018.

[7] H. D. R. Albonda and J. Pérez-Romero, "An Efficient RAN Slicing Strategy for a Heterogeneous Network With eMBB and V2X Services," *IEEE Access*, vol. 7, pp. 44771–44782, 2019.

[8] T. Guo and A. Suárez, "Enabling 5G RAN Slicing With EDF Slice Scheduling," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2865–2877, 2019.

[9] 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. 55765–55779, 2018.

[10] P. L. Vo, M. N. H. Nguyen, T. A. Le, and N. H. Tran, "Slicing the Edge: Resource Allocation for RAN Network Slicing," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 970–973, 2018.

[11] D. Marabissi and R. Fantacci, "Heterogeneous Public Safety Network Architecture Based on RAN Slicing," *IEEE Access*, vol. 5, pp. 24668–24677, 2017.

[12] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Pérez, "Multi-Tenant Radio Access Network Slicing: Statistical Multiplexing of Spatial Loads," *IEEE ACM Trans Netw.*, vol. 25, no. 5, pp. 3044–3058, 2017.

[13] S. Parsaeefard, R. Dawadi, M. Derakhshani, T. Le-Ngoc, and M. Baghani, "Dynamic Resource Allocation for Virtualized Wireless Networks in Massive-MIMO-Aided and Fronthaul-Limited C-RAN," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9512–9520, 2017.

[14] G. Tseliou, F. Adelantado, and C. Verikoukis, "Scalable RAN Virtualization in Multitenant LTE-A Heterogeneous Networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6651–6664, 2016.

[15] 3GPP TS 28.530 V.16.1.0, "Management and orchestration; Concepts, use cases and requirements (Release 16)," Dec. 2019.

[16] R. Ferrús, O. Sallent, J. Pérez-Romero, and R. Agustí, "On the automation of RAN slicing provisioning: solution framework and applicability examples," *EURASIP J. Wirel. Commun. Netw.*, vol. 2019, no. 1, pp. 1–12, 2019.

Chapter 7

Paper D. Radio Access Network Slicing Strategies at Spectrum Planning Level in 5G and Beyond

Authors:

Pablo Munoz, Oscar Adamuz-Hinojosa, Jorge Navarro-Ortiz, Oriol Sallent, and Jordi Perez-Romero.

The paper has been published in the IEEE Access, April, 2020.

Abstract

The new Fifth Generation (5G) era has transformed previous mobile generations into fast, smart networks that will be more responsive and customizable. With network slicing, 5G networks can be dynamically adapted to the different needs of specific vertical industries. This capability has opened the opportunity to new business models whereby infrastructure owners can monetize their investment by leasing resources to third parties (i.e., tenants). In this respect, a challenging task for the owner of the radio access network infrastructure (i.e., the network provider) is the spectrum planning of multi-tenant scenarios. This paper proposes different alternatives of hiring capacity to the provider as well as a set of spectrum planning strategies, each giving a certain degree of flexibility to allocate resources per tenant. These strategies are evaluated in a 5G small cell multi-tenant network through snapshot-based simulations. The performance of the strategies is assessed in terms of scalability, spectrum isolation, utilization and efficiency.

7.1 Introduction

The enormous growth of subscribers' data traffic in the last years has stressed the need of a substantial change on current mobile networks. Likewise, the growing industrial digitalization has boosted a wide range of novel applications with stringent and business critical requirements. To meet these rising and diverse demands, the Fifth Generation (5G) mobile network has introduced innovative architectural and technological features [1], such as network slicing, network softwarization, massive Multiple-input Multiple-output (MIMO) and device-to-device communications.

Currently, the cost of upgrading the infrastructure is extraordinary for most Mobile Network Operators (MNOs) since they rely on relatively low-cost flat rates. Deploying new infrastructure entails large delays due to site acquisition and installation, spectrum leasing, etc. Furthermore, in an operational network, there are underutilized resources due to traffic demand variations. Under these premises, the traditional scenario with independently deployed networks is unfeasible to embrace the 5G network evolution. Instead, to provide cost-efficient

solutions with a shorter time-to-market, new business models based on cooperation and infrastructure sharing are needed [2]. In this way, services can be deployed faster while reducing Capital Expenditures (CAPEXs) and Operational Expenditures (OPEXs).

Network sharing is a paradigm that enables MNOs to act as infrastructure providers, leasing the infrastructure to other MNOs or Mobile Virtual Network Operators (MVNOs) for entering the market or extending coverage/capacity. Multi-tenancy is an extension of this concept where a third-party making use of the infrastructure as a tenant becomes a service provider, such as those offering Over-The-Top (OTT) applications (e.g. streaming) or vertical industries (e.g. manufacturing, entertainment, public safety) [3].

Service providers or tenants impose diverse technical and business requirements to the network. To provide efficient deployment of these services, the network should be flexible and scalable. Network slicing has been proposed as an efficient solution to provide flexibility and scalability in the 5G mobile networks [4], [5]. This feature consists in creating multiple logical, self-contained networks on top of a common shared physical infrastructure and, therefore, it can be used to support multi-tenancy on the 5G network. In this case, each network slice is specifically built to meet the service requirements of a certain tenant (e.g. in terms of speed, capacity, connectivity and coverage). Technologies such as Software-Defined Networks (SDN) and Network Function Virtualization (NFV) are key enablers to the implementation of network slicing [6]. They enable the use of common resources such as storage and processors to run logical (software-based) elements that can be controlled programmatically.

Network slicing also provides adequate resource isolation, independent scaling and increased statistical multiplexing. Creating an independent virtualized End-to-End (E2E) network involves the configuration of the Radio Access Network (RAN), transport, and core network [7]. However, the complexity of the configuration in the RAN is greater due to difficulties in partitioning radio resources and virtualizing functionalities with tight latency requirements [8, 9]. On the one hand, slices cannot interfere with each other to ensure isolation. A strict resource isolation implies orthogonal spectrum allocation between slices. This may result in inefficient resource utilization, especially in large service areas with varying traffic demands. On the other hand, the lower layers of the radio protocol stack

have a large number of interfaces and varying capabilities that operate on a very fast timescale. This certainly complicates the virtualization and limits the functional split options between a Centralized Unit (CU) and Distributed Unit (DU) in a 5G RAN node.

The benefits of network slicing in the 5G RAN, or Next Generation Radio Access Network (NG-RAN), will rely on the flexibility and scalability offered by the lower layers of the radio protocol stack. In this way, the Third Generation Partnership Project (3GPP) has defined service and operational requirements for 5G network slicing [10] and technical specifications for the 5G air interface, known as New Radio (NR) [11]. The latter includes key technology features in the physical layer such as scalable numerology to support multiple Bandwidths (BWs) and spectrum and flexible frame structure to provide low latency and high efficiency. Thus, the high degree of configurability offered by the NR enables better resource sharing between tenants and better customization of slices according to service requirements.

In the 5G-RAN, the spectrum planning is in charge of allocating spectrum resources to each slice before its operation based on capacity and isolation requirements. There can be different ways to perform spectrum planning depending on the Service Level Agreement (SLA) between the network provider and tenant. This SLA determines how the provider can allocate spectrum resources over the network for each slice. Depending on the required level of isolation, for example, a slice can require exclusive (i.e. non-shared) use of a resource in the entire network or, alternatively, the resource can be shared in different cells with other slices, so that exclusive use is limited to the cell area. Each level of radio isolation determines the multiplexing gains and gives the provider some degree of flexibility for allocating spectrum resources. In this way, a slice with exclusive use of radio resources will prevent other slices from using them. The impact of the radio isolation on the spectrum planning have not been analyzed yet with the required depth.

In this work, we propose different ways of hiring capacity to the network provider and, based on them, we analyze a set of spectrum planning strategies with different degrees of flexibility for allocating spectrum resources. The proposed strategies are evaluated in terms of scalability, spectrum isolation, utilization and efficiency in a 5G Small Cell (SC) network. SCs can help satisfy

the increasing traffic demand while they facilitate the adoption of network slicing [12]. However, these low-power devices entail more complex spectrum planning than macro-cells because they facilitate extensive spatial reuse.

The remainder of this paper is organized as follows. In Section 7.2, the literature related to RAN slicing is discussed. Section 7.3 describes the system model. In Section 7.4, the proposed strategies for spectrum planning are presented. Section 7.5 provides the performance results for the different strategies. Finally, Section 7.6 summarizes the conclusions.

7.2 Related Works

The importance of network slicing has been widely recognized, becoming a fundamental topic in many research initiatives. Diverse standard organizations such as 3GPP, European Telecommunications Standards Institute (ETSI), International Telecommunication Union (ITU) and Internet Engineering Task Force (IETF) are spending much effort to network slicing, offering different views of it. There is a broad consensus that the SDN/NFV paradigm is a key enabler to provide functional customization over the same infrastructure. Comprehensive reviews on SDN/NFV-based solutions related to network slicing are discussed in [13, 14]. The work in [15] analyzes a proposal from ETSI that incorporates the capabilities of SDN into the NFV architecture to enable the realization of network slices. In [16], a slicing-enabled SDN core network architecture is proposed for the automotive vertical use case. From the management viewpoint, the work in [17] proposes a SDN/NFV-based framework to manage E2E network slices, including their lifecycle and context management, monitoring and configuration. The creation process of network slices is addressed in [18], where network slice descriptors are used to make this process more agile and automatic. The idea of a network and application store is introduced in [19] to simplify the procedure of defining the network slice. It provides a marketplace for delivering customized network functions and service templates tailored to specific use cases.

RAN slicing [20] poses many interesting challenges related to the management of the slice's lifecycle, as well as the abstraction and sharing of radio resources. In [21], the issue of RAN resource allocation is addressed considering resources of a base station such as radio BW, caching, and backhaul components. The

support for latency-sensitive and time-critical applications through RAN slicing is investigated in [22], where the number of radio resources and their relative position in the time domain are considered to satisfy the latency requirements. Similarly, the work in [23] presents a novel slice resource allocation approach that introduces the concept of mini-slots to support low latency communications. The issue of spectrum allocation to minimize inter-slice interference is analyzed in [24], where various algorithms are developed to guarantee orthogonality among RAN slices. However, such orthogonality may lead to inefficient resource usage. With a special focus on the E2E isolation, a systematic overview of existing isolation techniques is provided in [25]. Nevertheless, an exhaustive analysis of the radio isolation is still missing.

The concept of RAN slicing at different levels is introduced in [26], where each defined level provides a specific degree of granularity in the assignment of radio resources, isolation and customization. The spectrum allocation at the scheduler level is investigated in several works [27, 28]. The scope of these works lies on the dynamic resource allocation (operating at a faster time scale than the spectrum planning level) to cope with the traffic dynamics. The dynamic slice scheduling using a centralized approach (i.e. a SDN-enabled controller) for heterogeneous networks has been addressed in [29]. In [9], the link-layer scheduler is partitioned into two levels to perform inter- and intra-slice scheduling. In [30], the two-level hierarchy is implemented by giving priority to the different slices (e.g. prioritizing enhanced mobile broadband) and the users within the slices. In [31], this hierarchy is internal to the base station and supported by a centralized entity that controls spectrum sharing between tenants. At the spectrum planning level, the work in [32] proposes a spectrum planning scheme that maximizes the spectrum utilization. However, the isolation issue is not considered as part of the optimization problem.

Artificial Intelligence (AI) has also been successfully applied to network slicing. In [33], an AI-enabled 5G network architecture is proposed to adjust service configuration and control based on changes in user needs, environmental conditions and business goals. Some interesting AI techniques that have recently been applied to resource allocation of slices are Reinforcement Learning (RL) [34], deep RL [35, 36, 37, 38], deep learning neural networks [39, 40] and evolutionary algorithms [41]. These techniques are particularly effective in handling complicated

control problems.

There are still few works analyzing the impact of the realization of RAN slicing from a management perspective. In [42], a framework is proposed for the specification of RAN slices based on a set of configuration descriptors that characterize features, policies and resources. Such a framework has been extended in [43] with the specification of certain radio resource management functionalities (e.g. admission control and packet scheduler) as part of the RAN slice configuration.

Although there have been substantial research on RAN slicing, there is still little work on analyzing in detail different business-driven models for multi-slice/tenant spectrum sharing. The work in [44] proposes two spectrum sharing models and algorithms with different level of flexibility, depending on whether a set of dedicated and shared resources per tenant are predefined or not. However, the algorithms assign spectrum resources per user, acting as a link-layer scheduler. Thus, this algorithm may not scale well in large networks, as traffic demand variations among cells are not considered. In addition, it is hard to measure whether the offered capacity conforms or not to the contract because it would require extensive metric monitoring in the network. The present work tackles the problem of spectrum sharing at the planning level, which provides a wider view of the network, enabling optimal resource allocation per cell. In addition, this level facilitates the mapping of high-level capacity specifications to lower-level constraints, ensuring fairer resource allocation among tenants.

This work further develops the functional framework proposed in [45], [46] to include isolation as a key feature for slice specification and propose appropriate business-driven models for spectrum sharing. Such a framework enables self-planning of the radio access capacity in a NG-RAN, including automatic cell re-configuration mechanisms in order to facilitate the realization of slices. Under this framework, the contributions and novelties of this paper are the following:

- Proposing an effective business-driven model for capacity specification, simplifying the later capacity compliance analysis.
- Introducing isolation as a part of the slice specification for RAN slices to ensure that the traffic load of one slice does not negatively affect other slices.
- Proposing different spectrum sharing strategies at the planning level for

RAN slicing, each giving a certain degree of flexibility to allocate resources per slice.

- Providing an exhaustive analysis of the proposed strategies in terms of scalability, spectrum isolation, utilization and efficiency in a NG-RAN. In essence, there is a trade-off between maximizing resource usage and avoiding/reducing co-channel interference between slices.

7.3 System Model

7.3.1 Network Model

Consider a NG-RAN consisting of a set B of 5G NR SCs that are owned by a certain infrastructure provider. The SCs are conceived to satisfy high traffic demands in localized areas. Multiple tenants (e.g. OTT providers or industry vertical market players) can request and lease resources from the infrastructure provider to deploy a set S of network slices. The slice s provides a service over a certain area specified through a subset $B_s \subseteq B$ of SCs. The aggregated service demand D of the slices is non-uniformly distributed over the considered area. Accordingly, the network topology is assumed irregular (i.e. with cell service areas of different size) to absorb the service demand with maximum resource usage efficiency. Under this demand distribution, a set U of User Equipments (UEs) exist in the scenario, being U_s , $s \in S$ the subset of UEs belonging to slice s.

The UEs of a slice should be provided with enough resources to satisfy a guaranteed bit rate or service demand. In particular, the accessing scheme is Orthogonal Frequency-Division Multiple Access (OFDMA), since 3GPP agreed to adopt it for 5G NR [11]. Specifically, the system supports scalable numerologies with subcarrier spacing of $2^{\mu} \cdot 15$ KHz (μ= 0, 1, ..., 4). As shown in Fig. 7.1, the system BW is divided into a set of Resource Blocks (RBs), each consisting of 12 consecutive subcarriers in the frequency domain. In Release 16 [47], the number of RBs ranges from 11 to 273 units. Depending on the numerology, the number of RBs is mapped to a specific BW. For example, a single NR carrier with 133 RBs would require 25 MHz BW for $\mu = 0$ or 50 MHz for $\mu = 1$. The maximum allowable BW depends on the spectrum band where NR operates. In particular,

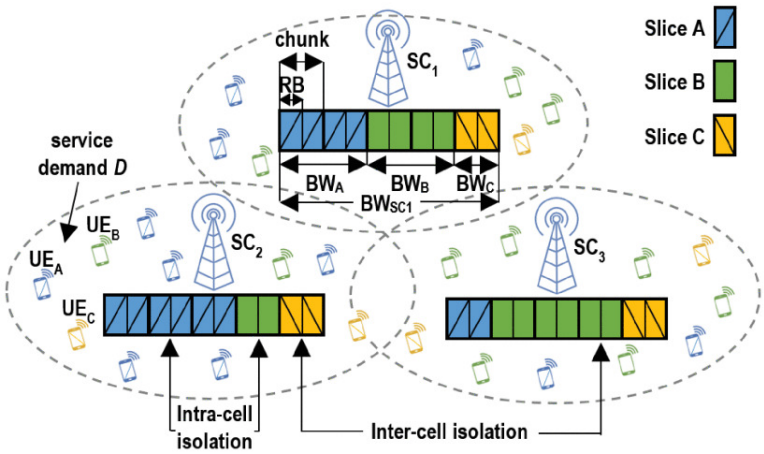


Figure 7.1: Resource model of the NG-RAN and an example with different levels of isolation among slices

this limit is 100 MHz for the sub-6 GHz band and 400 MHz for the millimeter wave band. At frequencies below 6 GHz, the cell size is larger and subcarrier spacing of 15 and 30 kHz are appropriate, while at higher frequencies, subcarrier spacing of 60 and 120 kHz are more suitable.

The number of RBs is typically high for most system BWs. Then, from the perspective of allocating the spectrum resources to the different slices, it becomes advantageous to reduce the management complexity by grouping the RBs into spectrum chunks, which are allocated to the slices as an indivisible unit. This can be done through the concepts of BW part and RB group defined in [48] and [49], respectively. The BW part is a subset of contiguous common RBs for a given numerology. This new feature will enable the coexistence of multiple slices with different physical layer requirements. The RB group is a collection of RBs within a given BW part that can be allocated to the scheduled UEs. The size of the spectrum chunk can be used to establish the minimum allocation unit size. This parameter may serve to reduce the signaling overhead at the expense of a loss of flexibility, which could be critical when the number of slices is large.

Bearing in mind these considerations, the RBs are grouped into a set R of spectrum chunks (see Fig. 7.1). Each chunk is composed of a number of RBs equal to the minimum allocation unit size. From the set of chunks, a subset R_b of the available chunks in the system BW is allocated to the SC b, $b \in B$. In

addition, among these chunks, $R_{b,s}$ are allocated to the slice s, $s \in S$. Depending on how these chunks are allocated to the slices, intra-cell or inter-cell isolation can be provided as will be explained in detail in the next section. In any case, to provide a slice with full coverage within its service area B_s, at least one chunk is allocated in every SC, i.e. $|R_{b,s}| > 0$, $b \in B_s$.

The subset of R_b allocated chunks provides the SC BW BW_b, which in turn determines the required transmit power, P_b^{TX} of the SC b. In particular, the transmit power must ensure a targeted Signal-to-Interference-plus-Noise Ratio (SINR) at the cell coverage range, i.e.:

$$P_b^{TX} = \min(P_N \cdot G_{PL,b}(d_{edge}) \cdot BW_b \cdot SINR_{edge}, P_{max}^{TX}) \tag{7.1}$$

where P_N is the noise power measured in one chunk,, $G_{PL,b}(d_{edge})$ is the path gain (loss) evaluated at the distance d_{edge} between the SC and the cell-edge, $SINR_{edge}$ is the target value at that distance and P_{max}^{TX} is the maximum transmit power. The cell-edge is determined by the distance to the closest adjacent SC.

The received power $P_b^{RX}(d)$ at a certain distance d when served by the SC b is given by:

$$P_b^{RX}(d) = P_b^{TX} \cdot G_b(d) \tag{7.2}$$

where $G_b(d)$ is the overall gain at the distance d including the antenna gain, the shadow fading and the path loss. The fast fading is not modelled as the channel gain is measured over a large time scale.

The $SINR(u, r)$ experienced by the UE u when transmitting on the chunk r is defined as:

$$SINR(u, r) = \frac{P_b^{RX}(d_{b,u})}{\left(\sum_{j \in B \setminus \{b\}} L_j \cdot \pi_j(r) \cdot P_j^{RX}(d_{j,u})\right) + P_N} \tag{7.3}$$

where $d_{b,u}$ is the distance between the SCs b and the UE u, L_j is the cell load factor of the SC j and $\pi_j(r)$ is a function that takes the value 1 when the chunk r is allocated to the SC j and the value 0 otherwise. The cell load factor is determined from the relation between the service demand and cell capacity, i.e.:

$$\hat{L}_j = \frac{\sum\limits_{u|j=\Gamma(u)} D_u}{\sum\limits_{u|j=\Gamma(u)} BW_u \cdot SE_u} \tag{7.4}$$

and

$$L_j = \min\left(\hat{L}_j, 1\right) \tag{7.5}$$

where D_u and SE_u are the service demand and spectral efficiency of the UE u, respectively, BW_u is the fraction of the cell BW allocated to this UE according to the slice's constraints and the resource scheduling policy and $\Gamma(u)$ is a function that returns the serving SC based on the strongest SINR. Then, the cell overload factor for the SC j is defined as:

$$OL_j = \hat{L}_j - L_j \tag{7.6}$$

This variable serves as an indicator of the congestion level in the network. Thus, under congested situations, it will be greater than zero.

The spectral efficiency $SE(u, r)$ of the UE u in the chunk r is derived from the $SINR(u, r)$ according to the following SINR mapping [50]:

$$SE = \begin{cases} 0, & SINR < SINR_{min} \\ \alpha \cdot \log_2(1 + SINR), & SINR_{min} \leq SINR < SINR_{max} \\ SE_{max}, & SINR \geq SINR_{max}, \end{cases} \tag{7.7}$$

(7.8)

where SE_{max} is the maximum achievable spectral efficiency with link adaptation, $SINR_{min}$ and $SINR_{max}$ are the minimum and maximum SINR values, respectively, and α stands for the attenuation factor, which represents implementation losses. Lastly, the UE throughput $T(u)$ is given by:

$$T(u) = \min\left(\frac{BW_u}{|R_{b,s}|} \cdot \sum_{r \in R_{b,s}} SE(u, r), D_u\right) \tag{7.9}$$

where $b = \Gamma(u)$ is the serving SC. The UE throughput depends on the resource scheduling scheme through the variable BW_u. For example, assuming a Round-Robin scheme, this variable is given by:

$$BW_u = \frac{BW_{ch} \cdot |R_{b,s}|}{|U_{b,s}|} \tag{7.10}$$

where BW_{ch} is the BW of one chunk and $U_{b,s}$ stands for the subset of UEs connecting to the same SC b and slice s that fairly share the spectrum, i.e. $U_{b,s} = \{v | v \in U_s \wedge b = \Gamma(v)\}$.

7.3.2 RAN Slicing Framework

A RAN slice defines a particular behavior of the NG-RAN in terms of capabilities and parameters configuration to meet the service requirements specified by the tenant. A key aspect in the orchestration and configuration of the RAN is how the radio spectrum is allocated and shared between the slices. The infrastructure provider is responsible for deploying and operating a number of concurrent RAN slices, including procedures such as slice instantiation, scaling and termination. These procedures should be carried out in an automated and agile way, allowing rapid adaptation to the business needs.

In order to address these issues, the general framework for network slicing with focus on the RAN is depicted in Fig. 7.2. This framework is based on a layered architectural approach and it is well aligned with most proposals from the literature [13, 27, 46], as well as standardization bodies (e.g. 3GPP, ETSI). Going into details, the service layer acts as the interface between the tenant and the infrastructure provider through a set of management functions to support several tasks such as SLA negotiation or performance monitoring. To describe the service requirements with a high abstraction level, the tenants has at its disposal the Generic Slice Template (GST) [51], which is a set of attributes that characterize a type of network slice (e.g., a mobile broadband slice). The tenant is asked to modify the GST to include its particular service requirements, giving place to the generation of the Individual Slice Template (IST). This template may contain requirements from the SLA, e.g. key performance indicators for throughput or latency, and other aspects such as demand patterns and additional services. The service layer delivers the IST containing the service level description to the

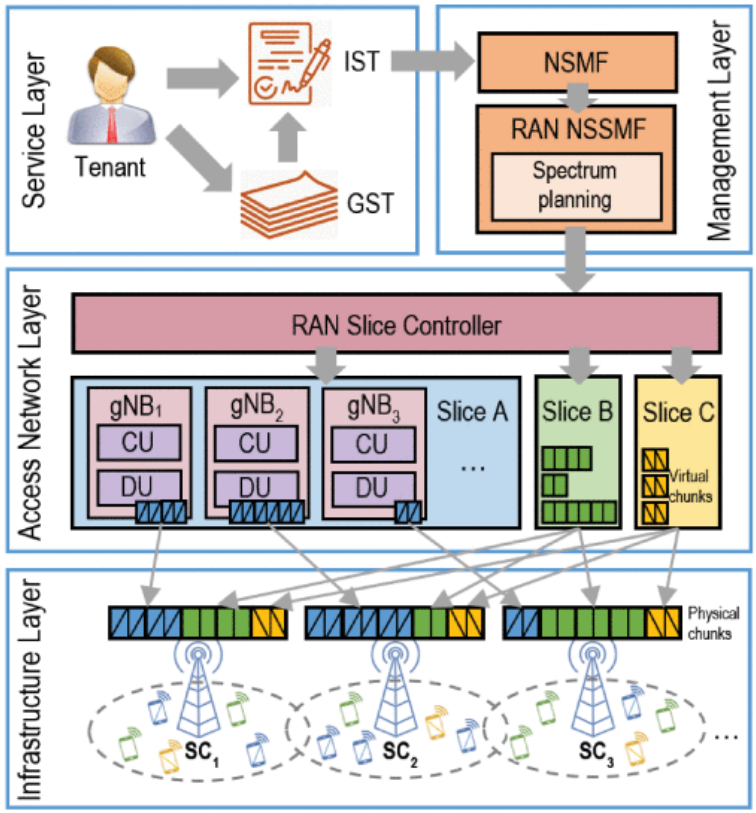


Figure 7.2: Architectural framework for network slicing

management layer. This latter comprises the Network Slice Management Function (NSMF), which is responsible for the creation and operation of the E2E slice. The NSMF relies on several Network Slice Subnet Management Function (NSSMF), each of which covers a particular network domain (e.g., access or core). To reduce the operational complexity, each resource domain that constitutes the E2E slice may have its own IST. Consequently, the RAN IST allows customizing the functions, policies and resources within the radio protocol stack for the RAN slice configuration. The RAN NSSMF is in charge of translating the slice requirements included within the IST to the configuration parameters in the RAN. To achieve this, the RAN NSSMF may support a wide range of internal functions for provisioning and performance/fault management, such as spectrum planning, admission control, SLA conformance monitoring or traffic forecasting.

Among them, the spectrum planning manages the long-term allocation of spectrum chunks for each slice given its capacity requirement and the desired level of resource isolation.

The access network layer comprises multiple instances of different combinations of logical resources grouped as slices. The main logical unit in the RAN is the next Generation NodeB (gNB), which hosts the full radio protocol stack functionality and it is decomposed into the CU and DU. This logical division provides deployment flexibility to split and move NR functions between the CU and DU entities. There is a broad consensus on the suitability of the SDN/NFV framework to implement these network functions. The logical slices are managed by a programmable RAN slice controller that is responsible for short-term decisions (e.g., inter-slice scheduling) considering the traffic dynamics of the slices and the guidance, parameters and policies provided by the RAN NSSMF.

Lastly, the infrastructure layer comprises the set of physical network resources in the RAN including SCs, edge data centers and interconnections through fiber or wireless-based transport networks. The RAN resources are located in strategic Points of Presence, e.g. the cell sites for SCs and the central offices for edge data centers. The virtualized resources (e.g., a certain number of virtual chunks) need to be mapped to actual physical resources (e.g., a set of frequencies in the spectrum band). To facilitate this task, the 5G NR physical layer incorporates new features such as the BW part [48].

7.4 Spectrum Planning Strategies for RAN Slicing System Model

The first part of this section introduces a business-driven model to specify network capacity requirements. Then, it analyzes how these specifications are translated into different spectrum allocation strategies.

7.4.1 Slice Specification

The description of the slice requirements made by the tenant involves, besides quantitative definitions of the required capacity, a set of conditions under which the leased capacity is operated and managed (e.g., the isolation). This specifica-

tion should be simple and expressed with a high abstraction level, avoiding details that might complicate other management tasks, such as the capacity conformance testing or the BW throttling. With the aim of automating this procedure, and following the same principles as in Fig. 7.2, the tenant is invited to build an IST containing the necessary parameters and their configuration for the capacity provisioning.

Specifically, the proposed IST should include the following attributes or parameters that are related to the required capacity and isolation (i.e. application-specific attributes are out of the scope of this paper).

1. **Requested Capacity Value(s)**: It specifies the capacity value(s) $v^{(s)}$ that satisfies the requirements of the slice s. This parameter can accept multiple definitions or values according to other parameters. For example, depending on the link direction, it can be downlink or uplink capacity; depending on the QoS class, it can be maximum or guaranteed capacity. It can also depend on the reservation type or the granularity level. Since these latter characteristics deserve special attention, they will discussed with more details below. In any case, the infrastructure provider must guarantee the requested capacity value(s) under the conditions given by the additional parameters.

2. **Capacity Reservation Type**: It determines whether the requested capacity is expressed in terms of throughput (e.g., in Mbps) or in terms of number of resources (e.g., in chunk units). While the former entails more accuracy in specifying the required capacity at the service layer, the latter can simplify the SLA compliance analysis and provide a fair resource usage between slices, especially under strong interference and congestion situations. In case of a throughput-based specification, the capacity conformance testing is carried out at the service layer, regardless of the underlying RB utilization. This may result in unfair resource sharing for low-traffic slices, where the UEs are receiving higher interference from heavily loaded slices. The choice on the reservation type may introduce additional parameters. For example, a throughput-based specification could also include a capacity margin to accept some excess traffic while there are available resources.

3. **Capacity Granularity Level**: Besides the various definitions of capacity

already explained, the requested capacity value(s) can also be defined with a certain granularity level in the network. In particular, this parameter specifies whether the requested capacity value is defined on a per-UE (i.e. the service demand), per-cell or per-network basis. The per-UE capacity represents the lowest level of granularity. In case the requested capacity is only given per-UE, the maximum number of sessions or the maximum number of UEs (either per cell or per network) should be specified to quantify the aggregated capacity that is required in the network. The per-cell capacity implies guaranteeing an exact amount of capacity in every SC. Lastly, the per-network capacity enables a more flexible capacity allocation among SCs according to the spatial traffic distribution. In this case, 'network' refers to the cluster of SCs, B_s , which provide the service. Note that the per-network capacity requirement is equivalent to targeting an average cell capacity, calculated over the number of SCs in the cluster.

4. **Resource Isolation Level**: This parameter indicates the degree of resource isolation with other slices. Consider the chunk i to be allocated in the SC j for a given slice. The following isolation levels are distinguished: (i) no isolation: other slices can use the chunk i in any SC, including the SC j; (ii) intra-cell isolation: other slices can use the chunk i in a SC other than SC j; and (iii) inter-cell isolation: other slices cannot use this chunk in the entire network, ensuring radio-electrical isolation. Introducing isolation may simplify capacity management and supervision. For example, due to the individual usage of resources, capacity conformance testing would not be necessary. With no isolation, accounting for the resource consumption over the time is required for ensuring equal RB sharing between slices or, alternatively, the packet scheduler could adopt a fair-share scheduling policy.

5. **Spatio-Temporal Constraints**: The spatial constraints limit the extent of the service area (e.g. a factory, a stadium, a mall, etc.) and, therefore, the cluster of SCs, B_s , serving the UEs. The time constraints define the time window(s) during which the service is offered to the UEs. The requested capacity value could be dependent on the spatial and time domains (e.g. to assign greater values in peak periods or high-traffic areas) at the expense

of a more complex management.

From an economic viewpoint, the resource-based specification with intra- or inter-cell isolation enables a fairer sharing model than the throughput-based specification, since the tenants are charged based on their resource consumption while the inter-slice interference can be avoided or limited. Additionally, this model provides better protection against congestion situations, thus being an attractive solution in multi-tenant environments.

The throughput-based specification is challenging for the infrastructure provider since a certain throughput should be guaranteed regardless of the underlying resources. The management system can continuously monitor the network throughput to detect a lack of capacity and provide the required changes in the network infrastructure. This problem, which is out of the scope of this work, is addressed in [45], where a self-planning entity runs an iterative algorithm to derive planning actions such as adding or relocating SCs in order to meet the capacity needs. On the other hand, the resource-based options require simpler cell planning, since there is a direct mapping between the number of RBs and the number of SCs. However, the existence of different levels of isolation entails a complex scenario with a greater variety of resource allocation strategies. Each strategy represents a particular degree of flexibility for allocating spectrum chunks. The infrastructure provider can leverage this flexibility to apply its own resource allocation policies. Example policies are minimizing inter-cell interference or maximizing slice isolation. The latter can facilitate the application of interference-mitigating strategies separately for each slice.

Since the throughput-based specification is analyzed in [45], this work focuses on the different resource-based specification approaches, analyzing their impact on isolation, performance and flexibility for enforcing operator's policies. With respect to other conditioning parameters, for the sake of simplicity, in this work the requested capacity is expressed as a guaranteed downlink capacity. In addition, the per-UE granularity level is excluded from the analysis since the requested per-UE capacity could be expressed on a per-cell or network basis by simply considering the maximum number of sessions or UEs.

7.4.2 Resource-Based Spectrum Sharing Strategies

Considering the capacity reservation type, granularity level, and isolation level parameters, there are six representative resource-based capacity specifications that are mapped to different strategies of spectrum sharing between slices. These strategies determine an average resource allocation at the planning phase. Thus, they are compatible with elastic resource allocation performed during the operation phase [33].

- **Strategy 1 (S1): Per-Cell Resource-Based Capacity Planning With Inter-Cell Isolation**: In this strategy, the requested value of spectrum chunks, $v^{(s)}$, is a constant value per cell. In this way, all the SCs should allocate the same number of chunks for the slice. Furthermore, due to the inter-cell isolation requirement, other slices cannot use the selected chunks. From the infrastructure provider' interests, concentrating the required resources over the network into a minimum number of chunks is an attractive solution to leave room for other slices demanding inter-cell isolation. This represents the most likely situation in a multi-tenant scenario. For example, the provider can simply select a number of chunks per slice from the set R. However, in case there is a plethora of available chunks (i.e. not allocated to other slices), the resource allocation could target the reduction of inter-cell interference in the network by selecting disjoint sets of chunks among SCs.

  This strategy of resource allocation does not provide flexibility to allocate resources due to the hard constraint on the number of chunks per cell. Consequently, S1 will not be able to adapt to spatial variations of the traffic demand. On the contrary, the high isolation level enables good protection against inter-slice interference and congestion situations.

- **Strategy 2 (S2): Per-Network Resource-Based Capacity Planning With Inter-Cell Isolation**: The requested value of spectrum chunks in S2 is defined on a network basis. It represents a soft constraint since a number $v^{(s)}$ of chunks in the network is to be freely distributed among the SCs. This problem is equivalent to targeting a specific per-cell average of the number of chunks. S2 differs from S1 because the former allows some

variations in the number of chunks per SC in order to cope with spatial traffic variations.

The S2 is divided into three steps:

1. Select a set $R^{(s)}$ of chunks for the slice (e.g. chunk #1 and #2). This number of chunks should be high enough to allocate $v^{(s)}$ chunks among the SCs. As in S1, the provider typically concentrates the required resources into a minimum number of chunks in the network.
2. Determine the number of chunks $|R_{b,s}|$ to allocate at each SC considering the spatial traffic variations and the requested value $v^{(s)}$. This number is chosen to be proportional to the estimated value of the slice's service demand in the cell area, whose actual value is defined as:

$$D^{(b,s)} = \sum_{\substack{u|b=\Gamma(u) \\ b\in B_s, u\in U_s}} D_u \tag{7.11}$$

where D_u is the service demand of the UE u and $\Gamma(u)$ returns the serving SC. The estimated value, $\hat{D}^{(b,s)}$, is based on the method proposed in [45], considering the service demand D_u and the correlation that can be expected between the slice's service demand and the actual network's service demand. To ensure accessibility from any location, at least one chunk is allocated per SC. In addition, the total number of allocated chunks should be equal to the required value $v^{(s)}$, i.e.:

$$v^{(s)} = \sum_{b\in B_s} |R_{b,s}| \tag{7.12}$$

Algorithm 1 describes the procedure to determine the number of chunks per SC, where $\hat{r}_{b,s}$ stands for the actual number of chunks and $r'_{b,s}$ is the targeted number of chunks, calculated as:

$$r'_{b,s} = \frac{\hat{D}^{(b,s)}}{K} \tag{7.13}$$

where K is the constant of proportionality between the cell demands

and the number of chunks. Specifically, the chunks are allocated following an iterative process where only one chunk is allocated for each SC within the same iteration. If, at a certain iteration in the loop starting at the line 4, the number of allocated chunks in a SC reaches the targeted value $r'_{b,s}$ calculated at the line 3, no more chunks will be allocated in that SC at this stage. However, if there are still chunks to allocate to the slice after the loop, the process continues allocating chunks consecutively in each SC until the stopping condition at the line 12 is satisfied.

3. Allocate a set $R_{b,s}$ of chunks to every SC b given the set $R^{(s)}$ from which they are selected and the required number of chunks per SC, $r'_{b,s}$. The selection is made according to the algorithm proposed in [45], which minimizes inter-cell interference. In this algorithm, the chunk allocation in a SC is performed so that the SC-to-SC distance between the given SC and the closest neighboring SC using the same chunk is the maximum possible. After the execution of the algorithm, the cardinality of the set $R_{b,s}$ for each SC should be equal to $r'_{b,s}$.

Applying S2 to a given slice leads to a more efficient resource usage than S1 since it fits better the spatial demands. However, in global terms, this efficiency could be small if other slices that are more loaded cannot share RBs with this slice. Such an effect is consequence of the high isolation level of S2, which also occurs in S1.

- **Strategy 3 (S3): Per-Cell Resource-Based Capacity Planning With Intra-Cell Isolation**: In S3, the per-cell definition of the requested value $v^{(s)}$ means that the number of chunks is the same for all the SCs. Consequently, this strategy is not adequate, like S1, to fit the spatial traffic variations. Unlike the previous strategies, the intra-cell isolation requirement in S3 enables that other slices can use the same chunk in a different SC, thus ensuring isolation between slices only within the area of a SC.

  The chunk allocation can be targeted to minimize co-channel interference by following the algorithm proposed in [45] or, alternatively, it can be oriented to maximize isolation by reusing the same chunks across the SCs if they are available. The former approach is assumed in this work since it has a

Algorithm 1: Calculation of the Number of Chunks for SCs

Inputs: $\hat{D}^{(b,s)}$, B_s, $v^{(s)}$;
Initialize $\hat{r}_{b,s} = 0$; $r'_{b,s} = 0$;
Compute $r'_{b,s}$, $b \in B_s$;
while $\sum_{b \in B_s} \hat{r}_{b,s} < v^{(s)}$ ***and*** $\sum_{b \in B_s} r'_{b,s} > 0$ **do**
  for $b \in B_s$ **do**
    if $r'_{b,s} > 0$ ***and*** $\sum_{b \in B_s} r'_{b,s} < v^{(s)}$ **then**
      $\hat{r}_{b,s} = \hat{r}_{b,s} + 1$;
      $r'_{b,s} = r'_{b,s} - 1$;
    end
  end
end
while $\sum_{b \in B_s} \hat{r}_{b,s} < v^{(s)}$ **do**
  for $b \in B_s$ **do**
    if $\sum_{b \in B_s} \sum_{b \in B_s} \hat{r}_{b,s} < v^{(s)}$ **then**
      $\hat{r}_{b,s} = \hat{r}_{b,s} + 1$;
    end
  end
end
return $\hat{r}_{b,s}$

better impact on the network performance. Consequently, the algorithm proposed in [45] is applied considering that, at each SC, the set of chunks belonging to other slices are not eligible. While in S2 the set of candidate chunks (given by $R^{(s)}$) is the same for all SCs, in S3 it may differ from SC to SC. Accordingly, $R^{(b,s)}$ represents the candidate chunks at each SC b for the slice s.

This strategy gives providers more flexibility to distribute RBs among slices since the required isolation level is lower. However, it is limited by the requirement of a fixed number of chunks per cell, which may lead to a suboptimal matching between demand and resources.

- **Strategy 4 (S4): Per-Network Resource-Based Capacity Planning With Intra-Cell Isolation**: In this case, the per-network definition of the requested value $v^{(s)}$ provides flexibility to adapt to the spatial traffic variations, i.e. the number of chunks per SC can vary to meet the particular

traffic demand at each cell. The only constraints are that each SC has at least one chunk allocated and that the total number of chunks allocated in the network (or cluster) is equal to $v(s)$, i.e. the condition in (7.12) is satisfied. Since isolation is only required within the cell area, the provider has even greater flexibility than S2 to perform the chunk allocation. Based on this, the process is composed of the following steps:

1. Determine the number of chunks $|R_{b,s}|$ to allocate at each SC based on the spatial demand distribution and the requested value $v^{(s)}$. This step is the same as the step 2 in S2.
2. Allocate a set $R_{b,s}$ of chunks to every SC b given the set $R^{(b,s)}$ from which they are selected and the required number of chunks per SC, $\hat{r}_{b,s}$, obtained in the step 1. To minimize the effect of inter-cell interference, the algorithm proposed in [45] is applied. It may happen that $|R_{b,s}|$ is lower than $\hat{r}_{b,s}$, meaning that the number of candidate chunks is not enough to reach the targeted value in a certain SC. In this case, the provider has enough flexibility to allocate the missing chunk(s) in a different SC without affecting the requested (per-network) value $v^{(s)}$.

A major advantage of this strategy is the greater flexibility to allocate RBs to slices with varying demands in a successful way. The gain will be larger as the number of slices planned with this strategy increases.

- **Strategy 5 (S5): Per-Cell Resource-Based Capacity Planning With No Isolation**: This strategy establishes a specific number of chunks per SC. Therefore, the number of chunks cannot be adapted to the cell demand at each SC. The main difference with the previous strategies is that, in this case, resource isolation is not mandatory. The providers can exploit this idea to leave room for other slices demanding resource isolation. In particular, those slices that do not require isolation will share RBs within the same SC. The selection of chunks is based on the algorithm proposed in [45] in order to minimize co-channel interference. The algorithm takes as input the set $R^{(b,s)}$ of candidate chunks at each SC. Among these chunks, the algorithm prioritizes the ones shared with other slices. However, it is

necessary to evaluate whether the SC is able to support or not the estimated $\hat{D}^{(b,s)}$. If not, the chunk is discarded from the candidate set.

Compared to the previous strategies, the S5 provides better resource usage as the traffic from different slices is aggregated into the same chunks. However, it is not optimal because the cell resources cannot be adapted to the spatial traffic variations. From the tenant's perspective, this strategy is suitable for services with no stringent requirements in terms of capacity or latency. In these cases, the tenant can achieve significant cost savings at the expense of greater uncertainty in performance due to the traffic load variations of other slices.

- **Strategy 6 (S6): Per-Network Resource-Based Capacity Planning With No Isolation**: The requested value in S6 is defined on a per-network basis. It means that the number of chunks at each SC can be adjusted to meet its traffic needs while guaranteeing a total number of chunks, $v^{(s)}$, allocated in the network. In addition, this strategy does not require resource isolation, which gives the provider the possibility to allocate the chunks being shared with other slices.

  The process behind S6 is similar to the strategies S2 and S4, which also define a network-wide capacity value. First, the number of chunks to allocate at each SC is calculated based on the spatial demand distribution. Second, a set of chunks is selected from a set of available chunks based on the algorithm proposed in [45], giving priority to the shared ones without causing overload. Note that, if the traffic of two slices is assumed highly correlated, enforcing S6 is similar to applying it to a single slice carrying the aggregated traffic of both slices.

  This strategy provides more efficient resource usage than S5 since the number of chunks at each SC can be fitted to the spatial traffic conditions. However, since no isolation is allowed, it retains the same disadvantages with regard to the impact of the traffic load variations of other slices.

7.5 Performance Evaluation

7.5.1 Simulation Scenario

In order to evaluate the proposed spectrum planning strategies, numerical examples and simulation results are presented in this section. The service area is 1.5 km x 1.5 km. It covers an urban environment with a set of deployed SCs. More specifically, the deployment scenario comprises the following: (i) for each slice, the statistical characterization of the traffic demand, which is non-uniformly distributed over the considered area and spatially cross-correlated with other slices; and (ii) a set of SCs deployed in the scenario according to the spatial variations of the aggregated traffic demand. The deployment scenario is simulated following a snapshot-based model, where each snapshot represents a random realization of the demand distribution. The different realizations of the same traffic probability distribution (i.e. varying the positions of the UEs) ensure reliable statistical significance analysis. The deployment scenario for 95% correlated demand between the slices A and B is shown in Fig. 7.3, where the triangles represent the location of the deployed SCs and the colored contour lines indicate the aggregated traffic demand density. As observed, the areas with higher traffic densities are provided with more SCs that are strategically located to serve the demand maximizing resource usage efficiency. The crosses in the figure represent the UE locations for a certain realization of the traffic probability distributions. The color of the crosses indicates the slice (A or B) to which the UE is connected. Finally, Table 7.1 summarizes the main parameters of the simulations. The requested value $v^{(s)}$ is calculated as a function of the considered RB occupancy in the network.

One issue regarding the implementation of the network model is that, due to the mutual interference between SCs, there is a dependency relation between the cell loads [54]. To reduce the computational complexity of the cell load factor [see (7.4) and (7.5)], the per-UE spectral efficiency SE_u is substituted by a cell-specific average value, which is taken from previous evaluations (i.e. from other snapshots). There is also a dependency relation between the $SINR(u, r)$ and $\Gamma(u)$, since the latter is based on the SINR to determine the serving SC. To simplify the procedure, the $SINR(u, r)$ is first estimated using $\Gamma(u)$ based on the strongest received power, which is calculated in (7.3). Then, the obtained values are used to compute $\Gamma(u)$ based on the SINR and, lastly, the $SINR(u, r)$.

Table 7.1: Simulation Parameters

Parameter	Configuration
Cellular Environment	Urban, 1.5 Km x 1.5 Km
Number of SCs	20
Operating Frequency	5 GHz
5G Numerology (μ)	0
System BW	120 MHz
Minimum allocation unit size (chunk)	20 MHz (106 RBs)
Propagation (path loss, shadowing)	UMi model [52]
SC antenna directivity	omni-directional
SC antenna height	6 m
UE antenna height	1.5 m
SC antenna gain	2 dBi
UE thermal noise	-174 dBm/Hz
UE noise figure	9 dB
Target SINR at cell-edge	9 dB [53]
UE minimum SINR	-10 dB [50]
UE maximum SINR	30 dB [50]
SC TX power range	[25-33] dBm
Number of UEs	250
UE service demand	5 Mbps
Resource scheduling scheme	Round-Robin
Number of slices	2
Proportion of UEs per slice	[50, 50] %
Traffic correlation between slices	95 %, 5 %
RB occupancy in the network	[50-100] %
Number of demand realizations	100

7.5.2 Performance Analysis of the Spectrum Planning Strategies

The first experiment provides a comparison between the different planning strategies regarding both network and service performance aspects. Specifically, the network performance is assessed using the cell overload factor and spectral efficiency metrics as defined in (7.6) and (7.8), respectively. The service performance is evaluated through the unsatisfied UE rate, which is defined as the fraction of UEs experiencing a throughput below the 25% of the UE service demand. The study is performed for two levels of correlation, 95% and 5%, between the traffic demands of the slices in the spatial domain. The high correlation value may represent slices that provide different services from the same tenant or the same service from different tenants. The low correlation value is more likely for slices owned by different tenants providing disparate services. To ensure a fair compar-

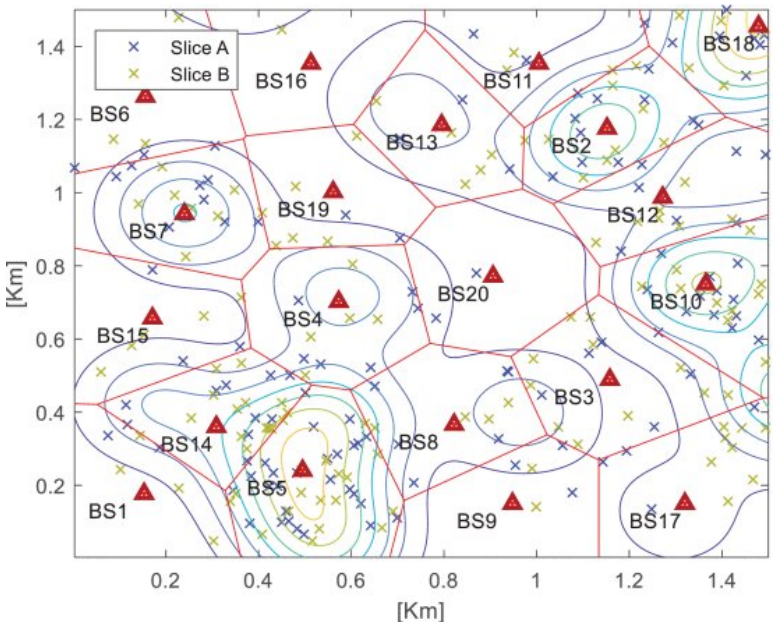
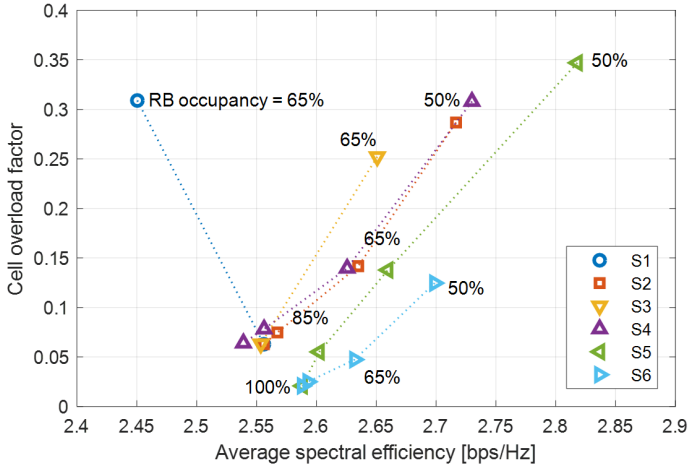


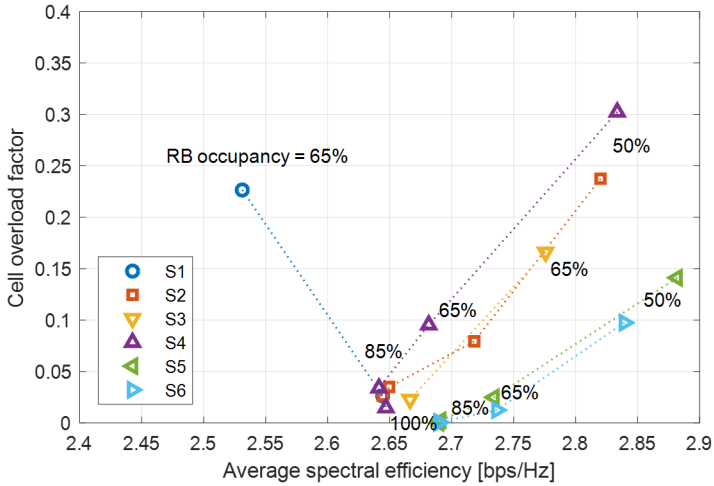
Figure 7.3: SC deployment with non-uniform traffic demand distribution and 95% correlated between slices.

ison, the strategies are evaluated for the same percentages of allocated spectrum chunks in the network. The RB occupancy affects network performance in the sense that, with high occupancy, the network will have more capacity, thus reducing the cell overload factor for the existing slices; however, it also reduces the possibilities to accept new slices, especially the ones that require resource isolation. The considered RB occupancy levels in the simulations are 50, 65, 85 and 100%, except for S1 and S3 for which the 50 and 85% values are not feasible. The latter is due to the simultaneous occurrence of the following factors: the proportion of UEs (and chunks) per slice is 50%; the specified number of chunks per cell must be an integer; and the 50% and 85% of the number of chunks in the system BW (i.e. three and five chunks, respectively) is not divisible by the number of slices.

The network performance metrics are represented against each other in Figs. 7.4(a) and 7.4(b) for the two correlation levels. The dotted lines connect the performance values for the different RB occupancy levels following a sequential order (50-65-85-100% for S2, S4, S5, S6 and 65-100% for S1, S3). The two



(a) 95% correlation between slices



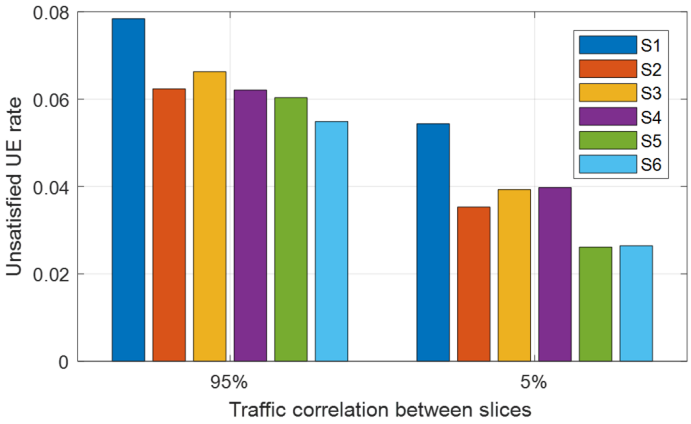
(b) 5% correlation between slices

Figure 7.4: Evaluation of network metrics for different percentages of chunk allocation and correlation levels between slices.

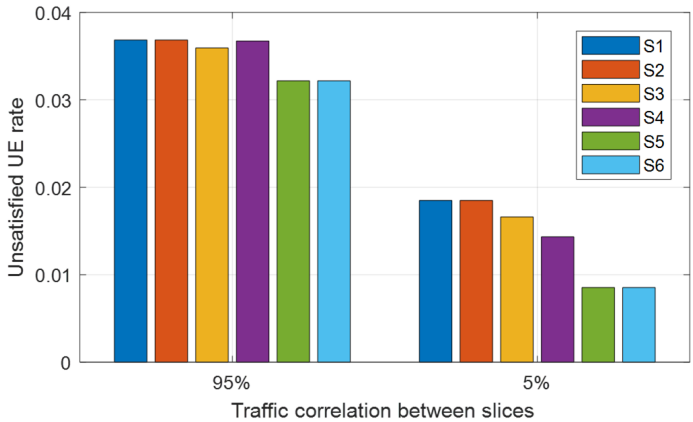
scenarios are evaluated with the same number of UEs; however, the values of the metrics are better with a lower correlation because the aggregated load of the two slices is more regularly distributed in the scenario. As observed, the trade-off between the cell overload factor and the average spectral efficiency is applicable to all the strategies except for S1 due to its poor matching between

traffic demand and network resources. In fact, such inefficiency is reflected in Fig. 7.4 by the extremely high values of the cell overload factor for a 65% of RB occupancy. The S3 also results in a high cell overload factor since the number of allocated chunks per cell cannot be adapted to the cell load. However, it provides a better average spectral efficiency than S1 because the S3 utilizes the entire system BW to allocate the 65% of chunks in the network, while the S1 only uses a fraction of the system BW with reuse-1. The strategies with no isolation, S5 and S6, provide the best trade-off between the two network metrics as their dotted lines in the figure are closer to the bottom right corner, especially for the case of a low demand correlation. This is a reasonable result since resource sharing typically leads to optimal network performance. However, as discussed later, the increased performance is at the expense of no isolation between slices. Behind these strategies, the S2 and S4 show good network performance while at the same time they provide resource isolation. Since the required capacity in these strategies is defined on a per-network basis, they enable some adaptation to the spatial traffic variations. Therefore, the resource usage is more efficient than in S1 and S3, where the slice specification entails a more rigid chunk allocation in the SCs. Such flexibility in allocating chunks is especially useful when the traffic demand of the slices is poorly correlated. Specifically, a slice can benefit from allocating additional chunks in overloaded areas where other slices are unloaded. The S2 and S4 achieve similar performance for the case of 95% correlated traffic, as shown in Fig. 7.4(a). However, in the case of 5% correlated traffic [see Fig. 7.4(b)] and high RB occupancy (above 85%), the S4 overcomes the performance of S2. In particular, for a 100% of RB occupancy, the S4 obtains a cell overload factor of 1.4%, while the S2 obtains 2.6%. Thus, the greater flexibility offered by the S4 due to the intra-cell isolation (as opposed to the inter-cell isolation of S2) entails an increased performance only if the number of allocated chunks in the network is sufficiently high.

Regarding service performance, Figs. 7.5(a) and 7.5(b) shows the mean of the unsatisfied UE rate obtained by each planning strategy for two correlation levels of the traffic demand, 95% and 5%, and two levels of RB occupancy, 65% and 100%. With 65% of RB occupancy, it is observed that the S1 results in the worst service performance level. However, with a 100% of RB occupancy [see Fig. 7.5(b)], the fraction of spectrum used by the S1 matches the system



(a) 65% RB occupancy



(b) 100% RB occupancy

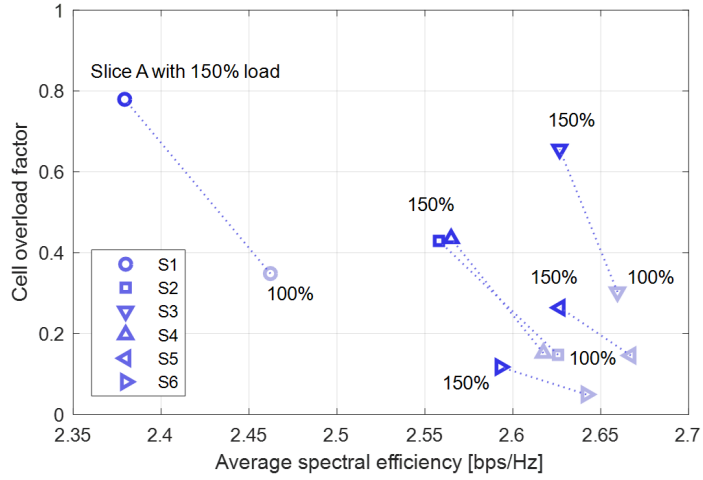
Figure 7.5: Evaluation of the unsatisfied UE rate for different percentages of chunk allocation and correlation levels between slices.

BW. Consequently, its unsatisfied UE rate is similar to other strategies. The S2, S3 and S4 provide similar unsatisfied UE rate, being better than S1 but worse than S5 and S6. With a 100% of RB occupancy and 5% of correlated traffic, the S4 provides better performance than S2 and S3 because, in this situation, the network further benefits from a more flexible chunk allocation. Lastly, the S5 and S6 result in the best performance level as they share chunks between slices. This gain is more pronounced with a 5% correlated traffic, reducing to the half the unsatisfied UE rate obtained by the S1.

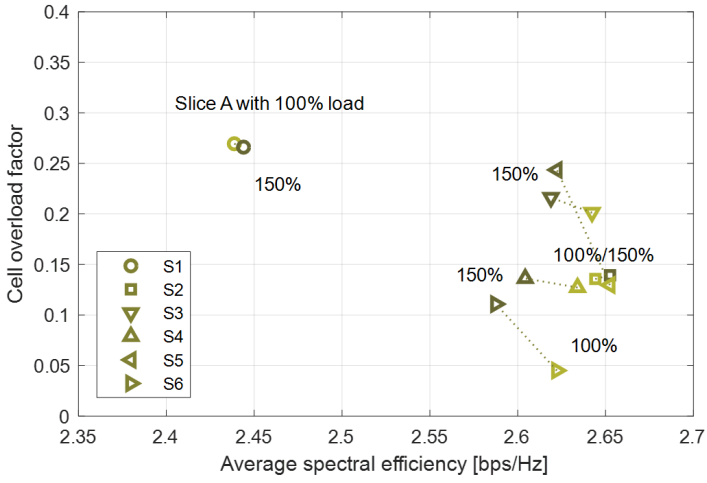
7.5.3 Analysis of the Impact on Resource Isolation

The following experiment evaluates the sensitivity of the performance metrics to the load variations of the slices. The scenario assumes a 95% of correlated traffic between slices and a 65% of RB occupancy. Initially, the network load is given by the configuration in Table 7.1, i.e. 250 UEs equally distributed between slices. After the initial stage, the load of the slice A increases by 50% (reaching 187 UEs), maintaining the same spatial distribution. The load of the slice B is not modified. Figs. 7.6(a) and 7.6(b) show the performance comparison between both situations concerning network metrics. The results are given per slice in order to highlight the impact of the increased load in the slice A on each slice separately. As observed in Fig. 7.6(a), the marks indicating the performance level with a 100% of load are shifted towards the upper left corner of the figure when such a load increases by 50%, meaning degradation of both metrics. Roughly, the cell overload factor is doubled in all the strategies except for the S2 and S4, where this increase is even greater. In Fig. 7.6(b), the results for the slice B depend on the degree of isolation enforced by the planning strategy. In the case of inter-cell isolation, the S1 and S2 maintain the same performance level before and after the load increase, thus providing full protection against the traffic variations of other slices. On the contrary, the other strategies are impacted to an extent that depends on the isolation level. The S3 and S4 result in a slight degradation since they perform intra-cell isolation. The S5 and S6 lead to a significantly worse performance, particularly, in terms of the cell overload factor, whose value is approximately doubled. As these strategies do not perform resource isolation, the impact of traffic variations from other slices is the greatest possible.

Fig. 7.7 represents the unsatisfied UE rate for the slice B before and after the load of the slice A increases by 50%. With a 100% of load, the results are in line with Fig. 7.6(b). Specifically, the S1 provides the worst service performance, while the S6 is slightly better than the other strategies. However, with a 150% of load, the degradation in the S5 and S6 leads to an unsatisfied UE rate that is similar to the performance of the S1, while the other strategies are hardly affected.



(a) Slice A



(b) Slice B

Figure 7.6: Evaluation of network metrics when the slice A increases its traffic demand by 50%. A 95% of correlated traffic between slices and a 65% of RB occupancy in the network are assumed.

7.5.4 Analysis of the Scalability of the Strategies

In this experiment, the impact of varying the chunk size is analyzed. To avoid a larger computational load, the system BW is not modified, maintaining the same value as in Table 7.1, i.e. 120 MHz. In return, the minimum allocation

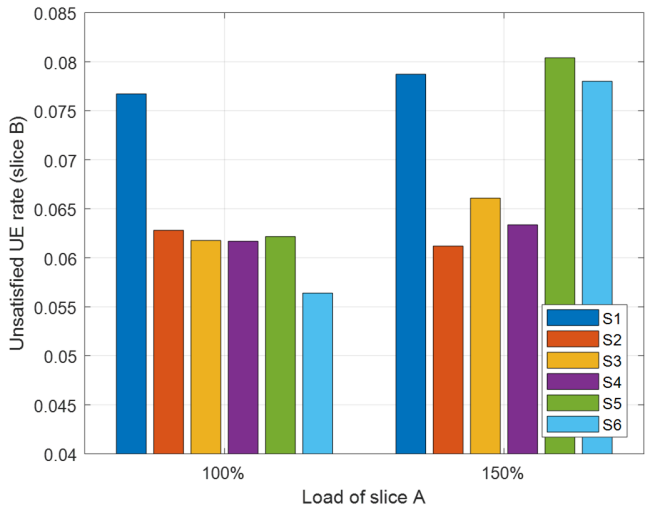


Figure 7.7: Evaluation of the unsatisfied UE rate for the slice B before and after the slice A increases its traffic demand by 50%.

unit size (i.e. the chunk size) is modified by sweeping the parameter through the following values: 60, 40, 30, 20, 15, 10 and 5 MHz, which are equivalent to 2, 3, 4, 6, 8, 12 and 24 chunks in the network. Each chunk comprises 324, 216, 160, 106, 79, 52 and 25 RBs, respectively. For example, the previous experiments are carried out with 20 MHz of minimum allocation unit size, which means 6 chunks for the system BW and 100 RBs per chunk. Depending on the planning strategy, some values of the minimum allocation unit size may not be feasible given the configuration in Table 7.1. In addition, the evaluated cases assume a 5% of correlated traffic between slices and a 65% of RB occupancy.

The results are shown in Fig. 7.8 with focus on the strategies for which the slice capacity is specified on a per-network basis. The dotted lines connect the performance values for the different minimum allocation unit sizes following a sequential order (5-10-15-20-30 MHz for S2, 5-10-15-20-30-40 MHz for S4 and 5-10-15-20-30-40-60 MHz for S6). Such strategies enable flexible chunk allocation, so that they can further benefit from increasing the number of chunks. It is observed that, as the number of chunks increases (i.e. the minimum allocation unit size decreases), the strategies saturate beyond a certain value. In particular, the cell overload factor saturates when the minimum allocation unit size is below 20 MHz. This value, used in the previous experiments, is applicable to the

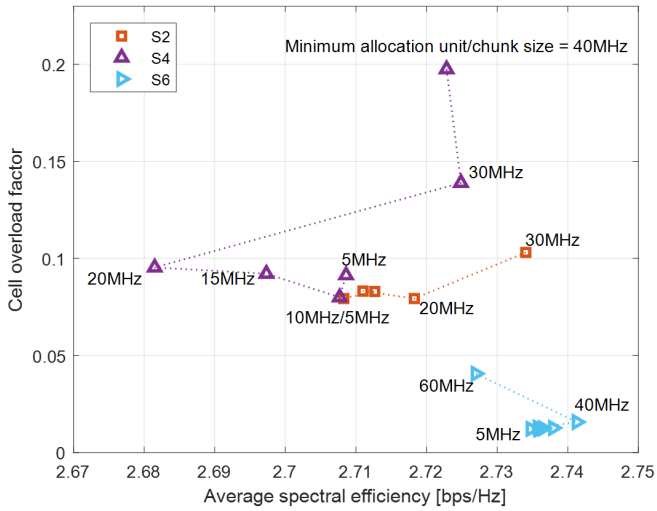


Figure 7.8: Evaluation of the impact of modifying the minimum allocation unit size for different planning strategies. A 5% of correlated traffic between slices and a 65% of RB occupancy in the network are assumed.

three planning strategies, as shown in Fig. 7.8. Since there is no substantial gain for chunk sizes below 20 MHz, this limit provides a good trade-off between performance and complexity. Lastly, it is also observed that the S4 is more sensitive to the variations of the chunk size than the other strategies.

7.6 Conclusions

This work has addressed the problem of spectrum planning for a 5G sliced network. To facilitate the transition to the new 5G paradigm, a business-driven model has been proposed to define tenant's requirements from a perspective of network resources. Then, following this model, different spectrum planning strategies for RAN slicing have been developed based on various levels of resource isolation and granularity. These strategies focus on the resource-oriented (as opposed to throughput-oriented) capacity specification. Each possible strategy gives the infrastructure provider different degrees of flexibility to allocate resources. By leveraging this flexibility, the provider can efficiently adapt the network resources to the traffic demands of the slices. The proposed spectrum planning strategies have been evaluated in a 5G sliced network of SCs through

snapshot-based simulations. The results show that the strategies with no isolation provide the best network performance due to a more efficient resource usage. The strategies based on a per-network capacity specification, as opposed to a per-cell definition, enable better adaptation to the spatial traffic variations, resulting in higher performance for low network resource occupancy and high traffic correlation between slices. The strategies with intra-cell or inter-cell isolation provide similar protection against inter-slice interference. However, for high resource occupancy and low traffic correlation, the intra-cell isolation results in better performance because of the greater flexibility for adapting resources to the slices' demands.

An interesting direction for future work extensions is the reallocation of resources in the network when a new slice request arrives. The new slice, based on its capacity specifications, may require some changes on the current resource allocation for its successful deployment. Moreover, when the slice becomes operative, the resources could also be dynamically allocated to cope with temporal variations of the traffic demands. AI techniques can also be applied to automate and optimize these tasks.

Acknowledgments

This work was supported in part by the H2020 Research and Innovation Project 5G-CLARITY under Grant 871428, in part by the Andalusian Knowledge Agency under Project A-TIC-241-UGR18, in part by the Spanish Ministry of Education, Culture and Sport (FPU) under Grant 17/01844, and in part by the Spanish Research Council and FEDER funds through SONAR 5G under Grant TEC2017-82651-R.

References

[1] R. N. Mitra and D. P. Agrawal, "5G mobile technology: A survey," *ICT express*, vol. 1, no. 3, pp. 132–137, 2015.

[2] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 32–39, 2016.

[3] 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 Wirel. Commun.*, vol. 24, no. 5, pp. 118–125, 2017.

[4] S. A. Kazmi, L. U. Khan, N. H. Tran, and C. S. Hong, *Network slicing for 5G and beyond networks*. Springer, 2019.

[5] P. Rost *et al.*, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, 2017.

[6] Y. Zhang, *Network Function Virtualization: Concepts and Applicability in 5G Networks*. John Wiley & Sons, 2018.

[7] 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 Wirel. Commun.*, vol. 26, no. 5, pp. 104–112, 2019.

[8] R. Su *et al.*, "Resource Allocation for Network Slicing in 5G Telecommunication Networks: A Survey of Principles and Models," *IEEE Netw.*, vol. 33, no. 6, pp. 172–179, 2019.

[9] A. Ksentini and N. Nikaein, "Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, 2017.

[10] 3GPP, "Feasibility Study on New Services and Markets Technology Enablers," 2015.

[11] 3GPP TS 38.300 V.16.6.0, "NR; NR and NG-RAN Overall description; Stage 2 (Release 15)," Dec. 2019.

[12] O. Bulakci and E. Pateromichelakis, "Slice-aware 5G Dynamic Small Cells," in *IEEE WCNC, Marrakech, Morocco*, pp. 1–6, 2019.

[13] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 3, pp. 2429–2453, 2018.

[14] 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," *Comput. Netw.*, vol. 167, p. 106984, 2020.

[15] 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 Commun. Mag.*, vol. 55, no. 5, pp. 80–87, 2017.

[16] D. A. Chekired, M. A. Togou, L. Khoukhi, and A. Ksentini, "5G-Slicing-Enabled Scalable SDN Core Network: Toward an Ultra-Low Latency of Autonomous Driving Service," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 8, pp. 1769–1782, 2019.

[17] L. Ma, X. Wen, L. Wang, Z. Lu, and R. Knopp, "An SDN/NFV based framework for management and deployment of service based 5G core network," *China Commun.*, vol. 15, no. 10, pp. 86–98, 2018.

[18] J. Ordonez-Lucena *et al.*, "The Creation Phase in Network Slicing: From a Service Order to an Operative Network Slice," in *EuCNC, Ljubljana, Slovenia*, pp. 1–36, 2018.

[19] K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini, and T. Braun, "Network Slices toward 5G Communications: Slicing the LTE Network," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 146–154, 2017.

[20] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5g ran slicing for verticals: Enablers and challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, 2019.

[21] P. L. Vo, M. N. H. Nguyen, T. A. Le, and N. H. Tran, "Slicing the Edge: Resource Allocation for RAN Network Slicing," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 970–973, 2018.

[22] J. Garcia-Morales, M. C. Lucas-Estañ, and J. Gozalvez, "Latency-Sensitive 5G RAN Slicing for Industry 4.0," *IEEE Access*, vol. 7, pp. 143139–143159, 2019.

[23] J. J. Escudero-Garzas, C. Bousoño-Calzon, and A. Garcia, "On the Feasibility of 5G Slice Resource Allocation With Spectral Efficiency: A Probabilistic Characterization," *IEEE Access*, vol. 7, pp. 151948–151961, 2019.

[24] S. D'Oro, F. Restuccia, A. Talamonti, and T. Melodia, "The Slice Is Served: Enforcing Radio Access Network Slicing in Virtualized 5G Systems," in *IEEE INFOCOM*, pp. 442–450, 2019.

[25] Z. Kotulski, T. W. Nowak, M. Sepczuk, and M. A. Tunia, "5G networks: Types of isolation and their parameters in RAN and CN slices," *Comput. Netw.*, vol. 171, p. 107135, 2020.

[26] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective," *IEEE Wirel. Commun.*, vol. 24, no. 5, pp. 166–174, 2017.

[27] T. Guo and A. Suárez, "Enabling 5G RAN Slicing With EDF Slice Scheduling," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2865–2877, 2019.

[28] D. Marabissi and R. Fantacci, "Highly Flexible RAN Slicing Approach to Manage Isolation, Priority, Efficiency," *IEEE Access*, vol. 7, pp. 97130–97142, 2019.

[29] Q. Ye, W. Zhuang, S. Zhang, A.-L. Jin, X. Shen, and X. Li, "Dynamic Radio Resource Slicing for a Two-Tier Heterogeneous Wireless Network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, 2018.

[30] S. O. Oladejo and O. E. Falowo, "5G network slicing: A multi-tenancy scenario," in *GWS, Cape Town, South Africa*, pp. 88–92, 2017.

[31] S. N. Khan, L. Goratti, R. Riggio, and S. Hasan, "On active, fine-grained RAN and spectrum sharing in multi-tenant 5G networks," in *IEEE PIMRC, Montreal, QC, Canada*, pp. 1–5, 2017.

[32] O. Al-Khatib, W. Hardjawana, and B. Vucetic, "Spectrum Sharing in Multi-Tenant 5G Cellular Networks: Modeling and Planning," *IEEE Access*, vol. 7, pp. 1602–1616, 2019.

[33] D. M. Gutierrez-Estevez *et al.*, "Artificial Intelligence for Elastic Management and Orchestration of 5G Networks," *IEEE Wirel. Commun.*, vol. 26, no. 5, pp. 134–141, 2019.

[34] V. Sciancalepore, X. Costa-Perez, and A. Banchs, "RL-NSB: Reinforcement Learning-Based 5G Network Slice Broker," *IEEE ACM Trans. Netw.*, vol. 27, no. 4, pp. 1543–1557, 2019.

[35] H. Xiang, S. Yan, and M. Peng, "A Realization of Fog-RAN Slicing via Deep Reinforcement Learning," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 4, pp. 2515–2527, 2020.

[36] G. Sun, Z. T. Gebrekidan, G. O. Boateng, D. Ayepah-Mensah, and W. Jiang, "Dynamic Reservation and Deep Reinforcement Learning Based Autonomous Resource Slicing for Virtualized Radio Access Networks," *IEEE Access*, vol. 7, pp. 45758–45772, 2019.

[37] X. Chen *et al.*, "Multi-Tenant Cross-Slice Resource Orchestration: A Deep Reinforcement Learning Approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2377–2392, 2019.

[38] R. Li *et al.*, "Deep Reinforcement Learning for Resource Management in Network Slicing," *IEEE Access*, vol. 6, pp. 74429–74441, 2018.

[39] A. Thantharate, R. Paropkari, V. Walunj, and C. Beard, "DeepSlice: A Deep Learning Approach towards an Efficient and Reliable Network Slicing in 5G Networks," in *IEEE UEMCON, Columbia University, New York, USA*, pp. 0762–0767, 2019.

[40] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Pérez, "Deepcog: Optimizing Resource Provisioning in Network Slicing With AI-Based Capacity Forecasting," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 361–376, 2020.

[41] B. Han, J. Lianghai, and H. D. Schotten, "Slice as an Evolutionary Service: Genetic Optimization for Inter-Slice Resource Management in 5G Networks," *IEEE Access*, vol. 6, pp. 33137–33147, 2018.

[42] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, 2018.

[43] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí, "On the configuration of radio resource management in a sliced RAN," in *IEEE/IFIP Network Operations and Management Symposium, Taipei, Taiwan*, pp. 1–6, 2018.

[44] J. Gang and V. Friderikos, "Inter-Tenant Resource Sharing and Power Allocation in 5G Virtual Networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7931–7943, 2019.

[45] P. Muñoz, O. Sallent, and J. Pérez-Romero, "Self-Dimensioning and Planning of Small Cell Capacity in Multitenant 5G Networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4552–4564, 2018.

[46] R. Ferrús, O. Sallent, J. Pérez-Romero, and R. Agusti, "On the Automation of RAN Slicing Provisioning and Cell Planning in NG-RAN," in *EuCNC, Ljubljana, Slovenia*, pp. 37–42, 2018.

[47] 3GPP TS 38.101-1 V.16.3.0, "User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (Release 16)," Mar. 2020.

[48] 3GPP TS 38.211 V.15.4.0, "NR; Physical channels and modulation (Release 15)," Dec. 2018.

[49] 3GPP TS 38.214 V.16.1.0, "NR, Physical layer procedures for data (Release 16)," Mar. 2020.

[50] 3GPP TS 38.803 V.14.2.0, "Study on New Radio Access Technology: Radio Frequency (RF) and co-Existance Aspects Version (Release 14)," Mar. 2017.

[51] GSM Association, "Generic Network Slice Template (Version 3.0)," May 2020.

[52] 3GPP TS 38.901 V.16.1.0, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 16)," Dec. 2019.

[53] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbp-s/UE in Cellular Systems: Understanding Ultra-Dense Small Cell Deploy-ments," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 4, pp. 2078–2101, 2015.

[54] I. Siomina and D. Yuan, "Analysis of cell load coupling for lte network planning and optimization," *IEEE Trans. Wirel. Commun.*, vol. 11, no. 6, pp. 2287–2297, 2012.

Chapter 8

Paper E. Analytical Model for the UE Blocking Probability in an OFDMA Cell providing GBR Slices

Authors:

Oscar Adamuz-Hinojosa, Pablo Ameigeiras, Pablo Munoz, Juan M. Lopez-Soler.

The paper has been published in the IEEE Wireless Communications and Networking Conference (WCNC), March, 2021.

Abstract

When a network operator designs strategies for planning and operating Guaranteed Bit Rate (GBR) slices, there are inherent issues such as the under(over)-provisioning of radio resources. To avoid them, modeling the User Equipment (UE) blocking probability in each cell is key. This task is challenging due to the total required bandwidth depends on the channel quality of each UE and the spatio-temporal variations in the number of UE sessions. Under this context, we propose an analytical model to evaluate the UE blocking probability in an Orthogonal Frequency-Division Multiple Access (OFDMA) cell. The main novelty of our model is the adoption of a multi-dimensional Erlang-B system which meets the reversibility property. This means our model is insensitive to the holding time distribution for the UE session. In addition, this property reduces the computational complexity of our model due to the solution for the state transition probabilities has product form. The provided results show that our model exhibits an estimation error for the UE blocking probability below 3.5%.

8.1 Introduction

Nowadays, the industry digitalization has boosted a wide variety of unprecedented services with stringent requirements. To economically provide them over a common infrastructure, network slicing has emerged as a solution [1]. Implemented as slices, most of these services are envisioned to rely on data transmissions with a strict Guaranteed Bit Rate (GBR) for each User Equipment (UE). When a network operator designs planning and operational strategies for GBR slices, it must consider the specific bandwidth consumption of each active UE per slice as well as the spatio-temporal variations in the number of UE sessions. Designing these strategies is challenging due to the bandwidth consumption of each UE is conditioned to its channel quality. With the aim of maintaining the Block Error Rate (BLER) for the UE's data below a certain threshold, the cells adopt Link Adaptation (LA). This technique enables each cell to adapt the UEs' Modulation and Coding Scheme (MCS) according to the experienced channel effects (i.e., path loss, shadowing, fast fading, inter-cell interference)[2].

To avoid inherent issues such as the under(over)-provisioning of radio resources, modeling the UE blocking probability in each cell is crucial for designing planning and operational strategies for GBR slices. Thereby, the network operator could decide the required number of cells, including their bandwidths, to deploy/scale GBR slices while the UE blocking probability is below a certain threshold.

Valuable models for evaluating the UE blocking probability have already been proposed in the literature (reported in Section 8.2). Most of them rely on Markov chains and queue theory. However, they are not appropriate for GBR slices due to they consider a variable rate per each UE. Additionally, some of these models are only valid for UE session durations following an exponential distribution, thus they cannot model the traffic behavior in real scenarios [3].

In this article, we focus on modeling the radio resource consumption of a GBR slice. Specifically, we have proposed an analytical model for assessing the UE blocking probability in a GBR slice for an Orthogonal Frequency-Division Multiple Access (OFDMA) cell under Poisson session arrivals. Using our model, the network operator can decide the number of radio resources required by a cell to provide a GBR slice while the UE blocking probability is below a given threshold. The main novelty of our model is the employment of a Multi-dimensional Erlang-B system which meets the reversibility property. It means our model allows the adoption of an arbitrary distribution for the UE session duration. Additionally, this property involves the solution for the state probabilities has product form, thus it eases their computation. Another innovation is the consideration of the average Signal-to-Interference-plus-Noise Ratio (SINR) for each UE. This allows a more precise characterization of the UEs' channel quality within the cell. The provided results show that our model exhibits an estimation error for the UE blocking probability below 3.5%.

The article is organized as follow. Section 8.2 summarizes the relevant literature. Section 8.3 presents the system model. In Section 8.4, we present the proposed Multi-dimensional Erlang-B model. This model is validated in Section 8.5. Finally, Section 8.6 draws the main conclusions and the future work.

8.2 Related Work

The existing literature for modeling the UE blocking probability in a cell is vast. In [4], the authors model a Code Division Multiple Access (CDMA)-High Data Rate (HDR) cell with a multi-class processor sharing queue. Since the processor sharing discipline is insensitive to the holding time distribution, arbitrary distributions can be adopted for the UE session duration. This discipline also forces an equal distribution of radio resources among the UEs, thus this model properly captures the behavior of Variable Bit Rate (VBR) services. In [5], this work was extended by including intra-cell UE mobility. However this improvement involves loosing the insensitivity property, thus only the exponential distribution is valid for the UE session duration.

In both works, the authors have considered concentric rings to model the channel quality distribution within the cell. This approximation easily enables the model to capture the behavior of LA techniques. For this reason, others authors have adopted this approach. In [6, 7, 8], the authors also focus on scenarios with intra-cell UE mobility but considering other medium access techniques such as Wideband Code Division Multiple Access (WCDMA) and OFDMA. Other works concentrate on non-3GPP access technologies. For instance, the authors of [9] define a model for WiMAX cells. This model is then used by a Quality of Service (QoS)-oriented resource allocation strategy for streaming flows that require a constant bit rate. Additionally, other authors consider services beyond mobile broadband. For example, the authors of [10, 11] adapt their models to IPTV services.

Despite these works present valuable contributions, the consideration of concentric rings limits the accuracy for modeling the distribution of the channel quality. For instance, two UEs located to the same distance from an access node could not perceive the same channel quality. The reason is there could be different obstacles and geographical features between each UE and the access node, involving a different impact of the channel effects such as shadowing or fast fading.

In an attempt to improve the model for the channel quality distribution, the authors of [12] consider a combination of indicators such as the RSRP, RSRQ, RSSI and SINR for each UE. Then, they use these indicators in a Markov chain which models the operation of a cell with intra-cell UE mobility. Notwithstanding,

the authors assume the reduction of the UE data rate when the total required bandwidth exceed the available bandwidth in the cell, thus this model is more appropriate for VBR services.

8.3 System Model

In this work, we focus on the downlink operation of one OFDMA cell. It provides a GBR service to their UEs, which dynamically request and release data sessions. This cell also supports LA, thus it must consider the channel quality perceived by each UE to allocate them radio resources. Based on this scenario, we first present the model for the cell. Then, we define the characteristics of the offered traffic. Finally, we describe the model for the radio resources.

8.3.1 Cell Model

To measure the channel quality within the cell, we adopt the SINR. This indicator depends on the radio environment and varies over time mainly due to (a) the path loss propagation, (b) shadowing, (c) fast fading and (d) inter-cell interference. In this work, we assume UEs have reduced mobility within the cell (e.g., semi-static people in live events such as sport events or concerts, IoT sensors, equipment for industry 4.0), thus (a)-(b) remain constant throughout the session duration while (c)-(d) vary over time. Let us define $\gamma_{u,n}^{(t)}$ as the instantaneous SINR measured by the UE u in the Resource Block (RB) n (see section 8.3.3) from the cell c. This parameter is provided by Eq. (8.1), where P_c^{RX} denotes the received power from the access node. This power results from the transmitted power less the attenuation suffered by the channel effects. L_j is the cell load factor and $\alpha_{j,n}$ is a function that takes the value 1 when the RB n is allocated to the neighbor cell j and the value 0 otherwise [1]. Note that the value for $\alpha_{j,n}$ will depend on the radio resource allocation algorithm implemented in each neighbor cell. Finally P_N is the noise power measured in one RB.

$$\gamma_{u,n}^{(t)} = \frac{P_c^{RX}}{\sum_{j \in \mathcal{C} \setminus \{c\}} L_j \alpha_{j,n} P_j^{RX} + P_N} \tag{8.1}$$

In this work, we focus on the average SINR for each UE $\overline{\gamma}_u$. To obtain this parameter, we average $\gamma_{u,n}^{(t)}$ over all the RBs consumed by the UE u, i.e.,

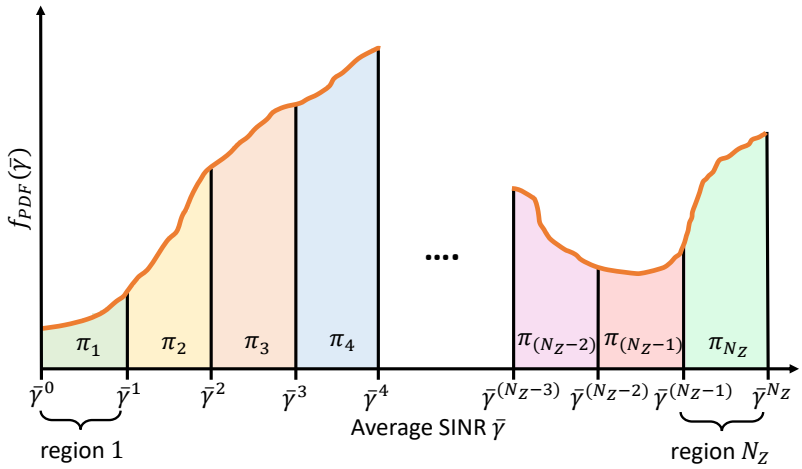


Figure 8.1: Splitting $f_{PDF}(\overline{\gamma})$ into N_Z regions

$\forall n \in \mathcal{N}_{RB}^{u,(t)}$, and the time period T_{ff} as Eq. (8.2) shows. This period is a time window over the fast fading is distinguishable. We also assume if $\overline{\gamma}_u$ is measured several times throughout the UE session, its remains constant. Note that the operator $|\cdot|$ denotes the cardinality of a set (i.e, the number of elements).

$$\overline{\gamma}_u = \frac{1}{T_{ff}} \int_{\tau}^{\tau+T_{ff}} \frac{1}{|\mathcal{N}_{RB}^{u,(t)}|} \sum_{n \in \mathcal{N}_{RB}^{u,(t)}} \gamma_{u,n}^{(t)} dt \tag{8.2}$$

Considering $\overline{\gamma}_u$ is measured for a considerable amount of UEs, we can derive the Probability Density Function (PDF) for the average SINR $f_{PDF}(\overline{\gamma})$ to model the channel quality within the cell. Since this feature could take a huge number of values, we split them into N_Z regions to make it treatable. Depicted in Fig. 8.1, each region z is defined as the set of values for the average SINR such as $\overline{\gamma} \in \left[\overline{\gamma}^{(z-1)}, \overline{\gamma}^{z}\right)$. For simplicity, we assume the session of an active UE takes place in one of these N_Z regions with probability π_z, which is provided by Eq. (8.3). Note that $\sum_{z=1}^{N_Z} \pi_z = 1$.

$$\pi_z = \int_{\overline{\gamma}^{(z-1)}}^{\overline{\gamma}^{z}} f_{PDF}(\overline{\gamma}) d\overline{\gamma} \tag{8.3}$$

Finally, we can also derive the average data rate per bandwidth unit $\overline{SE}_z$ (i.e., spectral efficiency) for each region z by using Eq. (8.4). The function

$f_{SINR \to SE}(\overline{\gamma})$ maps the SINR to the spectral efficiency under the assumption each UE achieves all the required radio resources. This function depends on the LA technique employed in the cell.

$$\overline{SE}_z = \int_{\overline{\gamma}^{(z-1)}}^{\overline{\gamma}^z} f_{SINR \to SE}(\overline{\gamma}) f_{PDF}(\overline{\gamma}) d\overline{\gamma} \tag{8.4}$$

8.3.2 Traffic Model

To model the traffic demands within a cell, we consider the statistical distributions and the average values for the arrival rate of UE sessions and the session duration.

For the arrival rate, we assume an average of λ UE session requests per unit time following a Poisson distribution. Since a Poisson process can be split into N_Z independent process[13], we can also express the average arrival rate for each region as $\lambda_z = \lambda \pi_z$. Note that $\lambda = \sum_{z=1}^{N_Z} \lambda_z$.

With respect to the session duration $t_{s,u}$ for each UE u, we assume a random variable extracted from an arbitrary distribution. Additionally, we define $\mu = 1/E[t_{s,u}]$ as the average rate for releasing UE sessions per unit time. Note that the release rate is independent from the region z.

8.3.3 Radio Resource Model

We assume an OFDMA cell with a total bandwidth W. This bandwidth is divided into N OFDM sub-carriers. In turn, these sub-carriers are grouped in groups of N_{SC} sub-carriers. Each group defines a RB, which is the smallest unit of resources that can be allocated to a UE. The number of available RBs on average during a slot is given by Eq. (8.5). The parameter Δf is the bandwidth between sub-carriers whereas OH denotes the overhead factor due to control plane data.

$$N_{RB}^{slot} = \left\lfloor \frac{W}{N_{SC} \Delta f} (1 - OH) \right\rfloor \tag{8.5}$$

Assuming all the UEs require an average data rate equal to the service GBR D_{GBR}, we need to compute the average number of RBs for each UE u in a time slot. Since we assume each UE is born in a specific region z, there exist only N_Z values for the average number of RBs required by a single UE within the cell.

These values are given by Eq. (8.6) and must satisfy Eq. (8.7) for any UE which is born in a region z (i.e., $\mathcal{U}^z$). In Eq. (8.7), $L_{n,u}$ denotes the specific amount of RBs allocated to the UE u in each time slot, whereas N^u_{slots} denotes the number of time slots during the session duration $t_{s,u}$. We assume $L_{n,u}$ is determined by a scheduler which aims to meet the GBR requirements of each UE.

$$N^{slot}_{RB,z} = \left\lceil \frac{D_{GBR}}{\overline{SE}_z N_{SC} \Delta f} \right\rceil \tag{8.6}$$

$$N^{slot}_{RB,z} = \frac{1}{N^u_{slots}} \sum_{n=1}^{N^u_{slots}} L_{n,u} \ \ \forall u \in \mathcal{U}^z \tag{8.7}$$

8.4 Capacity Model of an OFDMA Cell

This section explains the proposed model for an OFDMA cell, including the methodology used for deriving the UE blocking probability, the average RB utilization, and the cell capacity.

8.4.1 Multi-dimensional Erlang-B Model

Let us consider a cell where the values of $f_{PDF}(\overline{\gamma})$ are grouped into N_Z regions. To model this system we employ a multi-dimensional Erlang-B system. In this model, we assume each UE session takes place into one region z, defined by the tuple (λ_z, μ). The offered traffic intensity in each region becomes $\rho_z = \lambda_z/\mu$, and the total offered traffic intensity is $\rho = \sum_{z=1}^{N_Z} \rho_z$.

Let $s = (U_1, U_2, ..., U_{N_Z})$ denotes the state of the system, where U_z is the number of active UEs in the region z. To define the set of feasible states, we take into account (a) an active UE in the region z consumes $N^{slot}_{RB,z}$ RBs to meet its requirements; and (b) the available RBs in the cell are limited by N^{slot}_{RB}. These statements are gathered by Eq. (8.8), which provides the necessary condition to define a feasible state.

$$N^{slot}_{RB} - \sum_{z=1}^{N_Z} U_z N^{slot}_{RB,z} \geq 0 \ \ \forall s \tag{8.8}$$

Using this equation, we can build the state transition diagram as Fig. 2

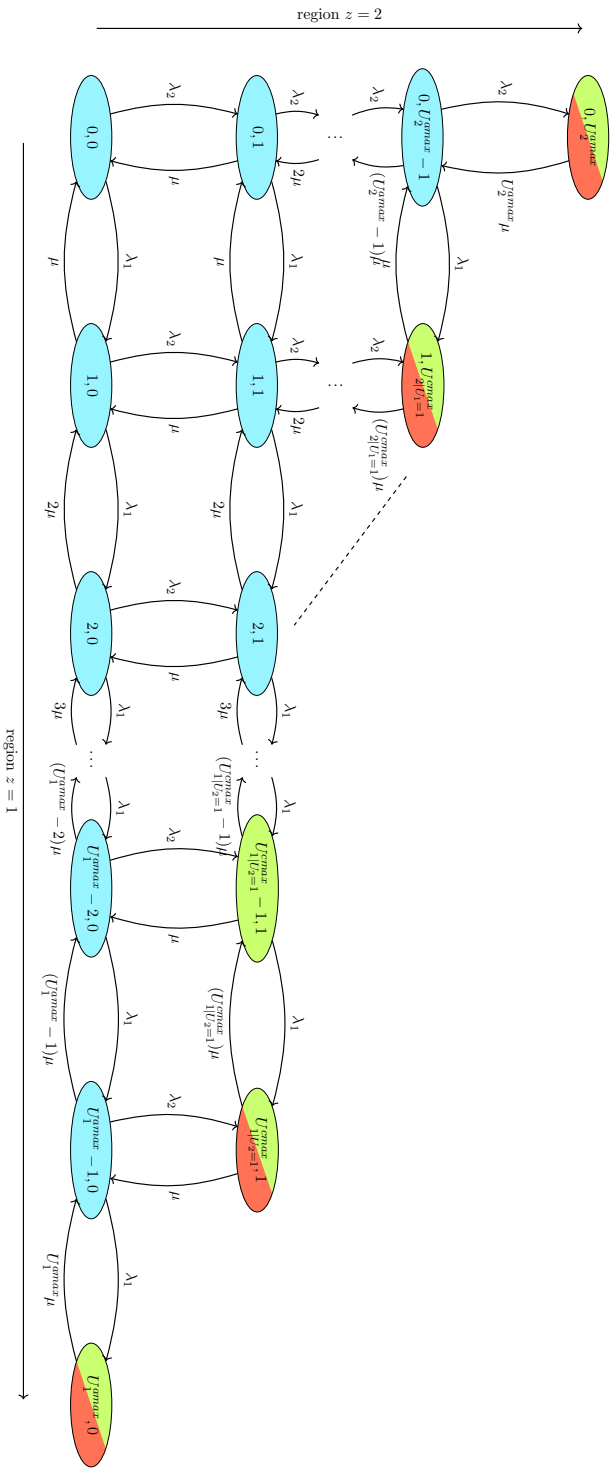


Figure 8.2: State transition diagram for a two-dimensional Erlang-B system. Note that red and green states correspond to $U_1 = U_{1|U_2}^{cmax}$ and $U_2 = U_{2|U_1}^{cmax}$, respectively.

shows. Note that for understandability purposes, the represented diagram only shows two dimensions. To define the upper bounds in each dimension, we use Eq. (8.9). In this equation, $U_{z|U_y}^{cmax}$ denotes the maximum number of UEs in the region z conditioned to the number of UEs in the remaining regions (e.g., red states for region 1, and green states for region 2). If the remaining regions have 0 UEs, we can define the absolute maximum number of UEs in region z as $U_z^{amax} = \left\lfloor \frac{N_{RB}^{slot}}{N_{RB,z}^{slot}} \right\rfloor$.

$$U_{z|U_y}^{cmax} = \left\lfloor \frac{N_{RB}^{slot} - \sum_{y \in \mathcal{Z} \setminus \{z\}} U_y N_{RB,y}^{slot}}{N_{RB,z}^{slot}} \right\rfloor \tag{8.9}$$

To clarify how the state transition diagram is built, Fig. 8.3 depicts a specific realization of the two-dimensional Erlang-B system presented in Fig. 8.2. In this example, each UE consumes $D_{GBR} = 2$ Mbps. Considering the average spectral efficiencies derived by Eq. (8.4) for each region are $\overline{SE}_1 = 2.778$ bps/Hz and $\overline{SE}_2 = 1.389$ bps/Hz, the required amount of RBs in each region are $N_{RB,1}^{slot} = 4$ and $N_{RB,2}^{slot} = 8$ (i.e., using Eq. (8.6) and considering $\Delta f = 15$ KHz and $N_{SC} = 12$ sub-carriers). If we assume the cell has available $N_{RB}^{slot} = 20$ RBs, we can check how Eq. (8.8) is met for all the states presented in Fig. 8.3. Furthermore, the absolute maximum number of UEs in each region (i.e., when the number of UEs in the remaining regions is 0) is $U_1^{amax} = 5$ UEs and $U_2^{amax} = 2$ UEs, respectively. Finally, focusing on the upper bounds in each dimension (i.e., red and/or green states), we can verify how Eq. (8.9) is met. For instance, considering region $z = 1$, the upper bounds (i.e., red states) are (a) $U_{1|U_2=0}^{cmax} = 5$ when $U_2 = 0$; (b) $U_{1|U_2=1}^{cmax} = 3$ when $U_2 = 1$; and (c) $U_{1|U_2=2}^{cmax} = 1$ when $U_2 = 2$.

The resulting multi-dimensional Erlang-B system corresponds to a reversible Markov process (see proof in appendix 8.6). This implies the proposed model is insensitive to the distribution of the UE session duration, which means the state probabilities depend only upon the mean service time [13]. Furthermore, the solution for the state probabilities has product form as Eq. (8.10) shows, where $p(U_z)$ is the one-dimensional truncated Poisson distribution for traffic stream in region z and K is a normalization constant.

$$\begin{aligned} p(U_1, U_2, ..., U_{N_Z}) =& K \cdot p(U_1) \cdot p(U_2) \cdot ... \cdot p(U_{N_Z}) \\ =& K \cdot \prod_{z=1}^{N_Z} \frac{\rho_z^{U_z}}{U_z!} \end{aligned} \tag{8.10}$$

To obtain the state probabilities, we need to derive K. This constant can be computed by summing all the state probabilities and equaling the resulting expression to 1, i.e., $\sum_{\forall s} p(U_1, U_2, ..., U_{N_Z}) = 1$. To that end, we recursively cover all the feasible states by using Eq. (8.11).

$$K^{-1} = \sum_{U_{N_Z}=0}^{U_{N_Z}^{amax}} \sum_{U_{N_Z-1}=0}^{U_{N_Z-1|U_{N_Z}}^{uplim}} ... \sum_{U_1=0}^{U_{1|U_{N_Z},U_{N_Z-1},...,U_2}^{uplim}} \left(\prod_{z=1}^{N_Z} \frac{\rho_z^{U_z}}{U_z!} \right) \tag{8.11}$$

In this equation, there is a summation per dimension with the aim of covering all the feasible states. The iterator of each summation is the number of UEs per a specific dimension, and it is bounded by the upper limit given in Eq. (8.12). This limit is an extension of $U_{z|U_y}^{cmax}$ (see Eq. (8.9)) which restricts its inter-dimensional dependence to those regions above z, i.e., the iterators of the outer summations. In this way, when the iterators of the outer summations increase, these summations will not cover those probabilities previously covered by the previous iterators. Considering the example provided in Fig. 8.3, this equation will iteratively cover the states of each row from the bottom to the top.

$$U_{z|U_{N_Z},U_{N_Z-1},...,U_x}^{uplim} = \left\lfloor \frac{N_{RB}^{slot} - \sum_{y=z+1}^{N_Z} U_y N_{RB,y}^{slot}}{N_{RB,z}^{slot}} \right\rfloor \quad \forall x > z \tag{8.12}$$

8.4.2 UE Blocking Probability

Assuming a new UE session is born in region z, it will be blocked if there not exists a transition from the current state $s^{(t)} = (U_1, U_2, \ldots, U_z, \ldots, U_{N_Z})$ to $s^{(t+1)} = (U_1, U_2, \ldots, U_z + 1, \ldots, U_{N_Z})$. This happens when $U_z + 1 > U^{cmax}_{z|U_y}$. In this way, the blocking states for each dimension are delimited by Eq. (8.9) (e.g., red states for region 1 in Fig. 8.3). If we iteratively cover the blocking states for each dimension, we can compute the UE blocking probability B_z conditioned to the region z where the new UE session is born by Eq. (8.13).

$$B_z = \sum_{U_{N_Z}=0}^{U^{amax}_{N_z}} \sum_{U_{N_Z-1}=0}^{U^{bplim}_{N_{Z-1}|z,U_{N_z}}} \cdots \sum_{U_1=0}^{U^{bplim}_{1|z,U_{N_z},U_{N_{z-1}},\ldots,U_2}} p\left(U_1, U_2, \ldots, U^{cmax}_{z|U_y}, \ldots, U_{N_z}\right) \quad (8.13)$$

This expression is composed by $N_Z - 1$ recursive summations, one per region excluding z. In turn, each summation is delimited by $U^{bplim}_{k|z,U_{N_z},\ldots,U_x}$, given by Eq. (8.14). The aim of this upper bound is similar to Eq. (8.12) with the difference the region z is not considered (i.e., U_z must be forced to $U^{cmax}_{z|U_y}$). Thereby, the summations only cover the blocking states in each iteration. The variable k denotes the specific region that a summation covers.

$$U^{bplim}_{k|z,U_{N_z},\ldots,U_x} = \left\lfloor \frac{N^{slot}_{RB} - \sum_{y>k, y\in\mathcal{Z}\setminus\{z\}} U_y N^{slot}_{RB,y}}{N^{slot}_{RB,k}} \right\rfloor \quad \forall x > k \quad (8.14)$$

Finally, the UE blocking probability in the cell is computed as the sum of the conditional blocking probabilities weighted by the probability of a UE session is born in each region (see Eq. (8.15)).

$$B = \sum_{z=1}^{N_Z} \pi_z B_z \quad (8.15)$$

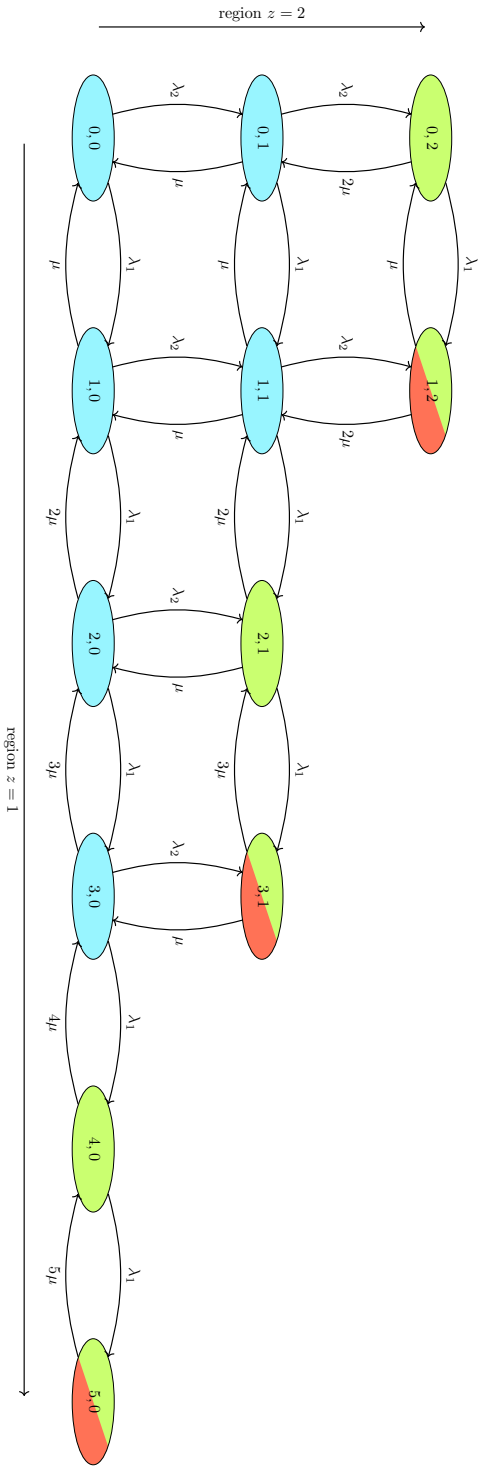


Figure 8.3: Specific realization of the state transition diagram for a two-dimensional Erlang-B system.

8.4.3 Mean Number of Consumed RBs and Cell Capacity

Another key parameters derived by our model are the mean number of RBs consumed in a cell $\overline{N}_{RB}$, and the cell capacity D_c.

The mean number of RBs can be computed as the average number of UEs $\overline{U}_z$ in each region z multiplied by the consumed RBs (see Eq. (8.16)). In turn, we compute $\overline{U}_z$ using Little's theorem[13], i.e., $\overline{U}_z = \lambda_z/\mu\,(1 - B_z) = \rho_z\,(1 - B_z)$.

$$\overline{N}_{RB} = \sum_{z=1}^{N_Z} \overline{U}_z N_{RB,z}^{slot} \tag{8.16}$$

The cell capacity is provided by Eq. (8.17). It is derived as the product of the mean number of UEs in each region multiplied by the data rate consumed by each UE.

$$D_c = \sum_{z=1}^{N_Z} \overline{U}_z D_{GBR} \tag{8.17}$$

8.5 Numerical Results

Since the state-of-the-art models for computing the UE blocking probability (see section 8.2) are not appropriate for GBR slices under our assumptions (i.e., reduced UE mobility and arbitrary distribution for the duration of the UE sessions), we cannot provide a fair comparison between these models and our model. For this reason, in this section we only focus on the validation of the proposed model. First, we present the experimental setup. Then, we analyze the aspects that impact on the execution time of our model. Finally, we evaluate the relative error for the UE blocking probability. Due to space limit, this section is focused on this parameter. Notwithstanding, we have also derived similar results for the mean number of consumed RBs and the cell capacity.

8.5.1 Experimental Setup

To validate the proposed model, we use a Matlab-based simulator that simulates the arrival and departure of UE sessions in a cell. This simulator receives the PDF for the average SINR $f_{PDF}(\overline{\gamma})$ as an input parameter. Table 8.1 summarizes the

Table 8.1: Configuration Parameters

Parameters	**Configuration**
Access Technology	5G NR
Sub-carrier Spacing Δf (OFDMA)	15 KHz
Sub-carriers per RB N_{SC} (OFDMA)	12
Access Node: Bandwidth W / Number of RBs N_{RB}^{slot}	10 MHz / 52 RBs; 15 MHz / 79 RBs; 20 MHz /106 RBs;
PDF average SINR in the cell: $f_{PDF}(\overline{\gamma})$	Built using a dataset from a live LTE network
Service GBR D_{GBR}	10 Mbps (e.g., on demand HD videostreaming)
PDF Regions	5, 9, 15 lineally spaced regions
Traffic Load ρ	From 0.2 to 1.5

configuration parameters. Regarding the access technology, we assume a Fifth Generation (5G)-New Radio (NR) cell implementing an OFDMA scheme with $\Delta f = 15$ KHz, and $N_{SC} = 12$. We also consider several cell bandwidths from 10 MHz to 20 MHz [14]. Additionally, the radio resource allocation in 5G-NR is carried out by multiples of 2, 4, 8 and 16 RBs [15]. In our evaluation, we consider each UE consumes multiples of 4 RBs, thus Eq. (8.6) was accordingly modified. With respect to $f_{PDF}(\overline{\gamma})$, we have derived it by using real dataset from a Long Term Evolution (LTE) network. Note that 5G dataset is not available due to the deployment of 5G networks are already in an early stage. Specifically, this dataset contained the probabilities of reporting a certain Channel Quality Indicator (CQI). This means we have directly used the Table 5.2.2.1-3 in [15] to map each probability into the spectral efficiency achieved by a specific CQI, i.e., this table correspond to the spectral efficiency provided by the 5G NR standard. For the GBR service, we have assumed a data rate of $D_{GBR} = 10$ Mbps.

Considering these configuration parameters, we have evaluated the UE blocking probability in function of the offered traffic intensity. Specifically from $\rho = 0.2$ to 1.5. Additionally, we have considered different values for the number of regions, from $N_Z = 5$ to 15. All the experiments have been carried out on a computer with 16 GB RAM and an Intel core i7-7700HQ @ 2.80 GHz.

8.5.2 Execution Time Evaluation

We have assessed the time complexity of our analytical model in two scenarios. In the former, we have covered several cell bandwidths, considering $N_Z = 9$. In the latter, we have considered different number of regions, with a cell bandwidth of 15 MHz (i.e., 79 RBs). The results for both scenarios are shown in Table 8.2. We observe the execution time grows exponentially with the number of regions and cell bandwidth. The reason is using higher values for both parameters involves an increment in the number of states in the Markov chain as Eq. (8.8) shows. Note that the execution time does not depend on the offered traffic intensity.

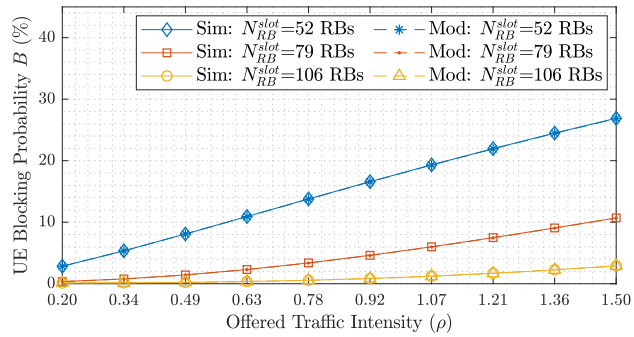
8.5.3 Model Validation

To validate our model, we have computed the relative error as $\epsilon_r(\%) = \frac{B_{sim} - B_{mod}}{B_{sim}} \cdot 100$, where B_{sim} and B_{mod} denote the UE blocking probability extracted from the model and simulator, respectively.

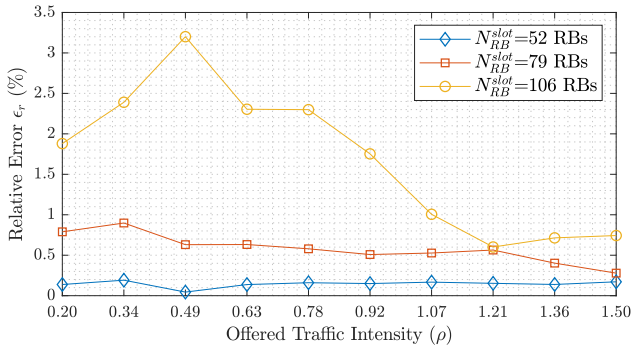
In Fig. 8.4(a), we depict the UE blocking probability derived from our model and the simulator. It shows how the UE blocking probability increases when (a) the available RBs in the cell are decreased and (b) the offered traffic intensity increases. This graph is useful for network operators because it allows to decide the bandwidth for each cell (i.e., N_{RB}^{slot}) while a threshold for B is provided, given certain conditions for the offered traffic intensity and the cell interference (i.e., a specific $f_{PDF}(\overline{\gamma})$). Due to the scale used for the vertical and horizontal axes in Fig. 8.4(a), the error between the simulation and the model cannot be observed. In Fig. 8.4(b), we represent the relative error, which is below 3.5 % for any case. We also notice this error is higher with higher cell bandwidths. This fact is induced by the simulator because it takes less samples for the highest states (e.g., blocking states) when the number of states increases. The reason is the probability of reaching the blocking states is lower, i.e., see Fig. 8.4(a), thus less

Table 8.2: Execution Time

$\boldsymbol{N_Z = 9}$			$\boldsymbol{N_{RB}^{slot} = 79}$		
$N_{RB}^{slot} = 52$	$N_{RB}^{slot} = 79$	$N_{RB}^{slot} = 106$	$N_Z = 5$	$N_Z = 9$	$N_Z = 15$
0.009 s	0.0036 s	0.171 s	0.006 s	0.032 s	0.343 s



(a) UE Blocking Probability: Model vs Simulation

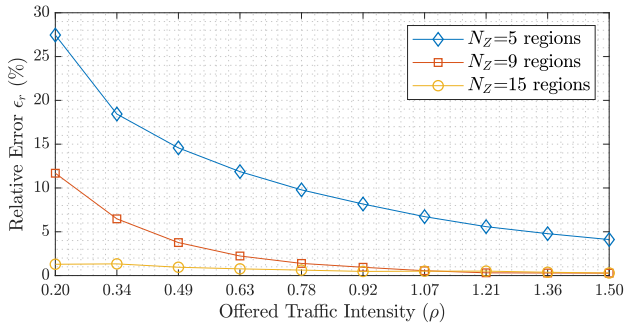


(b) Relative error

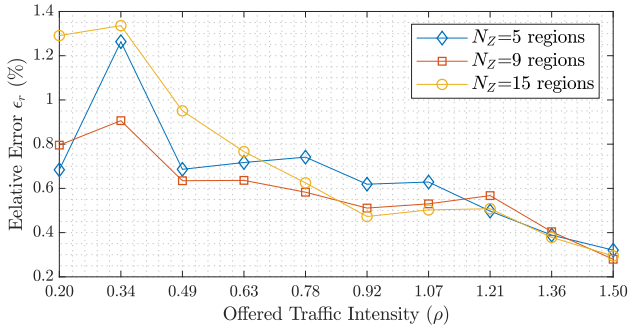
Figure 8.4: Evaluation of UE Blocking Probability for different cell bandwidths.

data samples for these states are taken during the simulation (i.e., number of times in a blocking state)

Finally, we evaluate the relative error when our model use a different number of regions. In 8.5(a), each case were compared with the simulator implementing $N_Z = 15$. We observe ϵ_r is below 1.5% when the model also implements $N_Z = 15$. However, this error increases when the number of regions decreases. This increment is not induced by our model but rather the fact of splitting the $f_{PDF}(\overline{\gamma})$ (see Fig. 8.1). When the number of regions is reduced and this number is small enough, each region could take a portion of the $f_{PDF}(\overline{\gamma})$ different from the case of using a higher number of regions. This results in a slightly different distribution of π_z, which involves an unfair comparison. In Fig. 8.5(b), we observe the relative error considerably decreases when each simulation implements the same number



(a) The simulator implements $N_Z = 15$



(b) The simulator implements the same number of regions as the model

Figure 8.5: Relative error in the evaluation of UE Blocking Probability for different number of regions

of regions as our model, proving it suits the simulator results.

8.6 Conclusions and Future Work

Network slicing is envisioned as a solution for providing emerging services over a common network infrastructure. Implemented as slices, most of these services will rely on data transmission with a strict GBR. Designing strategies for planning and operating GBR services could involve inherent issues such as the under(over)-provisioning of radio resources. To avoid that, it is crucial to model the UE blocking probability in each cell. Under this context, we propose an analytical model to evaluate this parameter. The main novelty is the consideration of a multi-dimensional Erlang-B system, which meets the reversibility property. This

means our model is valid for arbitrary distributions of the UE session duration. This property also reduces the computation complexity of the model due to the solution for the state probabilities has product form. Additionally, our proposal considers the PDF for the average SINR in the cell to model the distribution of the channel quality. The results show that our model exhibits an estimation error for the UE blocking probability below 3.5%.

Regarding the future work, several challenges lie ahead. One challenge is to include the effect of the scheduling discipline (e.g., proportional fair). Another challenge is considering services which simultaneously support VBR and GBR traffic.

Appendix A: Reversibility in a Markov Process

To proof the reversibility property of the proposed model, we follow the Kolmogorov cycle criteria [13]. This states that a necessary and sufficient condition for reversibility of a multi-dimensional Markov process is that for each dimension-pair, the circulation flow among four neighboring states in a square equals to zero (i.e., flow clockwise = flow counter-clockwise).

Considering four neighbor states from two arbitrary regions z and x (i.e., $z \neq x$): $s_1 = (U_1, ...U_z, ...U_x, ...U_{N_z})$, $s_2 = (U_1, ...U_z, ...U_x + 1, ...U_{N_z})$, $s_3 = (U_1, ...U_z + 1, ...U_x + 1, ...U_{N_z})$, and $s_4 = (U_1, ...U_z + 1, ...U_x, ...U_{N_z})$, we derive the clockwise and counter clockwise flows $f_{cw} = \lambda_x \cdot p_1 \cdot \lambda_z \cdot p_2 \cdot (U_x + 1)\mu \cdot p_3 \cdot (U_z + 1)\mu \cdot p_4$ and $f_{ccw} = \lambda_z \cdot p_1 \cdot \lambda_x \cdot p_4 \cdot (U_z + 1)\mu \cdot p_3 \cdot (U_x + 1)\mu \cdot p_4$, respectively. We denote p_y the probability of state s_y. If we compare both equations, we easily check that the clockwise and the counter clockwise flows are equal. Thus, the proposed Erlang-B model is reversible.

Acknowledgments

This work is partially supported by the H2020 research and innovation project 5G-CLARITY (Grant No. 871428); the Spanish Ministry of Economy and Competitiveness, the European Regional Development Fund (Project PID2019-108713RB-C53); and the Spanish Ministry of Education, Culture and Sport (FPU Grant 17/01844)

References

[1] P. Muñoz *et al.*, "Radio Access Network Slicing Strategies at Spectrum Planning Level in 5G and Beyond," *IEEE Access*, vol. 8, pp. 79604–79618, 2020.

[2] M. P. Mota *et al.*, "Adaptive Modulation and Coding Based on Reinforcement Learning for 5G Networks," *IEEE Globecom*, pp. 1–6, 2019.

[3] J. Navarro-Ortiz *et al.*, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Commun. Surveys Tuts*, vol. 22, no. 2, pp. 905–929, 2020.

[4] T. Bonald and A. Proutière, "Wireless Downlink Data Channels: User Performance and Cell Dimensioning," in *MobiCom, San Diego, California, USA*, pp. 339–352, 2003.

[5] T. Bonald *et al.*, "Flow-level performance and capacity of wireless networks with user mobility," *Queueing Systems*, vol. 63, no. 1-4, p. 131, 2009.

[6] S.-E. Elayoubi and T. Chahed, "Admission Control in the Downlink of WCDMA/UMTS," in *EuroNGI, Dagstuhl, Germany*, pp. 136–151, Springer, 2004.

[7] D. K. Kim *et al.*, "A novel ring-based performance analysis for call admission control in wireless networks," *IEEE Commun. Lett.*, vol. 14, no. 4, pp. 324–326, 2010.

[8] B. Sas *et al.*, "Modelling the time-varying cell capacity in LTE networks," *Telecommun Syst.*, vol. 55, no. 2, pp. 299–313, 2014.

[9] C. Tarhini and T. Chahed, "QoS-oriented resource allocation for streaming flows in IEEE802. 16e Mobile WiMAX," *Telecommun Syst.*, vol. 51, no. 1, pp. 65–71, 2012.

[10] A. Abdollahpouri and B. E. Wolfinger, "Measures to quantify the gain of multicast with application to IPTV transmissions via WiMAX networks," *Telecommun Syst.*, vol. 55, no. 2, pp. 185–198, 2014.

[11] M. Li, "Queueing Analysis of Unicast IPTV With Adaptive Modulation and Coding in Wireless Cellular Networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9241–9253, 2017.

[12] C. Kim *et al.*, "Mathematical Models for the Operation of a Cell With Bandwidth Sharing and Moving Users," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 744–755, 2020.

[13] V. B. Iversen, "Teletraffic engineering and network planning," 2015.

[14] 3GPP TS 38.101-1 V.16.3.0, "User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (Release 16)," Mar. 2020.

[15] 3GPP TS 38.214 V.16.1.0, "NR, Physical layer procedures for data (Release 16)," Mar. 2020.

Chapter 9

Paper F: UE Blocking Probability Model for 5G GBR Slices

Authors:

Oscar Adamuz-Hinojosa, Pablo Ameigeiras, Pablo Munoz, Juan M. Lopez-Soler.

The paper has been submitted to IEEE Transactions on Wireless Communications.

Abstract

Modeling the User Equipment (UE) blocking probability is key for planning the amount of radio resources required by a slice with Guaranteed Bit Rate (GBR) requirements. This task is challenging since the amount of these resources depends on the channel quality of each UE, the packet scheduler discipline, the number and locations of UEs, as well as the other cell interference. Under this context, we propose an analytical model to evaluate the UE blocking probability for a GBR slice in a 5G New Radio (NR) Orthogonal Frequency Division Multiple Access (OFDMA) cell. The main novelty of our model is the adoption of a multi-dimensional Erlang-B system which meets the reversibility property. This involves our model is insensitive to the holding time distribution for the UE sessions (i.e., the distribution can be arbitrary). The reversibility property also involves the state transition probabilities have product form, so that the computational complexity of our model is small. Our model admits as input any distribution for the average channel quality. Furthermore, it considers a channel-aware packet scheduler. We validate the proposed model by means of simulation, demonstrating an estimation error for the UE blocking probability below 1.5%.

9.1 Introduction

The emergence of Fifth Generation (5G) mobile networks will bring the digitalization of the industry verticals. It will boost a wide variety of unprecedented communication services with stringent requirements in terms of performance and functionalities [1]. Considering each communication service separately and building a Radio Access Network (RAN) accordingly would be unfeasible in terms of cost. RAN slicing is a technological solution to economically provide separate communication services over a common wireless infrastructure [2]. It consists of the provision of multiple logical networks, denominated RAN slices, each adapted to the requirements of a specific communication service. Implemented by RAN slices, some emerging communication services are envisioned to rely on data transmissions with strict Guaranteed Bit Rate (GBR) requirements for their User Equipments (UEs) [3].

When a Mobile Network Operator (MNO) performs activities for planning

in advance the deployment of multiple RAN slices with GBR requirements, it must consider the specific bandwidth consumption of each active UE session per RAN slice as well as the spatio-temporal variations in the number of UE sessions. Designing strategies for planning these RAN slices is challenging since the bandwidth consumption of each active UE session depends on (a) the UE channel quality; (b) the scheduler discipline of each serving cell, (c) the service guarantees in each RAN slice, and (c) the other cell interference.

To avoid the under(over)-provisioning of radio resources for each RAN slice, the MNO requires the design of accurate planning strategies. For instance, the under-provisioning of radio resources for a RAN slice could involve that a considerable amount of its UE sessions would be rejected since they would not be able to meet the GBR requirements. This means the probability of blocking UE sessions, hereinafter denominated as UE blocking probability, would be unacceptably high in one or more cells. To avoid that, it is crucial for the MNO to model the UE blocking probability for RAN slices with GBR requirements in each cell. Thereby, the MNO could properly carry out planning activities such as deciding the required number of cells, including their bandwidths, to deploy these RAN slices while the UE blocking probability for each RAN slice in each cell is below a certain threshold.

9.1.1 Related Works

The existing literature for modeling the UE blocking probability in a cell is vast. In [4], the authors provide an excellent model which is based on a multi-class processor sharing queue for a Code Division Multiple Access (CDMA)-High Data Rate cell. Since the processor sharing discipline is insensitive to the holding time distribution, arbitrary distributions can be adopted for the UE session duration. This discipline also forces an equal distribution of radio resources among the UEs, therefore this model is applicable to Round Robin and Proportional Fair (PF) scheduling disciplines. Considering these disciplines involve those UE sessions with better channel conditions could achieve a data rate equal or above the GBR, whereas those UE sessions with worse channel conditions could be rejected. This means this model is not appropriate for providing a GBR service to any UE regardless its channel conditions. In [5], this work was extended by

including intra-cell UE mobility. However this improvement involves loosing the insensitivity property, thus only the exponential distribution is valid for the UE session duration.

In both works, the authors build their model based on two scenarios: (a) a single cell in isolation, and (b) multiple cells following regular topologies. In the first scenario, the authors consider the intra-cell interference is constant over the considered cell. In the second scenario, the authors consider the cells are placed either equidistantly in an infinite line or following an hexagonal distribution. These assumptions easily enable their model to capture an approximate estimation of the average channel quality within the considered cell by using a discrete number of concentric rings. However, these assumptions limit the accuracy for modeling the distribution of the average channel quality. For instance, two UEs located at the same distance from an access node could not perceive the same channel quality. The reason is there could be different obstacles and geographical features between each UE and the serving access node, involving a different impact of the channel effects such as the shadowing.

Due to the simplicity for modeling the average channel quality by using concentric rings, other authors have also considered this approach in their models for the UE blocking probability (e.g., [6, 7, 8, 9, 10, 11]). In [6], the authors have proposed a Markov chain-based model to compute the UE blocking probability. The main novelty of this model is the consideration of UE mobility within a cell based on Wideband CDMA. Using this model, the paper proposes an admission control mechanism for UE sessions supporting voice and data calls with a minimum data rate. In [7], the authors use a ring-based model for modeling the coverage area of a base station which implements Orthogonal Frequency-Division Multiple Access (OFDMA). This model also considers the UE mobility within the cell as well as the effect of handovers from adjacent cells. Based on that, the model considers the equilibrium balance equations for the UEs in each ring to compute the UE blocking probability for a service with a constant data rate. In [8], the authors propose an analytical model to capture the time-varying capacity of an OFDMA cell. Based on a Markov chain, this model also considers concentric rings to capture the data rate achieved by each UE. Furthermore, their model captures the UE mobility within the cell by assuming an exponential distribution for moving an UE from one ring to another. Using this model, the authors design an ad-

mission control mechanism which considers the UE blocking probability. In [9], the authors propose a Quality of Service-oriented resource allocation strategy for streaming flows that require a constant bit rate in a WiMaX cell. The proposed strategy considers an analytical model to derive the UE blocking probability. This model is based on a Markov chain which assumes a set of concentric rings to model the channel quality within the cell. Additionally, all the previous models present valuable contributions which have been applied by others authors which go beyond traditional mobile broadband services. For instance, the authors of [10, 11] adapt their models to IPTV services which requires GBR requirements.

Despite these works contribute significantly to the state-of-the-art solutions for modeling the UE blocking probability for GBR services, they present some drawbacks as the assumption of an exponential distribution for the UE session duration and/or the limited modeling of the average channel quality within the cell.

In an attempt to consider a more precise characterization of the average channel quality within the cell, the authors of [12] consider several zones distinguished by the strength of the received signal. Specifically, they use a combination of indicators such as the Received Signal Received Power, the Received Signal Received Quality, the Received Signal Strength Indication and the Signal-to-Interference-plus-Noise Ratio (SINR) of each UE to compute the UE blocking probability in a scenario with intra-cell UE mobility. These indicators are taken as input for a Markov chain-based model. Despite this improvement, the authors assume a reduction of the data rate for each UE when the total required bandwidth exceed the available bandwidth in the cell, thus this model is not appropriate for a GBR service. Furthermore, this work assume the UE session duration follows an exponential distribution.

Finally, the state-of-the-art proposals which model the UE blocking probability for GBR services do not consider the impact on the channel gain when a packet scheduler dynamically allocates radio resources to the active UE sessions. In the literature, there exists a wide range of channel-aware strategies to schedule radio resources for GBR services. We recommend the readers to review the comprehensive survey presented in [13], where the most representative scheduling strategies for GBR services are analyzed.

9.1.2 Contributions

In this article, we assume a single OFDMA cell implements a RAN slice to provide a communication service with GBR requirements. Furthermore, this communication service does not present strict requirements in terms of latency. We also consider the UEs served by this RAN slice generate data sessions following a Poisson distribution. In addition, the cell implements a channel-aware scheduler to dynamically allocate radio resources to these UEs with the goal of satisfying their GBR requirements. Under this context, the main contributions of this article are:

- The proposal of an analytical model for assessing the UE blocking probability. This model is based on a Multi-dimensional Erlang-B system. It meets the reversibility property which means the proposed model allows the adoption of an arbitrary distribution for the UE session duration. Additionally, this property involves the solution for the state probabilities has product form, thus it reduces the complexity for their computation. Furthermore, the proposed model considers as input a generic distribution for the average SINR which an arbitrary UE session could perceive. Unlike the state-of-the-art proposals, this approach allows a more precise characterization of the UEs' channel quality within the cell. For example, our model may use as input a distribution of the average SINR obtained by either experimental measurements in a real network or simulation.

- The proposed model considers the cell implements a channel-aware scheduler to dynamically allocate the radio resources planned for this RAN slice to its active UE sessions. Specifically, we have provided the mathematical formulation which relates the GBR achieved by an active UE session and the number of allocated radio resources for such session. For comprehensibility purposes, we have considered a representative channel-aware scheduler based on the alpha-fair metric [14]. However, this does not detract the proposed model from considering any other channel-aware scheduler by following the proposed formulation.

A first version of this model was presented in our recent work [15], however we considered the cell implements a channel-agnostic scheduler. It means the previous version of the model did not capture the channel gain perceived by

a UE session if the scheduler had allocated to it those radio resources which provide it a better instantaneous SINR. In this paper, we take a crucial step towards a more realistic approach by considering the cell implements a channel-aware scheduler. This means the new version of the model considers this channel gain in the computation of the UE blocking probability.

In the provided results, we first validate the proposed multi-dimensional Erlang-B model by means of simulation, demonstrating that it exhibits an estimation error for the UE blocking probability below 1.5%. Then, based on the proposed model, we show the benefits of using a channel-aware scheduler to reduce the UE blocking probability.

The remainder of this article is organized as follow. Section 9.2 describes the system model. In Section 9.3, we present the proposed Multi-dimensional Erlang-B model. Then, we describe how the channel-aware scheduler meets the GBR requirements of each UE session in Section 9.4. In Section 9.5, we define the experimental setup and based on that, we validate the proposed model and evaluate the impact of using a channel-aware scheduler in the UE blocking probability. Finally, Section 9.6 draws the main conclusions.

9.2 System Model

In this work, we focus on the downlink operation of a single OFDMA cell. It implements a RAN slice providing a GBR service to their UEs, which dynamically request and release data sessions. This cell also implements a channel-aware scheduler which considers the channel quality perceived by the UEs to dynamically allocate them radio resources. Based on this scenario, we first describe the model for the radio resources in a OFDMA cell. Then, we define the channel model. Later, we present the characteristics of the offered traffic. Finally, we define the characteristic of the channel-aware scheduler.

9.2.1 Radio Resource Model

We assume a serving OFDMA cell $i \in \mathcal{I}$ with a total bandwidth W_i. This bandwidth is divided into N OFDM sub-carriers. In turn, these sub-carriers are arranged in groups of N_{SC} sub-carriers. Each group of sub-carriers defines a Resource Block (RB), which is the smallest unit of resources that can be allocated

to a single UE. The number of available RBs during a time slot is given by Eq. (9.1). The parameter Δf is the bandwidth between sub-carriers whereas OH denotes the overhead factor due to control plane data.

$$N_i^{RB} = \left\lfloor \frac{W_i}{N_{SC}\Delta f}(1 - OH) \right\rfloor \tag{9.1}$$

If the cell i employs a small sub-carrier spacing Δf and a large bandwidth W_i, the number of available RBs in a time slot could be too high. For instance, in a 5G-New Radio (NR) cell, the maximum number of available RBs could be 273 units [16] [17]. From the perspective of radio resource allocation, it becomes advantageous to reduce the management complexity by grouping the RBs into resource chunks, which are allocated to the UEs as indivisible units [18]. This can be done through the concept of Resource Block Group (RBG) defined in [19]. A RBG is a collection of consecutive RBs that can be allocated to a specific UE. The size of the RBG N_{size}^{RBG} (i.e., number of consecutive RBs) can be used for establishing the minimum allocation unit size. Increasing N_{size}^{RBG} may serve to reduce the signaling overhead at the expense of a loss of flexibility. Based on that, we can compute the available RBGs on a time slot in the cell i as $N_i^{RBG} = \lfloor N_i^{RB}/N_{size}^{RBG} \rfloor$.

9.2.2 Channel Model

In this work, we consider the UEs have reduced mobility within the cell (e.g., semi-static people in live events such as sport events or concerts, IoT sensors, or equipment for industry 4.0). This means we do not consider the time variation of the shadow fading and the path-loss for each UE. Based on such assumption, we adopt the SINR as the metric to measure the channel quality within the cell i.

Specifically, we define in Eq. (9.2) the instantaneous SINR $\gamma_{u,n}(m)$ for the UE u in the RBG n and the time slot m [20]. In Eq. (9.2), $X_{i,u,n}(m)$ denotes the fading component which is random and varies for each time slot m and RBG n. We adopt the well-known Rayleigh-fading model for the fading component $X_{i,u,n}(m)$, leading to an exponential distribution with unit mean. The parameters $\overline{P}_{i,u,n}$ and $\overline{P}_{j,u,n}$ denote the average received powers from the serving cell $i \in \mathcal{I}$ and the neighbor cell $j \in \mathcal{I}\backslash\{i\}$, respectively. The average received power

from a cell i depends on the path-loss $\overline{h}_{i,u}^{PL}$ and the shadowing $\overline{h}_{i,u}^{SH}$ as well as on the transmission power per RBG $P_{i,n}$, i.e., $\overline{P}_{i,u,n} = \overline{h}_{i,u}^{PL} \cdot \overline{h}_{i,u}^{SH} \cdot P_{i,n}$. We consider the cell transmits the same constant power for each RBG, i.e., $P_{i,n} = P_{i,m}$ $\forall n \in \mathcal{N}^{RBG}$, $\forall m \in \mathcal{N}^{RBG}$. Hereinafter, we omit the subscript n for the transmitted power and the average received power, i.e., P_i instead of $P_{i,n}$; and $\overline{P}_{i,u}$ instead of $\overline{P}_{i,u,n}$. Finally, P_N is the noise power measured in one RBG.

$$\gamma_{u,n}(m) = \frac{\overline{P}_{i,u,n} X_{i,u,n}(m)}{\sum_{j \in \mathcal{I} \setminus \{i\}} \overline{P}_{j,u,n} X_{j,u,n}(m) + P_N} \tag{9.2}$$

In addition to the instantaneous SINR, we consider the average SINR $\overline{\gamma}_{u,n}$ for each UE u and RBG n. To obtain this parameter, we average the instantaneous SINR $\gamma_{u,n}(m)$ in the last M time slots as Eq. (9.3) shows.

$$\overline{\gamma}_{u,n} = \frac{1}{M} \sum_{m=\tau-M}^{\tau-1} \gamma_{u,n}(m) \tag{9.3}$$

We also consider the number of interfering base stations goes to infinity and the average received interference power at UE u from each interference base station $j \in \mathcal{I} \setminus \{i\}$ is equally strong. Under such assumption, the instantaneous SINR $\gamma_{u,n}(t)$ follows an exponential distribution with mean $\tilde{\gamma}_u$ [20]. This results in the simplified Probability Density Function (PDF) $f_{\gamma_{u,n}}(\gamma)$ and Cumulative Distribution Function (CDF) $F_{\gamma_{u,n}}(\gamma)$ defined in Eqs. (9.4) and (9.5), respectively.

$$f_{\gamma_{u,n}}[\gamma] = \frac{1}{\tilde{\gamma}_u} \exp\left[\frac{-\gamma}{\tilde{\gamma}_u}\right] \tag{9.4}$$

$$F_{\gamma_{u,n}}[\gamma] = 1 - \exp\left[\frac{-\gamma}{\tilde{\gamma}_u}\right] \tag{9.5}$$

The parameter $\tilde{\gamma}_u$ is the SINR resulting from the average received power of the serving cell as well as the interfering powers omitting the fast-fading component as defined in Eq. (9.6). This means we are approximating the average SINR defined in Eq. (9.3) as $\overline{\gamma}_{u,n} \approx \tilde{\gamma}_u$. Since the average SINR does not depend on the RBG n, we omit the subscript n for such parameter in the remainder of this paper, i.e., $\overline{\gamma}_u$ instead of $\overline{\gamma}_{u,n}$.

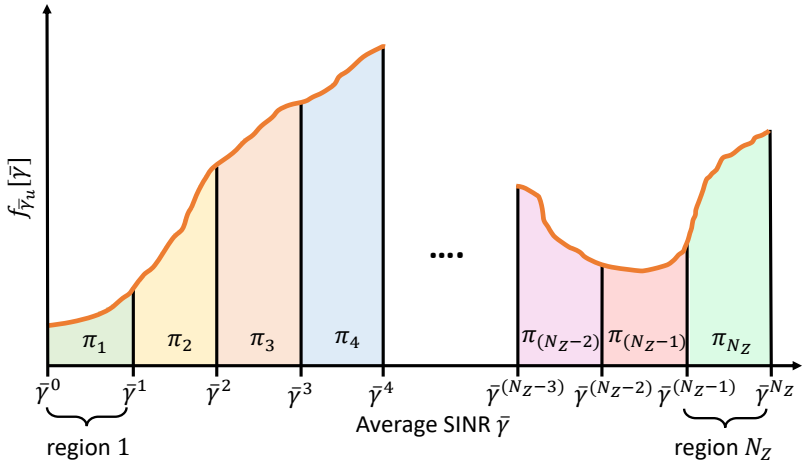


Figure 9.1: Splitting the PDF for the average SINR $f_{\overline{\gamma}_u}[\overline{\gamma}]$ into N_Z regions

$$\tilde{\gamma}_u = \frac{\overline{P}_{i,u}}{\sum_{j \in \mathcal{I} \setminus \{i\}} \overline{P}_{j,u}(t) + P_N} \tag{9.6}$$

Considering $\overline{\gamma}_u$ is measured for a considerable amount of active UEs, we can derive the PDF for the average SINR $f_{\overline{\gamma}_u}(\overline{\gamma})$ to model the average channel quality within the cell. Since this feature could take a huge number of values, we split them into N_Z regions to make it tractable. Depicted in Fig. 9.1, each region z is defined as the set of values for the average SINR such as $\overline{\gamma} \in \left[\overline{\gamma}^{(z-1)}, \overline{\gamma}^{z}\right)$. For simplicity, we assume the session of an active UE takes place in one of these N_Z regions with probability π_z, which is provided by Eq. (9.7). Note that $\sum_{z=1}^{N_Z} \pi_z = 1$. Furthermore, we assume the UE sessions which take place in a region $z \in \mathcal{Z}$ have an average SINR $\overline{\gamma}_z$. Finally, $\overline{\gamma}_z$ is the average of the average SINR $\overline{\gamma}$ within the interval $\left[\overline{\gamma}^{(z-1)}, \overline{\gamma}^{z}\right)$ as defined in Eq. (9.8).

$$\pi_z = \int_{\overline{\gamma}^{(z-1)}}^{\overline{\gamma}^{z}} f_{\overline{\gamma}_u}[\overline{\gamma}]\, d\overline{\gamma} \tag{9.7}$$

$$\overline{\gamma}_z = \frac{1}{\pi_z} \int_{\overline{\gamma}^{(z-1)}}^{\overline{\gamma}^{z}} \overline{\gamma} f_{\overline{\gamma}_u}[\overline{\gamma}]\, d\overline{\gamma} \tag{9.8}$$

9.2.3 Traffic Model

To model the traffic demands within a cell, we consider the statistical distributions and the average values for the arrival rate of UE sessions and the session duration.

For the arrival rate, we assume an average of λ UE session requests per unit time following a Poisson distribution. Since a Poisson process can be split into N_Z independent process[21], we can also express the average arrival rate for each region z as $\lambda_z = \lambda\pi_z$. Note that $\lambda = \sum_{z=1}^{N_Z} \lambda_z$.

With respect to the session duration $t_u^{session}$ for each UE u, we assume a random variable extracted from an arbitrary distribution. Additionally, we define $\mu = 1/E\left[t_u^{session}\right]$ as the average rate for releasing UE sessions per unit time. Note that the release session rate is independent from the region z.

9.2.4 Channel-Aware Scheduler

We focus on modeling the UE blocking probability for a RAN slice with GBR requirements considering the OFDMA cell implements a channel-aware scheduler. The analysis of which channel-aware scheduler provides a better performance for such RAN slice would deserve further investigation but is beyond the scope of this work. For this reason, we consider a representative channel-aware scheduler as the one depicted in Fig. 9.2. In the first step, the scheduler computes the metric $\hat{\gamma}_{u,n}$ defined in Eq. (9.9) for each UE $u \in \mathcal{U}$ in each RBG $n \in \mathcal{N}^{RBG}$. In Eq. (9.9), $\hat{\gamma}_{u,n}$ is based on the metric proposed in [22] and analyzed in [14]. For simplicity, we specifically consider the SINR instead of the data rate for an UE u. In this metric, the parameter z_u defines the region where the data session of the UE u was born. The fairness factor α is an adjustable parameter that controls the fairness of the RBG allocation.

$$\hat{\gamma}_{u,n} = \frac{\gamma_{u,n}}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}} \tag{9.9}$$

In this work we consider all the UE sessions require a GBR equal to D_{GBR}. However, the channel conditions for each UE session are different. This means each UE session needs a specific amount of RBGs to meet its GBR requirements. Note that if the MNO set $\alpha = 1.0$, the scheduler would implement the PF criteria.

It would involve all the UEs sessions have the same probability of being allocated with a RBG [23]. In this case, those UEs which perceived a low average SINR would have a lower data rate (it could be even below the GBR) than those UEs which perceived a higher average SINR. To avoid that, the MNO could set $\alpha > 1.0$ with the goal of increasing the amount of allocated RBGs to those UE which perceives a lower average SINR. In Section 9.5, we show the benefits of considering this approach.

Based on this metric, the scheduler decides in the second step to pre-allocate each RBG n to the UE u with the greatest metric $\hat{\gamma}_{u,n}$ as Eq. (9.10) shows. In Eq. (9.10), $S_{u,n}$ is a binary variable which defines with a value 1 if the RBG n is pre-allocated to the UE u and 0 otherwise. Note that $\sum_{u \in \mathcal{U}} S_{u,n} = 1$, i.e., the RBG n is pre-allocated to one UE. Furthermore, $\sum_{u \in \mathcal{U}} \sum_{n=1}^{N_i^{RBG}} S_{u,n} = N_i^{RBG}$, i.e., all the available RBGs in the cell are pre-allocated to the set of UEs $\mathcal{U}$.

$$S_{u,n} = \begin{cases} 1 & \hat{\gamma}_{u,n} \geq \max\limits_{\forall v \in \mathcal{U} \setminus \{u\}} \{\hat{\gamma}_{v,n}\} \\ 0 & \text{otherwise} \end{cases} \tag{9.10}$$

Despite the use of the fairness factor α in Eq. (9.9), this pre-allocation does not guarantee the UEs achieve a data rate equal to the GBR. Since we focus on a scenario where all the UEs only get the GBR, we consider the channel-aware

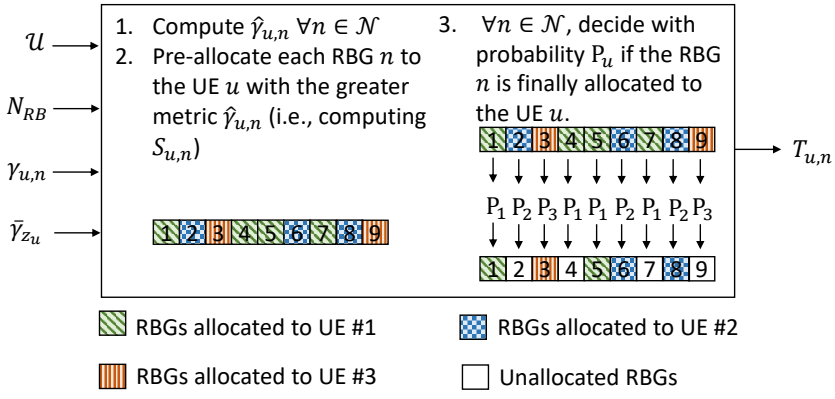


Figure 9.2: Tasks performed by the channel-aware scheduler during the time slot m. For simplicity, we show an illustrative example with three UEs and nine RBGs.

scheduler limits the amount of RBGs for those UEs with an achievable data rate above the GBR. For this reason, this scheduler decides in the third step which percentage P_u of RBGs are finally allocated for each UE u. Based on that, we define the probability $\mathrm{P}\left[T_{u,n}=1\right]$ that the RBG n is finally allocated to the UE u in Eq. (9.11). In this equation, the parameter $T_{u,n}$ is a binary variable which takes the value 1 if the RBG n is finally allocated to the UE u and 0 otherwise. $\mathrm{P}\left[S_{u,n}=1\right]$ denotes the probability that the RBG n is pre-allocated in the second step to the UE u. Note that P_u and $\mathrm{P}\left[S_{u,n}=1\right]$ are independent. Furthermore $\sum_{u\in\mathcal{U}} T_{u,n} \leq 1$, i.e., the RBG n could be (or not) allocated to one UE. In addition, $\sum_{u\in\mathcal{U}} \sum_{n=1}^{N_i^{RBG}} T_{u,n} \leq N_i^{RBG}$, i.e., all the available RBs may not be scheduled in a time slot.

$$\mathrm{P}\left[T_{u,n}=1\right] = \mathrm{P}_u \mathrm{P}\left[S_{u,n}=1\right] \tag{9.11}$$

To derive $\mathrm{P}\left[S_{u,n}=1\right]$, we perform the operations described in Eq. (9.12). The resulting expression depends on (a) the PDF of the instantaneous SINR for the UE u; and (b) the CDFs of the instantaneous SINR for the remaining UEs [24].

$$\begin{aligned}
\mathrm{P}\left[S_{u,n}=1\right] &= \mathrm{P}\left[\hat{\gamma}_{u,n} \geq \max_{\forall v\in\mathcal{U}\setminus\{u\}} \left\{\hat{\gamma}_{v,n}\right\} | \gamma_{u,n}=\gamma\right] \\
&= \int_0^\infty f_{\hat{\gamma}_{u,n}}\left[\gamma\right] \prod_{\forall v\in\mathcal{U}\setminus\{u\}} F_{\hat{\gamma}_{v,n}}\left[\gamma\right] d\gamma \\
&= \int_0^\infty \left(\overline{\gamma}_{z_u}\right)^\alpha f_{\gamma_{u,n}}\left[\left(\overline{\gamma}_{z_u}\right)^\alpha \cdot \gamma\right] \prod_{\forall v\in\mathcal{U}\setminus\{u\}} F_{\gamma_{v,n}}\left[\left(\overline{\gamma}_{z_v}\right)^\alpha \cdot \gamma\right] d\gamma
\end{aligned} \tag{9.12}$$

Since all the UE sessions are distributed into N_z regions where each region gathers U_z UE sessions (i.e., $U_1+U_2+\ldots+U_{N_z}=|\mathcal{U}|$), we can rewrite $\mathrm{P}\left[S_{u,n}=1\right]$ as Eq. (9.13) shows.

$$\begin{aligned}
\mathrm{P}\left[S_{u,n}=1\right] = \int_0^\infty &\left(\overline{\gamma}_{z_u}\right)^\alpha f_{\gamma_{u,n}}\left[\left(\overline{\gamma}_{z_u}\right)^\alpha \cdot \gamma\right] \cdot \left(F_{\gamma_{1,n}}\left[\left(\overline{\gamma}_1\right)^\alpha \cdot \gamma\right]\right)^{U_1} \cdot \left(F_{\gamma_{2,n}}\left[\left(\overline{\gamma}_2\right)^\alpha \cdot \gamma\right]\right)^{U_2} \ldots \\
&\cdot \left(F_{\gamma_{u,n}}\left[\left(\overline{\gamma}_{z_u}\right)^\alpha \cdot \gamma\right]\right)^{U_z-1} \cdot \ldots \cdot \left(F_{\gamma_{N_z,n}}\left[\left(\overline{\gamma}_{N_z}\right)^\alpha \cdot \gamma\right]\right)^{U_{N_z}} d\gamma
\end{aligned} \tag{9.13}$$

Considering the PDF and the CDFs of the instantaneous SINR defined in Eqs. (9.4) and (9.5), we can rewrite $\mathrm{P}\left[S_{u,n}=1\right]$ as Eq. (9.14) shows. The definite integral described in this equation has an analytical solution which depends on the specific value of the fairness factor α and the number of active UE sessions in each region (i.e., U_1, U_2, ..., U_{N_z}). Note that we have removed the subscript n since the PDF and the CDFs do not depend on the RBG n. Furthermore, we have replaced the subscript u with z as the computed probability is the same for all the UE whose data session has been born in the region z. This also involves that P_u can be redefined as P_z and thus, the Eq. (9.11) can be redefined as $\mathrm{P}\left[T_z=1\right]=\mathrm{P}_z\mathrm{P}\left[S_z=1\right]$.

$$\begin{aligned}\mathrm{P}\left[S_z=1\right]=\int_0^{\infty}(\overline{\gamma}_z)^{\alpha-1}\exp\left[-(\overline{\gamma}_z)^{\alpha-1}\cdot\gamma\right]\cdot\left(1-\exp\left[-(\overline{\gamma}_1)^{\alpha-1}\cdot\gamma\right]\right)^{U_1}\\ \cdot\left(1-\exp\left[-(\overline{\gamma}_2)^{\alpha-1}\cdot\gamma\right]\right)^{U_2}\cdot\ ...\ \cdot\left(1-\exp\left[-(\overline{\gamma}_z)^{\alpha-1}\cdot\gamma\right]\right)^{U_z-1}\cdot\ ...\\ \cdot\left(1-\exp\left[-\left(\overline{\gamma}_{N_z}\right)^{\alpha-1}\cdot\gamma\right]\right)^{U_{N_z}}d\gamma\end{aligned}\quad(9.14)$$

We need to compute the average number of RBGs for each UE u in a time slot. Since we assume each UE session is born in a specific region z of the cell i, the average number of RBGs for a single UE depends on the scheduling probability $\mathrm{P}\left[T_z=1\right]$. In Eq. (9.15), we see $\mathrm{P}\left[T_z=1\right]$ depends on the state $s=(U_1,U_2,\ldots,U_{N_Z})\in\mathcal{S}$. This state defines the total amount of active UE sessions as well as their distribution along the N_z regions. $\mathcal{S}$ denotes the set of potential states of the system.

$$N_{z,s}^{RBG}=\left\lceil\mathrm{P}\left[T_z=1\right]N_i^{RBG}\right\rceil\quad(9.15)$$

In addition, the average number of RBGs obtained by a single UE session which is born in region z must satisfy Eq. (9.16) in which T_{slots}^{u} denotes the number of time slots throughout the session duration.

$$N_{z,s}^{RBG}=\frac{1}{T_{slots}^{u}}\sum_{t=1}^{T_{slots}^{u}}\sum_{n=1}^{N_i^{RBG}}T_{u,n}^{(t)}\ \ \forall u\in\mathcal{U}^z\quad(9.16)$$

9.3 UE Blocking Probability and Capacity Model of an OFDMA Cell

This section explains the proposed model for an OFDMA cell, including the methodology to derive the UE blocking probability, the average RBG utilization, and the cell capacity for a RAN slice with GBR requirements.

9.3.1 Multi-dimensional Erlang-B Model

Let us consider a cell where the values of $f_{\overline{\gamma}_u}[\overline{\gamma}]$ for a single RAN slice are grouped into N_Z regions. To model this system, we employ a multi-dimensional Erlang-B model. In this model, we assume each UE session takes place into one region z, defined by the tuple (λ_z, μ). The offered traffic intensity in each region becomes $\rho_z = \lambda_z/\mu$, and the total offered traffic intensity is $\rho = \sum_{z=1}^{N_Z} \rho_z$.

To define the set of potential states of the system $\mathcal{S}$, we take into account (a) an active UE session in the region z consumes $N_{z,s}^{RBG}$ RBGs on average; and (b) the available RBGs in the cell are limited by N_i^{RBG}. These statements are gathered by Eq. (9.17), which provides the necessary condition to define a state $s \in \mathcal{S}$.

$$N_i^{RBG} - \sum_{z=1}^{N_Z} U_z N_{z,s}^{RBG} \geq 0 \ \ \forall s \in \mathcal{S} \tag{9.17}$$

Focusing on a single state $s \in \mathcal{S}$, it could happen that the GBR requirements could not be met for the UE sessions which fall within one or more regions, i.e., their average data rates could be less than D_{GBR}. This means the state s is not valid and thus, one or more UE sessions should be rejected. For this reason, we define the set $\mathcal{S}' = \mathcal{S} \setminus \mathcal{S}^{nf}$ of feasible states. $\mathcal{S}^{nf}$ denotes those states belonging to $\mathcal{S}$ where at least one UE session does not meet the GBR requirements. In section 9.4, we provide details about determining a valid state $s \in \mathcal{S}'$, i.e., that one in which all the active UE sessions meet the GBR requirements.

Considering the set $\mathcal{S}'$ of feasible states, we can build the state transition diagram as Fig. 9.3 shows. Note that for simplicity, the represented diagram only shows two dimensions, corresponding to the regions $z = 1$ and $z = 2$. In this diagram, $U_{z|U_y}^{cmax}$ denotes the maximum number of UE sessions in the

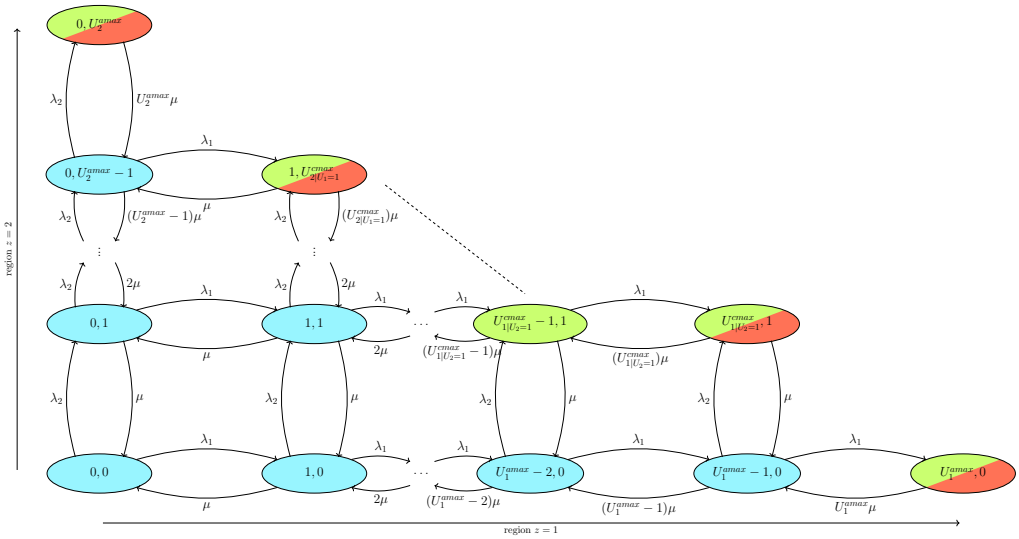


Figure 9.3: State transition diagram for a two-dimensional Erlang-B system. Note that red and green states correspond to $U_1 = U_{1|U_2}^{cmax}$ and $U_2 = U_{2|U_1}^{cmax}$, respectively.

region z conditioned to the number of UE sessions in the remaining regions (e.g., red states for region 1, and green states for region 2). This means that considering the state $s^{(t)} = \left(U_1, U_2, \ldots, U_{z|U_y}^{cmax}, \ldots, U_{N_z}\right) \in \mathcal{S}' \ \forall y \in \mathcal{Z} \setminus \{z\}$, then, the transition to the state $s^{(t+1)} = \left(U_1, U_2, \ldots, U_{z|U_y}^{cmax} + 1, \ldots, U_{N_z}\right) \notin \mathcal{S}'$ $\forall y \in \mathcal{Z} \setminus \{z\}$. If the remaining regions have 0 UEs, we can define the absolute maximum number of UE sessions U_z^{amax} in region z. This means that considering the state $s^{(t)} = (0, 0, \ldots, U_z^{amax}, \ldots, 0) \in \mathcal{S}'$, then the transition to the state $s^{(t+1)} = (0, 0, \ldots, U_z^{amax} + 1, \ldots, 0) \notin \mathcal{S}'$.

The resulting multi-dimensional Erlang-B system corresponds to a reversible Markov process (see proof in appendix 8.6). This implies the proposed model is insensitive to the distribution of the UE session duration, which means the state probabilities depend only upon the mean service time [21]. Furthermore, the solution for the probability of being in a state $s \in \mathcal{S}'$, i.e., $p_s\left[U_1, U_2, \ldots, U_{N_Z}\right]$, has product form as Eq. (9.18) shows. In this equation, $p\left[U_z\right]$ is the one-dimensional truncated Poisson distribution for traffic stream in region z and K is a normalization constant.

$$\begin{aligned} p_s\left[U_1, U_2, \ldots, U_{N_Z}\right] =& K \cdot p\left[U_1\right] \cdot p\left[U_2\right] \cdot \ldots \cdot p\left[U_{N_Z}\right] \\ =& K \cdot \prod_{z=1}^{N_Z} \frac{\rho_z^{U_z}}{U_z!} \end{aligned} \tag{9.18}$$

To obtain the state probabilities, we need to derive K. This constant can be computed by summing all the state probabilities and equaling the resulting expression to 1 as Eq. (9.19) shows.

$$K = \frac{1}{\sum_{s \in \mathcal{S}'} \left(\prod_{z=1}^{N_Z} \frac{\rho_z^{U_z}}{U_z!} \right)} \tag{9.19}$$

9.3.2 UE Blocking Probability

Assuming a new UE session is born in region z, it will be blocked if there not exists a transition from the current state $s^{(t)} = (U_1, U_2, \ldots, U_z, \ldots, U_{N_Z})$ to $s^{(t+1)} = (U_1, U_2, \ldots, U_z + 1, \ldots, U_{N_Z})$. This happens when $U_z + 1 > U_{z|U_y}^{cmax}$. We define $\mathcal{S}_z^B$ as the set of states where a transition is not possible in the region z. Based on that, we can compute the UE blocking probability B_z conditioned to the region z where the new UE session is born by Eq. (9.20).

$$B_z = \sum_{s \in \mathcal{S}_z^B} p_s \left[U_1, U_2, \ldots, U_{z|U_y}^{cmax}, \ldots, U_{N_z} \right] \tag{9.20}$$

Finally, the UE blocking probability B in the cell is defined in Eq. (9.21). This is computed as the sum of the conditional blocking probabilities weighted by the probability of a UE session is born in each region z.

$$B = \sum_{z=1}^{N_Z} \pi_z B_z \tag{9.21}$$

9.3.3 Mean Number of Consumed RBGs and Cell Capacity

Two key parameters derived by our model are the mean number of RBGs consumed in a cell $\overline{N}_{RBG}$, and the cell capacity D_i for a RAN slice with GBR requirements.

The mean number of RBGs $\overline{N}_{RBG}$ can be computed as Eq. (9.22) defines. In this equation, we multiply the probability of being in a state $s \in \mathcal{S}'$ by the amount of consumed RBGs in this state.

$$\overline{N}_{RBG} = \sum_{s \in \mathcal{S}'} p_s \left[U_1, U_2, \ldots, U_z, \ldots, U_{N_z}\right] \sum_{z=1}^{N_z} U_z N_{z,s}^{RBG} \tag{9.22}$$

The cell capacity D_i is provided by Eq. (9.23). It is derived as the product of the mean number of UEs $\overline{U}_z$ in each region z multiplied by the data rate consumed by each UE. We can compute $\overline{U}_z$ using Little's theorem[21], i.e., $\overline{U}_z = (\lambda_z/\mu)(1 - B_z) = \rho_z (1 - B_z)$.

$$D_i = \sum_{z=1}^{N_Z} \overline{U}_z D_{GBR} \tag{9.23}$$

9.4 UE Throughput with a Channel-Aware Scheduler

Considering the state $s = (U_1, U_2, \ldots, U_z, \ldots, U_{N_z})$, we compute in this section the expected throughput for a UE session in each region z and thus determining if the state s is feasible (i.e., $s \in \mathcal{S}'$) or not (i.e., $s \in \mathcal{S}^{nf}$).

Based on the channel-aware scheduler described in Section 9.2.4, we define the expected throughput for the UE session u as Eq. (9.24) shows,[20]. In this equation, $f_{\gamma_{u,n}|T_{u,n}=1}[\gamma]$ denotes the PDF of the instantaneous SINR $\gamma_{u,n}$ assuming the RBG n is allocated for the UE u. Furthermore, $C[\gamma]$ is a function which provides the data rate achieved by an UE when it perceives an instantaneous SINR γ.

$$R_u = \sum_{n=1}^{N_i^{RBG}} \int_0^{\infty} C[\gamma] \, f_{\gamma_{u,n}|T_{u,n}=1}[\gamma] \, \mathrm{P}[T_{u,n} = 1] \, d\gamma \tag{9.24}$$

In Eq. (9.24), we can define $\mathrm{P}[T_{u,n} = 1]$ in function of $\mathrm{P}[S_{u,n} = 1]$ and P_u. Furthermore, since $S_{u,n} = 1$ when $T_{u,n} = 1$, we can replace $f_{\gamma_{u,n}|T_{u,n}=1}[\gamma]$ with $f_{\gamma_{u,n}|S_{u,n}=1}[\gamma]$. Considering this changes, we can rewrite the expected throughput R_u as

$$R_u = \sum_{n=1}^{N_i^{RBG}} \int_0^{\infty} C\left[\gamma\right] f_{\gamma_{u,n}|S_{u,n}=1}\left[\gamma\right] \mathrm{P}\left[S_{u,n}=1\right] \mathrm{P}_u d\gamma \tag{9.25}$$

We can consider the Bayes' theorem, i.e., $f_{\gamma_{u,n}|S_{u,n}=1}\left[\gamma\right] = \frac{f_{S_{u,n}=1|\gamma_{u,n}}\left[\gamma\right] f_{\gamma_{u,n}}\left[\gamma\right]}{\mathrm{P}[S_{u,n}=1]}$. Furthermore, with the aim of maintaining the block error rate for the UE's data below a certain threshold, the cell adopts a link adaptation technique. This technique enables the cell to adapt the UEs' Modulation and Coding Scheme (MCS) according to the experienced channel effects. In our work, we consider a total of N_c MCSs. Hence, the range of the instantaneous SINR is split into N_c intervals $[\gamma_i,\ \gamma_{i+1}]$. In each interval, the achieved spectral efficiency takes a specific value c_i. Considering these statements, we rewrite the expected throughput as Eq. (9.26) defines. Note that P_u is out of the integral since it does not depend on the instantaneous SINR.

$$R_u = \mathrm{P}_u N_{SC} \Delta_f N_{size}^{RBG} \sum_{n=1}^{N_i^{RBG}} \sum_{i=1}^{N_c} c_i \int_{\gamma_i}^{\gamma_{i+1}} f_{S_{u,n}=1|\gamma_{u,n}}\left[\gamma\right] f_{\gamma_{u,n}}\left[\gamma\right] d\gamma \tag{9.26}$$

To compute the PDF of the allocation of the RBG n for UE u under the assumption of an instantaneous SINR $\gamma_{u,n}$, i.e., $f_{S_{u,n}=1|\gamma_{u,n}}\left[\gamma\right]$, we make use of the scheduling criteria defined in Eq. (9.9). Specifically, we perform the steps described in Eq. (9.27) [24]. Note that $\mathcal{U}^s$ denotes the set of UE sessions considered in state $s = (U_1, U_2, \ldots, U_z, \ldots, U_{N_z})$.

$$\begin{aligned} f_{S_{u,n}=1|\gamma_{u,n}}\left[\gamma\right] &= \mathrm{P}\left[\hat{\gamma}_{u,n} \geq \max_{\forall v \in \mathcal{U}^s \setminus \{u\}} \{\hat{\gamma}_{v,n}\} \,|\gamma_{u,n} = \gamma\right] \\ &= \mathrm{P}\left[\frac{\gamma}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}} \geq \max_{\forall v \in \mathcal{U}^s \setminus \{u\}} \{\hat{\gamma}_{v,n}\}\right] \\ &= \prod_{\forall v \in \mathcal{U}^s \setminus \{u\}} F_{\hat{\gamma}_{v,n}}\left[\frac{\gamma}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}}\right] \\ &= \prod_{\forall v \in \mathcal{U}^s \setminus \{u\}} F_{\gamma_{v,n}}\left[\frac{\left(\overline{\gamma}_{z_v}\right)^{\alpha} \gamma}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}}\right] \end{aligned} \tag{9.27}$$

If we include the result of Eq. (9.27) in Eq. (9.26), we obtain the expected throughput as

$$R_u = \mathrm{P}_u N_{SC} \Delta_f N_{size}^{RBG} \sum_{n=1}^{N_i^{RBG}} \sum_{i=1}^{N_c} c_i \int_{\gamma_i}^{\gamma_{i+1}} f_{\gamma_{u,n}}[\gamma] \prod_{\forall v \in \mathcal{U}^s \setminus \{u\}} F_{\gamma_{v,n}} \left[\frac{\left(\overline{\gamma}_{z_v}\right)^{\alpha} \gamma}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}} \right] d\gamma \tag{9.28}$$

Since the set $\mathcal{U}^s$ of UE sessions is split into N_z regions, we define in Eq. (9.29) the expected throughput R_z for a UE session which is born in the region z. Note that we have also replaced P_u with P_z since this probability is the same for all the UE sessions which are born in the same region z.

$$\begin{aligned} R_z = \mathrm{P}_z N_{SC} \Delta_f N_{size}^{RBG} \sum_{n=1}^{N_i^{RBG}} \sum_{i=1}^{N_c} c_i \int_{\gamma_i}^{\gamma_{i+1}} f_{\gamma_{u,n}}[\gamma] \left(F_{\gamma_{1,n}} \left[\frac{(\overline{\gamma}_1)^{\alpha} \gamma}{(\overline{\gamma}_z)^{\alpha}} \right] \right)^{U_1} \cdot \ldots \\ \cdot \left(F_{\gamma_{2,n}} \left[\frac{(\overline{\gamma}_2)^{\alpha} \gamma}{(\overline{\gamma}_z)^{\alpha}} \right] \right)^{U_2} \cdot \ldots \\ \cdot \left(F_{\gamma_{z,n}}[\gamma] \right)^{U_z - 1} \cdot \ldots \cdot \left(F_{\gamma_{N_z,n}} \left[\frac{\left(\overline{\gamma}_{N_z}\right)^{\alpha} \gamma}{(\overline{\gamma}_z)^{\alpha}} \right] \right)^{U_{N_z}} d\gamma \end{aligned} \tag{9.29}$$

If we consider the exponential distribution for the instantaneous SINR as shown in Eqs. (9.4) and (9.5), we can rewrite the expected throughput R_z for each region z as

$$\begin{aligned} R_z = \mathrm{P}_z N_{SC} \Delta_f N_{size}^{RBG} N_i^{RBG} \sum_{i=1}^{N_c} c_i \int_{\gamma_i}^{\gamma_{i+1}} \frac{1}{\overline{\gamma}_z} \exp\left[\frac{-\gamma}{\overline{\gamma}_z} \right] \left(1 - \exp\left[-\frac{(\overline{\gamma}_1)^{\alpha-1}}{(\overline{\gamma}_z)^{\alpha}} \gamma \right] \right)^{U_1} \\ \cdot \left(1 - \exp\left[-\frac{(\overline{\gamma}_2)^{\alpha-1}}{(\overline{\gamma}_z)^{\alpha}} \gamma \right] \right)^{U_2} \cdot \ldots \cdot \left(1 - \exp\left[-\frac{1}{\overline{\gamma}_z} \gamma \right] \right)^{U_z - 1} \cdot \ldots \\ \cdot \left(1 - \exp\left[-\frac{\left(\overline{\gamma}_{N_z}\right)^{\alpha-1}}{(\overline{\gamma}_z)^{\alpha}} \gamma \right] \right)^{U_{N_z}} d\gamma \end{aligned} \tag{9.30}$$

Since the expected throughput R_z must be the same for each region z, we

have $R_z = D_{GBR}$. Furthermore, we can denote as $I_z\left(\alpha, \gamma_i, \gamma_{i+1}, U_1, U_2, \ldots, U_{N_z}\right)$ the definite integral in Eq. (9.30). This integral has an analytical solution which depends on the value for the fairness factor α and the set of UEs $\mathcal{U}^s$ defined by the state $s = (U_1, U_2, \ldots, U_z, \ldots, U_{N_z})$. If we consider Eq. (9.30) for each region z, we can define a set of N_z equations with $N_z + 1$ unknown variables (i.e., α, $\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_{N_z}$) given by

$$\begin{cases} \mathrm{P}_1 N_{SC} \Delta_f N_{size}^{RBG} N_i^{RBG} \sum_{i=1}^{N_c} c_i I_1\left(\alpha, \gamma_i, \gamma_{i+1}, U_1, U_2, \ldots, U_{N_z}\right) & = & D_{GBR} \\ \mathrm{P}_2 N_{SC} \Delta_f N_{size}^{RBG} N_i^{RBG} \sum_{i=1}^{N_c} c_i I_2\left(\alpha, \gamma_i, \gamma_{i+1}, U_1, U_2, \ldots, U_{N_z}\right) & = & D_{GBR} \\ & \vdots & \\ \mathrm{P}_{N_z} N_{SC} \Delta_f N_{size}^{RBG} N_i^{RBG} \sum_{i=1}^{N_c} c_i I_{N_z}\left(\alpha, \gamma_i, \gamma_{i+1}, U_1, U_2, \ldots, U_{N_z}\right) & = & D_{GBR} \end{cases} \tag{9.31}$$

Based on the previous equation system, the state s will be only valid, i.e., $s \in \mathcal{S}'$, if all these equations are satisfied. To that end, the MNO must properly set the values for the fairness factor α and $\mathrm{P}_z \ \forall z \in \mathcal{Z}$. There exists multiple solutions for these two parameters. For this reason, the MNO could define an optimization criteria which provides it a benefit.

For example, an optimization criteria could be maximizing the efficiency of the RBG utilization for a RAN slice. To that end, each probability P_z must be as close as possible to the value 1, i.e., the scheduler defined in Section 9.2.4 discards the minimum amount of RBGs after applying the third scheduling step.

Despite the benefits of solving an optimization problem, its implementation in a real system is not practical. The reason is the optimization problem must be solved for each state $s \in \mathcal{S}$ to: (a) determine if this state is valid; and (b) in that case, establishing the optimal values of the fairness parameter α and the probabilities $\mathrm{P}_z \ \forall z \in \mathcal{Z}$ in such state. This is not scalable when the number of regions N_z and the number of available RBGs N_i^{RBG} are considerably high.

Due to these issues, we propose a sub-optimal solution in this work. Specifically, this solution consists of setting an unique value of the fairness factor α for all the states $s \in \mathcal{S}'$. Based on that value, the probabilities $\mathrm{P}_z \ \forall z \in \mathcal{Z}$ can be directly derived from the equation system defined in (9.31). The value for the fairness factor α must be computed by the MNO during the planning phase of the RAN slice (i.e., before deploying the corresponding GBR service). Motivated

by its interest, we evaluate how the fairness factor α impacts on the UE blocking probability, analyzing which aspects must be considered by the MNO to configure this parameter.

9.5 Numerical Results

Since the state-of-the-art models for computing the UE blocking probability (see section 9.1.1) are not appropriate for RAN slices with GBR requirements under our assumptions (i.e., reduced UE mobility, and arbitrary distributions for the duration of the UE sessions as well as for the average channel quality within the cell), we cannot provide a fair comparison. For this reason, in this paper we experimentally validate the proposed model by means of simulation. We also evaluate the UE blocking probability for different configurations of the channel-aware scheduler described in Section 9.2.4.

Specifically, we first analyze the aspects that impact the execution time of our model. Then, we evaluate the relative error for the UE blocking probability with respect to the one obtained by simulation. Finally, we evaluate UE blocking probability, the number of active UE sessions, and the radio resource utilization when the MNO sets different values for the fairness factor α considered by the channel-aware scheduler. These performance indicators are also evaluated in a baseline scenario where the cell implements a channel-agnostic scheduler as the one we assumed in our previous work [15].

9.5.1 Experimental Setup

To validate the proposed model, we use a Matlab-based simulator that resembles the arrival and departure of UE sessions for a RAN slice with GBR requirements in a single cell. This simulator generates UE sessions following a Poisson distribution. With respect to the UE session duration, we have carried out all the experiments considering an exponential distribution, an uniform distribution, and a constant duration. For all the cases, the results are equal since our model is insensitive to the holding time distribution. Focusing on a single UE session, the simulator considers (a) the region z where the session takes place and (b) the average number of allocated RBGs $N_{z,s}^{RBG}$ for such session. To determine if a new UE session can be admitted, the simulator first considers s as the state where

the new UE session is admitted. Then, it checks if this state belongs to the set of feasible states, i.e., $s \in \mathcal{S}'$. If true, the new UE session is admitted. Table 9.1 summarizes the configuration parameters.

Regarding the access technology, we assume a 5G-NR cell implementing an OFDMA scheme with $\Delta f = 15$ KHz, and $N_{SC} = 12$. We also consider different scenarios where the serving cell allocates 20, 25, 30, 35, 40 and 45 RBs for the RAN slice. Additionally, we consider each UE session consumes multiples of 2 RBs, i.e., the RBG size $N_{size}^{RBG} = 2$. With respect to $f_{\overline{\gamma}_u}(\overline{\gamma})$, we have derived it by using the distribution of the G-factor experimentally measured in a macro cell [25]. The G-factor distribution is defined as the average own cell power to the other-cell power plus noise ratio. With OFDMA in a wide system bandwidth,

Table 9.1: Configuration Parameters

Parameters	**Configuration**
Access Technology	5G-NR
Subcarrier spacing Δf (OFDMA)	15 KHz
Sub-carriers per RB N_{SC} (OFDMA)	12
Number of allocated RBs N_i^{RB}	20 RBs, 25 RBs, 30 RBs, 35 RBs, 40 RBs and 45 RBs
RBG Size N_{size}^{RBG}	2 RBs
Fast-fading distribution	Rayleigh with unit mean
PDF average SINR in the cell: $f_{\overline{\gamma}_u}(\overline{\gamma})$	Built using the distribution of the G-factor measured in a macro cell [25]
Regions for the average SINR, i.e., $[\overline{\gamma}^{z-1}, \overline{\gamma}^{z})$ (in dB)	$N_z = 4$: [-5, 1), [1, 7), [7, 13), [13, 19) $N_z = 5$: [-5, -0.2), [-0.2, 4.6), [4.6, 9.4), [9.4, 14.2), [14.2 19) $N_z = 6$: [-5, -1), [-1, 3), [3, 7), [7, 11), [11 15), [15 19) $N_z = 7$: [-5, -1.574), [-1.574, 1.857), [1.857, 5.286), [5.286, 8.714), [8.714 12.143), [12.143, 15.571), [15.571, 19) $N_z = 8$: [-5, -2), [-2, 1), [1, 4), [4, 7), [7 10), [10, 13), [13, 16), [16, 19) $N_z = 9$: [-5, -2.333), [-2.333, 0.333), [0.333, 3), [3, 5.667), [5.667 8.333), [8.333, 11), [11, 13.667), [13.667, 16.333), [16.333, 19)
Average SINR $\overline{\gamma}_z$ (in dB) and probability π_z in each region, i.e., $(\overline{\gamma}_z, \pi_z)$	$N_z = 4$: [-0.907, 0.280), [3.636, 0.372), [9.496, 0.223), [15.157, 0.125) $N_z = 5$: [-1.601, 0.186), [1.925, 0.337), [6.776, 0.232), [11.514, 0.151), [15.944, 0.095) $N_z = 6$: [-2.112, 0.129), [0.944, 0.296), [4.813, 0.227), [8.907, 0.162), [12.832, 0.112), [16.3364, 0.074) $N_z = 7$: [-2.449, 0.094), [0.159, 0.253), [3.411, 0.217), [6.860, 0.163), [10.336 0.121), [13.757, 0.094). [16.505, 0.058) $N_z = 8$: [-2.981, 0.070), [-0.402, 0.210), [2.374, 0.210), [5.402, 0.162), [8.402 0.122), [11.374, 0.101). [14.598, 0.080), [16.617, 0.046) $N_z = 9$: [-3.037, 0.053), [-0.879, 0.179), [1.589, 0.193), [4.252, 0.159), [6.888, 0.126), [9.383, 0.104), [12.271, 0.077), [15.0187, 0.071), [16.897, 0.038)
Service GBR D_{GBR}	0.8 Mbps
Distribution for the UE session arrival	Poisson
Distribution for the UE session duration	Exponential, Uniform and Constant
Offered Traffic Intensity ρ	From 0.5 to 1
Fairness Factor α	1.0, 1.3, 1.5 and 1.7

Table 9.2: Execution Time

$N_Z = 6$	$N_i^{RB} = 20$	$N_i^{RB} = 25$	$N_i^{RB} = 30$	$N_i^{RB} = 35$	$N_i^{RB} = 40$	$N_i^{RB} = 45$
	1.802 s	3.652 s	7.451 s	13.939 s	28.125 s	47.851 s
$N_i^{RB} = 35$	$N_Z = 4$	$N_Z = 5$	$N_Z = 6$	$N_Z = 7$	$N_Z = 8$	$N_Z = 9$
	1.299 s	3.742 s	13.939 s	57.981 s	260.926 s	912.471 s

the distribution of the G-factor corresponds to the distribution of the average SINR [26]. Additionally, we consider different values for the number of regions for such distribution, from $N_z = 4$ to $N_z = 9$. For the GBR service provided by the RAN slice, we assume a data rate of $D_{GBR} = 0.8$ Mbps for each active UE session. We consider a low data rate since we assume the UE sessions which are born in the region z' with the worst average SINR $\overline{\gamma}_{z'}$ must require an average number of RBGs less than the number of allocated RBGs for the RAN slice, i.e., $N_{z',s}^{RBG} \leq N_i^{RBG}\ \forall s \in \mathcal{S}'$. With respect to the channel-aware scheduler, we consider the constant values 1.0, 1.3, 1.5 and 1.7 for the fairness factor α.

Based on these configuration parameters, we have evaluated the UE blocking probability in function of the offered traffic intensity. Specifically, from $\rho = 0.5$ to $\rho = 1$.

All the experiments have been carried out on a computer with 16 GB RAM and an Intel core i7-7700HQ @ 2.80 GHz.

9.5.2 Execution Time Evaluation

We have assessed the time complexity of our analytical model in two scenarios. In the former, we have covered several amount of radio resources allocated for a RAN slice from 20 RBs to 45 RBs, considering $N_Z = 6$. In the latter, we have considered different number of regions from $N_Z = 4$ to $N_Z = 9$, with 35 RBs allocated for such RAN slice. In both scenarios, the fairness factor α is set to 1.3 and the offered traffic intensity $\rho = 1$. The results for both scenarios are shown in Table 9.2.

We observe the execution time grows exponentially with the number of regions and the number of RBs. The reason is using higher values for both parameters involves an increment in the number of states in the Markov chain as Eq. (9.17) shows. Furthermore, although it is not shown in Table 9.2, we have experimentally verified that setting a different value for the fairness factor could slightly

change the number of states in the Markov chain, thus the execution time could slightly increase or decrease. The impact of setting a different value for the fairness factor is evaluated in Sections 9.5.4, 9.5.5 and 9.5.6. Note that the execution time does not depend on the offered traffic intensity since it does not modify the number of states in the Markov chain.

9.5.3 Model Validation

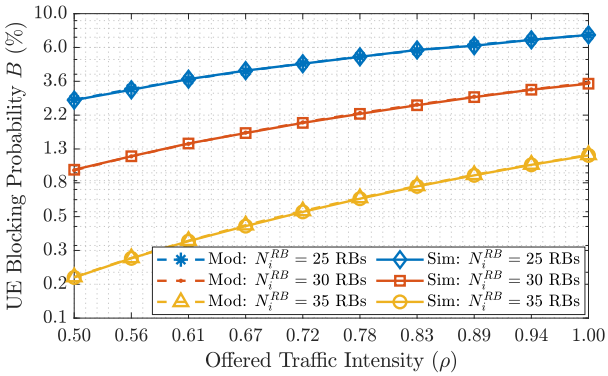
To validate our model, we have computed the relative error as $\epsilon_r(\%) = \frac{B_{sim}-B_{mod}}{B_{sim}} \cdot 100$, where B_{sim} and B_{mod} denote the UE blocking probability extracted from the proposed model and simulator, respectively. Furthermore, we have considered three scenarios with a specific RB allocation for a RAN slice: 25 RBs, 30 RBs and 35 RBs. In all the scenarios, the number of regions considered for $f_{\overline{\gamma}_u}(\overline{\gamma})$ is $N_z = 6$.

In Fig. 9.4(a), we depict the UE blocking probability derived from our model and the simulator. It shows how the UE blocking probability increases when (a) the available RBs for a RAN slice are decreased and (b) the offered traffic intensity increases. This graph is useful for MNOs to decide the bandwidth of each cell (i.e., N_i^{RB}) for RAN slices with GBR requirements while a threshold for B is provided, given certain conditions for the offered traffic intensity and the channel quality, i.e., a specific $f_{\overline{\gamma}_u}(\overline{\gamma})$.

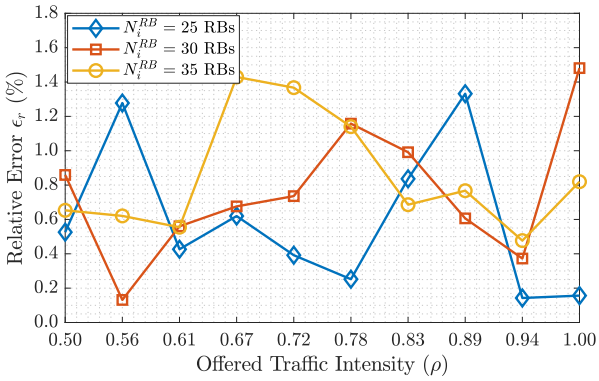
Due to the scale used for the vertical and horizontal axes in Fig. 9.4(a), the error between the simulation and the model cannot be observed. In Fig. 9.4(b), we represent the relative error, which is below 1.5 % for any case.

9.5.4 Evaluation of the UE Blocking Probability with a Channel-Aware Scheduler

In this experiment, we have evaluated the UE blocking probability for different scenarios where the channel-aware scheduler sets a specific value for the fairness factor α. Specifically, we have considered α takes the values 1.0, 1.3, 1.5 and 1.7 for each scenario, respectively. In addition, we have compared these scenarios with the case of implementing a channel-agnostic scheduler as the one we considered in our previous work [15]. For all the scenarios, we have assumed there are 35 RBs allocated for a RAN slice. Furthermore, we have set $N_z = 6$ as the



(a) UE Blocking Probability: Model vs Simulation



(b) Relative error

Figure 9.4: Evaluation of the UE Blocking Probability for different cell bandwidths.

number of regions for $f_{\overline{\gamma}_u}(\overline{\gamma})$.

We show these results in Fig. 9.5. It can be seen that the UE blocking probability B is usually lower when the cell implements a channel-aware scheduler. This is because the channel-aware scheduler improves the resource utilization by allocating RBGs to the UE sessions which are less affected by the fast-fading effect, as the metric defined in Eq. (9.9) describes. This improvement depends on how the MNO configures the fairness factor α in the channel-aware scheduler. Below, we show how setting different values for the fairness factor α impacts the UE blocking probability.

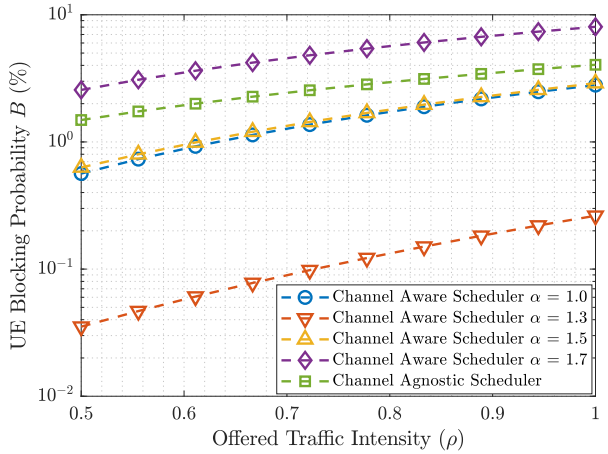


Figure 9.5: UE Blocking Probability when the cell implements a scheduler with a specific configuration.

Setting $\alpha = 1.0$ (i.e., using the PF criteria) may not be an appropriate option because in the second step of the channel-aware scheduler (see Fig. 9.2) all the RBGs are pre-allocated with the same probability for each UE session regardless their average SINRs. Then, some RBGs may not be finally allocated after the third step (see Fig. 9.2) in the regions which have the best average SINRs. This means that, after the scheduler operation, there could be free RBGs for admitting more UE sessions in the regions with the lowest average SINRs. However, these RBGs cannot be used for the UE sessions which have the worst average SINRs due to the equal RBG distribution of the PF criteria. To avoid this issue, the MNO must increase the fairness factor α. In this way, the UE sessions located in the regions which have lowest average SINRs receive a greater probability for being scheduled with more RBGs. This means more UEs sessions could be admitted and thus, the UE blocking probability would decrease. For instance, we observe this fact in Fig. 9.5 when the MNO sets $\alpha = 1.3$. In this case, the UE blocking probability is the lowest for all the values of the offered traffic intensity.

However, if we follow increasing the fairness factor α, for instance $\alpha = 1.5$, more RBGs are scheduled for those UE sessions which was born in the regions with the lowest average SINRs. This involves more UE sessions in the regions with the greatest average SINRs are rejected. Thus the cell is using more RBGs for admitting a less amount of UE sessions and the UE blocking probability increases.

Note that the UE blocking probability is very sensitive to small increments of α since this parameter is used as exponent of the average SINRs as Eq. (9.9) defines.

If this parameter is not properly configured, the MNO could not leverage the advantages of using a channel-aware scheduler. For instance, if we set $\alpha = 1.7$, the achieved UE blocking probability is greater than the one achieved by using a channel-agnostic scheduler.

9.5.5 Evaluation of the Number of Active UE Sessions with a Channel-Aware Scheduler

In this experiment, we have evaluated the number of active UE sessions in a RAN slice. To that end, we have considered the same scenarios presented in Section 9.5.4. Based on them, Fig. 9.6 depicts the Complementary Cumulative Distribution Functions (CCDFs) for the number of active UE sessions when the traffic intensity ρ varies from $\rho = 0.5$ to $\rho = 2$. It can be observed that increasing ρ yields a greater probability that more UE sessions are active, i.e., the CCDFs are right-shifted. This means the UE blocking probability increases as Fig. 9.5 depicts.

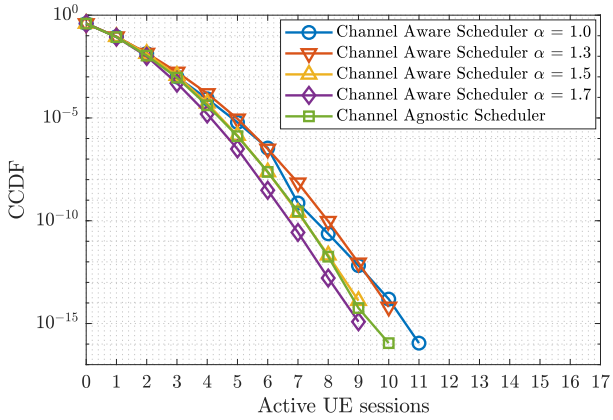
If we focus on the CCDFs for a specific value of the traffic intensity ρ, we observe how the probability of the number of active UE sessions is higher than a specific value (i.e., a x-axis value) is greater when the channel-aware scheduler sets the fairness factor $\alpha = 1.3$. This probability decreases (i.e., less active UE sessions) when the fairness factor α takes other values. For instance, when $\alpha = 1.3$ is more probable than there are more active UE sessions than the case of setting α with the values 1.0, 1.5 and 1.7. This means it is more difficult to reach a blocking state $\mathcal{S}^B$, thus the UE blocking probability is lower when the channel-aware scheduler sets $\alpha = 1.3$. The worst case happens when the MNO sets the fairness factor $\alpha = 1.7$.

Note that the values of the CCDF below 10^{-17} have been omitted because they are not significant.

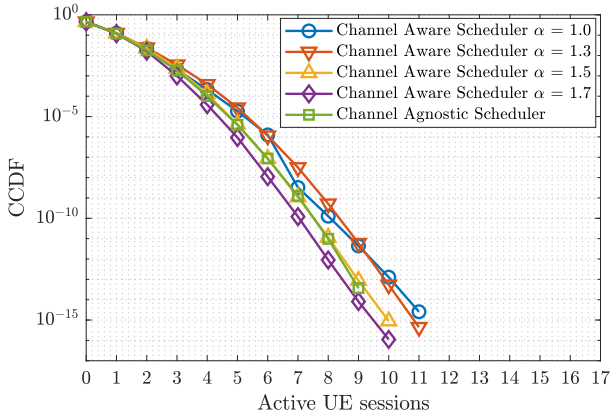
9.5.6 Analysis of the Radio Resource Utilization

In this experiment, we have evaluated the radio resource utilization for a RAN slice. To that end, we have considered the same scenarios presented in Section

9.5.4. To measure the radio resource utilization, we consider the probabilities P_z for each region z and for each valid state $s \in \mathcal{S}'$. The reason is these probabilities define the percentage of RBGs allocated for each UE session after the scheduling criteria in Eq. (9.9) is applied by the channel-aware scheduler in the steps 2-3 (see Fig. 9.2). These probabilities are depicted in Fig. 9.7. Each point represents the specific value of a probability P_z when there is a specific set of active UE sessions in the system, i.e., a specific state $s \in \mathcal{S}'$. For instance, the cyan points represent these probabilities for all the UE sessions which was born in region $z = 6$, i.e.,

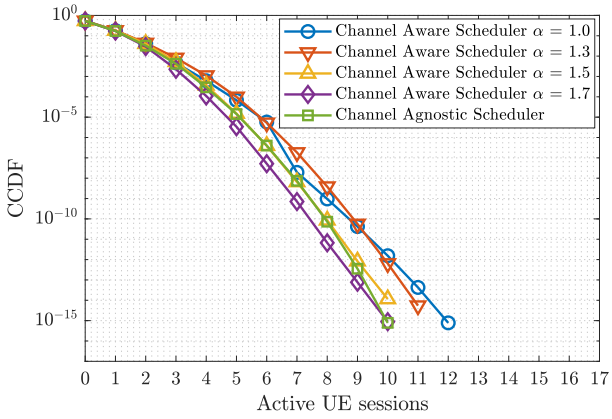


(a) Traffic Intensity $\rho = 0.5$

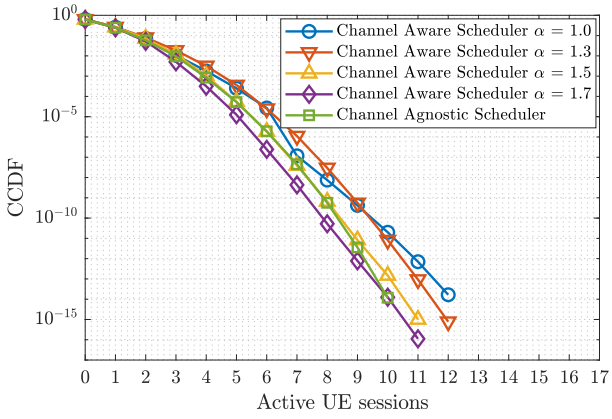


(b) Traffic Intensity $\rho = 1.0$

Figure 9.6: CCDFs of the number of active UE sessions



(c) Traffic Intensity $\rho = 1.5$



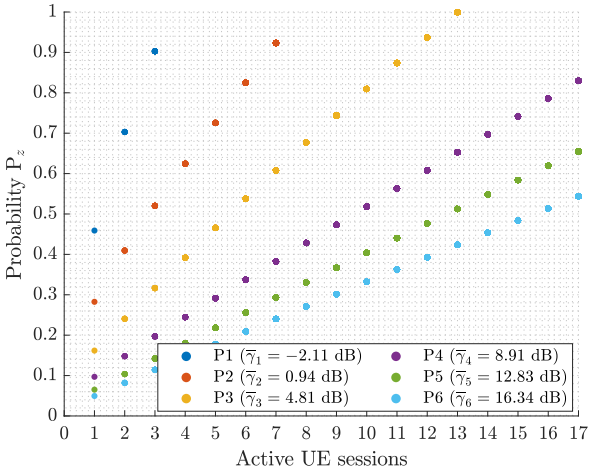
(d) Traffic Intensity $\rho = 2.0$

Figure 9.6: CCDFs of the number of active UE sessions

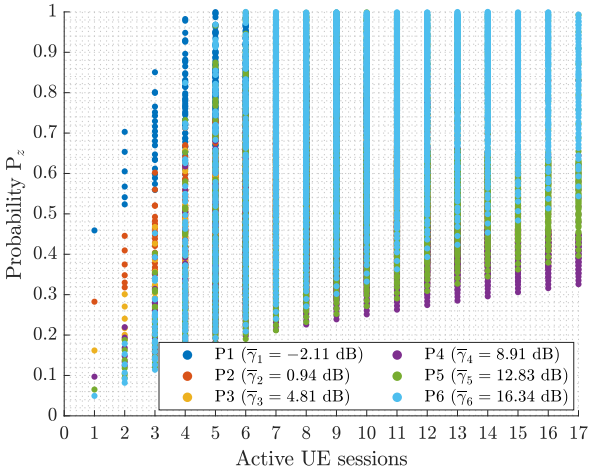
the region where $\overline{\gamma}_6 = 16.34$ dB.

Fig. 9.7(a) represents these probabilities in the case the cell implements a channel-aware scheduler with $\alpha = 1.0$. This means the scheduler uses the PF criteria and thus, the RBGs are equally distributed among the UE sessions before applying these probabilities in the step 2 (see Fig. 9.2). If we focus on a specific amount of UE sessions (i.e., see x-axis), the probability in the region $z = 1$ (i.e., $\overline{\gamma} = -2.11$ dB) is always the greatest. Thus, the channel-aware scheduler has to allocate a higher percentage of RBGs for a UE (i.e., step 3) with respect to

the amount of RBGs assigned following the PF criteria (i.e., step 2) in the region with the worst average SINR. As explained before, this means that using in a first attempt the PF criteria is not efficient because the percentage of RBGs which are not finally allocated for a UE in a region with a higher average SINR (for

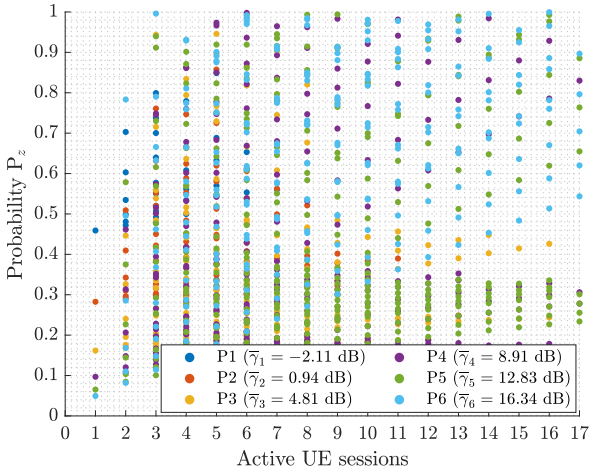


(a) Channel-Aware Scheduler with $\alpha = 1.0$



(b) Channel-Aware Scheduler with $\alpha = 1.3$

Figure 9.7: Values of P_z per each region z and for each valid state $s \in \mathcal{S}'$



(c) Channel-Aware Scheduler with $\alpha = 1.7$

Figure 9.7: Values of P_z per each region z and for each valid state $s \in \mathcal{S}'$

instance the region $z = 3$, where $\overline{\gamma} = 4.81$ dB) could be used to admit more UE sessions in regions which have a worst average SINR (for instance the region $z = 1$, where $\overline{\gamma} = -2.11$ dB).

Considering the MNO sets $\alpha = 1.3$ for all the valid states $\mathcal{S}'$, Fig. 9.7(b) shows how in some states the probabilities for region $z = 6$ (i.e., $\overline{\gamma} = 16.34$ dB) are closer to the value 1 in comparison with the probabilities obtained when the MNO sets $\alpha = 1.0$, i.e., the probabilities depicted in Fig. 9.7(a). However, almost all the probabilities in region $z = 1$ (i.e., $\overline{\gamma} = 5$ dB) are greater than the probabilities in the remaining regions in each single state. This phenomena can be observed in the left side of Fig. 9.7(b). This involves some RBGs are not used in the regions with a higher average SINR, and thus they could be used for admitting UE sessions in other states.

When the MNO sets $\alpha = 1.7$ for all the valid states $\mathcal{S}'$, the values of these probabilities have decreased in comparison with the values depicted in Fig. 9.7(b). This means that increasing the value of the fairness factor α involves that an excessive amount of RBGs are allocated for the lowest regions (i.e., those which have the lowest average SINRs). This involves not all the RBGs allocated for a UE session which falls in the worst region are used, thus these RBGs are wasted

and they are not used for admitting other UE sessions in regions with a better average SINR.

9.6 Conclusions

RAN slicing is envisioned as a solution for providing emerging communication services over a common wireless network infrastructure. Implemented as RAN slices, some of these communication services will rely on data transmission with requirements in terms of GBR. Designing strategies for planning GBR services could involve inherent issues such as the under(over)-provisioning of radio resources. For instance, the under-provisioning of radio resources could involve that a considerable amount of UE sessions would be rejected since they would not be able to meet the GBR requirements. This means the UE blocking probability would be excessively high. To avoid that, it is crucial for the MNO to model the UE blocking probability in each cell for such service.

Under this context, we propose an analytical model to evaluate this performance indicator. The main novelty is the consideration of a multi-dimensional Erlang-B system, which meets the reversibility property. This means our model is valid for arbitrary distributions of the UE session duration. This property also reduces the computation complexity of the model because the solution for the state probabilities has product form. Furthermore, our proposal considers as input an arbitrary distribution for the PDF of the average SINR in the cell. This allows the proposed model to consider a better characterization of the average channel quality within the cell.

Additionally, we formulate the GBR achieved by an UE session when the cell implements a channel-aware scheduler. This allows the proposed model to consider the impact of the channel gain of this scheduler on the UE blocking probability. The results show that our model exhibits an estimation error for the UE blocking probability below 1.5%. Furthermore, our model allows the MNO to determine in advance how a channel-aware scheduler must be configured to reduce the UE blocking probability when a GBR service supports a specific traffic intensity ρ. In the case of using an alpha-fair-based scheduler, the MNO can determine which value of α provides the lowest UE blocking probability.

Acknowledgments

This work is partially supported by the H2020 research and innovation project 5G-CLARITY (Grant No. 871428); the Spanish Ministry of Economy and Competitiveness, the European Regional Development Fund (Project PID2019-108713RB-C53); and the Spanish Ministry of Education, Culture and Sport (FPU Grant 17/01844)

References

[1] A. Aijaz, "Private 5G: The Future of Industrial Wireless," *IEEE Ind. Electron. Mag.*, vol. 14, no. 4, pp. 136–145, 2020.

[2] O. Adamuz-Hinojosa, P. Munoz, J. Ordonez-Lucena, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "Harmonizing 3GPP and NFV Description Models: Providing Customized RAN Slices in 5G Networks," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 64–75, 2019.

[3] I. Vision, "Framework and overall objectives of the future development of IMT for 2020 and beyond," *International Telecommunication Union (ITU), Document, Radiocommunication Study Groups*, 2015.

[4] T. Bonald and A. Proutière, "Wireless Downlink Data Channels: User Performance and Cell Dimensioning," in *MobiCom, San Diego, California, USA*, pp. 339–352, 2003.

[5] T. Bonald *et al.*, "Flow-level performance and capacity of wireless networks with user mobility," *Queueing Systems*, vol. 63, no. 1-4, p. 131, 2009.

[6] S.-E. Elayoubi and T. Chahed, "Admission Control in the Downlink of WCDMA/UMTS," in *EuroNGI, Dagstuhl, Germany*, pp. 136–151, Springer, 2004.

[7] D. K. Kim *et al.*, "A novel ring-based performance analysis for call admission control in wireless networks," *IEEE Commun. Lett.*, vol. 14, no. 4, pp. 324–326, 2010.

[8] B. Sas *et al.*, "Modelling the time-varying cell capacity in LTE networks," *Telecommun Syst.*, vol. 55, no. 2, pp. 299–313, 2014.

[9] C. Tarhini and T. Chahed, "QoS-oriented resource allocation for streaming flows in IEEE802. 16e Mobile WiMAX," *Telecommun Syst.*, vol. 51, no. 1, pp. 65–71, 2012.

[10] A. Abdollahpouri and B. E. Wolfinger, "Measures to quantify the gain of multicast with application to IPTV transmissions via WiMAX networks," *Telecommun Syst.*, vol. 55, no. 2, pp. 185–198, 2014.

[11] M. Li, "Queueing Analysis of Unicast IPTV With Adaptive Modulation and Coding in Wireless Cellular Networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9241–9253, 2017.

[12] C. Kim *et al.*, "Mathematical Models for the Operation of a Cell With Bandwidth Sharing and Moving Users," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 744–755, 2020.

[13] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 2, pp. 678–700, 2013.

[14] P. Ameigeiras, Y. Wang, J. Navarro-Ortiz, P. E. Mogensen, and J. M. Lopez-Soler, "Traffic models impact on OFDMA scheduling design," *EURASIP J. Wirel. Commun. Netw.*, vol. 2012, no. 1, pp. 1–13, 2012.

[15] O. Adamuz-Hinojosa, P. Ameigeiras, P. Muñoz, and J. M. Lopez-Soler, "Analytical Model for the UE Blocking Probability in an OFDMA Cell providing GBR Slices," in *IEEE WCNC, Nanjing, China*, pp. 1–7, 2021.

[16] 3GPP TS 38.101-1 V.16.3.0, "User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (Release 16)," Mar. 2020.

[17] 3GPP TS 38.101-2 V.16.4.0, "User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone (Release 16)," June 2020.

[18] P. Muñoz *et al.*, "Radio Access Network Slicing Strategies at Spectrum Planning Level in 5G and Beyond," *IEEE Access*, vol. 8, pp. 79604–79618, 2020.

[19] 3GPP TS 38.214 V.16.1.0, "NR, Physical layer procedures for data (Release 16)," Mar. 2020.

[20] D. Parruca and J. Gross, "Throughput Analysis of Proportional Fair Scheduling for Sparse and Ultra-Dense Interference-Limited OFDMA/LTE Networks," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 10, pp. 6857–6870, 2016.

[21] V. B. Iversen, "Teletraffic engineering and network planning," 2015.

[22] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, 1998.

[23] M. H. Ahmed, O. A. Dobre, and R. K. Almatarneh, "Analytical Evaluation of the Performance of Proportional Fair Scheduling in OFDMA-Based Wireless Systems," *J. Electr. Comput. Eng.*, vol. 2012, Jan. 2012.

[24] D. Parruca, M. Grysla, S. Gortzen, and J. Gross, "Analytical Model of Proportional Fair Scheduling in Interference-Limited OFDMA/LTE Networks," in *IEEE VTC Fall, Las Vegas, USA*, pp. 1–7, 2013.

[25] P. Ameigeiras, *Packet scheduling and quality of service in HSDPA*. Aalborg Universitetsforlag, 2003.

[26] P. Mogensen *et al.*, "LTE Capacity Compared to the Shannon Bound," in *IEEE VTC Spring, Dublin, Ireland*, pp. 1234–1238, 2007.

Chapter 10

Paper G. Potential-Game-Based 5G RAN Slice Planning for GBR Services

Authors:

Oscar Adamuz-Hinojosa, Pablo Munoz, Pablo Ameigeiras, Juan M. Lopez-Soler.

The paper has been submitted to IEEE Transactions on Mobile Computing.

Abstract

To deploy and operate Radio Access Network (RAN) slices for Guaranteed Bit-Rate (GBR) services, the Mobile Network Operator (MNO) typically performs Admission Control and Dynamic Resource Provisioning (DRP) procedures. Despite of their own utility, previously the MNO must also execute in advance a RAN slice planning procedure. Under this context, we propose a mathematical model for planning RAN slices. It specifically provides the minimum amount of radio resources which a DRP algorithm must use to meet the GBR requirements of each RAN slice for a given time window. Our model also ensures the User Equipment (UE) blocking probability for each RAN slice is below a certain threshold, given the worst-case inter-cell interference level. Particularly, we formulate multiple ordinal potential games and demonstrate the existence of a Nash Equilibrium solution which minimizes the average UE blocking probability for all the RAN slices. To reach our solution, we design novel strategies based on better response dynamics. This work also includes detailed simulations to demonstrate the effectiveness of the proposed solution in terms of performance, adaptability and renegotiation capability.

10.1 Introduction

Fifth Generation (5G) networks aim to boost the digital transformation of industry verticals. These verticals may bring a wide variety of communication services with diverging performance requirements. From the Mobile Network Operator (MNO) perspective, it would be unfeasible to deploy each communication service separately and build a dedicated Radio Access Network (RAN) accordingly. To economically provide these services, RAN slicing has emerged as a potential solution [1]. This technology consists of providing logically separated RANs, denominated RAN slices, each tailored to the requirements of a specific communication service over a common RAN infrastructure.

One of the main challenges of RAN slicing technology is how to allocate radio resources to each RAN slice. In this vein, most of the related contributions in the literature focus on Admission Control (AC) and Dynamic Resource Provisioning (DRP). The AC takes place when a RAN slice consumer, hereinafter referred as

tenant, requests the MNO a RAN slice with specific performance requirements. Then, the MNO verifies the feasibility of deploying such RAN slice under the current traffic conditions. If feasible, the RAN slice is accepted. After that, the assigned radio resources for such RAN slice are continuously adjusted by a DRP algorithm. Specifically, this algorithm adapts in runtime the amount of allocated radio resources for such RAN slice according to its traffic demands. Examples of RAN slicing solutions based on AC and DRP can be found in [2, 3, 4, 5, 6, 7].

As AC is executed on the spot, the sole use of AC to decide if a RAN slice is deployed would entail some limitations for the MNO. For instance, since AC is somehow a myopic procedure, the immediate admission of a new RAN slice could involve the rejection of more attractive (e.g., in economic terms) subsequent coming RAN slice requests.

Regarding DRP, the Third Generation Partnership Project (3GPP) has recently standardized a set of radio resource bounds, i.e., defined as policy ratios [8, 9]. Using these bounds, the MNO can define the minimum (or maximum) amount of radio resources which may be allocated by the DRP algorithm to each RAN slice while their performance requirements are met throughout their lifetimes. The utilization of these bounds has already been considered by some works (e.g., [10]), which assume they are conservatively established by the MNO before deploying the RAN slices. However, to the best of our knowledge, there are no any studies dealing with the computation of such acclaimed radio resource bounds.

To overcome the short-term nature of the AC and establish the radio resource bounds, the MNO must execute a long-term RAN slice planning procedure. It consists of deciding in advance the feasibility of deploying the requested RAN slices, the need to roll out new RAN resources (i.e., radio resources, cells, etc) and the adequate configuration of the RAN infrastructure to accommodate the feasible RAN slices, with the purpose of optimizing the utilization (e.g., in economic terms) of the infrastructure from the MNO's perspective.

Focusing on the RAN slice planning, we assume the MNO considers periodical time windows where new RAN slices may be deployed. For each time window, the MNO executes in advance a planning procedure which aims to accommodate the requested RAN slices along with the existing ones in its infrastructure. We consider the duration of the time window is fixed, and it may range from several

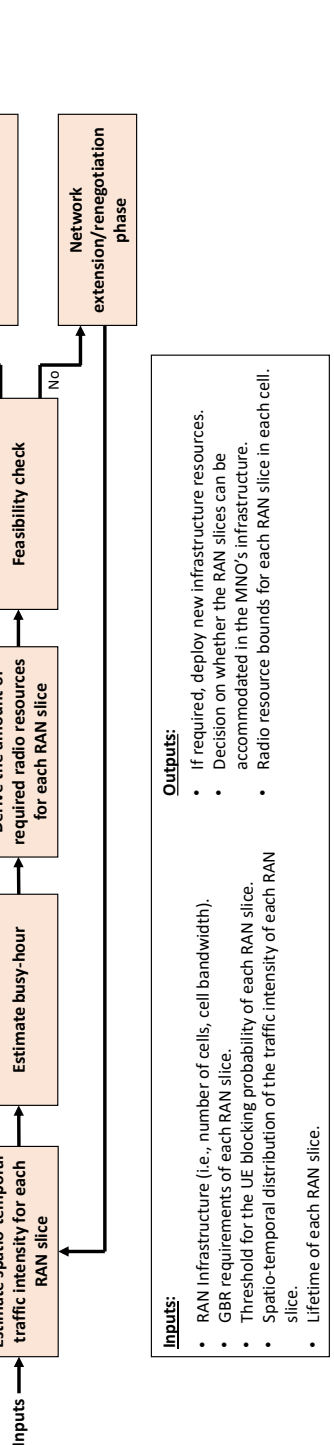


Figure 10.1: Planning procedure to accommodate the requested and already deployed RAN slices in a time window.

hours, to several days, or even weeks. Among other things, this duration mainly depends on how frequently the MNO receives RAN slice requests.

Considering a single time window, we illustrate a potential realization of the planning procedure for multiple RAN slices in Fig. 10.1. The MNO takes as inputs the performance requirements of all the RAN slices. In turn, the outputs are: (a) the amount of RAN slices which can be accommodated; (b) if required, the new infrastructure resources which would be instantiated before deploying the requested RAN slices; and (c) the configuration of the RAN infrastructure which includes, among other things, the radio resource bounds considered by the DRP algorithm.

To obtain these outputs, the MNO must translate the performance requirements of each RAN slice into RAN resources. To that end, the MNO needs to estimate the spatio-temporal traffic intensity experienced by each RAN slice in the considered time window. With this information, the MNO can determine the busy hour, i.e., the time period when the RAN infrastructure suffers the worst-case inter-cell interference.

Considering the busy hour, the MNO derives the amount of radio resources required by each RAN slice in each cell. The MNO then checks if the performance requirements of all the RAN slices are met with such allocation. If the checking procedure fails, a network extension/renegotiation phase starts. In this phase, the MNO should consider at least one of the following options: (a) adding more radio resources to the RAN infrastructure; (b) renegotiating the Service Level Agreements (SLAs) with one or more tenants; or (c) rejecting the least attractive RAN slices. This phase ends when a successful checking procedure is reached, and both the MNO and the tenants sign the SLAs. As a result, the MNO uses the radio resource allocation performed in the checking procedure to determine the radio resource bounds which will be used by the DRP algorithm.

Unlike the AC procedure, the RAN slice planning described before addresses more accurately the management tasks which a MNO must execute before deploying RAN slices. These management tasks have already been identified by the research community and they can be found, for instance, in [11].

This paper addresses the radio resource planning of multiple RAN slices. Specifically, it proposes a RAN slicing model which allows the MNO to plan in advance the deployment of requested RAN slices along with the existing ones

for a time window. We assume all the RAN slices accommodate communication services with Guaranteed Bit Rate (GBR) requirements. We consider each RAN slice supports User Equipment (UE) session arrivals following a Poisson distribution. In the proposed framework, the MNO should satisfy the GBR requirements of such sessions with an upper bound on the UE blocking probability. Under this context, the specific contributions of this work are:

- We address the radio resource allocation problem from the perspective of RAN slice planning. To that end, we provide a step-by-step description about the role of the RAN slicing architectural framework in the radio resource planning for multiple RAN slices.

- We propose a mathematical framework, denominated RAN Slice Planner, capable of translating the GBR requirements of the requested communication services into the minimum radio resource bound, hereinafter referred as quota, assigned for each RAN slice in each cell. Each quota guarantees the UE blocking probability for a RAN slice in a cell is below an upper bound under the inter-cell interference levels presented in the busy hour.

- We use game theory to model the radio resource planning in RAN slicing. Specifically, we formulate our problem as multiple ordinal potential games, one per RAN slice. In each game, the players are the cells and their actions are the allocation of radio resources for each considered RAN slice. The goal of each game is to guarantee the GBR requirements of each considered RAN slice, while its UE blocking probability in each cell is below the upper bound. We also demonstrate the existence of a Nash Equilibrium (NE) solution [12].

- We design novel strategies to solve the formulated problem. These strategies are based on better response dynamics and aim to minimize the UE blocking probability for all the RAN slices.

To evaluate the effectiveness of the proposed RAN Slice Planner, we perform detailed simulations and compare the obtained results with two reference solutions. Specifically, we analyze the performance of the RAN Slice Planer when handling multiple RAN slices with different traffic patterns in terms of spatio-temporal distribution and intensity. Furthermore, we analyze how the proposed

RAN Slice Planner is able to accommodate more RAN slices than the reference solutions. Finally, we also evaluate the renegotiation capability of the proposed solution by considering the requested RAN slices cannot be accommodated into the RAN infrastructure in a first attempt.

The remainder of this article is organized as follows. Section 10.2 provides the related works. Section 10.3 presents the RAN slicing framework on which our problem is formulated. Section 10.4 describes the system model. In Section 10.5, we formulate our problem as multiple ordinal potential games. Section 10.6 describes the proposed strategies for RAN slicing planning. Section 10.7 provides the performance results. Finally, Section 10.8 summarizes the conclusions.

10.2 Related Works

Most of the available literature on RAN slicing focus on AC and DRP.

Regarding AC, there exists multiple works in the literature, e.g., [2, 13, 14, 15, 16, 17, 18]. In such works, the authors assume the MNO receives requests for deploying RAN slices following a Poisson distribution. Furthermore, they also assume an exponential distribution for the lifetime of each RAN slice. In our work, instead of assuming the MNO triggers an AC when a new request arrives, we assume the MNO periodically plans in advance the deployment of one or more requested RAN slices.

Concerning DRP, the literature is vast e.g., [4, 6, 10, 19, 20, 21, 22, 23, 24, 25, 26, 27]. In these works, the authors assume the traffic demands of running RAN slices dynamically changes throughout their lifetimes. Under this scenario, the authors provide mechanisms to reallocate the available radio resources with the purpose of minimizing the SLA violations for each RAN slice. Despite their valuable contributions, these works omit how the MNO computes the radio resource quotas for such dynamic radio resource assignments.

The previous solutions analyze the radio resource allocation problem from the network operation perspective. This means they omit the RAN slice planning. From the network planning viewpoint, works such as [28, 29] address the network slice planning considering the core network. Focusing on RAN, there exists works such as [30, 31] which address the cell planning considering a multi-tenant environment, and others as [32, 33] which analyze spectrum sharing strategies for

different RAN slices. However the problem of deploying RAN slices following a network planning approach has not been addressed yet.

Regarding the scenarios considered by the works which address the radio resource allocation problem in RAN slicing, works such as [2, 5, 14, 18, 19, 20, 24, 25, 27, 34, 35] only consider single-cell environments. This means their solutions do not capture the impact of inter-cell interference when radio resources are reallocated in the neighbor cells. Despite others works such as [4, 6, 22, 26] focus on multi-cell environments, they also omit the impact of inter-cell interference by considering just gaussian noise. In [4, 6, 22, 26] the inter-cell interference is considered. However, the provided expressions for the Signal-to-Interference-plus-Noise Ratio (SINR) do not capture the changes in the load of each cell when radio resources are reallocated for every RAN slice.

Focusing on GBR services, several works as [17, 19, 20] provides novel solutions to allocate radio resources among several RAN slices while GBR requirements are met. These proposals assume the UEs could consume more data rate than the GBR. In our proposal, we assume each UE just consumes the GBR defined in the SLA. Some of these works have also evaluated the UE blocking probability for each RAN slice. To that end, they have modeled the UE session generation and release by a Markov process. This means these solutions must assume an exponential distribution for the UE session duration. In our work, we go beyond by considering generic distributions for the UE session duration.

Game theory has been widely used for modeling the radio resource allocation in RAN slicing. In [22], the authors use a weighted congestion game to perform user-cell association and distribute the available radio resources among several tenants based on the level of their financial contribution to the wireless network infrastructure. In [4, 6, 36], the same authors use Fisher market to model the radio resource allocation for non-GBR and GBR RAN slices. Unlike our proposal, their solutions do not guarantee the UE blocking probability for each RAN slice in each cell is below a certain upper bound. In [37], the authors use matching theory to address the radio resource allocation in RAN slicing. Despite its valuable contributions, this work does not consider the impact of inter-cell interference levels and the establishment of an upper bound for the UE blocking probability.

Finally, there exist solutions based on potential games to allocate radio resources e.g., [38, 39], however their do not consider RAN slicing. To the best of

our knowledge, we are the first of using potential games as a mechanism to solve the radio resource planning in RAN slicing.

10.3 RAN Slicing Framework

To compute the minimum radio resource quotas for the requested RAN slices and recompute them for the existing RAN slices, the MNO relies on the architectural framework depicted in Fig. 10.2. This framework is well aligned with most proposals from the literature [2, 40, 41, 42], as well as the leading Standards Developing Organizations (SDOs) on RAN slicing, e.g., 3GPP, or Global System for Mobile Communications Alliance (GSMA) [43, 44]. Considering this framework, we provide a step-by-step description of the main procedures which impact on the radio resource planning executed for a planning window.

When a planning period starts, the MNO must process the deployment requests for one or more communications services. At this point, it is crucial for the MNO to define an unified ability to (a) interpret the requirements from different tenants, and (b) represent them in a common language. In this regard, the GSMA has developed a universal network slice blueprint that provides a point of convergence between the MNO and the tenants on network slicing understanding. This blueprint, known as Generic Slice Template (GST), contains a set of attributes that can be used to characterize the communication service to be accommodated by a network slice [44, 45].

Focusing on Fig. 10.2, we assume the tenants have available the GST's attributes to fill them in a customized way (step 1). Alternatively, these attributes could be totally or partially filled by the MNO. In any case, when all the attributes are filled (step 2), the requirements of a specific communication service are gathered in the Network Slice Type (NEST). Different NESTs allow describing different types of network slices, which are registered and published in the MNO's service catalog.

Once the MNO has available the NESTs associated to the requested communication services (step 3), the Product Order Manager located in the Business Support System (BSS) has to map the NEST attributes with the slicing information models defined by the 3GPP. Specifically, the S/P-NEST attributes are translated into the service profile (step 4) [8]. The service profile is just an

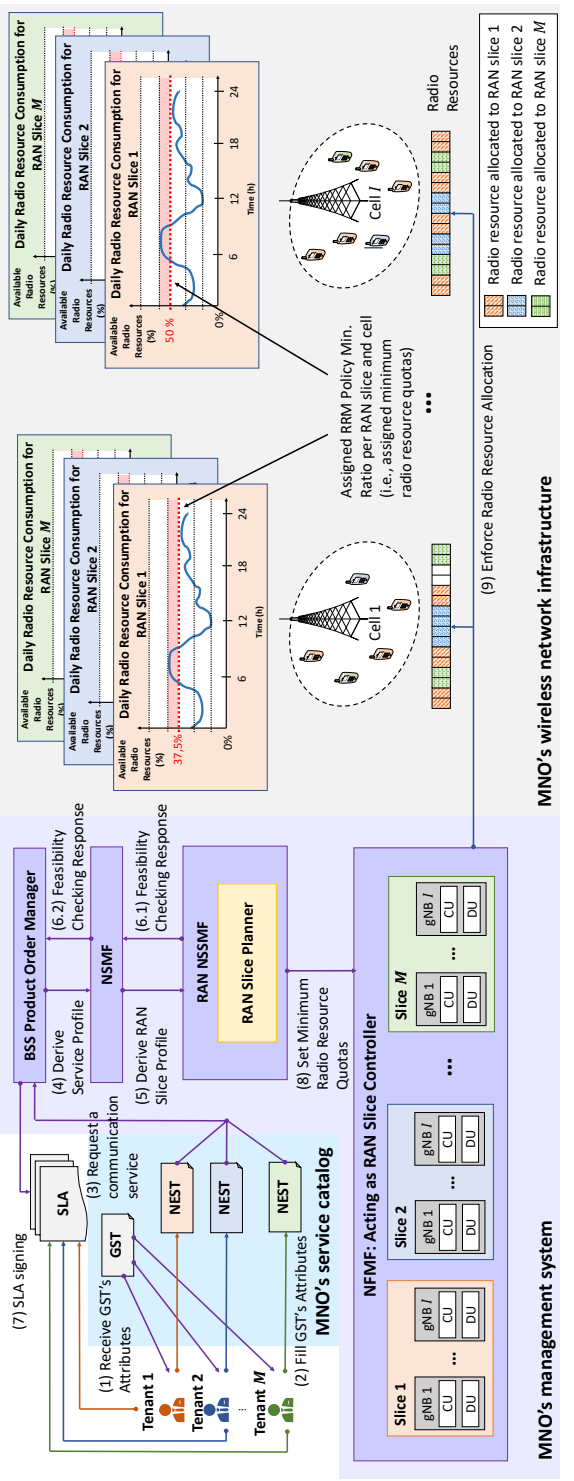


Figure 10.2: Role of the RAN Slicing architectural framework in a scheduled-based radio resource planning for RAN slices

Table 10.1: RAN slice profile's attributes considered as inputs for the proposed RAN Slice Planner

RAN slice profile's attributes defined in [8]	Definition	Aspects considered by the proposed RAN Slice Planner
Service Area	This parameter specifies the area where the UEs can access a particular communication service. There are two ways of describe the coverage area: a) based on cell location or b) based on geographical partitioning.	Geographical partitioning is more reasonable because the MNO has not to expose tenants the cell location. The disadvantage is the MNO has to translate the geographical partitioning into the specific cells. In our work, the RAN Slice Planner performs this translation.
Downlink (DL) throughput per UE	This parameter defines the downlink data rate for a UE served by a specific network slice. Two types are defined: a) guaranteed downlink throughput and b) maximum downlink throughput.	In this work, we assume all the UEs served by a RAN slice consume a specific GBR for downlink traffic. Specifically, we consider these UEs generate sessions following a Poisson distribution. If a UE session cannot consume the GBR, it is rejected (i.e., blocked).
UE density	This parameter describes the maximum number of connected and/or accessible devices per unit area supported by the network slice.	The proposed RAN Slice Planner goes beyond by considering the UE spatial distribution instead of the UE density. The reason is UE density could be different along the cells which comprise the entire coverage area. A single value for the UE density is insufficient to describe this variability. To measure the accessibility, we also consider an upper bound for the UE blocking probability of a RAN slice in each cell. Using this parameter, the RAN Slice Planner can guarantee a percentage of accessible UE sessions for each RAN slice. We also assume this parameter is set up by the MNO and is the same for all the RAN slices.
Maximum Number of UEs	This parameter describes the maximum number of UEs that can be connected to a specific network slice simultaneously.	In a first instance, the MNO considers this parameter to estimate the busy hour. Later, the RAN Slice Planner limits the maximum number of UEs by imposing the upper bound for the UE blocking probability, This means that the maximum number of UEs will depend on (a) the radio channel conditions and (b) the radio resource quotas assigned for each RAN slice.

adaptation from the description language used in the GST. In turn, the Network Slice Management Function (NSMF) has to translate the attributes of the service profile into the requirements supported in each network segment. Focusing on the RAN, this procedure results in the definition of the RAN slice profile (step 5). The information gathered in the RAN slice profile will be used by the RAN Network Slice Subnet Management Function (NSSMF) to manage and orchestrate the RAN slices throughout their lifetimes. Within this management entity, the proposed RAN Slice Planner could perform those tasks required to derive (or recompute) the minimum radio resource quotas for the requested (or already deployed) RAN slices. To that end, the RAN Slice Planner needs to take as input the parameters defined in the RAN slice profile. Table 10.1 describes those parameters which we consider in this work.

Regarding the radio resource quotas, the RAN Slice Planner could use different types in function of the policies imposed by the MNO. The 3GPP has standardized three policies denoted as Radio Resource Management (RRM) Policy Ratios [8, 9]:

- RRM Policy Dedicated Ratio (optional policy): It defines the dedicated radio resource quota for the associated RAN slice, i.e., its dedicated radio resources. These radio resources cannot be shared even if the associated RAN slice does not use them throughout its lifetime.

- RRM Policy Minimum Ratio (mandatory policy): It defines the minimum radio resource quota for the associated RAN slice, including prioritized radio resources and dedicated radio resources. Prioritized radio resources are those which are preferentially used by the associated RAN slice. When prioritized radio resources are not used by the associated RAN slice throughout its lifetime, other RAN slices could use them.

- RRM Policy Maximum Ratio (mandatory policy): It defines the maximum radio resource quota for the associated RAN slice, including shared radio resources, prioritized radio resources and dedicated radio resources. Shared radio resources are those which are shared among all RAN slices. This means the shared radio resources are not guaranteed for use by the associated RAN slice throughout its lifetime.

In this work, we focus on the RRM Policy Minimum Ratio. Specifically, the RAN Slice Planner translates the RAN slice profile's attributes defined in Table 10.1 for each RAN slice into the minimum radio resource quotas adopted in each cell. To that end, the RAN Slice Planner needs to consider the lifetimes of all the RAN slices, which could be partially overlapped over the planning window; and estimate the spatio-temporal traffic intensity experienced by each RAN slice in such window. With this information, the RAN Slice Planner determines the busy hour and performs a radio resource allocation to derive the amount of radio resources required by each RAN slice in each cell. Then, the RAN Slice Planner checks if the performance requirements for all the RAN slices are met with this allocation and sends the corresponding feasibility checking response to the tenants which requested a RAN slice (step 6).

At this point, there exists two scenarios. In the first scenario, all the RAN slices are perfectly accommodated into the RAN infrastructure. In this case, each tenant and the MNO signs the SLA (step 7). In the second scenario, one or more RAN slices cannot be accommodated. This means the derived amount of radio resource involves UE blocking probabilities above the upper bound defined by the MNO. In such case, the correspond tenants either re-defines the requirements of their communication services (i.e., less ambitious requirements), or tries to request other MNOs their communication services. At this point, note that the MNO could decide to add new resources in its RAN infrastructure. In any case, the steps 2-6 are re-executed in this scenario until obtaining a successful checking procedure. Then, the MNO and the tenants signs the SLAs (step 7).

When the SLAs are signed, the MNO uses the radio resource allocation performed in the checking procedure to determine the minimum radio resource quotas. Then, the RAN NSSMF sends these quotas (step 8) to the Network Function Management Function (NFMF). Finally, the NFMF enforces the radio resource allocation performed in each cell during the operation of each RAN slice meets the bounds imposed by these quotas (step 9). This means each RAN slice will have available at least the amount of radio resources defined in these quotas if the traffic demand will required them.

10.4 System Model

In this work, we focus on the downlink operation of a 5G-New Radio (NR) multi-cell environment with several RAN slices. Each RAN slice provides a GBR service to their UEs, which dynamically request and release data sessions. Furthermore, each cell supports Link Adaptation (LA), thus these cells consider the channel quality perceived by each UE to allocate them radio resources. Under this scenario, we first describe the network model. Then, we present the model for the radio resources. Next, we describe the channel model of a single cell. Finally, we define the characteristics of the offered traffic.

10.4.1 Network Model

We consider a MNO owns a RAN consisting of a set $\mathcal{C}$ of 5G NR cells. Before the MNO initiates a planning period, we assume: (a) multiple tenants have requested in advance one or more communication services; and (b) there exist RAN slices which are currently running in the RAN infrastructure. Defining $\mathcal{M}$ as the set of requested and deployed RAN slices, the MNO will execute a RAN slice planning procedure with the aim of checking the feasibility of accommodating these RAN slices, each with specific GBR requirements over a certain subset $\mathcal{C}^m \subseteq \mathcal{C}$ of cells.

The traffic demand of each RAN slice is non-uniformly distributed over the considered RAN. Accordingly, the area of each cell has a different size to absorb the aggregated traffic demand from all the RAN slices with a maximum usage efficiency. In this work, we consider the cell location has been already established by the MNO. Specifically, we have adopted the algorithm proposed in [46] to determine the location and size of each cell. Under this scenario, a set $\mathcal{U}$ of UEs exist, being (a) $\mathcal{U}^m \subseteq \mathcal{U}$ the subset of UEs served by the RAN slice m; (b) $\mathcal{U}_i \subseteq \mathcal{U}$ the subset of UEs served by the cell $i \in \mathcal{C}$; and c) $\mathcal{U}_i^m = \mathcal{U}^m \cap \mathcal{U}_i$ the intersection of both subsets.

10.4.2 Radio Resource Model

We assume Orthogonal Frequency-Division Multiple Access (OFDMA) as accessing scheme. Focusing on a single cell i, it supports a total bandwidth W_i. In turn, this bandwidth is divided into N_i OFDM sub-carriers, which are grouped

in groups of $N_{SC} = 12$ sub-carriers. Each group defines a Resource Block (RB), which is the smallest unit of resources that can be allocated to a UE. The number of available RBs on average during a time slot is given by Eq. (10.1). Since a 5G NR cell supports scalable numerologies ($\mu = 0,1, ..., 4$), the subcarrier spacing is computed as $\Delta_f = 2^{\mu} \cdot 15$ KHz. The parameter OH denotes the overhead factor due to control plane data [47].

$$N_{RB,i}^{slot} = \left\lfloor \frac{W_i}{N_{SC}\Delta f} (1 - OH) \right\rfloor . \tag{10.1}$$

In a single carrier, the number of RBs could range from 11 to 273 units [48] [49]. This means $N_{RB,i}^{slot}$ could be too high if the cell i employs a small numerology and a large bandwidth. Then, from the perspective of radio resource allocation in RAN slicing, it becomes advantageous to reduce the management complexity by grouping the RBs into resource chunks, which are allocated to the RAN slices as indivisible units [33]. This can be done through the concepts of Bandwidth Part (BWP) and Resource Block Group (RBG) defined in [50] and [51], respectively. A BWP is a continuous set of RBs for a given numerology. A RBG is a collection of consecutive RBs within a given BWP that can be allocated to a specific UE. The size of the RBG, i.e., herein denoted as R_{size}, can be used for establishing the minimum allocation unit size. Increasing R_{size} may serve to reduce the signaling overhead at the expense of a loss of flexibility, which could be critical when the number of RAN slices to be planned is large. Under these considerations, we denote (a) $\mathcal{R}_i$ as the set of RBGs in cell i, (b) $\mathcal{R}_i^m \subseteq \mathcal{R}_i$ as the subset of RBGs allocated to the slice m in cell i; and (c) $\mathcal{R}^u \subseteq \mathcal{R}_i^m$ as the subset of RBGs allocated to an UE u which is served by the RAN slice m in the cell i. Finally, we can compute the available RBGs on average during a time slot in the cell i as $R_i^{slot} = \left\lfloor N_{RB,i}^{slot}/R_{size} \right\rfloor$. Note that the sum of RBGs allocated for each RAN slice, i.e., $R_{i.m}^{slot}$, must be less or equal than R_i^{slot}.

10.4.3 Channel Model

To measure the channel quality within each cell, we consider the average SINR. Specifically, we define $\gamma_{u,r}$, as the average SINR measured by the UE $u \in \mathcal{U}$ in the RBG $r \in \mathcal{R}_i^m$ (see Eq. (10.2)). The parameter P_i^{RX} denotes the received power. This power results from the transmitted power minus the attenuation suffered

by the shadow fading and the path loss. The fast fading is not modeled since the average SINR is measured over a large time scale. Note that we assume the same transmitted power for all the RBGs. The parameter $\Gamma(u)$ is a function that returns the cell $i \in \mathcal{C}^m$ where the UE u served by the RAN slice m is attached. This cell is the one where this UE receives the strongest average SINR. Finally, the parameter $I_{u,r,i}$ denotes the interference suffered by the UE u in the RBG r, and P_N is the noise power measured in one RBG.

$$\gamma_{u,r} = \frac{P_i^{RX}}{I_{u,r,i}+P_N} \quad | \, i = \Gamma(u). \tag{10.2}$$

The interference $I_{u,r,i}$ is provided in Eq. (10.3). This parameter is split into two summations, each gathering the intra-slice and inter-slice interference terms, respectively. An interference term j is intra-slice when the RBG r from neighbor cell j is allocated to the same RAN slice m which serves the user u in the cell i. An interference term j is inter-slice when the RBG r from neighbor cell j is allocated to a RAN slice n different from the slice m. To identify these terms, we use the binary variable $\delta_{u,r,j}$. It takes the value 1 when the interference term is intra-slice and the value 0 otherwise.

$$I_{u,r,i} = \sum_{j \in \mathcal{C} \setminus \{i\}} L_{j,r} \alpha_{j,r} P_j^{RX} \delta_{u,r,j} + \sum_{j \in \mathcal{C} \setminus \{i\}} L_{j,r} \alpha_{j,r} P_j^{RX} (1 - \delta_{u,r,j}). \tag{10.3}$$

The parameter $\alpha_{j,r}$ is also a binary variable that takes the value 1 when the RBG r is allocated to the neighbor cell j and the value 0 otherwise. The value for $\alpha_{j,r}$ will depend on the radio resource allocation performed by the RAN Slice Planner in each neighbor cell. Finally, $L_{j,r}$ denotes the cell load factor, which is given by Eq. (10.4). In this equation, $\beta(j,r)$ is a function that indicates the RAN slice m for which the RBG r from cell j has been allocated. The parameter th_m denotes the data rate consumed by an UE attached to this RAN slice. In this work, we assume the same GBR for each UE of a specific RAN slice. Note that the number of UEs served by the RAN slice m in cell j is given by $|\mathcal{U}_j^m|$. The parameter $SE_{u,r}$ is the average data rate per bandwidth unit (i.e., spectral

efficiency) for the UE u in the RBG r. Unlike our previous work [33], we consider that only the RBGs allocated to a specific RAN slice can be scheduled to the UEs attached to this RAN slice. This means each RAN slice m produces a different load in a specific cell in function of its GBR requirements and the number of allocated RBGs.

$$\hat{L}_{j,r} = \frac{|\mathcal{U}_j^m| th_m}{N_{SC}\Delta_f \sum_{u\in\mathcal{U}_j^m}\sum_{r\in\mathcal{R}^u} SE_{u,r}} \mid m = \beta(j,r). \tag{10.4a}$$

$$L_{j,r} = \min\left(\hat{L}_{j,r}, 1\right). \tag{10.4b}$$

The average spectral efficiency $SE_{u,r}$ is recursively derived from $\gamma_{u,r}$ as Eq. (10.5) shows. The parameter SE_{max} denotes the maximum achievable spectral efficiency with LA, γ_{min} and γ_{max} the minimum and maximum average SINR values, respectively. Finally, σ is an attenuation factor due to implementation losses [52].

$$SE_{u,r} = \begin{cases} 0, & \gamma_{u,r} < \gamma_{min}; \\ \sigma \cdot \log_2\left(1+\gamma_{u,r}\right), & \gamma_{min} \leq \gamma_{u,r} < \gamma_{max}; \\ SE_{max}, & \gamma_{u,r} > \gamma_{max}; \end{cases} \tag{10.5}$$

10.4.4 Traffic Model

Regarding the arrival rate of UE sessions for RAN slice m, we assume an average of λ_m requests per unit time following a Poisson distribution. Since a Poisson process can be split into independent processes [53], we can also express the average arrival rate for each cell as $\lambda_{i,m} = \omega_{i,m}\lambda_m$. The variable $\omega_{i,m}$ denotes the probability an UE $u \in \mathcal{U}^m$ is served by the cell i. This probability will depend on (a) the UE spatial distribution in the entire RAN; and (b) the average SINR perceived by each UE from each cell.

With respect to the session duration $t_{u,m}^{ses}$ for each UE u served by the RAN slice m, we assume a random variable extracted from an arbitrary distribution. This means we could consider a different distribution for each RAN slice. Additionally, we define $\mu_m = 1/E\left[t_{u,m}^{ses}\right]$ as the average rate for releasing UE sessions per unit time of the RAN slice m.

Defined $\lambda_{i,m}$ and μ_m, we compute the average offered traffic intensity for the RAN slice m in each cell i as $\rho_{i,m} = \lambda_{i,m}/\mu_m$. Furthermore, the total averaged offered traffic intensity for this RAN slice is also computed as $\rho_m = \sum_{i'=1}^{|\mathcal{C}^m|} \rho_{i',m}$.

Finally, to model the probability of blocking the data session of an UE $u \in \mathcal{U}_i^m$ served by the RAN slice m in cell i, i.e., $B_{i,m}$, we use the analytical model which we proposed in [54]. In this model, we consider a multi-dimensional Erlang-B system where each dimension represents a region of the cell with a specific average SINR. Specifically, this model considers as inputs: (a) a discrete set of values for $\gamma_{u,r}$ (and thus, for the spectral efficiency $SE_{u,r}$); (b) the probability that an UE session perceives a specific value for $\gamma_{u,r}$; and (c) the amount of radio resources allocated for the RAN slice to compute the UE blocking probability. For more detailed information about how these input parameters impact on the UE blocking probability, see [54].

10.5 Radio Resource Planning Based on Ordinal Potential Games

In this work, we analyze the radio resource planning for RAN slices providing GBR services. In this procedure, each RAN slice in each cell needs a greater amount of radio resources to obtain a better performance, e.g., a lower UE blocking probability. If some radio resources are allocated for a RAN slice in a cell, these radio resources should not be allocated to the other RAN slices in the nearest cells because of interference. If they are allocated, such RAN slices will have performance degradation, e.g., a greater UE blocking probability. This makes the MNO must consider each RAN slice in each cell as a selfish entity because RAN slices need more resources to obtain a better performance.

To model the considered scenario, game theory is suitable. Specifically, in our proposal a selfish entity is a single cell instead of the tuple defined by one cell and one RAN slice. The reasons are: (a) this simplifies the game, i.e., less players; and (b) we have a disjoint resource allocation for each RAN slice in each cell, i.e., two or more RAN slices cannot have allocated the same radio resources in a cell.

10.5.1 Problem Formulation

As Eq. (10.6a) shows, the goal of our planning process is to minimize the average UE blocking probability $\overline{B}_{m'}$ of the RAN slice m' which has the highest value for this parameter. For each RAN slice m, the average UE blocking probability can be computed as $\overline{B}_m = \sum_{i \in \mathcal{C}^m} \omega_{i,m} B_{i,m}$. The constraints given in Eq. (10.6b) enforce the UE blocking probability $B_{i,m}$ for each RAN slice m in each cell i is below the upper bound B^{th}. We assume this bound on the UE blocking probability is established by the MNO before receiving any RAN slice request and it is the same for all the RAN slices.

$$\min_{\mathcal{R}_i^m} \quad \max\left(\overline{B}_1, \overline{B}_2, \ldots, \overline{B}_m, \ldots, \overline{B}_M\right) \quad \forall i \in \mathcal{C};\ \forall m \in \mathcal{M}. \tag{10.6a}$$

$$s.t. \quad B_{i,m} \leq B^{th}. \tag{10.6b}$$

Solving the formulated problem can be see as a combinatorial optimization, i.e., allocating specific radio resources for all RAN slices in each cell while the cost function is minimized. Performing an exhaustive search to find the optimal solution is not computationally tractable. As an alternative, searching a local optimum is a better option. By using game theory to model the formulated problem, we can find a local optimum by determining a NE solution. In this work, we model our problem as multiple ordinal potential games and demonstrate the existence of a NE solution.

10.5.2 Game Formulation

In game theory, a game is defined as $\mathcal{G} = [\mathcal{C}, \{S_i\}_{i \in \mathcal{C}}, \{\Phi_i\}_{i \in \mathcal{C}},]$ where $\mathcal{C}$ is the set of players participating in the game, S_i is the strategy selected by player i, and $\Phi_i : S \rightarrow \mathbb{R}$ is the utility function of that player, with S the strategy profile of the game (i.e., the set of strategies selected by all the players). If we refer to a single player, i.e., the ith player, then S can be rewritten as $S = (S_i, S_{-i})$, where S_{-i} denotes the joint strategy adopted by player i's opponents. In a game $\mathcal{G}$, each player will selfishly choose a new strategy T_i in its turn with the aim of improving its utility function considering the current strategies of the other

players. A game is an ordinal potential game if and only if a potential function $F(S)$ exists such that Eq. (10.7) is met, where sgn[·] denotes the signum function [55].

$$sgn\left[\Phi_i\left(T_i, S_{-i}\right) - \Phi_i\left(S_i, S_{-i}\right)\right] = sgn\left[F\left(T_i, S_{-i}\right) - F\left(S_i, S_{-i}\right)\right] \quad \forall i \in \mathcal{C}. \tag{10.7}$$

In our game, the set of players $\mathcal{C}$ are the cells where the requested RAN slices will be deployed and the existing RAN slices are running. For every ith cell, a strategy S_i consists of a specific RBG allocation for all the RAN slices which require the coverage of this cell. In turn, the utility function for each ith cell Φ_i is given by Eq. (10.8). Since the utility function is the same for all the players, our game could be seen as identical-interest game or perfect coordination game [56].

$$\Phi_i = \left(\max\left(\overline{B}_1, \overline{B}_2, \ldots, \overline{B}_m, \ldots, \overline{B}_M\right)\right)^{-1} \quad \forall m \in \mathcal{M}. \tag{10.8}$$

The potential function $F(S)$ is defined by Eq. (10.9). Since the utility function of each cell i is equal to the potential function, i.e., $\Phi_i\left(S_i, S_{-i}\right) = F\left(S_i, S_{-i}\right) \ \forall i \in \mathcal{C}$, it is easy to check that Eq. (10.7) is always met, thus an unconstrained game with the potential function $F(S)$ and these utility functions Φ_i is an ordinal potential game. Consequently, the proposed game always reaches a NE solution. Note that game theory states that if a game is a potential game, it always has a NE solution [55].

$$F(S) = \left(\max\left(\overline{B}_1, \overline{B}_2, \ldots, \overline{B}_m, \ldots, \overline{B}_M\right)\right)^{-1} \quad \forall m \in \mathcal{M}. \tag{10.9}$$

Defined the utility functions and the potential function, we formulate our constrained game $\mathcal{G}$ as Eq. (10.10) shows. The goal of this game is to determine the set of strategies S, i.e., the RBG allocation for each RAN slice in each cell, which maximize the potential function.

$$\begin{aligned} (\mathcal{G}): \quad \forall i \in \mathcal{C} \quad & \max_{S_i \in \mathcal{S}^i} \Phi_i\left(S_i, S_{-i}\right). \\ s.t. \quad & g_{i,m}\left(S_i, S_{-i}\right) \leq 0. \end{aligned} \tag{10.10}$$

We also assume there are $|\mathcal{C}| \cdot |\mathcal{M}|$ inequalities constrains in the form of $g_{i,m}(S) \leq 0$. Specifically, these constrains are expressed by Eq. (10.11). In [55], the authors proof that a constrained game is an ordinal potential game only if the equivalent game without constraints is also an ordinal potential game. This means that the proposed constrained game $\mathcal{G}$ is an ordinal potential game.

$$g_{i,m}(S) = B_{i,m} - B_m^{th} \quad \forall m \in \mathcal{M}, \forall i \in \mathcal{C}. \tag{10.11}$$

To perform the radio resource planning, the proposed RAN Slice Planner could follow two approaches: one-game-all, and consecutive games.

In the one-game-all approach, the RAN Slice Planner executes the game $\mathcal{G}$ with $M' = |\mathcal{M}|$ RAN slices. In this game, the starting point is the allocation of one RBG for each RAN slice in the cells where they require coverage.

In the consecutive games approach, the RAN Slice Planner executes $|\mathcal{M}|$ consecutive games as the one formulated in Eq. (10.10). In each mth game, only $M' = m$ RAN slices participate. Focusing on the first game, i.e., $M' = 1$, the RAN Slice Planner performs the RBG allocation for one RAN slice. In this game, the starting point is the allocation of one RBG in each cell where this RAN slice requires coverage. When the RAN Slice Planner executes the first game, it consider the derived RBG allocation as the starting point of the next game, where $M' = 2$ RAN slices participate. The RAN Slice Planner repeats this procedure until it executes the $|\mathcal{M}|$th game.

As we demonstrate in Section 10.7.2, the proposed game tends to equal the average UE blocking probabilities of all the RAN slices. This means that selecting a starting point of the game where the average UE blocking probabilities of all the RAN slices are closer could involve reaching a NE solution where all the average UE blocking probabilities are equaled and minimized, but the constraints given by Eq. (10.11) are not met. This scenario is frequent when the RAN Slice Planner follows the one-game-all approach. To avoid this issue, we consider the RAN Slice Planner follows the consecutive game approach.

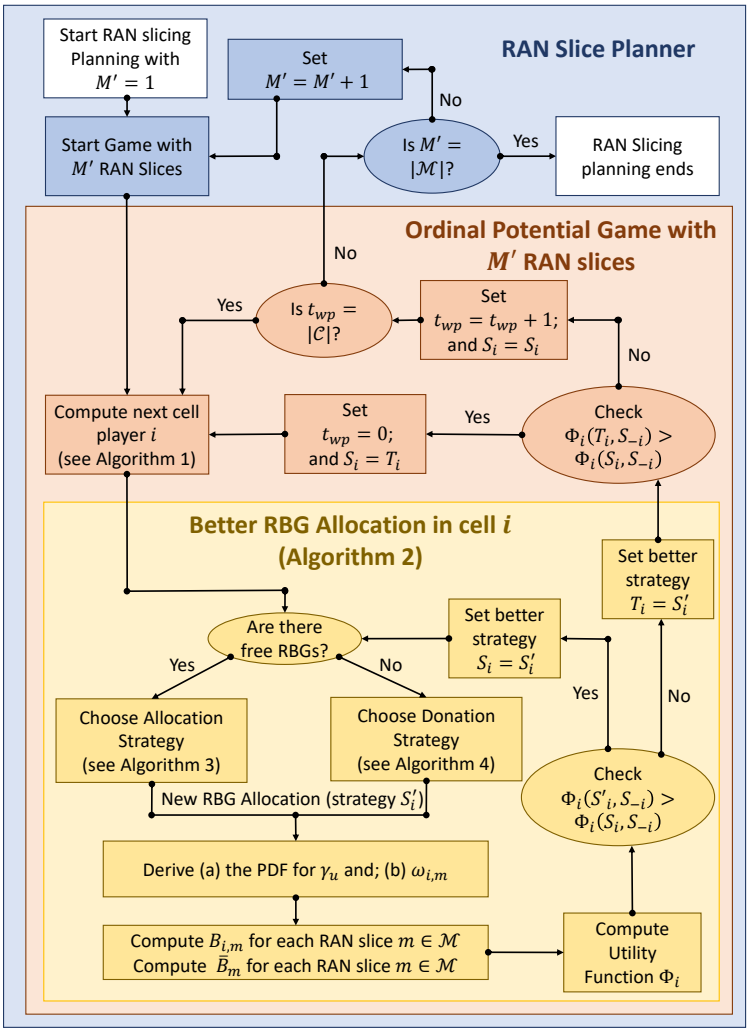


Figure 10.3: High-level view of the methods implemented by the RAN Slice Planner to perform the planning of $|\mathcal{M}|$ RAN slices

10.6 Planning Method Based on Better Response Dynamics

In Fig. 10.3, we illustrate a block diagram which summarizes the behavior of the proposed RAN Slice Planner. In a planning period, this mathematical framework executes $|\mathcal{M}|$ consecutive ordinal potential games. Focusing on a specific game $\mathcal{G}$

with $M' \leq |\mathcal{M}|$ RAN slices[1], the RAN Slice Planner selects the next cell player i and determines the strategy T_i which provides the better RBG allocation for each RAN slice in such cell. The method to determine such allocation is based on better response dynamics, i.e., the players proceed toward a NE solution via a local search method. Then, if the utility function Φ_i of this cell improves with respect to the previous strategy S_i, i.e., $\Phi_i(T_i, S_{-i}) > \Phi_i(S_i, S_{-i})$, the RAN Slice Planner considers T_i as the new strategy S_i in this cell and computes the next cell player. If the utility function Φ_i does not improve with respect to the previous strategy S_i, it remains as the better strategy for this cell. In such case, the RAN Slice Planner also increases by one a counter t_{wp} that considers the consecutive cell players which cannot improve their utility functions. The game $\mathcal{G}$ will end when none of the cell players can improve its utility function, i.e., $t_{wp} = |\mathcal{C}|$. If this happens, the RAN Slice Planner will execute more games until the number of executed games is $|\mathcal{M}|$. Then, the planning of RAN slices in this planning period will have ended. At this point, the RAN Slice Planner will have determined the minimum radio resource quotas for each RAN slice in each cell by considering the strategies $S_i \, \forall i \in \mathcal{C}$ derived in the $|\mathcal{M}|$th game.

In the following subsections, we provide details about the method used by the RAN Slice Planner to define the order in which each cell selects its better RBG allocation as well as the method to determine such allocation.

10.6.1 Method to Decide the Next Cell Player

In our game, we aim to maximize the potential function $F(S)$ using the minimum number of iterations, where each iteration correspond to the set of actions taken by a cell player to determine its better RBG allocation. When better response dynamics is used for computing the NE solution, the computational time to reach this solution strongly depends on the order by which players are chosen to perform their actions. Better response dynamics leaves unspecified the rules to define this order [57].

To minimize the required number of iterations for reaching the NE solution, the RAN Slice Planner executes the Algorithm 1, which decides the next cell player. First, this algorithm determines the RAN slice m' which provides the

[1]The order in which the RAN slices enter into a game is out of the scope.

Algorithm 1: Computing the next cell player i'

1 **Inputs:** $\overline{B}_m$, $B_{i,m}$, and $\omega_{i,m}$ $\forall m \in \mathcal{M}$ $\forall i \in \mathcal{C}$;
2 **Determine** $m' = \arg\max\big[\big(\overline{B}_1, ..., \overline{B}_{|\mathcal{M}|}\big)\big]$;
3 **Compute** $i' = \arg\max\big[\big(\omega_{1,m'}B_{1,m'}, ..., \omega_{|\mathcal{C}|,m'}B_{|\mathcal{C}|,m'}\big)\big]$;

maximum average UE blocking probability. Then, considering the weighted UE blocking probability for the selected RAN slice m' in each cell (i.e., $\omega_{i,m'}B_{i,m'}$), the algorithm selects as the next player i' the cell where the weighted UE blocking probability is maximum. Using the proposed algorithm, the RAN Slice Planner selects the cell where the RAN slice which provides the value for the potential function, see Eq. (10.9), has the worst weighted UE blocking probability. In this way, the RAN Slice Planner can reallocate RBGs in such cell with the goal of maximizing the potential function much faster.

10.6.2 Better RBG Allocation in the Cell Player

The Algorithm 2 provides the steps performed by the RAN Slice Planner to select the better RBG allocation (i.e., strategy T_i) in the cell player i. These steps are also depicted in Fig. 10.3. First, the RAN Slice Planner checks if there are free RBGs in the cell i, i.e., those RBGs which has not been allocated for any RAN slice. If yes, one free RBG will be allocated to RAN slice m' (i.e., the RAN slice derived by Algorithm 1). The Algorithm 3 details how the RAN Slice Planner selects this RBG (see Section 10.6.3). If there are not free RBGs, the only way to reduce the average UE blocking probability for RAN slice m' is to allocate it one RBG from another RAN slice. To than end, the RAN Slice Planner determines the RAN slice m'' which has the lowest weighted UE blocking probability in cell i. Then, one RBG is donated from RAN slice m'' to RAN slice m'. The Algorithm 4 details how the RAN Slice Planner determines the donated RBG (see Section 10.6.4).

After the RAN Slice Planner uses Algorithm 3 or Algorithm 4, it derives a new strategy S_i' resulted from the RBG reallocation in cell i. Based on that, the RAN Slice Planner computes $|\mathcal{C}| \cdot |\mathcal{M}|$ Probability Density Functions (PDFs) of the average SINR experienced by an arbitrary UE (i.e., $f_{PDF}\left(\overline{\gamma}_u\right)$), one per each pair of RAN slice and cell. During this procedure, the probability that an

Algorithm 2: Computing the better RBG allocation, i.e., strategy T_i, for cell player i

Initialization: RBG allocation for each RAN slice in cell i, i.e., S_i;
found_better_strategy = false;
while *found_better_strategy == false* **do**
  if *free_RBGs == true* **then**
    Allocate one RBG to RAN slice m' (see Algorithm 3) $\rightarrow$ New strategy S'_i;
  else
    Compute $m'' = \arg\min(\omega_{i,m} B_{i,m})\ \forall m \in \mathcal{M} \setminus \{m'\}$;
    Donate one RBG from RAN slice m'' to RAN slice m' (see Algorithm 4)$\rightarrow$ New strategy S'_i;
  end
  From $S = (S'_i, S_{-i})$, derive $f_{PDF}(\overline{\gamma}_u)$ for each pair of RAN slice an cell. $\omega_{i,m}$ is also derived $\forall i \in \mathcal{C}$ and $\forall m \in \mathcal{M}$;
  Compute $B_{i,m}\ \forall i \in \mathcal{C}, \forall m \in \mathcal{M}$;
  Compute $\overline{B}_m\ \forall m \in \mathcal{M}$;
  Compute $\Phi_i(S'_i, S_{-i})$;
  if $\Phi_i(S'_i, S_{-i}) > \Phi_i(S_i, S_{-i})$ **then**
    $S_i = S'_i$;
  else
    $T_i = S_i$;
    found_better_strategy = true;
  end
end
return: T_i

arbitrary UE is attached to a specific cell $\omega_{i,m}$ is also recomputed. The proposed algorithm uses the strongest SINR as the criteria to attach each UE to a specific cell.

After deriving these PDFs, the RAN Slice Planner computes the UE blocking probability for each RAN slice in every cell, i.e., $B_{i,m}$ by using the model we proposed in [54]. Then, the RAN Slice Planner computes the mean UE blocking probability for each RAN slice, and thus it derives the new value for the utility function $\Phi_i(S'_i, S_{-i})$ of the cell i. Next, the RAN Slice Planner compares the new value of the utility function with respect to the previous one, i.e., when cell i uses the old RBG allocation S_i. If the utility function improves, then the

Algorithm 3: Allocation of one RBG r' to the RAN slice m'

Initialization: $dist_{cells}(i') = \infty \ \forall i' \in \mathcal{C} \setminus \{i\}$;
for $r \in \mathcal{R}_i^{free}$ **do**
  for $i' \in \mathcal{C} \setminus \{i\}$ **do**
    if $r \notin \mathcal{R}_{i'}^{free}$ **then**
      Compute $dist_{cells}(i') = ED\,(i, i')$;
    end
  end
  Compute $dist_{RBG}(r) = \min(dist_{cells}(i'))$;
end
Compute $r' = \arg\max(dist_{RBG}(r))$;
return: r'

new RBG allocation is considered as the valid strategy (i.e., $S_i = S_i'$). In this case, these steps (i.e., from step 4 to step 19) are repeated until the RAN Slice Planner cannot improve the utility function. When this happens, the RAN Slice Planner ends the execution of Algorithm 2 and the better RBG allocation is the one derived in the previous iteration, i.e., T_i.

In the following subsections, we provide details about the steps performed by Algorithms 3 and 4 to reallocate the RBGs in a cell player.

10.6.3 RBG Allocation

To allocate one RBG r' to the RAN slice m', the RAN Slice Planner executes the steps described in Algorithm 3. First, the RAN Slice Planner initializes to infinity the vector $dist_{cells}$. Focusing on one RBG of those available in the cell, i.e., $r \in \mathcal{R}_i^{free}$, the RAN Slice Planner checks if this RBG has been allocated in the neighbor cells. If yes, $dist_{cells}$ stores the euclidean distance, given by the function $ED\,(\cdot)$, between the cell i and the neighbor cell i'. When this task is performed for all the neighbor cells, the RAN Slice Planner selects the minimum distance gathered in $dist_{cells}$. The goal is to determine the closest neighbor cell which induces an interference term into the RBG r. Repeating this procedure for all the free RBGs, the RAN Slice Planner determines the RBG r' which suffers the most significant interference term (i.e., from the same RBG in the closest neighbor cell) with the lowest received power.

Algorithm 4: Donation of one RBG r' from RAN slice m'' to RAN slice m'

Initialization: $dist_{cells}(i') = \infty \ \forall i' \in \mathcal{C} \setminus \{i\}$;
for $r \in \mathcal{R}_i^{m''}$ **do**
  for $i' \in \mathcal{C} \setminus \{i\}$ **do**
    if $r \notin \mathcal{R}_{i'}^{free}$ **then**
      Compute $dist_{cells}(i') = ED\,(i, i')$;
    end
  end
  Compute $dist_{RBG}(r) = \min(dist_{cells}(i'))$;
end
Compute $r' = \arg\max(dist_{RBG}(r))$;
return: r'

10.6.4 RBG Donation

To donate one RBG r' from the RAN slice m'' to the RAN slice m', the RAN Slice Planner executes the steps described in Algorithm 4. First the RAN Slice Planner initializes to infinity the vector $dist_{cells}$. Focusing on one RBG of those allocated for RAN slice m'', i.e., $r \in \mathcal{R}_i^{m''}$, the RAN Slice Planner checks if this RBG has been allocated in the neighbor cells. At this point, the behavior of the proposed algorithm is equal to Algorithm 3. The reason is this algorithm aims to donate the RBG r' which impacts with less strength in the interference terms suffered by the neighbor cells. In this way, if the RAN slice m' would induce a higher cell load in the RBG r' in comparison with the current cell load induced by the RAN slice m'', see Eq. (10.4), the interference term induced in RBG r' would be minimum.

10.7 Numerical Results and Discussions

In this section, we evaluate the performance of the proposed RAN Slice Planner and we compare it with two reference solutions. The reference solution 1 computes the minimum radio resource quota for a RAN slice m in a cell i as $R_{i,m}^{slot} = N_{RB,i}^{slot}/|\mathcal{M}|$, i.e., the radio resources are equally distributed between the RAN slices. The reference solution 2 computes the minimum radio resource quota

as $R_{i,m}^{slot} = \left\lfloor \left(\rho_{i,m} / \sum_{m' \in \mathcal{M}} \rho_{i,m}\right) N_{RB,i}^{slot} \right\rfloor$, i.e., the radio resources are distributed in proportion to the average offered traffic intensity of each RAN slice in such cell. For both reference solutions, we assume the amount of radio resources determined by these quotas are randomly allocated in the resource grid. In addition to the performance analysis, we also evaluate the adaptation and renegotiation capabilities of the proposed solution.

10.7.1 Experimental Setup

We consider a RAN infrastructure which comprises a set of $|\mathcal{C}| = 20$ cells deployed over an urban area of 1.5 Km x 1.5 Km. We also assume the traffic demand for each RAN slice is non-uniformly distributed over the considered area. To characterize the channel conditions of an UE served by a specific RAN slice, we use a snapshot-based model [33]. Each snapshot represents a random realization of the demand distribution for each RAN slice (i.e. varying the positions of the UEs). The different realizations of the same traffic probability distribution ensure reliable statistical significance analysis. Finally, Table 10.2 summarizes the parameters used for the simulations.

10.7.2 Performance Analysis of the Proposed RAN Slice Planner

The first experiment provides the performance analysis of the proposed RAN Slice Planner, as well as a comparison with the reference solutions 1 and 2. In this experiment, we assume the RAN Slice Planner must plan the deployment of three RAN slices with specific average offered traffic intensities, i.e., ρ_m. Specifically, we consider $\rho_1 = \rho_3 = \rho_0$, whereas ρ_2 could take different values from $0.25\rho_0$ to $4\rho_0$.

In Fig. 10.4, we depict the evolution of the average UE blocking probability $\overline{B}_m$ for each RAN slice when the RAN Slice Planner executes the multiple ordinal potential games. In this specific realization, we assume $\rho_2 = 2\rho_0$. When the game $M' = 1$ starts, only the RAN slice $m = 1$ participates. In this game, the RAN Slice Planner iteratively adds radio resources in each cell for this RAN slice until a NE solution is reached. In the game $M' = 2$, the RAN Slice Planner first adds the remaining free RBGs (if available after finishing the previous game) and then it donates RBGs from the RAN slice $m = 1$ to the RAN slice $m = 2$. The aim

Table 10.2: Simulation parameters

Parameter	Configuration
Cellular Environment	Urban, 1.5 Km x 1.5 Km
Number of Cells $\|\mathcal{C}\|$	20
Carrier frequency	2.14 GHz (i.e., within band n1 [48])
5G Numerology μ	0
Number of available RBs in a cell $N_{RB,i}^{slot}$ (same for all the cells)	106 RBs
RBG size R_{size}	4 RBs
Propagation (path loss, shadowing)	Umi model [58]
Cell antenna directivity	Omni-directional
Cell antenna height	6 m
UE antenna height	1.5 m
UE thermal noise	-174 dBm/Hz
UE noise figure	9 dB
UE minimum SINR γ_{min}	-10 dB [52]
UE maximum SINR γ_{max}	30 dB [52]
Attenuation factor σ	0.6
Maximum spectral efficiency SE_{max}	7.4063 bps/Hz
UE Downlink (DL) data rate per RAN slice th_m (same for all the RAN slices)	0.8 Mbps
Upper bound for the UE blocking probability B^{th}	1%
UE blocking probability model for each pair RAN slice - cell	See [54]
Number of RAN slices	From 3 to 6
Reference average offered traffic intensity ρ_0	20
Average offered traffic intensity per RAN slice ρ_m	From 0.25 to 4

of these procedures is to minimize the average UE blocking probability of the RAN slice which present the highest value for this parameter. We observe how the average UE blocking probabilities of these RAN slices tend to be equal in the first iterations of this game (i.e., iteration 850 aprox.). Then, the RAN Slice Planner slightly reduces these average UE blocking probabilities until reaching the NE solution (i.e., iteration 950 aprox.). Finally, the periodical peaks in the

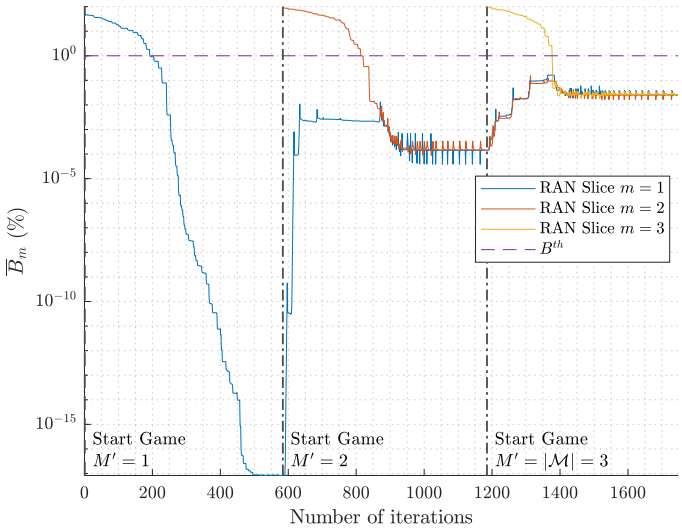
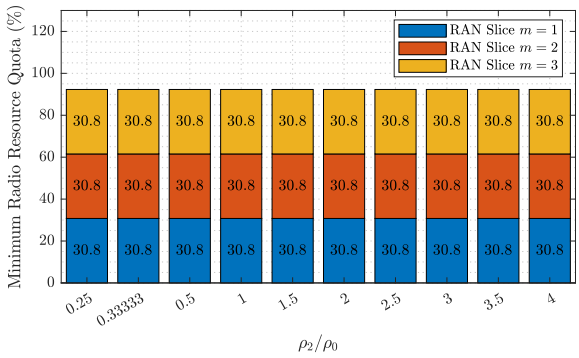


Figure 10.4: Evolution of the average UE blocking probabilities when the RAN Slice Planner executes the multiple ordinal potential games for $\mathcal{M} = 3$ RAN slices.

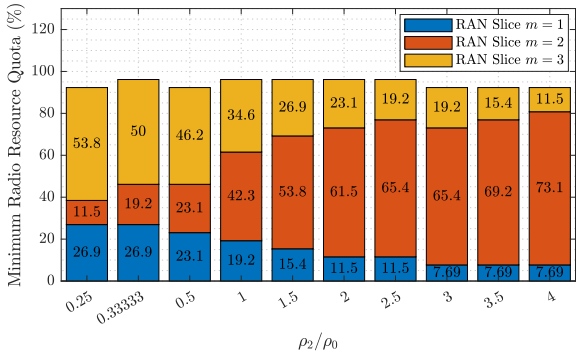
last iterations correspond to the situation where the unilateral reduction of the average UE blocking probability in one RAN slice involves an increment in the average UE blocking probability of the other RAN slice, which is above the value for the potential function in the previous iteration (i.e., turns where the cells cannot improve their utility functions). In the last game, i.e., $M' = 3$, the RAN Slice Planner acts in the same way as the previous game with the difference that RAN slices $m = 1$ and $m = 2$ donates radio resources to the new RAN slice.

Once the behavior of the proposed RAN Slice Planner is known, we compare its performance with the reference solutions 1 and 2. In Fig. 10.5, we show the minimum radio resource quota computed for each RAN slice in a specific cell when the RAN slice $m = 2$ has a different value for the average offered traffic intensity. Furthermore, Fig. 10.6 depicts the average UE blocking probability for the three RAN slices.

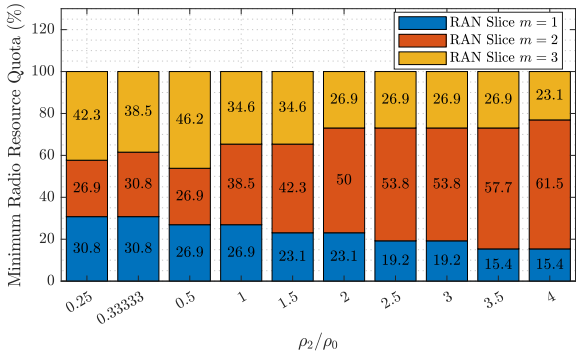
Focusing on the reference solution 1, the MNO under(over)-provision radio resources for the three RAN slices due to it always allocates them the same amount of radio resources regardless their traffic demands (see Fig. 10.5(a)). If we observe Fig. 10.6, we notice RAN slice $m = 2$ has a much lower average



(a) Reference solution 1



(b) Reference solution 2



(c) Proposed RAN Slice Planner

Figure 10.5: Minimum radio resource quotas computed in a specific cell. Note that $\rho_1 = \rho_3 = \rho_0$

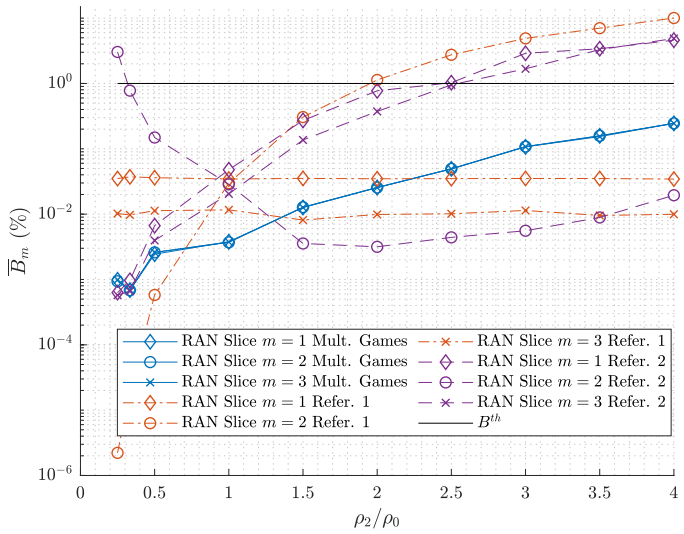


Figure 10.6: Evaluation of the average UE blocking probability $\overline{B}_m$ per RAN slice when $\rho_1 = \rho_3 = \rho_0$ and ρ_2 takes different values

UE blocking probability than the remaining RAN slices when ρ_2/ρ_0 is low. The opposite happens for higher values of ρ_2/ρ_0. This involves that (a) RAN slices $m = 1$ and $m = 3$ have higher values for the average UE blocking probability than they should have; and (b) the RAN slice $m = 2$ has not enough radio resources to achieve an average UE blocking probability below the imposed upper bound.

As Fig. 10.5(b) shows, the reference solution 2 determines the amount of radio resources allocated for each RAN slice in proportion to their average offered traffic intensities. This approach does not guarantee the average UE blocking probabilities are below the upper bound. This is mainly due to the inter-cell interference levels are not considered neither to compute the amount of required radio resources nor to allocate these resources in specific RBGs. We notice in Fig. 10.6 how the average UE blocking probability for RAN slice $m = 2$ is above the imposed upper bound for lower values of ρ_2/ρ_0. We also observe a similar behavior for RAN slice $m = 1$ and $m = 3$ when ρ_2/ρ_0 is higher. This is due to the reference solution 2 under-provisions radio resources for RAN slice $m = 2$ when ρ_2/ρ_0 is low, and over-provisions it for higher values of ρ_2/ρ_0.

In the case of using the proposed RAN Slice Planner, we observe our solution

outperforms the reference solutions 1 and 2. Specifically, we notice the average UE blocking probabilities of all the RAN slice are practically the same and are always below the upper bound. The reason is our solution does not only consider the average UE blocking probability for these RAN slices to allocate them radio resources, but also the inter-cell interference conditions in the RAN infrastructure.

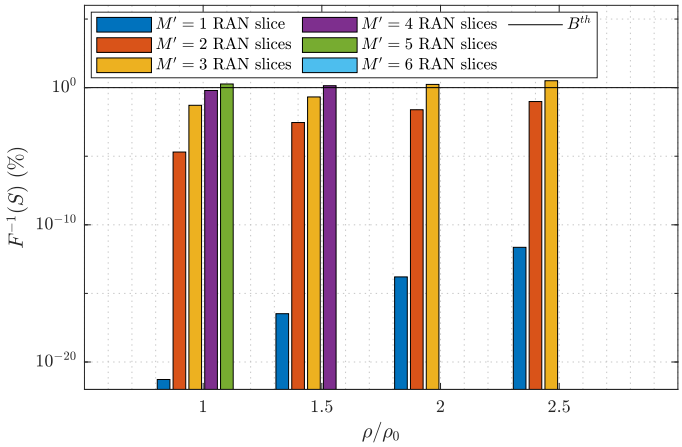
10.7.3 Analysis of Adaptability

In the second experiment, we aim to demonstrate how the proposed RAN Slice Planner can be adapted to accommodate a higher amount of RAN slices in comparison with the reference solutions 1 and 2. Specifically, we evaluate several scenarios where the offered traffic intensity for each RAN slice is the same, i.e., $\rho_m = \rho$. In Fig. 10.7, we observe the inverse of the potential function $F^{-1}(S)$ (i.e., the highest average UE blocking probability of the considered RAN slices) when the RAN infrastructure. We notice how the proposed RAN Slice Planer outperforms the reference solutions by allowing more RAN slices can be deployed with an average UE blocking probability below the upper bound.

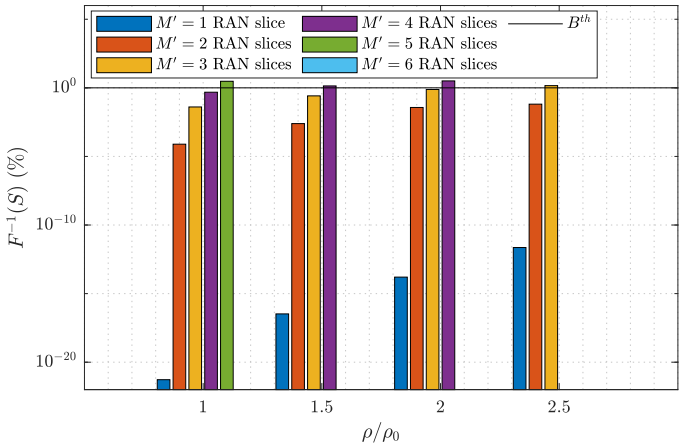
10.7.4 Analysis of the Renegotiation Capability

In the last experiment, we evaluate the renegotiation capability provided by the proposed RAN Slice Planner. To that end, we assume an scenario where (a) three RAN slices are currently running in the RAN and (b) the MNO receives the deployment requests of three new RAN slices. For simplicity, we consider all the RAN slices offer the same average traffic intensity and it is equal to the reference average traffic intensity $\rho_m = \rho_0$.

Under this scenario, the proposed RAN Slice Planner executes a planning procedure to determine if all the RAN slices can be accommodated into the RAN. This procedure results in the blue bars depicted in Fig.10.8(a). We notice the average UE blocking probability for all the RAN slices are above the imposed upper bound, thus the MNO cannot accommodate the six RAN slices. To solve that, the MNO should (a) renegotiate the SLA with the tenants which request the new RAN slices or (b) add more resources in the RAN infrastructure. In this experiment, we consider the MNO renegotiates the SLA with the tenants. Specifically,



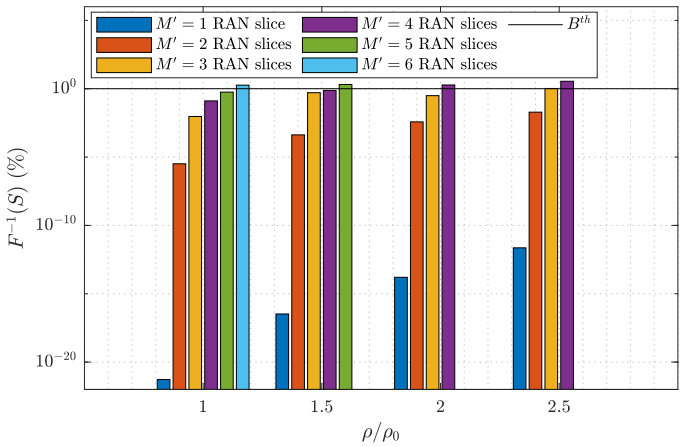
(a) Reference solution 1



(b) Reference solution 2

Figure 10.7: Maximum number of RAN slices which the RAN Slice Planner can accommodate into the RAN infrastructure

we assume the MNO negotiates a reduction in the number of subscribers for each RAN slice. This means the average offered traffic intensity for each RAN slice is reduced. In the first renegotiation, the MNO reduces the available subscribers by a 20 %. This means $\rho_4 = \rho_5 = \rho_6 = 0.8\rho_0$. After the RAN Slice Planner re-executes the planning procedure, the average UE blocking probability for each RAN slice (i.e., orange bars) is still above the upper bound despite the effective



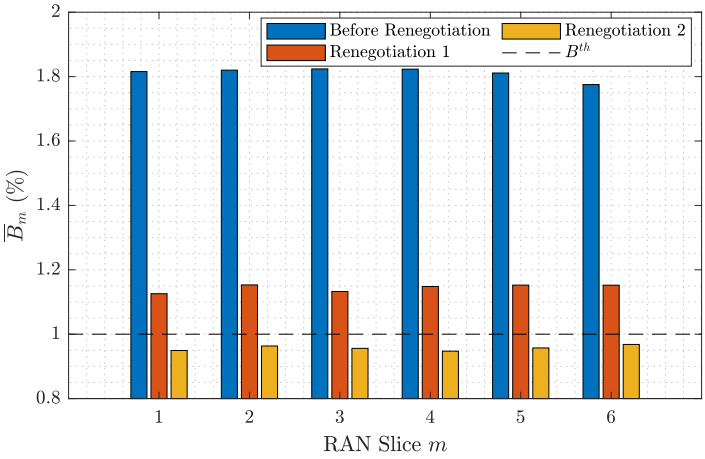
(c) Proposed RAN Slice Planner

Figure 10.7: Maximum number of RAN slices which the RAN Slice Planner can accommodate into the RAN infrastructure

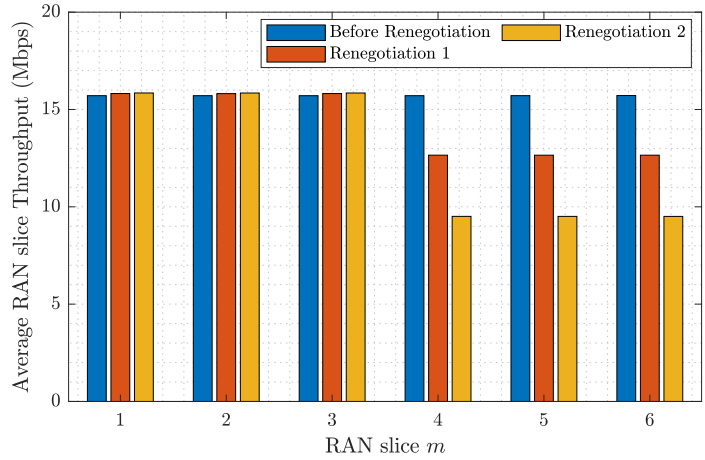
average throughput for the requested RAN slices is reduced as Fig. 10.8(b) shows. If this happens, the MNO tries to reduce more the amount of subscribers for the requested RAN slices. In the case of reducing the number of subscribers by a 40 % (i.e., $\rho_4 = \rho_5 = \rho_6 = 0.6\rho_0$) in these RAN slices, we observe (i.e., yellow bars) how the average UE blocking probability of each RAN slice is below the imposed upper bound.

10.8 Conclusions

During the operation phase of each RAN slice, the amount of assigned radio resources in each cell must be in compliance with the minimum and maximum radio resource quotas. Established during the RAN slice planning, these bounds aim to ensure the SLAs for all the deployed RAN slices are met throughout the planning window. Assuming the MNO periodically performs in advance a planning procedure for several RAN slices, we propose a mathematical framework to define (or recompute) the minimum radio resource quotas which guarantee the UE blocking probability for each requested (or already deployed) RAN slice in each cell is below an upper bound. We formulate the problem using game



(a) Average UE blocking probability for each RAN slice



(b) Effective Average Throughput for each RAN slice

Figure 10.8: RAN slice planning: (i) Before renegotiating the SLA, (ii) Renegotiation 1; and (iii) Renegotiation 2.

theory. Specifically, we use multiple ordinal potential games and demonstrate the existence of a NE solution. To solve the formulated games, we design novel strategies based on better response dynamics with the goal of minimizing the average UE blocking probability for all the RAN slices. The simulation results demonstrate the effectiveness of the proposed solution in terms of performance,

adaptability and renegotiation capability.

Acknowledgments

This work is partially supported by the H2020 research and innovation project 5G-CLARITY (Grant No. 871428); the Spanish Ministry of Economy and Competitiveness, the European Regional Development Fund (Project PID2019-108713RB-C53); and the Spanish Ministry of Education, Culture and Sport (FPU Grant 17/01844)

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 Commun. Mag.*, vol. 55, no. 5, pp. 80–87, 2017.

[2] T. Guo and A. Suárez, "Enabling 5G RAN Slicing With EDF Slice Scheduling," *IEEE Trans. Veh. Technol*, vol. 68, no. 3, pp. 2865–2877, 2019.

[3] 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, 2019.

[4] J. Zheng, P. Caballero, G. de Veciana, S. J. Baek, and A. Banchs, "Statistical Multiplexing and Traffic Shaping Games for Network Slicing," *IEEE ACM Trans. Netw.*, vol. 26, no. 6, pp. 2528–2541, 2018.

[5] H. D. R. Albonda and J. Pérez-Romero, "An Efficient RAN Slicing Strategy for a Heterogeneous Network With eMBB and V2X Services," *IEEE Access*, vol. 7, pp. 44771–44782, 2019.

[6] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Pérez, "Multi-Tenant Radio Access Network Slicing: Statistical Multiplexing of Spatial Loads," *IEEE ACM Trans. Netw.*, vol. 25, no. 5, pp. 3044–3058, 2017.

[7] B. Han, V. Sciancalepore, X. Costa-Pérez, D. Feng, and H. D. Schotten, "Multiservice-Based Network Slicing Orchestration With Impatient Tenants," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 5010–5024, 2020.

[8] 3GPP TS 28541 V.17.0.0, "Management and orchestration; 5G Network Resource Model (nrm); Stage 2 and stage 3 (Release 17)," Sept. 2020.

[9] J. Ordonez-Lucena, P. Ameigeiras, L. M. Contreras, J. Folgueira, and D. R. López, "On the Rollout of Network Slicing in Carrier Networks: A Technology Radar," *Sensors*, vol. 21, no. 23, 2021.

[10] I. Vilà, J. Pérez-Romero, O. Sallent, and A. Umbert, "A Novel Approach for Dynamic Capacity Sharing in Multi-tenant Scenarios," in *IEEE PIMRC (Virtual Conference)*, pp. 1–6, 2020.

[11] J. Ordonez-Lucena *et al.*, "The Creation Phase in Network Slicing: From a Service Order to an Operative Network Slice," in *EuCNC, Ljubljana, Slovenia*, pp. 1–36, 2018.

[12] M. J. Osborne and A. Rubinstein, "A course in game theory," *Cambridge, MA: MIT Press [Google Scholar]*, 1994.

[13] B. Han, D. Feng, and H. D. Schotten, "A Markov Model of Slice Admission Control," *IEEE Netw. Lett.*, vol. 1, no. 1, pp. 2–5, 2019.

[14] M. Vincenzi, E. Lopez-Aguilera, and E. Garcia-Villegas, "Maximizing Infrastructure Providers' Revenue Through Network Slicing in 5G," *IEEE Access*, vol. 7, pp. 128283–128297, 2019.

[15] V. Sciancalepore, X. Costa-Perez, and A. Banchs, "RL-NSB: Reinforcement Learning-Based 5G Network Slice Broker," *IEEE ACM Trans. Netw.*, vol. 27, no. 4, pp. 1543–1557, 2019.

[16] M. R. Raza, C. Natalino, P. Öhlen, L. Wosinska, and P. Monti, "Reinforcement Learning for Slicing in a 5G Flexible RAN," *J. Light. Technol.*, vol. 37, no. 20, pp. 5161–5169, 2019.

[17] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, and X. Costa-Pérez, "A Machine Learning Approach to 5G Infrastructure Market Optimization," *IEEE Trans. Mobile Comput*, vol. 19, no. 3, pp. 498–512, 2020.

[18] B. Han, J. Lianghai, and H. D. Schotten, "Slice as an Evolutionary Service: Genetic Optimization for Inter-Slice Resource Management in 5G Networks," *IEEE Access*, vol. 6, pp. 33137–33147, 2018.

[19] J. Pérez-Romero and O. Sallent, "Optimization of Multitenant Radio Admission Control Through a Semi-Markov Decision Process," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 862–875, 2020.

[20] I. Vilà, O. Sallent, A. Umbert, and J. Pérez-Romero, "An Analytical Model for Multi-Tenant Radio Access Networks Supporting Guaranteed Bit Rate Services," *IEEE Access*, vol. 7, pp. 57651–57662, 2019.

[21] Y. L. Lee, J. Loo, T. C. Chuah, and L. Wang, "Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks," *IEEE Trans. Wireless Commun*, vol. 17, no. 4, pp. 2146–2161, 2018.

[22] 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 Trans. Wireless Commun*, vol. 17, no. 10, pp. 6419–6432, 2018.

[23] A. Aijaz, "Hap – SliceR: A Radio Resource Slicing Framework for 5G Networks With Haptic Communications," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2285–2296, 2018.

[24] R. Li *et al.*, "Deep Reinforcement Learning for Resource Management in Network Slicing," *IEEE Access*, vol. 6, pp. 74429–74441, 2018.

[25] Y. Abiko, T. Saito, D. Ikeda, K. Ohta, T. Mizuno, and H. Mineno, "Flexible Resource Block Allocation to Multiple Slices for Radio Access Network Slicing Using Deep Reinforcement Learning," *IEEE Access*, vol. 8, pp. 68183–68198, 2020.

[26] H. Zhang and V. W. S. Wong, "A Two-Timescale Approach for Network Slicing in C-RAN," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6656–6669, 2020.

[27] I. Vilà, J. Pérez-Romero, O. Sallent, and A. Umbert, "Characterization of Radio Access Network Slicing Scenarios With 5G QoS Provisioning," *IEEE Access*, vol. 8, pp. 51414–51430, 2020.

[28] J. Prados-Garzon, A. Laghrissi, M. Bagaa, T. Taleb, and J. M. Lopez-Soler, "A Complete LTE Mathematical Framework for the Network Slice Planning of the EPC," *IEEE Trans. Mob. Comput.*, vol. 19, no. 1, pp. 1–14, 2020.

[29] B. Xiang, J. Elias, F. Martignon, and E. Di Nitto, "Joint Planning of Network Slicing and Mobile Edge Computing in 5G Networks," *arXiv preprint arXiv:2005.07301*, 2020.

[30] P. Muñoz, O. Sallent, and J. Pérez-Romero, "Self-Dimensioning and Planning of Small Cell Capacity in Multitenant 5G Networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4552–4564, 2018.

[31] F. Bahlke, O. D. Ramos-Cantor, S. Henneberger, and M. Pesavento, "Optimized Cell Planning for Network Slicing in Heterogeneous Wireless Communication Networks," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1676–1679, 2018.

[32] O. Al-Khatib, W. Hardjawana, and B. Vucetic, "Spectrum Sharing in Multi-Tenant 5G Cellular Networks: Modeling and Planning," *IEEE Access*, vol. 7, pp. 1602–1616, 2019.

[33] P. Muñoz *et al.*, "Radio Access Network Slicing Strategies at Spectrum Planning Level in 5G and Beyond," *IEEE Access*, pp. 1–1, 2020.

[34] C. Chang and N. Nikaein, "Ran Runtime Slicing System for Flexible and Dynamic Service Execution Environment," *IEEE Access*, vol. 6, pp. 34018–34042, 2018.

[35] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN Resource Slicing Mechanism for Multiplexing of eMBB and URLLC Services in OFDMA Based 5G Wireless Networks," *IEEE Access*, vol. 8, pp. 45674–45688, 2020.

[36] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Pérez, "Network Slicing Games: Enabling Customization in Multi-Tenant Mobile Networks," *IEEE ACM Trans. Netw.*, vol. 27, no. 2, pp. 662–675, 2019.

[37] D. H. Kim, S. M. A. Kazmi, A. Ndikumana, A. Manzoor, W. Saad, and C. S. Hong, "Distributed Radio Slice Allocation in Wireless Network Virtualization: Matching Theory Meets Auctions," *IEEE Access*, vol. 8, pp. 73494–73507, 2020.

[38] S. Zazo, S. Valcarcel Macua, M. Sánchez-Fernández, and J. Zazo, "Dynamic Potential Games With Constraints: Fundamentals and Applications in Communications," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3806–3821, 2016.

[39] D. Della Penda, A. Abrardo, M. Moretti, and M. Johansson, "Distributed Channel Allocation for D2D-Enabled 5G Networks Using Potential Games," *IEEE Access*, vol. 7, pp. 11195–11208, 2019.

[40] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Commun. Surveys Tuts*, vol. 20, no. 3, pp. 2429–2453, 2018.

[41] R. Ferrús, O. Sallent, J. Pérez-Romero, and R. Agusti, "On the Automation of RAN Slicing Provisioning and Cell Planning in NG-RAN," in *EuCNC, Ljubljana, Slovenia*, pp. 37–42, 2018.

[42] O. Adamuz-Hinojosa, P. Munoz, J. Ordonez-Lucena, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "Harmonizing 3GPP and NFV Description Models: Providing Customized RAN Slices in 5G Networks," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 64–75, 2019.

[43] 3GPP TS 28533 V.16.5.0, "Management and orchestration; Architecture framework (Release 16)," Mar. 2020.

[44] GSM Association, "Generic Network Slice Template (Version 3.0)," May 2020.

[45] 3GPP TS 28531 V.16.7.0, "Management and Orchestration; Provisioning; (Release 16)," Sept. 2020.

[46] P. Muñoz, O. Sallent, and J. Pérez-Romero, "Self-Dimensioning and Planning of Small Cell Capacity in Multitenant 5G Networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4552–4564, 2018.

[47] 3GPP TS 38.306 V.16.2.0, "NR; User Equipment (UE) radio access capabilities (Release 16)," Oct. 2019.

[48] 3GPP TS 38.101-1 V.16.3.0, "User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (Release 16)," Mar. 2020.

[49] 3GPP TS 38.101-2 V.16.4.0, "User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone (Release 16)," June 2020.

[50] 3GPP TS 38.211 V.16.2.0, "NR, Physical chnanels and modulation (Release 16)," June 2020.

[51] 3GPP TS 38.214 V.16.1.0, "NR, Physical layer procedures for data (Release 16)," Mar. 2020.

[52] 3GPP TR 38.803 V.14.2.0, "Study on new radio access technology: Radio Frequency (RF) and co-existence aspects (Release 14)," Sept. 2017.

[53] V. B. Iversen, "Teletraffic engineering and network planning," 2015.

[54] O. Adamuz-Hinojosa, P. Ameigeiras, P. Munoz, and J. Lopez-Soler, "Analytical Model for the UE Blocking Probability in an OFDMA Cell providing GBR Slices," in *IEEE WCNC, Nanjing, China*, 2020.

[55] Y. H. Chew, B.-H. Soong, *et al.*, "Potential Game Theory: Applications in Radio Resource Allocation," 2016.

[56] D. Bauso, *Game theory with engineering applications*. SIAM, 2016.

[57] M. Feldman, Y. Snappir, and T. Tamir, "The efficiency of best-response dynamics," in *SAGT 2017, L'Aquila, Italy*, pp. 186–198, Springer, 2017.

[58] 3GPP TS 38.901 V.16.1.0, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 16)," Dec. 2019.

Chapter 11

Paper H. A Delay-driven RAN Slicing Orchestrator to support B5G uRLLC Services

Authors:

Oscar Adamuz-Hinojosa, Vincenzo Sciancalepore, Pablo Ameigeiras, Juan M. Lopez-Soler, Xavier Costa-Perez.

The paper has been submitted to IEEE Transactions on Wireless Communications.

Abstract

The well-established network slicing concept has been successfully applied to the Radio Access Network (RAN) involving a number of technical challenges. In particular, the Mobile Network Operator (MNO) must ensure corresponding slice requirements throughout slice lifetimes. When dealing with ultra-Reliable Low Latency Communication (uRLLC) RAN slices, the MNO must guarantee the packet transmission delay for each slice within a delay budget with a certain probability. In this paper, we propose a Stochastic Network Calculus (SNC)-based model, which given *i*) the amount of dedicated radio resources for a uRLLC RAN slice, *ii*) the probability the packet transmission delay violates a bound and *iii*) the distribution of its traffic demand, provides the delay bound for such conditions.

Based on this model, we finally propose a RAN slice orchestrator to plan in advance the deployment of multiple uRLLC RAN slices in a multi-cell environment. Specifically, this orchestrator computes the amount of radio resources to be assigned to each slice while fulfilling their performance requirements in the long term. We validate the proposed SNC-based model, demonstrating it provides an upper and conservative estimation of the bound for the packet transmission delay of a slice, given a target violation probability: this shows significant benefits in a scenario with resource scarcity.

11.1 Introduction

The Fifth Generation (5G) of mobile networks brings digitalization and exploitation of the industrial vertical segment. It is supposed to boost a wide variety of unprecedented communication services with stringent requirements in terms of performance and functionalities. There is a general consensus identifying three main categories: enhanced Mobile Broadband (eMBB), massive Machine Type Communication (mMTC) and ultra-Reliable Low Latency Communication (uRLLC) [1]. The latter is of particular interest in industrial scenarios and comprises communication services with stringent requirements in terms of latency and reliability [2]. uRLLC services can perform monitoring and control of cyber-physical systems, deliver industrial Augmented Reality (AR)/Virtual Reality (VR) appli-

cations, or continuous maintenance operations on plant through massive wireless sensor networks.

However, considering each communication service as an independent monolithic network instance and building a dedicated Radio Access Network (RAN) infrastructure to accommodate such stringent requirements would be costly and unaffordable. Therefore, RAN slicing has emerged as a potential solution [3] to economically provide the means to manage and successfully orchestrate such services onto a shared network infrastructure. It consists of providing logically-separated self-contained RANs, denominated RAN slices, tailored onto the requirements of a specific communication service over a common physical infrastructure.

Differently from core network slicing, where virtual machines or containers might be instantiated to accommodate chains of virtual network functions, the RAN slicing concept comes at no negligible costs: *how to assign radio resources in advance to multiple RAN slices in such a way the Mobile Network Operator (MNO) can ensure their performance requirements will be completely satisfied throughout their lifetimes?* In particular, RAN slices providing uRLLC services can further exacerbate the resource assignment problem as the MNO must guarantee that the packet transmission delay is within the overall delay budget with a certain probability. In this regard, the Third Generation Partnership Project (3GPP) has recently standardized a set of policy ratios [3, 4] that must be considered by the packet scheduler while dynamically allocating radio resources for each RAN slice throughout their lifetimes. In this paper, we focus on the Radio Resource Management (RRM) Policy Dedicated Ratio. It defines the dedicated radio resource quota for an instantiated RAN slice, i.e., its dedicated radio resources, in a single cell. Such radio resources cannot be shared even though the corresponding RAN slice will not fully use those throughout its lifetime.

When the MNO executes the planning procedure for multiple uRLLC RAN slices in advance, it computes a dedicated radio resource quota per RAN slice and cell. To this end, the MNO must rely on *i*) a mathematical model, which automatically checks whether dedicated radio resources for a RAN slice in a cell can meet its performance requirements, and *ii*) an algorithm, based on the suggested model, which plans the dedicated radio resource quotas for each RAN slice.

11.1.1 Related Works

In the literature, there exists queueing theory-based models to derive the transmission delay of an uRLLC packet in the radio interface. Examples can be found in [5, 6, 7, 8, 9]. Despite their valuable contributions, the proposed models present some drawbacks, which do not make them suitable for planning uRLLC RAN slices. One drawback is that some models assume well-known but not enough-accurate statistical distributions for modeling the packet arrival rate and the packet transmission rate in the radio interface. For instance, works such as [7, 8, 9] consider an exponential distribution for the packet transmission rate of a uRLLC service. This assumption along with the consideration of a Poisson distribution for the packet arrival rate has the advantage that some models provide close-form expressions for the Cumulative Distribution Function (CDF) of the packet transmission delay. However, if more complex distributions are considered [10], computing the CDF using queueing theory is mathematically intractable. In such cases, queueing theory-based models can only provide average values for the packet transmission delay.

Another mathematical tool to derive the transmission delay is Stochastic Network Calculus (SNC). This tool allows us to compute non-asymptotic statistical performance bounds of the type $\mathrm{P}[\text{delay} > x] \leq \epsilon$ considering complex stochastic processes [11]. This generality comes at the expense of exacts solutions for such bounds. Instead, SNC provides a conservative estimation for such bounds. In the literature, there exists a wide variety of solutions based on SNC to compute the transmission delay of an uRLLC packet. Since this tool is ideal for scenarios where there exists multiple network nodes in tandem, most of the solutions focus on computing the upper bound of the packet delay in multi-hop cellular networks. Examples can be found in [12, 13, 14]. Specifically, they do not deepen into the impact of the channel effects (e.g., fast fading) and/or the main uRLLC traffic characteristics in the packet transmission delay. Conversely, other works focus on the radio interface considering specific access schemes. For instance [15] considers an Orthogonal Frequency-Division Multiple Access (OFDMA) scheme and [16, 17] assume non-orthogonal multiple access schemes. However, these works do not consider the RAN slicing technology. This means that the impact of this technology in the cell capacity is completely omitted. There exist SNC-

based solutions, which consider network slicing such as [18], however they focus on the transport and core networks, simplifying the behavior of the radio interface. To the best of our knowledge, there are no works that analyze the packet delay bound in the radio interface for an uRLLC RAN slice.

Focusing on solutions for provisioning uRLLC services, we can find works such as [7, 8, 9, 19]. In these works the authors consider latency and reliability as the main service performance requirements, however these solutions omit the RAN slicing technology. Others works such as [6, 20, 21, 22, 23, 24] consider RAN slicing to offer uRLLC services. However, they are mainly focused on the operation of each RAN slice instead of its planning. In these works, the authors provide an algorithm to dynamically allocate radio resources for each RAN slice in every transmission time interval or within a short time windows. However, how the latency requirements could be guaranteed in the long term for each RAN slice is out of their scope. Additionally, some of these works do not consider any model for estimating the packet transmission delay. Instead, these works propose online algorithms, which directly observe the packet buffers to decide the amount of radio resources allocated for each uRLLC slice. Therefore, the solutions proposed in these works omit the radio resource quotas, which must previously be derived during a planning phase.

11.1.2 Contributions

In this article, we assume a multi-cellular environment where the MNO must plan the deployment of multiple uRLLC RAN slices. Furthermore, each RAN slice has specific requirements in terms of latency and reliability. Specifically, each RAN slice requires its packet transmission delay is below a target bound with a certain probability. To summarize:

- we have proposed a SNC-based model which provides the packet delay bound of an uRLLC RAN slice in a single cell. To that end, this model considers as inputs: *i*) the amount of dedicated radio resources for this RAN slice, *ii*) the probability the packet delay is above the delay bound, *iii*) the CDF for the Signal-to-Interference-plus-Noise Ratio (SINR) experienced by the users served by this RAN slice, and *iv*) the traffic demand of this RAN slice, i.e., the distribution of the packet arrival rate, and the distribution

of the packet size. In our model, the packet size distribution could be arbitrary;

- to compute the CDF for the SINR perceived by the user, which is served by a specific uRLLC RAN slice, we use a novel model based on stochastic geometry. This model considers the impact of the interference incurred by multiple RAN slices deployed in neighbor cells on the capacity the serving cell offers to the serving RAN slice;

- we have proposed a RAN slice orchestrator that—using the proposed SNC-based model—plans in advance the deployment of multiple RAN slices with different requirements in terms of traffic demand and latency. It aims to derive the dedicated radio resource quotas for each RAN slice which satisfy their requirements throughout its lifetime. Specifically, the proposed RAN slice orchestrator relies on a novel heuristic to derive the radio resource quotas, which minimize the difference between the delay bounds achieved with such quotas and the target delay bounds.

In the provided results, we first validate the proposed SNC-based model by means of an exhaustive simulation campaign, demonstrating that it always provides an upper estimation of the real delay bound of an uRLLC RAN slice. This makes our model suitable for planning uRLLC RAN slices. Then, based on this model, we show the benefits of using the proposed RAN slice orchestrator to plan the deployment of multiple uRLLC RAN slices over a multi-cellular scenario with radio resource scarcity.

The remainder of this article is organized as follows. Section 11.2 provides a background on SNC. Section 11.3 describes the system model. In Section 11.4, we propose a SNC-based model for modeling the packet delay bound of an uRLLC RAN slice in a single cell. In Section 11.5, we formulate the radio resource planning for several RAN slices in a multi-cell environment. To solve this problem, we propose a novel heuristic. In Section 11.6, we validate the proposed SNC-based model and provide the performance results for the proposed heuristic. Finally, Section 11.7 summarizes the conclusions.

11.2 Background on Stochastic Network Calculus (SNC)

11.2.1 Fundamentals

Network calculus has been developed along two tracks: Deterministic Network Calculus (DNC) and SNC [25]. Focusing on DNC, its main principles are [26, 27]:

- For each individual network node, DNC considers: (a) the accumulative arrival process $A(\tau, t)$ of a specific service; (b) the accumulative service process $S(\tau, t)$; and (c) the accumulative departure process $D(\tau, t)$. These stochastic processes are considered in the time interval $(\tau, t]$.
- Characterizing $A(\tau, t)$ and $S(\tau, t)$ by upper and lower bounds. These bounds are known as arrival curve $\alpha(\tau, t)$ and service curve $\beta(\tau, t)$, respectively. $A(\tau, t)$ has an arrival curve if $A(\tau, t) \leq A(\tau, t) \otimes \alpha(\tau, t)\ \forall \tau \in [0, t]$. $S(\tau, t)$ has a service curve if $D(\tau, t) \geq A(\tau, t) \otimes \beta(\tau, t)\ \forall \tau \in [0, t]$. The operator $\otimes$ defines the convolution under the min-plus algebra.
- Using $\alpha(\tau, t)$ and $\beta(\tau, t)$, performance parameters such as the backlog and the delay bounds of each network node can be analyzed. The definition of backlog comprises the number of bits which are stored in the network node's buffer, whereas the delay is the waiting time for each bit in the buffer. The backlog bound B in a network node is defined in Eq. (11.1).

$$B = \max_{\tau \in [0,t]} \{\alpha(\tau, t) - \beta(\tau, t)\} \tag{11.1}$$

  Assuming First-come First-served (FCFS) order, the delay bound W in a network node is given by Eq. (11.2).

$$W = \min\{\omega \geq 0 : \max_{\tau \in [0,t]} \{\alpha(\tau, t) - \beta(\tau, t + \omega)\}\} \tag{11.2}$$

DNC considers the worst-case scenario to compute $\alpha(\tau, t)$ and $\beta(\tau, t)$, i.e., considering the greatest arrival rate and the lowest service rate. This means DNC ignores the effects of statistical multiplexing and therefore, it leads to an overestimation of the resource requirements for the service to be deployed in the

network node. This fact has motivated the development of SNC, which extends the DNC to a probabilistic setting. In SNC, $\alpha(\tau,t)$ and $\beta(\tau,t)$ are commonly known as arrival and service envelopes, respectively. $A(\tau,t)$ has a arrival envelope $\alpha(\tau,t)$ with bounding function $f(x)$, if for any $x \geq 0$, the Eq. (11.3) holds for all $t \geq 0$.

$$P\{A(\tau,t) - A(\tau,t) \otimes \alpha(\tau,t) > x\} \leq f(x) \tag{11.3}$$

$S(\tau,t)$ has a service envelope $\beta(x,t)$ with bounding function $g(x)$, if Eq. (11.4) holds for all $t \geq 0$:

$$P\{A(\tau,t) \otimes \beta(\tau,t) - D(\tau,t) > x\} \leq g(x) \tag{11.4}$$

A widely practice in SNC to compute the backlog and delay bounds in a network node consists of assuming affine functions to define $\alpha(\tau,t)$ and $\beta(\tau,t)$. In sections 11.2.2, 11.2.3 and 11.2.4, we summarize the methods defined in [11] to obtain these envelopes, and compute the bounds for the backlog and the delay.

11.2.2 Affine Arrival Envelope

Some stochastic traffic models use Moment Generating Functions (MGFs) to uniquely determine the distribution of a random process. The MGF of $A(\tau,t)$ is defined as $M_A(\theta) = E[e^{\theta A(\tau,t)}]$ with free parameter $\theta \geq 0$ [28]. Considering an affine arrival envelope model, Eq. (11.5) defines an upper bound for $M_A(\theta)$. The variables $\rho_A > 0$ and $\sigma_A \geq 0$ are the rate and burst parameters.

$$M_A(\theta) \leq e^{\theta[\rho_A(t-\tau)+\sigma_A]} \tag{11.5}$$

Other approach is to use the Exponentially Bounded Burstiness (EBB) model to provide the guarantee expressed in Eq. (11.6). This model is based on Eq. (11.3), and $\alpha(\tau,t)$ is defined in Eq. (11.7), where $\rho_A > 0$ and $b_A \geq 0$ are the rate and burst parameters. The EBB model relaxes the deterministic arrival curve described in section 11.2.1 by defining an overflow profile $\varepsilon_A \geq 0$.

$$P[A(\tau,t) > \alpha(\tau,t)] \leq \varepsilon_A \tag{11.6}$$

$$\alpha(\tau,t) = \rho_A(t-\tau) + b_A \tag{11.7}$$

Comparing the arrival envelope defined in Eq. (11.6) with the analogous deterministic arrival curve described in section 11.2.1, we notice a difference arises with respect to the computation of the backlog bound. The deterministic arrival curve can be immediately applied in Eq. (11.1), however the EBB envelope cannot. The reason is Eq. (11.1) evaluates all $\tau \in [0, t]$, where the $\tau = \tau^*$ that attains the maximum is a random variable. In contrast, the EBB envelope only provides a guarantee for an arbitrary, yet, fixed $\tau \in [0, t]$. To overcome this problem, a sample path argument as Eq. (11.8) shows is required. In [11], the authors demonstrate that $\rho'_A = \rho_A + \delta$, where the free parameter $\delta > 0$ is known as the slack rate.

$$P\left[\exists \tau \in [0,t] : A(\tau, t) > \rho'_A(t-\tau) + b_A\right] \leq \varepsilon'_A \tag{11.8}$$

The EBB and MGF models are directly connected by the Chernoff bound [28] as Eq. (11.9) shows [11]. In this expression, we obtain the overflow profile ε'_A.

$$\varepsilon'_A = \frac{e^{\theta\sigma_A} e^{-\theta b_A}}{1 - e^{-\theta\delta}} \tag{11.9}$$

Finally, we can obtain a mathematical expression for b_A as Eq. (11.10) shows.

$$b_A = \sigma_A - \frac{1}{\theta}\left[\ln(\varepsilon'_A) + \ln\left(1 - e^{-\theta\delta}\right)\right] \tag{11.10}$$

11.2.3 Affine Service Envelope

Considering an affine service envelope model, Eq. (11.11) defines a upper bound for the negative MGF of $S(\tau, t)$, i.e., $M_S(-\theta)$.

$$M_S(-\theta) \leq e^{-\theta(\rho_S(t-\tau)-\sigma_S)} \tag{11.11}$$

Using the analogous EBB model for the service envelope, also known as the Exponentially Bounded Fluctuation (EBF), we provide the guarantee defined in Eq. (11.12). $\beta(\tau, t)$ is defined in Eq. (11.13), where $[x]_+$ denotes $\max\{0, x\}$, $\rho_S(t-\tau) > 0$ and $b_S \geq 0$ the the rate and burst parameters, respectively. This model also includes the concept of sample path as Eq. (11.8). In this case,

$\rho'_S = \rho'_S - \delta$ for all $\delta > 0$.

$$P\left[\exists \tau \in [0,t] : S(\tau,t) < \beta(\tau,t)\right] \leq \varepsilon'_S \tag{11.12}$$

$$\beta(\tau,t) = \rho'_S \left[t - \tau - \frac{b_S}{\rho'_S}\right]_+ \tag{11.13}$$

In a similar way as the affine arrival envelope, the Chernoff bound is used in Eq. (11.12) to derive the deficit profile ε'_S (i.e., equivalent to ε'_A in the affine arrival envelope) as Eq. (11.14) describes. Finally, using this expression the burst parameter b_S is derived as Eq. (11.15) shows.

$$\varepsilon'_S = \frac{e^{\theta\sigma_S} e^{-\theta b_S}}{1 - e^{-\theta\delta}} \tag{11.14}$$

$$b_S = \sigma_S - \frac{1}{\theta}\left[\ln(\varepsilon'_S) + \ln\left(1 - e^{-\theta\delta}\right)\right] \tag{11.15}$$

11.2.4 Backlog and Delay bounds considering Affine Envelopes

Considering $\alpha(\tau,t)$ and $\beta(\tau,t)$, defined in Sections 11.2.2 and 11.2.3, respectively, and Eqs. (11.1) and (11.2), we obtain the backlog and delay bounds in Eqs. (11.16) and (11.17), respectively.

$$B = \frac{\rho_A + \delta}{\rho_S - \delta} b_S + b_A \tag{11.16}$$

$$W = \frac{b_A + b_S}{\rho_S - \delta} \tag{11.17}$$

Fig. 11.1 depicts a graphical representation of the derived backlog and delay bounds. The backlog bound B is the vertical deviation between $\alpha(\tau,t)$ and $\beta(\tau,t)$ whereas the delay bound W is the horizontal deviation between these envelopes.

11.3 System Model

In this work, we focus on the Downlink (DL) operation of a 5G-New Radio (NR) multi-cell environment with several RAN slices. Each RAN slice provides an uRLLC service to their User Equipments (UEs). Furthermore, each next Generation NodeB (gNB) supports Link Adaptation (LA), thus these gNBs consider

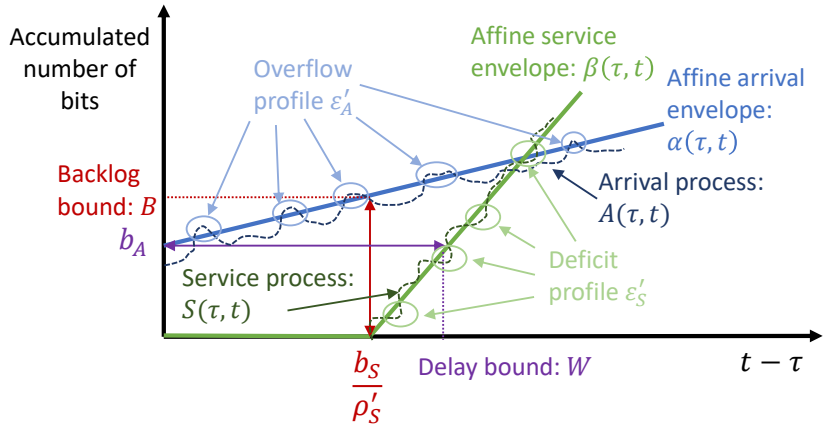


Figure 11.1: Graphical representation of backlog bound B and delay bound W.

the channel quality perceived by each UE to allocate them radio resources. Under this scenario, we first describe the network model. Then, we define the characteristics of the offered DL traffic for an uRLLC service. Next, we present the radio resource model. Finally, we describe the channel model for a single cell.

11.3.1 Network Model

We consider a MNO owns a RAN infrastructure consisting of a set $\mathcal{I}$ of cells. Defining $\mathcal{M}$ as the set of RAN slices, the RAN slice orchestrator will execute a radio resource planning procedure to accommodate them, each with specific uRLLC requirements, in the long term. Under this scenario, a set $\mathcal{U}$ of UEs coexist, being *i)* $\mathcal{U}^m \subseteq \mathcal{U}$ the subset of UEs served by the RAN slice m, *ii)* $\mathcal{U}_i \subseteq \mathcal{U}$ the subset of UEs served by the cell $i \in \mathcal{I}$, and *iii)* $\mathcal{U}_i^m = \mathcal{U}^m \cap \mathcal{U}_i$ the intersection of both subsets. Finally, we consider each UE is attached to the nearest cell.

11.3.2 uRLLC Traffic Model

We assume the traffic demand of each RAN slice is non-uniformly distributed over the considered RAN infrastructure. Focusing on the traffic demand of a single RAN slice $m \in \mathcal{M}$, we consider statistical distributions for its arrival packet rate and the size of their packets.

Regarding the packet arrival, we assume the RAN slice m transmits packets to their UEs in batches (i.e., simultaneous transmission of packets for multiple UEs). Specifically, we consider the RAN slice m has an average of λ_m batch arrivals per unit time following a Poisson distribution. Since a Poisson process can be split into independent processes [29], we can also express the average batch arrival for each gNB as $\lambda_{i,m} = \omega_{i,m}\lambda_m$. The variable $\omega_{i,m}$ denotes the probability an UE $u \in \mathcal{U}^m$ is served by the gNB $i \in \mathcal{I}$.

Focusing on a individual batch arrival, a set of $N_{i,m}^{pkt}$ packets are simultaneously generated. $N_{i,m}^{pkt}$ is a discrete random variable which ranges from 1 to $|\mathcal{U}_i^m|$, and follows an arbitrary distribution with Probability Mass Function (PMF) $p_{i,m,u}$. With respect to the transmitted packets in a RAN slice m, each packet presents a specific size of $L \in \mathcal{L}^m$ bits. L is also a discrete random variable which follows an arbitrary distribution with PMF $p_{m,l}$.

We assume each cell uses a packet scheduler per RAN slice. The scheduling policy used by each scheduler is the Earliest Deadline First (EDF). This discipline is the optimal policy for scheduling delay-sensitive traffic [30]. We also assume all the packets for the RAN slice m have the same deadline W_m^{th}. Since W_m^{th} is the same for all the packets, the EDF discipline is equivalent to the First In First Out (FIFO) policy.

Finally, we consider the MNO also imposes a value ε'_m for the violation probability before signing the Service Level Agreement (SLA) with the tenant which requests the RAN slice m. The violation probability is the probability that an individual uRLLC packet suffers a transmission delay above the packet delay budget W_m^{th}.

11.3.3 Radio Resource Model

We assume OFDMA as the accessing scheme for the cells. Focusing on the cell i, it supports a total bandwidth W_i. In turn, this bandwidth is divided into N_i OFDM subcarriers, which are grouped in groups of $N_{SC} = 12$ subcarriers. Each group defines a Resource Block (RB), which is the smallest unit of resources that can be allocated to a UE. The set of RBs available in the cell i during each timeslot t^{slot} is denoted by $\mathcal{R}_i$, and the amount of these RBs is given by Eq. (11.18). Since a cell supports scalable numerologies ($\mu_{5G} = 0, 1, \ldots, 4$), i)

the duration of a timeslot is computed as $t^{slot} = 10^{-3}/2^{\mu_{5G}}$ seconds, and *ii*) the subcarrier spacing is derived as $\Delta_f = 2^{\mu_{5G}} \cdot 15$ KHz. The parameter OH denotes the overhead factor due to control plane data [31].

$$|\mathcal{R}_i| = \left\lfloor \frac{W_i}{N_{SC}\Delta f}\,(1 - OH) \right\rfloor \tag{11.18}$$

Finally, we denote *i*) $\mathcal{R}_i^m \subseteq \mathcal{R}_i$ as the subset of RBs allocated to the RAN slice m in cell i, and *ii*) $\mathcal{R}_u \subseteq \mathcal{R}_i^m$ as the subset of RBs allocated for the UE u.

11.3.4 Channel Model

To measure the channel quality perceived by an arbitrary UE $u \in \mathcal{U}_i^m$ in the RB r, we consider the instantaneous SINR, i.e., $\gamma_{u,r}^{(t)}$. This parameter is described in Eq. (11.19), where P_{TX} is the transmitted power in a single RB. We assume all the gNBs transmit the same power for each RB. The parameter h denotes the gain due to the fast fading. We assume h follows a Rayleigh distribution with mean one. The parameter χ is the gain due to the shadowing and follows a log-normal distribution characterized by the mean μ_χ and standard deviation σ_χ, both in dB. The parameter $d_{u,i}$ denotes the distance from the gNB i to the UE u, with α the pathloss exponent. The summation terms in the denominator gather the interference suffered by the UE u in the RB r, and P_N is the noise power measured in a single RB. Focusing on a specific interference term j, the parameter $\xi_{j,r}$ denotes a binary variable that takes the value 1 when the neighbor cell j transmits data in the RB r and the value 0 otherwise. The value for $\xi_{j,r}$ will depend on the dynamic radio resource allocation performed by the corresponding gNBs in the neighbor cells. We consider that only the RBs allocated to a specific RAN slice can be scheduled to its UEs.

$$\gamma_{u,r}^{(t)} = \frac{P_{TX} h \chi d_{u,i}^{-\alpha}}{\sum_{j \in \mathcal{I} \setminus \{i\}} \xi_{j,r} P_{TX} h \chi d_{u,j}^{-\alpha} + P_N} \tag{11.19}$$

Considering $\gamma_{u,r}^{(t)}$ is measured for a considerable amount of time for the UEs attached in the cell i and served by the RAN slice m, we can obtain the CDF of the SINR to model the channel quality for this RAN slice in this cell. In the literature, stochastic geometry has been used as a powerful analytical tool for

modeling the CDF of the SINR [32]. In this work, we rely on several state-of-the-art works on stochastic geometry (i.e., [33, 34, 35, 36, 37]) to provide a closed-form expression for the CDF of the SINR. In [33], the authors have defined this CDF as Eq. (11.20) shows. The parameter γ_{th} is the target SINR. The parameter p_{suc} is the probability that an arbitrary UE perceives a SINR $\gamma_{u,r}$ higher than γ_{th}, whereas p_{sel} is the probability the gNB schedules a RB to transmit data for this UE.

$$f_{CDF}(\gamma_{u,r}, \gamma_{th}) = P[\gamma_{u,r} \leq \gamma_{th}] = 1 - p_{suc} \cdot p_{sel} \tag{11.20}$$

From the results of [34], and considering the pathloss exponent is $\alpha = 4$, we get p_{suc} by Eq. (11.21). In this equation $\kappa_{gNB,\chi} = \kappa_{gNB} E\left[\sqrt{\chi}\right]$, where κ_{gNB} is the gNB density and $E\left[\sqrt{\chi}\right]$ is the fractional moment of the log-normal distribution. $E\left[\sqrt{\chi}\right]$ models the shadowing channel power as Eq. (11.22) shows. The authors of [36] use this fractional moment to incorporate the shadowing effect in Eq. (11.20). The parameter $\kappa_{j,\chi} = \kappa_j E\left[\sqrt{\chi}\right]$, where κ_j denotes the density of the neighbor cells interfering in an arbitrary RB. It is defined as $\kappa_j = \kappa_{gNB}\left(1 - p_{off}\right)$, where the parameter p_{off} is the probability that a generic cell does not transmit in a RB. The parameter $\upsilon\left(\gamma_{th}\right)$ is given by Eq. (11.23). Finally, the function $Q(\cdot)$ denotes the Gaussian Q-function.

$$p_{suc} = \sqrt{\frac{\pi P_{TX}}{\gamma_{th} P_N}} \exp\left(\frac{\left(\pi\left[\kappa_{gNB,\chi} + \kappa_{j,\chi}\upsilon\left(\gamma_{th}\right)\right]\right)^2 P_{TX}}{4\gamma_{th} P_N}\right) \cdot Q\left(\frac{\pi\left[\kappa_{gNB,\chi} + \kappa_{j,\chi}\upsilon\left(\gamma_{th}\right)\right]}{\sqrt{\frac{2\gamma_{th} P_N}{P_{TX}}}}\right) \tag{11.21}$$

$$E\left[\sqrt{\chi}\right] = \exp\left(\frac{\ln\left(10\right)\mu_\chi}{20} + \frac{1}{2}\left(\frac{\ln\left(10\right)\sigma_\chi}{20}\right)^2\right) \tag{11.22}$$

$$\upsilon\left(\gamma_{th}\right) = \sqrt{\gamma_{th}}\left[\frac{\pi}{2} - \arctan\left(\frac{1}{\sqrt{\gamma_{th}}}\right)\right] \tag{11.23}$$

To compute p_{sel} and p_{off}, we need to assume a specific model for the cell load (i.e., the fraction of RBs that are being scheduled on average for the attached UEs). In [33], the authors assume each cell has available only one RB whereas the authors of [35] extend the definition of p_{sel} and p_{off} given in [33] by considering

an arbitrary number of RBs in each cell. Assuming the density of UEs κ_{UEs} is much higher than the density of cells (i.e., $\kappa_{UEs} \gg \kappa_{gNBs}$), we have in Eqs. (11.24) and (11.25) the mathematical expressions for p_{sel} and p_{off}. Note that $\Gamma(\cdot)$ denotes the gamma function.

$$p_{sel} = |\mathcal{R}_i| \left(\frac{\kappa_{UEs}}{\kappa_{gNB,\chi}} \right)^{-1} \tag{11.24}$$

$$p_{off} = \frac{4}{63} \frac{3.5^{3.5}}{\Gamma(3.5)} \frac{\Gamma(4.5 + |\mathcal{R}_i|)}{\Gamma(1 + |\mathcal{R}_i|)} \left(\frac{\kappa_{UEs}}{\kappa_{gNB,\chi}} \right)^{-3.5} \tag{11.25}$$

These parameters were formulated under the assumption the UE density is the same in the entire RAN. However, a RAN slice could present a different amount of UEs in each cell. Furthermore, the UE density for each RAN slice could be different. For this reason, we define the UE density for a RAN slice m in cell i as $\kappa_{UE,i,m}$. In addition, the authors of [35] consider the UEs attached to a cell are randomly chosen for transmission in an arbitrary RB. This means a cell can schedule one RB to transmit data for a single UE. In our work, we consider each UE could receive data from multiple RBs in a timeslot. Depending the RAN slice which serves this UE, the length of each packet could take a different value (see Sec. 11.3.2). Furthermore, each transmitted packet could require a specific amount of RBs according to the channel quality. All this involves computing the density of the required amount of RBs for transmitting a packet in a specific RAN slice and cell, i.e., $\kappa_{RBs,i,m}$ instead of $\kappa_{UEs,i,m}$. For simplicity, we consider the average number of required RBs, i.e., $\overline{R}^{pkt}_{i,m}$, to compute $\kappa_{RBs,i,m}$ as Eq. (11.26) shows.

$$\kappa_{RBs,i,m} = \kappa_{UEs,i,m} \overline{R}^{pkt}_{i,m} \tag{11.26}$$

Under the above assumptions, we reformulate p_{sel} and p_{off} as Eqs. (11.27) and (11.28) show, respectively. In Eq. (11.28), the second summation gathers the probability that a neighbor gNB j does not transmit in a RB for each RAN slice. Since the result of the second summation is different for each neighbor gNB j, we sum the weighted result of each neighbor gNB. To consider the impact of each neighbor gNB, we define the weight ι_j according to the pathloss from the

neighbor gNB j to the considered gNB i, i.e., $\iota_j = d_{i,j}^{-\alpha} / \sum_{j \in \mathcal{I} \setminus \{i\}} d_{i,j}^{-\alpha}$.

$$p_{sel} = |\mathcal{R}_i^m| \left(\frac{\kappa_{RBs,i,m}}{\kappa_{gNB,\chi}} \right)^{-1} \tag{11.27}$$

$$p_{off} = \sum_{j \in \mathcal{I} \setminus \{i\}} \iota_j \sum_{m \in \mathcal{M}} \frac{4}{63} \frac{3.5^{3.5}}{\Gamma(3.5)} \frac{\Gamma\left(4.5 + |\mathcal{R}_j^m|\right)}{\Gamma\left(1 + |\mathcal{R}_j^m|\right)} \left(\frac{\kappa_{RBs,j,m}}{\kappa_{gNB,\chi}} \right)^{-3.5} \tag{11.28}$$

Once the CDF for the SINR is known, we can compute the PMF for the spectral efficiency achieved by an UE served by the RAN slice m in cell i. To that end, we first translate $\gamma_{u,r}$ into the equivalent spectral efficiency $SE_{u,r}$. Since uRLLC packets are very short, the achievable spectral efficiency cannot be accurately capture by Shannon's capacity [22]. Instead, the spectral efficiency in uRLLC falls in the finite blocklength channel coding regime, which is derived as Eq. (11.29) shows [38]. In this equation, $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function. The parameter ϵ_{dec} is the decoding error probability, which could range from 10^{-3} to 10^{-7} [39]. Finally, n_{block} denotes the blocklength.

$$SE_{u,r} = f_{\gamma \to SE}(\gamma_{u,r}) = \log_2(1 + \gamma_{u,r}) - \sqrt{\frac{1 - \frac{1}{(1+\gamma_{u,r})^2}}{n_{block}}} Q^{-1}(\epsilon_{dec}) \log_2(e) \tag{11.29}$$

Then, we consider N_z values for the spectral efficiency $SE_{z,i,m}$ which an arbitrary UE can achieve. This finite set of values depends on the Modulation and Coding Scheme (MCS) pair selected by the gNB after this UE reports the Channel Quality Indicator (CQI) [40, Table 5.2.2.1-3]. As Eq. (11.30) shows, we can compute the probabilities $\pi_{z,i,m}$ of reporting certain CQIs, i.e., the PMF of $SE_{z,i,m}$, by using the CDF of the SINR provided in Eq. (11.20).

$$\pi_{z,i,m} = \begin{cases} f_{CDF}(\gamma_{u,r}, f_{SE \to \gamma}(SE_{1,i,m})) & \text{if } z = 1 \\ f_{CDF}(\gamma_{u,r}, f_{SE \to \gamma}(SE_{z+1,i,m})) & \\ -f_{CDF}(\gamma_{u,r}, f_{SE \to \gamma}(SE_{z,i,m})) & \text{if } 1 < z < N_z \\ 1 - f_{CDF}(\gamma_{u,r}, f_{SE \to \gamma}(SE_{N_z,i,m})) & \text{if } z = N_z \end{cases} \tag{11.30}$$

Since $\pi_{z,i,m}$ is independent from the PMFs for the packet length, i.e., $p_{m,l}$ $\forall m \in \mathcal{M}$, we can compute the amount of required RBs for transmitting a packet

as Eq. (11.31) defines. The variable r' denotes a specific combination of a packet size $L \in \mathcal{L}^m$ and a spectral efficiency value $SE_{z,i,m}$. The PMF of the random variable $R^{pkt}_{r',i,m}$, i.e., $p_{r'}$ is the joint PMF of the random variables L and $SE_{z,i,m}$, whose PMFs are $p_{m,l}$ and $\pi_{z,i,m}$, respectively.

$$R^{pkt}_{r',i,m} = \left\lceil \frac{L_{r'}}{t^{slot} N_{SC} \Delta f SE_{r'}} \right\rceil \tag{11.31}$$

Finally, we can compute the average number of required RBs for transmitting a packet as Eq. (11.32) shows. Note that we previously needed $\overline{R}^{pkt}_{i,m}$ to derive $\pi_{z,i,m}$ as Eq. (11.26) defines. For this reason, we need to compute $\pi_{z,i,m}$ with at least two iterations. For example, in the first iteration we could use arbitrary values for $\pi_{z,i,m}$ to estimate $\overline{R}^{pkt}_{i,m}$. Then, in the second iteration, we can recompute a more accurate version of $\pi_{z,i,m}$ and $\overline{R}^{pkt}_{i,m}$.

$$\overline{R}^{pkt}_{i,m} = \sum_{r' \in \mathcal{R}^{req}_m} p_{r'} R^{pkt}_{r',i,m} \tag{11.32}$$

11.4 SNC-based Model for an uRLLC RAN slice in a cell

In this section, we propose a SNC-based model to determine the amount of RBs $|\mathcal{R}^m_i|$, which the RAN slice orchestrator defines as dedicated radio resource quota for a RAN slice m in a cell i. Considering a specific amount of RBs and the violation probability ε'_m, this model provides the delay and backlog bounds which guarantee ε'_m for this RAN slice in this cell. To that end, the proposed model considers different stochastic processes, which characterize: *i*) the DL traffic of an uRLLC RAN slice, and *ii*) the capacity the serving gNB provides for such RAN slice.

Under this context, we first describe the traffic model for an uRLLC RAN slice. Then, we present the service model for the gNB capacity available for such RAN slice. Finally, we use SNC to derive the mathematical expressions for the backlog and delay bounds.

11.4.1 Traffic Model for an uRLLC RAN slice

The arrival process $A_{i,m}(\tau, t)$ is the accumulative number of bits which arrive from the core network to the cell i for the RAN slice m. This stochastic process is defined in Eq. (11.33), where $N_{i,m}^{batch}(t)$ denotes the number of batch arrivals during the interval $[0, t]$. The random variable y_b denotes the amount of bits generated in a batch arrival.

$$A_{i,m}(\tau, t) = \sum_{b=1}^{N_{i,m}^{batch}(t)} y_b \tag{11.33}$$

In a batch arrival, the number of bits which arrives to the corresponding gNB is given by Eq. (11.34). The parameter $N_{i,m,b}^{pkt}$ defines the number of packets which simultaneously arrive for different UEs. For each packet, the parameter $L_n \in \mathcal{L}^m$ represents its size in bits.

$$y_b = \sum_{n=1}^{N_{i,m,b}^{pkt}} L_n \tag{11.34}$$

We compute the MGF of $A_{i,m}(\tau, t)$ as Eq. (11.35) shows. By substitution of $\ln\left(M_{y_b}(\theta)\right) = \nu$, we have the MGF of the Poisson process $N_{i,m}^{batch}(t)$ in function of the free parameter ν. This MGF can be expressed as Eq. (11.36) shows [28].

$$\begin{aligned} M_{A_{i,m}}(\theta) = E[e^{\theta A_{i,m}(\tau,t)}] = E[(M_{y_b}(\theta))^{N_{i,m}^{batch}(t)}] = \\ E[e^{N_{i,m}^{batch}(t)\cdot\ln\left(M_{y_b}(\theta)\right)}] = E\left[e^{\nu N_{i,m}^{batch}(t)}\right] = M_{N_{i,m}^{batch}}(\nu) \end{aligned} \tag{11.35}$$

$$M_{N_{i,m}^{batch}}(\nu) = e^{\lambda_{i,m} t(e^{\nu}-1)} \tag{11.36}$$

If we express the resulting MGF in function of the free parameter θ as Eq. (11.37) shows, we observe which the MGF of $A_{i,m}(\tau, t)$ depends on the MGF of y_b, i.e., $M_{y_b}(\theta)$.

$$M_{A_{i,m}}(\theta) = e^{\lambda_{i,m}\left(M_{y_b}(\theta)-1\right)t} \tag{11.37}$$

We compute the MGF of y_b as Eq. (11.38) shows. The resulting expression

is equal to the MGF of $N_{i,m,b}^{pkt}$ with free parameter ν, i.e., $M_{N_{i,m,b}^{pkt}}(\nu)$.

$$M_{y_b}(\theta) = E[e^{\theta y_b}] = E[(M_{l_n}(\theta))^{N_{i,m,b}^{pkt}}] = E[e^{N_{i,m,b}^{pkt} \cdot \ln(M_{l_n}(\theta))}] = E\left[e^{\nu N_{i,m,b}^{pkt}}\right] = M_{N_{i,m,b}^{pkt}}(\nu) \quad (11.38)$$

Using the definition of MGF (see section 11.2.2), we define the MGF of $N_{i,m,b}^{pkt}$ as Eq. (11.39) shows. If we express the resulting MGF in function of the free parameter θ, we observe which the MGF of y_b depends on the MGF of L_n, i.e., M_{L_n} (see Eq. (11.40)).

$$M_{N_{i,m,b}^{pkt}}(\nu) = \sum_{u=1}^{|\mathcal{U}_i^m|} e^{\nu u} p_{i,m,u} \quad (11.39)$$

$$M_{y_b}(\theta) = \sum_{u=1}^{|\mathcal{U}_i^m|} [M_{L_n}(\theta)]^u \, p_{i,m,u} \quad (11.40)$$

Similarly to Eq. (11.39), we define the MGF of L_n in Eq. (11.41).

$$M_{L_n}(\theta) = \sum_{l \in \mathcal{L}^m} e^{\theta l} p_{m,l} \quad (11.41)$$

If we include this expression in Eq. (11.40), and then we replace the resulting expression in Eq. (11.37), we obtain the MGF of $A_{i,m}(\tau, t)$ as Eq. (11.42) shows.

$$M_{A_{i,m}}(\theta) = e^{\lambda_{i,m}\left(\sum_{u=1}^{|\mathcal{U}_i^m|}\left[\sum_{l \in \mathcal{L}^m} e^{\theta l} p_{m,l}\right]^u p_{i,m,u} - 1\right)t} \quad (11.42)$$

Finally, by equaling the right side of Eq. (11.5) with Eq. (11.42), we obtain Eq. (11.43). In this expression, we define the parameters $\rho_{A_{i,m}}$ and $\sigma_{A_{i,m}}$ of the affine arrival envelope $\alpha_{i,m}(\tau, t) = \left(\rho_{A_{i,m}} + \delta\right)[t - \tau] + \sigma_{A_{i,m}}$ which bounds the arrival process $A_{i,m}(\tau, t)$.

$$\sigma_{A_{i,m}} = 0 \tag{11.43a}$$

$$\rho_{A_{i,m}} = \frac{\lambda_{i,m}}{\theta}\left[\sum_{u=1}^{|\mathcal{U}_i^m|}\left[\sum_{l\in\mathcal{L}^m} e^{\theta l} p_{m,l}\right]^u p_{i,m,u} - 1\right] \tag{11.43b}$$

11.4.2 Service Model for a RAN slice

The service process $S_{i,m}(\tau, t)$ is the accumulative number of bits which could be processed by the cell i for the RAN slice m. This process is described in Eq. (11.44).

$$S_{i,m}(t) = \sum_{n=0}^{N^{slot}(t)} C_{i,m}(n) \tag{11.44}$$

The deterministic variable $N^{slot}(t) = t/t^{slot}$ represents the number of accumulated timeslots in t. The random variable $C_{i,m}(n)$ denotes the amount of bits which the corresponding gNB could process in the timeslot n if all the allocated RBs for the RAN slice m were used. This variable depends on (a) the amount of allocated RBs for this RAN slice, i.e., $|\mathcal{R}_i^m|$; and (b) the amount of RBs consumed by each transmitted uRLLC packet, i.e., $R_{r',i,m}^{pkt}$ and its PMF, i.e., $p_{r'}$. Considering (a)-(b), we can obtain all the possible values for $C_{i,m}(n)$, i.e., $c_q\ \forall q \in \mathcal{Q}_i^m$, and its PMF p_q. Note that the definition of a model to obtain c_q and p_q is out of the scope. In our work, we estimate them by Montecarlo simulations.

Using the definition of MGF (see section 11.2.2), we define the negative MGF for $C_{i,m}(n)$ as Eq. (11.45) shows.

$$M_{C_{i,m}}(-\theta) = E\left[e^{-\theta C_{i,m}(n)}\right] = \sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q \tag{11.45}$$

Considering $M_{C_{i,m}}(-\theta)$, we can compute the negative MGF of the service process $S_{i,m}(\tau, t)$ as Eq. (11.46) shows. By substitution of $\ln\left(M_{C_{i,m}}(-\theta)\right) = -\nu$, we have the negative MGF of the deterministic variable $N^{slot}(t)$ in function of the free parameter $-\nu$.

$$M_{S_{i,m}}(-\theta) = E\left[e^{-\theta S_{i,m}(\tau,t)}\right] = E\left[\left(M_{C_{i,m}}(-\theta)\right)^{N^{slot}(t)}\right] = E\left[e^{N^{slot}(t)\ln\left(M_{C_{i,m}}(-\theta)\right)}\right] = E\left[e^{-\nu N^{slot}(t)}\right] = M_{N^{slot}}(-\nu) \quad (11.46)$$

Furthermore, we can obtain the negative MGF of $S_{i,m}(\tau,t)$ as Eq. (11.47) shows.

$$M_{S_{i,m}}(-\theta) = e^{\frac{\ln\left(\sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q\right)}{t^{slot}} t} \quad (11.47)$$

Finally, by equaling the right side of Eq. (11.11) with Eq. (11.47), we obtain Eq. (11.48). In this expression, we define the parameters $\rho_{S_{i,m}}$ and $\sigma_{S_{i,m}}$ of the affine service envelope $\beta_{i,m}(\tau,t) = \left(\rho_{S_{i,m}} - \delta\right)[t-\tau] + \sigma_{S_{i,m}}$ which bound the service process $S_{i,m}(\tau,t)$.

$$\sigma_{S_{i,m}} = 0 \quad (11.48a)$$

$$\rho_{S_{i,m}} = \frac{-1}{\theta t^{slot}} \ln\left(\sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q\right) \quad (11.48b)$$

11.4.3 Backlog and Delay bound for a RAN Slice

To compute the backlog and delay bounds for the RAN slice m in the cell i, we assume the violation probability ε'_m is equally distributed into the overflow and deficit profiles, i.e., $\varepsilon'_{A_m} = \varepsilon'_{S_m} = \varepsilon'_m/2$. Considering that, we obtain in Eqs. (11.49) and (11.50) the backlog and delay bounds by (a) applying Eqs. (11.43a) and (11.48a) into Eqs. (11.10) and (11.15), respectively; and (b) using them along with Eqs. (11.43b) and (11.48b) into Eqs. (11.16) and (11.17), respectively.

$$B_{i,m} = \frac{1}{\theta}\left[\ln\left(\frac{\varepsilon'_m}{2}\right) + \ln\left(1 - e^{-\theta\delta}\right)\right] \cdot \left[\frac{\lambda_{i,m} t^{slot}\left(\sum_{u=1}^{|\mathcal{U}|_i^m}\left[\sum_{l\in\mathcal{L}^m} e^{\theta l} p_{m,l}\right]^u p_{i,m,u} - 1\right) + \delta\theta t^{slot}}{\ln\left(\sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q\right) + \delta\theta t^{slot}} - 1\right] \quad (11.49)$$

$$W_{i,m} = \frac{2t^{slot}\left[\ln\left(\frac{\varepsilon'_m}{2}\right) + \ln\left(1 - e^{-\theta\delta}\right)\right]}{\ln\left(\sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q\right) + \delta\theta t^{slot}} \tag{11.50}$$

11.5 Radio Resource Planning for RAN slices

Considering the proposed SNC-based model, we design a radio resource planning scheme. The RAN slice orchestrator will execute it to decide the dedicated radio resource quotas assigned for each RAN slice in each cell. First, we present the problem formulation. Then, we provide a heuristic for solving this problem.

11.5.1 Problem Formulation

When the RAN slice orchestrator plans the deployment of several uRLLC RAN slices, it must guarantee that their performance requirements are met in the long term. Specifically, this means the probability that a packet k transmitted for the RAN slice m suffers a delay W_k above the packet delay budget is less than or equal to the violation probability, i.e., $\mathrm{P}\left[W_k > W_m^{th}\right] \leq \varepsilon'_m$. To meet this condition, the RAN slice orchestrator must set a dedicated radio resource quota per RAN slice and cell.

Assuming the RAN slice orchestrator establishes a specific set of RBs for each pair of RAN slice and cell, i.e., $|\mathcal{R}_i^m|$, each RAN slice m will present in each cell i an upper bound for the packet latency $W_{i,m}$ in such a way that $\mathrm{P}\left[W_k > W_{i,m}\right] = \varepsilon'_m$. With the aim of characterizing how this upper bound is close to the packet delay budget W_m^{th}, we define the parameter $\Delta W_{i,m}$ in Eq. (11.51). Note that $\Delta W_{i,m}$ is zero if the upper bound $W_{i,m}$ is less than the packet delay budget.

$$\Delta W_{i,m} = \max\left(W_{i,m} - W_m^{th},\, 0\right) \tag{11.51}$$

If we consider $\Delta W_{i,m}$ in all the cells where the RAN slice m must provide the uRLLC service, we can compute the average for this parameter as Eq. (11.52) shows.

$$\overline{\Delta W}_m = \sum_{i\in\mathcal{I}^m} \omega_{i,m} \Delta W_{i,m} \tag{11.52}$$

The RAN slice orchestrator aims to compute the dedicated radio resource quotas in such a way that $\overline{\Delta W}_m$ was zero for all the RAN slices. If this pa-

rameter was not zero for one or more RAN slices using all the available RBs, this would mean that the entire RAN infrastructure has not enough capacity to satisfy the latency requirements of these RAN slices. In this scenario, the MNO must prioritize which RAN slices will achieve a delay bound closer to their required packet delay bounds. To that end, we define the RAN slice priority ψ_m as a potential parameter which the MNO may tune for instance, considering an economic-based policy. With the purpose of minimizing $\psi_{m'}\overline{\Delta W}_{m'}$ for the RAN slice m' which presents the greatest value for such parameter, we formulate our problem as follows.

$$\underset{|\mathcal{R}_i^m|}{\text{minimize}} \quad \max\left(\psi_1\overline{\Delta W}_1, \ldots \psi_m\overline{\Delta W}_m, \ldots \psi_{|\mathcal{M}|}\overline{\Delta W}_{|\mathcal{M}|}\right), \tag{11.53}$$

$$\text{subject to:} \quad \sum_{m\in\mathcal{M}} \psi_m = 1, \tag{11.54}$$

$$\sum_{m\in\mathcal{M}} |\mathcal{R}_i^m| \leq |\mathcal{R}_i| \tag{11.55}$$

11.5.2 Heuristic Algorithm Design

To solve the formulated problem, we propose the heuristic described in Algorithm 1. Given the performance requirements and the traffic demand of each RAN slice $m \in \mathcal{M}$ (line 1), Algorithm 1 provides the steps performed by the RAN slice orchestrator to determine the dedicated radio resource quotas for these RAN slices.

First, the algorithm equally distributes the available RBs in each cell among all the RAN slices (line 2). Furthermore the parameter $N^{ite}_{not\_imp}$ is set to zero. This parameter indicates the number of consecutive iterations which are not valid (see while loop). Based on the initial RB allocation, the algorithm derives (line 3): the PMF of $SE_{z,i,m}$ (i.e., the probabilities of reporting a certain CQI); the possible amount of required RBs to transmit a packet; and the PMF for this random variable. These parameters are used by the algorithm as inputs for the proposed SNC-based model (line 4). Using this model, the algorithm estimates the delay bound $W_{i,m}$ for each RAN slice and each cell. Additionally, during the SNC-based model execution, the algorithm (a) estimates the gNB capacity for a RAN slice (i.e., computing p_q and c_q $\forall q \in \mathcal{Q}_i^m$); and (b) optimizes the free parameters θ and δ to obtain $W_{i,m}$. Next, the algorithm computes the difference

Algorithm 1: Radio Resource Planning for $|\mathcal{M}|$ RAN slices

1 **Inputs:** Performance requirements (i.e., W_m^{th} and ε'_m) and traffic distribution (i.e., $|\mathcal{U}_i^m|$, $p_{i,m,u}$, λ_m, $L \in \mathcal{L}^m$, and $p_{m,l}$) for each RAN slice $m \in \mathcal{M}$;
2 **Initialization:** Equal distribution of the RBs among the RAN slices. Set $N_{not\_imp}^{ite} = 0$;
3 Compute $\pi_{z,i,m}$, $R_{r',i,m}^{pkt}$, and $p_{r'}$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$;
4 Compute $W_{i,m}$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$ by using the proposed SNC-based model;
5 $\Delta W_{i,m}$ and $\overline{\Delta W}_m$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$. Evaluate $curr\_func =$ Eq. (11.53);
6 **while** $curr\_func > 0$ *and* $N_{not\_imp}^{ite} < |\mathcal{I}|$ **do**
7 Select $m' = \arg\max \left(\psi_1 \overline{\Delta W}_1, \ldots \psi_m \overline{\Delta W}_m, \ldots \psi_{|\mathcal{M}|} \overline{\Delta W}_{|\mathcal{M}|}\right)$ $\forall m \in \mathcal{M}$;
8 Set $a = \text{sort}_{desc} \left(\omega_{1,m'} \Delta W_{1,m'}, \ldots, \omega_{i,m'} \Delta W_{i,m'}, \ldots, \omega_{|\mathcal{I}|,m'} \Delta W_{|\mathcal{I}|,m'}\right)$ $\forall i \in \mathcal{I}$. Remove the first $N_{not\_imp}^{ite}$ elements of a. Select $i' = \arg(a(1))$;
9 Select $m'' = \arg\min \left(\omega_{i',1} \Delta W_{i',1}, \ldots, \omega_{i',m} \Delta W_{i',m}, \ldots, \omega_{i',m} \Delta W_{i',m}\right)$ $\forall m \in \mathcal{M} \setminus \{m'\}$;
10 Redistribute one RB from RAN slice m'' to RAN slice m', i.e., $|\mathcal{R}_{i'}^{m'}| = |\mathcal{R}_{i'}^{m'}| + 1$ and $|\mathcal{R}_{i''}^{m''}| = |\mathcal{R}_{i''}^{m''}| - 1$;
11 Compute $\pi_{z,i,m}$, $p_{r'}$, and $R_{r',i,m}^{pkt}$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$;
12 Set $prev\_func = curr\_func$;
13 Compute $W_{i,m}$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$ by using the proposed SNC-based model;
14 $\Delta W_{i,m}$ and $\overline{\Delta W}_m$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$. Evaluate $curr\_func =$ `SNC-BASED OPT`;
15 **if** $curr\_func < prev\_func$ **then**
16 Set $N_{not\_imp}^{ite} = 0$. Variables computed in lines 10-14 are valid;
17 **else**
18 Set $N_{not\_imp}^{ite} = N_{not\_imp}^{ite} + 1$. Variables computed in lines 10-14 are invalid;
19 **end**
20 **end**
21 **return:** $\mathcal{R}_i^m$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$

between the estimated delay bounds and the packet delay budget for each RAN slice (line 5). Furthermore, the algorithm averages these differences for each RAN slice. With these parameters, the algorithm evaluates the function defined in Eq.

(11.53).

In the following steps (lines 6-20), the algorithm iteratively redistributes the RBs allocated for each RAN slice in each cell with the aim of minimizing the function defined in Eq. (11.53). Focusing on a single iteration, the algorithm redistributes one RB between two RAN slices in a single cell. To that end, the algorithm first determines the RAN slice m' which will receive a new RB (line 7). It is the one which maximize Eq. (11.53). Then, the algorithm decides the cell where one RB will be redistributed (line 8). This cell will be the one where the weighted difference between the estimated delay bound and the target packet delay bound is greater. Note that if one or more consecutive iterations cannot improve the current value of Eq. (11.53), the cells selected in the previous iterations cannot be considered in the new iteration. This means the $N^{ite}_{not\_imp}$ cells which provide the greatest values for the weighted difference are not considered. Later, the algorithm decides the RAN slice m'' which will donates one RB to the RAN slice m' (line 9). This RAN slice will be the one which has the lowest weighted difference between the estimated delay bound and the packet delay budget. When the cell and the RAN slices involved in the RB redistribution are determined, the algorithm computes the new amount of RBs in such RAN slices (line 10). After that, the algorithm recomputes (lines 11-14): the PMFs for the spectral efficiency and the required amount of RBs per transmitted packet; the delay bound using the SNC-based model; and the value of the function defined in Eq. (11.53). Then, the algorithm checks if the value of this function has decreased with respect to its value in the previous iteration (lines 15-19). If yes, the algorithm considers the variables computed in the lines 10-14 are valid. If not, these variables are invalid, and the algorithm does not update them in this iteration. The algorithm stops when the value of Eq. (11.53) is either equal to zero or can no longer be minimized.

11.6 Numerical Results and Discussions

In this section, we (a) validate the proposed SNC-based model and (b) evaluate the performance of the proposed heuristic, comparing it with two reference solutions. The reference solution #1 consists of the RAN slice orchestrator establishes the dedicated radio resource quotas by allocating proportionally the RBs

according to the traffic demand of each RAN slice. Focusing on a specific cell, this traffic demand depends on the number of attached UEs, the average arrival of packets and the size of each packet. This means the reference solution #1 is agnostic to the latency requirements of each RAN slice. In the reference solution #2, the RAN slice orchestrator periodically recomputes the dedicated radio resource quotas for each RAN slice. Specifically, it redistributes the RBs available for each RAN slice in each cell according to its average buffer size (i.e, packets in queue waiting to be transmitted). This means the RAN slice orchestrator provides more RBs for the RAN slice which presents the greatest average buffer size. This reference solution indirectly considers the latency experienced by the packets of a RAN slice.

11.6.1 Experimental Setup

To validate the SNC-based model, we use a Matlab-based simulator that resembles the packet arrival and their transmission for an uRLLC RAN slice in a single cell. Furthermore, we consider an arbitrary set of values for the probabilities of reporting certain CQIs, i.e., $\pi_{z,i,m}$. Table 11.1 summarizes the main parameters used in the model validation.

Regarding the evaluation of the proposed heuristic, we consider a RAN infrastructure composed of $|\mathcal{I}| = 7$ cells deployed over an area of 0.95 Km x 0.95 Km. We also consider the traffic demand for each RAN slice is non-uniformly distributed over this area. This means each RAN slice (a) serves a different amount of UEs in each cell; (b) has a specific batch arrival rate for its packets; and (c) has a specific distribution for the packet size. In addition, each RAN slice accommodates an uRLLC service with specific performance requirements in terms of packet delay budget and violation probability. In Table 11.2, we summarize the parameters used for evaluating the proposed heuristic.

11.6.2 Validation of the proposed SNC-based model

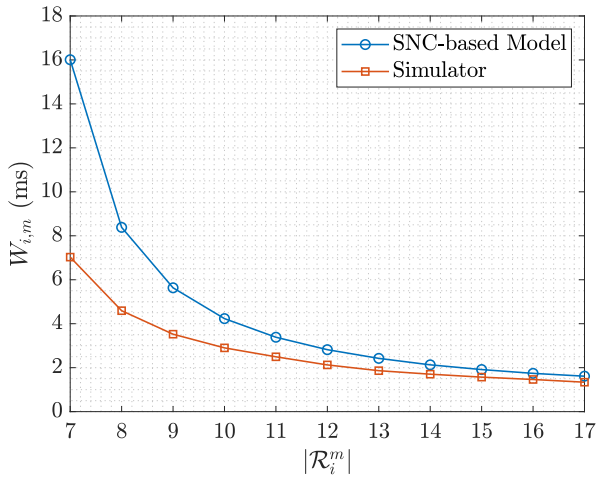
In Fig. 11.2, we have evaluated the delay bound $W_{i,m}$ in function of the dedicated radio resource quota $|\mathcal{R}_i^m|$ assigned for the RAN slice m. We observe the SNC-based model always overestimates the amount of required RBs to obtain a specific delay bound, given a specific value for the violation probability ε'_m. This makes

Table 11.1: Simulation Parameters for SNC-based model validation

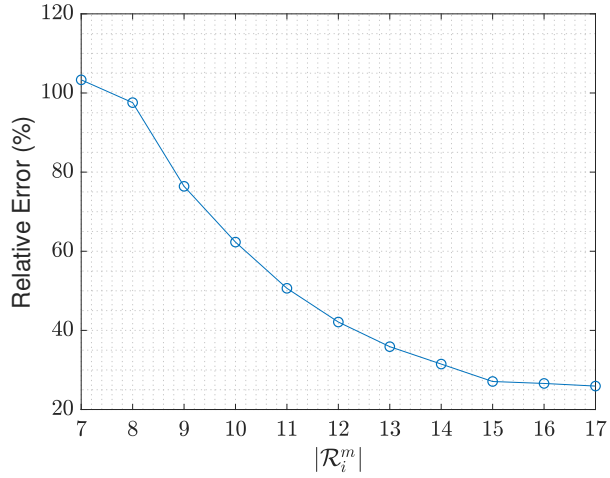
Parameter	**Configuration**	**Parameter**	**Configuration**
5G Numerology μ_{5G}	2	Average batch arrival rate $\lambda_{i,m}$	From 3000 to 5000 batches/s Default: 3400 batches/s
Carrier bandwidth W_i ($\lvert\mathcal{R}_i\rvert$)	20 MHz (24 RBs)	Number of UEs in the cell $\lvert\mathcal{U}_i^m\rvert$	5
RBs for the RAN slice $\lvert\mathcal{R}_i^m\rvert$	From 7 to 17 RBs Default: 10 RBs	Probability of simultaneous transmission of packets $p_{i,m,u}$	Equiprobable
Packet delay budget W_m^{th}	5 ms	Packet size $L \in \mathcal{L}^m$	[256 512 1024 2048] bits
Violation probabilty ε_m'	From 0.001% to 10 % Default: 0.1 %	Packet size distribution $p_{m,l}$	Equiprobable

Table 11.2: Simulation Parameters for RAN slice planning

Parameter	Configuration	Parameter	Configuration
Cellular Environment	0.95 Km x 0.95 Km	Decoding Error Probability ϵ_{dec}	10^{-5}
Number of Cells $\lvert\mathcal{C}\rvert$	7	Blocklength n_{block}	168
Cell areas $\forall i \in \mathcal{C}$	[0.1415; 0.1413; 0.1081; 0.1419 0.1255; 0.1399; 0.1060] Km^2	Number of RAN slices	3
gNB density κ_{gNBs}	$7.756 \cdot 10^{-6}$ gNBs/Km^2	Number of UEs per RAN slice and cell $\lvert\mathcal{U}_i^m\rvert$	$\lvert\mathcal{U}_i^1\rvert$ = [8; 6; 7; 10; 5; 4; 7] $\lvert\mathcal{U}_i^2\rvert$ = [10; 8; 7; 5; 8; 7; 9] $\lvert\mathcal{U}_i^3\rvert$ = [9; 10; 12; 10; 10; 13; 11]
5G Numerology μ_{5G}	2	Packet Delay Budget W_m^{th}	[15; 25; 5] ms
Carrier Bandwidth W_i ($\lvert\mathcal{R}_i\rvert$)	40 MHz (51 RBs)	Violation Probability ϵ_m'	[0.1; 1; 0.05] %
Cell Transmited Power $P_{TX}^{cell} = P_{TX} \cdot \lvert\mathcal{R}_i\rvert$	30 dBm	RAN Slice Priority ψ_m	[0.667; 0.444; 0.889]
Shadowing Parameters	$\mu_\chi = 0$ dB $\sigma_\chi = 4$ dB	Packet length $l \in \mathcal{L}^m$	[256; 512; 1024; 2048] $\in \mathcal{L}^1$ bits [512] $\in \mathcal{L}^2$ bits [512; 2048] $\in \mathcal{L}^3$ bits
UE noise figure	10 dB	PMF packet length p_l $\forall l \in \mathcal{L}^m$	[0.25 0.25 0.25 0.25] $\forall l \in \mathcal{L}^1$ [1] $\forall l \in \mathcal{L}^2$ [0.75 0.25] $\forall l \in \mathcal{L}^3$
UE thermal noise	-174 dBm/Hz	Batch arrival rate λ_m	λ_1 = 24500 batches/s λ_2 = 26950 batches/s λ_3 = 22750 batches/s
Pathloss exponent α	4	Probability of simultaneous transmission of packets p_u	Equiprobable



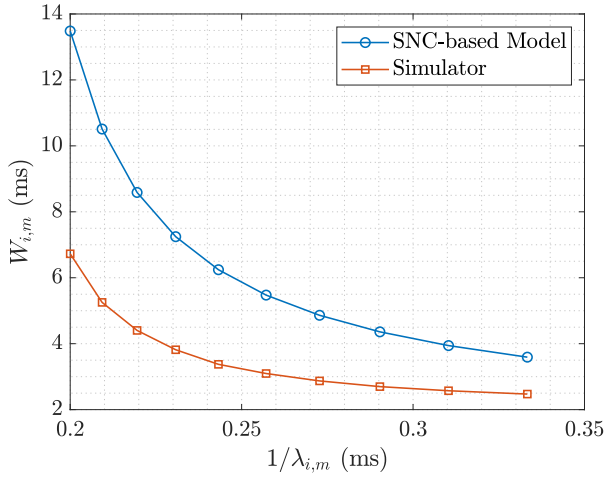
(a) Delay bound $W_{i,m}$



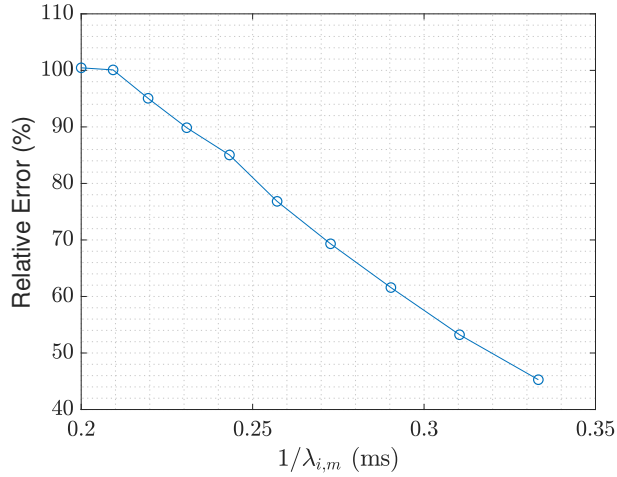
(b) Relative Error (%)

Figure 11.2: Evaluation of the delay bound $W_{i,m}$ in function of the number of RBs allocated to the RAN slice m, i.e., $|\mathcal{R}_i^m|$.

the proposed model suitable for ensuring the performance requirements of uRLLC RAN slices when the MNO plan them in advance. We also notice the relative error between the SNC-based model and the simulator decreases when the number of assigned RBs increases. Using these graphics, the RAN slice orchestrator can estimate the dedicated radio resource quota for a RAN slice in a cell which



(a) Delay bound



(b) Relative Error (%)

Figure 11.3: Evaluation of the delay bound $W_{i,m}$ in function of average time between batch generations, i.e., $1/\lambda_{i,m}$.

guarantee the $(1-\varepsilon'_m)\cdot 100$ % of the packets meet the packet delay budget W_m^{th}. Although they are not depicted due to space limits, we have obtained similar results for the backlog bound $B_{i,m}$ in function of the dedicated radio resource quota $|\mathcal{R}_i^m|$ assigned for the RAN slice m.

We have also evaluated the delay bound $W_{i,m}$ in function of the average time

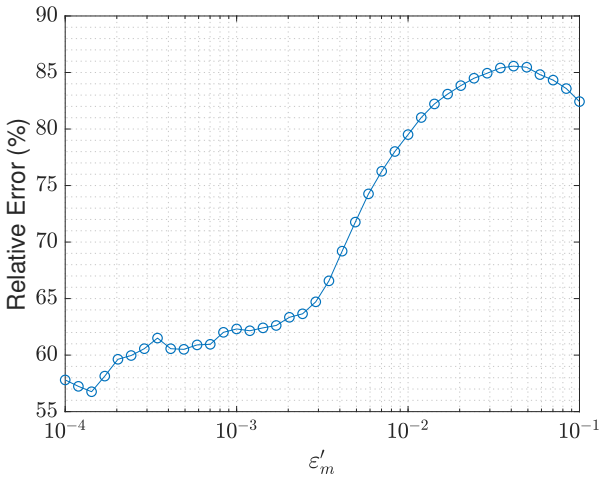


Figure 11.4: Evaluation of the relative error in function of the violation probability ε'_m.

between batch generations (i.e., $1/\lambda_{i,m}$) as Fig. 11.3 shows. Given a specific value for $1/\lambda_{i,m}$, we observe the proposed SNC-based model overestimates the packet delay budget. We also notice the relative error between the proposed model and the simulator decreases when the average time between batch generations increases.

Finally, we have evaluated the relative error in function of the violation probability as Fig. 11.4 shows. We observe the SNC-based model presents a better accuracy when the violation probability takes low values. This means our proposed model is ideal for uRLLC services, which have extreme requirements in terms of reliability.

11.6.3 Performance Analysis of the proposed heuristic

We have analyzed the performance of the proposed heuristic in terms of scalability and we have compared the obtained delay bounds with the ones obtained when the RAN slice orchestrator uses the reference solution #1 or #2.

With respect to the scalability analysis, we study the convergence of the proposed heuristic when the number of cells varies from 6 to 10. In Fig. 11.5, we depict the time required by the proposed heuristic (i.e., x-axis) to reach the solution for Eq. (11.53). If we focus on a single curve, each dot represents the

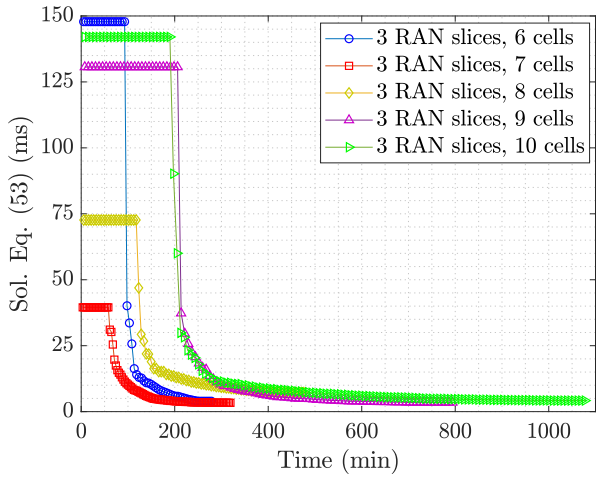
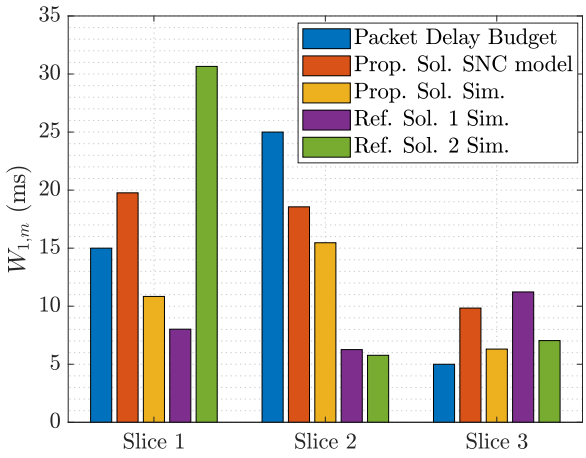


Figure 11.5: Evolution of the solution for Eq. (11.53) along the iterations performed by the proposed heuristic.

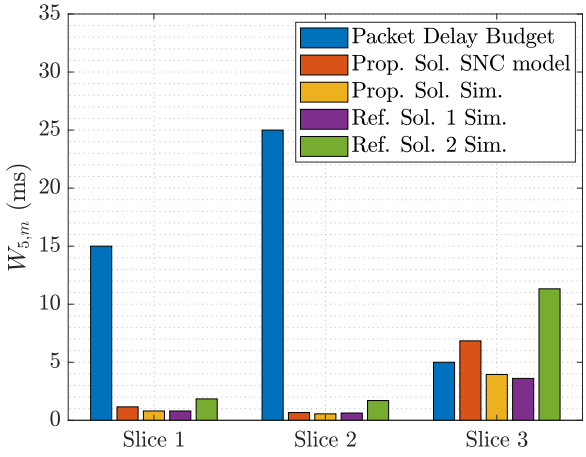
value of such equation in a specific iteration (line 14 in Algorithm 1). We observe the curve is flat in the first iterations. The meaning of the flat region is the parameter $\overline{\Delta W}_m$ is infinite for one or more RAN slices. This means these RAN slices cannot accommodate the DL traffic with the amount of allocated RBs in this iteration (line 10 in Algorithm 1), i.e., with this RB allocation, the amount of packets in the buffer of one or more cells will dynamically increase up to infinity. For simplicity, we have represented this phenomena as a flat region in the curve. When all the RAN slices have allocated enough resources to accommodate their DL traffic, the curve exponentially decreases up to the heuristic converges to a sub-optimal solution for the formulated problem.

If we compare the heuristic performance for the different scenarios, we observe the time between two consecutive iterations increases when the number of considered cells increases. The reason is the proposed heuristic has to call more times, one per cell, the SNC-based model to compute the delay bound for each RAN slice (line 13 in Algorithm 1). This makes the heuristics needs more time for reaching the sub-optimal solution in scenarios with more cells.

When the RAN slice orchestrator computes the dedicated radio resource quota for each RAN slice $m \in \mathcal{M}$ in each cell $i \in \mathcal{I}$, the obtained delay bounds (i.e., $W_{i,m}$) will depend on the adopted RB allocation strategy. In Fig. 11.6, we



(a) Cell $i = 1$. For the proposed solution, $|\mathcal{R}_1^m| = (19, 15, 17)$ RBs. For the the reference solution #1, $|\mathcal{R}_1^m| = (20, 16, 15)$ RBs. Dynamic allocation in the reference solution #2.



(b) Cell $i = 5$. For the proposed solution, $|\mathcal{R}_5^m| = (11, 17, 23)$ RBs. For the the reference solution #1, $|\mathcal{R}_5^m| = (11, 16, 24)$ RBs. Dynamic allocation in the reference solution #2.

Figure 11.6: Delay bounds obtained for each RAN slice in two arbitrary cells.

compare the delay bounds when the RAN slice orchestrator executes the proposed heuristic and the reference solutions #1 and #2. The blue bars represent the

different packet delay budgets W_m^{th} $\forall m \in \mathcal{M}$. Although it is not showed in Fig. 11.6, the violation probability associated to each packet delay budget is also different for each RAN slice as Table 11.2 defines. The orange bars represent the delay bounds considering our proposed solution. In this case, the delay bounds are computed using the proposed SNC-based model. In Fig. 11.6(a), the delay bounds $\Delta W_{1.m} > 0$ $\forall m \in \mathcal{M}$, which means there are not enough RBs in cell $i = 1$ for all the RAN slices. Using the proposed solution, the RAN slice orchestrator computes the RB allocation which provides the delay bounds closer to the packet delay budgets (i.e., blue bars). In Fig. 11.6(b), the computed delay bounds are far below the corresponding packet delay budgets for RAN slices $m = 1$ and $m = 2$. Notwithstanding, the delay bound for RAN slice $m = 3$ is slightly above the target packet delay bound. In this cell, we could think that adding some RBs from RAN slices $m = 1$ and $m = 2$ to the RAN slice $m = 3$ would involve: (a) a delay bound below the target packet delay bound for RAN slice $m = 3$ and (b) the delay bounds for the remaining RAN slice increase but keeping below the corresponding packet delay budgets. However, this RB redistribution is not possible because it would increase the value of the function defined in Eq. (11.53). The reason is this redistribution would reduce the cell load for RAN slices $m = 1$ and $m = 2$ in cell $i = 5$, which would involve the interference increment to the neighbors cells, i.e, $j \in \mathcal{I} \setminus \{i\}$, and thus, the decrement of the channel quality perceived by the UEs attached to the neighbor cells.

Focusing on the yellow bars, they also represent the delay bounds considering our proposed solution. Unlike orange bars, the delay bounds defined by these bars have been computed by simulation. Since the proposed SNC-based model provides a conservative value for the delay bound (as we demonstrate in Section 11.6.2), all the yellow bars are below the orange bars. Even we can observe in Fig. 11.6(a) the delay bounds for RAN slices $m = 1$ and $m = 2$ are below the corresponding packet delay budgets by using the proposed heuristic.

The purple bars represent the obtained delay bounds when the RAN slice orchestrator uses the reference solution #1. In Fig. 11.6(a), we observe the RAN slice $m = 3$ presents a delay bound greater than the provided by our proposed solution. The reason is the traffic demands in RAN slices $m = 1$ and $m = 2$ are greater than the traffic demand in RAN slice $m = 3$, thus the RAN slice orchestrator assigns more RBs for RAN slices $m = 1$ and $m = 2$. However,

the reference solution #1 omits the RAN slice $m = 3$ has the most stringent requirements in terms of latency (i.e., the lowest packet delay budget and violation probability). The same arguments explain the similar behavior observed for the delay bounds in Fig. 11.6(b).

Finally, the green bars correspond to the derived delay bounds when the RAN slice orchestrator uses the reference solution #2. In Fig. 11.6(a), we observe the delay bounds for RAN slices $m = 1$ and $m = 3$ are greater than the delay bounds obtained by our proposed solution. Furthermore in Fig. 11.6(b), the delay bounds for all the RAN slices are also greater. Despite the reference solution #2 involves the RAN slice orchestrator performs a dynamic computation of the dedicated radio resource quotas (i.e., in periods of N_{all} timeslots), our proposed solution outperforms this reference solution. The reason is the RAN slice orchestrator only considers the buffer status in the last N_{all} timeslots and omits the traffic dynamics in the next N_{all} timeslots when it uses the reference solution #2. For instance, in the last N_{all} timeslots one RAN slice could present the greatest average buffer size, thus the RAN slice orchestrator would allocate it the greatest amount of RBs. Then, in the following N_{all} timeslots, the traffic demand for this RAN slice could be considerable lower in comparison with the last period. This would involve this RAN slice would be wasting RBs whereas the remaining RAN slices could require them because of their traffic demands would have increased.

11.7 Conclusions

Deploying and running RAN slices providing uRLLC services would require an ad-hoc analysis of expected performance in terms of delay and reliability thereby driving the network resources orchestration process. Based on that, the MNO can properly take decisions on the dedicated radio resource quota to be assigned to each RAN slice within a given cell.

Under this context, we have first proposed a SNC-based model, which given *i*) the dedicated radio resource quota for a uRLLC RAN slice in a single cell, *ii*) the target violation probability, *iii*) its traffic demand and *iv*) the CDF for the SINR experienced by its attached UEs, provides the delay bound for the packet transmission delay. To derive such CDF, we relied on a model based on stochastic geometry. This model considers the impact of the interference incurred

by multiple RAN slices deployed in neighbor cells on the capacity the serving cell offers to a serving RAN slice. Additionally, we have proposed a RAN slice orchestrator to plan in advance the deployment of multiple uRLLC RAN slices in a multi-cell environment. Specifically, our orchestration solution relies on a novel heuristic to compute the dedicated radio resource quotas which ensure the performance requirements of each RAN slice in the long term.

We have validated the proposed SNC-based model by means of an exhaustive simulation campaign, demonstrating it provides a conservative upper estimation of the delay bound for the packet transmission delay of an uRLLC RAN slice, given its target violation probability. This makes the proposed model suitable for ensuring the performance requirements of uRLLC RAN slices when the MNO plan them in advance. Additionally, we have showed the benefits of using the proposed heuristic when the RAN slice orchestrator simultaneously plans the deployment of several uRLLC RAN slice in a multi-cellular scenario under the assumption of radio resource scarcity.

Acknowledgments

This work is partially supported by the H2020 research and innovation project 5G-CLARITY (Grant No. 871428); the Spanish Ministry of Economy and Competitiveness, the European Regional Development Fund (Project PID2019-108713RB-C53); and the Spanish Ministry of Education, Culture and Sport (FPU Grant 17/01844)

References

[1] I. Vision, "Framework and overall objectives of the future development of IMT for 2020 and beyond," *International Telecommunication Union (ITU), Document, Radiocommunication Study Groups*, 2015.

[2] J. Prados-Garzon, P. Ameigeiras, J. Ordonez-Lucena, P. Muñoz, O. Adamuz-Hinojosa, and D. Camps-Mur, "5G Non-Public Networks: Standardization, Architectures and Challenges," *IEEE Access*, vol. 9, pp. 153893–153908, 2021.

[3] J. Ordonez-Lucena, P. Ameigeiras, L. M. Contreras, J. Folgueira, and D. R. López, "On the Rollout of Network Slicing in Carrier Networks: A Technology Radar," *Sensors*, vol. 21, no. 23, 2021.

[4] 3GPP TS 28541 V.17.0.0, "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3 (Release 17)," Sept. 2020.

[5] X. Ge, "Ultra-Reliable Low-Latency Communications in Autonomous Vehicular Networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5005–5016, 2019.

[6] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu, and M. Kadoch, "Dynamic Resource Allocation With RAN Slicing and Scheduling for uRLLC and eMBB Hybrid Services," *IEEE Access*, vol. 8, pp. 34538–34551, 2020.

[7] C. Guo, L. Liang, and G. Y. Li, "Resource Allocation for Vehicular Communications With Low Latency and High Reliability," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3887–3902, 2019.

[8] C. Guo, L. Liang, and G. Y. Li, "Resource Allocation for High-Reliability Low-Latency Vehicular Communications With Packet Retransmission," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6219–6230, 2019.

[9] M. Patra, R. Thakur, and C. S. R. Murthy, "Improving Delay and Energy Efficiency of Vehicular Networks Using Mobile Femto Access Points," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1496–1505, 2017.

[10] L. Chinchilla-Romero, J. Prados-Garzon, P. Ameigeiras, P. Muñoz, and J. M. Lopez-Soler, "5G Infrastructure Network Slicing: E2E Mean Delay Model and Effectiveness Assessment to Reduce Downtimes in Industry 4.0," *Sensors*, vol. 22, no. 1, 2022.

[11] M. Fidler and A. Rizk, "A Guide to the Stochastic Network Calculus," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–105, 2015.

[12] K. Katsaros, M. Dianati, R. Tafazolli, and X. Guo, "End-to-End Delay Bound Analysis for Location-Based Routing in Hybrid Vehicular Networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7462–7475, 2016.

[13] J. P. Champati, H. Al-Zubaidy, and J. Gross, "Transient Analysis for Multihop Wireless Networks Under Static Routing," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 722–735, 2020.

[14] S. Ma, X. Chen, Z. Li, and Y. Chen, "Performance evaluation of URLLC in 5G based on stochastic network calculus," *Mob. Netw. Appl.*, vol. 26, no. 3, pp. 1182–1194, 2021.

[15] Á. A. Cardoso, M. V. G. Ferreira, and F. H. T. Vieira, "Delay bound estimation for multicarrier 5G systems considering lognormal beta traffic envelope and stochastic service curve," *Trans. Emerg. Telecommun. Technol.*, p. e4281, 2021.

[16] C. Xiao *et al.*, "Downlink MIMO-NOMA for Ultra-Reliable Low-Latency Communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 780–794, 2019.

[17] S. Schiessl, M. Skoglund, and J. Gross, "NOMA in the Uplink: Delay Analysis With Imperfect CSI and Finite-Length Coding," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3879–3893, 2020.

[18] Q. Xu, J. Wang, and K. Wu, "Learning-Based Dynamic Resource Provisioning for Network Slicing with Ensured End-to-End Performance Bound," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 28–41, 2020.

[19] C.-F. Liu, M. Bennis, and H. V. Poor, "Latency and Reliability-Aware Task Offloading and Resource Allocation for Mobile Edge Computing," in *IEEE Globecom , Singapore*, pp. 1–7, 2017.

[20] J. García-Morales, M. C. Lucas-Estañ, and J. Gozalvez, "Latency-Sensitive 5G RAN Slicing for Industry 4.0," *IEEE Access*, vol. 7, pp. 143139–143159, 2019.

[21] T. Guo and A. Suárez, "Enabling 5G RAN Slicing With EDF Slice Scheduling," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2865–2877, 2019.

[22] 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, 2019.

[23] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, and X. Costa-Pérez, "A Machine Learning Approach to 5G Infrastructure Market Optimization," *IEEE Trans. Mobile Comput.*, vol. 19, no. 3, pp. 498–512, 2020.

[24] L. Zanzi, V. Sciancalepore, A. Garcia-Saavedra, H. D. Schotten, and X. Costa-Pérez, "LACO: A Latency-Driven Network Slicing Orchestration in Beyond-5G Networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 667–682, 2021.

[25] Y. Jiang, Y. Liu, *et al.*, *Stochastic network calculus*, vol. 1. Springer, 2008.

[26] C.-S. Chang, *Performance guarantees in communication networks*. Springer Science & Business Media, 2012.

[27] J.-Y. Le Boudec and P. Thiran, *Network calculus: a theory of deterministic queuing systems for the internet*, vol. 2050. Springer Science & Business Media, 2001.

[28] S. Ross, *A First Course in Probability*. Pearson, 2014.

[29] V. B. Iversen, "Teletraffic engineering and network planning," 2015.

[30] A. Karamyshev, E. Khorov, A. Krasilov, and I. Akyildiz, "Fast and accurate analytical tools to estimate network capacity for URLLC in 5G systems," *Comput. Netw.*, vol. 178, p. 107331, 2020.

[31] 3GPP TS 38.306 V.16.2.0, "NR; User Equipment (UE) radio access capabilities (Release 16)," Oct. 2019.

[32] Y. Hmamouche, M. Benjillali, S. Saoudi, H. Yanikomeroglu, and M. D. Renzo, "New Trends in Stochastic Geometry for Wireless Networks: A Tutorial and Survey," *Proceedings of the IEEE*, pp. 1–53, 2021.

[33] S. M. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," in *WiOpt, Tsukuba Science City, Japan*, pp. 119–124, 2013.

[34] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A Tractable Approach to Coverage and Rate in Cellular Networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, 2011.

[35] M. Di Renzo, W. Lu, and P. Guan, "The Intensity Matching Approach: A Tractable Stochastic Geometry Approximation to System-Level Analysis of Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5963–5983, 2016.

[36] J. G. Andrews, A. K. Gupta, and H. S. Dhillon, "A primer on cellular network analysis using stochastic geometry," *arXiv preprint arXiv:1604.03183*, 2016.

[37] H. S. Dhillon and J. G. Andrews, "Downlink rate distribution in heterogeneous cellular networks under generalized cell selection," *IEEE Commun. Lett.*, vol. 3, no. 1, pp. 42–45, 2014.

[38] 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, 2010.

[39] Shirvanimoghaddam *et al.*, "Short Block-Length Codes for Ultra-Reliable Low Latency Communications," *IEEE Commun. Mag.*, vol. 57, no. 2, pp. 130–137, 2019.

[40] 3GPP TS 38.214 V.16.5.0, "NR; Physical layer procedures for data (Release 16)," Mar. 2021.

Part IV

Conslusions

Chapter 12

Conclusions

This thesis has focused on management mechanisms to automate the deployment and orchestration of Radio Access Network (RAN) slices under an architectural framework based on the Third Generation Partnership Project (3GPP) and European Telecommunications Standards Institute (ETSI)-Network Function Virtualization (NFV) specifications. In this final chapter, the main findings of this thesis are summarized. Furthermore, it also provides possible avenues for future research which arise from the presented conclusions.

12.1 Main Findings

Regarding the design of a RAN slicing framework based on the 3GPP and ETSI-NFV standards, the most relevant conclusions are:

- The options for automatically scaling the Virtualized Network Functions (VNFs) of a RAN slice with the NFV-Management and Orchestration (MANO) are limited by the way the NFV descriptors are constructed. During their lifecyles, these VNFs only can move among the instantiation levels defined in the NFV descriptors, so their design is critical to ensure an effective automated scaling. How these instantiation levels are built is analyzed in this thesis. Furthermore, an illustrative example where these levels are included for scaling the virtualized part of a RAN slice is provided.

- To automate the scaling procedure, the NFV-MANO runs a Dynamic Re-

source Provisioning Algorithm (DRPA) that, (a) taking the content of the NFV descriptors; and (b) the information of the operative VNFs, determines the optimum instantiation level towards the VNFs must be scaled to. This output forces the way the scaling procedures are performed with NFV-MANO. To clarify the interactions and information exchanges between the NFV-MANO blocks in the scaling procedure, this thesis has provided an ETSI-NFV-compliant workflow for such procedure.

- This thesis has proposed a description model that harmonizes the 3GPP and ETSI-NFV viewpoints with the aim of enabling the customization and deployment of the constituent next Generation NodeBs (gNBs) of a RAN slice. The proposed solution benefits from the reusability provided by NFV descriptors to define the underlying resources of the Centralized Unit (CU) and Distributed Units (DUs) of the gNBs for each RAN slice. To customize the behavior of each RAN slice, the most representative radio parameters to configure their gNBs have been identified. Furthermore, to facilitate the comprehension of the proposal, an example of the description of three RAN slices for enhanced Mobile Broadband (eMBB), massive Machine Type Communication (mMTC) and ultra-Reliable Low Latency Communication (uRLLC) communication services has been provided.

- This thesis sheds light on the key aspects for sharing gNB components between RAN slices. If the Radio Resource Management (RRM) algorithms at intra-slice level and the Radio Resource Control (RRC) layer are slice specific or slice-aware, the gNB components could be shared because the treatment of the Fifth Generation (5G)-New Radio (NR) functionalities for the Data Radio Bearers (DRBs) of each RAN slice could be specifically configured. This thesis has also identified that controlling the number of radio resources allocated to each RAN slice and the Modulation and Coding Schemes (MCSs) assigned to their User Equipments (UEs), the isolation between RAN slices can be guaranteed in a gNB component implemented as VNF. Finally a description model to define the lifecycle management of shared gNB components using the 3GPP and the NFV management templates has been proposed.

Regarding the design, implementation and evaluation of management mech-

anisms for planning RAN slices, the most relevant conclusions are:

- This thesis has analyzed different ways of hiring capacity to the Mobile Network Operator (MNO) and, based on them, a set of radio resource planning strategies with different degrees of flexibility for allocating radio resources have been analyzed. These strategies have been evaluated in a 5G multi-tenant network through snapshot-based simulations. The performance of the strategies have been assessed in terms of scalability, isolation, utilization and efficiency.

- An analytical model to evaluate the UE blocking probability for a RAN slice with Guaranteed Bit Rate (GBR) requirements in an Orthogonal Frequency-Division Multiple Access (OFDMA) cell has been proposed. This model is based on a multi-dimensional Erlang-B system. It meets the reversibility property which means the proposed model allows the adoption of an arbitrary distribution for the UE session duration. Additionally, this property involves the solution for the state probabilities has product form, thus it reduces the complexity for their computation. Furthermore, the proposed model considers as input a generic distribution for the average Signal-to-Interference-plus-Noise Ratio (SINR) which an arbitrary UE session could perceive. This approach allows a more precise characterization of the UEs' channel quality within the cell. Another innovation is the proposed model considers the cell implements a channel-aware scheduler to dynamically allocate the radio resources planned for this RAN slice to its active UE sessions. Specifically, this thesis has provided the mathematical formulation for the GBR achieved by an active UE session when the cell implements a representative channel-aware scheduler. The results show that our model exhibits an estimation error for the UE blocking probability below 1.5%.

- Based on the previous model, a mathematical framework for planning the radio resources of RAN slices offering Guaranteed Bit Rate (GBR) services has been provided. This is capable of translating the GBR requirements of the requested communication services into the minimum radio resource quota assigned for each RAN slice in each cell. Each quota guarantees the

UE blocking probability for a RAN slice in a cell is below an upper bound under the inter-cell interference levels presented in the busy hour. This framework uses game theory to model the radio resource planning in RAN slicing. Specifically, the radio resource planning is formulated as multiple ordinal potential games, one per requested or already deployed RAN slice. The goal of each game is to guarantee the GBR requirements of each considered RAN slice, while its UE blocking probability in each cell is below the upper bound. To solve the formulated problem, novel strategies have been designed. These strategies are based on better response dynamics and aim to minimize the UE blocking probability for all the RAN slices. Detailed simulations have been performed to demonstrate the effectiveness of the proposed solution in terms of performance, adaptability and renegotiation capability.

- A Stochastic Network Calculus (SNC)-based model is provided to compute the delay bound for a RAN slice. To that end, this model considers as inputs: a) the amount of radio resources allocated for such RAN slice; (b) the target violation probability; and (c) its traffic characteristics. Furthermore, this thesis has proposed a delay-driven orchestrator that using the SNC-based model, plans the allocation of radio resources for multiple RAN slices, each providing an uRLLC service with specific requirements in terms of latency a reliability. This orchestrator aims to accommodate all the RAN slices whereas the differences between the achieved delay bounds and the target delay bounds are minimized. The model has been validated and detailed simulations have been performed to demonstrate the effectiveness of the proposed orchestrator in terms of performance and scalability.

12.2 Future Work

The work presented herein opens many possibilities for future research on RAN slicing. Some of them are a direct consequence of the findings stated above, others are aspects that could not be addressed due to the limited time resources of the PhD study and, finally, there are points which fall outside the scope of the thesis.

With respect to the proposed management framework for RAN slicing, several

challenges lie ahead:

- Using the proposed management templates in a real scenario. This means describing real software implementations of gNB components with the proposed NFV-based management templates. Specifically, the idea is to define the set of Instantiation Levels (ILs) to scale on runtime the virtualized components of real gNBs. This task is specially challenging if some of these virtualized components must be shared among two or more RAN slices.

- Integrating the proposed NFV templates in a real implementation of the NFV-MANO (e.g., Open-Source MANO and/or OpenBaton) and use them to deploy one or more RAN slices. The challenge lies on deploying on demand and in an automated way these RAN slices over a cellular environment.

- Implement the proposed scaling workflow in a real implementation of the NFV-MANO. Then, based on such implementation, analyzing the impact of the proposed scaling procedure in the performance of the RAN slice. For instance, measuring the time spent by the real NFV-MANO implementation to scale the RAN slice, and how the performance of such RAN slice is affected by such procedure.

- Analyzing how the proposed mechanisms for sharing gNB components among RAN slices impact on their performance in quantitative terms. This analysis could also involve the design of a mechanism to look for a trade-off solution between (a) reducing the required virtual resources due to most of them are shared among gNB components; and (b) reaching the performance requirements required by each RAN slice.

Focusing on the proposed solutions for planning RAN slices, several challenges lie ahead:

- In the queueing-based model to evaluate the UE blocking probability, and the SNC-based model to derive the packet dealy bound, we assume a RAN slice is guaranteed a specific amount of radio resources. However, these models omit that additional radio resources from other RAN slices could

be occasionally allocated for a target RAN slice if the remaining RAN slices are not using them. These models could be improved by considering this approach.

- Focusing on the queueing-based model to evaluate the UE blocking probability, a challenge lies on defining a mechanism to determine the value of the fairness parameter α which provides the lowest UE blocking probability. This mechanism could be leveraged by the MNO to establish lower minimum radio resource quotas for each RAN slice and thus doing a better radio resource usage.

- Extending the proposed SNC-based model to consider the computational part of a gNB, i.e., considering the time spent by the gNB to process an uRLLC packet before its transmission via the radio interface. This extension will allow the MNO to completely characterize the delay bound of network slice in the RAN.

- Integrating the solutions for planning GBR slices and uRLLC slices in a single mathematical framework. The goal of this framework will be simultaneously planning of uRLLC and GBR slices in the same geographical area. The challenge lies on ensuring the requirements of both type of slices are met in the long term. Furthermore, this framework should also consider the priority of these RAN slices (e.g., in terms of cost) in scenarios with radio resource scarcity.

Appendices

Appendix A

Resumen

El presente apéndice incluye un amplio resumen en castellano de la memoria de tesis con el objetivo de cumplir con la normativa de la Escuela de Posgrado de la Universidad de Granada referente a la redacción de tesis doctorales cuando éstas son escritas en inglés.

A.1 Introducción

Las redes móviles de quinta generación (5G) han emergido como una solución tecnológica para impulsar la digitalización de la sociedad. Concretamente, la industria y la academia han materializado su esfuerzo en estándares y contribuciones que permitirán a los operadores de redes móviles desplegar nuevos servicios de comunicación con requisitos muy divergentes entre sí en términos de rendimiento y funcionalidad. Hay un consenso general en agrupar estos servicios de comunicación en tres categorías:

- Servicios de banda ancha mejorados (-*enhanced Mobile Broadband*- eMBB): Esta categoría comprende principalmente servicios de acceso a contenido multimedia. Debido al continuo incremento de demanda de estos servicios, la red móvil 5G deberá de proporcionar una mayor capacidad. Ejemplos de servicios: realidad aumentada, transmisión de vídeos en ultra alta definición (4K).

- Comunicaciones masivas de tipo máquina (-*massive Machine Type Commu-*

nication- mMTC): Esta categoría comprende aquellos servicios de comunicación en los cuales los dispositivos se comunican entre sí sin la intervención de humanos. Además, estos servicios se caracterizan por una gran cantidad de dispositivos de bajo coste y equipados con baterías de gran duración, que transmiten datos de forma infrecuente y no sensibles al retardo. Ejemplos de servicios: ciudad inteligente, agricultura inteligente.

- Comunicaciones ultra fiables de baja latencia (-*ultra-Reliable Low Latency Comunication*- uRLLC): Esta categoría comprende servicios de comunicación con requisitos estrictos en términos de latencia y fiabilidad. En estos servicios, tanto humanos como máquinas intervienen en el proceso de comunicación. Ejemplos de servicios: cirugía remota, automatización de la industria, vehículos autónomos.

Estos nuevos servicios imponen requisitos de rendimiento que son divergentes entre sí. Por ello, es difícil implementar simultáneamente estos servicios en una arquitectura de red única, como por ejemplo la definida para las redes de cuarta generación (4G). Para abordar este problema, *network slicing* ha sido propuesto como una solución tecnológica. Esta solución consiste en particionar lógicamente la infraestructura física de un operador de red móvil en un conjunto de redes virtuales independientes, llamadas *network slices*, cada una adaptada a los requisitos específicos de un servicio de comunicación.

Para llevar *network slicing* a una red móvil 5G, el operador de red tiene que satisfacer una serie de principios y condiciones. Los más importantes se resumen a continuación:

- **Automatización de la gestión**: Este principio habilita al operador de red a configurar, desplegar, operar y terminar *network slices* sin la intervención humana. Para ese fin, estas tareas de gestión tienen que depender de mecanismos de señalización que continuamente comprueben que los requisitos de rendimiento y de funcionalidad de cada *network slice* se cumplen a lo largo de su tiempo de vida.

- **Aislamiento entre *network slices***: Este principio se define en términos de rendimiento, seguridad y gestión:

  - Rendimiento: El operador de red tiene que asegurar que los requisitos específicos de rendimiento de un *network slice* se cumplen siempre, independientemente de los niveles de congestión y rendimiento de los otros *network slices*.
  - Seguridad: Los ataques de seguridad en un *network slice* no pueden impactar en otros *network slices*. Además, cada *network slice* debería tener funciones de seguridad independientes que prevengan a entidades no autorizadas de leer o escribir la configuración específica de este *network slice*.
  - Gestión: Desde el punto de vista de un consumidor de *network slices*, cada *network slice* podría ser gestionado como una red independiente. Para ese fin, el operador de red tiene que definir una serie de mecanismos que permitan a cada consumidor de *network slices* acceder individualmente a ciertas capacidades del *network slice* que están consumiendo.

- **Personalización**: Este principio garantiza que el operador de red use eficientemente las funciones de red y los recursos de infraestructura asignados a cada *network slice* mientras sus requisitos de funcionalidad y rendimiento se cumplen a lo largo de su tiempo de vida.

- **Elasticidad**: Este principio asegura que el operador de red satisface los requisitos de rendimiento de un *network slice* considerando variaciones temporales de: (a) las condiciones del núcleo de red y la red de acceso radio, (b) la cantidad de usuarios servidos por dicho *network slice*, (c) la demanda de tráfico y (d) el área geográfica donde sirve dicho *network slice* a causa de la movilidad de sus usuarios.

- **Programabilidad**: Esta condición permite a los consumidores de *network slices* controlar los recursos asignados a cada *network slice* y su configuración a través de interfaces de programación de aplicaciones (-*Application Programming Interfaces*- APIs) que exponen las capacidades de estos *network slices*. Esto facilita la personalización de servicios bajo demanda y la elasticidad de recursos.

- **Despliegue extremo a extremo**: Este principio permite al operador de red desplegar un *network slice* a lo largo de diferentes dominios administrativos (es decir, ubicaciones de infraestructura gestionadas por diferentes proveedores) y de red (núcleo de red, red de acceso radio y red de transporte).

- **Abstracción jerárquica**: Este principio permite al operador de red ejecutar un procediemiento de abstracción, que podría repetirse sucesivamente en niveles más altos. El propósito de este procedimiento es que los recursos de un *network slice*, asignados a un consumidor específico, puedan ser negociados para ser asignados a un tercero.

A.2 Tecnologías Vinculadas a *Network Slicing*

A.2.1 Virtualización de Funciones de Red

Tradicionalmente ha existido una relación muy fuerte entre las funciones de una red y el hardware que el operador de red usa para implementar dichas funciones, el cual es específico de un proveedor. Este enfoque presenta algunos inconvenientes. Por un lado, cada vez que el operador de red requiere actualizar la red o incluir una nueva función, tiene que: (a) adquirir nuevo hardware específico, el cual suele ser caro, y (b) encontrar lugar para alojarlo, su fuente de alimentación y los sistemas de refrigeración para este hardware. Por otro lado, el operador requiere de personal altamente cualificado para diseñar, integrar y operar este hardware en su infraestructura de red. Todo esto dificulta considerablemente al operador de red incluir nuevos servicios y características de red con gran agilidad y a un precio reducido.

Para superar estos problemas, la virtualización de funciones de red (-*Network Functions Virtualization*- NFV) ha emergido como una solución tecnológica. Esta consiste en desacoplar las funciones de red del hardware propietario, para ejecutarlas como software en servidores de propósito general. Con esta tecnología, las funciones de red que requiere un *network slice* se pueden construir con uno o más contenedores de virtualización, es decir con máquinas virtuales o contenedores Linux.

A.2.2 Redes Definidas por Software

Las redes definidas por software (-*Software Defined Networking*- SDN) son una tecnología que se basa principalmente en los siguientes principios: (a) el desacoplo de los planos de control y de datos, (b) la centralización lógica del plano de control, (c) la programabilidad de la red, (d) el uso de interfaces abiertas.

Esta tecnología separa completamente los planos de control y de datos de una infraestructura de red, habilitando de esa forma la programabilidad de la red. Concretamente, el plano de control está compuesto por una entidad externa conocida como controlador SDN, el cual está implementado mediante software. El propósito de este controlador es controlar un conjunto de dispositivos de red de bajo coste, los cuales comprenden el plano de datos de la red. En este paradigma, el controlador SDN tiene una vista global de la red, y puede tomar decisiones de gestión del tráfico del plano de datos acorde las políticas de operación impuestas por el operador de red.

La tecnología SDN proporcionará beneficios sustanciales al operador de red a la hora de implementar *network slices*, concretamente en la red de transporte. El uso de un controlador centralizado facilitará al operador de red automatizar la gestión de los recursos de la red de transporte asignados a cada *network slice*. Esto significa que las reglas de reenvío en los dispositivos del plano de datos se pueden ajustar dinámicamente para acomodar las demandas de tráfico de los *network slices* desplegados en la red.

A.2.3 Computación perimetral de acceso múltiple

La computación perimetral de acceso múltiple (-*Multi-access Edge Computing*- MEC) es una tecnología emergente en las redes móviles 5G. Esta tecnología tiene como objetivo extender los recursos de cómputo, almacenamiento y red desde los centros de datos localizados en el núcleo de red hacía el borde de la red, es decir cerca de la red de acceso radio. Esto permitirá al operador de red: (a) reducir los cuellos de botella en el núcleo de red, (b) ayudar en la ejecución de tareas computacionalmente costosas de los usuarios finales, que pasarán a ser ejecutadas en los servidores del borde de la red, y (c) llevar las funcionalidades del plano de datos del núcleo de red a los servidores del borde para reducir el retardo experimentado en la transmisión de paquetes.

El uso de MEC junto a *network slicing* puede proporcionar beneficios potenciales a la hora de implementar un amplio abanico de nuevos servicios de comunicación. Concretamente, MEC es de gran interés para implementar *network slices* que acomoden servicios uRLLC y mMTC. En el caso de servicios uRLLC, el operador de red puede aprovechar los servidores alojados en el borde de la red para reducir el retardo de transmisión de los paquetes experimentado por los usuarios. En el caso de un servicio mMTC, el operador de red puede usar los servidores del borde de la red para recoger información de los usuarios, procesarla, y usar la información resultante para ofrecer otros tipos de servicios.

A.3 Principales Desafíos en *Network Slicing*

Para realizar completamente una red móvil 5G basada en *network slicing*, varios desafíos deben de ser abordados. Los más significativos se resumen a continuación:

- **Virtualización de la red de acceso radio**: La virtualización de la red de acceso radio debería de proporcionar al operador de red la capacidad de crear instancias de redes de acceso radio virtuales, cada una con un conjunto específico de funcionalidades y con requisitos diferentes. La tecnología de virtualización ya ha sido aplicada al núcleo de red. Sin embargo, el uso de esta tecnología en la red de acceso radio es desafiante ya que la capacidad de los nodos de acceso inalámbricos dependen principalmente de efectos de canal como los desvanecimientos rápidos y lentos, e interferencia entre nodos de acceso; ambos efectos variantes en el tiempo. Esto significa que los métodos de virtualización usados en el núcleo de red no pueden ser directamente implementados en la red de acceso radio. Por tanto, se tienen que diseñar nuevos mecanismos para virtualizar los nodos de acceso inalámbricos.

- **Automatización de la gestión**: Para evitar labores manuales y errores, la gestión de *network slices* debería de ser automática, es decir sin la intervención humana. Esto es desafiante ya que el operador de red tiene que considerar múltiples dimensiones y tecnologías para (a) instanciar, activar, operar y terminar *network slices*, (b) ajustar el balanceo de carga,

políticas de tarificación, seguridad y calidad de servicio para cada *network slice*; (c) abstraer y aislar los recursos asignados a cada *network slice*, y (d) implementar mecanismos de compartición de recursos entre *network slices*.

- **Compartición de recursos**: Esto puede ser implementado mediante particiones estáticas o dinámicas. Debido a que la demanda de tráfico de cada *network slice* podría cambar a lo largo de su tiempo de vida, la compartición dinámica de recursos haría más eficiente la utilización de los recursos de la infraestructura del operador de red. Implementar técnicas de compartición dinámica de recursos es desafiante porque el operador de red tiene que garantizar al mismo tiempo que los requisitos de rendimiento de cada *network slice* se cumplen a lo largo de su tiempo de vida.

- **Seguridad**: Esto es un problema crítico ya que los diferentes *network slices* comparten la misma infraestructura de red. Además, cada *network slice* podría tener requisitos específicos en políticas de seguridad. Todo esto implica que el operador de red tiene que diseñar mecanismos de seguridad para un *network slice* considerando el impacto de dichos mecanismos en la seguirdad del resto de *network slices*.

- **Gestión de la movilidad**: Por un lado, el soporte de movilidad podría ser opcional para algunos *network slices*. Por ejemplo, *network slices* para servicios mMTC podrían no requerir movilidad debido a la posición fija de sus usuarios (ejemplo: sensores). Por otro lado, *network slices* que requieren movilidad podrían diferir en sus requisitos. Por ejemplo, los requisitos de movilidad de *network slices* para servicios eMBB y uRLLC podrían diferir. Por ende, el diseño de un protocolo de movilidad orientado a cada *network slice* es necesario para abordar los desafíos de movilidad en *network slicing*.

- **Composición del servicio con funciones de red muy granulares**: Desde el punto de vista del operador de red, es más sencillo componer las funciones de red de un *network slice* con una menor granularidad, debido a que se tiene que definir menos interfaces para encadenar estas funciones. El principal inconveniente de este enfoque es la flexibilidad reducida para adaptar los *network slices* a sus demandas de tráfico, las cuales cambian a

lo largo del tiempo. Por este motivo, es más deseable para el operador que las funciones de red sean más granulares.

- **Traducción de requisitos de servicio a requisitos de recursos**: Un desafío importante para la realización de *network slices* es como ir desde una descripción a alto nivel de los requisitos de servicio hasta la cantidad concreta de recursos de infraestructura que el operador de red necesita para satisfacer tales requisitos de servicio a lo largo del tiempo de vida de los *network slices*. Para abordar este problema, el operador de red necesita lenguajes de descripción específicos para cada dominio de red. Estos se usarán para describir tanto los requisitos de cada *network slice* así como para describir los recursos de infraestructura que se necesitan para implementar dicho *network slice*. Esto es desafiante, especialmente en la red de acceso radio, donde la capacidad ofrecida a un *network slice* depende de los efectos de canal como los desvanecimientos rápidos y lentos, y la interferencia entre nodos de acceso.

- **Orquestación y gestión de *network slices* extremo a extremo**: Un *network slice* podría ser desplegado en múltiples dominios administrativos y de red. Como cada dominio tiene requisitos específicos, sería deseable definir múltiples entidades de gestión, cada una responsable de gestionar el ciclo de vida de un *network slice* en un dominio específico. El desafío radica en coordinar las tareas de gestión y orquestación de cada dominio mientras los requisitos de rendimiento de cada *network slice* se cumplen a largo plazo.

A.4 Alcance y Objetivos de la Tesis Doctoral

Network slicing es una solución tecnológica que tendrá un rol clave en las redes móviles 5G. Concretamente, esta solución permitirá al operador de red proporcionar económicamente servicios de comunicación emergentes, cada uno con un conjunto específico de requisitos funcionales y de rendimiento, sobre una infraestructura física de red común.

En un primer intento, la comunidad investigadora ha dedicado esfuerzos en la integración de *network slicing* en el núcleo de red. Luego, considerando los

beneficios proporcionados por esta tecnología, el análisis de *network slicing* ha sido extendido a otros dominios de red, como la red de transporte y la red de acceso radio. En esta tesis, nos centramos en la red de acceso radio.

Uno de los principales retos en *network slicing* es la virtualización de la red de acceso radio. Para abordarla, primero se necesita construir conocimiento en NFV y analizar en detalle como esta tecnología habilita al operador de red desplegar y operar *netowrk slices* en la red de acceso radio. La organización responsable de estandarizar NFV es el Instituto Europeo de Normas de Telecomunicaciones (-*European Telecommunications Standards Institute*- ETSI), concretamente el grupo NFV. Esta organización ha definido un marco de referencia, conocido como NFV-*Management and Orchestration* (MANO), para gestionar el ciclo de vida de las funciones de red que se implementan mediante software, es decir las funciones de red virtualizadas. Por tanto, es clave comprender que funcionalidades de la red de acceso radio podrían ser virtualizadas y como el marco de referencia NFV-MANO podría instanciar, escalar hacia arriba/abajo y liberar los recursos virtuales asignados a las funciones de red virtualizadas.

La organización ETSI-NFV también ha definido un conjunto de plantillas de gestión, conocidas como descriptores NFV, que pueden ser usados por el marco de referencia NFV-MANO para instanciar, escalar y liberar los recursos virtuales asignados a cada función de red virtualizada de forma automática. El desafío radica en describir con los descriptores NFV los recursos virtuales de las funciones de red virtualizadas que acomodarán las fluctuaciones temporales y espaciales de la demanda de tráfico de un *network slice* en un entorno celular.

Otro aspecto interesante en NFV es como compartir las mismas instancias de funciones de red entre múltiples *network slices*. Este enfoque conllevaría ganancias de multiplexación estadística a la hora de utilizar los recursos virtualizados. Sin embargo, es un desafío lograr el nivel de personalización requerido por cada *network* slice en la red de acceso radio. Esto significa que los impactos de compartir las funcionalidades de la red de acceso radio entre múltiples *network slices* tiene que ser analizado en términos de personalización y aislamiento.

Otro desafío a ser abordado radica en la traducción de requisitos de servicio de un *network slice* a la cantidad de recursos de infraestructura que el operador de red requiere para desplegar este *network slice* y operarlo a lo largo de su tiempo de vida. Además de ETSI-NFV, el Proyecto Asociación de Tercera Generación

(-*3rd Generation Partnership Project*- 3GPP) tendrá también un rol clave en este contexto. Esta organización es responsable de definir las funciones de red que un *network slice* necesita en la red de acceso radio, incluyendo sus funcionalidades y capacidades. Por tanto, es clave entender desde un punto de vista de gestión, como los puntos de vista en *network slicing* de ETSI-NFV y 3GPP pueden ser usados por un operador de red para proceder desde la solicitud de despliegue de un *network slice* hasta la provisión de dicho *network slice* en la red de acceso radio.

Por un lado, se tiene que analizar como los requisitos de servicio de un *network slice* en la red de acceso radio pueden mapearse a la compresión de *network slicing* del 3GPP, es decir a los modelos de información que describen las funcionalidades de la red de acceso radio. Por otro lado, es necesario trabajar en cómo el operador de red debería mapear los modelos de información del 3GPP a los descriptores de NFV que definen los recursos virtuales para desplegar las funciones de red virtualizadas de los *network slices* en la red de acceso radio.

Además de enalzar los modelos de información del 3GPP y de ETSI-NFV para *network slicing*, el operador de red debe de confiar en un marco de referencia matemático para calcular la cantidad de recursos requeridos para desplegar múltiples *network slices* sobre un entorno celular común. Concretamente, este marco de referencia matemático tiene que planificar la cantidad de recursos que satisfacen simultáneamente los requisitos de rendimiento de cada *network slice* en la red de acceso radio a largo plazo.

Bajo este contexto, el objetivo principal de este proyecto de tesis es estudiar mecanismos de gestión para automatizar el despliegue y orquestación de *network slices* en la red de acceso radio bajo un marco de referencia arquitectónico basado en las especificaciones del 3GPP y ETSI-NFV. Para ese fin, esta tesis aborda los siguientes objetivos específicos:

- **Objetivo 1**: Diseño de un marco de referencia arquitectónico basado en los estándares de ETSI-NFV y 3GPP para *network slicing* en la red de acceso radio. Este objetivo se divide en los siguientes sub-objetivos:
  - **Sub-objetivo 1.1**: Diseñar procedimientos de gestión para desplegar y operar la parte virtualizada de un *network slice* en la red de acceso radio con el marco de referencia NFV-MANO.

  - **Sub-objetivo 1.2**: Proponer un marco de referencia basado en los estándares 3GPP y ETSI-NFV para gestionar *network slices* en la red de acceso radio a lo largo de sus ciclos de vida.
  - **Sub-objetivo 1.3**: Definir un mecanismo para la compartición de las funcionalidades de la red de acceso radio y de sus recursos subyacentes entre múltiples *network slices*.

- **Objetivo 2**: Diseñar, implementar y evaluar mecanismos de gestión para la planificación de *network slices* en la red de acceso radio. Esta tesis se centra en *network slices* para servicios con requisitos en términos o bien de tasa de datos garantizada o de latencia y fiabilidad. Bajo este contexto, este objetivo se divide en los siguientes sub-objetivos:

  - **Sub-objetivo 2.1**: Diseñar, implementar y evaluar una solución para planificar *network slices* en la red de acceso radio que proporcionan servicios con requisitos de tasa de datos garantizada.
  - **Sub-objetivo 2.2**: Diseñar, implementar y evaluar una solución para planificar *network slices* en la red de acceso radio que proporcionan servicios con requisitos de latencia y fiabilidad.

A.5 Conclusiones

Con respecto al diseño de un marco de referencia basado en los estándares 3GPP y ETSI-NFV para gestionar *network slices* en la red de acceso radio, las principales conclusiones son las siguientes:

- Las opciones para escalar de forma automática las funciones de red virtualizadas de un *network slice* con el marco de referencia NFV-MANO están limitadas por la forma en la que los descriptores de NFV están construidos. Durante el ciclo de vida de las funciones de red virtualizadas, la cantidad de recursos virtuales que estás necesitan dependerán de los niveles de instanciación definidos en los descriptores de NFV, por lo que el diseño de dichos descriptores es crítico para asegurar que el escalado automático sea efectivo. En esta tesis, se ha analizado como estos niveles de instanciación se tiene que diseñar. Además, se ha mostrado un ejemplo ilustrativo donde

estos niveles son usados para escalar la parte virtualizada de un *network slice* en la red de acceso radio.

- Para automatizar el procedimiento de escalado, el marco de referencia NFV-MANO tiene que ejecutar un algoritmo de provisión dinámica de recursos que (a) usando el contenido de los descriptores de NFV, y (b) la información de las funciones de red virtualizadas que están operativas, determinará el nivel de instanciación óptimo a través del cual estas funciones de red virtualizadas se tienen que escalar. Esta salida fuerza la forma en la que los procedimientos de escalado son ejecutados con el marco de referencia NFV-MANO. Para clarificar las interacciones y la información intercambiada por los diferentes bloques funcionales del marco de referencia NFV-MANO, esta tesis ha proporcionado un flujo de trabajo compatible con los estándares de ETSI-NFV para el procedimiento de escalado.

- Esta tesis arroja luz en los aspectos clave para compartir los componentes de un gNB entre *network slices* en la red de acceso radio. Si los algoritmos para la gestión de recursos radio a nivel intra-*slice* y la capa de control de recursos radio son específicos para un *network slice* o consciente de *network slicing*, los componentes de un gNB podrían ser compartidos. El motivo es que debido a que el trato de las funcionalidades radio de 5G para las portadoras de datos radio podrían configurarse de forma específica para cada *network slice*. Esta tesis también ha identificado que controlando el número de recursos radio asignados a cada *network slice* y los esquemas de modulación y codificación asignados a los usuarios servidos por dichos *network slices*, el aislamiento entre *network slices* puede ser garantizado en un componente virtualizado de un gNB. Finalmente se ha propuesto un modelo de descripción que usa las plantillas de gestión definidas por el 3GPP y ETSI-NFV para gestionar el ciclo de vida de componentes de gNB que son compartidos entre *network slices*.

Con respecto al diseño, implementación y evaluación de mecanismos de gestión para planificar *network slices* en la red de acceso radio, las conclusiones más relevantes son las siguientes:

- Esta tesis ha analizado diferentes formas de contratar capacidad con el oper-

ador de red y, basándose en estas, se ha analizado un conjunto de estrategias para planificar recursos radios con diferentes grados de flexibilidad. Estas estrategias han sido evaluadas en una red de acceso radio con múltiples *network slices* a través de simulaciones basadas en capturas. El rendimiento de estas estrategias ha sido evaluado en términos de escalabilidad, aislamiento, utilización y eficiencia.

- Se ha propuesto un modelo analítico para evaluar la probabilidad de bloquear una sesión de datos de usuario para un *network slice* en la red de acceso radio que proporciona un servicio de tasa de datos garantizada. Este modelo está basado en un sistema Erlang-B multidimensional. Este sistema es reversible, lo que significa que el modelo propuesto permite adoptar una distribución arbitraria para la duración de la sesión de datos de usuario. Adicionalmente, esta propiedad implica que la solución para las probabilidades de estado tienen forma de producto, por lo que se reduce la complejidad al calcular dichas probabilidades. Además el modelo propuesto considera como entrada una distribución genérica para el promedio de la relación señal a ruido más interferencia que un usuario podría percibir. Este enfoque permite una caracterización más precisa de la calidad del canal para un usuario dentro de una celda. Otra innovación es que el modelo propuesto considera que la celda implementa un planificador de paquetes consciente del canal para asignar los recursos radio a los usuarios correspondientes. Concretamente, esta tesis ha proporcionado una formulación matemática para la tasa de datos garantizada que una sesión activa de datos de usuario podría alcanzar cuando la celda implementa un caso representativo de planificador de paquetes consciente del canal. Los resultados muestran que el modelo tiene un error de estimación para la probabilidad de bloquear sesiones de datos de usuario por debajo de 1.5%.

- En base al modelo anterior, esta tesis propone un marco de referencia matemático para planificar los recursos radio de varios *network slices* en la red de acceso radio ofreciendo servicios con tasa de datos garantizada. Este marco de referencia es capaz de traducir los requisitos de tasa de datos garantizada de cada *network slice* en la cuota mínima de recursos radio que se asigna a cada *network slice* en una celda. Cada cuota garantiza que la

probabilidad de bloquear una sesión de datos de usuario esta por debajo de un límite superior bajo los niveles de interferencia inter-celda presentes en la hora cargada. El marco de referencia propuesto usa teoría de juegos para modelar la planificación de recursos radio en *network slicing*. Concretamente, la planificación de recursos radio se formula mediante múltiples juegos potenciales ordinales, uno por *network slice* solicitado o desplegado. El propósito de cada juego es garantizar que se cumplen los requisitos de tasa de datos garantizada de cada *network slice* al mismo tiempo que la probabilidad de bloquear una sesión de datos de usuario para cada *network slice* esté por debajo de un umbral superior. Para resolver el problema formulado, se han diseñado un conjunto de estrategias novedosas. Estas están basadas en dinámicas de mejor respuesta y buscan minimizar la probabilidad de bloqueo de datos de usuario para todos los *network slices*. Se han llevado a cabo simulaciones detalladas para demostrar la efectividad de la solución propuesta en términos de rendimiento, adaptabilidad y capacidad de re-negociación.

- Se ha propuesto un modelo basado en *Stochastic Network Calculus* para calcular el umbral de retardo de un *network slice* en la red de acceso radio. Para ese fin, este modelo considera como entradas: (a) la cantidad de recursos radio asignados a dicho *network slice*, (b) la probabilidad de que el retardo de transmisión de un paquete esté por encima de un umbral, y (c) las características de tráfico de dicho *network slice*. Además, esta tesis ha propuesto un orquestador que usa este modelo para planificar los recursos radio de múltiples *network slices*, cada uno proporcionando un servicio uRLLC con requisitos específicos en términos de latencia y fiabilidad. Concretamente, este orquestador busca acomodar todos los *network slices* en la infraestructura del operador de red al mismo tiempo que las diferencias entre los umbrales de retardo logrados y los umbrales de retardo objetivos son mínimas. Se ha validado el modelo basado en *Stochastic Network Calculus* y se han realizado simulaciones detalladas para demostrar la efectividad del orquestador propuesto en términos de rendimiento y escalabilidad.

This thesis has been supported by the following projects and grants:

- Programa de Formación de Profesorado Universitario (FPU) grant from the Spanish Ministry of Education, Culture and Sport (FPU Grant ref. **17/01844**).

- National Research Project **TEC2016-76795-C6-4-R** "5G-CITY: Adaptive Management of 5G Services to Support Critical Events in Cities" funded by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund.

- National Research Project **PID2019-108713RB-C53** "TRUE5G: Towards Zero Touch Network and Services for beyond 5G" funded by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund.

- European Research Project **871428** "5G-CLARITY: Beyond 5G multi-tenant private networks integrating Cellular, WiFi and LiFi, Powered by Artificial Intelligence and Intent Based Policy" funded by H2020 research program.