UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Luiz Eduardo Hupalo

Physical Layer Network Slicing for eMBB and mMTC with Distributed Power Allocation

Florianópolis
2023

Luiz Eduardo Hupalo

Physical Layer Network Slicing for eMBB and mMTC with Distributed Power Allocation

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Engenharia Elétrica.
Supervisor:: Prof. Richard Demo Souza, Dr.
Co-supervisor:: Prof. João Luiz Rebelatto, Dr.

Florianópolis
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Hupalo, Luiz Eduardo
Physical Layer Network Slicing for eMBB and mMTC with Distributed Power Allocation / Luiz Eduardo Hupalo ; orientadora, Richard Demo Souza, coorientador, João Luiz Rebelatto, 2023.
50 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia Elétrica, Florianópolis, 2023.

Inclui referências.

1. Engenharia Elétrica. 2. Beyond 5G. 3. Network Slicing. 4. mMTC. 5. eMBB. I. Souza, Richard Demo. II. Rebelatto, João Luiz. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Elétrica. IV. Título.

Luiz Eduardo Hupalo

Physical Layer Network Slicing for eMBB and mMTC with Distributed Power Allocation

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Ohara Kerusauskas Rayel, Dr.
Universidade Tecnológica Federal do Paraná - UTFPR

Profa. Victoria Dala Pegorara Souto, Dra.
Instituto Nacional de Telecomunicações - Inatel

Prof. Bartolomeu Ferreira Uchôa Filho, Ph.D (Suplente)
Universidade Federal de Santa Catarina - UFSC

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Engenharia Elétrica.

Documento assinado digitalmente
Eduardo Augusto Bezerra
Data: 30/06/2023 15:40:02-0300
CPF: ***.851.577-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Eduardo Augusto Bezerra, Ph.D.
Coordenador em Exercício do Programa de Pós-Graduação

Documento assinado digitalmente
Richard Demo Souza
Data: 30/06/2023 13:27:14-0300
CPF: ***.267.379-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Richard Demo Souza, Dr.
Orientador

Florianópolis, 2023.

ACKNOWLEDGEMENTS

Quero agradecer a minha família por todas as formas de apoio que já recebi em minha vida. Aos meus pais, Luiza e Alexandre, pelos valores a mim ensinados, e ao meu irmão Leandro, quem tenho como exemplo. Ao professor Richard Demo Souza pela orientação e pelas suas excelentes aulas que tive a oportunidade de assistir. Certamente o professor Richard teve um enorme impacto em minha carreira profissional através da disciplina de Comunicações Sem Fio, que cursei na graduação. Ao meu coorientador João Luiz Rebelatto, pela ajuda com o trabalho. Por fim, agradeço a Deus e a boa sorte, que juntos me deram tanta fortuna.

“Rust never sleeps.”

RESUMO

Este trabalho aborda o conceito de fatiamento de uma rede de acesso (RAN) 5G para diferentes serviços com diferentes requisitos, detalhando-o na camada física e de rede. Na primeira etapa, onde o estudo compreende a camada de rede, são detalhados os requisitos de rede e funcionalidades que são fundamentais para a existência do fatiamento de rede, como *Redes Definidas por Software* (SDN), *Virtualização de Função de Rede* (NFVs) e *Virtualized Network Functions* (VNF). É discutida uma abordagem para um caso de uso específico do fatiamento de rede numa arquitetura *Open RAN*, evidenciando detalhes de seu funcionamento. Na segunda etapa, para a camada física, é estudado o compartilhamento de recursos de rádio entre *enhanced mobile broadband* (eMBB) e *massive machine-type communications* (mMTC) de forma ortogonal (H-OMA – *Heterogeneous Orthogonal Multiple Access*) e não ortogonal (H-NOMA – *Heterogeneous Non-Orthogonal Multiple Access*) num cenário de uplink. Para este cenário, visando obter um aumento da quantidade de dispositivos mMTC ativos, é proposto um modelo de alteração da potência de transmissão destes dispositivos sem a necessidade de coordenação entre o usuário e a *Base Station* (BS). Os resultados para este cenário específico mostram ganhos médios de 60% para o caso ortogonal para todos os padrões de tráfego eMBB e de 40% para o caso não ortogonal quando o tráfego eMBB é baixo.

Palavras-chave: 5G. eMBB. mMTC. Network Slicing.

ABSTRACT

This work addresses the concept of slicing a 5G radio access network (RAN) for different services with different requirements, detailing it in the physical and network layer. In the first stage, where the study comprises the network layer, network requirements and features that are fundamental for the existence of network slicing are detailed, such as *Software Defined Networks* (SDN), *Network Function Virtualization* (NFVs) and *Virtualized Network Functions* (VNFs). An approach for a specific use case of network slicing in an *Open RAN* architecture is discussed, showing details of its operation In the second stage, for the physical layer, it is studied the sharing of radio resources between *enhanced mobile broadway* (eMBB) and *massive machine-type communication* (mMTC) in an orthogonal way (H-OMA – Heterogeneous Orthogonal Multiple Access) and non-orthogonal (H-NOMA – Heterogeneous Non-Orthogonal Multiple Access), for an uplink scenario. For this scenario, aiming to increase the number of active mMTC devices, a model is proposed to change the transmission power of these devices without the need for coordination between the user and the *Base Station* (BS). The results for this specific scenario show average gains of 60% for the orthogonal case for all eMBB traffic patterns and 40% for the non-orthogonal case when eMBB traffic is low.

Keywords: 5G. eMBB. mMTC. Network Slicing.

LIST OF FIGURES

Figure 1 – 5G network slicing framework containing the service layer, network function layer, infrastructure layer, and the network slice controller. 18

Figure 2 – Evolution of the Radio Access Network. 21

Figure 3 – Example of an Open RAN architecture when considering Network Slicing. The sharing of some network functions depends on the service type and on its requirements. 22

Figure 4 – Information flow for the RAN SLA assurance use case. The enumerated procedures are: 1) collection of E2 performance; 2) enriched information based on operating KPIs, 3) policies for dynamic SLA assurance, and 4) E2 control messages for resource allocation. 24

Figure 5 – Illustration of O1 RAN action control loop for the described RAN SLA assurance use case between slices. 26

Figure 6 – Time-frequency grid of the aforementioned H-OMA and H-NOMA allocation methods, with $S = 6$ mini-slots and $F = 5$ resources. In H-OMA, eMBB and mMTC services have dedicated resources, and the word "heterogeneous" comes from the fact that the orthogonal allocation is made between devices of different services, and not of the same service. For $f = 2$, the H-OMA is shown with 50% of the time allocated to each service. On the other hand, for $f = 4$, both services access the resources at the same time, characterizing the H-NOMA. 29

Figure 7 – H-NOMA histograms of mMTC SNRs seen by the BS with and without the proposed method. When $\beta = \{1.9\ 0.1\}$, one can clearly see the existence of "clusters of instantaneous SNR". This behavior becomes more evident as the distance between the elements of the vector increases. 36

Figure 8 – H-NOMA histograms of mMTC SNRs seen by the BS with and without the proposed method. It is depicted in this figure how the configuration of the elements present in the β power allocation vector impacts the distribution of the SNRs. 37

Figure 9 – Arrival rate λ_M^{orth} for the H-OMA case, as a function of eMBB rate r_B. The simulation parameters are $\Gamma_M = 5$ dB, $\Gamma_B = 20$ dB, $\varepsilon_M = 10^{-1}$, $A_M = 200$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$. 39

Figure 10 – Arrival rate λ_M^{orth} for the H-OMA case, considering the best β values from Figure 9 for each case, as a function of eMBB rate r_B. The simulation parameters are $\Gamma_M = 5$ dB, $\Gamma_B = 20$ dB, $\varepsilon_M = 10^{-1}$, $A_M = 200$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$. 40
Figure 11 – Arrival rate $\lambda_M^{non-orth}$ for the H-NOMA case, with the clustering procedure, as a function of eMBB rate r_B. The simulation parameters are $\Gamma_M = 5$ dB, $\Gamma_B = 20$ dB, $A_M = 200$, $\varepsilon_M = 10^{-1}$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$. 42
Figure 12 – Arrival rate $\lambda_M^{non-orth}$ for the H-NOMA case, with the envelope containing the β values with best performance from Figure 11, as a function of eMBB rate r_B. The simulation parameters are $\Gamma_M = 5$ dB, $A_M = 200$, $\Gamma_B = 20$ dB, $\varepsilon_M = 10^{-1}$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$. 43
Figure 13 – Arrival rate $\lambda_M^{non-orth}$ for both H-OMA and H-NOMA case, with the envelope containing the β values that allows the best performance, as a function of eMBB rate r_B. It is shown the comparison between the orthogonal and non-orthogonal methods when using the power allocation technique. The simulation parameters are $\Gamma_M = 5$ dB, $\varepsilon_M = 10^{-1}$, $A_M = 200$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$. 44

LIST OF ABBREVIATIONS AND ACRONYMS

1G	First generation of mobile wireless communication system
2G	Second generation of mobile wireless communication system
3G	Third generation of mobile wireless communication system
3GPP	Third Generation Partnership Project
4G	Fourth generation of mobile wireless communication system
5G	Fifth generation of mobile wireless communication system
AD/DA	Analog-to-Digital/Digital-to-Analog
BBU	RAN Baseband Processing Unit
BS	Base Station
CI/CD	Continuous Integration/Continuous Delivery
CN	Core Network
CP	Cyclic Prefix
CQI	Channel Quality Indicator
CSI	Channel State Information
CU	Central Unit
DL	Downlink
DU	Distributed Unit
E2AP	E2 Application Protocol of the Open RAN E2 interface
eMBB	Enhanced Mobile Broadband Communications
FFT	Fast Fourier Transform
gNB	Next Generation Node Base
IaaS	Infrastructure as a Service
IoT	Internet of Things
KPI	Key Performance Indicator
KPM	Key Performance Metrics
LPWAN	Low Power Wide Area Network
LTE Cat-M	Long Term Evolution for Machine Type Communications
MAC	Medium Access Control
mMTC	Massive Machine-Type Communications
MNO	Mobile Network Operator
NB-IoT	Narrowband Internet of Things
near-RT RIC	Near real time RIC
NFV	Network Functions Virtualization
NG-RAN	Next Generation Radio Access Networks

non-RT RIC	Non real time RIC
Open RAN	Open Radio Access Network
PDCP	Packet Data Convergence Protocol
PDU	Packet Data Unit
PRB	Physical Resource Block
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RIC	RAN Intelligent Controller
RLC	Radio Link Control
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RU	Radio Unit
SDAP	Service Data Application Protocol
SDN	Software Defined Networking
SIC	Successive Interference Cancellation
SINR	Signal-to-Noise Ratio plus Interference
SLA	Service Level Agreement
SMO	Open RAN Service Management and Orchestration
SNR	Signal-to-Noise Ratio
UE	User Equipment
UL	Uplink
URLLC	Ultra Reliable Low-Latency Communications
VNF	Virtualized Network Function

LIST OF SYMBOLS

f_B	Frequency channel where the eMBB device transmit
f_M	Frequency channel where the set of mMTC devices transmit
λ_M	Number of connected mMTC devices
H_B	eMBB channel coefficients
Γ_B	Average SNR of eMBB users
H_m	mMTC channel coefficients
Γ_M	Average SNR of mMTC users
A_M	Number of all active mMTC devices
$\mathbf{G}_m$	Vector containing all the instantaneous SNRs of the mMTC devices seen at the BS
G_B	Instantaneous channel gain of eMBB user
P_B	Average power
ε_B	eMBB reliability requirement
G_B^{min}	Threshold SNR for the eMBB user transmission
G_B^{tar}	Target SNR for the eMBB user transmission
r_B^{orth}	eMBB transmission rate for the orthogonal coexistence
r_M	mMTC transmission rate
ε_M	mMTC reliability requirement
$\sigma_{[m_0]}^{orth}$	SINR of the m_0-th mMTC device seen at the BS for the orthogonal coexistence
D_M	Number of mMTC devices in outage in H-OMA
α	Fraction of time allocated for the eMBB device use of the channel in the H-OMA
$\sigma_{[m_0]}^{non-orth}$	SINR of the m_0-th mMTC device seen at the BS for the non-orthogonal coexistence
D_B	Number of eMBB devices in outage in H-NOMA
β	mMTC power allocation vector

CONTENTS

1 INTRODUCTION 14
1.1 GOALS AND CONTRIBUTIONS 15
1.1.1 Main Contributions 16
2 NETWORK SLICING ON THE NETWORK LAYER DOMAIN . . 17
2.1 RAN SLICING 19
2.2 OPEN RAN 19
2.2.1 RAN Functional Splits 20
2.2.2 Use case: RAN Slice SLA Assurance 22
2.3 END-TO-END SLICING 26
3 NETWORK SLICING ON THE PHYSICAL LAYER DOMAIN . . 28
3.1 SYSTEM MODEL 30
3.1.1 eMBB and mMTC Orthogonal Coexistence 31
3.1.2 eMBB and mMTC Non-Orthogonal Coexistence 33
3.2 PROPOSED METHOD 34
3.2.1 Simulation Results for H-OMA 38
3.2.2 Simulation Results for H-NOMA 41
3.2.3 H-OMA and H-NOMA Results Comparison 44
4 CONCLUSION AND FUTURE WORKS 45
REFERENCES 47

1 INTRODUCTION

Since the genesis of wireless mobile communications, all the so-called generations have served a specific purpose and were developed based on particular concerns. The first and the second generation – 1G and 2G – were designed to supply voice services as a primary goal, starting from analogical communications and reaching improvements by using digital techniques, respectively. The third generation (3G) introduced the massive broadband concept as a result of the popularization of smartphones around the world. At that time, the main goal was to provide high throughput data rates to the end user instead of just voice calls and text messaging. Next, the fourth generation (4G), in addition to improving the broadband connections has also introduced the low power wide area networks (LPWAN) concept. The introduction of long term evolution for machine-type communications (LTE Cat-M) and narrowband internet of things (NB-IoT) wireless technologies made it possible the communication between machines, and not only people (ASGHAR; MEMON; HÄMÄLÄINEN, 2022).

Looking at the evolution pattern of wireless mobile communications, one can see that there is an odd generation that introduced disruptive technologies and an even generation that improved the systems that were conceived in the past generation. So, while the main concern of all generations until 4G was to connect people (ZHANG, Z. *et al.*, 2019), the 5G brings a new disruptive concept to connect people and machines based on three pillars: mMTC (Massive Machine-Type Communications), eMBB (Enhanced Mobile Broadband Communications) and URLLC (Ultra-Reliable Low-Latency Communications) (SHAFI *et al.*, 2017).

Thus, due to the wide range of use cases foreseen in 5G, the traditional one-size-fits-all network solution is not viable any longer owing to the heterogeneity of the services. Therefore, one of the key enablers of 5G is network slicing: an approach to slice a network – on the physical and/or network layers – into several virtual layers according to the service that is being provided, allowing different customizations to satisfy different (and sometimes conflicting) services requirements (ZHANG, S., 2019).

The SDNs (Software Defined Networks), together with NFVs (Network Function Virtualization) and VNFs (Virtualized Network Functions), play an important role in this scenario as they allow high programmability and modularization of the network (ZHANG, S., 2019). In this line, the Open RAN movement uses the aforementioned concepts to build an open radio access network that encompasses intelligent controllers capable of housing machine learning models to automate and optimize the whole network (ALBERTI *et al.*, 2022).

On the other hand, in the physical layer domain, the seminal work of (POPOVSKI *et al.*, 2018a) brings theoretical foundations on how devices belonging to different services can coexist in the same radio resource. Such study is carried out following two resource-sharing

schemes: orthogonal and non-orthogonal. Through the concept of reliability diversity, even with the possibility of interference between devices, it is possible to gain new insights into techniques that can be used in 5G and beyond systems in conjunction with network slicing at the network layer domain, and improve the overall system performance.

Recent works consider the use of network slicing in 5G and beyond (B5G) networks. An overview of intra and interslice resource allocation methods inside a network slicing context is provided in (DEBBABI *et al.*, 2022). Some approaches discuss the use of machine learning techniques in slices resource management (WU *et al.*, 2022) and to dynamically allocate, schedule and orchestrate resources (SÁNCHEZ; CASILIMAS; RENDON, 2022). Specifically in the physical layer, (POPOVSKI *et al.*, 2018a) provides the theoretical background regarding the slicing between the three aforementioned 5G services and introduces the concept of heterogeneous orthogonal multiple access (H-OMA) and heterogeneous non-orthogonal multiple access (H-NOMA), where the term *heterogeneous* comes from the different classes of services. Later, (SANTOS *et al.*, 2020) elaborated on the model from (POPOVSKI *et al.*, 2018a) by considering the slicing between eMBB and URLLC services using the max-matching diversity (MMD) algorithm (BAI *et al.*, 2010) to allocate channels to the eMBB devices. The use of the MMD algorithm introduces frequency diversity, simultaneously increasing the eMBB achievable rate and the URLLC reliability. In the context of eMBB and mMTC slicing, (TOMINAGA *et al.*, 2021) considers the use of multiple receiving antennas in the uplink. The results show that the space diversity provided by the multiple antennas are more beneficial to H-NOMA than to H-OMA, and that such difference increases with the number of antennas.

1.1 GOALS AND CONTRIBUTIONS

The objective of this work is twofold: 1) to carry out a study on network slicing and its necessary implementation requirements from the point of view of the upper layers of communication and network functions; 2) to propose a method based on the adjustment of mMTC transmission powers to leverage SIC (Successive Interference Cancellation) performance and improve the number of active mMTC devices sharing resources with eMBB devices at the same physical layer infrastructure.

For the network layer approach, the importance of the concepts of virtualization and programmability of a wireless network to support the enabling of the network slicing is discussed, showing how future architectures should behave to cover the new use cases proposed by research bodies. As for the physical layer approach, the concepts and modeling of the work system proposed by (POPOVSKI *et al.*, 2018a) were used as a basis for this dissertation. Specifically, a method was proposed to modify the received powers of mMTC devices seen by the BS (Base

Station) when competing for transmission resources with the eMBB service. This proposed method shows that an adaptation of the transmission power of the mMTC devices can induce a spread of the average SNR (Signal-to-Noise Ratio) seen by the BS, generating the so-called clusters and thus impacting the number of mMTC devices capable of coexisting with eMBB devices because of the SIC procedure that leverages the differences of received powers at the BS.

1.1.1 Main Contributions

- A study on network slicing at the network layer, showing in practical terms how the implementation can be done through the Open RAN architecture. A use case involving the RAN (Radio Access Network) SLA (Service Level Agreement) assurance for different network slices is shown, addressing the details of its operation.
- Elaboration of a new method where the received powers of the mMTC devices seen by the BS are changed through a clustering procedure based on the modification of the transmission power of each mMTC device of the active set. Considering a particular scenario, this new method can bring gains of approximately 60% for H-OMA considering low eMBB traffic, and an average of 40% for H-NOMA considering low eMBB traffic (up to 1.5 bits/s/Hz) when compared to the performances without this method.
- A bird's-eye view of network slicing is presented, allowing a global understanding of the factors that can impact its end-to-end implementation.

2 NETWORK SLICING ON THE NETWORK LAYER DOMAIN

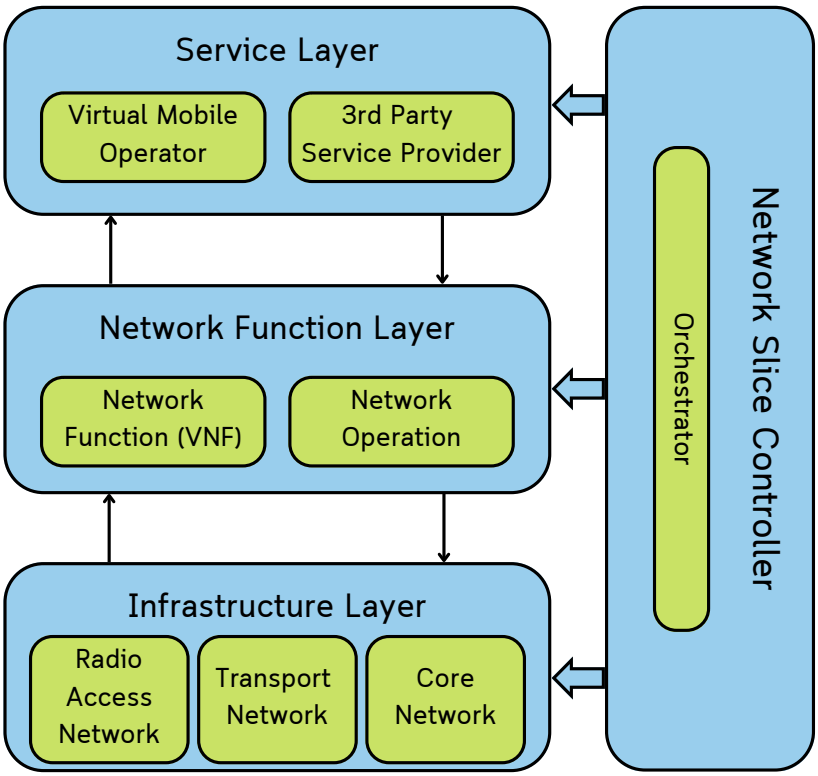
Future wireless mobile networks – starting from the 5G – are expected to be scalable, flexible, agile, and fully programmable, as a manner to support the coexistence of different services with different QoS (Quality of Service) requirements (ALBERTI *et al.*, 2022). Through an approach based on the softwarization of the different levels of the network, the features cited above can be fulfilled using software programming to implement, deploy and maintain the networks, providing end-to-end service management through closed-loop processes enabled by network intelligent controllers (BARAKABITZE *et al.*, 2020).

The Next Generation 5G Mobile Networks (NGMN) introduced network slicing as a technique to build separated logical networks (namely slices) with different operating purposes but using the same infrastructure. On the lower layers, the slices share the same radio resources to attend to different QoS requirements, thus it is of extreme importance the openness of the interfaces between the functions of the control and user planes (NGMN, 2015). As traditional RANs (Radio Access Networks) are closed and delivered by a unique vendor in a black-box way, the Open RAN proposes an aperture of the interfaces to enable multi-vendor supplies and intelligent networks with the RIC (RAN Intelligent Controller). These features are based on an architecture that pretends to be fully programmable, virtualized, and disaggregated (POLESE *et al.*, 2023).

As a crucial characteristic of network slicing, the virtualization process is designed and operated by means of NFV, SDN, and VNF. In SDN, the control and data plane of the network are separated. The control plane, managed by software, determines how traffic is handled while the data plane consists of the data flowing between nodes. This separation leads to centralized management, a comprehensive network overview, increased reliability, and flexibility, and the ability to make changes to policies and routing quickly (VMWARE, 2019). This way, NFV is defined as the infrastructure platform that guides the VNF. Finally, a VNF is nothing more than a network function that runs software that performs a specific task, such as a load balancer, a router, a firewall, etc.

There is a very close relationship between network slicing and NFVs because it is through the NFVs architecture that the slicing can be enabled. According to (FOUKAS *et al.*, 2017), a generic 5G framework for network slicing is composed of three layers: a service layer, a network layer, and an infrastructure layer, all managed by a network slice controller, as illustrated in Figure 1. Starting from the infrastructure layer, there is a need to move towards the IaaS (Infrastructure as a Service) paradigm, in which CN (Core Network) and RAN are present. Several studies have been made about where and how the infrastructure should be located – such as a centralized cloud or an edge cloud, but the fact is that these

Figure 1 – 5G network slicing framework containing the service layer, network function layer, infrastructure layer, and the network slice controller.



Source: Author, adapted from (FOUKAS *et al.*, 2017).

infrastructure resources must be available in a virtualized manner. Specifically talking about the virtualization of the RAN part, one can cite the Open RAN architecture, which brings the concept of programmable networks. Following on the framework, the network function layer encompasses all the operations associated with the configuration and lifecycle of the VNFs that run over the virtual infrastructure (i.e. NFVs), having as enabling technology the SDN. The network function layer supplies the network characteristics needed by the service layer for normal operation. Finally, the service layer is where the business rules and the specifications of a slice lifecycle come into play, for example, the SLAs, traffic specifications, etc. The services on the service layer can be offered by a mobile operator or a third party. Tied to the three layers is the network slice controller, which translates service models and use cases into end-to-end network slices through the connection of network functions, associating them to infrastructure resources, and configuring and monitoring each slice during its lifecycle.

2.1 RAN SLICING

As mentioned, the overall concept of network slicing is to obtain an end-to-end slice regarding a specific service type, respecting a SLA. The NG-RAN (Next Generation Radio Access Networks) are expected to support network slicing under some principles and requirements that span different network functions, procedures, and the management of radio resources. For example, a slice that provides V2X services has to perform handover procedures, mobility management, and also resource allocation on the physical layer through different configurations of MAC (Medium Access Control) scheduling and RRM (Radio Resource Management) procedures. On the other hand, an IoT (Internet of Things) slice may not need mobility and handover procedures at all, meaning that network functions present in one slice may not appear in others.

According to (3GPP, 2022b), one can summarize the key features needed to enable NG-RAN support for network slicing:

- NG-RAN awareness of slices means that it is necessary to ensure isolation, availability, and suitable allocation of resources through radio resource management;
- NG-RAN needs to support different slices with different behaviors;
- As each slice is usually tied to a SLA, one important aspect is cross-slice resource management as there is a need to support cross-slice policy enforcement per SLA to ensure this for each individual slice. In other words, NG-RAN should be able to decide when and how the radio resources can be allocated through different slices to maintain the defined SLAs.

The implementation of the slicing in the RAN domain is very challenging, because it involves the sharing of radio resources in different forms for each slice while keeping them isolated and attending to the different configured SLAs. In fact, the challenge lies in the need to find resource allocation schemes that ensure fairness in the resource allocation between the slices, QoS within a specific slice, and overall resource efficiency. For example, questions like "Will it be needed to reserve resources? how much?" arise, and further study is necessary; in this case, due to the random nature of traffic, the reservation process can be solved by a probabilistic approach. Anyway, the scheduling method used to allocate the resources also needs to be studied and analyzed inside the context of each slice (ELAYOUBI *et al.*, 2019).

2.2 OPEN RAN

Nowadays, when a MNO (Mobile Network Operator) designs a new cellular network, they need to buy software and hardware from vendors in a complete solution: this is what we

call a closed RAN, and the MNO does not have the power to make personal changes to your own network, because all functions and applications are delivered in "black box" mode. This approach has been used and consolidated over the years as a traditional RAN approach, but its use results in a network that prevents on-demand self-configuration, fine-tuning, and possible optimizations of the RAN components, leading to vendor lock-in that do not allow the MNO to customize the network according to its particular business models (POLESE *et al.*, 2023).

To overcome this limitation, a new access network architecture paradigm emerged based on virtualization, disaggregation, and programmability of RAN functions – thus allowing the use of modular, containerized software development, in addition to enabling continuous integrations through CI/CD aiming to increase the quality of the developed software. This network transformation for Open RAN also brings the opening of interfaces, which, in an open and virtualized architecture that uses generic hardware and software – the so-called "commercial off-the-shelf" – will facilitate interoperability between combinations including open distributed units (O-DU) and open radio units (O-RU) from different vendors. The process of standardization, coupled with the decoupling of software and hardware, will promote fair competition and drive innovation. It will enable MNOs to implement diverse strategies for optimizing their operations in a cost-effective and efficient manner (GAVRILOVSKA; RAKOVIC; DENKOVSKI, 2020).

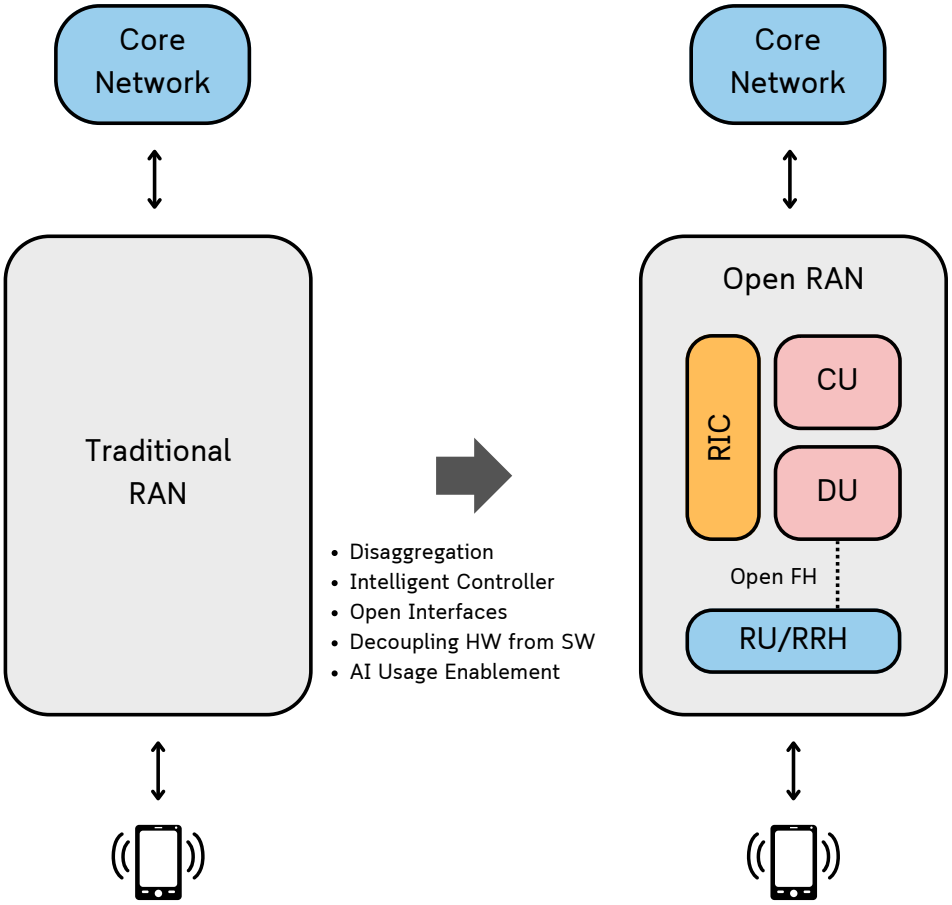
As shown in Figure 2, it is possible to summarize this transformation from the black box architecture towards a modular, virtualized, programmable, and open architecture. Also, as a new feature of Open RAN, the Ran Intelligent Controller (RIC) appears as a key enabler of the use of artificial intelligence and machine learning (AI/ML). RIC is divided into near-real-time RIC (or near-RT RIC) and non-real-time RIC (or, non-RT RIC); the near-RT RIC contains the xApps – applications that operate in control loops with a periodicity between 10 ms and 1 s, while the non-RT RIC contains the rApps – applications that operate in control loops with periodicity greater than 1 s.

2.2.1 RAN Functional Splits

In the 5G RAN architecture, besides the existence of the RU (Radio Unit), the BBU (Baseband Unit) is divided into a DU (Distributed Unit) responsible for physical and MAC layer tasks and a CU (Central Unit) responsible for network and packet forwarding tasks (3GPP, 2022a). Due to the fact that the network functions belong to different layers, the assignment between network functions of the architecture must also be optimized.

3GPP (Third Generation Partnership Project) defined several options to accomplish this separation of network functions between the DU and the RU and the one chosen by the

Figure 2 – Evolution of the Radio Access Network.



Source: Author.

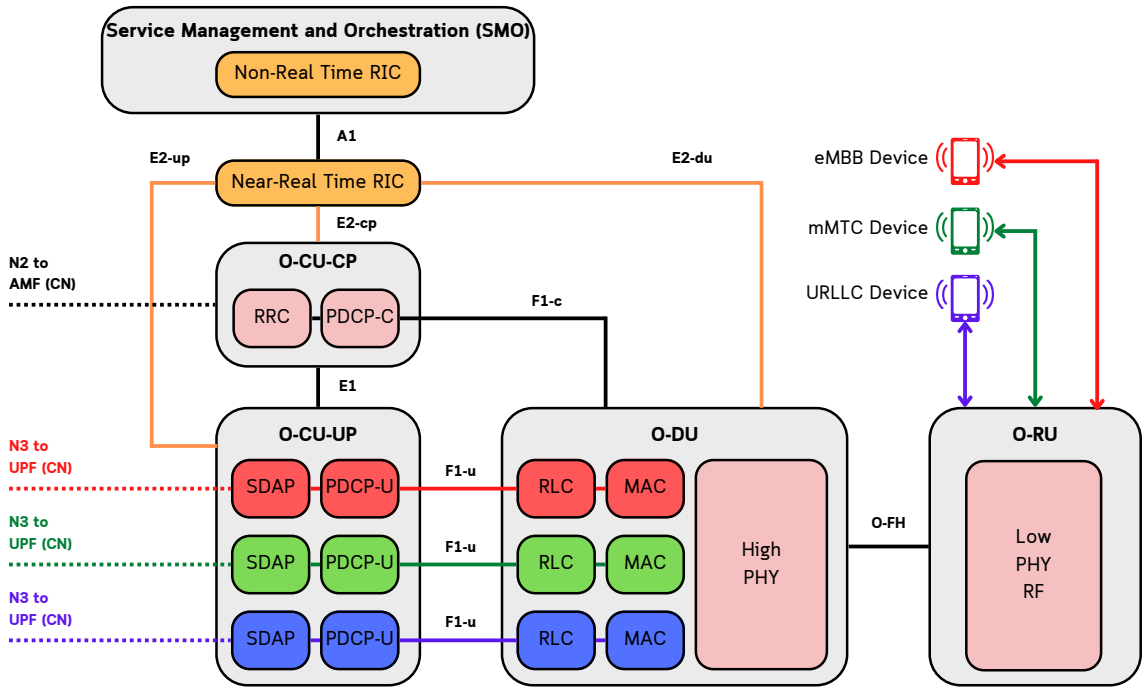
O-RAN Alliance for Open RAN was the so-called 7.2x split, which provides a balance between latency and throughput across layers (POLESE *et al.*, 2023). As shown in Figure 2, we can see the existence of the RAN units cited above.

In split 7.2x, the RU performs RF signal handling and operations that belong to the lower physical layer such as precoding, beamforming, FFT (Fast Fourier Transform), a CP (Cyclic Prefix) addition/removal, and AD/DA (Analog-to-Digital/Digital-to-Analog) signal conversions. The DU takes care of the lasting higher physical layer functions, such as modulation, and also of the RLC (Radio Link Control) and MAC layers. Finally, the CU holds the higher layers of the stack, like PDCP (Packet Data Convergence Protocol) – which performs tasks such as reordering of packets, compression, and decompression of headers, data integrity verification, etc., SDAP (Service Data Application Protocol) – that manages the QoS of the data radio bearers of specific PDU (Packet Data Unit) sessions, and RRC (Radio Resource Control), which connects and configures all the radio resources. The way functions are divided in this

functional split can greatly affect the performance of network slicing, and the best division depends on the specific characteristics of the targeted service. For instance, to meet low latency requirements, a URLLC slice may require that most RAN functions run on the DU. Conversely, greater centralization in an eMBB slice can increase throughput by aggregating RRHs (Remote Radio Heads).

When considering network slicing, certain RAN functions can be shared among different slices. For example, Figure 3, it is shown the relevant components of an example of network slicing inside an Open RAN architecture. It can be seen that the entire physical layer is shared between the three types of slices but functions like PDCP and RLC are specific according to each slice. To bring an example, one could think about a low latency requirement slice: the RLC function in this slice would not use header compression, and the function should be tailored to run in a transparent mode – which is a different configuration when compared to a broadband service, that has QoE/QoS requirements to meet.

Figure 3 – Example of an Open RAN architecture when considering Network Slicing. The sharing of some network functions depends on the service type and on its requirements.



Source: Author, adapted from (O-RAN ALLIANCE, 2022b).

2.2.2 Use case: RAN Slice SLA Assurance

To understand how Open RAN can leverage the concepts of disaggregation and virtualization of the network functions inside the context of network slicing, we address an use case

related to the allocation of resources to meet the SLA requirements for each slice based on the O-RAN Alliance Working Group 2 specification (O-RAN ALLIANCE, 2022a).

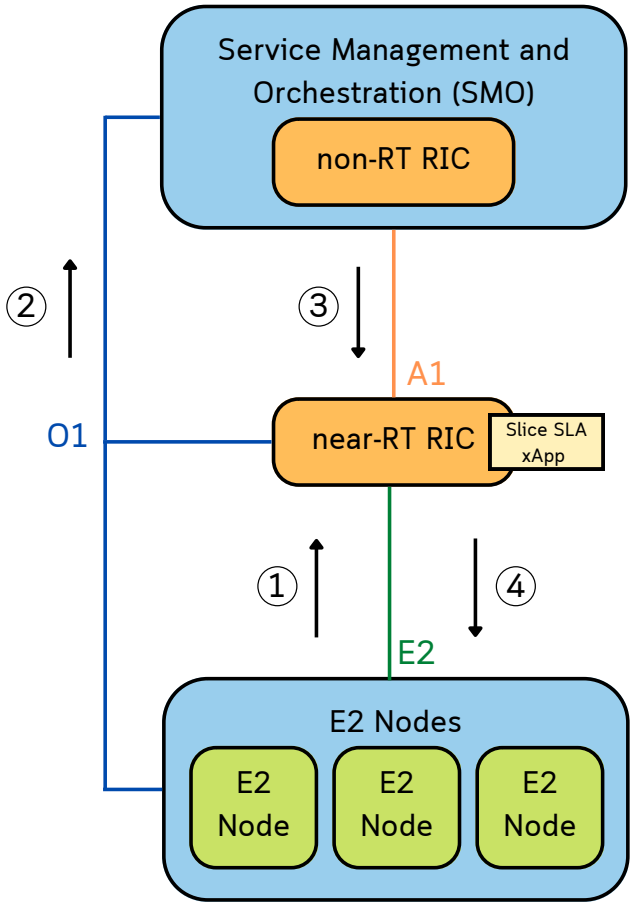
The aim of this use case is to dynamically allocate the radio resources and control the slice configurations based on the current performance of the slice, ensuring that the SLA will not trespass. This can be done through the Open RAN components and open interfaces shown in Figure 3; the near/non-RT RIC role is to monitor the SLA by capturing the KPM data provided by the E2 nodes (i.e. a gNB) so that trained machine learning models can send control messages via A1 policies through the E2 interface to adapt the resource allocation performed by the O-DU and O-CU network functions.

For this use case, the role of the entities of the Open RAN architecture are:

- **SMO and Non-RT RIC:**
  - get the SLA slice target to be achieved by accessing the NSSMF (Network Slice Management Function), in which the slice informations are stored;
  - realize the long-term monitoring of performance metrics of the E2 nodes;
  - train the ML models that will be deployed both on non-RT RIC (slow loop optimizations) and on near-RT RIC (fast loop optimizations);
  - create, update and send A1 policies to the near-RT RIC
- **Near-RT RIC:**
  - performs near real-time monitoring of performance metrics of a specific RAN slice;
  - support deployment and execution of ML models sent by the non-RT RIC;
  - performs RAN actions through the execution of xApps by sending E2AP protocol PDUs with control procedures directed to the E2 nodes.
- **E2 Nodes:**
  - support E2 reports of performance metrics both through O1 and E2 interfaces;
  - support the execution of the E2AP PDU control messages sent by the near-RT RIC to realize the proper resource allocation.

All of the entities cited above can be seen in Figure 4, within a brief description of the information flow of the control loop. The first step of the SLA assurance is to collect all the information of the E2 nodes with the purpose to feed the future procedures that will have to make decisions on, for example, what will be the action delivered by the RIC to ensure the slice SLA. The statistics collected here can be per UE and/or per slice, such

Figure 4 – Information flow for the RAN SLA assurance use case. The enumerated procedures are: 1) collection of E2 performance; 2) enriched information based on operating KPIs, 3) policies for dynamic SLA assurance, and 4) E2 control messages for resource allocation.



Source: Author, adapted from (O-RAN ALLIANCE, 2022b).

as CQI measurements, average DL/UL (Downlink/Uplink) throughput, statistics about RRC connections and disconnections, and the number of successful or failed PDU sessions, average PDCP PDU data volume, packet drop rate, etc. Other important aspects that have to be collected are the informations about the gNB capabilities on its CU and DU functions, so the E2AP control message can be properly built and sent back from the xApp present on the near-RT RIC to the E2 node.

The second and the third steps are tied to the SMO and near-RT RIC, which will receive the fine-grained information collected by the E2 nodes and group and count into KPIs to send to the non-RT RIC through the O1 interface. When such information arrives at the SMO and at the non-RT RIC, they are stored in persistent databases allowing the long-term

monitoring of the related performance metrics and KPIs. This brings intelligence to the RIC by enabling the training and deployment of ML models, together with the delivery of A1 policies. An A1 policy is a feature that supports non-real-time radio resource management and higher layer procedure optimization while providing guidance, parameters, policies, and AI/ML models to support the operation of near-RT RIC functions in the RAN to achieve higher-level non-real-time objectives.

Finally, the fourth step is the delivery of the action that will guarantee the SLA assurance for a determined slice based on measurements collected on the same E2 nodes that started the entire procedure; this action can be performed both through the E2 or O1 interfaces. Specifically, the way the near-RT RIC xApp uses to send control messages to the E2 nodes is by the E2AP protocol. These messages are assembled in an XML (Extensible Markup Language) type and encoded in an ASN.1 scheme and contain all the required information so that the E2 node can realize the action established. The O-RAN Alliance Working Group 3 defined a set of E2 procedures that can carry different action contents, such as reports, inserts, policies deployment, etc. (O-RAN ALLIANCE, 2022c).

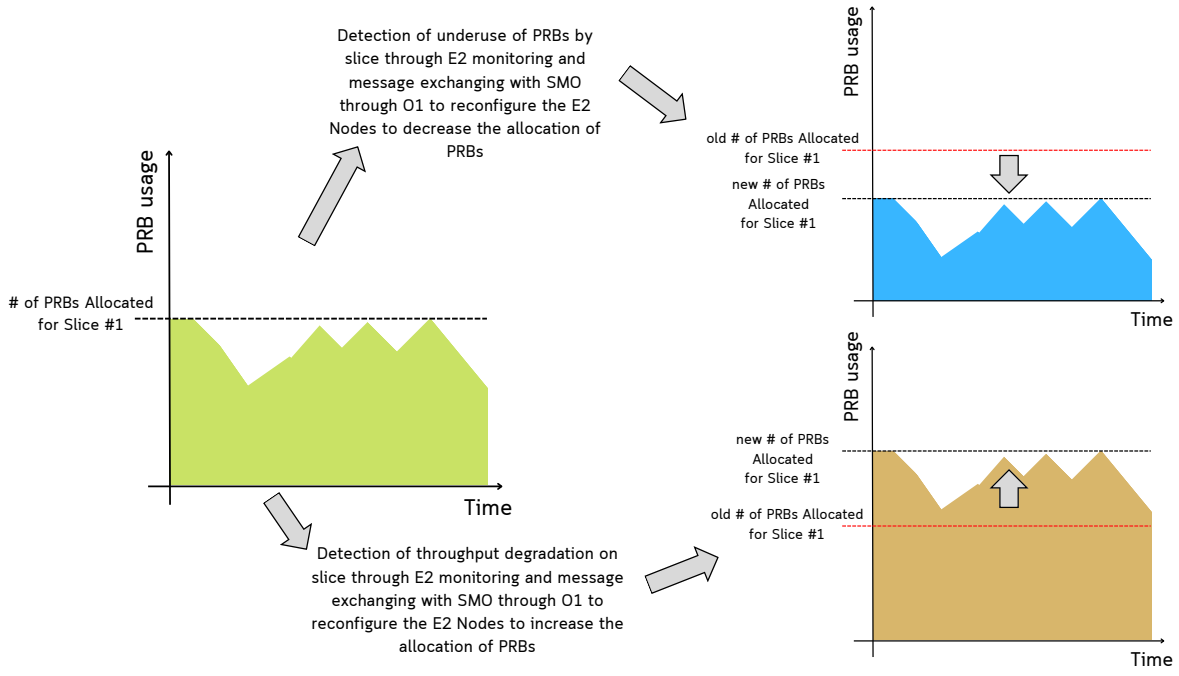
The code snippet presented below is an example of A1 policy transmitted from the non-RT RIC to the near-RT RIC to assure the SLA to slice #1 – in this case, UL and DL throughput values described in a data rate metric *e.g.* Mbps, where the slice is deployed over a geographic area covered by cells #1, #2 and #3. After receiving the policy, the near-RT RIC enforces it and provides guidance to the RAN behavior over the E2 interface to meet the imposed SLA.

```
{
    "PolicyId": "1",
    "scope": {
        "sliceId":  "1",
        "cellId":   "1", "2", "3",
},
"statement": {
    "uLThptPerSlice": "30"
    "dLThptPerSLice": "30"
    }
}
```

For example, to perform the RAN actions over the O1 interface according to this policy, it can be leveraged the direct connection between the E2 nodes and the SMO entity, shown in Figure 4. As described before, the flow statistics and KPIs of the connections are sent to

the SMO where it is performed the PRB (Physical Resource Blocks) evaluations, as a way to reconfigure the E2 nodes to meet the proposed throughput. If the monitoring process detects an underuse of PRBs by the E2 nodes, the SMO/non-RT RIC sends a new reconfiguration message to the E2 nodes decreasing the number of PRBs allocated to them. Else, if the monitoring process detects degradation of throughput on the monitored E2 nodes, the SMO sends a reconfiguration message increasing the number of PRBs allocated to these E2 nodes. This whole process can be seen in Figure 5.

Figure 5 – Illustration of O1 RAN action control loop for the described RAN SLA assurance use case between slices.



Source: Author, adapted from (O-RAN ALLIANCE, 2022b).

2.3 END-TO-END SLICING

OpenRAN appears to be a technology with important features to enable slicing and meet the requirements of next-generation RANs. Because it has a closed-loop control with access to resource allocation functions (for example, MAC scheduling) and performance monitoring, its use becomes interesting for implementing optimization algorithms in near real-time. The result obtained is composed of several logical networks that seek to meet specific requirements for different services.

On the other hand, it is also possible to implement network slicing in the physical layer, through methods and algorithms that run directly on physical resources, such as time

slots, PRBs, waveform numerologies, etc. Through the correct allocation of these resources to different services, the air, the physical means of wireless transmission, is subdivided into abstract slices. Finally, the whole system composed of slicing on the network and physical layer can be defined as end-to-end slicing.

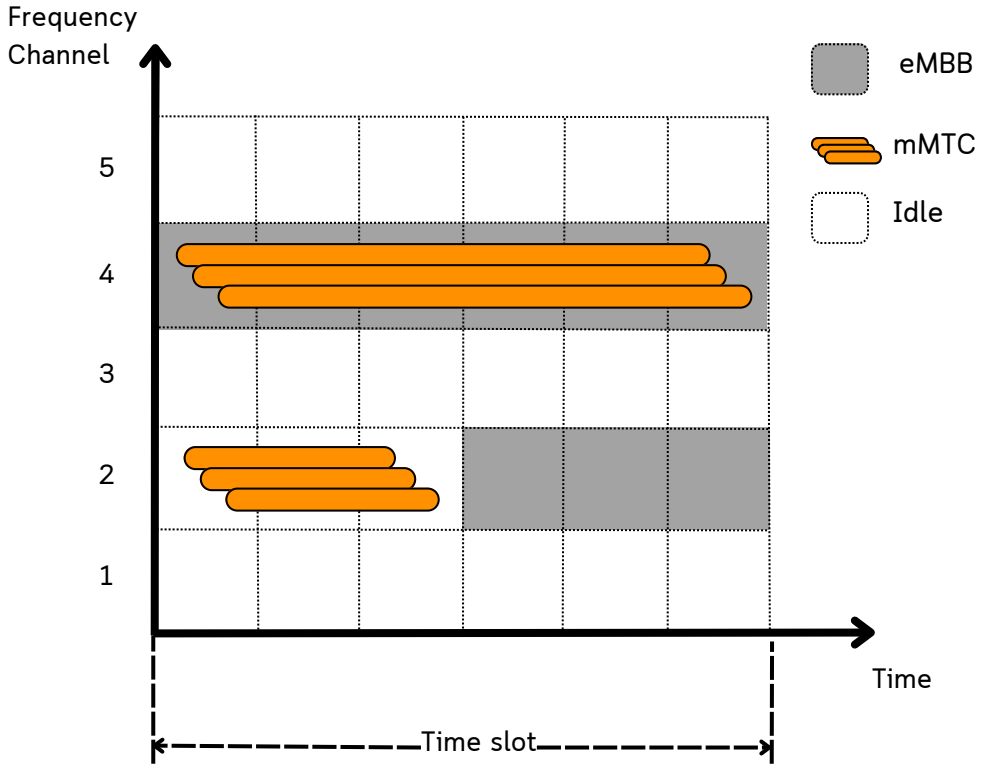
Chapter 3 will discuss network slicing over physical resources in wireless transmissions. Its objective is to give details of the physical layer slicing problem definition and its performance metrics for a specific case, besides discussing a power allocation method that can be used for mMTC devices in a slicing context between eMBB and mMTC services.

3 NETWORK SLICING ON THE PHYSICAL LAYER DOMAIN

Consider the following scenario: a set of mMTC, URLLC, and eMBB devices aiming at performing uplink data transmission to a common BS. Because of the characteristics of each service, the lack of coordination among devices belonging to different services makes the uplink access procedure a complex task. For example, mMTC devices are numerous and transmit intermittently at low transmission rates. The large amount of devices that can be active simultaneously introduces a random characteristic to the problem of resource allocation for this service. The ultimate goal is to maximize the mMTC traffic arrival rate, defined as the number of supported devices (MEHMETI; PORTA, 2022). The URLLC service also has an intermittent characteristic, but with a smaller set of devices able to transmit. Hence, the resources granted to URLLC devices are time restricted and spanned over more than one frequency channel to achieve diversity (ZHANG, H. *et al.*, 2017), due to the rigorous latency and reliability constraints required by the service (POPOVSKI *et al.*, 2018b). Finally, the eMBB service seeks to maximize its throughput and, within the context of slicing, its performance gains are directly related to the number of radio resources and channel conditions available for its transmission and the amount of interference caused by other services.

In (POPOVSKI *et al.*, 2018a), the concept of heterogeneity for the sharing of resources between devices from different services is introduced, and the idea of reliability diversity for network slicing is proposed. The idea behind the heterogeneity relies on the multiple access between devices that belongs to different services, and not between devices with the same traffic pattern, error requirements, and reliabilities. This way, H-OMA stands for the orthogonal multiple access between the different 5G services, while H-NOMA stands for the non-orthogonal multiple access. Figure 6 shows how the patterns of resource usage by services are arranged in a time-frequency grid for both H-NOMA and H-OMA cases. Notice that for channel $f = 2$ the division between the mMTC and eMBB services is done orthogonally in a temporal multiplexing scheme, where the use of the channel for each service is assigned in 50% of the time slot. It is important to mention that orthogonality occurs between services, and the mMTC devices transmit in a non-orthogonal way in the same service. The opposite case, H-NOMA, is depicted on the channel $f = 4$ where one can see the overlap between devices from different services in the use of resources. For the slicing between mMTC and eMBB, it is expected that the intermittent traffic characteristic of mMTC devices can be used to increase the performance of the eMBB service, even if in this configuration there is interference between the services.

Figure 6 – Time-frequency grid of the aforementioned H-OMA and H-NOMA allocation methods, with $S = 6$ mini-slots and $F = 5$ resources. In H-OMA, eMBB and mMTC services have dedicated resources, and the word "heterogeneous" comes from the fact that the orthogonal allocation is made between devices of different services, and not of the same service. For $f = 2$, the H-OMA is shown with 50% of the time allocated to each service. On the other hand, for $f = 4$, both services access the resources at the same time, characterizing the H-NOMA.



Source: Author, adapted from (POPOVSKI *et al.*, 2018a).

3.1 SYSTEM MODEL

It is considered a single-cell uplink scenario with eMBB and mMTC devices transmitting to a common BS. As mentioned in the related works section, most of the studies give great importance to alternative techniques for the slicing between URLLC and eMBB. However, the main idea of this work is to investigate how the nature of the mMTC traffic can be leveraged to improve eMBB transmission rates.

For both mMTC and eMBB services, transmission occurs at the same specific frequency $f_B = f_M$, with the mMTC service having a number of active mMTC devices equal to A_M which follows a Poisson distribution with mean λ_M (i.e. the arrival rate). The disposal of the services through the channels can be seen in Figure 6, for the orthogonal and non-orthogonal cases.

Also, at each transmission, the channel is considered constant in the time comprised in one time slot, with independent and Rayleigh-distributed channel coefficients (POPOVSKI *et al.*, 2018a). The eMBB channel coefficient is denoted by $H_B \in \mathcal{CN} \sim (0, \Gamma_B)$, with Γ_B being the average channel gains resulting from path loss and transmission power from the user. Likewise, the mMTC channel coefficients are denoted by $H_m \in \mathcal{CN} \sim (0, \Gamma_M)$ for $m \in \{1, 2, ..., A_M\}$.

Two scenarios are elaborated for the study of network slicing: orthogonal and non-orthogonal coexistence between eMBB and mMTC services, seeking to optimize both the number of active mMTC devices and the eMBB data transmission rate by a maximum value. In both orthogonal and non-orthogonal cases, and as described in the paragraph above, mMTC devices use the same frequency resource f_M to transmit. As their channel coefficients are independent and follow a Rayleigh distribution, BS sees the SNRs of these devices floating around an average Γ_M. The vector containing all the instantaneous SNRs of the mMTC devices trying to connect to the BS is denoted in this work as $\mathbf{G}_m = \left[G_{[0]}, G_{[1]}, \cdots, G_{[A_M]}\right]$.

In our system, the SIC procedure is employed to perform the decoding and subtraction of the strongest component of the received signal for the mMTC traffic (LIEN *et al.*, 2017). It is important to emphasize here that the SIC procedure for mMTC decoding is enabled by the fading that occurs over the channel, allowing the instantaneous SNR of the devices to vary around the average Γ_M.

Moreover, in this work, a method is proposed to make the BS see more than one average SNR in the uplink procedure for the A_M active mMTC devices, aiming to maximize the SIC performance and enable the decoding of more devices.

3.1.1 eMBB and mMTC Orthogonal Coexistence

In (POPOVSKI *et al.*, 2018a) H-OMA is introduced as a manner to perform the network slicing in the physical layer between the services. Considering the eMMB and the mMTC services, in the H-OMA uplink transmission they coexist on a time-sharing scheme, where when one service is transmitting the other remains silent. The channel conditions (*i.e.,* the CSI) of the eMBB and the mMTC devices are denoted by, respectively, G_B and G_M. However, because of the physical characteristics of the mMTC devices – such as the concern of embedded implementations that keep a low battery consumption – and also of their traffic pattern and relatively low-reliability target, it is not required or practical for this kind of service to exploit CSI information prior to transmission. On the other hand, for the eMBB service, the transmission is based on the following assumptions:

- each transmission occurs at a radio resource f_B to the eMBB user;
- the eMBB user is aware of the CSI information G_B, which is used to select its transmission power $P_B(G_B)$.

The main objective is to maximize the eMBB rate r_B, which depends on the outage probability requirement ε_B of the service and is subject to a long-term average power constraint. This forms the following optimization problem (POPOVSKI *et al.*, 2018a):

$$\begin{aligned} \max \quad & r_B \\ \text{s.t.} \quad & \mathbb{P}\left[\log_2\left(1 + G_B P_B(G_B) < r_B\right)\right] \leqslant \varepsilon_B \\ & \mathbb{E}\left[P_B(G_B)\right] = 1 \end{aligned} \tag{1}$$

As mentioned in (POPOVSKI *et al.*, 2018a), the solution to (1) is the truncated power inversion, where the eMBB user compensates the path loss using the inverse of the CSI as the transmit power. However, this depends on a threshold G_B^{min} that guarantees minimum conditions necessary for transmission and depends only on the eMBB reliability condition ε_B.

$$G_B^{min} = \Gamma_B \ \ln\left(\frac{1}{1-\varepsilon_B}\right). \tag{2}$$

The user decision to transmit or not will be based on this threshold value – if the SNR is higher than the threshold the transmission occurs, otherwise not.

After the power-inversion scheme, the target SNR that will be used to evaluate the transmission power is given by

$$G_B^{tar} = \frac{\Gamma_B}{\gamma\left(0, \frac{G_B^{min}}{\Gamma_B}\right)}, \tag{3}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function. This implies that the eMBB rate can be obtained from (3) as

$$r_B^{orth} = \log_2\left(1 + G_B^{tar}\right) \text{ [bps/Hz]}. \tag{4}$$

The mMTC transmissions have the characteristic to be sporadic, thus, random and unknown. Assuming a fixed transmission rate r_M and a reliability constraint ε_M, our goal is to maximize the mMTC arrival rate λ_M which a frequency resource can support. We also assume the use of a SIC decoder on the BS side, which makes it possible to leverage the power imbalance and improve the overall decoding.

In this scenario, each mMTC device transition will arrive with a different instantaneous SNR (POPOVSKI *et al.*, 2018a), being possible to order the devices as $G_{[1]} \geq G_{[2]} \geq ... \geq G_{[A_M]}$. In the absence of eMBB interference, the decoding of an mMTC with index m_0 depends only on its instantaneous channel gain and on the channel gains of the remaining mMTC devices. To be correctly decoded, the following inequality must be satisfied:

$$\log_2\left(1 + \sigma_{[m_0]}^{orth}\right) \geq r_M, \tag{5}$$

with $\sigma_{[m_0]}^{orth}$ being the SINR of the m_0–th device given by

$$\sigma_{[m_0]}^{orth} = \frac{G_{[m_0]}}{1 + \sum_{m=m_0+1}^{A_M} G_{[m]}}. \tag{6}$$

Note that the SINR of the m_0–th device given by (6) considers as noise all the mMTC devices yet to be decoded. In fact, this is what SIC does: if the device is correctly decoded, its component is subtracted from the whole received signal and the procedure follows until all devices are decoded or an outage is declared. Considering D_M to be the number of mMTC devices in outage, we can evaluate the error rate of the mMTC devices as

$$\mathbb{P}(E_M) = \frac{\mathbb{E}[D_M]}{\lambda_M}, \tag{7}$$

where $\mathbb{E}[D_M]$ is the average number of users in outage and λ_M is the average number of active users. Finally, the maximum number of active users which can be supported under the reliability constraint $\mathbb{P}(E_M) = \varepsilon_M$ can be obtained using Monte Carlo numerical methods and is formulated as

$$\lambda_M^{orth}(r_M) = \max\left\{\lambda_M : \mathbb{P}(E_M) \leq \varepsilon_M\right\}. \tag{8}$$

As mentioned before, there is no interference between the two types of services using the same radio resource – *i.e.* when the eMBB device is active, the mMTC device is not transmitting. This can be modeled as a time-sharing scheme: let α be the fraction of time in

which the resources are allocated to the eMBB device and $1-\alpha$ the fraction of time allotted to the mMTC devices, with $\alpha \in [0,1]$.

With the above assumptions, the main objective of the coexistence between the two services is to maximize the pair (r_B, λ_M), composed by the eMBB rate r_B given by $r_B = \alpha r_B^{orth}$ and the mMTC arrival rate given by

$$\lambda_M = \lambda_M^{orth}\left(\frac{r_M}{1-\alpha}\right). \tag{9}$$

while meeting a given mMTC transmission rate requirement r_M and error requirement ε_M. Note that λ_M^{orth} is the evaluation of (8) as a function of $r_M(1-\alpha)^{-1}$.

3.1.2 eMBB and mMTC Non-Orthogonal Coexistence

In Non-Orthogonal Heterogeneous Multiple Access (H-NOMA), both the eMBB and the mMTC users share the same radio resource in the same transmission, then the BS must decide whether to attempt to decode the eMBB or the mMTC devices first. In (POPOVSKI *et al.*, 2018a), it is assumed that at each decoding step, the BS first tries to decode the eMBB device – if it has not yet been decoded – or the next active mMTC device, in a decreasing order of their instantaneous channel gains. This approach is guided by the idea that an mMTC device can have high channel gains, causing interference in the eMBB traffic and making it difficult to decode the latter.

For the non-orthogonal case, the eMBB rate is given by

$$r_B = \log_2\left(1 + G_B^{tar}\right), \tag{10}$$

where the G_B^{tar} is the eMBB target SNR. Unlike the orthogonal case given by (2), this target SNR does not depend only on the eMBB reliability constraint ε_B. In fact, in the non-orthogonal slicing, in order to reduce the interference it can be appropriate to transmit at a lower target SNR than given by (3). This new value can be modeled as

$$G_B^{tar} \leq \frac{\Gamma_B}{\gamma\left(0, \frac{G_B^{min}}{\Gamma_B}\right)}. \tag{11}$$

If the eMBB is inactive due to insufficient SNR, then the mMTC devices do not suffer from interference and are decoded in the order to its decreasing channel gains. But if there is an active eMBB device, the receiver starts by evaluating the SNR from the m_{0-th} device using

$$\sigma_{[m_0]}^{non-orth} = \frac{G_{[m_0]}}{1 + G_B^{tar} + \sum_{m=m_0+1}^{A_M} G_{[m]}}. \tag{12}$$

If the m_{0-th} mMTC device is correctly decoded, it is subtracted from the received signal and the procedure continues until an outage occurs or the last device is decoded. Then, the BS evaluates the eMBB instantaneous SNR as

$$\sigma_B^{non-orth} = \frac{G_B^{tar}}{1 + \sum_{m=m_0}^{A_M} G_{[m]}} \tag{13}$$

and tries to decode it; if successful, the procedure runs as in the orthogonal case – as now there is no more eMBB traffic on this time slot.

In order to maximize the achievable pair $(r_B,\ \lambda_M)$, let $D_M \in \{1, 2, ..., A_M\}$ and $D_B \in \{0, 1\}$ be the random variables denoting the number of mMTC and eMBB devices in outage, respectively. The solution can be obtained from the following optimization problem

$$\begin{aligned} \lambda_M^{non-orth}(r_B) = \max \quad & \left\{\lambda_M \geq 0 : \exists G_B^{tar} \text{ and } G_B^{min}\right\} \\ \text{s.t.} \quad & \frac{\mathbb{E}[D_M]}{\lambda_M} \leq \varepsilon_M, \\ & \mathbb{E}[D_B] \leq \varepsilon_B \end{aligned} \tag{14}$$

3.2 PROPOSED METHOD

In (POPOVSKI *et al.*, 2018a), the SIC is used to decode a mixture of signals concurrently received in the BS. In practice, the overall received signal contains several independent components, which allows the decoding of the strongest component and its cancellation in the received signal, thus enabling the decoding of the weakest signal that previously suffered interference from the strongest components. In the context of this work, an important parameter is the average SNR seen by the BS for devices that belong to each service. Here, it is said "average" because the instantaneous powers of the mMTC devices received by the BS vary around Γ_M due to fading, while the instantaneous power of the eMBB device varies around the mean Γ_B. Therefore, SIC operates around an average SNR for each service, and it is the difference in the instantaneous SNRs around this average SNR that will enable its operation.

This work proposes the production of more than one average SNR in the BS for the mMTC devices; this means that the instantaneous SNRs of part of these devices will fluctuate around an average value, while the SNRs of other devices will fluctuate around one or more different averages. This can be done without coordination between the devices and the BS. For that sake, we propose that the devices randomly define the average received SNR they want to generate in the BS among a set of predefined values. In order to yield this average SNR in the BS the devices need to set their transmission power accordingly. This procedure is performed at the time of transmission when the transmitting power of the m-th mMTC device is modified in such a way that its instantaneous gain $G_{[m-th]}$ to be seen in the BS at the

moment of reception will fluctuate around different averages. This multiplication factor comes from a power allocation vector defined as β, composed of $N-1$ real elements and with the restriction that their mean is equal to 1 so that it is not introduced any long-term amplification or attenuation of power in the devices when compared with the original case.

As stated before, the introduction of β changes the original equations proposed in (POPOVSKI *et al.*, 2018a). For the H-OMA case, note that now (6) must take into account the vector containing the instantaneous SNRs of the mMTC devices:

$$\sigma^{orth}_{[m_0]} = \frac{G^{\beta}_{[m_0]}}{1 + \sum_{m=m_0+1}^{A_M} G^{\beta}_{[m]}}. \tag{15}$$

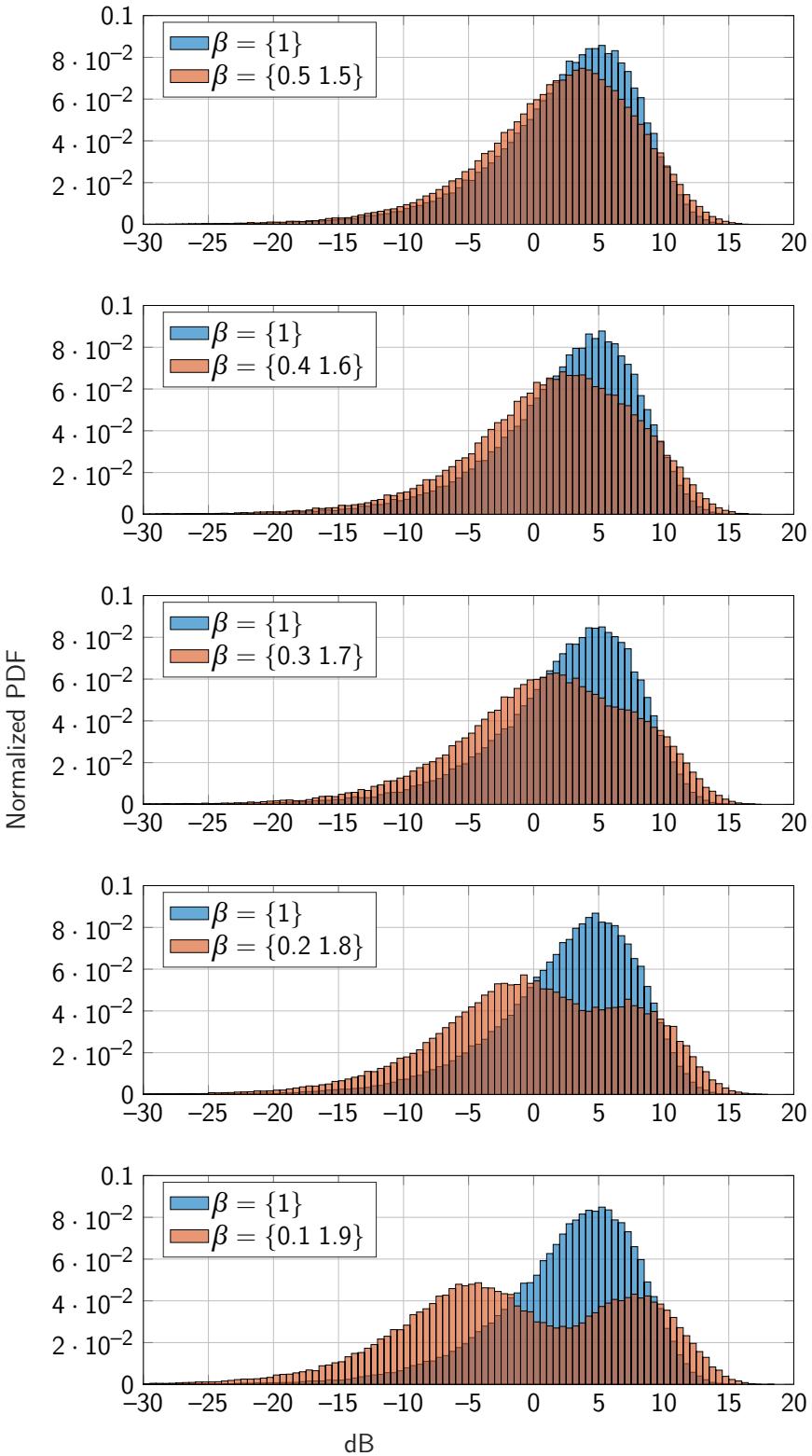
As for the H-NOMA case, (12) will have a similar change, with the modification of the instantaneous SNR of the m_0-th to be decoded, given by

$$\sigma^{non-orth}_{[m_0]} = \frac{G^{\beta}_{[m_0]}}{1 + G^{tar}_{B} + \sum_{m=m_0+1}^{A_M} G^{\beta}_{[m]}}. \tag{16}$$

In order to illustrate the potential of the proposed approach, let us consider the following power allocation vectors: $\beta = \{1.5\ 0.5\}, \{1.6\ 0.4\}, \{1.7\ 0.3\}, \{1.8\ 0.2\}, \{1.9\ 0.1\}$. These power allocation vectors were chosen in an arbitrary manner to study how the distance between the elements affects the distribution of the SNR of the devices seen at the BS. Figure 7 shows the probability density function (PDF) of mMTC devices instantaneous channel gains, considering the aforementioned values of β. One can see that the PDF considerably changes upon changing β and that the probability density becomes organized in clusters that are more visible when the values from the power allocation vector β are more distant.

The creation of the so-called clusters enables the occurrence of more events of successful SIC due to the larger difference in the instantaneous SNRs of the mMTC devices. Other power allocation vectors could be used to create more than two average SNRs at the BS. However, the constraint of having unitary mean into the elements of β makes it difficult to obtain well-separated clusters without the average SNR of one of them being very low, leading to good performance only in case the target rate is also very low. Thus, in this work we consider the cardinality of β to be two. By enforcing this imbalance of the average channel gains seen at the BS, the dynamics of the interference patterns are also changed. This imbalance creates more than just a point at which the instantaneous SNRs of the mMTC devices fluctuate, potentially benefiting the SIC performance.

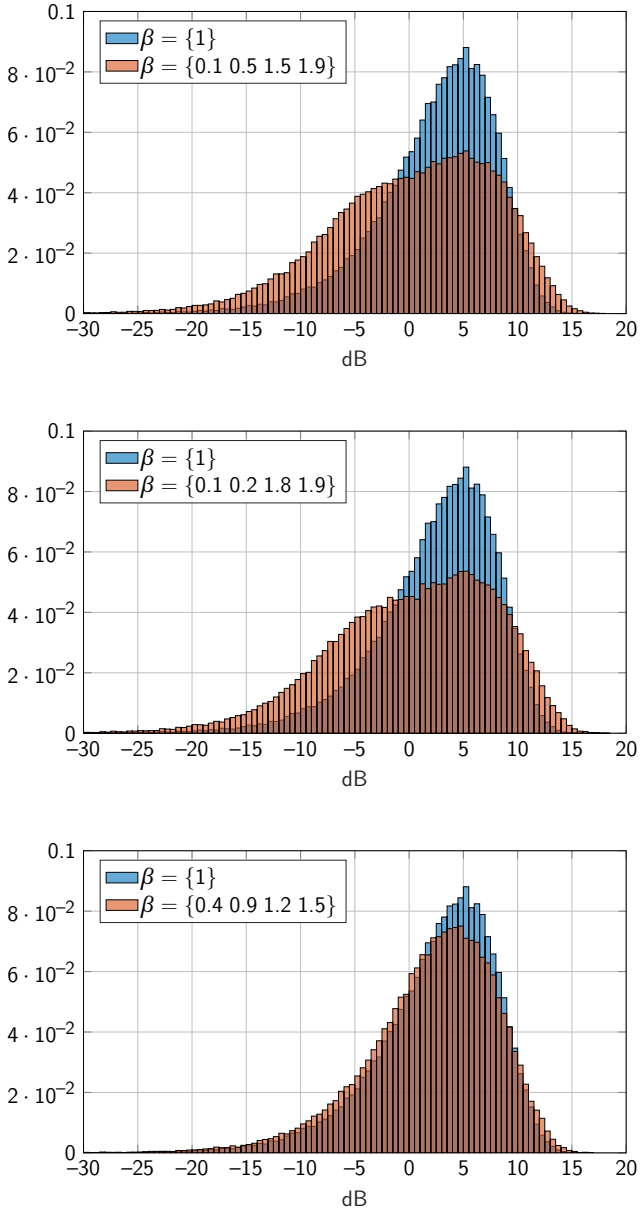
Figure 7 – H-NOMA histograms of mMTC SNRs seen by the BS with and without the proposed method. When $\beta = \{1.9\ 0.1\}$, one can clearly see the existence of "clusters of instantaneous SNR". This behavior becomes more evident as the distance between the elements of the vector increases.



Source: Author.

One could think of other ways to build the power allocation vector β by increasing the number of elements, but following the unitary mean restriction. Figure 8 shows how a β composed of four elements would change the mMTC SNRs seen at the BS; when comparing with the power allocation vectors from Figure 7, it can be seen that what really impacts the spreading is the amount of difference between the big and small vector elements. With just two elements in β is easier to achieve a complete visualization of the clusters than when using four elements, due to the unitary mean restriction.

Figure 8 – H-NOMA histograms of mMTC SNRs seen by the BS with and without the proposed method. It is depicted in this figure how the configuration of the elements present in the β power allocation vector impacts the distribution of the SNRs.



Source: Author.

3.2.1 Simulation Results for H-OMA

The results for the H-OMA in terms of the number of active mMTC devices *vs.* how eMBB traffic rate can be seen in Figure 9 and in Figure 10, both for the orthogonal case. For the simulation trials, it was used a set of $A_M = 200$ mMTC devices trying to connect at the BS, and 10^3 runs with independent channel coefficients. When taking into account the H-OMA case, the channel access multiplexing is temporal – a time division multiple access. This means that a fraction α of the time slot is allocated to the eMBB service, while the remaining $\alpha - 1$ is concurrently used by the mMTC devices.

If the channel is fully available for mMTC ($\alpha = 0$) with a low transmission rate r_M, a high number of mMTC devices can be decoded successfully; that is, the available radio resource allows devices with precarious channel gains to be decoded. However, as the eMBB traffic increases – that is, the dedicated time for eMBB increases and for mMTC decreases – α tends to 1, so the effective mMTC rate would have to increase. The result is that only devices with better channels can be decoded, and thus the "cut" is higher and the λ_M is lower in this region.

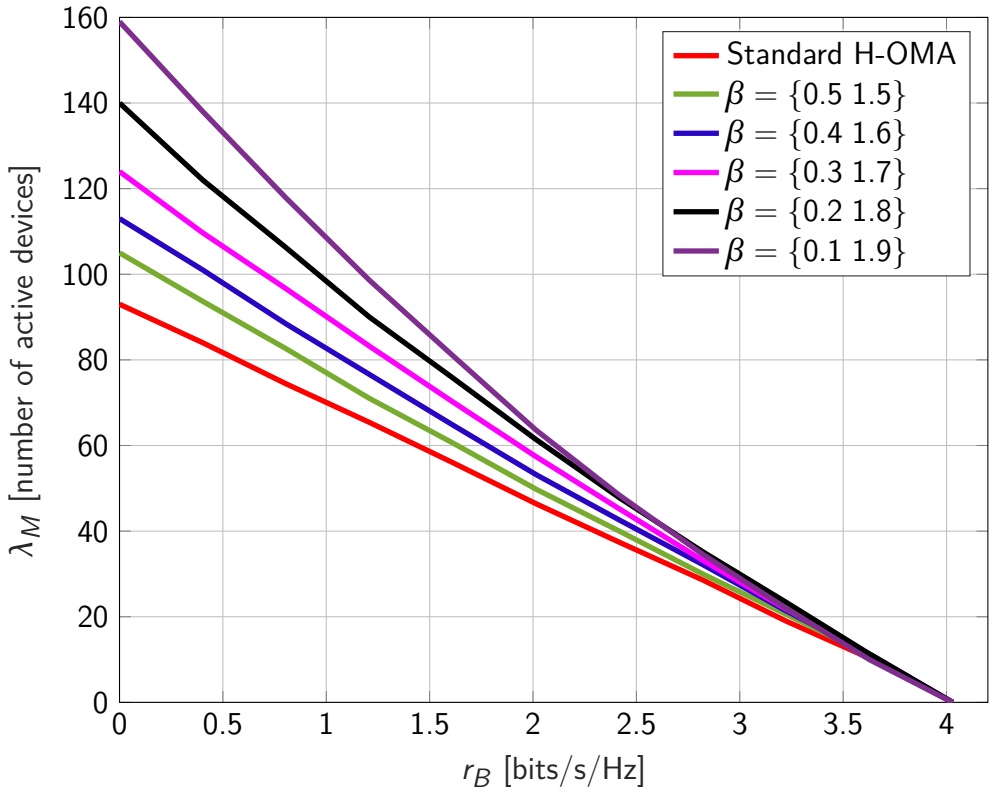
In practice, this behavior can be understood from (9), in which as α tends to 1, the value of its argument tends to infinity. In the process of determining the number of mMTC devices coexisting with the eMBB service, it is necessary to ensure that the reliability requirement is met through the ratio given by (7); for this, it is necessary to count the average number of mMTC devices active and in outage. It is precisely in this count that the introduction of α in the determination of λ_M is implied by the fact that an mMTC device will be correctly decoded if the inequality below is satisfied.

$$\log_2\left(1 + \sigma_{[m_0]}\right) \geq \frac{r_M}{1-\alpha}. \tag{17}$$

In other words, as α moves from 0 to 1, the higher the effective mMTC transmission rate should be so that the information rate r_M is achieved and the mMTC devices continue to be decoded correctly.

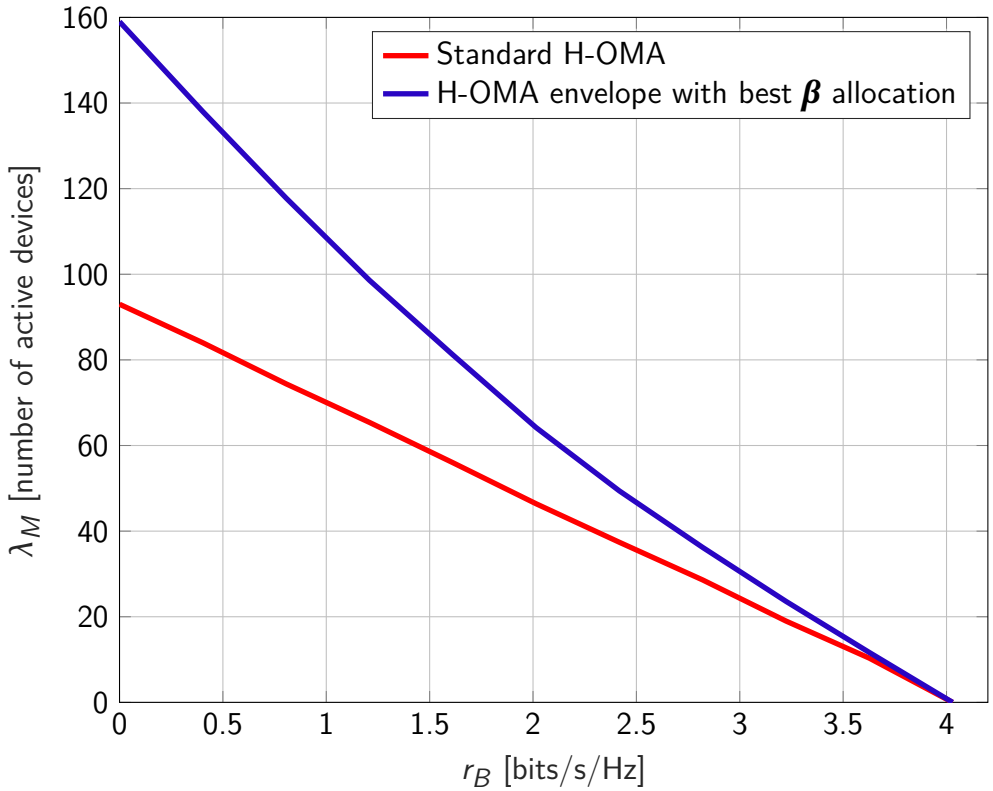
In Figure 10 it is also possible to see that, considering the power allocation vectors for each case, it is possible to obtain an average gain of 60% in the number of mMTC devices using the channel in time-sharing with eMBB when compared to the standard method without the proposed scheme. This gain is evaluated as the ratio between the active devices of both methods, among the values of r_B.

Figure 9 – Arrival rate λ_M^{orth} for the H-OMA case, as a function of eMBB rate r_B. The simulation parameters are $\Gamma_M = 5$ dB, $\Gamma_B = 20$ dB, $\varepsilon_M = 10^{-1}$, $A_M = 200$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$.



Source: Author.

Figure 10 – Arrival rate λ_M^{orth} for the H-OMA case, considering the best β values from Figure 9 for each case, as a function of eMBB rate r_B. The simulation parameters are $\Gamma_M = 5$ dB, $\Gamma_B = 20$ dB, $\varepsilon_M = 10^{-1}$, $A_M = 200$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$.



Source: Author.

3.2.2 Simulation Results for H-NOMA

In the non-orthogonal case, all devices transmit simultaneously, in the same channel. Here, in addition to performing the decoding of the active mMTC devices, the SIC also performs a separation and decoding of the eMBB user, if active. In this way, the influence of broadband traffic as interference in the mMTC decoding process varies as its occupancy in the channel increases.

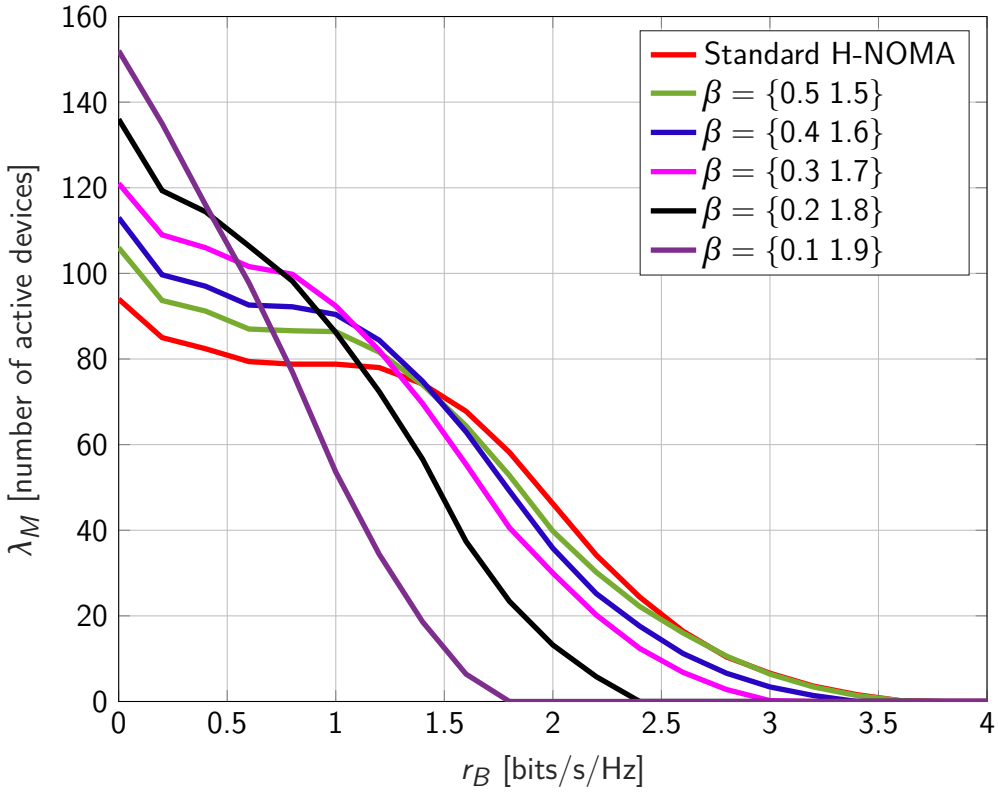
Figure 11 shows the result of using the power allocation proposed method to find the pairs (r_B, λ_M) of the H-NOMA case. For the simulation trials, it was used a set of $A_M = 200$ mMTC devices trying to connect at the BS, and 10^3 runs with independent channel coefficients. The figure can be divided into two regions. *i)* The region where there is little influence of eMBB traffic and *ii)* The region where interference from broadband traffic does not allow gains when using the power allocation method.

In the first region, when the rate r_B is low and there is little use of the channel by the eMBB device, the behavior is very similar to the H-OMA case. Here, the eMBB interference can be removed by the SIC and the mMTC performance remains high with the power allocation procedure. As the rate r_B increases, the eMBB device can only be decoded after the mMTC devices with better channels are decoded and canceled by the SIC. This occurs due to the fact that when r_B is low, the interference between services is not high enough to cause degradation; then, the most part of the mMTC devices can be decoded without effort. However, when r_B is high, the eMBB device introduces a strong component of interference and can only be decoded after some of mMTC devices with the best channels. In the standard case, when the proposed power allocation method is not used, after decoding the eMBB device, the process fails due to interference from mMTC devices that had not yet been decoded. When using the proposed method, after eMBB decoding, the mMTC devices that remain have channels with different enough SNRs for the SIC to be able to decode and cancel the best ones for a few more steps, explaining the better mMTC performance in this initial range of r_B.

In the second region, as r_B increases, the mutual interference between mMTC and eMBB devices causes the performance of both services to be low. With the use of β power allocation vectors, the more unbalanced the gains of the devices, the sooner the degradation of the two services occurs. Figure 11 shows that there is a trade-off between the channel usage by the eMBB service and the imbalance SNR level of the mMTC devices. In a high imbalanced scenario – for example, $\beta = [0.1, 1.9]$ – the minimum channel usage by the eMBB service causes the two services to degrade by r_B around 1.5 bits/s/Hz. On the other hand, one can forego having a high amount of λ_M for low values of r_B and thus extend the point where service degradation occurs, just using $\beta = [0.5, 1.5]$.

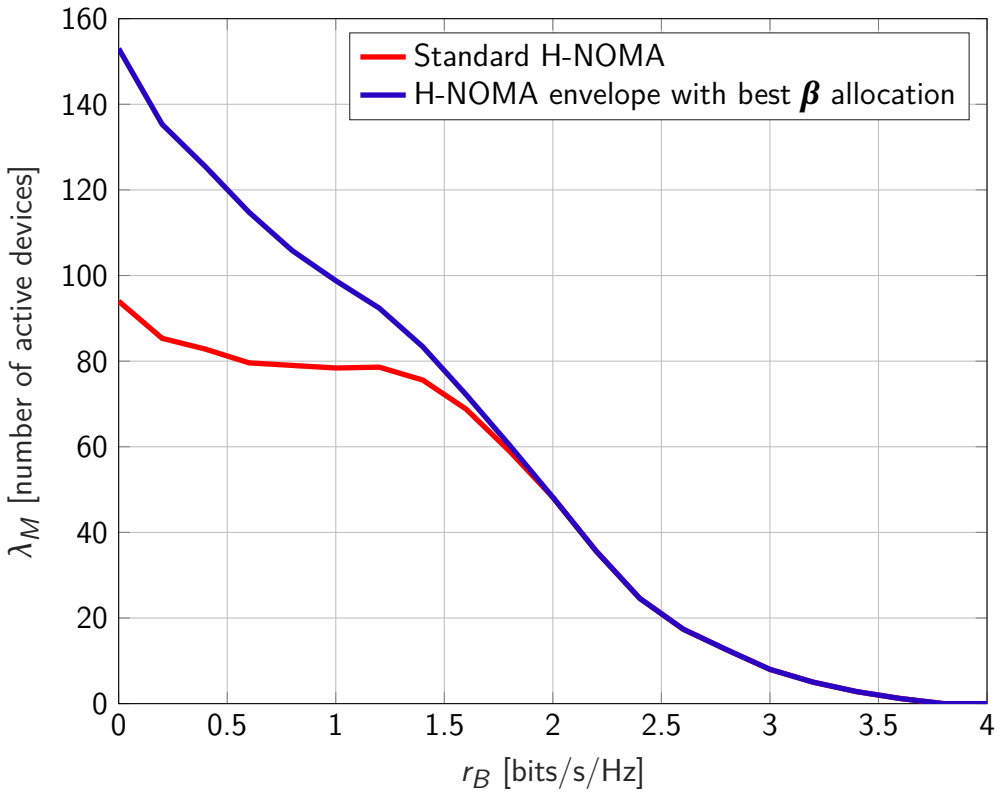
Considering a system that can choose the power allocation vector according to the network operating point in order to always maximize the eMBB transmission rate and the number of mMTC devices connected to the BS, its performance curve can be obtained through the envelope of Figure 11, as can be seen in Figure 12. As an average performance gain of 40% can be achieved for regions with low eMBB traffic – up to about $r_B = 1.5$ bps/Hz. For rates above this value, the interference between the devices of the different services stands out, and the envelope follows the curve of the standard procedure.

Figure 11 – Arrival rate $\lambda_M^{non-orth}$ for the H-NOMA case, with the clustering procedure, as a function of eMBB rate r_B. The simulation parameters are $\Gamma_M = 5$ dB, $\Gamma_B = 20$ dB, $A_M = 200$, $\varepsilon_M = 10^{-1}$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$.



Source: Author.

Figure 12 – Arrival rate $\lambda_M^{non-orth}$ for the H-NOMA case, with the envelope containing the β values with best performance from Figure 11, as a function of eMBB rate r_B. The simulation parameters are $\Gamma_M = 5$ dB, $A_M = 200$, $\Gamma_B = 20$ dB, $\varepsilon_M = 10^{-1}$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$.



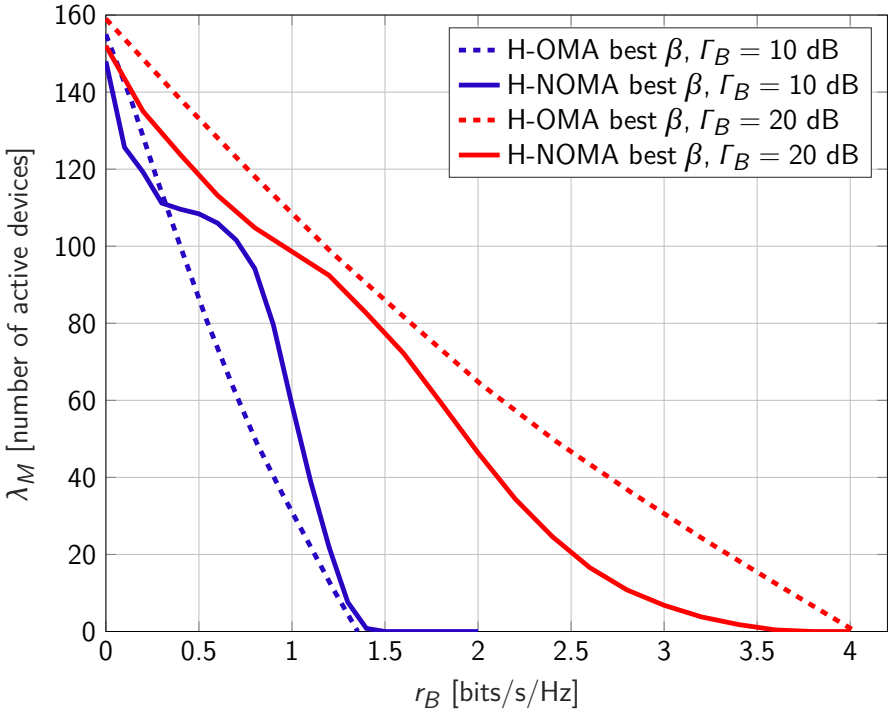
Source: Author.

3.2.3 H-OMA and H-NOMA Results Comparison

To compare the performance between H-OMA and H-NOMA using the power allocation method in mMTC devices, two average SNR scenarios were simulated for the eMBB service, with the results shown in Figure 13. For $\Gamma_B = 10$ dB, it is noted that the H-OMA outperforms the H-NOMA results only when the channel is almost fully allocated for mMTC devices. As time is divided equally until the moment when only the eMBB service uses the channel, H-NOMA is the alternative that allows reaching the largest pairs (r_B, λ_M).

However, this same behavior does not occur when $\Gamma_B = 20$ dB. In this scenario, the use of the power allocation method allows greater gains to the H-OMA when compared to the H-NOMA. In fact, there is a lower bound for the H-NOMA pairs causing saturation of the result as Γ_B increases (POPOVSKI *et al.*, 2018a). At the beginning of the values of r_B, the envelope of the H-NOMA is stretched upwards using the power allocation method, but even so, it cannot overcome the performance of the orthogonal method.

Figure 13 – Arrival rate $\lambda_M^{non-orth}$ for both H-OMA and H-NOMA case, with the envelope containing the β values that allows the best performance, as a function of eMBB rate r_B. It is shown the comparison between the orthogonal and non-orthogonal methods when using the power allocation technique. The simulation parameters are $\Gamma_M = 5$ dB, $\varepsilon_M = 10^{-1}$, $A_M = 200$, $\varepsilon_B = 10^{-3}$ and $r_M = 0.04$.



Source: Author.

4 CONCLUSION AND FUTURE WORKS

This work discussed two approaches to network slicing: one in the network layer domain and another in the physical layer domain. We sought to obtain a high-level view of network slicing in a wireless network, understanding from the requirements for its operation (*i.e.,* infrastructure, architectures, etc.) to the elaboration of a method to improve the performance when eMBB and mMTC traffic coexists.

In the review of the network layer, it was described what kind of architecture must exist so that the mobile communication systems of the future can house functionalities such as network slicing. In this sense, the existence of virtualized infrastructure – called NFVs – that run on generic hardware resources and house virtualized VNF functions is essential for creating network slices. Moreover, in the implementation of slicing in the access network, the role of Open RAN is addressed, a fully programmable and virtualized architecture that proposes the opening of interfaces and the introduction of intelligence in the network through RIC. A use case proposed by the O-RAN Alliance working group was discussed in detail.

Then, the coexistence of eMBB and mMTC services in the physical layer was also addressed, based on the model proposed by (POPOVSKI *et al.*, 2018a). Through the concept of heterogeneity between services, two forms of coexistence can be used: H-OMA and H-NOMA, for orthogonal and non-orthogonal methods, respectively. A method of changing the transmission powers of mMTC devices through power allocation vectors is proposed. In this way, the instantaneous SNR of the mMTC devices, seen at the BS, fluctuates around more than one average, improving SIC performance. It is shown that for the orthogonal case (H-OMA) there is an average gain of approximately 60% of active mMTC devices when using the same frequency resource of an eMBB device in time-sharing, in contrast to the standard procedure without the proposed power allocation method. For the non-orthogonal case (H-NOMA), an average gain of approximately 40% is verified for eMBB rates of up to $r_B = 1.5$ bits/s/Hz, that is, for low eMBB traffic; as traffic increases, interference between services becomes too high and overall performance drops dramatically.

Future works on slicing concepts at the network layer can study the use of testbeds such as (BREEN *et al.*, 2021) and (BONATI *et al.*, 2021), which can simulate network slicing scenarios and build datasets to train machine learning models and build applications capable of running on an Open RAN architecture such at in (JOHNSON; MAAS; MERWE, 2021). It is possible to build a system in practice showing the details of the implementation and architecture of the network, something that is still incipient due to the difficulty of simulating and integrating the software and codes available in the open-source Open RAN platforms.

In the case of the slicing in the physical layer domain, future approaches can follow in

the direction of studying other alternatives for the power allocation vectors considered in this work. It can be evaluated how these vectors can be built in a deterministic way to maximize the performance of the coexistence of the services, instead of the use of heuristics as done in the present work.

REFERENCES

3GPP. **Architecture description (3GPP TS 38.401 version 17.2.0 Release 17)**. 3GPP: ETSI, 2022. Available from: https://www.etsi.org/deliver/etsi_ts/138400_138499/138401/17.02.00_60/ts_138401v170200p.pdf.

3GPP. **NR and NG-RAN overall description (3GPP TS 38.300)**. 3GPP: ETSI, 2022. Available from: https://www.etsi.org/deliver/etsi_ts/138300_138399/138300/17.02.00_60/ts_138300v170200p.pdf.

ALBERTI, Antonio; SILVA, Daniely; FIGUEIREDO, Felipe; CARMO, Francisco do; AQUINO, Guilherme; BRITO, Jose. **White Paper - OpenRAN, a conexão do futuro**. Inatel: Instituto Nacional de Telecomunicações – INATEL, 2022.

ASGHAR, Muhammad Zeeshan; MEMON, Shafique Ahmed; HÄMÄLÄINEN, Jyri. Evolution of Wireless Communication to 6G: Potential Applications and Research Directions. **Sustainability**, MDPI AG, v. 14, n. 10, p. 6356, May 2022.

BAI, Bo; CHEN, Wei; CAO, Zhigang; LETAIEF, Khaled. Max-matching diversity in OFDMA systems. **IEEE Transactions on Communications**, Institute of Electrical and Electronics Engineers (IEEE), v. 58, n. 4, p. 1161–1171, Apr. 2010.

BARAKABITZE, Alcardo Alex; AHMAD, Arslan; MIJUMBI, Rashid; HINES, Andrew. 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. **Computer Networks**, Elsevier BV, v. 167, p. 106984, Feb. 2020.

BONATI, Leonardo *et al.* Colosseum: Large-Scale Wireless Experimentation Through Hardware-in-the-Loop Network Emulation. *In*: 2021 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN). IEEE: IEEE, Dec. 2021.

BREEN, Joe *et al.* Powder: Platform for Open Wireless Data-driven Experimental Research. **Computer Networks**, Elsevier BV, v. 197, p. 108281, Oct. 2021.

DEBBABI, Fadoua; JMAL, Rihab; FOURATI, Lamia Chaari; AGUIAR, Rui Luis. An Overview of Interslice and Intraslice Resource Allocation in B5G Telecommunication Networks. **IEEE Transactions on Network and Service Management**, Institute of Electrical and Electronics Engineers (IEEE), v. 19, n. 4, p. 5120–5132, Dec. 2022.

ELAYOUBI, Salah Eddine; JEMAA, Sana Ben; ALTMAN, Zwi; GALINDO-SERRANO, Ana. 5G RAN Slicing for Verticals: Enablers and Challenges. **IEEE Communications Magazine**, Institute of Electrical and Electronics Engineers (IEEE), v. 57, n. 1, p. 28–34, Jan. 2019.

FOUKAS, Xenofon; PATOUNAS, Georgios; ELMOKASHFI, Ahmed; MARINA, Mahesh K. Network Slicing in 5G: Survey and Challenges. **IEEE Communications Magazine**, Institute of Electrical and Electronics Engineers (IEEE), v. 55, n. 5, p. 94–100, May 2017.

GAVRILOVSKA, Liljana; RAKOVIC, Valentin; DENKOVSKI, Daniel. From Cloud RAN to Open RAN. **Wireless Personal Communications**, Springer Science and Business Media LLC, v. 113, n. 3, p. 1523–1539, Mar. 2020.

JOHNSON, David; MAAS, Dustin; MERWE, Jacobus Van Der. NexRAN. *In*: PROCEEDINGS of the 15th ACM Workshop on Wireless Network Testbeds, Experimental evaluation Characterization. ACM: ACM, Oct. 2021.

LIEN, Shao-Yu; SHIEH, Shin-Lin; HUANG, Yenming; SU, Borching; HSU, Yung-Lin; WEI, Hung-Yu. 5G New Radio: Waveform, Frame Structure, Multiple Access, and Initial Access. **IEEE Communications Magazine**, Institute of Electrical and Electronics Engineers (IEEE), v. 55, n. 6, p. 64–71, 2017.

MEHMETI, Fidan; PORTA, Thomas F. La. Modeling and Analysis of mMTC Traffic in 5G Base Stations. *In*: 2022 IEEE 19th Annual Consumer Communications Networking Conference (CCNC). IEEE: IEEE, Jan. 2022.

NGMN. **White Paper - NGMN 5G**. NGMN: NGMN Alliance, 2015. Available from: `https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf`.

O-RAN ALLIANCE. **Non-RT RIC & A1 Interface: Use Cases and Requirements**. O-RAN Alliance: O-RAN.WG2.Use-Case-Requirements-v06.00, 2022. Available from: `https://orandownloadsweb.azurewebsites.net/specifications`.

O-RAN ALLIANCE. **O-RAN Architecture Description**. O-RAN Alliance: O-RAN.WG1.O-RAN-Architecture-Description-v07.00, 2022. Available from: `https://orandownloadsweb.azurewebsites.net/specifications`.

O-RAN ALLIANCE. **O-RAN Working Group 3, Near-Real-time RAN Intelligent Controller, E2 Application Protocol (E2AP)**. O-RAN Alliance: O-RAN.WG3.E2AP-v02.03, 2022. Available from: `https://orandownloadsweb.azurewebsites.net/specifications`.

POLESE, Michele; BONATI, Leonardo; D'ORO, Salvatore; BASAGNI, Stefano; MELODIA, Tommaso. Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges. **IEEE Communications Surveys & Tutorials**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 2023.

POPOVSKI, Petar; TRILLINGSGAARD, Kasper Floe; SIMEONE, Osvaldo; DURISI, Giuseppe. 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 6, p. 55765–55779, 2018.

POPOVSKI, Petar *et al.* Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks. **IEEE Network**, Institute of Electrical and Electronics Engineers (IEEE), v. 32, n. 2, p. 16–23, Mar. 2018.

SÁNCHEZ, Johanna Andrea Hurtado; CASILIMAS, Katherine; RENDON, Oscar Mauricio Caicedo. Deep Reinforcement Learning for Resource Management on Network Slicing: A Survey. **Sensors**, MDPI AG, v. 22, n. 8, p. 3031, Apr. 2022.

SANTOS, Elco Joao dos; SOUZA, Richard Demo; REBELATTO, Joao Luiz; ALVES, Hirley. Network Slicing for URLLC and eMBB With Max-Matching Diversity Channel Allocation. **IEEE Communications Letters**, Institute of Electrical and Electronics Engineers (IEEE), v. 24, n. 3, p. 658–661, Mar. 2020.

SHAFI, Mansoor; MOLISCH, Andreas F.; SMITH, Peter J.; HAUSTEIN, Thomas; ZHU, Peiying; SILVA, Prasan De; TUFVESSON, Fredrik; BENJEBBOUR, Anass; WUNDER, Gerhard. 5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice. **IEEE Journal on Selected Areas in Communications**, Institute of Electrical and Electronics Engineers (IEEE), v. 35, n. 6, p. 1201–1221, June 2017.

TOMINAGA, Eduardo Noboro; ALVES, Hirley; LOPEZ, Onel L. Alcaraz; SOUZA, Richard Demo; REBELATTO, Joao Luiz; LATVA-AHO, Matti. Network Slicing for

eMBB and mMTC with NOMA and Space Diversity Reception. *In*: 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). IEEE: IEEE, Apr. 2021.

VMWARE. **A Guide to SDN, SD-WAN, NFV, and VNF**. VMWare: VMWare, 2019. Available from: `https://www.vmware.com/content/dam/digitalmarketing/vmware-sase/pdfs/208805aq-so-vcloud-guide-sd-wan-nfv-vfn-uslet-web.pdf`.

WU, Wen; ZHOU, Conghao; LI, Mushu; WU, Huaqing; ZHOU, Haibo; ZHANG, Ning; SHEN, Xuemin Sherman; ZHUANG, Weihua. AI-Native Network Slicing for 6G Networks. **IEEE Wireless Communications**, Institute of Electrical and Electronics Engineers (IEEE), v. 29, n. 1, p. 96–103, Feb. 2022.

ZHANG, Haijun; LIU, Na; CHU, Xiaoli; LONG, Keping; AGHVAMI, Abdol-Hamid; LEUNG, Victor C. M. Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges. **IEEE Communications Magazine**, Institute of Electrical and Electronics Engineers (IEEE), v. 55, n. 8, p. 138–145, Aug. 2017.

ZHANG, Shunliang. An Overview of Network Slicing for 5G. **IEEE Wireless Communications**, Institute of Electrical and Electronics Engineers (IEEE), v. 26, n. 3, p. 111–117, June 2019.

ZHANG, Zhengquan; XIAO, Yue; MA, Zheng; XIAO, Ming; DING, Zhiguo; LEI, Xianfu; KARAGIANNIDIS, George K.; FAN, Pingzhi. 6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies. **IEEE Vehicular Technology Magazine**, Institute of Electrical and Electronics Engineers (IEEE), v. 14, n. 3, p. 28–41, Sept. 2019.