



FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING  
DEGREE PROGRAMME IN WIRELESS COMMUNICATIONS ENGINEERING

## MASTER'S THESIS

# SCALABLE COEXISTENCE OF eMBB, URLLC AND mMTC ENABLED BY NON-ORTHOGONAL MULTIPLE ACCESS AND NETWORK SLICING

|                 |                             |
|-----------------|-----------------------------|
| Author          | Eduardo Noboro Tominaga     |
| Supervisor      | Prof. Hirley Alves          |
| Second Examiner | Prof. Carlos Morais de Lima |

January 2021

**Tominaga E. N. (2021) Scalable Coexistence of eMBB, URLLC and mMTC Enabled by Non-Orthogonal Multiple Access and Network Slicing.** University of Oulu, Faculty of Information Technology and Electrical Engineering, Degree Programme in Wireless Communications Engineering, Master's Thesis, 54 p.

## **ABSTRACT**

The 5G systems feature three use cases: enhanced Mobile BroadBand (eMBB), massive Machine-Type Communications (mMTC) and Ultra-Reliable and Low-Latency Communications (URLLC). The diverse requirements of the corresponding services in terms of achievable data-rate, number of connected devices, latency and reliability can lead to sub-optimal use of the 5G resources, thus network slicing emerges as a promising alternative that customizes slices of the network specifically designed to meet specific requirements. By employing network slicing, the radio resources can be shared via orthogonal and non-orthogonal schemes. Motivated by the Industrial Internet of Things (IIoT) paradigm where a large number of sensors may require connectivity with stringent requirements of latency and reliability, we propose and evaluate the joint use of network slicing and Non-Orthogonal Multiple Access (NOMA) with Successive Interference Cancellation (SIC) in two different uplink scenarios. In the first scenario, eMBB coexists with URLLC in the same Radio Access Network (RAN) and, in order to improve the number of concurrent URLLC connections to the same base station (BS), they transmit simultaneously and across multiple frequency channels. In the second scenario, eMBB coexists with mMTC and, to provide connectivity to a massive number of devices, the BS has multiple receive antennas. In both cases, we set the reliability requirements for the services and compare the performance of both orthogonal and non-orthogonal network slicing schemes in terms of maximum achievable data rates and connected users. Our results show that, even with overlapping transmissions from multiple devices, network slicing, NOMA and SIC techniques allow us simultaneously satisfy all the heterogeneous requirements of the 5G services.

**Keywords:** 5G, Network Slicing, eMBB, URLLC, mMTC, IIoT, NOMA, SIC.

# TABLE OF CONTENTS

|   |    |
|---|----|
| ABSTRACT  |    |
| TABLE OF CONTENTS   |    |
| FOREWORD  |    |
| LIST OF ABBREVIATIONS AND SYMBOLS                               |    |
| 1 INTRODUCTION  | 8  |
| 1.1 5G, Beyond-5G and 6G Networks                               | 8  |
| 1.2 Network Slicing   | 9  |
| 1.3 URLLC Principles and building blocks                        | 9  |
| 1.4 Industrial IoT and Cyberphysical Systems                    | 11 |
| 1.5 Non-Orthogonal Multiple Access and SIC Decoding             | 12 |
| 1.6 Antenna Diversity   | 13 |
| 1.7 Related Works   | 13 |
| 1.8 Main Contributions  | 14 |
| 1.9 Work Outline  | 16 |
| 2 PRELIMINARIES   | 17 |
| 2.1 Capacity of the AWGN channel                                | 17 |
| 2.2 Capacity of fading channels                                 | 18 |
| 2.3 Slow Fading Channel and $\epsilon$ -Outage Capacity         | 19 |
| 2.4 Capacity of fading channels with frequency diversity        | 20 |
| 2.5 Capacity of SIMO fading channel                             | 22 |
| 2.6 Multiuser Capacity, NOMA and SIC decoding                   | 23 |
| 2.7 Comparison between OMA and NOMA schemes                     | 25 |
| 2.8 Extension to the $K$ -user capacity                         | 26 |
| 2.9 Uplink fading channel                                       | 27 |
| 3 SYSTEM MODEL  | 29 |
| 3.1 Network Slicing for the Coexisting eMBB and URLLC Use Cases | 29 |
| 3.2 Network Slicing for Coexisting eMBB and mMTC Use Cases      | 31 |
| 3.3 Analysis of the eMBB Performance                            | 33 |
| 3.4 Analysis of the URLLC Performance                           | 34 |
| 3.5 Analysis of the mMTC Performance                            | 37 |
| 4 NETWORK SLICING FOR COEXISTING EMBB AND URLLC SCENARIOS       | 41 |
| 4.1 Orthogonal Slicing For Coexisting eMBB and URLLC            | 41 |
| 4.2 Non-Orthogonal Slicing Between Coexisting eMBB and URLLC    | 41 |
| 4.3 Performance Evaluation                                      | 42 |
| 5 NETWORK SLICING FOR COEXISTING EMBB AND MMTC SCENARIOS        | 45 |
| 5.1 Orthogonal Slicing Between Coexisting eMBB and mMTC         | 45 |
| 5.2 Non-Orthogonal Slicing Between Coexisting eMBB and mMTC     | 45 |
| 5.3 Performance Evaluation                                      | 47 |
| 6 CONCLUSIONS   | 50 |
| 7 REFERENCES  | 52 |

## FOREWORD

This research was completed in the Centre of Wireless Communications (CWC) of the University of Oulu and was financially supported by 6Genesis Flagship project (grant 318927), FIREMAN project (grant 326201) and Academy Professor project from Academy of Finland (grant 307492).

I would like to thank professors Hirley Alves, Richard Demo Souza, Onel Luis Alcaraz López and João Luiz Rebelatto for their guidance during the research work I performed in the past two years. I also would like to thank professor Carlos Morais de Lima for reviewing this thesis and for giving me many valuable tips about scientific writing. Specially, I would like to express my most sincere gratitude to Prof. Hirley for giving me the opportunity to work as an Research Assistant at the CWC and also pursue my master's degree in Wireless Communications Engineering (WCE).

Oulu, 1st February, 2021

Eduardo Noboro Tominaga

## LIST OF ABBREVIATIONS AND SYMBOLS

|         |   |
|---------|---|
| 3GPP    | Third Generation Partnership Project                |
| 5G      | Fifth Generation of Wireless Communications Systems |
| 6G      | Sixth Generation of Wireless Communications Systems |
| ACK     | Acknowledgment                                      |
| AWGN    | Additive White Gaussian Noise                       |
| BS      | Base Station  |
| CDMA    | Code-Division Multiple Access                       |
| CPS     | Cyber-Physical System                               |
| CSI     | Channel-State Information                           |
| eMBB    | enhanced Mobile BroadBand                           |
| FDMA    | Frequency-Division Multiple Access                  |
| HTC     | Human-Type Communication                            |
| IID     | Independent and Identically Distributed             |
| IIoT    | Industrial Internet of Things                       |
| IoT     | Internet of Things                                  |
| MMSE    | Minimum Mean Square Error                           |
| mMTC    | massive Machine-Type Communication                  |
| MRC     | Maximum Ratio Combining                             |
| MTC     | Machine-Type Communication                          |
| MU-MIMO | Multiuser Multiple-Input Multiple-Output            |
| mURLLC  | massive Ultra-Reliable Low-Latency Communication    |
| NOMA    | Non-Orthogonal Multiple Access                      |
| OFDMA   | Orthogonal Frequency-Division Multiple Access       |
| OMA     | Orthogonal Multiple Access                          |
| RAN     | Radio Access Network                                |
| SIC     | Successive Interference Cancellation                |
| SIMO    | Single-Input Multiple-Output                        |
| SINR    | Signal-to-Interference-plus-Noise Ratio             |
| SNR     | Signal-to-Noise Ratio                               |
| TDMA    | Time-Division Multiple Access                       |
| URLLC   | Ultra-Reliable Low-Latency Communication            |
| ZF      | Zero-Forcing  |

### Symbols:

|                   |  |
|-------------------|--|
| $a_B$             | eMBB activation probability                            |
| $\alpha$          | Fraction of the radio resource allocated for an user   |
| $B$               | Bandwidth  |
| $C_{\text{AWGN}}$ | Capacity of the AWGN channel                           |
| $C_{\text{SIMO}}$ | Capacity of the SIMO channel                           |
| $C_{\text{sum}}$  | Sum capacity of the multiuser channel                  |
| $C_{\text{sym}}$  | Symmetric capacity of the multiuser channel            |
| $\mathcal{C}$     | Capacity region of the multiuser wireless channel      |
| $C_\epsilon$      | $\epsilon$ -outage capacity of the slow fading channel |
| $D_M$             | Number of decoded mMTC devices in a timeslot           |

|                             |  |
|-----------------------------|--|
| $\epsilon_B$                | Reliability requirement for eMBB   |
| $\epsilon_M$                | Reliability requirement for mMTC   |
| $\epsilon_U$                | Reliability requirement for URLLC  |
| $f$                         | Frequency channel index  |
| $F$                         | Number of frequency channels in the time-frequency grid                    |
| $F_B$                       | Number of frequency channels allocated for eMBB                            |
| $F_U$                       | Number of frequency channels allocated for URLLC                           |
| $g$                         | Wireless channel gain  |
| $g_{B,f}$                   | Instantaneous wireless channel gain in the frequency channel $f$ for eMBB  |
| $g_{M,f}$                   | Instantaneous wireless channel gain in the frequency channel $f$ for mMTC  |
| $g_{U,f}$                   | Instantaneous wireless channel gain in the frequency channel $f$ for URLLC |
| $\gamma$                    | Received SNR   |
| $\bar{\gamma}$              | Average received SNR   |
| $\gamma_{B,f}^{\min}$       | Threshold SNR for the eMBB device antennas                                 |
| $\gamma_{B,f}^{\text{tar}}$ | Target SNR for the eMBB device   |
| $\Gamma_B$                  | Average channel gain for eMBB  |
| $\Gamma_M$                  | Average channel gain for mMTC  |
| $\Gamma_U$                  | Average channel gain for URLLC   |
| $h[m]$                      | Wireless channel coefficient at time instant $m$                           |
| $h_{B,f}$                   | Wireless channel coefficient in the frequency channel $f$ for eMBB         |
| $h_{M,f}$                   | Wireless channel coefficient in the frequency channel $f$ for mMTC         |
| $h_{U,f}$                   | Wireless channel coefficient in the frequency channel $f$ for URLLC        |
| $I_u^{\text{sum}}$          | Sum of mutual information for the $u$ -th URLLC user                       |
| $k$                         | Index of a user  |
| $K$                         | Total number of users  |
| $\kappa$                    | Non-empty subset of active users   |
| $l$                         | Index of the receive antenna element                                       |
| $L$                         | Number of receive antenna elements   |
| $m$                         | Index of the mMTC device   |
| $M$                         | Total number of mMTC devices   |
| $n_U$                       | Total number of URLLC users  |
| $N_0$                       | Variance of the AWGN noise   |
| $P$                         | Transmit power constraint  |
| $P_B$                       | eMBB instantaneous transmit power  |
| $\mathcal{P}_{\text{out}}$  | Outage probability   |
| $\Pr(E_B)$                  | Probability of error for eMBB  |
| $\Pr(E_M)$                  | Probability of error for mMTC  |
| $\Pr(E_U)$                  | Probability of error for URLLC   |
| $r_B$                       | eMBB data rate (in bits/s/Hz)  |
| $r_B^{\text{out}}$          | eMBB outage rate   |
| $r_B^{\text{sum}}$          | eMBB sum rate  |

|                    |   |
|--------------------|---|
| $r_U$              | URLLC data rate (in bits/s/Hz)                                      |
| $r_U^{\text{out}}$ | URLLC outage rate   |
| $r_M$              | mMTC data rate (in bits/s/Hz)                                       |
| $R_k$              | Target data rate of the $k$ -th user                                |
| $s$                | Minislot index  |
| $S$                | Number of minislots in the timeslot                                 |
| $\sigma_B$         | Available SINR while decoding the signal of the eMBB device         |
| $\sigma_{[m]}$     | Available SINR while decoding the signal of the $m$ -th mMTC device |
| $u$                | Index of the URLLC device   |
| $w[m]$             | AWGN sample at time instant $m$                                     |
| $x[m]$             | Transmitted complex symbol at time instant $m$                      |
| $y[m]$             | Received complex symbol at time instant $m$                         |

### Mathematical Operators:

|                     |   |
|---------------------|---|
| $\mathbb{C}$        | Set of complex numbers  |
| $\mathbb{Z}^+$      | Set of positive integer numbers   |
| $\mathbb{E}(\cdot)$ | Expectation operator  |
| $f_X(x)$            | Probability density function of the random variable $x$                           |
| $\Gamma(a, z)$      | Upper incomplete Gamma function, $\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t} dt$ |
| $\gamma(a, z)$      | Lower incomplete Gamma function, $\gamma(a, z) = \int_0^z t^{a-1} e^{-t} dt$      |
| $\log_b(\cdot)$     | Logarithm to base $b$   |
| $\Pr\{A\}$          | Probability of the event $A$  |
| $\max(\cdot)$       | Maximum of all input arguments  |
| $(\cdot)^H$         | Conjugate transpose operation   |
| $(\cdot)^T$         | Transpose operation   |
| $ \cdot $           | Absolute value of a scalar  |
| $\ \cdot\ $         | Euclidean norm of a vector  |

### Distributions of Random Variables:

|                      |   |
|----------------------|---|
| $\mathcal{CN}(a, b)$ | Circularly symmetric complex Gaussian distribution with mean $a$ and variance $b$ |
|----------------------|---|

# 1 INTRODUCTION

## 1.1 5G, Beyond-5G and 6G Networks

The deployment of the Fifth Generation (5G) of wireless communications systems has already started around the world. In addition to the traditional Human-Type Communication (HTC) services (voice calls, text messaging and mobile internet) that were provided by the previous generations, the 5G is the first generation to provide native Machine-Type Communication (MTC) services, which are the key enablers of the Internet of Things (IoT). More specifically, the 5G introduces three use cases [1], [2]:

- **enhanced Mobile Broadband (eMBB):** provides increased data rates for HTC services. The peak and moderate rates are expected to be on the order of gigabits and megabits per second, respectively, while the latter case should also have very high availability. Some promising applications are augmented/virtual reality and remote presence;
- **massive Machine-Type Communications (mMTC):** provides wireless connectivity for a very large number of devices with low software and hardware complexity and low-energy operation. These device are rarely active, usually require low data rates and are deployed in wide area setups with hundreds or even thousands of devices per square kilometer of coverage;
- **Ultra-Reliable and Low-Latency Communications (URLLC):** also known as critical MTC, should provide ultra-reliable wireless connectivity while operating in short block lengths, a requirement to achieve low latency demanded by time-critical applications. Promising URLLC applications are, for example, vehicle-to-vehicle/infrastructure communications and critical industrial automation.

5G enables many new business opportunities in a variety of different areas, including automotive industry, construction, energy systems, healthcare, manufacturing, media, retail and transportation [2].

Nowadays 5G is maturing as a global standard, but it is already recognized that present-day technologies and network infrastructure are not yet capable of fully meeting the very stringent requirements in terms of latency and reliability of the target URLLC applications. Thus, the research community around the globe has already started defining new key performance indicators and developing technical solutions for Beyond-5G systems, hereafter referred to as Sixth Generation (6G) wireless communication systems, whose deployment is expected to start worldwide in 2030 [3]. Regarding the mMTC use case, it is predicted that the number of connected devices will increase substantially, up to hundreds of devices per cubic meter, which poses very stringent requirements on the spatial spectral efficiency and respective frequency bands for reliable connectivity [3]. One of the use cases for mMTC towards 6G are the connected industries, that is, the evolution from Industry 4.0 to Industry 5.0. The massive connectivity in industrial setups will enable data-driven solutions with unprecedented levels of personalization and customization of products, as well as the improvement of the operation and performance efficiency [4].

In this context, the industrial control for wireless factory automation is one of the most challenging deployments envisioned for 6G networks. For instance, it may

require only one erroneous bit in a billion and a latency lower than 0.1 ms [3]. Such stringent requirements demand highly-flexible networks with (re)configurable radios, where artificial intelligence and machine learning techniques can be used to obtain knowledge about the static and dynamic characteristics of the radio environment, and so perform dynamic resource allocation for base stations and users [3]. From [5], it is worthy emphasizing that 6G is expected to be the first wireless standard to completely replace existing industry-specific standards by a single global solution enabling seamless connectivity across different vertical industries.

## 1.2 Network Slicing

The coexistence of the different 5G services and applications with their diverse and sometimes conflicting requirements can lead to a sub-optimal use of the wireless network. An efficient solution is to slice the network in multiple virtual and isolated logical networks running on a common physical infrastructure in an efficient and economic way [6]. A network slice can span across multiple parts of the network, be deployed across multiple operators and comprise dedicated and/or shared resources e.g. in terms of processing power, storage and bandwidth, and be individually customized with respect to, e.g., latency, energy efficiency, mobility, massive connectivity and throughput [6]. A factory, for example, could order from a operator network slices for URLLC that could be used by its critical control systems.

Most of the recent works about network slicing for 5G, for example [7], [8] and [9], deal with the problem of the slicing of communication and computing resources for the upper layers of communications systems, which is out of the scope of this work. However, some recent works like [10] and [11] also present the concept of network slicing for the physical layer, that is, the partitioning of radio resources among heterogeneous 5G services, which is the definition adopted on this thesis.

Some future 6G applications may require a dynamic and/or multiple service-type resource allocation. Different from the fixed categorization of eMBB, URLLC and mMTC, some 6G applications may require wireless communication services with requirements of data rates, latency, reliability and number of connected devices that can dynamically change over time. In this new context, it will be needed to establish an automatic and dynamic provisioning of the required network slices and resource allocation according to the needs of each application [5]. In [12], for example, the authors introduce the concept of massive URLLC (mURLLC), which is the merge of the URLLC and mMTC services, and refers to applications that features a reliability-latency-scalability trade-off.

## 1.3 URLLC Principles and building blocks

Due to the stringent requirements imposed by ultra-low latency and very high reliability, the design of a network slice for the URLLC service is overly challenging. In this regard, Popovski et al. propose the basic building blocks of a wireless communication system for supporting the URLLC in two different works [13], [14]. Equally important, metadata (i.e. control information) and payload sizes become comparable in a URLLC packet,

such that they should be jointly encoded so as to reduce the required transmission bandwidth. Moreover, authors also state that the concepts of reliability and latency are tightly coupled, meaning that the former corresponds to the probability that the latter does not exceed a pre-described deadline when delivering a packet.

Therein, one of the basic foundations of URLLC is the exploitation of different types of diversity, including access point diversity (i.e. network densification), spatial diversity obtained with the use of a massive number of antennas, and interface diversity through the use of independent paths to transmit a packet (that may include different wireless technologies and/or different mobile network operators). In [13], authors present an interesting discussion about the importance of frequency diversity for URLLC. The number of available channel uses for a URLLC transmission is limited by the latency constraints, and is approximately proportional to the product of the time duration and the bandwidth of the transmitted signal. Hence, if we increase the bandwidth of the URLLC transmission, we obtain more available channel uses, enabling us to decrease the channel use time duration. In contrast, if we fix the time duration for the URLLC transmission, increasing the bandwidth enables us to achieve higher reliability due to the frequency diversity [13].

In [14], authors discuss about access networking for URLLC. They propose the use of static allocation of resources in scenarios with deterministic traffic arrivals, and three different schemes in scenarios with stochastic traffic arrivals:

- Four-step access: the device sends a transmission request, and then waits for an access grant. Next, it sends the data on the uplink, and finally waits to receive an ACK. This scheme is suitable when the URLLC devices have a very low probability of activation;
- Three-step access: the request is skipped and the BS sends directly the access grant to the device. This scheme is indicated when the BS can accurately predict the activation of the URLLC devices;
- Grant-free access (or two-step access): the BS skips the access grant transmission. This scheme is indicated in scenarios with very stringent latency requirements or when the URLLC packets are very short, such that the overhead due to request/grant is very significant and impacts the system efficiency.

The need for a grant-free or coordinated grant-free access depends on the latency constraints. These random and non-orthogonal mechanisms aim to skip the reservation phase at the expense of collisions and interference among users [13]. To avoid that, URLLC device can use more bandwidth and/or transmission power than would be required if no collisions occurred. In this approach, multi-packet reception can be achieved, and the base station could use Successive Interference Cancellation (SIC) to recover the multiple overlapping packets. To avoid many collisions in scenarios where the users have high activation probability, coordinated grant-free access can be adopted. On this approach, the scheduling of URLLC users is performed by the base station, which specifies an access pattern to them. This access pattern can be, for example, the assignment of some URLLC users to transmit only in specific slots [13].

Mahmood et. al. [5] present six key enablers for MTC in 6G networks. Regarding the enablers for the efficient, fast and massive access, they propose to use agile numerology such as minislots, that can be flexibly adjusted to the channel conditions so as to improve

the latency and reliability of the wireless links. They also state the grant-free access is a solution to reduce the latency of URLLC, since a URLLC user promptly transmit its packet without the need of receiving a channel grant. From [5], semi-persistent scheduling technique can efficiently reduce the transmission latency for industrial applications with periodic traffic arrival. By employing this technique, radio resources are periodically allocated to nodes thus reducing the latency associated with grant-acquisition.

#### 1.4 Industrial IoT and Cyberphysical Systems

In a recent white paper [15], Ericsson defines four different 5G-based IoT segments that jointly enable the deployment of a industrial network:

- *Massive IoT* focuses on extending coverage for a massive number of low-complexity devices that infrequently exchange messages and have extreme requirements on battery life. Their traffic is often delay tolerant, and include applications such as metering, wearables and trackers.
- *Critical IoT* enables applications that require extremely low latency and ultra-high reliability, such as real-time coordination between autonomous vehicles and transportation infrastructure, detection and restoration of faults in smart grids, remote surgeries and remote driving.
- *Industrial automation IoT* aims at providing connectivity in industrial environments with extremely demanding requirements, very accurate indoor positioning and distinct architecture and security attributes.
- *Broadband IoT* is the solution for IoT applications that require higher data rates and latency lower than Massive IoT, but that also require extended coverage and very long battery life. Some use cases are augmented/virtual reality, autonomous cars and drone control.

The combination of industrial automation IoT and critical IoT is the key enabler of the Industry 4.0, a concept that envisions a factory where all devices and assets are fully connected. From [15], all the aforementioned IoT segments can be supported in the same Radio Access Network (RAN) with the effective use of techniques such as network slicing and radio resource partitioning.

The wireless connectivity required by the different IoT segments will be provided by the generic 5G services. The wireless communication systems can be used to add redundancy or replace faulty wired solutions in harsh industrial environments, where temperature, pressure, vibration, radiation or atmospheric corrosion may make wired communication unreliable. They can also be used in safety systems to detect or prevent injuries to humans or to the environment, and can also generate useful data that can be used to optimize factory operations, machine scheduling and maintenance, reduce the production costs and improve the quality of the factory output [16].

The three basic characteristics of an industrial network are reliability, latency and scale. The latency, which corresponds to the delay between the time when an event happened and the time in which knowledge of that event is made available for an application, can vary from the order of 1 second to 10 ms, depending on the application. The reliability

measures the probability of data loss in the network in terms of the packet error rate and can vary from  $10^{-5}$  to  $10^{-9}$ . Finally, the scale measures the number of devices that may be deployed without sacrificing the reliability and the latency, and can vary from tens to tens of thousands [16].

In [17], authors also discuss the concepts of Industrial IoT (IIoT) and Industry 4.0. The former refers to connecting all the industrial assets, including machines and control systems, with information systems and business processes. The latter is represented as the union of the IIoT paradigm with the employment of CyberPhysical Systems (CPSs), systems that extend real-world, physical objects by interconnecting them altogether and providing their digital description. Some potential applications of CPSs are energy systems (managing of energy consumption in smart buildings, power production and power distribution), healthcare (systems that continuously monitor patients and their medications), autonomous vehicles, smart traffic management, smart manufacturing, smart agriculture, etc [18]. The CPSs enable the interconnection of systems across distances in distributed applications and processes, in a reliably and low-cost way, allowing a remote management with data collection for analysis and optimization [18].

### 1.5 Non-Orthogonal Multiple Access and SIC Decoding

The previous generations of wireless communications systems were mostly based on the utilization of different Orthogonal Multiple Access (OMA) schemes: Frequency-Division Multiple Access (FDMA) for 1G, Time-Division Multiple Access (TDMA) for 2G, Code-Division Multiple Access (CDMA) for 3G and Orthogonal Frequency-Division Multiple Access (OFDMA) for 4G. In those schemes the users are allocated with radio resources which are orthogonal in time, frequency or code domain, and ideally no interference exists among them. However, one drawback of OMA schemes is that the maximum number of users is limited by the total amount of available orthogonal resources [19].

To meet the diverse requirements in terms of improved spectral efficiency, ultra-reliability, low latency and massive connectivity, Non-Orthogonal Multiple Access (NOMA) emerges as a promising technology for 5G and 6G networks. According to [19], NOMA allows controllable interference by non-orthogonal resource allocation with the tolerable increase in receiver complexity, and can be classified into two different categories: power and code domain multiplexing. In the former case, different users are allocated with distinct power levels according to their channel conditions, and they are separated at the receiver through the SIC decoding. In the latter case, different users are assigned distinct codes and then multiplexed over the same time-frequency resources.

NOMA was being considered by Third Generation Partnership Project (3GPP) to be included in the first releases of 5G, at least for the uplink, in addition to OMA. However, this study-item was discarded because the studies did not demonstrate significant performance gains of current NOMA techniques over other technologies such as Multi-User Multiple-Input Multiple-Output (MU-MIMO) systems. For this reason, 3GPP decided to leave NOMA for beyond-5G scenarios where new use cases with a very large numbers of users could motivate the use of NOMA [20]. The future 6G wireless systems, for example, are expected to support ever higher number of connected devices with much more stringent requirements in terms latency and reliability when compared to 5G [3].

Hence, the joint use of multiple techniques such as NOMA, MU-MIMO and interface diversity could become necessary.

It is expected that 6G will feature dynamic multiple access protocols that can dynamically switch between OMA and NOMA schemes, as well between random or scheduled access, depending on the need of each specific application and the current state of the network [12].

In [21], the authors introduce the concept of modern random access protocols for 6G IoT scenarios. Due to the massive number of nodes with sporadic transmission of short data packets, it is very difficult to implement an efficient resource allocation policy. By employing modern random access protocols, transmitters access the medium in an uncoordinated fashion to send multiple copies of their packets, and the receiver performs SIC decoding alongside other signal processing techniques to recover information.

## 1.6 Antenna Diversity

Another important technique used to enhance the performance of 5G systems is multiple antennas. Since the separation necessary to ensure independence between antennas decreases with the carrier frequency, for higher-frequencies, e.g. mmWave band, a massive number of antennas may be available, which increases the capability for beamforming. On the other hand, for the lower frequency bands, the number of antennas is typically low to moderate, e.g. up to 32 active antennas [22]. Nonetheless, the available bandwidth in the lower frequency bands is scarce, which may require the combination of multi-antenna techniques with other solutions to increase the number of connected users and the spectral efficiency, e.g. NOMA.

## 1.7 Related Works

Aiming to allow the three generic 5G services to coexist in the same RAN when the number of available radio resources is limited, Popovski et. al. [10] proposed an information-theoretic framework for the slicing of radio resources on the uplink using NOMA techniques. Similar to OMA opposed to NOMA, they proposed the concepts of orthogonal and non-orthogonal network slicing on the physical layer. Considering the orthogonal slicing, the three services are allocated on radio resources that are orthogonal in time and/or frequency. Conversely, the different 5G services share the same radio resources when employing the non-orthogonal slicing.

In [10], the authors studied two different scenarios: the slicing between eMBB and URLLC, and the slicing between eMBB and mMTC. By using orthogonal slicing, some frequency channels are allocated exclusively for the eMBB, while other frequency channels are allocated exclusively for the URLLC, that is, both services coexist in a FDMA manner. On the other hand, the same frequency channels are shared by eMBB and URLLC services with non-orthogonal slicing configuration. However, authors did not consider NOMA for multiple URLLC devices, such that the maximum number of URLLC devices connected to the same Base Station (BS) was limited by the number of minislots available in the timeslot. Regarding the specific case of network slicing between eMBB and mMTC, authors in [10] proposed a framework where an eMBB device and multiple

MTC devices share the same radio resource composed of one timeslot in a single frequency channel. In the case of orthogonal slicing, they coexist in a TDMA manner. Conversely, both services overlap their traffic in the same radio resource allocation when non-orthogonal slicing is used. In both cases, multiple MTC devices are allowed to transmit concurrently by means of NOMA, and the BS performs SIC to decode the multiple overlapping signals. Moreover, in [10], all the eMBB, URLLC and mMTC devices and the serving BS were equipped with only a single antenna.

The coexistence of eMBB and URLLC services has been extensively studied in the literature. Joint scheduling of eMBB and URLLC traffic has been studied in, for example, [23], [24] and [25]. In [24], authors studied the joint scheduling of eMBB and URLLC traffics by investigating different eMBB rate loss models associated with URLLC superposition/puncturing. The slicing of resources for eMBB and URLLC has been also studied in [26], where authors proposed a risk-sensitive based formulation to allocate resources to URLLC users and ensure their reliability while minimizing the risk of eMBB users having a low data rates. In [27], authors adopted a time/frequency resource blocks approach to address the sum rate maximization problem subject to latency and slicing isolation constraints while guaranteeing the reliability requirements with the use of adaptive modulation coding. In [11], authors analyze the coexistence of eMBB and URLLC in fog-radio architectures where the URLLC traffic is processed at the edge while eMBB traffic is handled at the cloud. In [28], authors also study the orthogonal and non-orthogonal slicing of radio resources for eMBB and URLLC using a max-matching diversity algorithm to allocate the frequency channels for the eMBB users.

The coexistence between eMBB and mMTC is also studied in [29], [30] and [31], but considering single antenna receivers. The uplink scenario where multiple MTC devices are allowed to communicate with one or multiple receivers using NOMA schemes has also been studied lately. In [32], authors studied the uplink mMTC in a large-scale cellular network overlaid with data aggregators using a stochastic geometry analytical framework. In [33], authors studied a multi-cell scenario with single cell BSs for a ultra-narrow band low power wide area network. They considered two different SIC mechanisms: SIC performed locally at each BS without information exchange between BSs, and SIC performed across multiple BSs where BSs can send decoded packets to neighboring cells. The performance of a LoRa network with multiple devices connected in the uplink with a single antenna BS is studied in [34]. Therein, the BS is allowed to perform a SIC decoding with one iteration to avoid packet losses due to collisions.

In the scope of multiple antenna receivers, Liu et. al. [35] studied the performance of a single-cell large scale MU-MIMO uplink system in terms of outage probability of three linear receivers: Maximum Ratio Combining (MRC), Zero-Forcing (ZF) and Minimum Mean Square Error (MMSE). In their model, all users were allowed to communicate with the BS simultaneously in the same time-frequency resource. However, they did not consider the coexistence of devices with heterogeneous performance requirements.

## 1.8 Main Contributions

Motivated by IIoT scenarios and based on recent works that addresses the coexistence between eMBB and URLLC and between eMBB and mMTC, in this thesis we develop

two extensions for the communication-theoretic framework proposed in [10]. Both contributions are based on our previous works [36, 37], and are summarized as follows:

- Following [36], we aim at increasing the number of URLLC devices that may be connected to a common BS in the uplink of 5G deployment scenarios, while still guaranteeing their performance requirements. To achieve this goal, we extend the setup of [10] for the network slicing between eMBB and URLLC to allow the NOMA for multiple URLLC devices, such that they can share the same radio resources with eMBB users in a scalable manner. To achieve this, our innovative approach relies on the joint utilization of three different techniques: frequency diversity, NOMA and SIC decoding at the BS. That is, we allow multiple URLLC users to transmit simultaneously in the same minislot and across multiple frequency channels and the serving BS employs SIC decoding to recover packets belonging to them and the coexisting eMBB users as well. To characterize the performance trade-offs between eMBB and URLLC, we evaluate the pairs of achievable sum rates under pre-defined reliability requirements in orthogonal and non-orthogonal scenarios. We show that, even with overlapping transmissions from multiple URLLC users, the use of frequency diversity, NOMA and SIC guarantee the reliability requirements of eMBB and URLLC services in both orthogonal and non-orthogonal slicing of radio resources;
- Following [37] and inspired by the recent results on MTC uplink scenarios with receive diversity (specially the setup from [35]), we extend the framework for the network slicing between eMBB and mMTC proposed in [10] to consider a BS equipped with multiple antennas. More specifically, we study the performance of orthogonal and non-orthogonal network slicing in a single-cell scenario where one eMBB device and multiple MTC devices communicate in the uplink with a multi-antenna BS. The BS utilizes an iterative MRC-SIC<sup>1</sup> receiver to decode multiple packets that arrive simultaneously. Differently from [35], we consider a scenario where heterogeneous devices transmit in the uplink. Besides, while in [10] the authors considered only a single-antenna BS, we evaluate the performance gains provided by multiple receive antennas operating using MRC. The performance is evaluated in terms of achievable data rates and number of connected MTC devices for given reliability requirements of both services. We show, through Monte Carlo simulations, that despite the space diversity reception improving the performance of both slicing schemes, the performance gains are more pronounced with the non-orthogonal slicing, which makes it more attractive than the orthogonal slicing when the BS is equipped with multiple antennas. Given a number of connected MTC devices, the advantage of non-orthogonal slicing over its orthogonal counterpart increases as the number of receive antennas increases. Moreover, non-orthogonal slicing allows us to improve significantly the number of MTC devices that can be connected to the BS as the number of receive antennas increases, for a given target mMTC data rate.

---

<sup>1</sup>The BS utilizes a MRC receiver based on the assumption that the number of MTC devices may be much larger than the number of receive antennas in mMTC scenarios.

## 1.9 Work Outline

This thesis is organized as follows. In Chapter 2, we present the theoretical background in wireless communications that is required for the understanding of the contributions of this thesis. Chapter 3 presents the 5G uplink scenario and network slicing strategies for the coexistence between eMBB and URLLC and between eMBB and mMTC. We also present the individual performance analysis of the eMBB, URLLC and mMTC services. In Chapter 4, we show how eMBB and URLLC devices can share the same RAN under orthogonal and non-orthogonal slicing, and then we present numerical results illustrating the performance trade-off between the services. We do the same for the coexistence scenario between eMBB and mMTC in Chapter 5. Finally, the conclusions and directions for future works are presented in Chapter 6.

## 2 PRELIMINARIES

In this section, we revisit the theoretical foundations of wireless communication systems which are necessary to carry out our investigations. The contents of this section are mostly based on the Chapters 5 and 6 of [38].

First we study the concepts of channel capacity and outage probability. We start the discussion with the simple Additive White Gaussian Channel (AWGN) model and then extend our investigations by incorporating fading channel as well. Next, we study how frequency and spatial reception diversity techniques can enhance the performance of such wireless communication systems. Finally, we present the concepts of multiuser capacity, NOMA and SIC decoding, which will be essential to study how an increasing number of uplink users can simultaneously connect to a common BS.

### 2.1 Capacity of the AWGN channel

The essential performance metric of a communication channel is its capacity: the maximum rate of communication for which an arbitrarily small error probability can be achieved.

The discrete time complex baseband model of the AWGN channel is

$$y[m] = x[m] + w[m], \quad (1)$$

where  $x[m] \in \mathbb{C}$  is the transmitted symbol,  $w[m] \sim \mathcal{CN}(0, N_0)$  is the AWGN with zero mean and variance  $N_0$ , and  $y[m] \in \mathbb{C}$  is the received symbol corrupted by noise. We assume a transmission power constraint  $\mathbb{E}\{|x[m]|^2\} \leq P$ .

The capacity of the AWGN channel in bits/s/Hz, also denominated as the maximum achievable spectral efficiency, is defined as

$$C_{\text{AWGN}} = \log_2(1 + \gamma), \quad (2)$$

where  $\gamma = \frac{P}{N_0 B}$  is the instantaneous Signal-to-Noise Ratio (SNR).

Fig. 1 shows the capacity of the AWGN channel versus the SNR  $\gamma$ . As we increase the SNR, the capacity is also increased. But the function is concave, which means that the higher the SNR is, the smaller is the corresponding increase on the capacity.

Let  $R$  (in bits/s/Hz) be the target data rate of a communication system. The concept of capacity says that, if the transmitter encodes data at rate  $R < C$ , the error probability of the system can be made arbitrarily small.

Now let us define  $\gamma_{\min}$  as the minimum SNR required to satisfy the target data rate  $R$ . Then we have

$$R = \log_2(1 + \gamma_{\min}). \quad (3)$$

Rewriting the above equation, we obtain the minimum SNR

$$\gamma_{\min} = 2^R - 1. \quad (4)$$

We define the outage probability of the system as the probability that the instantaneous SNR is less than the minimum SNR required to achieve the target data rate, that is,

$$\mathcal{P}_{\text{out}} = \Pr \{\gamma < \gamma_{\min}\} = \Pr \{\log_2(1 + \gamma) < R\}. \quad (5)$$

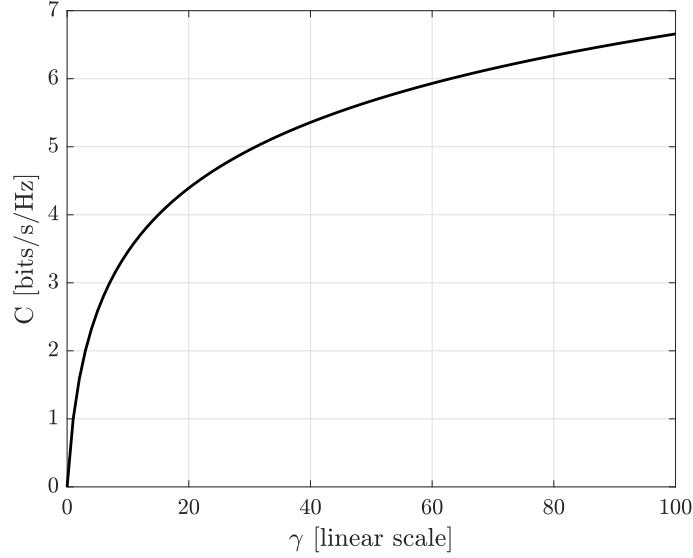


Figure 1. Spectral efficiency of the AWGN channel

## 2.2 Capacity of fading channels

The most basic representation of a wireless communication channel is the fading channel. Its discrete time complex baseband model is given by

$$y[m] = h[m]x[m] + w[m], \quad (6)$$

where  $h[m] \in \mathbb{C}$  yields the wireless channel coefficient, and  $x[m]$ ,  $w[m]$  and  $y[m]$  are the same as in (1). For normalization, we consider  $\mathbb{E}\{|h[m]|^2\} = 1$ . In this work, we assume that the channels coefficients are complex Gaussian distributed with zero mean and unit variance, that is,  $h[m] \sim \mathcal{CN}(0, 1)$ . As a result, the envelop  $|h[m]|$  is Rayleigh distributed, and the received power  $|h[m]|^2$  is exponentially distributed.

Herein we assume that the signal bandwidth is considerably less than the channel bandwidth, that is, the transmitted signal is narrowband. In other words, the fading is approximately equal across the entire signal bandwidth, such that the wireless channel can be characterized by a single filter tap. This assumption is known as the flat fading channel model [39, 38].

We assume that the receiver has perfect Channel State Information (CSI), while the transmitter has no CSI. The received SNR in this case is  $\gamma = \frac{|h|^2 P}{N_0}$ , where  $P$  is the power constraint and  $N_0$  is the variance of the AWGN samples. Moreover, we denote  $\bar{\gamma} = \frac{P}{N_0}$  as the average received SNR.

The maximum rate of reliable communication supported for a given realization of the fading channel is  $\log_2(1 + |h|^2 \bar{\gamma})$ . We assume that the transmitter encodes data at rate  $R$  bits/s/Hz. If the channel realization  $h$  is such that  $\log_2(1 + |h|^2 \bar{\gamma}) < R$ , the decoding error probability cannot be made arbitrarily small independently of the code used by the receiver, and the system is said to be in outage.

Thus, the outage probability of the flat fading channel reads

$$\mathcal{P}_{\text{out}} = \Pr \left\{ \log_2(1 + |h|^2 \bar{\gamma}) < R \right\}. \quad (7)$$

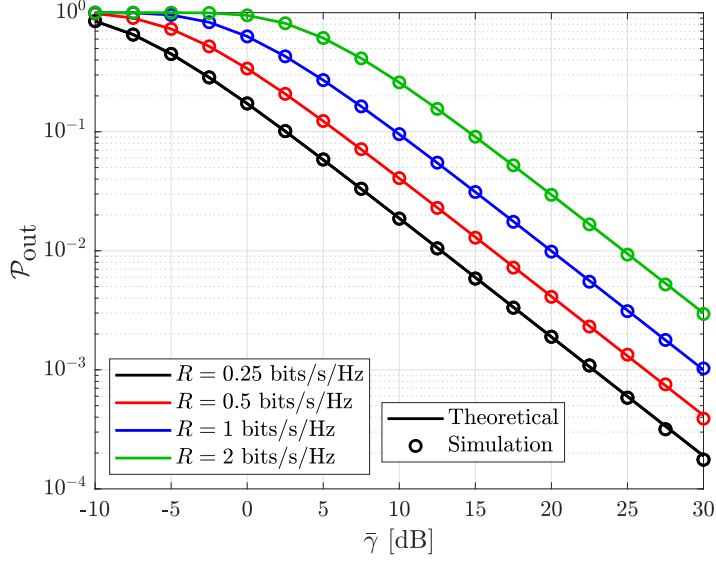


Figure 2. Outage probabilities of the flat fading channel versus average received SNR for increasing values of spectral efficiency.

As the channel gains are exponentially distributed, the outage probability is

$$\mathcal{P}_{\text{out}} = 1 - \exp\left(-\frac{2^R - 1}{\bar{\gamma}}\right). \quad (8)$$

Fig. 2 shows the outage probabilities of the flat fading channel versus the average SNR  $\bar{\gamma}$  for increasing values of the spectral efficiency  $R$ . The simulation results, which were computed using (7), match with the theoretical results computed using (8). As the target data rate  $R$  is increases, the outage probability becomes higher, as expected. This happens because it becomes less probable that the wireless channel will have favorable conditions to satisfy a more stringent data rate requirement. We also observe that, for a given value of  $R$  and in the high SNR regime, the outage probability decays like  $1/\bar{\gamma}$ .

### 2.3 Slow Fading Channel and $\epsilon$ -Outage Capacity

Another important concept for the study of wireless communication channels is the concept of slow fading, also known as quasi-static fading channel. Under this model, the channel coefficient is random but remains constant for all time, i.e.,  $h[m] = h \forall m$ .

In the case of an AWGN channel, the channel coding averages out the effect of the random AWGN noise, so one can send data at any rate  $R < C$  while making the error probability as small as desired. However, coding cannot average out the effect of channel fading, which affects all the coded symbols. Since the probability that the channel is in a deep fade is non-zero, the capacity of the slow fading channel in the strict sense is zero. An alternative performance measure is the  $\epsilon$ -outage capacity  $C_\epsilon$ , which is defined as the maximum achievable transmission rate such that the outage probability is less than  $\epsilon$ .

If we set  $\mathcal{P}_{\text{out}} = \epsilon$  in (7), we obtain the  $\epsilon$ -outage capacity as

$$C_\epsilon = \log_2(1 + \mathcal{F}^{-1}(1 - \epsilon)\bar{\gamma}), \quad (9)$$

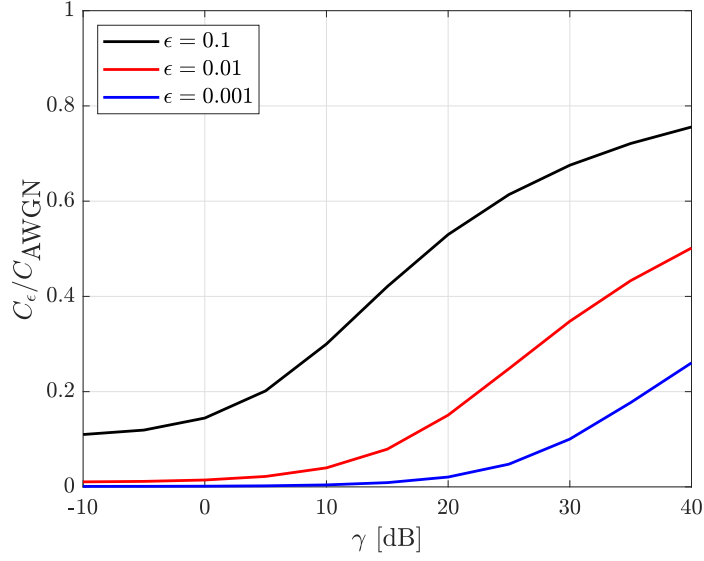


Figure 3. Ratio between the  $\epsilon$ -outage capacity and the capacity of the AWGN channel versus the SNR  $\gamma$  in dB, for different values of the outage probability requirement.

where  $\mathcal{F}$  is the complementary cumulative distribution function of  $|h|^2$ , that is  $\mathcal{F}(x) := \Pr\{|h|^2 > x\}$ . For Rayleigh fading, we have that  $\mathcal{F}^{-1}(1 - \epsilon) = -\ln(1 - \epsilon)$ .

Fig 3 shows the ratio between the  $\epsilon$ -outage capacity and the capacity of the AWGN channel versus the SNR  $\gamma$  in dB, for different values of the outage probability requirement. It is possible to observe that  $\epsilon$ -outage capacity is just a fraction of the capacity of the AWGN channel, and that the impact is much more significant in the low SNR regime.

In order to enhance the spectral efficiency of wireless communication systems under reliability requirements, we can adopt different techniques. In the following subsections we study how the frequency and spatial reception diversity can enhance the received SNR and thus the performance of the system.

## 2.4 Capacity of fading channels with frequency diversity

In this section, we study the performance of the fading channel when employing frequency diversity, i.e. the transmitter encodes data across a set of  $F$  parallel frequency channels. The analysis presented here is also valid for the case when the transmitter encodes data across a set of coherent periods, that is, when it explores the time diversity.

The parallel channel is defined as a collection of  $F$  subchannels, where each such subchannel is denoted by

$$y_f[m] = h_f[m]x_f[m] + w_f[m] \quad f = 1, \dots, F, \quad (10)$$

where  $h_f[m] \sim \mathcal{CN}(0, 1)$ .

The total power constraint of the parallel channel is given by  $FP$ , where  $P$  corresponds to the average power constraint per subchannel. Moreover, we assume that the transmitter has no CSI, and so it allocates equal powers  $P$  to each of the subchannels.

Given the channel realizations across the set of subchannels, the maximum rate of reliable communication in bits/s/Hz is

$$\sum_{f=1}^F \log_2(1 + |h_f|^2 \bar{\gamma}). \quad (11)$$

Given a target data rate  $R$  per subchannel, an outage event occurs when

$$\sum_{f=1}^F \log_2(1 + |h_f|^2 \bar{\gamma}) < FR. \quad (12)$$

Using a capacity achieving AWGN code with rate  $\log_2(1 + |h_f|^2 \bar{\gamma})$  for each of the subchannels, we obtain the average rate as

$$\frac{1}{F} \sum_{f=1}^F \log_2(1 + |h_f|^2 \bar{\gamma}). \quad (13)$$

Finally, the outage probability of the parallel channel is

$$\mathcal{P}_{\text{out}} = \Pr \left\{ \frac{1}{F} \sum_{f=1}^F \log_2(1 + |h_f|^2 \bar{\gamma}) < R \right\}. \quad (14)$$

Fig. 4 shows the outage probabilities versus the average received SNR for different numbers of frequency channels and  $R = 1$  bit/s/Hz. As can be seen from this figure, the higher the number of frequency channels, the better the system reliability becomes for a given SNR threshold. However, it is worth noticing that progressively allocating more frequency channels, gradually reduces its impact on the system reliability.

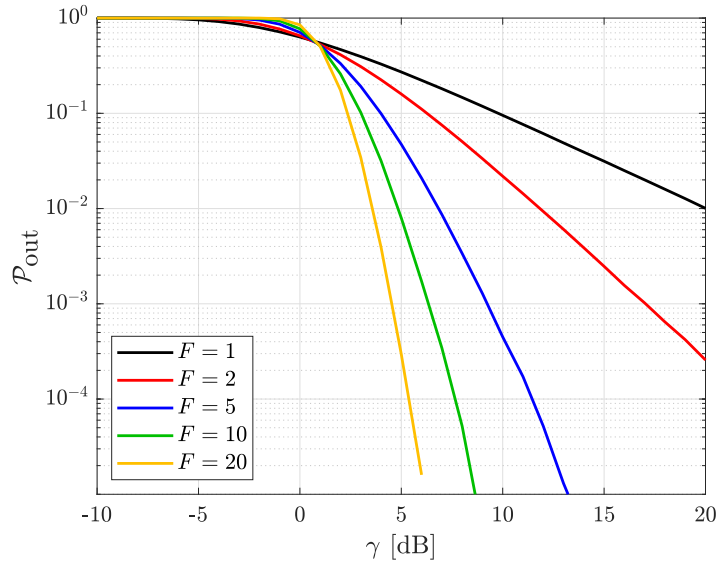


Figure 4. Outage probabilities versus the average received SNR for different numbers of frequency channels and for  $R = 1$  bit/s/Hz

## 2.5 Capacity of SIMO fading channel

Now we consider a Single-Input Multiple-Output (SIMO) system, that is, a system with one transmit antenna and  $L$  receive antennas. The baseband signal model for this system is

$$y_l[m] = h_l[m]x[m] + w_l[m] \quad l = 1, \dots, L, \quad (15)$$

where  $x[m] \in \mathbb{C}$  is the transmitted symbol,  $h_l \in \mathbb{C}$  is the wireless channel coefficient from the transmit antenna to the  $l$ -th receive antenna,  $w_l[m] \sim \mathcal{CN}(0, N_0)$  is the AWGN sample at the  $l$ -th receive antenna and  $y_l[m] \in \mathbb{C}$  is the received signal sample at the  $l$ -th receive antenna.

Herein, we assume that the receiver has perfect CSI, that is, the receiver knows all the channel gains between the transmitter each receive antenna, whereas the transmitter is assumed to not have CSI.

The detection of  $x[m]$  from  $y[m]$  is performed through a linear combination that maximizes the SNR, also called receive beamforming, and is given by

$$\hat{y}[m] = \mathbf{h}^H \mathbf{y}[m] = \|\mathbf{h}\|^2 x[m] + \mathbf{h}^H \mathbf{w}[m], \quad (16)$$

where  $\mathbf{h} = [h_1, \dots, h_L]^T$  is the vector of channel realizations between the transmit and the  $L$  receive antennas,  $\mathbf{w} = [w_1, \dots, w_L]^T$  is the vector of AWGN samples at the  $L$  receive antennas, and the superscript  $H$  denotes the conjugate transpose.

The capacity of the SIMO channel is

$$C_{\text{SIMO}} = \log_2(1 + \gamma), \quad (17)$$

where  $\gamma = \frac{\|\mathbf{h}\|^2 P}{N_0}$  is the received SNR. Comparing the SNR expressions of the AWGN channel and the SIMO channel, we observe that the use of multiple receive antennas increase the effective SNR and thus provide a power gain, which yields a significant increase in the capacity.

The outage probability of the SIMO fading channel is then given by

$$\mathcal{P}_{\text{out}} = \Pr \left\{ \log_2(1 + \|\mathbf{h}\|^2 \bar{\gamma}) < R \right\}. \quad (18)$$

We rewrite (18) as follows

$$\mathcal{P}_{\text{out}} = \Pr \left\{ \|\mathbf{h}\|^2 < \frac{2^R - 1}{\bar{\gamma}} \right\}. \quad (19)$$

Note that  $\|\mathbf{h}\|^2$  yields the sum of the squares of  $2L$  independent Gaussian random variables and follows a Chi-square distribution with  $2L$  degrees of freedom whose probability density function is

$$f_X(x) = \frac{1}{(L-1)!} x^{L-1} e^{-x}, \quad x \geq 0. \quad (20)$$

The outage probability of the SIMO fading channel is then given by [39, Eq. 7.17]

$$\mathcal{P}_{\text{out}} = 1 - \exp \left( -\frac{2^R - 1}{\bar{\gamma}} \right) \sum_{l=1}^L \left( \frac{[(2^R - 1)/\bar{\gamma}]^{l-1}}{(l-1)!} \right). \quad (21)$$

Fig. 5 presents the outage probabilities versus the average received SNR for increasing numbers of receive antennas and  $R = 1$  bit/s/Hz. Similar to the frequency diversity case in Fig. 4, we observe that the higher the number of receive antennas, the better the system reliability becomes for a given SNR threshold.

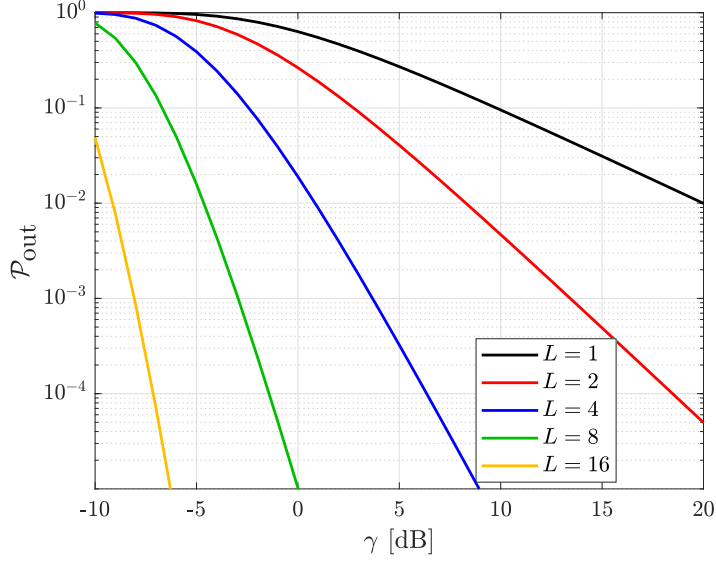


Figure 5. Outage probabilities versus the average received SNR for increasing numbers of receive antennas and  $R = 1$  bit/s/Hz

## 2.6 Multiuser Capacity, NOMA and SIC decoding

Now we study the NOMA uplink scenario where multiple single-antenna transmitters aim at transmitting independent packets to a common single-antenna receiver while sharing the same communication channel. We start the discussion with the simple AWGN channel, and then extend our formulation to the fading channel.

The baseband discrete-time model for the uplink AWGN channel with only two users is

$$y[m] = x_1[m] + x_2[m] + w[m], \quad (22)$$

where  $x_k[m] \in \mathbb{C}$  corresponds to the symbol transmitted by the  $k$ -th user,  $k \in \{1, 2\}$ ,  $P_k$  yields the respective power constraint, and  $w[m] \sim \mathcal{CN}(0, N_0)$  corresponds to the AWGN samples.

In the single-user case, the concept of capacity says that reliable communication can be attained at any rate  $R < C$ . For the multiuser case, this concept has to be extended to the concept of capacity region  $\mathcal{C}$ : the set of rate pairs  $(R_1, R_2)$  such that users 1 and 2 can simultaneously and reliably communicate at rates  $R_1$  and  $R_2$ , respectively. Since both users share the same radio resources, there is a natural tradeoff between them – if one wants to communicate at a higher rate, the other has to communicate at a lower rate. The capacity region  $\mathcal{C}$  characterizes the optimal tradeoff achievable by any multiple-access scheme.

Two other performance metrics of interest are the symmetric capacity

$$C_{\text{sym}} = \max_{(R_1, R_2) \in \mathcal{C}} R, \quad (23)$$

that is the maximum common rate at which both users can simultaneously and reliably communicate, and the sum capacity

$$C_{\text{sum}} = \max_{(R_1, R_2) \in \mathcal{C}} R_1 + R_2, \quad (24)$$

that is the maximum total throughput that can be achieved.

The capacity region  $\mathcal{C}$  of two users with AWGN channel, as illustrated on Fig. 6, must satisfy the following constraints:

$$R_1 < \log_2 \left( 1 + \frac{P_1}{N_0} \right) \quad (25)$$

$$R_2 < \log_2 \left( 1 + \frac{P_2}{N_0} \right) \quad (26)$$

$$R_1 + R_2 < \log_2 \left( 1 + \frac{P_1 + P_2}{N_0} \right) \quad (27)$$

The first two constraints say that the rate of each individual user cannot exceed the capacity of the single-user case when there is no multiuser interference. The third constraint says that the sum rate (total throughput) cannot exceed the capacity of the single user case with the sum of the received powers of the two users.

The most surprising fact about the capacity region showed in Fig. 6 is that user 1 can achieve its channel capacity while at the same time the user 2 can get a non-zero rate, which is indicated by point A in Fig. 6. Conversely, user 2 can also achieve its channel capacity while at the same time the user 1 can get a non-zero rate, which is indicated by the point B at the same figure.

Now we explain how this fact can be achieved. Each user encodes its own data using a capacity-achieving AWGN code. Then, the receiver decodes the information of both users in two stages. First, it decodes the data from user 2 treating the signal from user 1 as interference. Then the maximum rate achieved by user 2 is

$$R_2^* = \log_2 \left( 1 + \frac{P_2}{P_1 + N_0} \right). \quad (28)$$

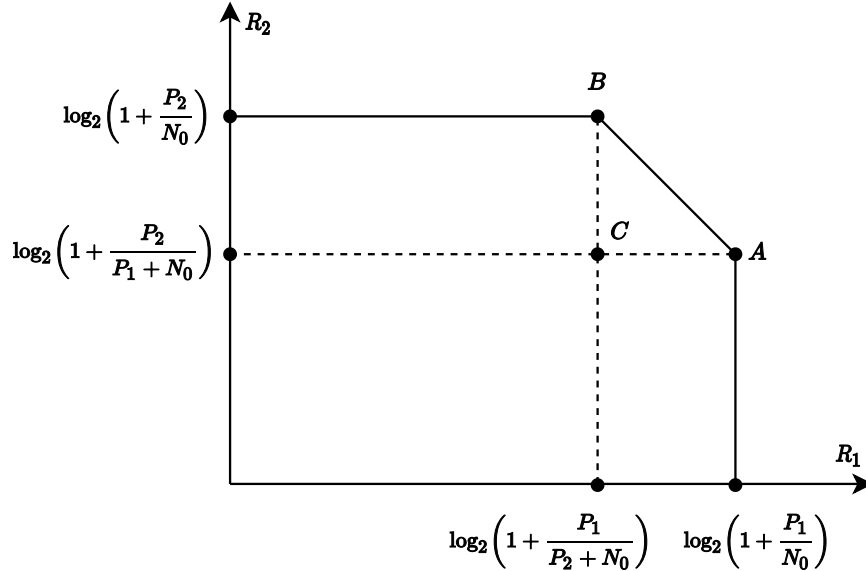


Figure 6. Capacity region  $\mathcal{C}$  of the two-user uplink AWGN channel (reproduced from [38]).

Once the receiver decodes the data from user 2, it can reconstruct its signal and then subtract it from the total received signal. Afterwards, only AWGN degrades the received signal. As a result, user 1, denoted as SIC receiver, achieves its capacity

$$R_1 = \log_2 \left( 1 + \frac{P_1}{N_0} \right). \quad (29)$$

If the decoding order is inverted, the point B on Fig. 6 is achieved.

## 2.7 Comparison between OMA and NOMA schemes

Considering the two-user uplink AWGN channel, we now compare the performance of OMA against NOMA by analyzing the capacity regions achieved by using both schemes. By employing OMA, a fraction  $\alpha$  of the radio resources (on time and/or frequency) is allocated to user 1, while the remaining  $1 - \alpha$  is assigned to user 2. The total energy received from user 1 is  $P_1/\alpha$ , thus the maximum rate user 1 can achieve (in bits/s/Hz) is

$$R_1 \leq \alpha \log_2 \left( 1 + \frac{P_1}{\alpha N_0} \right). \quad (30)$$

Similarly, the maximum rate user 2 can achieve (in bits/s/Hz) is

$$R_2 \leq (1 - \alpha) \log_2 \left( 1 + \frac{P_2}{(1 - \alpha)N_0} \right). \quad (31)$$

Notice that we obtain all the achievable OMA data rate pairs by varying alpha from 0 to 1.

Figs. 7 and 8 compare the capacity regions obtained by the OMA in NOMA schemes. In Fig. 7 we consider a scenario where  $P_1/N_0 = P_2/N_0 = 10$  dB, that is, the received powers from the two users are the same. In Fig. 8, the second user is assumed to be 10 dB stronger than the first one, i.e., we build a scenario where  $P_1/N_0 = 10$  dB and  $P_2/N_0 = 20$  dB.

From both figures, we observe that the NOMA scheme outperforms the OMA one for the most part, aside from one point where both curves intercept. This point corresponds to the case when  $\alpha = \frac{P_1}{P_1+P_2}$ , that is, when the amount of resources allocated to each user is proportional to their received power. In the case of Fig. 7, this means that both users will have the same data rates since they have the same received power. However, in the case of Fig. 8, the OMA scheme renders a highly unfair operating point wherein the user 2 experiences higher received power which in its turn achieves much higher rate than that of user 1.

In summary, using a NOMA scheme with SIC decoding, the stronger user is decoded first and the weak user is decoded next. In fact, this scheme allows the weakest user to achieve the highest possible rate and, as a result, to work at the fairest operating point. On the other hand, the OMA scheme can only achieve such performance for the weakest user at the cost of sacrificing the rate of the strongest user.

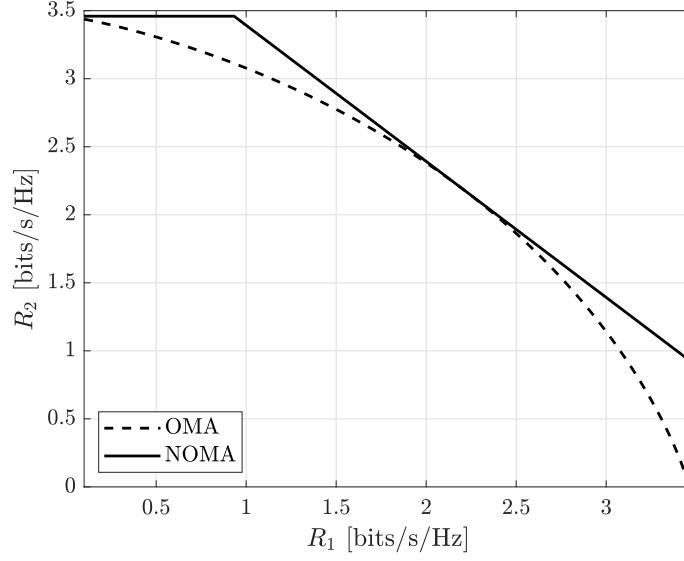


Figure 7. Performance comparison of OMA and NOMA schemes on the two user uplink AWGN channel, for  $P_1/N_0 = P_2/N_0 = 10$  dB.

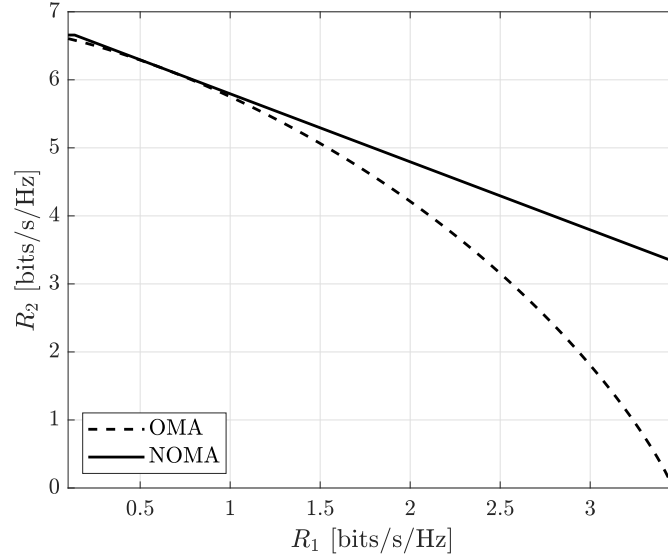


Figure 8. Performance comparison of OMA and NOMA schemes on the two user uplink AWGN channel, for  $P_1/N_0 = 10$  dB and  $P_2/N_0 = 20$  dB.

## 2.8 Extension to the $K$ -user capacity

In this section, we extend the results of the two-user uplink AWGN channel to a scenario where  $K$  users communicate simultaneously to a common receiver. The capacity region is now described by  $2^K - 1$  constraints, one for each possible non-empty subset  $\kappa \subset \{1, \dots, K\}$  of users:

$$\sum_{k \in \kappa} R_k < \log_2 \left( 1 + \frac{\sum_{k \in \kappa} P_k}{N_0} \right), \forall \kappa \subset \{1, \dots, K\}. \quad (32)$$

The sum capacity of the system (in bits/s/Hz) is

$$C_{\text{sum}} = \log_2 \left( 1 + \frac{\sum_{k \in \kappa} P_k}{N_0} \right). \quad (33)$$

When all the  $K$  users have the same received power, that is,  $P_k = P \forall k \in \{1, \dots, K\}$ , the sum capacity is simply

$$C_{\text{sum}} = \log_2 \left( 1 + \frac{KP}{N_0} \right). \quad (34)$$

Finally, the symmetric capacity if this case is

$$C_{\text{sym}} = \frac{1}{K} \log_2 \left( 1 + \frac{KP}{N_0} \right). \quad (35)$$

## 2.9 Uplink fading channel

We now include the effect of fading on the uplink channel. The discrete-time baseband representation of the  $K$ -user uplink fading channel is

$$y[m] = \sum_{k=1}^K h_k[m] x_k[m] + w[m], \quad (36)$$

where  $h_k[m]$  is the fading process of user  $k$ ,  $x_k[m] \in \mathbb{C}$  is the symbol transmitted by the user  $k$ , and  $w[k] \sim \mathcal{CN}(0, N_0)$  corresponds to AWGN samples. We assume that all the fading processes are independent and identically distributed and  $\mathbb{E}\{|h_k[m]|\} = 1$ . Besides, we also assume that all the  $K$  users are subject to the sample average power constraint  $P$ .

When each  $h_k[m]$  is a time-varying ergodic process (fast fading), the sum capacity is

$$C_{\text{sum}} = \mathbb{E} \left\{ \log_2 \left( 1 + \frac{\sum_{k=1}^K |h_k|^2 P}{N_0} \right) \right\}. \quad (37)$$

By using the Jensen's inequality to compare the uplink capacity of the fading channel against that of the AWGN channel, we obtain

$$\begin{aligned} \mathbb{E} \left\{ \log_2 \left( 1 + \frac{\sum_{k=1}^K |h_k|^2 P}{N_0} \right) \right\} &\leq \log_2 \left( 1 + \frac{\mathbb{E} \left\{ \sum_{k=1}^K |h_k|^2 \right\} P}{N_0} \right) \\ &= \log_2 \left( 1 + \frac{KP}{N_0} \right). \end{aligned} \quad (38)$$

From (38), the presence of fading always harms the performance of the channel. However, when the number of users becomes very large and by the law of large numbers,  $\frac{1}{K} \sum_{k=1}^K |h_k|^2 \rightarrow \mathbb{E}\{|h_k|^2\} = 1$  with probability 1, thus the effect of fading vanishes. In other words, the users with favorable channel conditions (and consequently high capacities) compensate for the users with poor channel conditions (and low channel capacities) in the total sum capacity.

Recall the study of NOMA with SIC decoding to understand how the effect of fading vanishes as the number of users connected to the uplink grows large. Consider the  $k$ -th

step of the SIC decoding procedure, where the user  $k$  is being decoded and the subsequent users  $\{k+1, \dots, K\}$  have not been decoded and canceled yet. The received signal at this step can be written as

$$y[m] = h_k[m]x_k[m] + \sum_{i=k+1}^K h_i[m]x_i[m] + w[m]. \quad (39)$$

The rate the  $k$ -th user then becomes

$$R_k = \mathbb{E} \left\{ \log_2 \left( 1 + \frac{|h_k|^2 P}{\sum_{i=k+1}^K |h_i|^2 P + N_0} \right) \right\}. \quad (40)$$

Since there are many users sharing the channel, the Signal-to-Interference-plus-Noise Ratio (SINR) of  $k$ -th user is small. However, resorting again to the law of large numbers, the summation in (40) corresponds to an averaging of the interference from other users, and, as a result, the fading effect vanishes. The users with favorable channel conditions generate high levels of interference, whereas the users with poor channel conditions generate low levels of interference. When we sum the interference contributions from all the users and assuming that the number of interfering users is large, the total interference seen by the  $k$ -th users become deterministic and is given by the total number of users times their transmit power. The rate of user  $k$  can then be approximated as

$$\begin{aligned} R_k &\approx \mathbb{E} \left\{ \frac{|h_k|^2 P}{\sum_{i=k+1}^K |h_i|^2 P + N_0} \right\} \log_2 e \\ &\approx \mathbb{E} \left\{ \frac{|h_k|^2 P}{(K-k)P + N_0} \right\} \log_2 e \\ &= \frac{P}{(K-k)P + N_0} \log_2 e, \end{aligned} \quad (41)$$

which represents the same rate that the user  $k$ -th would have achieved in the uplink AWGN channel. The first approximation in (41) comes from  $\log_2(1+x) \approx x$  for small  $x$ . The second approximation in (41) comes from the law of large numbers.

We conclude that a large number of users  $K$  in the uplink fading channel provides a spatial diversity that averages out the harmful effect of fading to the total sum capacity. As a result, the data rate seen by each individual user tends to the data rate corresponding to the uplink AWGN channel.

### 3 SYSTEM MODEL

In this chapter, we first establish the uplink scenario and the orthogonal and non-orthogonal network slicing strategies for the coexistence between eMBB and URLLC and between eMBB and mMTC. Then we present the performance analysis of the three 5G use cases when they are considered in isolation, i.e., in scenarios where they do not share the same RAN.

#### 3.1 Network Slicing for the Coexisting eMBB and URLLC Use Cases

The system model and network slicing strategies for coexisting eMBB and URLLC uses cases are based in the previous work [36]. We consider the uplink of 5G networks with multiple eMBB and URLLC devices independently communicating to a common BS as illustrated in Fig. 9. The eMBB and URLLC devices and the BS are equipped with a single-antenna. The radio resources to be shared between the two services consist of a time-frequency grid composed of  $F$  frequency channels indexed by  $f \in \{1, \dots, F\}$  and  $S$  minislots indexed by  $s \in \{1, \dots, S\}$ . The set of  $S$  minislots composes a timeslot. We define a resource block as the whole timeslot in a single frequency channel  $f$ .

The radio resource slicing schemes for the coexistence between eMBB and URLLC are based on [10] and illustrated in Fig. 10. We consider  $F = 10$  available frequency channels with  $F_U = 5$  frequency channels allocated for the URLLC traffic spanning  $S = 6$  minislots. Differently from [10], we allow more than one URLLC user to transmit within the same minislot, that is, NOMA for URLLC, as indicated by the darker blue resource blocks in Fig. 10.

The transmission of an eMBB user occupies a whole resource block. Moreover, assuming that radio access and competition among eMBB users have been solved prior to the considered time slot, we only model their standard scheduled transmission phase. A URLLC user, in turn, transmits within a single minislot across a subset of  $F_U \leq F$  frequency channels, as a means of achieving frequency diversity and meet the reliability requirements [10]. Owing to the low latency requirement, each URLLC packet must be decoded within the duration of a minislot and cannot span over multiple minislots. We

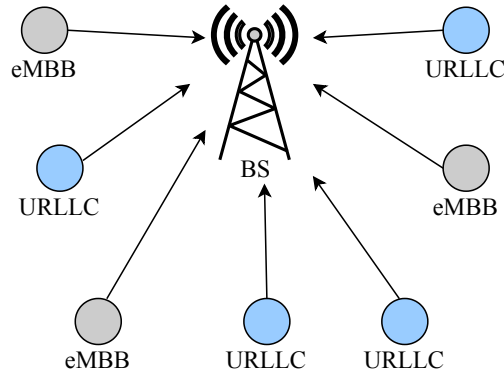


Figure 9. Uplink transmissions to a common base station (BS) from multiple eMBB and URLLC users. The gray circles represent the eMBB users, whereas the blue circles represent the URLLC users (reproduced from [36]).

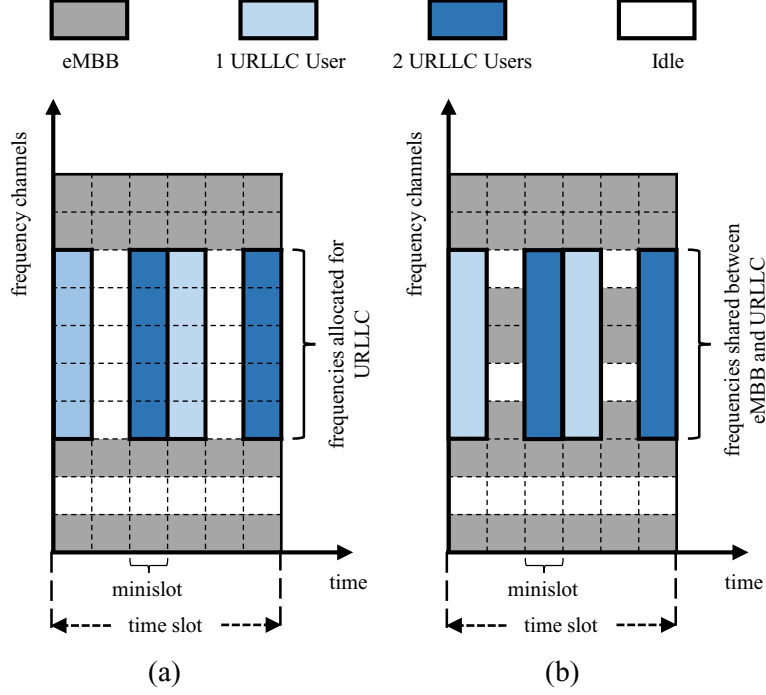


Figure 10. Illustration of the time-frequency grid used for the network slicing for the eMBB and URLLC services in the (a) orthogonal and (b) non-orthogonal scenarios. The darker blue tone indicates the overlap of URLLC transmissions (reproduced from [36]).

assume that a large number of URLLC users are connected to the same BS, but only a subset of them are active simultaneously in the same minislot. Each radio resource  $f$  is assumed to be within the time- and frequency-coherence interval of the wireless channel, so that the wireless channel coefficients are constant within each minislot, and they also fade independently across different minislots. The channel coefficients seen by the eMBB and URLLC devices at frequency channel  $f$ , which we denote by  $h_{i,f}$  with  $i \in \{B, U\}$ , are Independent and Identically Distributed (IID), and follow a complex Gaussian distribution with unit variance, that is,  $h_{i,f} \sim \mathcal{CN}(0, 1)$ . As a result, the channel gains  $g_{i,f} = \Gamma_i |h_{i,f}|^2$  affecting by the eMBB and URLLC users at frequency channel  $f$  are exponentially distributed with average  $\Gamma_i$ .

The average transmission power of all devices is normalized to one. We also assume that the variance  $N_0$  of the noise samples at the BS is normalized to one, such that the average received SNR  $\bar{\gamma} = \Gamma_i / N_0$  equals the average channel gain  $\Gamma_i$ . Moreover, no CSI is assumed at the URLLC devices, whereas the eMBB devices and BS are assumed to have perfect CSI as in [10]. The outage probabilities of the eMBB and URLLC devices are denoted by  $\Pr(E_B)$  and  $\Pr(E_U)$ , respectively, and must satisfy the reliability requirements  $\Pr(E_B) \leq \epsilon_B$  and  $\Pr(E_U) \leq \epsilon_U$ .

Let us denote by  $n_U \in \{\mathbb{N}^+\}$  the number of URLLC devices transmitting simultaneously in the same minislot. For the orthogonal slicing, the baseband signal received at the serving BS in minislot  $s$  and frequency channel  $f$  is

$$y_f^{\text{ort}}[s] = \begin{cases} h_{B,f}[s]x_{B,f}[s] + w_f[s], & \text{if } f \text{ is allocated for eMBB,} \\ \sum_{u=1}^{n_U} h_{U_u,f}[s]x_{U_u,f}[s] + w_f[s], & \text{if } f \text{ is allocated for URLLC,} \end{cases} \quad (42)$$

where  $h_{B,f}[s]$  is the wireless channel coefficient seen by the eMBB user in frequency channel  $f$ ,  $x_{B,f}[s] \in \mathbb{C}$  is the symbol transmitted by the eMBB user in frequency channel  $f$ ,  $w_f[s] \sim \mathcal{CN}(0, 1)$  is the AWGN sample at the receive antenna,  $h_{U_u,f}[s]$  is the wireless channel coefficient seen by the  $u$ -th URLLC device in frequency channel  $f$  and  $x_{U_u,f}[s] \in \mathbb{C}$  is the symbol transmitted by the  $u$ -th URLLC device in frequency channel  $f$ .

For the non-orthogonal slicing, the baseband signal vector received at the BS in the minislot  $s$  and frequency channel  $f$  is

$$y_f^{\text{non}}[s] = h_{B,f}[s]x_{B,f}[s] + \sum_{u=1}^{n_U} h_{U_u,f}[s]x_{u,f}[s] + w_f[s] \quad (43)$$

where all terms are defined as in (42). In this work, we study the performance of the worst case scenario where there is always an eMBB user transmitting in each frequency channel  $f$  and  $n_U$  URLLC users active in all minislots.

### 3.2 Network Slicing for Coexisting eMBB and mMTC Use Cases

The system model and network slicing strategies for coexisting eMBB and mMTC use cases are based in the previous work [37]. We assess the uplink performance of 5G networks where a single eMBB device and multiple MTC devices transmit independent packets to a common BS as illustrated in Fig. 11. Both eMBB and mMTC devices use single antenna, whereas the BS is equipped with  $L$  receive antennas, indexed by  $l \in \{1, \dots, L\}$ .

The eMBB and  $M$  MTC devices share the same radio resource under orthogonal or non-orthogonal slicing schemes as illustrated in Fig. 12. By using the orthogonal slicing, a fraction  $\alpha$  of the radio resource is allocated exclusively to the mMTC traffic, while the remaining part is allocated exclusively to the eMBB traffic. The orthogonal slicing means that the eMBB and MTC devices share the channel in a TDMA manner. On

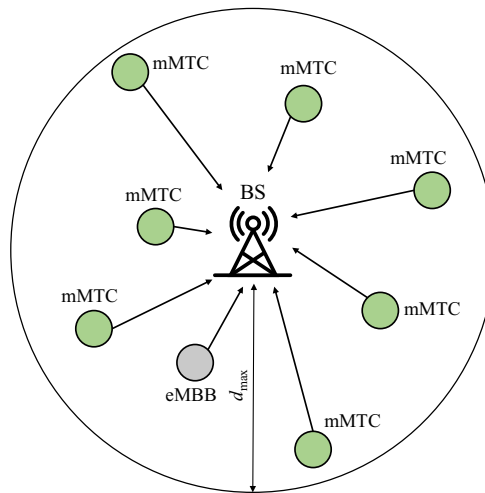


Figure 11. The uplink of a 5G network where an eMBB and multiple MTC devices are connected to a common BS. The gray circles represent the eMBB devices, whereas the green circles represent the MTC devices (reproduced from [37]).

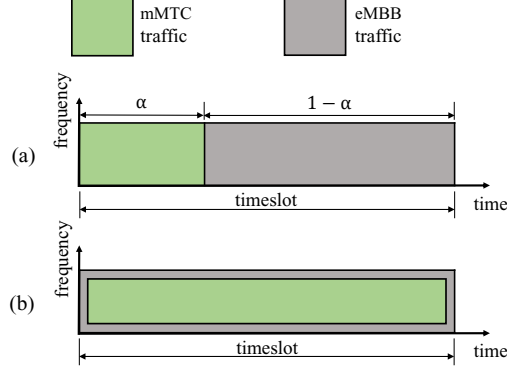


Figure 12. Orthogonal (a) and non-orthogonal (b) slicing of radio resources between eMBB and mMTC (reproduced from [37]).

the other hand, the whole radio resource is allocated to eMBB and mMTC under the non-orthogonal slicing, thus the traffic of both services overlap during the whole timeslot.

As in [10], we assume a standard scheduled transmission phase for the eMBB traffic, where the scheduling of the eMBB device has been solved prior to the considered timeslot. The frequency channel  $f$  is assumed to be within the time- and frequency-coherence interval, so that the wireless channel coefficients are constant within each timeslot and also fade independently across different timeslots. The channel gains of the eMBB and MTC devices at the receive antenna  $l$ ,  $g_{i,l}$  with  $i \in \{B, M\}$ , are IID and follow a zero-mean complex Gaussian distribution with variance  $\Gamma_i$ , i.e., Rayleigh fading. In other words,  $g_{i,l} \sim \mathcal{CN}(0, \Gamma_i)$ , where  $\Gamma_i$  is the average channel gain.

Let us denote by  $\mathbf{g}_i = [g_{i,1}, \dots, g_{i,L}]^T$  the vector of the wireless channel gains between eMBB or MTC device and the serving BS at the  $L$  receive antennas. In the case of interference-free transmissions, the received SNR obtained after applying MRC is given by

$$\gamma_i = \|\mathbf{g}_i\|^2. \quad (44)$$

In our model, we assume that the average transmit power of all devices is normalized to one, and that differences in the actual transmit power of devices and path loss exponents are accounted for in the different channel gains experienced by each device. Moreover, we also consider that the noise power at the receiver is normalized to one, such that the received power equals the SNR for all the devices.

No CSI is assumed at the MTC devices, whereas the eMBB device and BS are assumed to have perfect CSI as in [10]. As a result, the eMBB device can adapt its transmit power according to the channel conditions such that its achievable data rate equals a predefined value. Since the MTC devices operate without CSI, they all transmit with the same fixed data rate.

The outage probabilities of the eMBB and mMTC services are denoted as  $\Pr(E_B)$  and  $\Pr(E_M)$ , respectively, and must satisfy the reliability requirements  $\Pr(E_B) \leq \epsilon_B$  and  $\Pr(E_M) \leq \epsilon_M$ .

### 3.3 Analysis of the eMBB Performance

In this section, we extend the performance analysis of eMBB presented in [10] to a scenario where the BS is equipped with multiple receive antennas. This extension is based on the previous work [37].

The eMBB device adapts its transmit power  $P_B(\gamma_B)$  according to the instantaneous channel gains such that the received SNR always equals a predefined value. Following [10], the eMBB device aims to transmit at the largest rate  $r_B$  that is compatible with the outage probability requirement  $\epsilon_B$  under a long-term average power constraint<sup>1</sup>. This can be formulated as the following optimization problem

$$\begin{aligned} & \text{maximize } r_B \\ & \text{subject to } \Pr \{ \log_2[1 + P_B(\gamma_B)\gamma_B] \leq r_B \} \leq \epsilon_B \\ & \quad \text{and } \mathbb{E}[P_B(\gamma_B)] = 1. \end{aligned} \quad (45)$$

The optimal solution to this problem is given by the truncated power inversion scheme: the eMBB device chooses a transmit power that is inversely proportional to the received SNR  $\gamma_B$  if the latter is above a given threshold  $\gamma_B^{\min}$ , while it refrains from transmitting otherwise [10]. Thus, the activation probability of the eMBB device can be written as [39, Eq. 7.17]

$$\begin{aligned} a_B &= \Pr \{ \gamma_B \geq \gamma_B^{\min} \} \\ &= \exp \left( -\frac{\gamma_B^{\min}}{\Gamma_B} \right) \sum_{l=1}^L \frac{(\gamma_B^{\min}/\Gamma_B)^{l-1}}{(l-1)!} \\ &= \frac{\Gamma(L, \gamma_B^{\min}/\Gamma_B)}{(L-1)!}, \end{aligned} \quad (46)$$

where  $\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t} dt$  is the upper incomplete gamma function.

In the absence of interference from the mMTC traffic, the only source of outage for an eMBB transmission is the failed transmission event because of extremely poor channel conditions. In this case, the outage probability of the eMBB device can be written as

$$\Pr(E_B) = \Pr \{ \gamma_B < \gamma_B^{\min} \} = 1 - a_B, \quad (47)$$

where  $a_B$  is given by (46).

Imposing the reliability requirement  $\Pr(E_B) = \epsilon_B$  on (47), we obtain the threshold SNR as

$$\gamma_B^{\min} = \Gamma_B \gamma^{-1}(L, \epsilon_B(L-1)!), \quad (48)$$

where  $\gamma(a, z) = \int_0^z t^{a-1} e^{-t} dt$  is the lower incomplete gamma function. Moreover, in the case of  $L = 1$ , (48) reduces to

$$\gamma_B^{\min} \Big|_{L=1} = \Gamma_B \ln \left( \frac{1}{1 - \epsilon_B} \right). \quad (49)$$

---

<sup>1</sup>Notice we also assume eMBB users to have full CSI as in [10]. Since eMBB transmissions are scheduled, devices have sufficient time to undergo CSI acquisition procedures [10, 25].

Based on the truncated power inversion scheme, the instantaneous power  $P_B(\gamma_B)$  chosen as a function of the received SNR  $\gamma_B$  is

$$P_B(\gamma_B) = \begin{cases} \frac{\gamma_B^{\text{tar}}}{\gamma_B} & \text{if } \gamma_B \geq \gamma_B^{\min} \\ 0 & \text{if } \gamma_B < \gamma_B^{\min} \end{cases}, \quad (50)$$

where  $\gamma_B^{\text{tar}}$  is the target SNR, which is obtained by imposing the average power constraint as [10]

$$\begin{aligned} \mathbb{E}[P_B(\gamma_B)] &= \int_{\gamma_B^{\min}}^{\infty} \frac{\gamma_B^{L-1} e^{-\gamma_B/\Gamma_B}}{\Gamma_B^L (L-1)!} P_B(\gamma) d\gamma \\ &= \frac{\gamma_B^{\text{tar}}}{\Gamma_B (L-1)!} \Gamma\left(L-1, \frac{\gamma_B^{\min}}{\Gamma_B}\right) = 1. \end{aligned} \quad (51)$$

This implies that the target SNR is

$$\gamma_B^{\text{tar}} = \frac{\Gamma_B (L-1)!}{\Gamma\left(L-1, \frac{\gamma_B^{\min}}{\Gamma_B}\right)}. \quad (52)$$

In the case of  $L = 1$ , the target SNR becomes

$$\gamma_B^{\text{tar}}|_{L=1} = \frac{\Gamma_B}{\Gamma\left(0, \frac{\gamma_B^{\min}}{\Gamma_B}\right)}. \quad (53)$$

Finally, the outage rate experienced by the eMBB device is [10]

$$r_B^{\text{out}} = \log_2(1 + \gamma_B^{\text{tar}}). \quad (54)$$

Fig. 13 presents the outage rates experienced by the eMBB device as a function of the number of receive antennas at the BS for different values of the average received SNR and  $\epsilon_B = 10^{-3}$ . We observe that increasing the average received SNR yields significant increases on the achievable outage rates in the whole range of  $L$ . However, the benefits of increasing the number of receive antennas at the BS are more significant only when the number of antennas is low to moderate.

Notice that the outage probability of the eMBB user is uniquely determined by imposing the reliability requirement  $\epsilon_B$ . As a result, in the orthogonal slicing between eMBB and either URLLC or mMTC, we assume that all the eMBB users transmit with the outage rate  $r_B^{\text{orth}}$ . In the case of non-orthogonal slicing, if we impose, for example,  $\epsilon_B = 10^{-3}$ , we have  $a_B = 0.999$  and so very close to one, thus we conservatively assume that the interference from eMBB is always seen by either URLLC or mMTC devices.

### 3.4 Analysis of the URLLC Performance

We present in this section the performance analysis of URLLC based on [36]. The URLLC devices and the serving BS are equipped with a single antenna. The URLLC user transmits across the  $F_U$  frequency channels. Moreover, due to the NOMA behavior

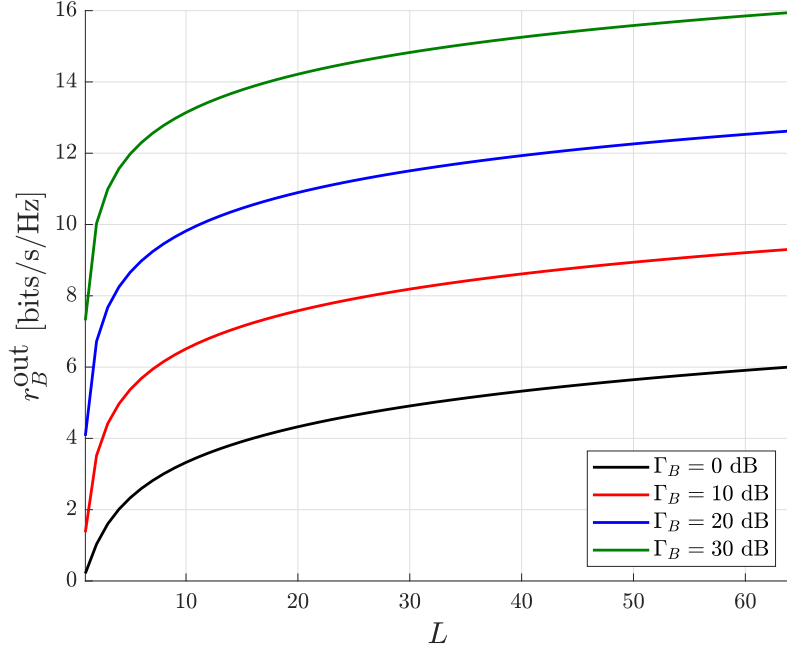


Figure 13. Outage rates achieved by the eMBB device as a function of the number of receive antennas at the BS, and for increasing values of the average received SNR

of URLLC devices, there is always multi-access interference when there is more than one URLLC device connected to the serving BS, that is,  $n_U > 1$ . Herein, without lack of generality and to simplify our mathematical treatment, we consider that all of the URLLC devices transmit with the same data rate  $r_U$  since they experience the same channel conditions and do not have any CSI.

As aforementioned, we consider the worst case assumption where there are always  $n_U \in \{\mathbb{Z}^+\}$  URLLC devices transmitting in the same minislots. The BS combines the signals received in all  $F_U$  frequency channels and then performs SIC<sup>2</sup> to decode the multiple URLLC packets that arrive in the same minislot. Let  $u \in \{1, \dots, n_U\}$  index an URLLC device with channel gain  $G_{U,u,f}$  for the allocated frequency channel  $f$ , and assume (without loss of generality) a SIC decoding ordering  $\{U_1, \dots, U_{n_U}\}$ . The SINR for the URLLC user with index  $u$  in the frequency channel  $f$  is

$$\sigma_{u,f} = \frac{g_{U,u,f}}{1 + \sum_{j=u+1}^{n_U} g_{U,j,f}}. \quad (55)$$

The outage probability for the URLLC user with index  $u$  is

$$\Pr(E_U) = \Pr \left\{ \frac{1}{F_U} \sum_{f=1}^{F_U} \log_2(1 + \sigma_{u,f}) < r_U \right\}, \quad (56)$$

<sup>2</sup>Note that, as presented in [10], SIC outperforms other multi-user detection techniques, such as puncturing.

where  $r_U$  is the data rate in bits/s/Hz. In other words, the URLLC device is decoded successfully if

$$\frac{1}{F_U} \sum_{f=1}^{F_U} \log_2(1 + \sigma_{u,f}) \geq r_U. \quad (57)$$

During the SIC decoding procedure, the BS first attempts to decode the strongest user among all the active URLLC users in a minislot. If this user is correctly decoded, its interference is subtracted from the received signal, then the BS attempts to decode the next user in the descending order, and so on. The SIC decoding procedure ends when the decoding of one active URLLC user fails or after all the active URLLC users have been correctly decoded. We assume that all SIC decoding steps can be realized within the time duration of a minislot.

We simulate the scenario where the serving BS combines and decodes the signals from the  $n_U$  URLLC users. To do that, we define the SIC ordering according to the sum of mutual information of the URLLC users in all frequency channels  $F_U$ . For the URLLC  $u$ -th user, this sum is given by

$$I_u^{\text{sum}} = \sum_{f=1}^{F_U} \log_2(1 + \sigma_{u,f}). \quad (58)$$

First we compute the sum of mutual information for all the  $n_U$  URLLC users transmitting simultaneously, and then decode the URLLC user with  $\max_n I_n^{\text{sum}}$ . If the decoding fails, the SIC decoding procedure ends. Otherwise, we compute again the sum of mutual information for the remaining  $n_U - 1$  URLLC users, but now without the interference from the device previously decoded, and then the procedure continues. The SIC decoding ends when one decoding step fail or after all the URLLC users have been correctly decoded.

Fig. 14 depicts the outage probabilities versus  $r_U$  for  $n_U = 2$ ,  $\Gamma_U = 10$  dB and different values of  $F_U$ . For comparison, we also plot the outage probability of the OMA scenario where there is only one URLLC device allocated in a single frequency channel. Similarly, Fig. 15 shows outage probability results for the case  $n_U = 4$ .

By comparing Figs. 14 and 15, we observe that, for a given data rate  $r_U$ , the outage probability increases with the number of URLLC users due to the higher interference levels. Conversely, by fixing the outage probability with few connected URLLC users, it is possible to achieve higher data rates. In both figures we also observe that, for the higher range of values of  $r_U$  ( $r_U > 1.2$  bits/s/Hz in Fig. 14 and  $r_U > 0.6$  bits/s/Hz in Fig. 15), the OMA scheme outperforms any of the NOMA schemes. This result was expected since, for high values of the data rate, the multiple-access interference degrades the achievable SINR for the decoding of the URLLC signals. Interestingly, in this higher range of  $r_U$ , the higher the number of frequency channels allocated to the URLLC traffic, the worse the system reliability becomes. On the other hand, for lower range of values of the data rate ( $r_U < 1.2$  in Fig. 14 and  $r_U < 0.5$  in Fig. 15), the allocation of more frequency channels for URLLC increases substantially the reliability. It is also worth noticing that the performance of the NOMA scheme with only one frequency channel equals that of OMA on the lower range of  $r_U$ . Nevertheless, the improvements in the system reliability are more significant when the number of frequency channels is low, that is, when we increase  $F_U$  from 1 to 2 and then to 5. In the case of  $n_U = 4$  (Fig. 15),

the performance of the NOMA with  $F_U = 10$  equals the one achieved with  $F_U = 5$  for  $r_U < 0.5$  bits/s/Hz, and is even worse for  $r_U > 0.5$  bits/s/Hz.

Given the reliability requirement  $\Pr(E_U) \leq \epsilon_U$ , we aim to obtain the maximum rate  $r_U$  as a function of the number of frequency channels allocated for URLLC traffic. The URLLC sum rate, which corresponds to the sum of the data rates of the  $n_U$  active URLLC devices transmitting in a minislot, is given by

$$r_U^{\text{sum}} = n_U r_U. \quad (59)$$

Increasing  $F_U$  enhances the frequency diversity and, hence, makes it possible to satisfy the reliability target  $\epsilon_U$  at a larger rate  $r_U$  [10].

### 3.5 Analysis of the mMTC Performance

In this section, we present the performance analysis of mMTC based on [37]. We assume that  $M$  MTC devices are connected to the serving BS. From [35] and considering the absence of interference from the eMBB traffic, the  $L \times 1$  baseband received vector at the BS is given by

$$\mathbf{y} = \sqrt{P_M} \mathbf{G}_M \mathbf{x}_M + \mathbf{w}, \quad (60)$$

where  $\mathbf{G}_M \in \mathbb{C}^{L \times M}$  is the matrix of channel gains between the MTC devices and the serving BS,  $\sqrt{P_M} \mathbf{x}_M \in \mathbb{C}^{M \times 1}$  is the vector of symbols transmitted by the MTC devices, and  $\mathbf{w} \in \mathbb{C}^{L \times 1}$  is the vector of AWGN samples with zero mean and unit variance. The  $m$ -th element of  $\mathbf{x}_M$ ,  $x_m$ , is zero if the  $m$ -th MTC device is inactive in the timeslot.

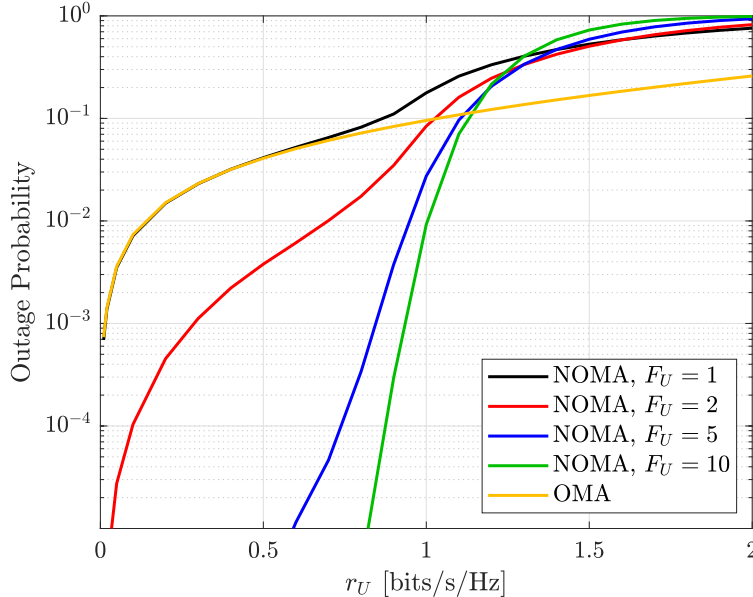


Figure 14. URLLC outage probabilities versus the data rate  $r_U$  for the NOMA and OMA scenarios. We consider  $n_U = 2$ ,  $\Gamma_U = 10$  dB and different values of  $F_U$  for the NOMA.

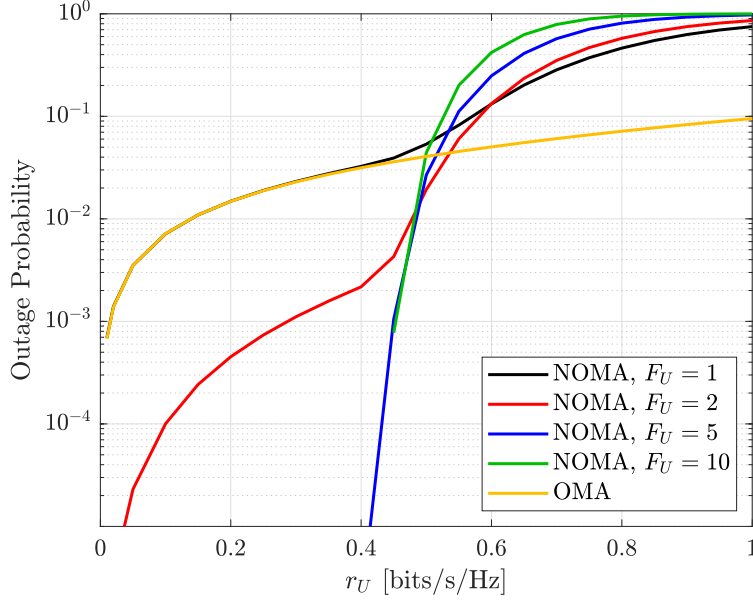


Figure 15. URLLC outage probabilities versus the data rate  $r_U$  for the NOMA and OMA scenarios. We consider  $n_U = 4$ ,  $\Gamma_U = 10$  dB and different values of  $F_U$  for the NOMA.

By exploiting the perfect CSI, the BS utilizes a MRC-SIC iterative receiver to decode the signals from the multiple MTC devices that arrive in the same timeslot. Following [35], the received signal vector after the MRC processing is given by

$$\begin{aligned}\hat{\mathbf{x}} &= \mathbf{G}_M^H \mathbf{y} \\ &= \sqrt{P_M} \mathbf{G}_M^H \mathbf{G}_M \mathbf{x}_M + \mathbf{G}_M^H \mathbf{w},\end{aligned}\tag{61}$$

where  $\hat{\mathbf{x}} \in \mathbb{C}^{M \times 1}$  and the superscript  $H$  indicates the conjugate transpose of the matrix  $\mathbf{G}_M$ .

Let  $\hat{x}_m$  denote the  $m$ -th element of the vector  $\hat{\mathbf{x}}$ , which corresponds to the signal transmitted by the  $m$ -th MTC device. As in [35], we have

$$\hat{x}_m = \sqrt{P_M} \mathbf{g}_m^H \mathbf{g}_m x_m + \sqrt{P_M} \mathbf{g}_m^H \sum_{m' \neq m}^M \mathbf{g}_{m'} x_{m'} + \mathbf{g}_m^H \mathbf{w},\tag{62}$$

where  $\mathbf{g}_m \in \mathbb{C}^{L \times 1}$  denotes the  $m$ -th column of the matrix  $\mathbf{G}_M$ . The first term in (62) represents the signal transmitted by the  $m$ -th MTC device, while the remaining terms correspond to the interfering signals from other MTC devices and the noise.

Since MTC devices operate without CSI, they all transmit with the same power and have the same target data rate  $r_M$ . During the MRC-SIC decoding, first the BS detects the strongest MTC device among the active ones, decodes its signal, subtracts its interference component from the received signal, then proceeds to the second strongest MTC device, and so on. The decoding procedure ends when the decoding of one MTC device fails or after all the active devices are correctly decoded.

The SIC decoding ordering is defined in the descending order of received SNRs of the active MTC devices. Let us denote a SIC decoding ordering by  $\{1, \dots, M\}$ , such that

$$\mathbf{g}_1^H \mathbf{g}_1 \geq \mathbf{g}_2^H \mathbf{g}_2 \geq \dots \geq \mathbf{g}_M^H \mathbf{g}_M.$$

The SINR while decoding the signal from the  $m$ -th MTC device, and assuming that the  $\{1, \dots, m-1\}$ -th MTC devices have already been correctly decoded, reads

$$\sigma_m = \frac{P_M \|\mathbf{g}_m\|^4}{P_M \sum_{m'=m+1}^M |\mathbf{g}_m^H \mathbf{g}_{m'}|^2 + \|\mathbf{g}_m\|^2}. \quad (63)$$

The outage probability for the  $m$ -th MTC device is

$$\Pr(E_M) = \Pr \{ \log_2(1 + \sigma_m) < r_M \}, \quad (64)$$

where  $r_M$  is the data rate in bits/s/Hz. In other words, the  $m$ -th MTC device is correctly decoded if the inequality  $\log_2(1 + \sigma_m) \geq r_M$  holds.

The mMTC outage probability must satisfy a reliability requirement  $\epsilon_M$ , that is,  $\Pr(E_M) \leq \epsilon_M$ . In Fig. 16, we set  $\epsilon_M = 0.1$ ,  $M = 10$  and compute the maximum achievable rate  $r_M$  (in bits/s/Hz) for different values of the average received SNR  $\Gamma_M$ . We observe that as we increase the number of receive antennas on the serving BS, we significantly increase the maximum achievable data rates. On the other hand, since all the mMTC devices are under the same channel conditions, increasing  $\Gamma_M$  does not yield significant improvements on the achievable data rates. The reason is that when the channel conditions of the MTC devices are improved due, for example, close proximity to the BS or higher transmit power, we are also increasing the levels of interference seen by each device.

In Fig. 17 we plot the outage probabilities of the mMTC versus data rate  $r_M$  for  $M = 10$ ,  $\Gamma_M = 0$  dB and different numbers of receive antennas at the BS. We can see in this figure that by increasing the number of receive antennas at the BS we significantly improve the reliability of the system because, fixing a data rate, we achieve much lower

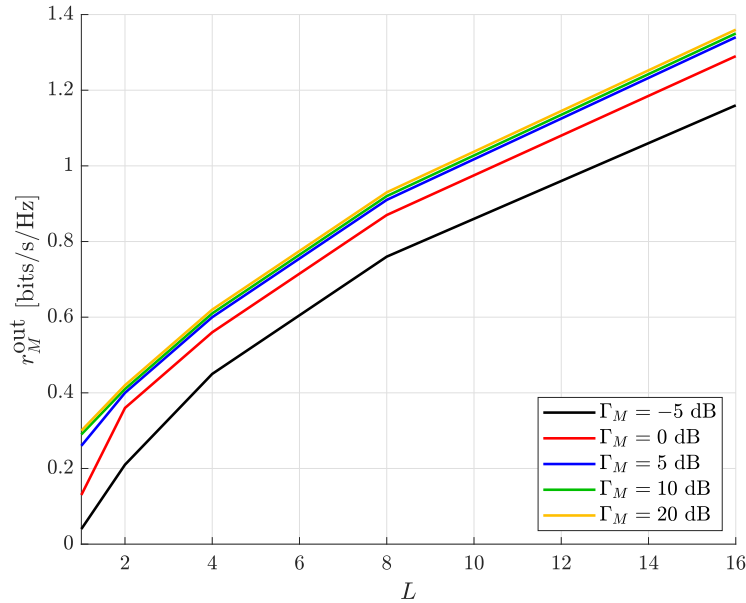


Figure 16. Outage rates of the mMTC devices versus number of antennas at the BS for  $\epsilon_M = 0.1$ ,  $M = 10$  and different values of  $\Gamma_M$ .

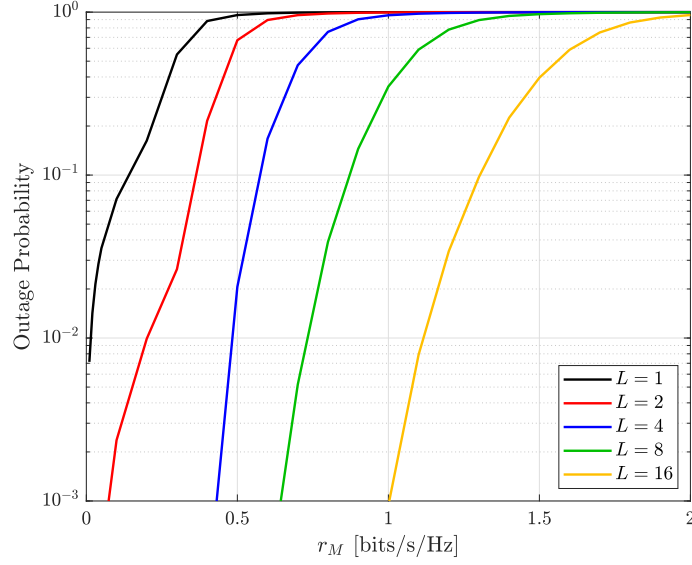


Figure 17. Outage probabilities of the mMTC versus data rate  $r_M$  for  $M = 10$ ,  $\Gamma_M = 0$  dB and different numbers of receive antennas at the BS.

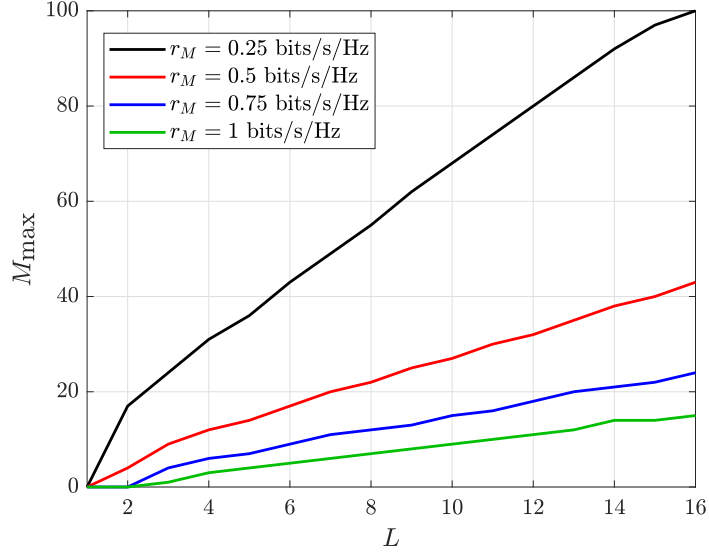


Figure 18. Maximum number of mMTC devices that can be connected to the BS as a function of the number of receive antennas  $L$ , for  $\epsilon_M = 0.1$ ,  $\Gamma_M = 0$  dB and different data rates  $r_M$

outage probabilities. Conversely, for a given outage probability, the more receive antennas at the BS, the higher is the data rate that can be achieved by the MTC devices.

Finally, Fig. 18 depicts the maximum number of MTC devices that can be connected to the BS as a function of the number of receive antennas, for  $\epsilon_M = 0.1$ ,  $\Gamma_M = 0$  dB and different data rates  $r_M$ . We observe that the maximum number of connected devices increases linearly with the number of receive antennas. However, as we increase the data rate  $r_M$ , we decrease severely the number of connected devices.

## 4 NETWORK SLICING FOR COEXISTING EMBB AND URLLC SCENARIOS

In this chapter, we assess the coexistence of eMBB and URLLC for both orthogonal and non-orthogonal slicing of radio resources. The slicing strategies, results and discussions presented here are based on the previous work [36].

### 4.1 Orthogonal Slicing For Coexisting eMBB and URLLC

Under the orthogonal slicing assumption,  $F_U$  out of  $F$  frequency channels over all minislots are allocated for URLLC traffic, while the remaining  $F_B = F - F_U$  channels are each allocated to one eMBB user. Given the reliability constraints  $\epsilon_U$  and  $\epsilon_B$  for each such services, the performance of the system is then evaluated in terms of the pair of the sum rates  $(r_B^{\text{sum}}, r_U^{\text{sum}})$ .

The eMBB sum rate is obtained by [10]

$$r_B^{\text{sum}} = (F - F_U)r_B^{\text{out}}, \quad (65)$$

where  $r_B^{\text{out}}$  is given by (54).

Condition on a a number of frequency channels allocated for URLLC traffic, we compute the maximum URLLC symmetric rate  $r_U$  that guarantees the reliability constraint  $\Pr(E_U) \leq \epsilon_U$  for all  $n_U$  (the number of URLLC devices simultaneously transmitting), as detailed in Section 3.4. Besides, when computing  $r_U$ , the error probabilities for all URLLC users are computed individually.

### 4.2 Non-Orthogonal Slicing Between Coexisting eMBB and URLLC

Under the non-orthogonal slicing assumption, all  $F$  frequency channels are used for both eMBB and URLLC services, such that  $F_U = F_B = F$ . Due to the latency constraints, the decoding of a URLLC transmission cannot wait for the decoding of the eMBB traffic. The eMBB latency requirements are less demanding, and hence its decoding can wait until the URLLC transmissions are decoded. Thus, during the SIC decoding procedure, the BS first attempts to successively decode the packets from all the  $n_U$  active URLLC users in a minislot. After decoding the packet from an URLLC device, it is subtracted from the total received signal, thus it no longer represents interference to other users. Only after successfully decoding all URLLC packets the BS attempts to decode the packets from the eMBB users. As a result, during the decoding of the URLLC, the interference from eMBB is always present. On the other hand, when the BS tries to decode the eMBB packets, there is no longer interference from URLLC.

In the orthogonal case, as shown in (53), the variable  $\gamma_{B,f}^{\text{tar}}$  is uniquely determined by the reliability requirement  $\epsilon_B$  and the threshold SNR  $\gamma_{B,f}^{\text{min}}$  defined in (49). For the non-orthogonal slicing, it may be beneficial to choose a smaller target SNR than the one given in (53), so as to reduce the interference inflicted to URLLC transmissions. This yields the following inequality [10]

$$\gamma_{B,f}^{\text{tar}} \leq \frac{\Gamma_B}{\Gamma\left(0, \frac{\gamma_{B,f}^{\text{min}}}{\Gamma_B}\right)}. \quad (66)$$

The maximum allowed SNR for the eMBB devices, which we denote by  $\gamma_{B,\max}^{\text{tar}}$ , is set by the inequality in (66). Thereby, the maximum allowed eMBB data rate is given by  $r_B^{\max} = \log_2(1 + \gamma_{B,\max}^{\text{tar}})$ .

Similar to the orthogonal slicing case, conditional on both the eMBB and URLLC reliability requirements, our objective is to determine the achievable pairs of sum rates  $(r_B^{\text{sum}}, r_U^{\text{sum}})$ . To this end, we initially fix an eMBB per device data rate  $r_B \in [0, r_B^{\max}]$  and then we compute the maximum achievable symmetric rate  $r_U$ . For a given value of  $r_U$ , we search for the minimum value of the SNR  $\gamma_B^{\text{tar}} \in [\gamma_B^{\min}, \gamma_{B,\max}^{\text{tar}}]$  that can be adopted for all eMBB devices. The error probabilities for all eMBB and URLLC users are computed individually.

Adapting (55) to include the interference from eMBB, and again assuming a decoding order  $\{U_1, \dots, U_{n_U}\}$ , the SINR of the URLLC user with index  $u \in \{1, \dots, n_U\}$  in the frequency channel  $f$  is given by

$$\sigma_{u,f} = \frac{g_{U_u,f}}{1 + \gamma_{B,f}^{\text{tar}} + \sum_{j=u+1}^{n_U} g_{U_j,f}}. \quad (67)$$

For the non-orthogonal slicing, the outage probability of the URLLC user is also given by (56), and it is correctly decoded if the condition in (57) holds, but now considering the SINR given by (67).

As stated in Section 3.3, we assume that the interference from eMBB is always present while decoding the packets from the URLLC users. As a result, the outage probability for the  $F$  eMBB users that share their frequency channels with the URLLC users is the same and equal to the probability that at least one URLLC packet is decoded incorrectly.

### 4.3 Performance Evaluation

In this section, we evaluate numerical results obtained using Monte Carlo simulations for the orthogonal and non-orthogonal slicing of radio resources between coexisting eMBB and URLLC. For the sake of tractability, we consider the cases of  $n_U \in \{1, 2, 3, 4\}$ , where  $n_U = 1$  is the scenario from [10]. Besides, we consider a time-frequency grid with  $F = 10$  frequency channels and reliability requirements  $\epsilon_U = 10^{-5}$  and  $\epsilon_B = 10^{-3}$  [10].

We plot the pair of sum rates for the orthogonal and non-orthogonal slicing scenarios considering average channel gains of  $\Gamma_U = 20$  dB and  $\Gamma_B = 10$  dB in Fig. 19 and  $\Gamma_U = 10$  dB and  $\Gamma_B = 20$  dB in Fig. 20. For  $\Gamma_U > \Gamma_B$ , the highest values of  $r_U^{\text{sum}}$  are achieved when only one URLLC user is active in the minislot, that is,  $n_U = 1$ , as showed by the black curves in Fig. 19. As we increase  $r_B^{\text{sum}}$ , the non-orthogonal slicing allows us to achieve pairs of sum rates that are not possible to achieve using the orthogonal slicing, as depicted by the dashed curves in Fig. 19. For both orthogonal and non-orthogonal slicing and  $\Gamma_U > \Gamma_B$ , the  $r_U^{\text{sum}}$  is inversely proportional to  $n_U$ , but interestingly, in the setups with  $n_U > 1$ , the values of  $r_U^{\text{sum}}$  do not vary much in the orthogonal slicing even for high increases of  $r_B^{\text{sum}}$ , while the non-orthogonal slicing makes them be almost constant for the whole range of  $r_B^{\text{sum}}$ .

In Fig. 20 we consider an opposite scenario where  $\Gamma_U < \Gamma_B$ . Differently from the case where  $\Gamma_U > \Gamma_B$ , now the setups with  $n_U > 1$  outperform the case with  $n_U = 1$ , as shown by the achieved pairs of sum rates in Fig. 20. When  $\Gamma_U < \Gamma_B$ , it is possible to achieve

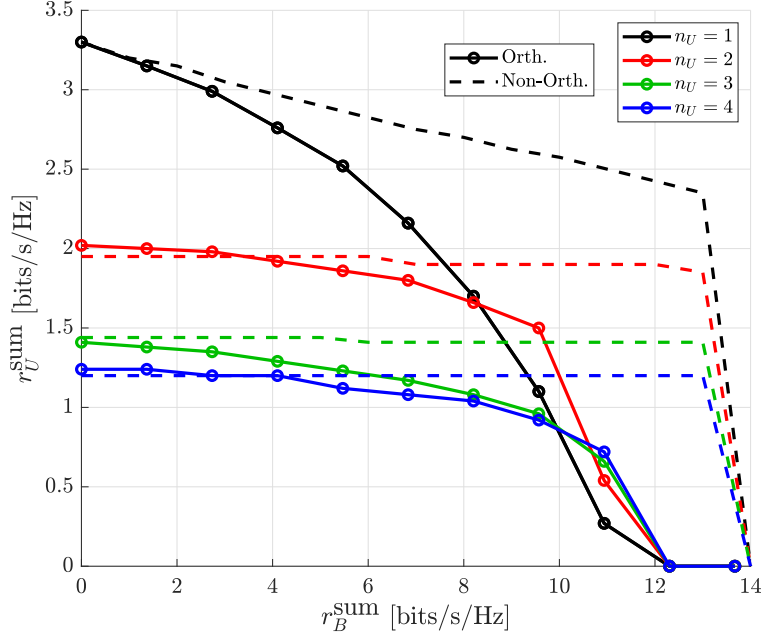


Figure 19. eMBB sum rate  $r_B^{\text{sum}}$  versus URLLC sum rate  $r_U^{\text{sum}}$  for the the orthogonal and non-orthogonal slicing when  $\Gamma_U = 20$  dB,  $\Gamma_B = 10$  dB,  $F = 10$ ,  $\epsilon_U = 10^{-5}$  and  $\epsilon_B = 10^{-3}$  (reproduced from [36]).

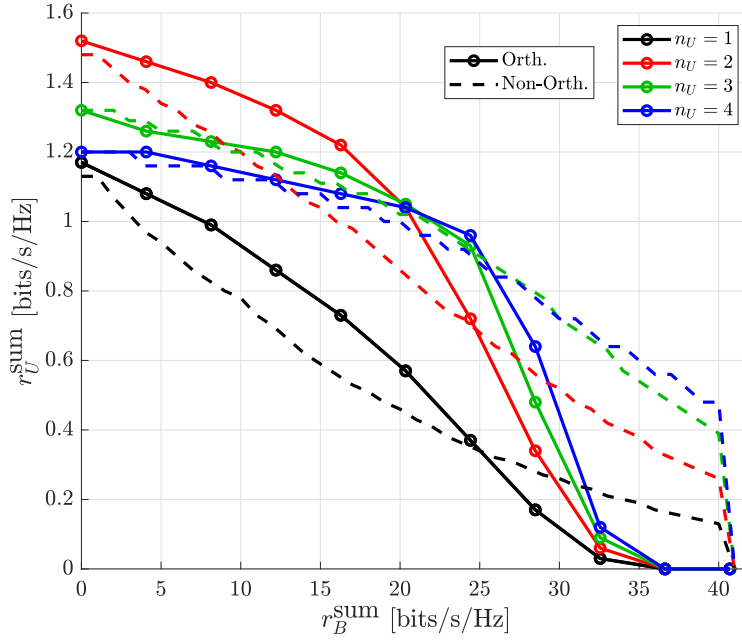


Figure 20. eMBB sum rate  $r_B^{\text{sum}}$  versus URLLC sum rate  $r_U^{\text{sum}}$  for the the orthogonal and non-orthogonal slicing when  $\Gamma_U = 10$  dB,  $\Gamma_B = 20$  dB,  $F = 10$ ,  $\epsilon_U = 10^{-5}$  and  $\epsilon_B = 10^{-3}$  (reproduced from [36]).

higher values of  $r_B^{\text{sum}}$ , but at the cost of lower values of  $r_U^{\text{sum}}$ . For a large range of  $r_B^{\text{sum}}$ , the orthogonal slicing outperforms the non-orthogonal approach, while the non-orthogonal

slicing outperforms the orthogonal one only when  $r_B^{\text{sum}}$  is high (see the dashed curves Fig. 20). In both Figs. 19 and 20, the curves obtained for  $n_U = 1$  match those in [10].

All in all, when  $\Gamma_U > \Gamma_B$  it is possible to achieve higher URLLC sum rates at cost of lower eMBB sum rates. However, the URLLC sum rate reduces as the number of connected devices increases. Therefore, under these conditions and for applications that impose strict rate constraints, it is better to limit  $n_U \leq 2$ , which provides the largest gains. On the other hand, the situation where  $\Gamma_U < \Gamma_B$  is favorable to eMBB traffic in terms of the achievable sum rates. However, under this condition the NOMA URLLC becomes advantageous, since it is possible to achieve higher sum rates when we have multiple URLLC users connected to the same BS<sup>2</sup>. Finally, the non-orthogonal slicing is the best choice only for applications that require very high eMBB sum rates.

---

<sup>2</sup>Overall, the results indicate that NOMA provides the largest gains for limited number of users, in this case  $n_U = 2$ , which corroborated with results in [40].

## 5 NETWORK SLICING FOR COEXISTING EMBB AND MMTC SCENARIOS

In this chapter, we present the orthogonal and non-orthogonal slicing schemes that allow the eMBB and mMTC services to coexist in the same RAN. For each such scheme we characterize the pair of maximum achievable data rates  $(r_B, r_M)$  given the corresponding reliability requirements  $\epsilon_B$  and  $\epsilon_M$  for eMBB and mMTC, respectively. The slicing schemes, results and discussions presented here are based on the previous work [37].

### 5.1 Orthogonal Slicing Between Coexisting eMBB and mMTC

Under the orthogonal slicing assumption, the eMBB and the MTC devices use the radio resource in a time-sharing manner. Let  $\alpha \in [0, 1]$  and  $1 - \alpha$  denote the fraction of time in which the frequency channel is allocated to the eMBB traffic and to the mMTC traffic, respectively. For a given time-sharing factor  $\alpha$ , the eMBB data rate is [10]

$$r_B = \alpha r_B^{\text{out}}, \quad (68)$$

where  $r_B^{\text{out}}$  is given by (54). Similarly, the mMTC data rate is

$$r_M = (1 - \alpha) r_M^{\text{out}}, \quad (69)$$

where  $r_M^{\text{out}}$  is the maximum achievable mMTC data rate in the absence of interference from the eMBB traffic.

To characterize the performance of the orthogonal slicing, for each value of  $\alpha$ , we set an eMBB data rate according to (68). Then we compute the maximum achievable mMTC data rate  $r_M$  for which the reliability requirements  $\epsilon_B$  and  $\epsilon_M$  are met.

### 5.2 Non-Orthogonal Slicing Between Coexisting eMBB and mMTC

By employing the non-orthogonal slicing, the coexisting eMBB and mMTC traffics overlap on the radio resource. The received signal vector at the BS is then

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{w}, \quad (70)$$

where

$$\mathbf{G} = [\mathbf{g}_{m,1} \ \mathbf{g}_{m,2} \ \cdots \ \mathbf{g}_{m,M} \ \mathbf{g}_B] \quad (71)$$

is a matrix containing the channel gains between all the devices and the serving BS,  $\mathbf{G} \in \mathbb{C}^{L \times (M+1)}$ ,

$$\mathbf{x} = \left\{ \sqrt{P_M} [x_{m,1} \ x_{m,2} \ \cdots \ x_{m,M}] \ \sqrt{P_B} x_B \right\}^T \quad (72)$$

is a complex vector containing the transmitted symbols by the MTC and eMBB devices, and  $\mathbf{w} \in \mathbb{C}^{L \times 1}$  is the vector containing the noise samples. As in the orthogonal case, the received signal vector after carrying out MRC at the BS is

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{G}^H \mathbf{y} \\ &= \mathbf{G}^H \mathbf{G} \mathbf{x} + \mathbf{G}^H \mathbf{w}. \end{aligned} \quad (73)$$

Let us denote by  $\hat{x}_m$  and  $\hat{x}_B$  the elements of the vector  $\hat{\mathbf{x}}$  corresponding to the signal of the  $m$ -th MTC and the eMBB devices, respectively, which are given by

$$\begin{aligned} \hat{x}_m = & \sqrt{P_M} \mathbf{g}_m^H \mathbf{g}_m x_m + \sqrt{P_M} \mathbf{g}_m^H \sum_{m' \neq m}^M \mathbf{g}_{m'} x_{m'} + \\ & \sqrt{P_B} \mathbf{g}_m^H \mathbf{g}_B x_B + \mathbf{g}_m^H \mathbf{w}, \end{aligned} \quad (74)$$

$$\hat{x}_B = \sqrt{P_B} \mathbf{g}_B^H \mathbf{g}_B x_B + \sqrt{P_M} \mathbf{g}_B^H \sum_{m=1}^M \mathbf{g}_m x_m + \mathbf{g}_B^H \mathbf{w}. \quad (75)$$

Assuming the SIC decoding ordering  $\{1, \dots, M\}$  is used as in the orthogonal case, the SINR of the  $m$ -th mMTC in the presence of the eMBB interferer reads

$$\sigma_m = \frac{P_M \|\mathbf{g}_m\|^4}{P_M \sum_{m'=m+1}^M |\mathbf{g}_m^H \mathbf{g}_{m'}|^2 + P_B |\mathbf{g}_m^H \mathbf{g}_B|^2 + \|\mathbf{g}_m\|^2}. \quad (76)$$

As in the orthogonal case, the mMTC is correctly decoded if the inequality  $\log_2(1 + \sigma_m) \geq r_M$  holds.

After correctly decoding the  $m$ -th MTC device, the BS attempts to decode the eMBB device if it has not been decoded yet. Then, the SINR of the eMBB device is

$$\sigma_B = \frac{P_B \|\mathbf{g}_B\|^4}{P_M \sum_{m'=m+1}^M |\mathbf{g}_B^H \mathbf{g}_{m'}|^2 + \|\mathbf{g}_B\|^2}. \quad (77)$$

Conditional on the data rate  $r_B$ , the eMBB device is correctly decoded only if the inequality  $\log_2(1 + \sigma_B) \geq r_B$  holds.

By employing the orthogonal slicing approach, the eMBB device adopts a fixed target SNR  $\gamma_B^{\text{tar}}$  that satisfies the power constraint  $\mathbb{E}\{P_B\} = 1$ . Conversely, aiming to minimize the interference that the eMBB traffic causes to the mMTC traffic, when using the non-orthogonal slicing we allow the eMBB device to adopt lower values for the target SNR, which yields the inequality [10]

$$\gamma_B^{\text{tar}} \leq \frac{\Gamma_B(L-1)!}{\Gamma\left(L-1, \frac{\gamma_B^{\min}}{\Gamma_B}\right)}. \quad (78)$$

Hence, we have  $\mathbb{E}\{P_B\} \leq 1$ . Nevertheless, this condition is acceptable when the eMBB device transmits with a data rate  $r_B \leq r_B^{\text{out}}$ .

Differently from the orthogonal slicing, the error probability for eMBB has two components in the non-orthogonal case: i) the probability of the eMBB device does not transmit due to insufficient SNR; and ii) the probability of the eMBB device transmits because it has sufficient SNR, but a decoding error occurs due to the interference from the mMTC traffic. In order to satisfy the same reliability requirement from the orthogonal case, we must allow the activation probability of the eMBB device to be higher, which yields  $a_B > 1 - \epsilon_B$ . If we adopt, for example,  $\epsilon_B = 10^{-3}$ , then  $a_B > 0.999$ . For the computation of the maximum achievable mMTC data rate, we conservatively assume

that the eMBB interference is always present, that is,  $a_B = 1$ , such that the error probability for eMBB is just the decoding error probability, that is,

$$\Pr(E_B) = \Pr \{ \log_2(1 + \sigma_B) < r_B \}. \quad (79)$$

The SIC decoding procedure for the non-orthogonal slicing works as follows. Initially, all the MTC devices suffer with the interference from eMBB traffic. First the BS attempts to decode the strongest MTC device. If the decoding succeeds, the signal from the decoded device is subtracted from the received signal, thereafter BS attempts to decode the second strongest MTC device, and so on. If the decoding of a MTC device fails, the BS tries to decode the signal of the interfering eMBB device. If its signal is correctly decoded, the eMBB interference component is subtracted from the received signal, then the BS returns to the decoding of the MTC devices, and the procedure continues as described in Section 3.5. Otherwise, if the decoding of the eMBB fails, the SIC decoding ends. Alternatively, the SIC decoding procedure may also end when all the MTC devices are correctly decoded, and so the last step is just to decode the eMBB signal without the interference from the mMTC traffic. It is important to note that the step when the eMBB device is decoded is random.

The performance evaluation of the non-orthogonal slicing is a two dimensional numerical search: first we set the eMBB data rate  $r_B \in [0, r_B^{\text{out}}]$  and then compute the maximum mMTC data rate  $r_M$  that is achievable by all the MTC devices connected to the BS while still satisfying the reliability requirements  $\epsilon_B$  and  $\epsilon_M$ ; during this computation, we seek for the minimum value of  $\gamma_B^{\text{tar}}$  that can be adopted by the eMBB device.

### 5.3 Performance Evaluation

In this section, we carry out Monte Carlo simulations to evaluate the performance of both orthogonal and non-orthogonal slicing of radio resources between coexisting eMBB and mMTC. We set the reliability requirements  $\epsilon_B = 10^{-3}$  and  $\epsilon_M = 10^{-1}$  for the eMBB and mMTC services, respectively, while the respective average channel gain equals  $\Gamma_B = 20$  dB for the former and  $\Gamma_M = 5$  dB for the latter.

Fig. 21 shows the pairs of achievable data rates  $(r_M, r_B)$  for  $M = 10$  MTC devices connected to the serving BS, and increasing number of receive antennas elements  $L \in \{1, 2, 4, 8, 16\}$ . Besides, Fig. 22 depicts the pairs  $(M_{\text{max}}, r_M)$  of maximum number of connected MTC devices versus the eMBB data rate given mMTC data rate  $r_M = 0.25$  bits/s/Hz, and also for  $L \in \{1, 2, 4, 8, 16\}$ . In both figures, the dashed and solid curves correspond to the orthogonal and non-orthogonal network slicing strategies, respectively.

From both figures, increasing  $L$  always improves the performance of the system for both slicing schemes. At each SIC decoding step the MRC receiver projects the received signal vector onto the direction of the signal of interest. By considering more antenna elements, we increase the power gain, i.e., the components of the received signal are amplified onto the direction of the signal of interest and attenuated on other directions, which maximizes available SINR at each SIC decoding step.

Meanwhile, we also observe that as the number of receive antenna elements  $L$  is increased, the non-orthogonal slicing becomes more advantageous compared to the orthogonal slicing both in terms of the maximum achievable mMTC data rate  $r_M$  and the maximum number of connected MTC devices  $M_{\text{max}}$ . For  $L \leq 2$ , the orthogonal

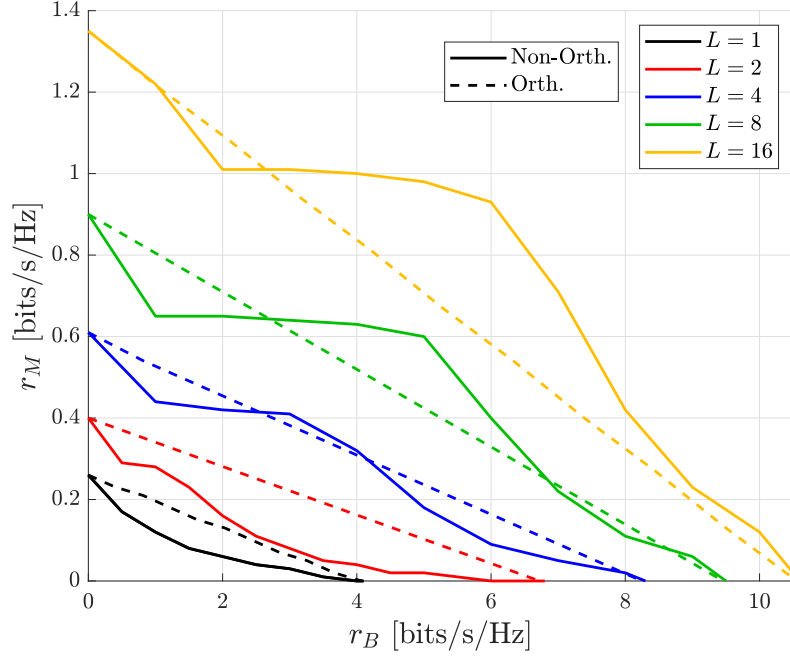


Figure 21. eMBB data rate  $r_B$  versus mMTC data rate  $r_M$  for the orthogonal and non-orthogonal slicing, considering different numbers of receive antennas and for  $\Gamma_B = 20$  dB,  $\Gamma_m = 5$  dB,  $\epsilon_B = 10^{-3}$ ,  $\epsilon_m = 10^{-1}$  and  $M = 10$  (reproduced from [37]).

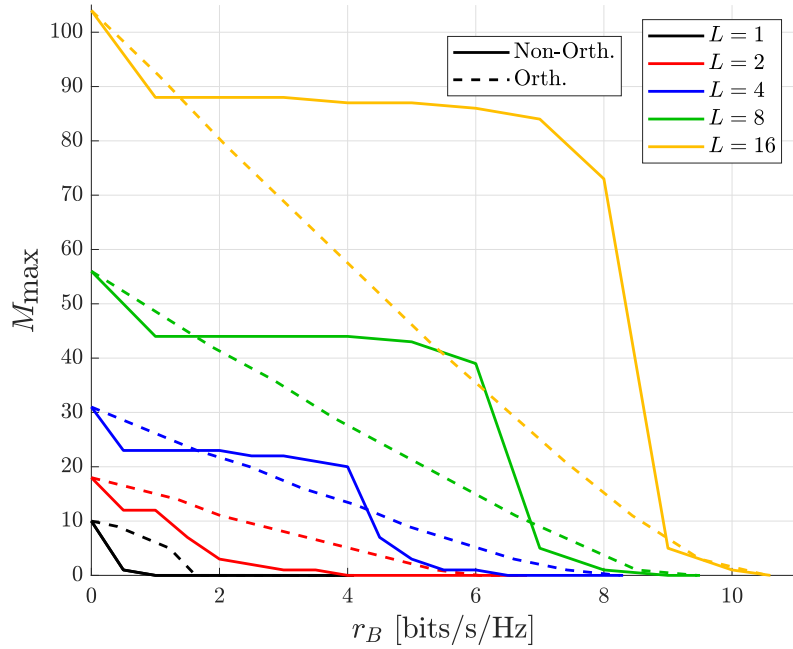


Figure 22. eMBB data rate  $r_B$  versus the maximum number of connected MTC devices  $M_{\max}$  for the orthogonal and non-orthogonal slicing, considering different numbers of receive antennas and for  $\Gamma_B = 20$  dB,  $\Gamma_m = 5$  dB,  $\epsilon_B = 10^{-3}$ ,  $\epsilon_m = 10^{-1}$  and  $r_M = 0.25$  bits/s/Hz (reproduced from [37]).

slicing outperforms the non-orthogonal approach for the whole range of  $r_B$ . However, as we adopt  $L \geq 4$ , the non-orthogonal slicing achieves higher values of  $r_M$  and  $M_{\max}$  compared to the orthogonal approach over an increasing range of  $r_B$ .

The orthogonal slicing curves are straight lines because  $r_M$  and  $r_B$  in Fig. 21 and  $M_{\max}$  and  $r_B$  in Fig. 22 scale linearly with the fraction of the radio resource  $\alpha$  that is allocated for each service. On the other hand, the non-orthogonal case exhibits non-linear curves because is conditional on the level of interference that eMBB causes to mMTC. Starting from  $r_B = 0$ , there is no interference from the eMBB traffic, so the mMTC performance for both slicing schemes remains the same. Then, when  $r_B > 0$ , there is an abrupt reduction in the performance of mMTC because of the presence of interference from eMBB traffic. For the lowest values of  $r_B$ , the interference generated by the eMBB service to the mMTC traffic is minimal, so that almost all the MTC devices are correctly decoded before decoding the eMBB device. As we increase  $r_B$ , we also increase the interference level of eMBB because high SNR threshold needs to be set in order to meet the target data rate. In this regime, the eMBB device starts to be decoded before some of the MTC devices have been decoded. As a result, after correctly decoding the eMBB, some of the MTC devices do not suffer with the interference anymore, and the decrease in the performance of mMTC keeps almost constant. Finally, for high values of  $r_B$ , the eMBB device adopts high SNR target, which causes severe interference levels to the mMTC. As a result, the performance of mMTC decreases abruptly.

It is important to note that although adopting spatial receive diversity substantially improves the performance of mMTC in terms of both achievable data rates and number of connected devices, it also increases the receiver complexity. Moreover, NOMA of a massive number of MTC devices also increases the receiver complexity and yields higher processing delay times. These aspects must be taken into account when implementing the network slicing of radio resources in practical situations.

## 6 CONCLUSIONS

In this thesis, we studied two different uplink scenarios for beyond-5G and 6G networks: i) the coexistence between eMBB and URLLC, and ii) the coexistence between eMBB and mMTC. In both cases, the coexistence of the heterogeneous services in the same RAN is enabled by means of physical layer network slicing, diversity schemes, and NOMA with SIC decoding.

When studying the coexistence between eMBB and URLLC services, we assume that multiple eMBB and URLLC devices are simultaneously connected to the same serving BS in the uplink. The URLLC transmissions explored frequency diversity and minislots structure to meet the very stringent requirements of reliability and latency, respectively. The eMBB and URLLC share a time-frequency radio resource grid with  $F$  frequency channels and  $S$  minislots. By employing the orthogonal slicing strategy, a given number of frequency channels are exclusively allocated to eMBB, while the remaining resources are allocated to URLLC. On the other hand, when using the non-orthogonal slicing scheme, both services share the same frequency channels. In both cases, aiming at increasing the number of URLLC devices that may be connected to the BS, multiple URLLC devices are allowed to transmit simultaneously in the same minislot by using NOMA. The serving BS performs SIC decoding to recover multiple overlapping URLLC packets, as well as the eMBB packets in the case of non-orthogonal slicing. Our simulation results show that, even with multiple overlapping eMBB and URLLC signals, the reliability requirements of both services are still met. However, the main drawbacks of the proposed approach are an increased receiver complexity and lower data rates for URLLC due to the increased interference levels. We also demonstrate through simulations that when the URLLC users have better channel conditions than the eMBB users, the non-orthogonal slicing outperforms the orthogonal scheme for the whole range of eMBB sum rates. However, when the eMBB users have better channel conditions than the URLLC ones, the non-orthogonal slicing is only the best option for high values of eMBB sum rates.

The coexistence of eMBB and mMTC is also enabled by employing orthogonal and non-orthogonal slicing schemes in the uplink of the RAN. Both services share a radio resource that is composed of a single timeslot in a single frequency channel. Under the orthogonal slicing assumption, a fraction of the timeslot is allocated exclusively for mMTC, while the remaining of the timeslot is allocated exclusively for eMBB. On the other hand, by using the non-orthogonal slicing, the traffic from both services overlap during the whole timeslot. In both schemes, the massive connectivity required by mMTC applications is achieved using NOMA along with SIC decoding. Multiple receive antennas mitigate the imperfections of the wireless channel and guarantee the spectral efficiency of both services. We set the reliability requirements and then evaluate the pairs of maximum achievable data rates through Monte Carlo simulations. Our simulation results show that, the higher the number of receive antennas, the more advantageous the non-orthogonal slicing becomes when compared against the orthogonal slicing in terms of both the achievable mMTC data rates for a given number of connected devices, and the number of connected MTC devices for a given mMTC data rate. Finally, although the spatial receive diversity increases substantially the performance of the system, it also increases the receiver complexity. Moreover, NOMA of a massive number of devices is also a complex task and yields higher processing delay. Such aspects must be considered in practical implementations.

The framework developed in this work can be used in the specification of beyond-5G and 6G networks for industrial setups. In such scenarios, a large number of devices used for the control and/or monitoring of critical processes may require URLLC connectivity in the coexistence with other applications that require the high data rates provided by eMBB (e.g. video surveillance in the industrial environment). At the same time, the monitoring a very large number of machines and assets may require the mMTC connectivity.

The scenarios addressed in this thesis are mathematically intractable, thus we resorted to Monte Carlo simulations. It is worth mentioning that the time required to run such simulations constitute the greatest difficulty faced during the compilation of this work. For example, since we investigate the performance of URLLC adopting a reliability requirement of  $10^{-5}$ , each Monte Carlo simulation must have a number of runs between  $10^6$  or  $10^7$  so that the results can sufficiently reliable [41]. The adopted step size for the variables also significantly affect the required execution times. Despite running the simulations on dedicated servers and using the parallel computing toolbox of MATLAB [42] to run independent iterations of simulations on multiple CPU cores simultaneously, each Monte Carlo simulation can take several hours to finish.

There are many possible directions for future works based on the proposed framework. One could relax the latency requirements of URLLC and allow the use of fixed retransmissions schemes to explore the trade-offs between data rate, reliability requirements and latency budget. Future works could also consider the heterogeneous requirements for URLLC, i.e., different classes of URLLC devices with different data rates and reliability requirements, such that user pairing can be adopted to maximize the sum rates [43]. It would be also interesting to study the performance gains in scenarios with interface diversity, that is, where an URLLC user can be connected to more than one BS, and also scenarios where the BS have multiple receive antennas. Artificial intelligence and machine learning solutions could perform a dynamic switching between the OMA and NOMA schemes based on the current network status, and also a dynamic channel allocations for eMBB and URLLC users based on their channel conditions. Finally, it would be interesting to develop the analytical models for the proposed framework, and also consider a traffic source models for URLLC.

## 7 REFERENCES

- [1] H. Tullberg et al. “The METIS 5G System Concept: Meeting the 5G Requirements”. In: *IEEE Commun. Mag.* 54.12 (Dec. 2016), pp. 132–139. DOI: 10.1109/MCOM.2016.1500799CM.
- [2] Ericsson. *5G Systems*. Tech. rep. 2017. URL: <https://www.ericsson.com/en/reports-and-papers/white-papers/5g-systems--enabling-the-transformation-of-industry-and-society>.
- [3] M. Latva-Aho and K. Leppänen. “Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence”. In: *6G Wireless Summit, Levi, Finland*. Mar. 2019.
- [4] N. H. Mahmood et al. *White Paper on Critical and Massive Machine Type Communications Towards 6G*. White Paper. University of Oulu, 2020.
- [5] N. H. Mahmood et al. “Six Key Features of Machine Type Communication in 6G”. In: *2020 2nd 6G Wireless Summit (6G SUMMIT)*. 2020, pp. 1–5.
- [6] GSM Association. *An Introduction to Network Slicing*. Tech. rep. 2017. URL: <https://www.gsma.com/futurenetworks/wp-content/uploads/2017/11/GSMA-An-Introduction-to-Network-Slicing.pdf>.
- [7] X. Foukas et al. “Network Slicing in 5G: Survey and Challenges”. In: *IEEE Communications Magazine* 55.5 (2017), pp. 94–100.
- [8] H. Chien et al. “End-to-End Slicing as a Service with Computing and Communication Resource Allocation for Multi-Tenant 5G Systems”. In: *IEEE Wireless Communications* 26.5 (2019), pp. 104–112.
- [9] H. Zhang et al. “Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges”. In: *IEEE Communications Magazine* 55.8 (2017), pp. 138–145.
- [10] P. Popovski et al. “5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View”. In: *IEEE Access* 6 (2018), pp. 55765–55779. DOI: 10.1109/ACCESS.2018.2872781.
- [11] R. Kassab et al. “Non-Orthogonal Multiplexing of Ultra-Reliable and Broadband Services in Fog-Radio Architectures”. In: *IEEE Access* 7 (2019), pp. 13035–13049. DOI: 10.1109/ACCESS.2019.2893128.
- [12] W. Saad, M. Bennis, and M. Chen. “A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems”. In: *IEEE Network* 34.3 (2020), pp. 134–142.
- [13] P. Popovski et al. “Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks”. In: *IEEE Network* 32.2 (Mar. 2018), pp. 16–23. DOI: 10.1109/MNET.2018.1700258.
- [14] P. Popovski et al. “Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)”. In: *IEEE Transactions on Communications* 67.8 (Aug. 2019), pp. 5783–5801. ISSN: 1558-0857. DOI: 10.1109/TCOMM.2019.2914652.

- [15] Ericsson. *Cellular IoT Evolution for Industry Digitalization*. Tech. rep. 2019. URL: <https://www.ericsson.com/en/reports-and-papers/white-papers/cellular-iot-evolution-for-industry-digitalization>.
- [16] R. Candell and M. Kashef. “Industrial wireless: Problem space, success considerations, technologies, and future direction”. In: *2017 Resilience Week (RWS)*. Sept. 2017, pp. 133–139. DOI: 10.1109/RWEEK.2017.8088661.
- [17] E. Sisinni et al. “Industrial Internet of Things: Challenges, Opportunities, and Directions”. In: *IEEE Trans. on Ind. Inf.* 14.11 (Nov. 2018), pp. 4724–4734. DOI: 10.1109/TII.2018.2852491.
- [18] D. Serpanos. “The Cyber-Physical Systems Revolution”. In: *Computer* 51.3 (2018), pp. 70–73.
- [19] L. Dai et al. “Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends”. In: *IEEE Commun. Mag.* 53.9 (Sept. 2015), pp. 74–81. DOI: 10.1109/MCOM.2015.7263349.
- [20] B. Makki et al. “A Survey of NOMA: Current Status and Open Research Challenges”. In: *IEEE Open Journal of the Communications Society* 1 (2020), pp. 179–189.
- [21] Federico Clazzer et al. “From 5G to 6G: Has the Time for Modern Random Access Come?” In: *arXiv e-prints*, arXiv:1903.03063 (Mar. 2019), arXiv:1903.03063. arXiv: 1903.03063 [eess.SP].
- [22] Alid A. Zaid et al. “Designing for the Future: the 5G NR Physical Layer”. In: *Ericsson Technology Review* (June 2017). URL: <https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/designing-for-the-future-the-5g-nr-physical-layer>.
- [23] Z. Wu, F. Zhao, and X. Liu. “Signal Space Diversity Aided Dynamic Multiplexing for eMBB and URLLC Traffics”. In: *2017 ICC*. Dec. 2017, pp. 1396–1400. DOI: 10.1109/CompComm.2017.8322772.
- [24] A. Anand, G. De Veciana, and S. Shakkottai. “Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks”. In: *IEEE INFOCOM 2018*. Apr. 2018, pp. 1970–1978. DOI: 10.1109/INFOCOM.2018.8486430.
- [25] A. A. Esswie and K. I. Pedersen. “Opportunistic Spatial Preemptive Scheduling for URLLC and eMBB Coexistence in Multi-User 5G Networks”. In: *IEEE Access* 6 (2018), pp. 38451–38463. DOI: 10.1109/ACCESS.2018.2854292.
- [26] M. Alsenwi et al. “eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach”. In: *IEEE Commun. Lett.* 23.4 (Apr. 2019), pp. 740–743. DOI: 10.1109/LCOMM.2019.2900044.
- [27] P. K. Korrai et al. “Slicing Based Resource Allocation for Multiplexing of eMBB and URLLC Services in 5G Wireless Networks”. In: *2019 IEEE CAMAD*. Sept. 2019, pp. 1–5. DOI: 10.1109/CAMAD.2019.8858433.
- [28] E. J. dos Santos et al. “Network Slicing for URLLC and eMBB With Max-Matching Diversity Channel Allocation”. In: *IEEE Communications Letters* 24.3 (2020), pp. 658–661.

- [29] O. Vikhrova et al. “Enhanced Radio Access Procedure in Sliced 5G Networks”. In: *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. 2019, pp. 1–6.
- [30] N. H. Mahmood et al. “Radio Resource Management Techniques for eMBB and mMTC Services in 5G Dense Small Cell Scenarios”. In: *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*. 2016, pp. 1–5.
- [31] M. Kamel, W. Hamouda, and A. Youssef. “Uplink Performance of NOMA-Based Combined HTC and MTC in Ultradense Networks”. In: *IEEE Internet of Things Journal* 7.8 (2020), pp. 7319–7333.
- [32] O. L. Alcaraz López et al. “Aggregation and Resource Scheduling in Machine-Type Communication Networks: A Stochastic Geometry Approach”. In: *IEEE Transactions on Wireless Communications* 17.7 (2018), pp. 4750–4765.
- [33] Y. Mo, C. Goursaud, and J. Gorce. “Uplink Multiple Base Stations Diversity for UNB based IoT networks”. In: *2018 IEEE Conference on Antenna Measurements Applications (CAMA)*. 2018, pp. 1–4.
- [34] J. M. d. S. Sant’Ana et al. “LORA Performance Analysis with Superposed Signal Decoding”. In: *IEEE Wireless Communications Letters* (2020), pp. 1–1.
- [35] M. Liu et al. “Non-Asymptotic Outage Probability of Large-Scale MU-MIMO Systems with Linear Receivers”. In: *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*. 2016, pp. 1–5.
- [36] E. N. Tominaga et al. *Non-Orthogonal Multiple Access and Network Slicing: Scalable Coexistence of eMBB and URLLC*. 2021. arXiv: 2101.04605 [eess.SP].
- [37] E. N. Tominaga et al. *Network Slicing for eMBB and mMTC with NOMA and Space Diversity Reception*. 2021. arXiv: 2101.04983 [eess.SP].
- [38] David Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*. USA: Cambridge University Press, 2005. ISBN: 0521845270.
- [39] Andrea Goldsmith. *Wireless Communications*. Cambridge University Press, 2005. DOI: 10.1017/CB09780511841224.
- [40] O. L. Alcaraz López et al. “Aggregation and Resource Scheduling in Machine-Type Communication Networks: A Stochastic Geometry Approach”. In: *IEEE Trans. on Wireless Commun.* 17.7 (July 2018), pp. 4750–4765. DOI: 10.1109/TWC.2018.2830767.
- [41] Gerardo Rubino, Bruno Tuffin, et al. *Rare event simulation using Monte Carlo methods*. Vol. 73. Wiley Online Library, 2009.
- [42] Mathworks®. *Parallel Computing Toolbox*. 2021. URL: <https://www.mathworks.com/products/parallel-computing.html> (visited on 01/28/2021).
- [43] H. Zhang et al. “User Pairing Algorithm with SIC in Non-Orthogonal Multiple Access System”. In: *2016 IEEE International Conference on Communications (ICC)*. 2016, pp. 1–6. DOI: 10.1109/ICC.2016.7511620.