Design and evaluation of technical components building the flexible 5G radio interface supporting heterogeneous use cases

Department of Electrical and Information Engineering
ELECTRICAL AND INFORMATION ENGINEERING
Ph.D. Program
SSD: ING-INF/03–TELECOMMUNICATIONS

Final Dissertation

Design and evaluation of technical components building the flexible 5G radio interface supporting heterogeneous use cases

---

by
Grassi Alessandro

Supervisor:

Prof. Ing. Gennaro Boggia

Coordinator of Ph.D. Program:
Prof. Luigi Alfredo Grieco

Cycle XXXI, 01/11/2015-31/10/2018

Department of Electrical and Information Engineering
ELECTRICAL AND INFORMATION ENGINEERING
Ph.D. Program
SSD: ING-INF/03–TELECOMMUNICATIONS

Final Dissertation

Design and evaluation of technical components building the flexible 5G radio interface supporting heterogeneous use cases

---

by
Grassi Alessandro

_____

Referees:
Prof. Andrea Abrardo
Prof. Periklis Chatzimisios

Supervisor:
Prof. Ing. Gennaro Boggia

_____

Coordinator of Ph.D Program:
Prof. Luigi Alfredo Grieco

_____

Cycle XXXI, 01/11/2015-31/10/2018

To my parents, who always support me in what I do.

And to You, who know everything about me and yet you love me.

Acknowledgments

I would like to acknowledge my supervisor, Prof. Ing. Gennaro Boggia, and my colleague, Giuseppe Piro, for all the efforts spent in assisting me on developing each of the research topics included in this thesis.

Contents

List of Figures iv

List of Tables vi

Introduction: Dissertation overview vii

Personal Scientific Contributions x

1 Overview of the 5G technology 1

1.1 The role of mobile technologies 1

1.2 The 5G Vision 3

1.2.1 Previous technologies and their limitations 3

1.2.2 5G principles and requirements 7

1.3 Use cases and challenges 14

1.3.1 Dense urban areas 14

1.3.2 Wide area coverage 15

1.3.3 Massive sensor networks 16

1.3.4 Tactile Internet 16

1.3.5 Self-driving cars 17

1.3.6 Broadband access at high mobility 17

1.3.7 Multicast and broadcast services 18

1.3.8 E-health 19

2 Massive MIMO for broadband access 20

2.1 Massive MIMO highlights 20

2.1.1 TDD Massive MIMO 21
2.1.2 FDD Massive MIMO 22
2.2 JSDM and beam coordination for Massive MIMO downlink 24
2.2.1 Formulation of JSDM and beam coordination 25
2.2.2 Performance evaluation methodology 33
2.2.3 Simulation results 37
2.3 Density-based user clustering for JSDM 47
2.3.1 Original and improved DBSCAN for JSDM 49
2.3.2 Numerical results 57
2.4 Game-theoretic resource allocation for uplink with Massive MIMO 62
2.4.1 Formulation of the resource allocation scheme 62
2.4.2 Performance evaluation methodology 70
2.4.3 Simulation results 73
3 Predictor antennas for high-mobility Internet access 84
3.1 Description of SRTA technique 85
3.2 System-level simulation and results 89
4 Enhanced Random Access for massive MTC communications 93
4.1 Basic and extended RACH procedure 94
4.2 Evaluation of the extended RACH procedure 96
5 Advancements for multicast and broadcast 101
5.1 Baseline and enhanced broadcasting techniques 102
5.2 Simulation and results 104
5.2.1 Local broadcast services 105
5.2.2 Regional and national broadcast services 108
6 Conclusions and Future Research Directions 110

Appendix.....112

References.....117

List of Figures

Figure 1.1: use of Bandwidth Parts to permit the coexistence of different, unknown, and/or still unspecified technologies 9
Figure 1.2: Example realization of the 5G air interface, with flexible allocation of multiple services. 11
Figure 1.3: High-level 5G architecture. Source: NGMN 5G White Paper. 12
Figure 1.4: Network slicing in 5G. Source: NGMN 5G White Paper. 14
Figure 2.1: JSDM block diagram 29
Figure 2.2: Example of coverage profile for beam coordination. 32
Figure 2.3: Example of the SINR map for LTE-A, Massive MIMO, and beam coordination 39
Figure 2.4: Cumulative distribution function of SINR (ISD = 200 m) 40
Figure 2.5: Cumulative distribution function of SINR (ISD = 600 m) 40
Figure 2.6: Cumulative distribution function of SINR (ISD = 1000 m) 41
Figure 2.7: User experienced data rate (ISD = 200 m) 42
Figure 2.8: User experienced data rate (ISD = 600 m) 43
Figure 2.9: User experienced data rate (ISD = 200 m) 43
Figure 2.10: Cumulative distribution function of user throughput (ISD = 200 m) 44
Figure 2.11: Cumulative distribution function of user throughput (ISD = 200 m) 44
Figure 2.12: Cumulative distribution function of user throughput (ISD = 200 m) 44
Figure 2.13: Effect of CSI acquisition threshold 47
Figure 2.14: Positions of the mobile users. 58
Figure 2.15: Number of detected groups and served users. 59
Figure 2.16: Average user SE and sum SE. 60
Figure 2.17: Cumulative Density Function (CDF) of user SE [bps/hz]. 61
Figure 2.18: Algorithm flow chart 67

Figure 2.19: Average energy efficiency. 75
Figure 2.20: Average application goodput. 79
Figure 2.21: Peak data rate. 80
Figure 2.22: Average spectral efficiency. 81
Figure 2.23: Outage probability (3 km/h user speed). 82
Figure 2.24: Outage probability (120 km/h user speed). 83
Figure 3.1: Example of the antenna configuration for SRTA-PI. 86
Figure 3.2: Illustration of the SRTA-PI approach. 87
Figure 3.3: Total throughput with 2 receiving units on the train. 90
Figure 3.4: Total throughput with 4 receiving units on the train. 90
Figure 3.5: Total throughput with 8 receiving units on the train. 91
Figure 3.6: Distribution of the MCS index for different speeds and technologies. 92
Figure 4.1: Baseline RACH procedure for LTE. 94
Figure 4.2: Enhanced RACH procedure. 96
Figure 4.3: Success probability of the complete transmission. 98
Figure 4.4: Average delay from initial request to successful data transmission. 99
Figure 4.5: Profile of the collision rate occurring during simulations. 100
Figure 5.1: layout of simulation area. 103
Figure 5.2: Results for user experienced data rate. 106
Figure 5.3: Results for latency. 106
Figure 5.4: CDF of average user throughput. 107
Figure 5.5: Number of additional RBs used for HARQ. 107
Figure 5.6: Results for Packet Loss Ratio. 108
Figure 5.7: breakout of large MBSFN area. 109

List of Tables

Table 2.1: Symbol reference 26
Table 2.2: simulation parameters 35
Table 2.3: Traffic density [Gbps/km^2] 45
Table 2.4: Cell throughput [Gbps] 46
Table 2.5: List of mathematical symbols 64
Table 2.6: Simulation parameters..... 72
Table 2.7: Average transmission power registered by mobile terminals [mW]. 74
Table 2.8: Average power consumption registered by the base station [W]... 75
Table 2.9: Average MCS..... 76
Table 2.10: Average number of sub-channels per time slot..... 78

Introduction: Dissertation overview

In the last decades, mobile communications evolved from a limited and expensive technology to a commodity that is available to billions of people and is used every day. Their scope changed as well: first it was limited to voice calls, then messaging was introduced, and later multimedia communications with images and video were added. As of 2018, the prevailing use of cellular networks is to connect people to the Internet, allowing them to work, look for information, communicate, and many other activities. This has been possible tanks to fourth generation (4G) technologies such as LTE and LTE-Advanced.

This landscape is ready to evolve once again: the fifth generation (5G) of cellular networks is almost ready for the roll-out of its first iteration, with the goal of further improving the performance of existing services and paving the road for completely new scenarios. Current plans include bringing broadband connectivity to remote rural areas, supporting huge numbers of small battery-powered sensors and IoT devices, providing fast Internet on very high-speed trains, and much more. But more importantly, 5G is being designed as a flexible, modular, and extensible technology, that can be improved at will when the need arises, to be able to support new scenarios that do not even exist today.

Given these premises, the research activity carried out during the Ph.D. consisted mostly in the identification, modeling, and evaluation of promising techniques suitable for specific use cases of future 5G networks. Most of the work has been conducted as part of the EU H2020 FANTASTIC-5G project (Flexible Air iNTerfAce for Scalable service delivery wiThin wIreless Communication networks of the 5th Generation), whose main objective has been the design of a flexible air interface for the 5G, with focus on the lowest layers of the protocol stack and the integration of many components in a unified framework.

Chapter 1 of this thesis offers an overall description of the 5G technology, of

the vision behind it, of the main use cases being considered at the present time, together with their corresponding challenges.

Chapter 2 introduces the Massive MIMO technology, which employs a large number of base station antennas and advanced signal processing to offer significant advantages over traditional systems with a limited number of antennas. First, the problem of implementing the technology in FDD mode is considered, as the original formulation requires the TDD mode, but it is inconvenient in most existing networks. The two-stage JSDM precoding technique is explored as a possible solution, together with a beam coordination technique to reduce the interference, and both techniques are shown to offer a large advantage over the legacy LTE-Advanced in urban and rural scenarios. Subsequently, a specific aspect of the JSDM technique is investigated, that is the identification of suitable user cluster for precoding. This activity has been pursued during a research period abroad, at the Fraunhofer Heinrich Hertz Institute in Berlin, Germany. An existing density-based clustering approach has been extended to cover a larger variety of operating conditions, ranging from small isolated user clusters to large crowds that need to be broken down. The resulting approach shows improved performance and compliance with the intended goals. The last part of this chapter is dedicated to the description of a distributed resource allocation algorithm suitable for the uplink of a Massive MIMO network, which allows the users to autonomously select a subset of resources and a power allocation profile, with the goal of maximizing the energy efficiency while respecting QoS constraints. The algorithm is shown to work as intended and provide significant improvements over two different reference strategies.

Next, Chapter 3 tackles the problem of providing broadband Internet access on very high-speed trains, despite a number of obstacles due to the mobility. Specifically, as the feedback about the channel state becomes outdated very quickly, the Predictor Antenna concept is exploited to provide a more reliable channel estimation, resulting in up to 100% gain at a speed of 500 km/h.

Chapter 4 deals with massive sensor deployments, where a large number of devices may try to access the network in a short time span, resulting in many collisions and failures during the initial connection phase. An extended version of

the LTE access protocol is described, which is able to support a larger number of devices, as shown in the numerical evaluation.

Chapter 5 proposes enhancements for the multicast/broadcast operation available in LTE, which are based on the introduction of an uplink channel associated with the downlink transmission. With this hypothesis, it is possible to implement a dynamic adaptation of the modulation and coding levels, as well as requesting retransmission of data packets that are not received correctly. As shown in the evaluation, these tools can make the broadcast transmission more effective and more reliable.

Finally, the Appendix describes the most important aspects and developments of the LTE-Sim simulation tool, that has been used in most of the works presented in this Ph.D. thesis.

Personal Scientific Contributions

Scientific contributions leading to publications during the PhD work are listed here. They have been accepted for publication in international journals and conferences, or they have been submitted and are still waiting for acceptance.

International journals

- A. Grassi, G. Piro, G. Bacci, and G. Boggia, “Uplink resource management in 5G: when a distributed and energy-efficient solution meets power and QoS constraints”, *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5176 - 5189, 2017.
- A. Grassi, G. Piro, and G. Boggia, “A look at Random Access for Machine-Type Communications in 5G cellular networks”, *Internet Technology Letters*, no. 1, Jan., 2018.
- A. Grassi, G. Piro, G. Boggia, M. Kurras, W. Zirwas, R. SivaSiva Ganesan, K. Pedersen, and L. Thiele," Massive MIMO Interference Coordination for 5G Broadband Access: Integration and System Level Study", *Computer Networks (Elsevier)*, vol. 147, pp. 191-203, October, 2018, To be published, doi: 10.1016/j.comnet.2018.10.012.

International conferences

- Alessandro Grassi, Giuseppe Piro, Domenico Striccoli, Roberto Fantini, and Gennaro Boggia, “Multicast and Broadcast in 5G Networks: What Lessons Learned From 4G Experience?”, *Proc. of European Conf. on Networks and Communications, EUCNC*, Athens, Greece, Jun., 2016.

- Sergio Martiradonna, Alessandro Grassi, Giuseppe Piro, Luigi Alfredo Grieco, and Gennaro Boggia, “An open source platform for exploring NB-IoT system performance”, *Proc. of IEEE European Wireless (EW)*, Catania, Italy, May, 2018.

- Alessandro Grassi, Martin Kurras, Giuseppe Piro, Gennaro Boggia,

Stephan Fahse, and Lars Thiele, "Density Based Clustering for Downlink User Grouping in FDD Massive MIMO", *Proc. of IEEE European Wireless (EW)*, Catania, Italy, May, 2018.

- Alessandro Grassi, Giuseppe Piro, Gennaro Boggia, and Dinh-Thuy Phan-Huy, "A system level evaluation of SRTA-PI transmission scheme in the high-speed train use case", *Proc. of IEEE International Conference on Telecommunications (ICT)*, Saint-Malo, France, June, 2018.

- D. Di Lecce, A. Grassi, G. Piro, and G. Boggia, "Boosting Energy Efficiency of NB-IoT Cellular Networks Through Cooperative Relaying", *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Bologna, Italy, Sep., 2018.

Contributions to books

- M. Maternia, J. F. Monserrat, D. Martin-Sacristan, Y. Wu, C. Yang, M. Boldi, Y. Bao, F. Pujol, G. Piro, G. Boggia, A. Grassi, H. Scheck, I. P. Belikaidis, A. Georgakopoulos, K. Demesticha, and P. Demestichas, "Performance and Energy-Efficiency and Techno-Economic Assessment", *5G System Design: Architectural and Functional Considerations and Long Term Research*, 2018.

1 Overview of the 5G technology

This chapter offers a general introduction to the 5G mobile technology. The main motivations for its development are presented, starting from the capabilities and limitations of the current cellular standards, as well as from established and emerging patterns in user demands. The key features envisioned for the upcoming 5G standard are then described, with an emphasis on the corresponding design objectives. The main use cases targeted at the current time are also listed and explained, along with their fundamental challenges.

1.1 The role of mobile technologies

One of the most remarkable achievements of the modern society is the ability of people to always be in contact, with potentially anyone else, at any time. This a result emerging from various layers of technology built on top of each other: first, the landline telephone allowed people to talk to each other in real time, even at great distances. Later, the first mobile phones offered the very same ability, but without being constrained to stay at home or at a public phone. At about the same time, the use of the Internet started to become commonplace in the home of an increasing number of people: this permitted the exchange of information other than just voice and fueled new forms of communication, such as chats and online forums. Modern cellular networks combine the advantages of mobile phones and the Internet: many kinds of communication are possible, from text to voice to exchanging pictures and videos, and all of this is available from commodity smartphones that can be used almost everywhere.

This kind of evolution may look quite linear and simple from an outsider's point of view, but nonetheless it required a great number of technical advances. Especially

regarding mobile networks, moving from the rather static requirements of circuit-switched voice to packet-switched broadband Internet, by using constrained spectrum resources, has been extremely challenging [1]. Moreover, the definition of "broadband Internet" is a moving target with increasingly higher requirements, thus requiring constant efforts just to avoid falling out-of-date. Thus, actually improving the user experience demands even more hard work.

On the plus side, pursuing such difficult targets is beneficial in many direct and indirect ways. For the last few decades, the manufacturing of mobile phones and corresponding network equipment saw a continuous improvement in design and production techniques, so that every new iteration is considerably more capable than the previous one [2]. Additionally, many scientific advancements were achieved for the purpose of developing better and faster wireless connectivity, e.g. in the areas of modulation, channel coding, MIMO, and so on. Moreover, the convenience of using similar and compatible cellular technologies across different countries encouraged a continuous collaboration among the involved parties, which is best represented by the 3GPP (3rd Generation Partnership Project) consortium.

One of the most relevant advantage of the ubiquitous connectivity provided by mobile networks is their role as a platform for a large number of digital services. Applications such as instant messaging, social networks, and music streaming are extremely more useful when they can be accessed from pocket devices during idle times, such as commutes and waiting lines, as opposed to requiring a full-fledged computer and a home Wi-Fi connection. Most questions that may arise during a conversation can be answered timely with a simple online search. Some other services, such as real-time directions with GPS positioning, only make sense on a mobile device.

In line with this latest observation, it is not surprising to know that the most important goal of 5G is to continue this trend, by extending the cellular technology's usefulness as a platform even more [3]. In this view, 5G has been designed to offer far better guarantees than any of the previous technologies, either in terms of data rate, coverage, reliability, and more. More importantly, 5G will be able to tune such advanced capabilities for each specific service, so as to always offer its best compromise in terms of performance and requirements.

1.2 The 5G Vision

To understand the design criteria and requirements of 5G, it is useful to first consider the previous technologies, their evolution, and their limitations. In general, every new generation of mobile standards was designed to overcome the shortcomings of its predecessors and steer usage patterns towards new directions, and the 5G is no exception to this rule.

1.2.1 Previous technologies and their limitations

The first generation (1G) of mobile technologies includes AMPS (Advanced Mobile Phone System) in north America and Australia, NMT (Nordic Mobile Telephone) in the northern and eastern Europe, and TACS (Total Access Communications Systems) in the United Kingdom and central Europe. They were launched in the early 1980-1990 decade and remained in operation throughout the 1990-2000 period. Voice transmission was entirely analog, using Frequency Division Multiple Access (FDMA) with a technology-dependent channel size, which resulted in large size and power consumption of the mobile equipment [2]. International roaming was not possible at all, as each operator's network was isolated from every other's. More importantly, there was no security guarantee, as both voice and signaling information were sent out unencrypted and could be easily intercepted with the help of a radio scanner. Over the years, this resulted in a growing number of frauds at the expense of the operators and their customers. Eventually, these networks were completely replaced by superior digital systems.

The second generation (2G) includes a few different systems: GSM (Global System for Mobile communication) and D-AMPS (Digital AMPS) were based on a combination of FDMA and Time Division Multiple Access (TDMA), while IS-95 (Interim Standard 95) was the first system to adopt Code Division Multiple Access (CDMA). They were entirely digital, meaning that the voice signal was digitized,

compressed, encrypted, and only then transmitted, putting an end to easy eavesdropping. 2G services were implemented on a large scale during the 1990-2000 decade.

GSM also supported roaming, meaning that a mobile user could make and receive calls when moving outside its operator's coverage, by using another operator's network. Roaming was possible on the condition that the two operators had a roaming agreement in place, and it incurred higher costs for the user, but it greatly extended the usefulness of mobile phones, so they could be used in international travels. Eventually GSM became the dominant mobile technology, replacing the alternative almost everywhere, and roaming became available worldwide as a consequence [4].

Still, 2G systems were initially conceived only for the purpose of placing voice calls. Actually, a GSM voice channel could be used to make an internet connection with the so-called Circuit-Switched Data (CSD) mode, but the speed was very limited (9,6 kbps), and the cost was high. It is only with the introduction of the General Packet Radio System (GPRS) that packet-switched data connection is introduced, along with many other innovations. In fact, GPRS allowed one user to receive data on more than one time slot when possible, and the channel coding is changed automatically to provide higher throughput or higher robustness depending on the current channel quality [5]. The Enhanced Data rates for GSM Evolution (EDGE) technology is a further improvement of GPRS which introduces higher-level modulations, resulting in even higher throughputs. Nevertheless, further improvements were difficult because of backwards compatibility, and the voice and data services were mutually exclusive.

As for the third generation (3G), it was conceived right from the beginning with the capability to provide voice and packet data services, even simultaneously, which was not possible at all in GSM-based networks. The main technologies are called Universal Mobile Telecommunication System (UMTS) and CDMA2000, and they are both based on CDMA [6]. The main design goal of 3G standards was to popularize multimedia services such as Multimedia Message Service (MMS), video telephony, video streaming and mobile Internet in general. In fact, subsequent iterations of these standards focused on increasing the peak throughput, which grew

from less than 1 Mbps to more than 100 Mbps, and the overall capacity. During the 2000-2010 decade, people started using mobile Internet at growing rates, to the point that 3G network approached saturation during peak hours of the day. The realization that 3G technology would quickly become insufficient for the exponential increase of mobile traffic prompted the development of a new and even more capable technology.

The fourth generation (4G) of mobile networks converged towards a single technology that has been adopted worldwide, that is the Long Term Evolution (LTE). It was designed to provide broadband Internet connectivity, greatly surpassing the capabilities of 3G technologies, and to realize a completely packet switched architecture, thus dropping support for circuit-switched operation. LTE is based on the Orthogonal Frequency Division Multiplexing (OFDM) modulation, which provides higher throughput than previous solutions. Already from the first release, identified as Release-8, it has been capable of peak data rates up to 300 Mbps per cell, using advanced concepts such as Adaptive Modulation and Coding (AMC) and Multiple-Input Multiple-Output (MIMO) [7]. The first networks appeared around 2010, with a rapid increase toward widespread adoption.

In LTE, the available radio resources are divided into small time-frequency units, called Resource Blocks, which can be flexibly assigned to any connected user. This is done by the packet scheduler, which decides which user gets which RBs based on a wide range of information: current throughput, past throughput, fairness, priority, real-time requirements, and so on. In fact, designing the packet scheduler is a crucial task for the successful operation of an LTE network.

Subsequent evolutions of the LTE technology introduced even more advanced capabilities, with the purpose of directly increasing the peak throughout, reducing the interference levels, or adapting the network's operation to new scenarios. Notable cornerstones are named LTE-Advanced (LTE-A) [8], corresponding to Release-10, and LTE-Advanced Pro (LTE-A Pro) [9], at Release-13. Some of the most important enhancements are:

- 8-antenna transmission: while LTE allows up to 4 antennas at both transmitter and receiver in the downlink direction, LTE-A increases that to 8 antennas, which doubles the potential throughput.
- Carrier Aggregation: it allows mobile terminals to connect to multiple carriers and download/upload data simultaneously from/to all of them, thus multiplying the peak throughput. LTE-A allows 5 aggregated carriers, and LTE-A Pro extends it to 32.
- Coordinated Multi-Point (CoMP): with this technique, two or more base stations can coordinate and send the same signal to a given user, so that it adds up constructively, resulting in a higher useful signal and a lower interference.
- Heterogeneous Networks: low-power and low-coverage nodes can be deployed alongside traditional macro base station, to improve capacity and coverage in selected areas and to offload the macro base stations at the same time.
- Relay Nodes: these nodes are deployed in specific areas to enhance the coverage, but instead of having a cabled connection to the operator's core network, they use a wireless link to an existing base station, thus acting as both a mobile and a base station.
- Enhancements for Machine-Type Communications (MTC): these allow small, autonomous, battery-powered device to operate with lower power consumptions, thus extending their lifetime, by allowing longer periods of inactivity. They also provide increased coverage for devices that are placed in difficult locations, and/or use a single receive antennas to reduce manufacturing costs.

Despite its impressive capabilities, LTE and its evolutions still have their shortcomings. Most notably, further evolution is increasingly limited by the requirement of backwards compatibility with legacy equipment. Additionally, many parameters of the physical layer have been fixed in the first release and can't be changed because of the aforementioned compatibility issue. They were chosen as a compromise between different competing requirements, such as latency, overhead, and channel estimation accuracy. However, some emerging use cases cannot be

adequately addressed with this approach, as they tend to prioritize one or few performance indicators above all the others. For example, LTE cannot accommodate services requiring extremely low latency and extremely high reliability without breaking all legacy equipment. This has been one of the main motivators behind the design of 5G.

1.2.2 5G principles and requirements

Mobile connectivity has fundamentally changed the way people interact with each other and with the vast amount of information available on the Internet, making everything more accessible and seamless. However, there are still many areas where universal digital connectivity is not much used yet, but it could make a huge difference. Some of these include home appliances that could be turned into a "smart home", vehicles which could communicate and cooperate for assisted or autonomous driving, and factories which could be automated and operated remotely, among many others. The main goal of 5G, which is expected to launch commercially around 2020, is to be an almost universal communication platform, able to adapt to many different scenarios [10]. However, this is not trivial at all, because all these scenarios have extremely challenging requirements even for the current state-of-the art technology, and each one is demanding in a different way. In some cases, the network is required to provide a very high throughput. In other cases, it is necessary to support a huge number of connected devices. For yet another application, the requirement could be that the latency is very low. In the end, designing a single Radio Access Technology (RAT) to meet all the requirements simultaneously (like LTE at its time) is simply not possible, for both economical and practical reasons. On top of that, new use cases will emerge in the future and will need to be considered as well.

Because of these compelling reasons, 5G has been designed as a general framework where multiple RATs can be instantiated and coexist in an organized

fashion. At the physical layer, this means that an operator will have a single 5G carrier, but this will be structured internally in one or more blocks with different characteristics, with each block dedicated to a given subset of devices with specific requirements, which are addressed using a suitable technology with appropriate operational parameters [11]. This is apparent both in related research project such as FANTASTIC-5G and in 3GPP specifications.

The FANTASTIC-5G project was part of the European Union Horizon-2020 program [12]. Its main goal was the definition of a unified air interface for the frequency range under 6 GHz, including the physical (PHY) and Medium Access Control (MAC) layers, with the capability to meet the aforementioned requirements of flexibility and extensibility. Five core services were identified, which represent typical patterns in the utilization of the mobile networks, so that the numerous use cases can be described as instances of one or more core services, possibly in specific environmental conditions. The five core services are: Mobile Broadband (MBB), Vehicular to Everything (V2X), Massive Machine Communications (MMC), Mission Critical Communications (MCC), Broadcast and Multicast Services (BMS). For each of these core service, a specific and optimized air interface was developed, including all the required technologies. Finally, all these dedicated air interfaces were integrated into a single framework where they can be flexibly instantiated depending on the instantaneous network conditions. This two-step design methodology was important, because it provided a backup strategy if the unified framework turned out to be not feasible or not efficient in practice: in that case, the individual air interfaces tailored for each core service would have been used separately, without integration.

The idea of providing multiple air interfaces within the same structure can also be found in Release-15 3GPP specifications, which define the first standardized version of the 5G technology [13]. Release-15 defines the concept of Bandwidth Parts (BWPs), shown in Figure 1.1, that are sub-parts of a 5G carrier dedicated to a specific service. The partitioning into BWPs is not fixed, but it can be changed frequently according to the instantaneous requirements. The BWP solution allows for a very flexible operation, as well as providing a high degree of future-proofness: supporting a new kind of service requires the definition and implementation of a

new BWP type, but it doesn't affect previously existing services, which can continue to work for as long as needed.

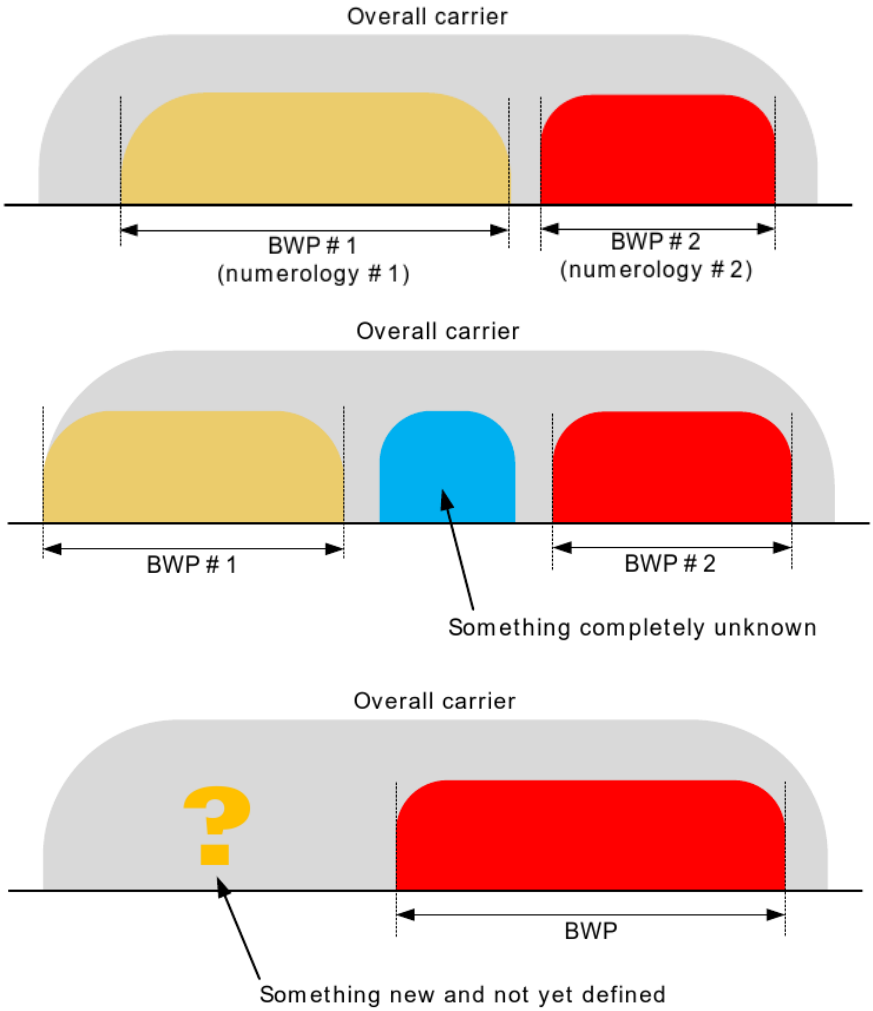


Figure 1.1: use of Bandwidth Parts to permit the coexistence of different, unknown, and/or still unspecified technologies
Source: "Understanding the 5G NR Physical Layer", Keysight Technologies, November 1st, 2017.

Some of the main parameters that characterize a BWP are the following:

- Subcarrier spacing: for OFDM and similar waveforms, this refers to the distance in frequency domain between the individual tones. As a reference, LTE usually allows a single value of 15 kHz. Lower values

are more robust to frequency-selective fading, and they may also allow better coverage in uplink for power-limited devices. On the other hand, higher values are more resistant to Doppler shift, and they are also more suitable at higher frequencies, especially millimeter waves.

- RB size: it defines the size of a RB in the frequency domain. Lower values allow more scheduling flexibility, but also incur a higher signaling overhead. The optimal value may also change depending on the subcarrier spacing.
- Sub-frame size: this indicates the duration of a sub-frame in the time domain. Together with RB size, it defines the minimum amount of resources that can be scheduled. A lower size permits a lower latency, but it has more overhead. 5G should allow a flexible composition of different sub-frame sizes even in the same BWP, to better meet the requirements of each user.
- Waveform: OFDM has been used with great success in LTE and many other wireless technologies, and it is thus considered as a baseline option for 5G. However, there are a number of other possibilities that are based on it, typically by performing some kind of filtering on individual pulses or entire sub-bands. Some examples include Filter-Bank Multi-Carrier (FBMC), Pulse-shaped OFDM (P-OFDM), and Universally Filtered OFDM (UF-OFDM).
- Channel estimation: it defines the resources and methods used for downlink channel estimation at the mobile terminal. It could include cell-specific or user-specific signals, with different densities of pilot symbols in frequency and time. Higher densities allow a better estimation even in frequency-selective or time-selective environments, but they also result in a higher overhead.

A possible realization of such a multi-part air interfaces, evolving over short time spans to accommodate different traffic types as needed, is shown in Figure 1.2. According to the requirements, different time-frequency blocks can be configured with diverse transmission parameters, techniques, or even entirely different communication paradigms, e.g. deploying a broadcast transmission alongside with

Device-to-Device (D2D) resources, while also providing typical internet access with high-order MIMO. Overlapping of two or more services is also possible, e.g. low-latency traffic can "steal" resources from regular traffic via puncturing.

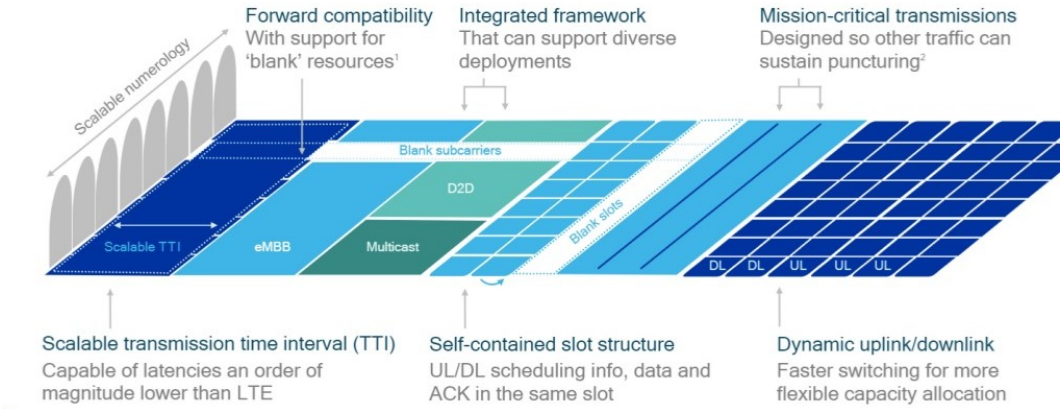


Figure 1.2: Example realization of the 5G air interface, with flexible allocation of multiple services.

The flexibility envisioned for 5G does not affect only the physical layer, but it encompasses the entire system. In fact, in many situations it is not convenient or effective to have a monolithic architecture like in LTE, where the network is managed exclusively by the operator. This is especially true for companies that need to offer specific services on top of the 5G network with strong and domain-specific requirements, such as vertical industries and Over-The-Top (OTT) providers. These parties should be granted the ability to define and deploy suitable configurations, and to do so autonomously and in a timely manner, without continuous and explicit intervention of the operator. The most promising way to achieve this goal is to have a completely virtualized network, where resources and capabilities can be instantiated on-demand using appropriate APIs that are made available to third parties.

A possible incarnation of this concept is shown in Figure 1.3, depicting the architecture envisioned by the NGMN (Next Generation Mobile Networks) Alliance. In this concept, the 5G network is organized in three layers and an End-

to-End management and orchestration entity [3]. The lowest layer, called "Infrastructure resources layer", includes all the resources such as computing nodes, network nodes, access nodes. 5G mobile devices may also act as relays, computing resources, or storage resources, depending on the situation, so they are also included. These are all exposed to the intermediate layer and to the management and orchestration entity via virtualization paradigms, most notably Software Defined Networking (SDN) and Network Functions Virtualization (NFV). The intermediate layer is the "Business enablement layer", and it combines and configures the virtualized resources to provide a library of modular and reusable higher-level functions. Some of these may also be present in more than one version, featuring different parameters and performance, and all the functions are exposed to the highest layer and to the management and orchestration entity via suitable APIs. The top layer, called the "Business application layer", contains actual applications that the operator or other third parties can use on the 5G network. Finally, the E2E management and orchestration entity coordinates all the layers to translate the use cases and business models into actual configurations, functions, and slices of the 5G network. For a given application, it selects, configures, and instantiates the relevant functions on the infrastructure resources. It is also entry point for 3rd parties to configure their own services and slices on a 5G system.

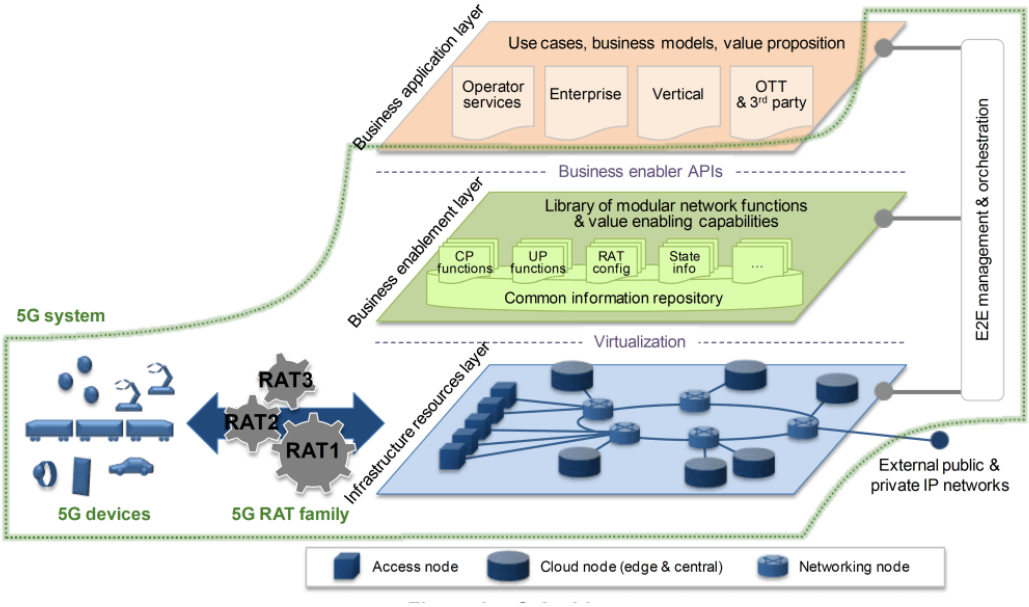


Figure 1.3: High-level 5G architecture. Source: NGMN 5G White Paper.

An important concept for this kind of architecture is that of “network slicing”. A 5G slice includes every element that is required to run a given application at all level of the architecture, including network functions, configurations, RATs, and resources. Crucially, a slice should offer only the functionality required for its use case, leaving out every unnecessary part. This kind of flexibility is the most important feature to handle many different use cases.

An example of a 5G network running multiple slices simultaneously is shown in Figure 1.4. Each slice uses a different subset of the available resources and different RATs. Some physical resources are also shared between slices: this is important for efficient utilization of the virtualized resources, but also to provide functionalities that are necessarily shared among slices, such as scheduling. Separation of the control plane and user plane functions is also required, with open interfaces between them, in accordance to the SDN paradigm. Similarly, the interfaces between most other functions should be open and well-defined to allow multi-vendor interoperability. Some examples include access-specific versus access-agnostic functions and the remote radio units versus the baseband processing.

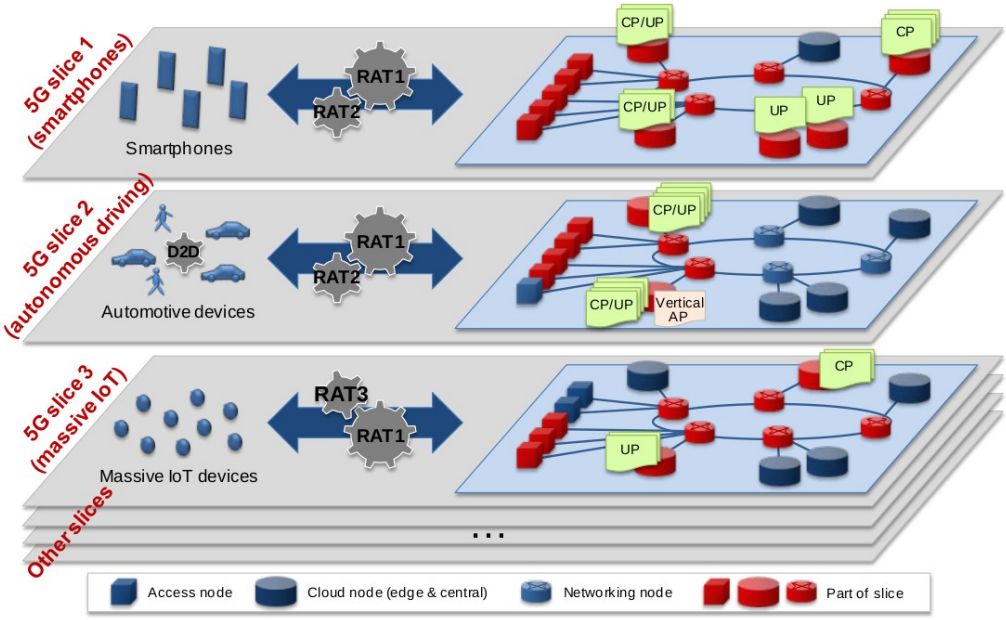


Figure 1.4: Network slicing in 5G. Source: NGMN 5G White Paper.

1.3 Use cases and challenges

This section presents some of the main applications and services that are expected to be delivered on 5G networks, along with the associated requirements and challenging aspects. For some of these, innovative solutions will be presented in the following chapters. Note, however, that the list is not exhaustive, and that completely new scenarios may open up in the future.

1.3.1 Dense urban areas

This use case relates to areas with high population density, such as city centers and offices. Since the introduction of mobile Internet, the amount of traffic has been growing exponentially, and it is expected to continue this trend in the 2020-2030 decade as well. The applications driving such growth are multiple: HD video streaming for personal entertainment, data-intensive cloud applications, video conferencing, and virtual presence, just to name a few. These increasing traffic requirements, together with the high population density, will results in exceptionally

high capacity requirements for 5G mobile networks. In order to provide a satisfactory and consistent user experience, the perceived bit rate should never fall below a minimum value, regardless of peak hours and network congestion. Around 2020, the minimum effective throughput to qualify a "broadband" connection as such will be in the range 50-100 Mbps. Note that this refers to the throughput effectively perceived by the user, not the theoretical peak throughput, which is already much higher today but totally unrealistic in describing the real performance. When combining this requirement with the user density, it is easy to come up with multiple Gbps/km^2 of expected traffic density.

So far, the usual approach to increase capacity has been to increase cell density, however this will not be possible or sufficient in the 5G era. Instead, a mix of new approaches will be used, such as small cells, Massive MIMO, and CoMP [14]. In this situation, proper radio planning will be extremely difficult and interference levels may become exceedingly high, unless proper management techniques are put into place.

1.3.2 Wide area coverage

New developments in mobile networks are often deployed in dense cities first, because of the larger number of potential customers. Conversely, locations with lower population density, such as rural areas, are only considered at later times, or sometimes not at all, because of the lower economic returns. Thus, they remain in a persistent state of digital divide. One of the goals of 5G is to be commercially viable in these remote areas as well [15].

Like in the dense urban scenario, the network should be able to consistently provide broadband services to all the active users, even at the cell edge. The effective perceived throughput should be at least 50 Mbps. The major challenge is due to the high distance from the base stations and the correspondingly large path loss, together with the necessity to maintain a low cost of the infrastructure to allow

a profitable deployment despite the low number of customers. Massive MIMO can be an important tool towards this goal.

1.3.3 Massive sensor networks

An increasingly larger number of small and autonomous devices is being deployed worldwide, and more will follow in the coming years. These include sensor for environmental monitoring (e.g. temperature, humidity, pollution), metering devices for utilities (such as water and electricity), and combined sensors/actuators for applications such as smart city lights and smart agriculture. Together, these and many other kinds of devices are building up the Internet of Things (IoT), which is expected to bring many changes and innovations in the modern society.

Many of these devices will be connected to cellular networks, but they have very different requirements from typical mobile devices. In most cases, they will only transmit small data packets with large idle times in-between, therefore throughput and latency are not a concern. Instead, it is crucial to ensure a minimal power consumption, as most devices will be battery-powered with a target lifetime of several years. Moreover, a huge number of devices is expected to come into operation, and 5G network may be overwhelmed unless they are properly designed for this [16].

1.3.4 Tactile Internet

This use case pertains interactions that require a strong interactivity and real-time component. Common examples include manipulation of remote or virtual objects. Some kind of factories can greatly reduce safety risks for its employees by performing some operations with remotely-controlled robots. Virtual presence applications can allow two or more people to interact in a shared virtual environment. Also, augmented reality applications can involve the manipulation of virtual objects that are stored on an online server.

In all of these cases, the interaction needs to be characterized by a high degree

of responsiveness and smoothness. The human mind perceives a reaction as immediate only if it happens within a few milliseconds from its cause, thus the end-to-end latency budget is extremely limited. Therefore, the radio access part should require less than 1 ms on its own, to leave room for processing and other tasks. Similarly, the performance should be extremely stable regardless of the external conditions, so as to maintain a feeling of concreteness. Packet losses and subsequent retransmissions should be reduced to an absolute minimum [17].

1.3.5 Self-driving cars

In the next few years, many applications involving semi-autonomous or fully autonomous cars will appear. These include, for example, assisted breaking or steering to prevent accidents, automatic traffic management to reduce jams, identification and reservation of parking spots. Platooning can be used to reduce fuel consumption and increase road safety and efficiency. Moreover, communication between vehicles and vulnerable road users (e.g. pedestrian and cyclists) can result in greater safety for the latter.

All these applications rely on timely and reliable information exchanges between the involved parties, so that the appropriate measures can be taken quickly. This is further complicated by the mobility of the vehicle, that can reach substantial speeds in some cases (150-200 km/h or even more) which require even faster response speeds. In many cases, the D2D communication paradigm is more appealing than the traditional infrastructure-based approach because of the hop gain and proximity gain. However, it is also more difficult to manage properly, especially in a dynamic environment such as urban traffic [18].

1.3.6 Broadband access at high mobility

As broadband mobile Internet becomes more and more commonplace, people

expect to find it anywhere they go. This includes, for example, trains connecting distant cities together. During travel or commutes, passengers should be able to work, watch high-definition video, download large files, and participate in conferences or video-conferences.

While local trains are usually fine in this regard, because they travel at a moderate speed, the situation is different for high-speed trains, from 350 km/h and up. At these speed, current mobile technology approaches or exceeds the limits of its operating conditions. Even when it remains functional, performance is seriously degraded due to inaccurate channel estimation and high levels of Doppler spread [19]. Moreover, long-distance trains often cross low-population areas where mobile infrastructure is sparse, thus worsening the problem.

Another interesting scenario is that of aircrafts, which currently have no service or only very expensive satellite service. Companies will want to offer cheaper terrestrial-based connectivity, but this requires very different approaches compared to traditional ground users. For instance, airplanes move in three dimensions, and they need to be tracked continuously with focused beams. Although their trajectory is predictable, coordinating the entirety of air traffic to avoid interference is challenging. Also, propagation conditions are different, as multipath is absent and atmospheric conditions can vary a lot with the altitude [20].

1.3.7 Multicast and broadcast services

While communications are increasingly personalized, there are also cases where a one-to-many or one-to-all distribution model is more appropriate. Moreover, the wireless medium is naturally a shared one, and multicast/broadcast schemes are the best way to take advantage of this characteristic. Notable applications of this idea include video broadcasting in local event such as concerts and sport matches, regional emergency warnings for dangerous weather conditions, and large-scale firmware updates for IoT or automotive devices [21].

The main objective for these scenarios is to achieve a large and consistent coverage. However, this has to be balanced with throughput, as a higher transmission rate (achieved via higher modulation orders and code rates) makes the

reception more difficult for cell-edge users. When multiple cells are involved, they need to be tightly synchronized, and managing multiple partially-overlapping broadcast areas may become difficult. Moreover, the network should have means to detect the presence of users interested in the same content, to decide whether a multicast transmission is appropriate, and to establish it on-the-fly.

1.3.8 E-health

Healthcare is going to take advantage of recent advances in electronics miniaturization and improved connectivity. Patients under observation will be monitored constantly for physiological parameters such as blood pressure, heart rate, body temperature, and glucose levels. Such data will be stored remotely for doctors to analyze and suggest actions. In some cases, countermeasures will be transmitted to wearable actuators, such as insulin injectors. Given the importance of the involved data, the communication layer must be highly reliable, have low latency, and ensure the users' privacy [22].

Another related scenario, that is even more challenging, is remote surgery. It is already done currently on wired connections, but in some situations, it can be required in places where only wireless infrastructure is available, e.g. on ambulances or in disaster rescue scenarios. Clearly, reliability and low latency are even more crucial in such context.

2 Massive MIMO for broadband access

As stated in Chapter 1, 5G is required to provide large performance improvements compared to the state of the art. Depending on the specific use cases, these can be related to higher capacity, increased spectral efficiency, and/or improved energy efficiency. A promising technology that can be able to meet such requirements is Massive MIMO, which improves traditional MIMO with a much larger number of antennas. Therefore, this chapter describes this new concept, its importance in future wireless networks, and its application to scenarios involving mobile broadband scenarios. Given the large variety of possible practical implementations, only some of them are described in detail, along with numerical simulation results conducted to prove their effectiveness.

2.1 Massive MIMO highlights

Massive MIMO is a promising transmission technique developed in the last decade, involving a large number of antennas at the base station side [23]. Crucially, these antennas need to be independently controllable by the baseband unit, so as to allow arbitrary precoding depending on the mobile users and their channels. Massive MIMO can be thought as an extension of existing Multiuser MIMO (MU-MIMO) techniques, with the condition of having about one order of magnitude more antennas at the base station side compared to the users' side, thus resulting in antenna arrays with hundreds of antennas. All these excess antennas result in some useful properties, hinted by asymptotic results: with proper channel estimation and precoding, the inter-user interference becomes vanishingly small, and channel variations are significantly reduced both in time and frequency, making advanced scheduling less of a concern. Thus, it becomes possible to serve tens of users simultaneously, compared to few users in current MIMO systems.

The radiated power can be reduced as well, resulting in simpler and cheaper hardware components such as power amplifiers. In the end, the use of Massive MIMO can result in great improvements in terms of spectral and energy efficiency [24].

The details regarding channel estimation and/or feedback, which is essential for proper precoding, may change a lot depending on the operating mode, which is either Time Division Duplexing (TDD) or Frequency Division Duplexing (FDD). The most important proposals in the literature are hereby described separately for the two modes.

2.1.1 TDD Massive MIMO

Many works focusing on Massive MIMO assume a TDD operation to only perform channel estimation in the uplink direction, and then exploit channel reciprocity and use the acquired Channel State Information (CSI) for the downlink as well [25]. This has the advantage that the CSI estimation overhead is proportional to the number of active users, rather than the much higher number of downlink antennas. The use of TDD has been proposed since the seminal papers on Massive MIMO, and the basics remain basically unchanged as long as the TDD mode is maintained. Various proposals are built on top of it, such as the use of 1-bit analog-to-digital converters to reduce hardware costs [26], hardware calibration to ensure true uplink-downlink reciprocity [27], and channel modeling with large antenna arrays [28].

When Massive MIMO is implemented in TDD mode, an important concern is the pilot contamination effect. Specifically, it arises from the limited number of available uplink training sequences, with the subsequent necessity to reuse them at relatively short distances [29]. Differently from other issues, such as fast fading and inter-user interference, pilot contamination persists even if the number of base station antennas grows to infinity. Therefore, it must be approached by other means

[30]. At the time of this writing, interesting techniques have been presented in [29], [31], [32] and [33]. In [29], the problem is solved by introducing a multi-cell Minimum Mean-Square Error (MMSE) precoding approach, which considers the training sequences allocated to all the terminals. The authors of [31] describe a pilot contamination precoding technique that involves multiple cells, but only requires knowledge of the slow-fading coefficients from the other cells. The work in [32] proposes a time-shifting of the pilot signals transmission in different cells, on the basis that non-overlapping transmissions (in time) of the same pilot do not cause contamination. Recently, [33] proposed software-defined Massive MIMO, which employs coordination of multiple remote radio heads, each one with a large number of antennas.

2.1.2 FDD Massive MIMO

Most of the 4G network deployments currently operate in FDD mode and converting them to TDD is not straightforward. Emerging 5G networks will still operate in FDD mode. Therefore, it can be desirable to adapt Massive MIMO to FDD operation as well, but this task is not easy to accomplish. First, with a typical pilot-based channel estimation, the overhead would be proportional to the number of base station antennas. Thus, a simple Massive MIMO implementation could quickly become infeasible [34]. Second, the CSI about the downlink channel should be explicitly fed back to the base station, and there is a trade-off between CSI precision and utilization of the limited uplink resources [35]. In fact, in FDD Massive MIMO, the problem of CSI acquisition becomes much more severe than pilot contamination. Thus, it is generally tackled first, and then pilot contamination is considered in the resulting system, if necessary. However, most strategies proposed to reduce the CSI overhead work by reducing the dimensionality of the channel, which has the positive side effect of reducing the number of required pilots, as well as the pilot contamination effect itself.

The literature proposes different methods to achieve an FDD-based Massive MIMO system. For example, [36] describes a technique to reduce the number of required training signals by exploiting spatially correlated channels. The work in

[37] presents a framework based on Compressive Sensing to collect only partial CSI information at the users and reconstruct the entire channel matrix at the base station, and the idea is further extended in [38]. Also, [39] proposes training schemes with memory, which use not only the last received training signal but also previous ones. In [40], the authors propose a codebook-based feedback quantization with linear (rather than exponential) search complexity. The works in [41] [42] describe a feedback scheme for beamforming-based networks where only relevant components of relevant beams are reported, resulting in high accuracy and low overhead.

The Joint Spatial Division and Multiplexing (JSDM) is described in [43]. It is a two-stage precoding approach: while the first-stage precoder captures the long-term second-order statistics of the channel and reduces the CSI size (both for estimation and uplink feedback), the second-stage precoder captures the short-term channel variations. As stated in [43] a near-optimal choice for the first-stage precoder is a set of unitary Discrete Fourier Transform (DFT) vectors. This configuration is usually referred to as a Grid-of-Beams (GoB), because it produces a number of spatially-oriented beams. As for the second-stage precoder, any standard Multi-User MIMO (MU-MIMO) strategy can be used, including linear filters such as Maximum Ratio Transmission (MRT) and Regularized Zero Forcing (RZF) [44]. With suitable parameters, this strategy reduces the size of the "effective" channel (i.e. as seen after the first-stage precoding) so that the amount of pilot signals and CSI feedback are decreased as well. Moreover, each user only receives a subset of the produced beams with relevant intensity, leaving room for further savings as explained in [45]. Initially studied for a single-cell scenario, the JSDM framework is investigated in a multi-cell fashion in [46] with promising results. However, the analysis only considers a cluster of 3 adjacent sectors from different sites, which does not directly extend to a complete multi-cell scenario.

Few studies can be found in the literature which evaluate the performance of Massive MIMO in a network based on FDD at the system level, in terms of user

and/or cell throughput. One such work is [47], where a multiple types of antenna arrays are evaluated against a baseline Release 8 Long-Term Evolution (LTE) system. However, it uses a simple MRT criterion for the transmission beamforming, and the problem of downlink training overhead is not discussed. Moreover, only a single scenario (in terms of user density and ISD) is investigated. Finally, the baseline setup is only a 2x2 Multiple-Input Multiple-Output (MIMO) system, which is quite out-of-date with respect to current LTE-A capabilities. A similar evaluation is presented in [48], which also compares multiple antenna arrays geometries. While this study adds the effect of different user densities, it presents limitations similar to the previous one. Specifically, no measure is taken to reduce the overhead of channel estimation, and the effect of different Inter-Site Distances (ISDs) is not discussed. Also, no baseline solution is used as a comparison term. 3GPP carried out a calibration study for Full Dimension MIMO (FD-MIMO) in LTE [49], using 2D for elevation as well as azimuth beamforming. However, the assumptions include a smaller number of antennas and a feedback scheme based on LTE codebooks. There is also an ongoing calibration study for beamforming in 5G [50], but only one User Equipment is scheduled at any given time.

2.2 JSDM and beam coordination for Massive MIMO downlink

This section discusses some techniques investigated for the use of Massive MIMO in the downlink direction. These are the JSDM, to allow the use of Massive MIMO in FDD mode, and the inter-cell beam coordination, which is built on top of JSDM and helps reducing the average interference level. The techniques are evaluated for the “Dense urban areas” and “Wide area coverage” use cases: in the former case, the use of Massive MIMO allows an aggressive reuse of the resources and an increment of the spectral efficiency, while in the latter case, Massive MIMO allows focusing the available power in the desired direction, thus fighting the increase in path loss. A simulation campaign for these techniques has been conducted as part of the FANTASTIC-5G project to verify their suitability for the

considered use cases.

Section 2.2.1 presents the theoretical formulation of the proposed techniques, whereas Section 2.2.2 describes the evaluation methodology and Section 2.2.3 comments on the obtained results.

2.2.1 Formulation of JSDM and beam coordination

Throughout this section, calligraphic uppercase letters designate a set of indices, where $\mathcal{A} = \{1, \dots, A\}$, and $A = |A|$ denotes the cardinality of the set. Consequently, upper-case and lower-case letters denote a scalar. In contrast to this, boldface lower-case letters and boldface upper-case letters are used to represent column vectors and matrices, respectively. The element of a matrix is indicated by $[\mathbf{A}]_{i,j}$ where the first subscript means the i-th row and the second subscript the j-th column of matrix $\mathbf{A}$. The element of a vector is indicated by $[\mathbf{a}]_i$. The A-dimensional identity matrix is given by $\mathbf{I}_A$. Moreover, $(\cdot)^*$, $(\cdot)^T$, and $(\cdot)^H$ denote complex conjugate, transpose, and complex conjugate transpose of a vector or matrix. The trace of a square matrix $\mathbf{A}$ of size $A \times A$ is defined as $\mathrm{trace}(\mathbf{A}) = \sum_{a=1}^{A} [\mathbf{A}]_{a,a}$. $\sqrt{(\cdot)}$ denotes the element-wise square root and $\mathbb{E}[\cdot]$ is the expectation operation. In order to improve the readability of this manuscript, the list of mathematical symbols adopted herein is summarized in Table 2.1.

In this work a multi-cell network architecture based on the OFDM physical layer is considered. More specifically, the communication between base stations and mobile terminals, simply referred to as downlink, is taken into account. The considered scenario integrates K mobile terminals and L base stations. Therefore, let $\mathcal{K} = \{1, \dots, K\}$ and $\mathcal{L} = \{1, \dots, L\}$ be the list of mobile terminals and base stations available in the network. Furthermore, each base station $l \in \mathcal{L}$ is equipped with M antennas and each device $k \in \mathcal{K}$ with N antennas.

Focusing the attention to the k-th mobile terminal, let $\mathbf{y}_k \in \mathbb{C}^N$ be the received

signals at the N antennas. It is expressed as:

$$\boldsymbol{y}_k = \sum^{l\in\mathcal{L}} \boldsymbol{H}_{k,l}\sqrt{\boldsymbol{P}_l}\boldsymbol{V}_l\boldsymbol{x}_l + \boldsymbol{n}_k \tag{1}$$

where, $\mathbf{H}_{k,l} \in \mathbb{C}^{N\times M}$ is the channel matrix between the l-th base station and the k-th device, $\mathbf{P}_l \in \mathbb{R}^{M\times M}$ is a diagonal power allocation matrix with the sum constraint $\mathrm{trace}(\mathbf{P}_l) \leq P_{\max}$, $\mathbf{V}_l \in \mathbb{C}^{M\times T_l}$ denotes the downlink precoding matrix of the l-th base station, $\mathbf{x}_l \in \mathbb{C}^{T_l}$ are the transmitted symbols of the l-th base station

Symbol	Meaning
$\mathcal{L}$	Set of base stations in the system
$\mathcal{K}$	Set of mobile terminals in the system
$\mathcal{T}_l$	Set of streams transmitted by the l-th base station
$\mathbf{x}_l$	Signal transmitted from the l-th base station
$\mathbf{V}_l$	Precoding matrix of the l-th base station
$\mathbf{B}_l$	First-stage precoder of the l-th base station
$\mathbf{b}_{l,i}(k)$	Precoder to produce the i-th beam using k antennas
$\mathbf{C}_l$	Second-stage precoder of the l-th base station
$\mathbf{D}_l$	Non-normalized RZF precoding matrix
$\mathbf{E}_l$	Power normalization matrix
$\mathbf{P}_l$	Power allocation matrix of the l-th base station
$\mathbf{H}_{k,l}$	Channel matrix from the l-th base station to the k-th mobile terminal
$\mathbf{H}_l$	Channel matrix from the l-th base station to all its served mobile terminals
$\mathbf{y}_k$	Signal received by the k-th mobile terminal
$\mathbf{n}_k$	Receiver noise at the k-th mobile terminal
$\tilde{\mathbf{h}}_{k,t}$	Effective signal of the t-th stream at the k-th mobile terminal
$\boldsymbol{\vartheta}_{k,t}$	Intra-cell interference for the t-th stream at the k-th mobile terminal
$\mathbf{z}_k$	Inter-cell interference at the k-th mobile terminal
$\mathbf{w}_{k,t}$	Receiver filter for the t-th stream at the k-th mobile terminal
$\mathbf{W}_l$	Estimated compound receive filter of all the mobile terminals served by the l-th base station
$\gamma_{k,t}$	Receive Signal-to-Interference-plus-Noise Ratio (SINR) for the t-th stream at the k-th mobile terminal
σ^2	Thermal noise variance

Table 2.1: Symbol reference

and $n_k \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_N)$ constitutes the Gaussian distributed uncorrelated noise with variance σ^2. The parameter T_l represents the number of spatial multiplexed stream/layers of the l-th base station. Under the assumption that stream t from the l-th base station is allocated to the k-th device, the received signal in Eq. (*1*) can be

rewritten as:

$$\mathbf{y}_{k,t} = \underbrace{\mathbf{H}_{k,l}\sqrt{p_t}\mathbf{v}_{t,l}x_{t,l}}_{\tilde{\mathbf{h}}_{k,t}} + \underbrace{\sum_{j\neq t}^{j\in\mathcal{T}_l}\mathbf{H}_{k,l}\sqrt{p_j}\mathbf{v}_{j,l}x_{j,l}}_{\boldsymbol{\vartheta}_{k,t}} + \underbrace{\sum_{m\neq l}^{m\in\mathcal{L}}\mathbf{H}_{k,m}\sqrt{\mathbf{P}_m}\mathbf{V}_m\mathbf{x}_m + \mathbf{n}_k}_{\mathbf{z}_k} \tag{2}$$

where $\mathcal{T}_l = \{1, \dots, T_l\}$ is the set of stream indices for the l-th base station. In Eq. (*2*), the received signal is divided into three parts: the effective signal corresponding the t-th stream is reported in $\tilde{\mathbf{h}}_{k,t}$; the intra-sector interference caused by streams $j \neq t$ from the l-th base station is reported in $\boldsymbol{\vartheta}_{k,t}$; and finally inter-sector interference plus noise is denoted by $\mathbf{z}_k$.

Under the assumption of Gaussian distributed symbols and an equalizer or combiner denoted by $\mathbf{w}_{k,t} \in \mathbb{C}^N$, the Signal to Interference and Noise Ratio (SINR) of the received signal in Eq. (*2*), that is $\gamma_{k,t}$, is expressed as:

$$\gamma_{k,t} = \frac{\mathbf{w}_{k,t}^{\mathrm{H}}\tilde{\mathbf{h}}_{k,t}\tilde{\mathbf{h}}_{k,t}^{\mathrm{H}}\mathbf{w}_{k,t}}{\mathbf{w}_{k,t}^{\mathrm{H}}\mathbf{Z}_{k,t}\mathbf{w}_{k,t}} \tag{3}$$

where $\mathbf{Z}_{k,t} = \boldsymbol{\vartheta}_{k,t}\boldsymbol{\vartheta}_{k,t}^{\mathrm{H}} + \mathbf{z}_k\mathbf{z}_k^{\mathrm{H}}$ denotes the covariance matrix of the intra- and the inter-cell interference plus noise. For the case of a linear MMSE receiver, $\mathbf{w}_{k,t}$ is calculated as:

$$\mathbf{w}_{k,t} = \mathbf{v}_{t,l}^{\mathrm{H}}\mathbf{H}_{k,l}^{\mathrm{H}}\left(\mathbf{H}_{k,l}\mathbf{v}_{t,l}\mathbf{v}_{t,l}^{H}\mathbf{H}_{k,l}^{H} + \mathbb{E}\left[\mathbf{Z}_{k,t}\right]\right)^{-1} \tag{4}$$

When a given mobile terminal is receiving more than one stream from its serving base station, it calculates a single effective SINR value starting from the SINR values associated to each stream. Specifically, the effective SINR, $\overline{\gamma_k}$, is obtained through an appropriate link-to-system model, that in this work is the Mutual Information Effective SINR Mapping (MIESM):

$$\bar{\gamma}_k = \beta I^{-1}\left(\frac{1}{N_s}\sum_{t\in\mathcal{T}_{l,k}} I\left(\frac{\gamma_{k,t}}{\beta}\right)\right) \tag{5}$$

where $\mathcal{T}_{l,k}$ is the set of streams sent to the k-th mobile terminal, N_s is the number of streams belonging to $\mathcal{T}_{l,k}$, $I(\cdot)$ is the mutual information function, and β is a tuning parameter [51].

2.2.1.1 JSDM

The notation and expressions introduced so far are equally valid for standard MIMO as well as Massive MIMO systems. However, there are some well-known challenges related to Massive MIMO, like the channel estimation and the corresponding CSI feedback for a massive number of antenna elements when working in FDD mode [34]. A first, simple way forward to down-scale the potentially very high number antennas to be considered, is the JSDM approach, recently introduced by [43]. It is a two-stage beamforming/precoding scheme designed to handle inter-group interference based on second-order channel statistics on a long-term time scale and multi-user interference inside of a group based on short-term (instantaneous) CSI. Its block diagram is shown in Figure 2.1.

The first-stage precoding can take the form of a Grid-of-Beams (GoB), where a beamforming network is used to send independent signals in a number of fixed spatial directions, spanning the entire area of a cell sector. This choice comes from the observation that the covariance matrix for a uniform linear array is a Toeplitz matrix [52], which can be asymptotically approximated by a circulant matrix and then diagonalized with a DFT basis [53]. This arrangement provides a significant beamforming gain as well as a great reduction of the effective channel matrix size [54]. This, in turn, reduces the number of pilot sequences required in the downlink, thus limiting the well-known pilot contamination issue.

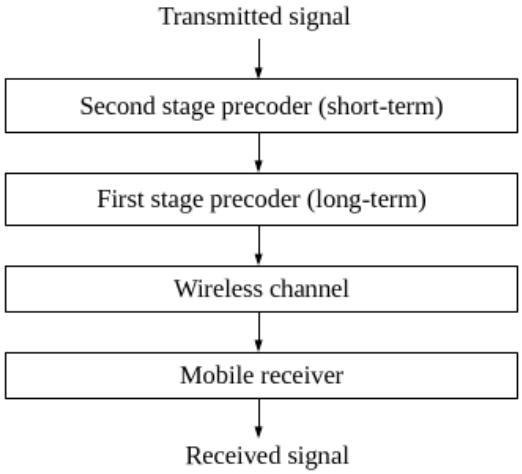


Figure 2.1: JSDM block diagram

Further analysis indicates that the number of relevant channel components, i.e., those beams being received by a mobile terminal within a certain power window (for example 20 dB), will be only a subset of the total set of beams, which is the result of the spatial structuring due to the GoB beamforming. In other terms, the effective channel matrix is sparse (i.e. many elements can be approximated as zero), leading to a number of benefits like the further reduction of CSI reporting overhead or the limited processing power for calculation of the precoder matrices. This concept is exploited in [45], resulting in a reporting scheme which reduces downlink training overhead to about 5%, while also preventing the pilot contamination problem. Such requirement is comparable with current LTE-A training overhead, and thus it is assumed to be the estimation and reporting overhead evaluated in this work.

To use a GoB configuration, the general precoding matrix V from Eq. (1), is split into two matrices by:

$$\mathbf{V}_l = \mathbf{B}_l \mathbf{C}_l \tag{6}$$

with dimensions $\mathbf{B}_l \in \mathbb{C}^{M \times n}$ and $\mathbf{C}_l \in \mathbb{C}^{n \times T_l}$. The matrices $\mathbf{B}_l$ and $\mathbf{C}_l$ denote the GoB beamformer (first stage) and the RZF precoder (second stage), respectively.

$n \in \mathbb{N}$, $T_l < n < M$ is the number of beams produced and is a design parameter. The contents of $\mathbf{B}_l$ depend, among other things, on the structure of the antenna array. For a uniformly spaced linear array, it is possible to write:

$$\mathbf{B}_l = [\mathbf{b}_{l,1}(M) \quad \mathbf{b}_{l,2}(M) \quad \cdots \quad \mathbf{b}_{l,n}(M)] \tag{7}$$

where $\mathbf{b}_{l,i}(k) \in \mathbb{C}^k$ is the precoder to produce the i-th beam using k antennas:

$$\left[\mathbf{b}_{l,i}(k)\right]_h = e^{j(h-1)\left(-\frac{2\pi}{\lambda}d\cos(\theta_i)\right)} \tag{8}$$

where λ is the wavelength, d is the distance between antenna elements and θ_i is the angle between the array broadside direction and the desired horizontal direction of the i-th beam.

For a uniformly spaced $M_v \times M_h$ rectangular array, 3D beamforming becomes possible and the GoB precoding matrix becomes:

$$\mathbf{B}_l = \begin{bmatrix} \mathbf{b}'_{l,1,1}(M_h) & \mathbf{b}'_{l,1,2}(M_h) & \cdots & \mathbf{b}'_{l,1,n}(M_h) \\ \mathbf{b}'_{l,2,1}(M_h) & \mathbf{b}'_{l,2,2}(M_h) & \cdots & \mathbf{b}'_{l,2,n}(M_h) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{b}'_{l,M_v,1}(M_h) & \mathbf{b}'_{l,M_v,2}(M_h) & \cdots & \mathbf{b}'_{l,M_v,n}(M_h) \end{bmatrix} \tag{9}$$

where:

$$\mathbf{b}'_{l,q,i}(k) = \mathbf{b}_{l,i}(k) e^{j(q-i)\left(-\frac{2\pi}{\lambda}d\cos(\phi_i)\right)} \tag{10}$$

and ϕ_i is the angle between the array broadside direction and the desired vertical direction of the i-th beam.

Finally, if the elements of the rectangular array are cross-polarized antenna pairs, then $\mathbf{B}_l$ can be simply rewritten as:

$$\mathbf{B}_l = \begin{bmatrix} \mathbf{B}_{l,1} & 0 \\ 0 & \mathbf{B}_{l,2} \end{bmatrix} \tag{11}$$

where both $\mathbf{B}_{l,1}$ and $\mathbf{B}_{l,2}$ take the same form as $\mathbf{B}_l$ in Eq. (9).

Here the focus is on carrier frequencies below 6 GHz, as they provide good overall coverage and all existing mobile networks already work at these frequencies. Note that, for below 6 GHz urban macro scenarios with multiple reflections and

diffractions, serving ten or more users simultaneously lead to severe inter-user interference, so that simple beam selection would not be a good choice. Instead, typical precoders rely on Zero Forcing (ZF) or, in case of mobile terminals with limited SINR, on Regularized Zero Forcing (RZF) precoders [44]. In this last case, the second-stage precoder $\mathbf{C}_l$ is calculated as:

$$\mathbf{C}_l = \mathbf{D}_l \mathbf{E}_l^{-\frac{1}{2}} \tag{12}$$

$$\mathbf{D}_l = [\mathbf{d}_{l,1} \quad \mathbf{d}_{l,2} \quad \dots \quad \mathbf{d}_{l,T_l}] = \mathbf{H}_l^{\mathrm{H}} \mathbf{W}_l^{\mathrm{H}} \left(\mathbf{W}_l \mathbf{H}_l \mathbf{H}_l^{\mathrm{H}} \mathbf{W}_l^{\mathrm{H}} + \sigma^2 \mathbf{I}_{T_l} \right)^{-1} \tag{13}$$

where $\mathbf{H}_l$ is the compound channel matrix obtained by concatenating the channel matrices $\mathbf{H}_{k,l}$ of all the users being served, and likewise the receiver matrix $\mathbf{W}_l$ is the concatenation of the (estimated) receive filters used at all the served receivers. $\mathbf{E}_l$ is a diagonal matrix used to enforce the total power constraint and improve the sum throughput, by assigning lower power to high-SINR users and high power to low-SINR users. It is calculated with the vector normalization method [55], which achieves a good balance between sum throughput and fairness among users:

$$[\mathbf{E}_l]_{i,i} = d_{l,i}^H d_{l,i}$$

2.2.1.2 Beam coordination

Motivated by the growing traffic demands, the densification of base stations in certain areas is expected to grow [56]. Such scenarios are often referred to as ultra-dense networks in literature [57] [14]. In this context, the inter-cell interference management is required also with Massive MIMO. The idea is to use the first-stage GoB beamformer introduced in Section 2.2.1.1 for beam-coordination between base stations.

The JSDM was originally designed for a single Massive MIMO base station. However, the concept is extended to coordinate interference between multiple Massive MIMO base stations [58]. Preliminary results already presented in [46]

promise significant performance gains. Still, the scenarios therein are simplified to only 2 and 3 base stations, which do not expose possible conflicts that arise in larger setups. This work extends that to a complete multi-site scenario.

In [46], the first stage precoders are selected out of DFT matrix which enables easy implementation into existing system level environments. This concept combines the advantage of GoB to reduce dimensionality of the Massive MIMO channel and, at the same time, to provide spatial interference mitigation among interfering base stations.

In this work, an approach derived from [42] is adopted to reduce the average interference level. Three different GoB beamformers are defined which only cover a subset of the sector, instead of the entire sector as described in Section 2.2.1.1. Specifically, the sector's azimuth span of 120 degrees is split into three equal parts of 40 degrees each, with the beams of the first-stage precoder confined to one of the three sub-sectors. The three beamformers are changed frequently with a fixed or adaptive pattern. With traditionally sectorized base stations, this result could only be obtained by physically replacing the antenna arrays. On the contrary, with Massive MIMO this only requires a re-configuration of the first-stage precoder by the scheduler.

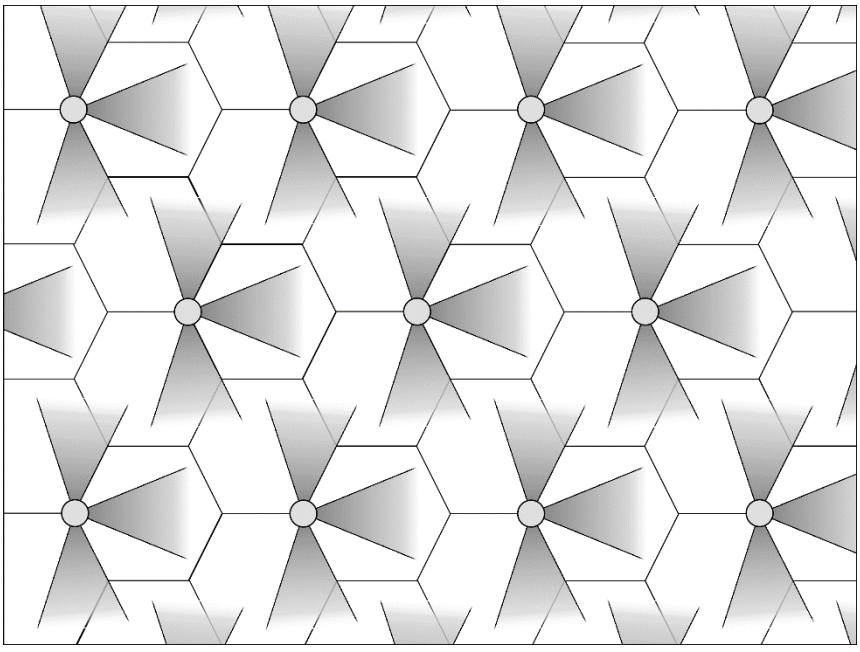


Figure 2.2: Example of coverage profile for beam coordination.

When considering the three sectors of one site, each sector uses a different beamformer and they are arranged so that in no case adjacent sub-sectors are served. Moreover, the same combination applied in one site is repeated verbatim on all the other sites. Accordingly, the use of a specific beamformer over multiple cells defines a coverage profile. One example profile is shown in Figure 2.2, the others are similar but rotated by ±120°. This configuration has a number of advantages:

- in the long term, all the sub-sectors can be served equally well;
- when looking at the complete multi-cell pattern, served sectors are never facing each other directly at a close distance, thus reducing the interference $\mathbf{z}_k$ from Equation (2) and raising the SINR;
- if the sequence of the beamformers to be used is known in advance, a User Equipment can go in a power saving state when his sub-sector is not being served;
- for inter-cell coordination, the only requirement is that the cells are time-synchronized and that they exchange the index of the beamformers configuration, which can only have three possible values;
- in sub-sectors which are not being served, pico/femto-cells (if present) can work with reduced interference levels. The short-term feedback for the second stage precoding is designed locally at each base station independently, based on the short-term CSI, e.g., at base station l from the users connected to it.

2.2.2 Performance evaluation methodology

The technologies presented in Section 2.2.1 were implemented in the open-source simulator LTE-Sim [59] and evaluated via computer simulations. This section describes the implementation details and parameters for each of the considered technologies.

2.2.2.1 Common parameters

The evaluation considers two different environments: the first one is the "Wide area coverage" scenario described in Section 1.3.2, and the other one is the "Dense urban areas" scenario. Both include a standard layout of 19 sites (57 cells) placed on a hexagonal grid. However, results are collected only for the center cell to obtain a realistic interference profile. The ISD is set to 200, 600, or 1000 m.

Mobile terminals are randomly distributed over the entire simulation area, with a uniform probability distribution. During the whole simulations, they move at a velocity of 3 km/h. According to the requirements of the considered use cases, the user density is set to 2500 users/km^2, 400 users/km^2, and 100 users/km^2 for the urban, suburban, and rural scenarios, respectively.

The total bandwidth is set to 100 MHz. This is achieved by using carrier aggregation on 5 component carriers of 20 MHz each. The center frequency is set to 2000 MHz for the urban/suburban case and 800 MHz for the rural case.

Regarding the channel model, the 3GPP 3D model, described in [60], is adopted for both large-scale path loss and small-scale fading. Shadow fading is modeled as a log-normal random variable with a standard deviation of 10 dB. The overall network embraces 80% of indoor users. They experience a penetration loss equal to 20 dB. The remaining 20% of users are outdoor and they do not experience any additional loss due to the penetration phenomenon.

The transmission power of the base station is set to 46 dBm for each 20 MHz carrier. The Modulation and Coding Scheme (MCS) and Transport Block Size (TBS) are selected according to procedures standardized in the LTE-A specifications, based on the predicted SINR at the receiver [61]. Specifically, the final user throughput is directly proportional to the selected TBS.

Mobile terminals send their feedback information (which can vary according to the adopted technology) with a periodicity of 5 ms. When receiving data at the physical layer, the SINRs of each sub-channel are calculated according to Eq. (3), which are then mapped to an effective SINR using the MIESM formula of Eq. (5). Finally, the probability of receiving a corrupted block, i.e. the Block Error Rate (BLER), is calculated according to pre-calculated SINR-BLER curves in an

equivalent AWGN channel.

To look for the upper bound on the throughput performance, an infinite-buffer traffic model is used at the application layer. Each simulation is repeated 30 times with a different initial seed, which affects random variables such as user position and fast fading realizations.

All these parameters are summarized in Table 2.2.

Parameter	**Value**
Cellular deployment	Hexagonal grid, 19 sites
ISD	200, 600, 1000 m
User density	100, 400, 2500 users/km^2
User distribution	Uniform over simulation area
User speed	3 km/h
Carrier frequency	800 MHz or 2 GHz
Operative bandwidth	100 MHz (5 x 20 MHz)
Channel model	3GPP 3D channel model
Link-to-system mapping	MIESM
Shadowing std. dev.	10 dB
Indoor penetration loss	20 dB
Indoor/outdoor ratio	80%
Transmission power	46 dBm for each 20 MHz carrier
CQI reporting period	5 ms
Traffic model	Infinite buffer
Duration	10 s
Number of realizations	30

Table 2.2: simulation parameters

2.2.2.2 LTE-A baseline

The baseline LTE-A technology, considered as a term of comparison, uses a physical layer based on Transmission Mode 9 and a MIMO transmission scheme with an 8x8 antenna configuration. Specifically, the antenna array at the base station is a Uniform Linear Array (ULA) with vertical polarization and $\lambda/2$ element spacing, while the mobile terminal has a cross-polarized array with $\lambda/2$ spacing as well.

The Single-User MIMO (SU-MIMO) mode is used, i.e. only one user is scheduled in a cell at any given time, with a round-robin criterion. Also, pre-coding is done according to the dual-index codebook defined in the LTE-A standard for operation with 8 transmission antennas.

At the receiving side, the mobile terminal calculates the SINR using a linear MMSE receiver and send back feedback information: the Channel Quality Indicator (CQI) for link adaptation, and Precoding Matrix Indicator (PMI) plus Rank Indicator (RI) for hints on the preferred MIMO processing at the base station [62]. PMI and RI are calculated via an extensive search over all the possible combinations, and the one which maximizes the predicted throughput is selected. CQI is finally calculated with respect to the resulting expected SINR with the chosen PMI and RI.

2.2.2.3 Massive MIMO with JSDM

The antenna array used for Massive MIMO contains 16 elements in the horizontal direction and 8 elements in vertical direction, with $\lambda/2$ spacing in both directions. Each element is actually a dual-polarized antenna, with $\pm 45^{\circ}$ orientations.

The GoB precoding creates 16 beams in the horizontal direction, with equally spaced boresight directions spanning the entire area of the cell (i.e. a 120° sector). The vertical dimension is exploited by giving different tilting angles to even-numbered and odd-numbered beams. Each beam is created by both left and right slanted antennas. Thus, there are 32 beams in total.

On top of GoB, the RZF precoding is applied to map data streams to beams with appropriate weights and phases. In this case a MU-MIMO configuration is employed, where up to 8 devices can be scheduled on the same physical resources. This amounts to a total of 16 data streams, against the maximum theoretical 32. Because of how the RZF precoder works, using more than half of the possible streams reduces the available degrees of freedom and degrades the performance rapidly. The users to be served are selected with a multi-user version of the round-robin scheduler. Specifically, the scheduler selects a new set of users for each Transmission Time Interval (TTI), shifting over the list of registered users. When it

reaches the end of the list, it just resumes from the beginning.

After the RZF, vector normalization as described in [55] is performed on the obtained precoding vectors, to partially compensate for the different path losses and provide a good compromise between total throughput and fairness.

Mobile terminals are equipped with two receive antennas in a cross-polarized configuration, and they are configured to receive two layers of spatially-multiplexed data. They use a linear MMSE receiver to evaluate the received SINR, which is then used for BLER estimation and CQI feedback. In this work, CQIs are calculated by using the same SINR-CQI mapping as in LTE-A.

Regarding the CSI reporting scheme, the proposed study initially assumes a perfectly knowledge of the CSI at the base station side. In practice, in fact, a high accuracy can be achieved with reasonable uplink data rates, by reporting only relevant beams/taps as shown in [45] or [42]. Then, to provide a further insight, the conducted study also evaluated the impact that a reduction of CSI feedbacks has on the overall system performance.

2.2.2.4 Beam coordination

For the coordinated beamforming scheme, three different profiles are used. They are obtained by dividing the area of a sector into three sub-sectors in the angular domain, as already described in Section 2.2.1.2.

For the scheduling, it is important to only consider users which are in sub-sectors with good coverage. Therefore, users have to observe the signal quality over all the coverage profiles and indicate the preferred one to the base station.

All the other parameters presented for Massive MIMO also apply here.

2.2.3 Simulation results

This section presents the results of the numerical evaluation and related comments. Specifically, Section 2.2.3.1 describes results related to the channel

quality, Section 2.2.3.2 deals with some Key Performance Indicators (KPIs), and Section 2.2.3.3 investigates CSI acquisition.

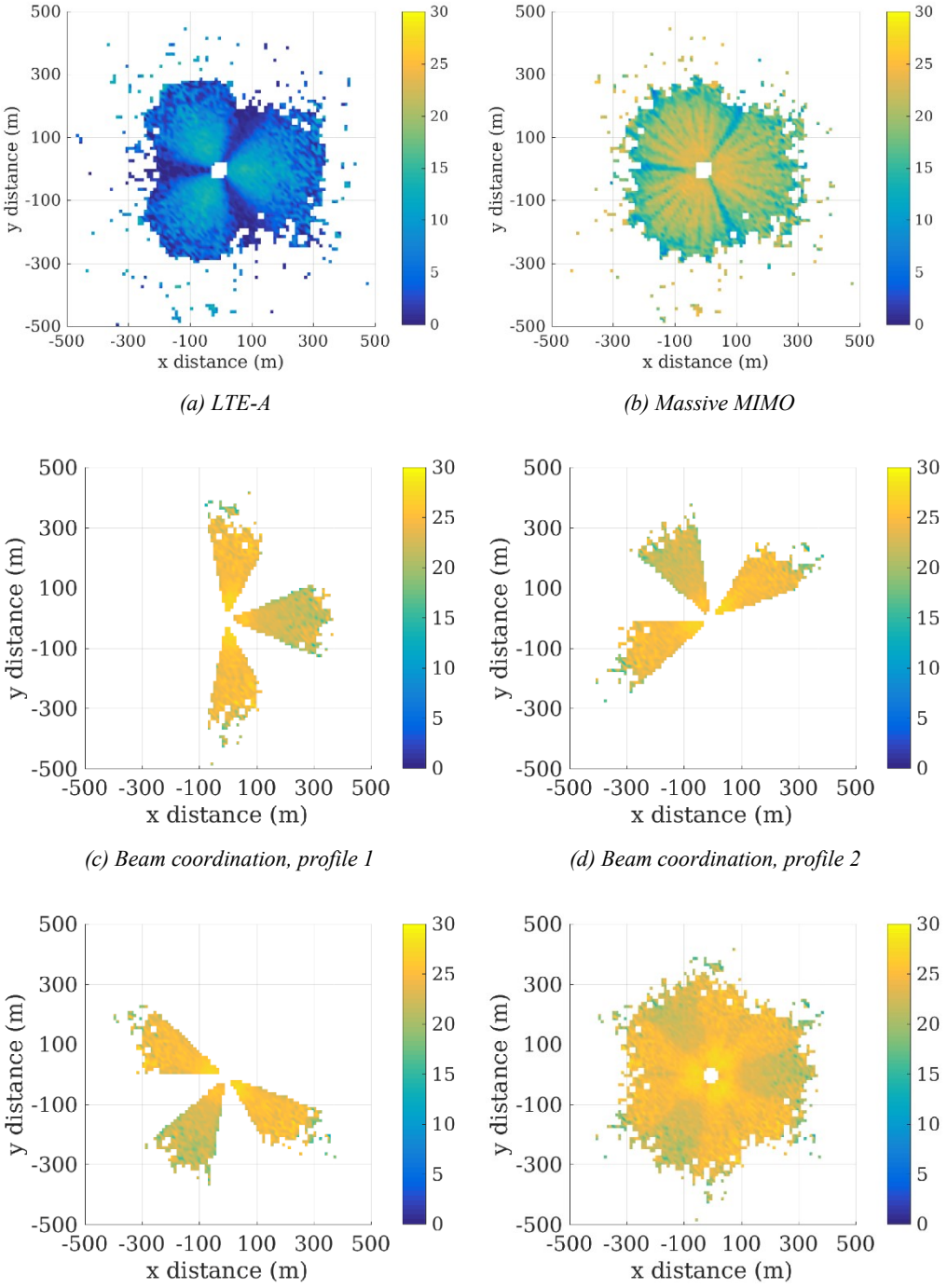
2.2.3.1 Impact on the channel quality

The performance of the proposed technologies is firstly evaluated in terms of channel quality experienced by mobile terminals. Such a performance index is expressed by means of the post-detection SINR, calculated accordingly to Eq. (3). In order to have a first look of the SINR within a cell, Figure 2.3 reports the SINR map measured in a scenario with ISD=500 m and user density equal to 100 users/km^2. Note that even if a multi-cell environment is simulated, reported results only show what is registered in the central cell.

Figure 2.3-a) shows the spatial SINR distribution obtained with the baseline LTE-A technology. Here, the main lobes of the three sectors are clearly visible, as well as the large gaps at the sector boundaries with low SINR ($\gamma < 5$ dB). The SINR ranges from 10-15 dB in the main lobes, down to 0 at the cell edges.

Figure 2.3-b) represents the same information for Massive MIMO with GoB. In this case, the SINR is much higher due to the high directivity gain, while the weak inter-sector areas are drastically reduced. This clearly reflects a large improvement of the perceived channel quality, which in turn permits greater throughputs.

Figure 2.3-c), Figure 2.3-d), and Figure 2.3-e) show the SINR distribution for the three coverage profiles of the coordinated beamforming, while Figure 2.3-f) picks the maximum of the three values for each point. This configuration provides a further increment of the SINR, due to the reduction of the interfering power, with a more uniform coverage as well.



(a) LTE-A

(b) Massive MIMO

(c) Beam coordination, profile 1

(d) Beam coordination, profile 2

(e) Beam coordination, profile 2

(f) Beam coordination, maximum values for all profiles

Figure 2.3: Example of the SINR map for LTE-A, Massive MIMO, and beam coordination Obtained in a scenario with an ISD of 500 m and user density equal to 100 users/km².

A more comprehensive analysis is reported in Figure 2.4, Figure 2.5, and Figure 2.6, which show the Cumulative Density Function (CDF) of the post-detection SINR for all the combinations of ISD and user density. In accordance with previous comments, LTE-A is always surpassed by Massive MIMO, which in turn is surpassed by Massive MIMO with coordinated beamforming. Note that the LTE-A curves are not smooth as an effect of the dynamic rank adaptation.

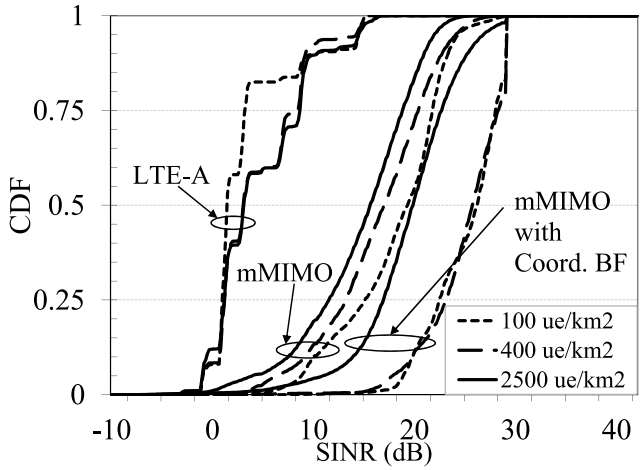


Figure 2.4: Cumulative distribution function of SINR (ISD = 200 m)

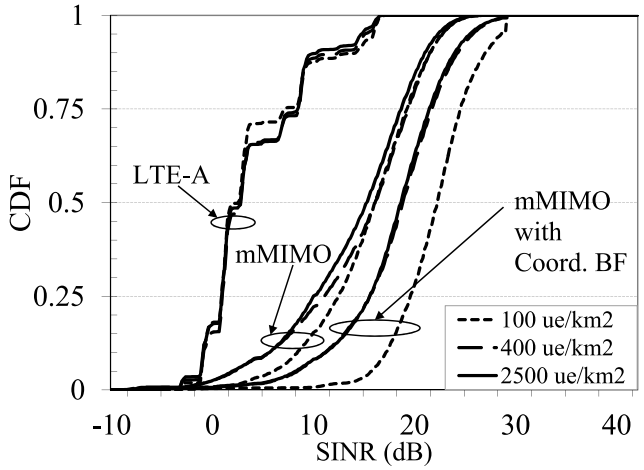


Figure 2.5: Cumulative distribution function of SINR (ISD = 600 m)

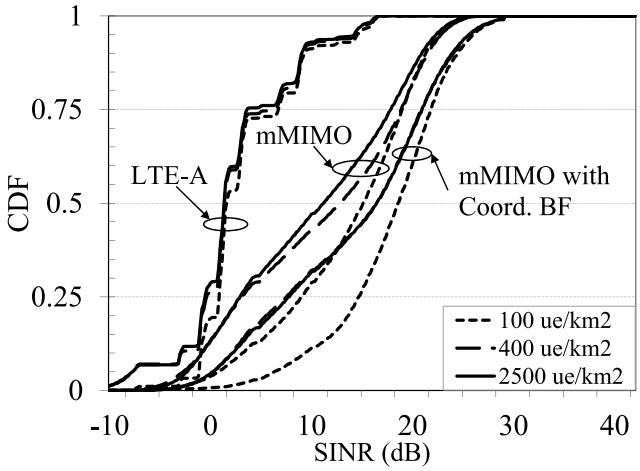


Figure 2.6: Cumulative distribution function of SINR (ISD = 1000 m)

For all the technologies, it is possible to observe that an increment in the ISD produces a shift towards lower SINRs. This effect is expected, as it is due the increment of the average path loss. On the contrary, the user density has a different impact on LTE-A compared to the proposed 5G solutions. With Massive MIMO and Massive MIMO with coordinated beamforming, for instance, a higher user density results in a higher average number of co-scheduled users. Thus, the power transmitted to each user is lower and the SINR is reduced as well. On the other hand, LTE-A has very little variation on the user density. The only exception is the curve for ISD of 200 m and density of 100 users/km^2. This is probably due to statistical fluctuations, because it is a scenario with a very low number of users per sector.

2.2.3.2 Investigation of KPIs

Figure 2.7, Figure 2.8, and Figure 2.9 report the results for the user experienced data rate, defined as the average data rate achieved by each user over an entire simulation run. It is reported for all the evaluated technologies, all ISDs, and all density values. The target value is assumed to be 50 Mbps per user. The LTE-A technology is able to meet the required value in only three instances, that is, when

the number of users is relatively low compared to the number of base stations: with an ISD of 200 m and 100/400 users/km^2, or with an ISD of 600 m and 100 users/km^2. These can be considered as oversized cases, therefore they are of limited practical interest. In the other (more challenging) scenarios, instead, LTE-A is not able to guarantee a sufficient data rate. On the other hand, Massive MIMO and Massive MIMO with beam coordination can reach nearly one order of magnitude of gain. Specifically, they provide more than 100 Mbps (i.e. twice the required value) also in these three scenarios, which have a higher ratio between users and base stations: 1000 m ISD + 100 users/km^2, 600 m ISD + 400 users/km^2, and 200 m ISD + 2500 users/km^2. These deployments can be called "balanced scenarios", as the number of users per cell is in a reasonable range (i.e., from 30 to 50). Finally, the remaining three combinations (1000 m ISD + 400 users/km^2, 600 m ISD + 2500 users/km^2, and 1000 m ISD + 2500 users/km^2) are the undersized cases, because there are so many mobile terminals in each cell that they cannot reach the required data rate.

It is interesting to note that, in the oversized cases only, beam coordination performs worse than Massive MIMO, despite the average SINR being higher. This happens because for this technology the number of users available for scheduling is further reduced by the sub-sectorization of the cells, and the full capacity cannot be exploited. In all the other cases, coordinated beamforming provides a gain of around 20% with respect to Massive MIMO.

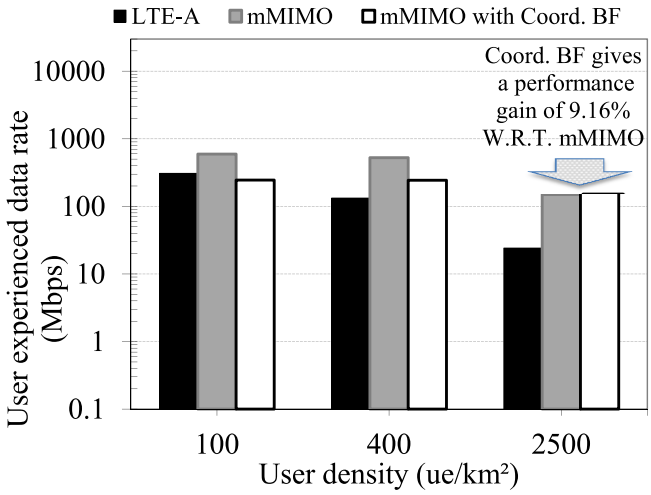


Figure 2.7: User experienced data rate (ISD = 200 m)

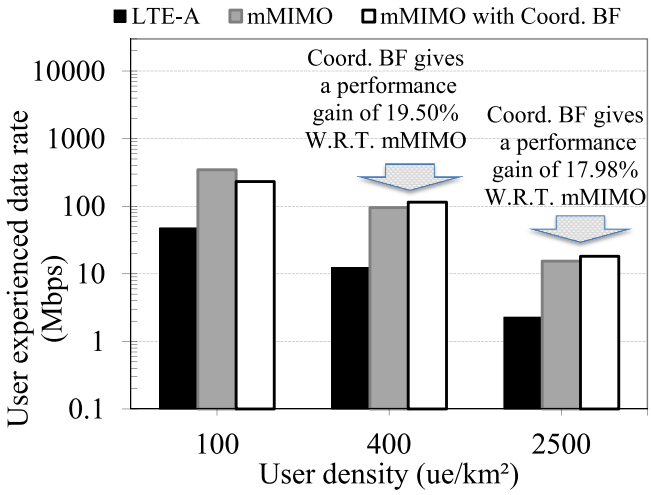


Figure 2.8: User experienced data rate (ISD = 600 m)

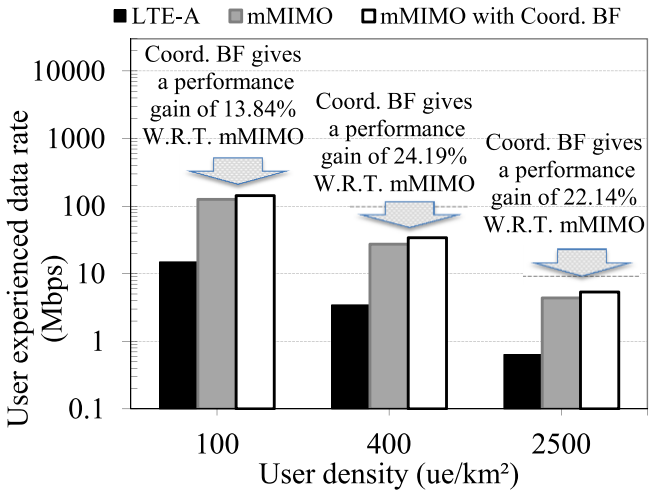


Figure 2.9: User experienced data rate (ISD = 200 m)

To provide a further insight, Figure 2.10, Figure 2.11, and Figure 2.12 show the CDF of the user experienced data rate for the three "balanced scenarios". They confirm that most of the transmissions performed with the newer schemes (i.e., Massive MIMO and Massive MIMO with beam coordination) can achieve a throughput much higher than the target value. Moreover, they show that the gain provided by the beam coordination scheme is quite consistent across the entire

curve, and it is effective at improving the 95 %-ile of the throughput.

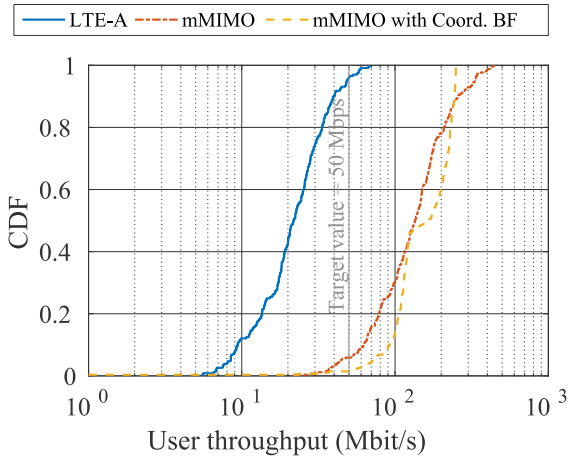


Figure 2.10: Cumulative distribution function of user throughput (ISD = 200 m)

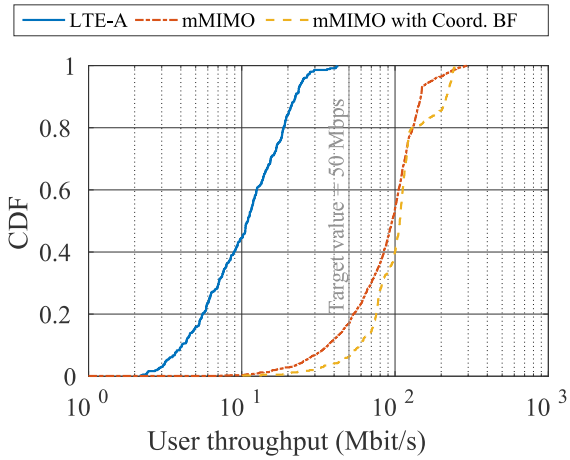


Figure 2.11: Cumulative distribution function of user throughput (ISD = 200 m)

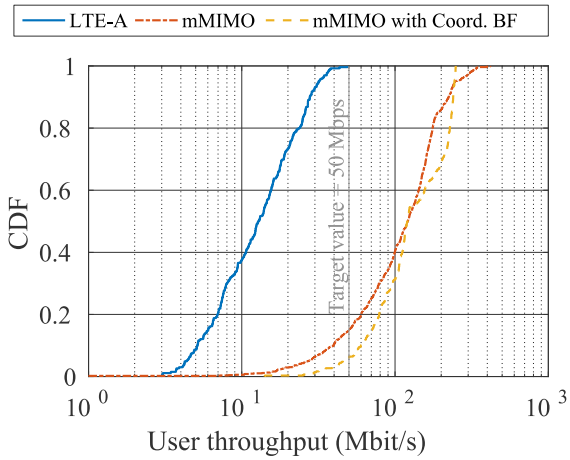


Figure 2.12: Cumulative distribution function of user throughput (ISD = 200 m)

Table 2.3 shows the measured traffic density, defined as the ratio of aggregate

traffic in a given area (e.g., a cell) over the size of such area. In this case the target values are different for each scenario: 5, 20, and 125 Gbps/km^2 for user density of 100, 400, and 2500 users/km^2, respectively. With respect to these target values, the behavior of the traffic density is very similar to the user experienced data rate: LTE-A meets the requirements only in the oversized cases, while Massive MIMO and beam coordination provide good results also in the balanced cases (but not in the undersized cases). Again, coordinated beamforming performs worse in the oversized cases, but provides around 20% gain in all the other situations.

			Inter-site distance (km)		
			0.2	0.6	1
User density (Users/km^2)	100	LTE-A	23.9	4.09	1.38
		Massive MIMO	65.0	36.2	12.4
		Beam coordination	21.2	20.8	14.7
	400	LTE-A	39.0	4.24	1.18
		Massive MIMO	205	37.0	10.8
		Beam coordination	78.1	44.7	13.0
	2500	LTE-A	48.4	4.95	1.42
		Massive MIMO	330	35.2	10.1
		Beam coordination	393	44.9	13.1

Table 2.3: Traffic density [Gbps/km^2]

It is interesting to note that the increment of the user density results in a higher area throughput, but only when oversized and balanced scenarios are considered. On the contrary, when going from balanced to undersized scenarios, the traffic density is fairly stable, meaning that the full potential of the network is already being used.

This is also confirmed by Table 2.4, which reports the total throughput of one cell, but without normalization with respect to the area. In this case it is even more apparent that balanced/undersized scenarios all reach similar results, corresponding to the practical limit of the system.

			Inter-site distance (km)		
			0.2	0.6	1
User density (Users/km^2)	100	LTE-A	0.276	0.425	0.397
		Massive MIMO	0.750	3.76	3.59
		Beam coordination	0.245	2.16	4.24
	400	LTE-A	0.450	0.441	0.341
		Massive MIMO	2.36	3.85	3.11
		Beam coordination	0.902	4.65	3.76
	2500	LTE-A	0.560	0.515	0.409
		Massive MIMO	3.81	3.66	2.90
		Beam coordination	4.54	4.66	3.78

Table 2.4: Cell throughput [Gbps]

2.2.3.3 Investigation on the CSI acquisition

As already mentioned, the two-stage precoding structure natively reduces the CSI overhead, as the number of pilot signals to transmit and feed back is equal to the number of beams, rather than the number of physical antennas. Moreover, additional savings can be obtained by only reporting the most relevant beams and discarding those with very low received power. This approach is usually implemented by comparing the quality of beams against the strongest one through a given threshold. For instance, adopting a threshold of 25 dB means that any beam that is more than 25 dB weaker than the stronger beam is completely neglected. Of course, this would cause an additional inter-user interference, because the effect of such beams is not considered anymore by the RZF precoder.

In this section, the effect of different threshold values for CSI reporting on the user throughput is investigated. The results are shown in Figure 2.13, for the three balanced scenarios identified in Section 2.2.3.2 and for the JSDM + beam coordination scheme only. A higher threshold value means that more beams are included in the CSI report and the precoding is more precise, while lower values reduce the set of reported beams but also incur more interference and lower SINRs. The results are expressed as a percentage of the throughput of the ideal case, i.e. the throughput is 100% when all the beams are reported, and decreases as more beams

are excluded. The overall effect is similar for the three scenarios: threshold values of 35 dB or higher give almost as much throughput as the ideal case, and with 30 dB more than 90% of the throughput is still maintained. Instead, with lower values of 25 dB and 20 dB, the throughput degrades more rapidly, thus they should probably be avoided, although the actual throughput is still within the target requirements. In general, 30 dB appears to be the most reasonable compromise between CSI precision and feedback reduction.

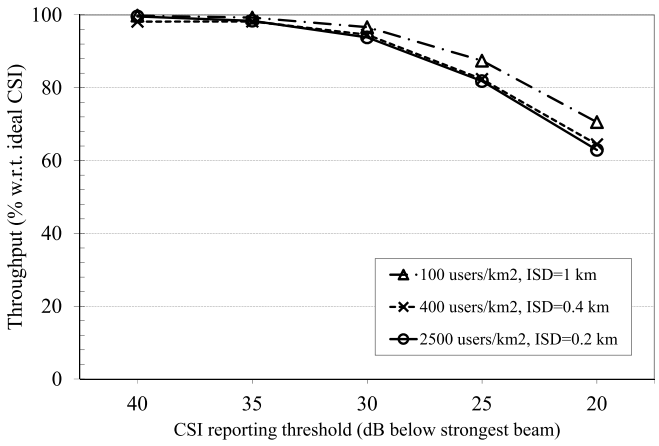


Figure 2.13: Effect of CSI acquisition threshold

2.3 Density-based user clustering for JSDM

In Section 2.2, the JSDM precoding technique was introduced and described in its overall structure. Instead, this section explores a specific topic related to the JSDM, namely the generation of user clusters for the first precoding stage.

As already stated, the first-stage precoder is computed based on long-term statistical CSI, and it is used for all the users inside a group. For a perfect matching, the users in one group should have the exact same covariance matrix, but this is not possible in practice. Instead, it is possible to find groups of users that have similar covariance matrices, so that the resulting inter-group interference is small enough

to have a limited impact. Thus, the problem at hand consists in constructing such groups from a larger set of users which may have very different channel covariances.

Formally, this is called a clustering problem, and it is typically encountered in data science. For the specific case in JSDM, a few approaches have been considered in the literature already. In [54], the problem is approached with k-means partitioning, which iteratively assigns each user to the closest cluster, or with fixed partitioning of the angular domain, with the latter being simpler and providing better results. Instead, [63] proposes an agglomerative clustering method as an alternative to k-means. While it is faster for low numbers of users, it becomes comparable or even slower when hundreds of users are considered. The work in [64] compares k-means, k-means++, and the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm for the same purpose, in a scenario where the clusters already exist and need to be detected. The results show that k-means and k-means++ are not well-suited, mainly because they forcefully assign all the users to some cluster, even isolated ones. Instead, the DBSCAN approach came out as the most promising, as it can exclude outliers and parameter selection is relatively easy. These findings are confirmed in [65], where DBSCAN is similarly compared to k-means and k-medoids methods.

However, DBSCAN presents important limitations in many scenarios. For example, when there is a large compact crowd of users with a mostly constant density, it would be detected as a single cluster. This poses serious problems for JSDM. From one side, it may not be possible to serve multiple groups, thus limiting the total throughput. From the other side, users in a large group may have very different covariance matrices, which results in intra-group interference at the first-stage precoder. For these reasons, after describing the basics of the original DBSCAN algorithm, Section 2.3.1 proposes an extension of the baseline DBSCAN approach with a configurable limit on the cluster size. This way, large clusters are split into smaller parts, which are more suitable for JSDM precoding. The resulting algorithm retains the desirable properties of DBSCAN identified in prior work, but it can be applied to a much broader set of scenarios. This is verified by numerical simulations in Section 2.3.2, which show an increase of both user and system

spectral efficiencies, and also highlight that the algorithm parameters do not need a very precise tuning.

2.3.1 Original and improved DBSCAN for JSDM

In the context of Massive MIMO and JSDM, some clustering approaches have been investigated in [64], where the density-based approach DBSCAN is shown by simulations to be better than the parametric-based methods, i.e. k-means and k-means++. Specifically, it was assumed that the users were actually located in well-separated clusters, except for a few outliers, and the clustering step had to detect such clusters and outliers. In this work, a different user distribution is targeted where DBSCAN performs poorly, and an improved version that can cope with the new situation is proposed. After introducing the system model in Section 2.3.1.1, the original DBSCAN algorithm is described in Section 2.3.1.2, while the new challenging scenario and the improved DBSCAN algorithm are presented in Section 2.3.1.3.

2.3.1.1 System model

Throughout this section, scalar quantities are denoted with italic letters (a, A), while vectors are written with lowercase bold letters (**a**) and matrices are assigned uppercase bold letters (**A**). Sets are represented by calligraphic letters ($\mathcal{A}$), and their elements are listed in square brackets ($[a_1, a_2, \dots, a_n]$). $(\cdot)^{\mathrm{H}}$ represents the Hermitian matrix, while $\mathbb{E}[\cdot]$ is the expectation operator, and $\|\cdot\|_F$ is the Frobenius norm of a matrix. In the considered scenario, the base station is equipped with N antennas and a MIMO downlink block-fading channel is taken into account. At the end user side, there are K single antenna receivers. Thus, the receive signal $\mathbf{y} \in \mathbb{C}^K$ is written as:

$$\mathbf{y} = \mathbf{H}^{\mathrm{H}}\mathbf{V}\mathbf{x} + \mathbf{n} \tag{14}$$

where $\mathbf{H} \in \mathbb{C}^{N\times K}$ is the radio channel between the N antennas of the base station and the K antennas of the end users, $\mathbf{V} \in \mathbb{C}^{N\times T}$ is the precoding matrix for T independent data streams, $\mathbf{x} \in \mathbb{C}^{T}$ is the data vector, and $\mathbf{n} \in \mathbb{C}^{K}$ denotes the additive white Gaussian noise assuming $\mathbf{n} \propto \mathcal{CN}\ (1, \mathbf{I}_K\ )$. Since linear precoding is used, the number of data streams is bounded by $T \leq \min(N, K)$ [4]. By following the assumptions in [5], the transmit power is equally split to all data streams, such that $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{H}}] = \frac{P}{T}\,\mathbf{I}_T$, where P is the total transmit power. The radio channel of the receiver $k \in \mathcal{K} = [1, \dots, K]$ follows $\mathbf{h}_k \propto \mathcal{CN}\ (0, \mathbf{R}_k\ )$, where the covariance matrix $\mathbf{R}_k$ is positive semi-definite. $\mathbf{R}_k$ is a second-order statistic of the radio channel with a longer coherence time than direct CSI of $\mathbf{h}_k$. The singular value decomposition of $\mathbf{R}_k$ is given as:

$$\mathbf{R}_k = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{U}_k^{\mathrm{H}} \tag{15}$$

where $\mathbf{U}_k \in \mathbb{C}^{N\times b_k}$ contains the b_k eigenvectors that correspond to the non-zero b_k eigenvalues in the diagonal matrix $\mathbf{\Lambda}_k$ of size $b_k \times b_k$. A subset of $\mathbf{U}_k$ is used later to obtain the input metric for receiver clustering.

Following the two stage precoding in [43], the precoder $\mathbf{V} = \mathbf{B}\mathbf{P}$ is split into a first and second stage precoder. According to a suitable clustering approach, the K receivers are divided into G clusters of size K_g, such that $K\ = \sum_{g=1}^{G} K_g$, where subscript $g \in [1, \dots, G]$ is the cluster index. In this work, the terms "group" and "cluster" are used interchangeably to denote the disjoint subsets of users produced by a clustering algorithm. Following the guidelines in [43], receivers within the same cluster should have similar covariance matrix eigenvalues whereas the eigenvalues of different clusters should be orthogonal. The first stage precoder $\mathbf{B} \in \mathbb{C}^{N\times b}$ is a function of second order statistics to coordinate inter-cluster interference. The second stage precoder $\mathbf{P} \in \mathbb{C}^{b\times T}$ is a function of direct CSI of the effective channel $\widetilde{\mathbf{H}} = \mathbf{B}^{\mathrm{H}}\mathbf{H}$. To take into account the G receiver clusters, $\mathbf{H}_g = \left[\mathbf{h}_{g1}, \dots, \mathbf{h}_{gK_g}\right]$ denotes the channel matrix for one group as the concatenation of the individual users' channels, and the first-stage precoder becomes $\mathbf{B} = [\mathbf{B}_1, \dots, \mathbf{B}_G\ ]$, where $\mathbf{B}_i$ is the i-th group's precoding matrix. Also, P can be written as $\mathbf{P} =$

$\mathrm{diag}(\mathbf{P}_1, \ldots, \mathbf{P}_G)$, where $\mathbf{P}_i$ is the i-th group's diagonal power allocation matrix, and $\mathbf{x}_g = \left[x_{g1}, \ldots, x_{gK_g}\right]$ is the matrix of all the user signals in group g.

With these definitions, the receive signal reported in Eq. (14) can be rewritten for the user k in cluster g as:

$$y_{gk} = \underbrace{\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{B}_g\mathbf{p}_k x_k}_{\hat{h}_{gk}} + \underbrace{\sum_{k' \in \mathcal{K}_g,\, k' \neq k} \mathbf{h}_{gk'}^{\mathrm{H}}\mathbf{B}_g\mathbf{p}_{k'}x_{k'}}_{z_{gk}} + \underbrace{\sum_{g' \neq g} \mathbf{H}_{g'}^{\mathrm{H}}\mathbf{B}_{g'}\mathbf{P}_{g'}\mathbf{x}_{g'}}_{\hat{z}_{gk}} + n_k \tag{16}$$

where subscript gk denotes receiver k in cluster g, $\mathcal{K}_g = [1, \ldots, K_g]$ the set of receivers in cluster g and $\hat{h}_{gk}$, z_{gk}, $\hat{z}_{gk}$ comprise the signal, intra-cluster and inter-cluster interference, respectively.

Thus, the resulting SINR, γ_{gk}, is given by:

$$\gamma_{gk} = \frac{\hat{h}_{gk}^{\mathrm{H}}\hat{h}_{gk}}{z_{gk}^{\mathrm{H}}z_{gk} + \hat{z}_{gk}^{\mathrm{H}}\hat{z}_{gk} + \mathbb{E}[n_k^{\mathrm{H}}n_k]} \tag{17}$$

and the corresponding capacity as normalized with the bandwidth assigned to receiver k in cluster g is obtained by:

$$C_{gk} = \log_2(1 + \gamma_{gk}) \tag{18}$$

2.3.1.2 DBSCAN

In the DBSCAN clustering algorithm [66], a cluster is intuitively defined as a region of the feature space where there is a higher density of data points, compared to regions outside of the cluster. A more formal definition starts with the notion of *Eps-neighborhood*: given a distance function $\mathrm{dist}(p, q)$ and a distance Eps, the Eps-neighborhood $N_{Eps}(p)$ of a data point p is defined as the set of points no farther than Eps from p:

$$N_{Eps}(p) = \{q | \mathrm{dist}(p, q) < Eps\}$$

(19)

A simple approach could define a cluster as a set of points whose Eps-neighborhood contain at least a certain number $minPts$ of data points. However, this would be sub-optimal, because points on the edge of the cluster (border points) typically have a lower number of points in their Eps-neighborhood, compared to internal points (core points), even if the density of the cluster is constant. Using a lower value for $minPts$ would underestimate the density of the cluster, which can be problematic in the presence of noise. A better solution is to require core points to have at least $minPts$ points in their Eps-neighborhood, and border points to be *directly density-reachable* from one or more core points. Given a core point q, a point p is said to be directly density-reachable from q if it is contained in the Eps-neighborhood of q.

The directly density-reachability property can only connect points in a cluster that are close to each other. The definition of an entire cluster requires two more concepts: *density-reachable* point and *density-connected* point. A point p is density-reachable from a point q if there exist a sequence of points $p_1 = q, p_2, \dots, p_{n-1}, p_n = p$ which connects q to p, so that p_{i+1} is directly density-reachable from p_i. This allows going from any core point to any other point of the cluster, but not from border points to other points. The notion of density-connected points fills this last gap: two points p and q are said to be density-connected if they are both density-reachable from some point o. Now, a cluster can be unambiguously identified by any one of its points, either a core point or a border point, together with all of the points that are density-connected to it.

DBSCAN relies on such definition of cluster. Ideally, given the right values of both Eps and $minPts$, which may be different for each cluster, and a starting point for each cluster, it is straightforward to assign each other point to the right cluster. However, this would require some previous knowledge of the data that is usually not available. Instead, DBSCAN uses a single value for Eps and for $minPts$, which are assumed to be valid for the thinnest cluster and thus would also work for more dense clusters.

The authors of [66] describe a procedure to determine such parameters with minimal user interaction. The algorithm starts from the first point in the database,

finds all the points that are density-connected to it, and marks them as belonging to the same cluster. Then it picks the next unclassified point and repeats the procedure. This is repeated until all points are either assigned to a cluster or marked as noise points, i.e. points that are not density-reachable from any other point.

The detection of noise points is particularly important in the context of JSDM: in fact, they represent users that have covariance matrices quite different from any other user, and thus cannot be assigned to any group. As already observed in [64], the best strategy for dealing with these users is to schedule them in dedicated slots.

The pseudo-code for most parts of DBSCAN is given in [66] and is used in this work without modification. The pseudo-code used for the *RegionQuery* function, that is not reported in the original work, is given here. This function returns the Eps-neighborhood ($Neps$ in the pseudo-code) of a given point p, and it is implemented by Algorithm 2.1. The improved version of this function will be described later. The distance function used here is the chordal distance between matrices, applied to the eigenspaces of the users' covariance matrices:

$$\mathrm{dist}(\mathbf{U}_1, \mathbf{U}_2) = \left\| \mathbf{U}_1\mathbf{U}_1^{\mathrm{H}} - \mathbf{U}_2\mathbf{U}_2^{\mathrm{H}} \right\|_F^2 \tag{20}$$

```
Input: p, Eps
  for i := 0 to SetOfPoints.size do
    q := SetOfPoints.get(i)
    // Check that q is not assigned to other clusters already
    if q.ClusterId IN {NONE,NOISE,p.ClusterId} then
      if dist(p,q)<Eps then
        Neps.append(q)
      end if
    end if
  end for
return Neps
```

Algorithm 2.1: RegionQuery for the original DBSCAN.

2.3.1.3 Improved DBSCAN

DBSCAN is effective at detecting clusters when they are delimited by a clear decrease of data points density, i.e. when the original data is already spontaneously clustered. This is assumed in [42], where the groups of users are separated by a certain distance. In this work, the totally opposite scenario is targeted, where there are no clearly identifiable groups, but it is necessary to artificially create them.

Suppose that there is a dense crowd of mobile users spread over a large area, i.e. at an outdoor concert or similarly crowded events, with no significant fluctuations in density, and JSDM is used for downlink transmission. The area of interest should be large enough to span the coverage area of the closest base stations, which is likely if micro/pico cells are deployed. With DBSCAN, each of these base stations would only detect a single large cluster of users in its coverage area, because their density stays essentially constant. However, this is detrimental for the use of JSDM for two reasons: (i) the number of groups that can be concurrently served is reduced to one, and (ii) users sufficiently distant from each other will have different covariance matrices, so it is not possible to find a first-stage precoder that applies to the entire group.

A better strategy to handle this situation is to split the large cluster into smaller clusters that have a limited “size”, measured through the chordal distance between the eigenspaces of the users. It should be tuned so that all the users of one group can be served with the same first-step precoding matrix. Then, the base station would select a set of groups that are suitable for co-scheduling, i.e. groups that are sufficiently separated in the angular domain. If such condition is not met, they would still experience significant mutual interference [54].

```
Input: p, Eps, Dmax
   // Calculate centroid of the cluster where p is assigned
   c := SetOfPoints.centroid(p.ClusterId)
   for i := 0 to SetOfPoints.size do
      q := SetOfPoints.get(i)
      // Check that q is not assigned to other clusters already
      if q.ClusterId IN {NONE,NOISE,p.ClusterId} then
         if dist(p,q)<Eps and dist(c,q)<Dmax/2 then
            Neps.append(q)
         end if
      end if
   end for
return Neps
```

Algorithm 2.2: RegionQuery for the improved DBSCAN.

It would be cumbersome to have different clustering methods for many scenarios, as it would be necessary to detect such scenarios as well. Instead, DBSCAN has been modified to work as desired in both these extreme cases. It should also cover all the intermediate cases, although that hasn't been tested yet. The key idea is to modify the *RegionQuery* function so that when it builds the Eps-neighborhood of point p, it also takes into account the centroid of the cluster where p belongs. Define $D_{max} \in R^+$ as the maximum acceptable diameter for a cluster. Therefore, if a point is further away than $D_{max}/2$ from the centroid, it is not included in the Eps-neighborhood of p. The extended *RegionQuery* function is reported in Algorithm 2.2. Given the eigenspaces $\mathbf{U}_1, \dots, \mathbf{U}_N \in \mathbb{C}^{M \times b}$ and the corresponding eigenvalues $D_1, \dots, D_N \in \mathbb{C}^b$ of the N users assigned to a cluster, the eigenspace of its centroid is calculated as:

$$\mathbf{U}_c = \mathrm{eig}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{U}_i \mathrm{diag}(D_i)\mathbf{U}_i^{\mathrm{H}}\right] \tag{21}$$

where $\mathrm{diag}(x)$ is the diagonal matrix with the elements of x on its diagonal. The centroid is updated for each invocation of *RegionQuery*, because new points could have been added between subsequent invocation.

This updated formulation introduces new properties, while retaining others that already existed in DBSCAN. First of all, the detection of border points with respect to Eps and $minPts$ is untouched, so the limits of clusters that already exist are still recognized. Thus, in the scenario of [64] with small and well-separated clusters, the modified DBSCAN works exactly as the original one, provided that the new parameter D_{max} is sufficiently large. In fact, this is true in any case: one can always find a D_{max} so large that the new DBSCAN behaves as the original one, with $D_{max} \to +\infty$ as the extreme case.

As for the new scenario with a single large cluster, the selection of D_{max} creates different cases. Note that as D_{max} decreases, the number of detected groups $G \in \mathbb{N}$ grows and their average user count $\hat{G} \in \mathbb{R}^+$ decreases. The possible outcomes, differentiated for their effect on the JSDM precoding, are listed here in the order of decreasing D_{max}:

- For $D_{max} \to +\infty$, the new DBSCAN behaves as the original DBSCAN, i.e. $G = 1$. The difference between user covariances can be large, and JSDM is not useful.
- For some value of D_{max} , the algorithm results in $G > 1$, but the clusters are still too large to meet the assumption of almost-identical covariance throughout one group. JSDM is still not usable at its full potential.
- For decreasing D_{max}, G gets larger and the groups get smaller with respect to the distance between users' eigenspaces. As long as $\hat{G} \geq b$, JSDM can be used with high performance.
- For even smaller D_{max}, the groups get so small that $\hat{G} < b$. JSDM can still be used, but less users can be served, and the sum throughput diminishes.
- For $D_{max} \to 0$, all the users are marked as outliers ($\hat{G} = 1$) and JSDM can't

be used anymore.

Thus, the value chosen for D_{max} is critical for the outcome of the clustering step, and it should be selected according to the specific application and its requirements. For use in the context of JSDM, it is possible observe the following: D_{max} is expressed with respect to the chordal distance between users' eigenspaces, which directly affects the performance of the two-step precoder. Thus, the optimal value for a given scenario is very likely to work in most other situations: when there are large groups, they would be split in smaller groups of reasonable size according to D_{max}, and when they are small, they would just be detected and used as they are. Ultimately, this modified version of DBSCAN would be able to handle both the extreme cases outlined here, intermediate situations, and scenarios with both small and large groups, without additional tuning.

2.3.2 Numerical results

The improved version of DBSCAN has been evaluated through computer simulations, by assuming a large-crowd scenario as described in Section 2.3.1.3. Specifically, mobile users are evenly distributed over a sub-region of a base station sector, as shown in Figure 2.14. The distance from the base station varies from 100 to 225 m, and the azimuth ranges from −60° to 60°. The angular spread varies with the distance, from 28° for the closest users to 13.5° for the farthest users.

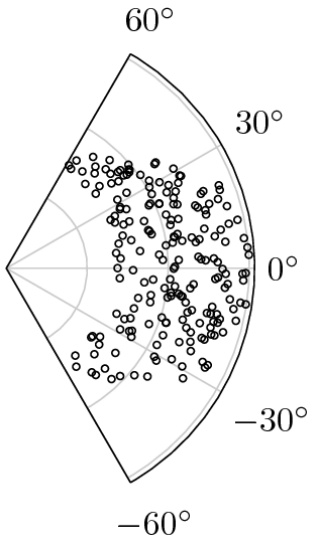


Figure 2.14: Positions of the mobile users.

Parameters from [64] are reused as much as possible to make comparison easier: the base station uses a Massive MIMO array with 256 antennas, with $\lambda/2$ spacing, center frequency of 2.5 GHz, and JSDM for precoding. For the first stage, the design parameter b of JSDM is set to 6, and the first- stage precoder, valid for all the users in the same group, is matched to the covariance matrix of one of the scheduled users, that is selected randomly [54]. As for the second-stage precoder, it uses RZF [44] and Per-Group Processing (PGP) [54]. For channel modeling, the Kronecker correlation model [67] together with the one-ring scattering model [68].

At the clustering step, the improved version of DBSCAN is employed, with parameters $Eps = 3.0$, $minPts = 3$, and D_{max} from 0.5 to 24. After clustering, an ideal selection is assumed for which clusters are actually served. Since the best results are obtained when the clusters are well-separated in the angular domain, the three clusters closest to azimuth directions −60°, 0°, and 60° are selected. After the group selection, there is an intra-group scheduling to randomly select $b = 6$ users within each group, unless the group already has 6 users or less. A more realistic selection of the clusters is left for future study.

The receivers are equipped with a single antenna, and the total equivalent Signal to Noise Ratio (SNR) at the end user side is 50 dB. This means that if a single user is served and no interference is considered, its SNR is 50 dB. With multiple users, the equivalent SNR is split among them in equal parts, and then their individual

intra-group interference and inter-group interference are also added. After all the precoding and scheduling steps, the final SINR experienced by each user is evaluated, and the Spectral Efficiency (SE) is derived using Shannon's capacity formula. The entire process is repeated for 30 independent channel realizations.

Figure 2 shows the number of groups detected during the clustering phase, as a function of D_{max} , together with the total number of users served in each case. For $D_{max} \geq 24$ there is only one group, which means that the algorithm is working like the classical DBSCAN, treating the entire population as a single density-connected cluster. Consequently, only one group can be scheduled, and only $b = 6$ users can be served. For lower values of D_{max}, more and more sub-clusters are created, and in most cases, it is possible to serve 3 groups with 6 users each, reaching 18 users in total. However, with $D_{max} < 4$, the number of served users starts decreasing, because groups get increasingly smaller, to the point of containing less than 6 users. As a consequence, the multiplexing gain is reduced. Also note that at the lowest considered value of $D_{max} = 0.5$, almost half of the users are marked as noise point and thus can't be served as part of a cluster.

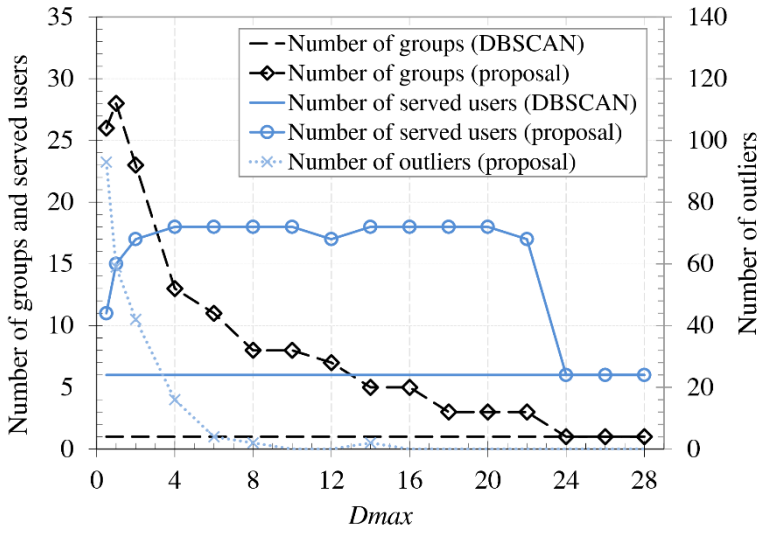


Figure 2.15: Number of detected groups and served users.

Figure 3 shows the sum SE and the user SE, both as a function of D_{max}. They are proportional for most of the values, i.e. for the entire range where the number of served users is high and stable. However, there is a sharp contrast between the left part ($4 \leq D_{max} \leq 12$) and the right part ($12 < D_{max} \leq 22$), as the left part exhibits consistently higher values than the right one. The reason is that for $D_{max} > 12$, the groups become so large that some adjacent groups must be scheduled together, which creates high inter-group interference and thus lower SE values. Please note that the performance in the left part is quite stable, thus the optimal value for D_{max} does not need to be estimated with extreme accuracy.

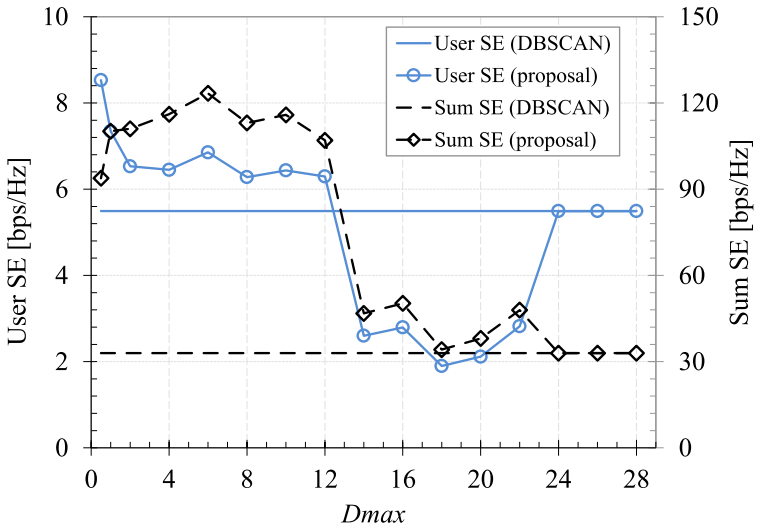


Figure 2.16: Average user SE and sum SE.

At the extremes of the investigated region, the user SE increases but the sum SE decreases. For the sum SE, it is easily noted that the lower values are due to the lower number of served users. As for the user SE, the reason is different for the two cases. With very low values of D_{max}, the chordal distance between the user eigenspaces is reduced, so the first- stage precoder is better matched to all the users in the group, leading to higher SE. Instead, when $D_{max} \geq 24$, the user SE suddenly increases because there is only one group, and thus the inter-group interference disappears. However, this comes at the cost of serving less users and is thus undesirable.

Figure 4 shows the CDF of the user spectral efficiency, for different values of

D_{max}. Again, there is a sharp difference between $D_{max} > 12$ region (dashed lines) and the $D_{max} \leq 12$ region (solid lines). In the first case, many users have low throughput and only few users have high throughput, similarly to a maximum-throughout scheduler with low fairness. This is due not only to both inter-group interference between adjacent groups, but also intra-group imbalance: recall that the first-stage precoder is based only on one of the users, and in large groups other users may not be well matched to it. Conversely, in the second case, the CDFs look more like a round-robin scheduler with higher fairness. In this regime, inter-group interference is more limited because the scheduled groups are not adjacent, and intra-group imbalance is also reduced due to the lower group size. This confirms that the ideal operating region is $4 \leq D_{max} \leq 12$ in the adopted scenario.

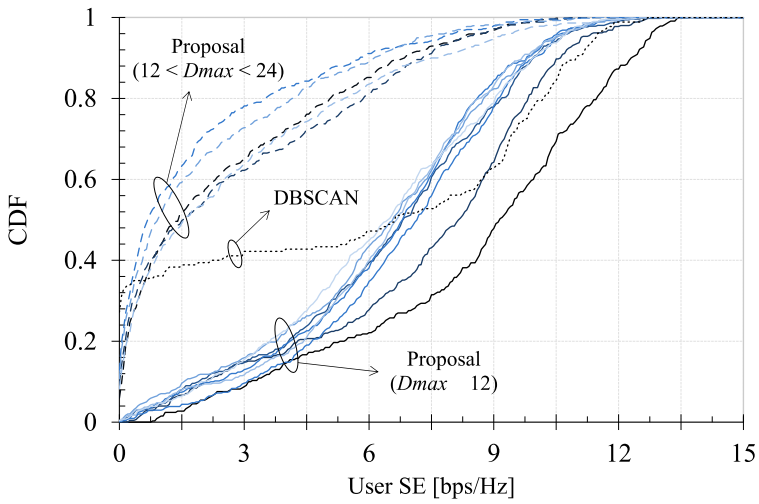


Figure 2.17: Cumulative Density Function (CDF) of user SE [bps/hz].

Regarding complexity, computing the additional distance from the cluster center can make *RegionQuery* twice as expensive. However, users that have been already classified in other groups are completely ignored by it, therefore creating smaller groups removes users from calculations earlier. In the optimal operating range for the considered scenario, this results in a net complexity gain compared to the original DBSCAN.

2.4 Game-theoretic resource allocation for uplink with Massive MIMO

This section describes a decentralized and energy-efficient resource allocation method for the uplink of a Massive MIMO cellular system. Typically, the resource assignment is performed by the packet scheduler, even for the uplink direction, and only one or few users are scheduled. With Massive MIMO multiple users can be scheduled at the same time, and the allocation problem can quickly become complicated. Using an approach based on non-cooperative game theory, it is possible to allow mobile users to autonomously select the most appropriate time-frequency resources, the MCS, and the power allocation profile in the frequency direction, while avoiding mutual interference as much as possible. The proposed method can consider a per-user Quality-of-Service (QoS) constraint in the form of a minimum average data rate, while also maximizing the energy efficiency (which is crucial for the lifetime of mobile terminals) and enforcing a maximum limit for the radiated power (as typically required by regulations).

Section 2.4.1 reports the formal description of the problem and the proposed method, while Section 0 explains the setup for numerical simulations and Section 2.4.3 presents and discusses the results.

2.4.1 Formulation of the resource allocation scheme

The reference scenario is a multi-cell system where cells share the same spectrum composed by B consecutive sub-channels. Let $\mathcal{B} = \{0, 1, \dots, B-1\}$ be the set of these sub-channels available in every cell, each one with the same width equal to Δf. The set of users available in a given cell of the scenario are denoted by $\mathcal{U} = \{0, 1, \dots, U-1\}$, where $U = |\mathcal{U}|$ is the total number of users. During the i-th TTI, a user u may transmit its data by (potentially) using the whole set of frequencies available in the cell (according to the multi-user transmission capability offered by Massive MIMO, however, many users can send packets by using the same set of radio resources). The power level in each sub-channel $b \in \mathcal{B}$ is equal to $\Pi_{i,u}(b)$.

Therefore, the power profile of user u during TTI i is denoted by $\mathbf{\Pi}_{i,u} = \{\Pi_{i,u}(0);\ \Pi_{i,u}(1);\ \dots, \Pi_{i,u}(B-1)\}$. It is also assumed that the total power level that each user can use for the transmission is limited to a given value P_M.

Also, $R_{i,u}$ represents the bit-rate of the user u during TTI i. Note that, according to the LTE-A behaviour, its value depends on the selected MCS scheme and on the number of sub-channels adopted for the transmission. At the application layer, each user $u \in \mathcal{U}$ is subject to a QoS constraint: the perceived throughput $R_{i,u}$, averaged over the time interval i, should not be lower than a target value, O_u. Note that, if this condition is not met, an *outage condition* occurs. Moreover, let $\Gamma_{i,u}(b)$ be the SINR measured by the base station for user u on sub-channel b during the TTI i.

Symbol	Meaning
b	Sub-channel index
B	Total number of sub-channels
$\mathcal{B}$	Set of available sub-channels
$E_{i,u}$	Energy efficiency during TTI i for user u
$\boldsymbol{h}_{i,u}(b)$	Vector of channel gains from user u to its base station on sub-channel b during TTI i
M	Number of receiving antennas at the base station
N_0	Thermal noise power
O_u	Target bit-rate of user u
P_M	Maximum allowed power for each mobile user
$R_{i,u}$	Bit-rate achieved during TTI i for user u
S	Number of sub-channels used to calculate the effective SINR $\Gamma_{i,u}^{eff}$
$\mathcal{S}$	Subset of sub-channels used to calculate the effective SINR $\Gamma_{i,u}^{eff}$
u	User index
$\mathcal{U}$	Set of available users
U	Total number of users
i	TTI index
$\Gamma_{i,u}(b)$	SINR that user u gets on sub-channel b during TTI i
$\Gamma_{i,u}^{eff}$	Effective SINR of user u on all sub-channels during TTI i
Δf	Sub-channel spacing in the frequency domain
$\lambda_{i,u}^{a}$	Waterfilling level for maximum energy efficiency of user u during TTI i

$\lambda_{i,u}^{b}$	Waterfilling level for maximum power of user u during TTI i
$\lambda_{i,u}^{c}$	Waterfilling level for maximum power of user u during TTI i
$\lambda_{i,u}^{*}$	Final waterfilling level for user u during TTI i
$\Pi_{i,u}(b)$	Transmission power level used by user u on sub-channel b during TTI i
$\widetilde{\mathbf{\Pi}}$	Mobile terminal non-radiative power
$\mathbf{\Pi}_{i,u}$	Power profile used by user u during TTI i
$\mathbf{\Pi}_{i,u}^{*}$	Optimal power profile used by user u during TTI i

Table 2.5: List of mathematical symbols

In line with [69], the energy efficiency for the user u during the TTI i can be defined as the ratio between the bit-rate and the total power used for transmission:

$$E_{i,u} = \frac{R_{i,u}}{\widetilde{\mathbf{\Pi}} + \sum_{b=0}^{B-1} \Pi_{i,u}(b)} \tag{22}$$

where $\widetilde{\mathbf{\Pi}}$ accounts for the non-radiative power dissipated in the electronic circuitry of the mobile terminal, assumed to be a constant value valid for all users during all TTIs. For easier reading, a summary of symbols is reported in Table 2.5.

The resource allocation technique formulated in this section aims at evaluating, for each user, the optimal power profile $\mathbf{\Pi}_{i,u}^{*}$ that is able to maximize the energy efficiency, while ensuring both QoS and power constraints. It can be formally expressed, for all $u \in \mathcal{U}$ and for each TTI i, through the optimization problem:

$$\mathbf{\Pi}_{i,u}^{*} = \arg\max_{\mathbf{\Pi}_{i,u}} E_{i,u} \tag{23}$$

Subject to: $R_{i,u} \geq O_u$ (24)

$$\sum_{b=0}^{B-1} \mathbf{\Pi}_{i,u}(b) \leq P_M \tag{25}$$

In [69], the optimization problem described in Eq. (23) is formulated as a non-cooperative game between users, where the utility function to maximize is $E_{i,u}$. The term non-cooperative is used to highlight that no explicit cooperation between users

is employed [70]. Nevertheless, the behavior of all active players (i.e., users involved in the game that aim at transmitting data over the same spectrum) is indirectly taken into account with the SINR values $\Gamma_{i,u}(b), \forall b \in \mathcal{B}$ that are periodically sent by the base station as discussed below.

The solution of the game-theoretic problem is investigated in the form of Generalized Nash Equilibrium (GNE) states [71]. In these states, each user adopts the power allocation profile which maximizes its utility function, given the power allocation profile of all the other users. Also, the power allocation profile is expected to remain unchanged if all users keep using the same power profiles, because every deviation would result in a lower utility outcome. Thus, a GNE state is intrinsically stable, as all the users will stick to it because they have no convenience in doing otherwise. This property is useful in the context of distributed resources allocation mechanisms, where the stability cannot be guaranteed by a centralized entity. The work presented in [69] demonstrates the existence and uniqueness conditions for GNE states of the aforementioned problem. The iterative method useful to reach such states is also presented.

Now, without demonstrating each single step of the algorithm already discussed in [69], it is possible to summarize that the optimal power profile is calculated by each user u as in the following:

- The set of SINRs $\Gamma_{i,u}(b), \forall b \in B$ is received from the serving base station.
- A WaterFilling (WF) formula is used to calculate the power profile. During TTI i, user u calculates the power level of the sub-channel b as:

$$\Pi_{i,u}(b) = \max\left\{\frac{1}{\lambda^*_{i,u}} - \frac{\Pi_{\mathrm{i-1,u}}(b)}{\Gamma_{\mathrm{i-i,u}}(b)}, 0\right\} \tag{26}$$

  where $\lambda^*_{i,u}$ is the water level of the WF operator. Note that in [69] it is proved that a power profile is in a GNE state if and only if it is calculated as in Eq. (26).
- The parameter $\lambda^*_{i,u}$ in Eq. (26) is set to $\lambda^*_{i,u} = \min\left(\lambda^a_{i,u}, \lambda^b_{i,u}\right)$, where $\lambda^a_{i,u}$ is

calculated by solving the optimization problem in Eq. (23) without considering the minimum rate constraint and $\lambda_{i,u}^{b}$ is calculated by imposing the minimum rate constraint, as in Eq. (24). More details about the methodology used for computing both $\lambda_{i,u}^{a}$ and $\lambda_{i,u}^{b}$ can be found in [69].

- With reference to the constraint described in Eq. (24), the achievable transmission bit-rate, $R_{i,u}$, with the described power profile, is ideally evaluated through the Shannon theorem:

$$R_{i,u} = \sum_{b=0}^{B-1} \Delta f \log_2\left(1 + \Gamma_{i,u}(b)\right) \tag{27}$$

- The algorithm is repeated multiple times until stability is reached. After that, no further updates are made.

Unfortunately, the aforementioned approach cannot be implemented on top of a real technology due to the presence of some limitations. First, since the optimization problem did not consider the power constraint, the final solution may bring to power levels higher than those really supported by mobile terminals, as well as imposed by normative regulations. Second, a time-invariant channel is considered, which means that the propagation loss used for computing $\Gamma_{i,u}(b)$ is only influenced by the path loss; on the contrary, fading effects are ignored at all. In these conditions, the WF algorithm is able to reach the stability after a few iterations and, as a consequence, its ability to adapt the power profile to network dynamics (elicited by both channel and application traffic) is not investigated in depth. Third, $R_{i,u}$ is evaluated without considering any physical details of a real technology. In LTE-A, for instance, the user has to transmit data by using the same MCS for all the selected sub-channels and the achieved bit-rate could be significantly different with respect to the Shannon bound (especially in the presence of intense frequency selective fading). Fourth, it is supposed that each user may potentially transmit over the entire spectrum, even if the achieved bit-rate is higher that the required one. This assumption can produce huge interference levels (and hence lower performance) in highly loaded scenarios. Finally, the WF algorithm described in Eq. (26) assumes that each user receives CSI indicators from the base

station every TTI, with no delays. Therefore, the impact that a real protocol may have on CSI reporting delays is completely ignored.

The conceived solution extends the work presented in [69] in order to solve all the aforementioned issues, while ensuring at the same time its feasibility in a real environment under the constraints provided by the underlying technology (including physical transmission rate and maximum transmission power).

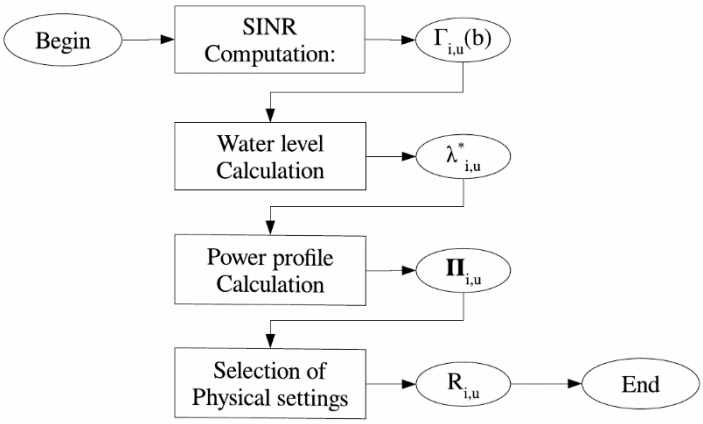


Figure 2.18: Algorithm flow chart

The set of tasks implemented by each user are reported in Figure 2.18. An in-depth discussion is, instead, reported below.

1. *SINR computations*: During the i-th TTI, the base station uses the MMSE receiver for calculating the SINR, $\Gamma_{i,u}(b)$, related to the user u and the sub-channel b (note that MMSE is used as the baseline decoder in LTE-A) as:

$$\Gamma_{i,u}(b) = \Pi_{i,u}(b)\boldsymbol{h}_{i,u}^{\mathrm{H}}(b)\left[N_0\mathbf{I}_M + \chi_{tot,u}(b)\right]^{-1}\boldsymbol{h}_{i,u}(b) \quad (28)$$

with $\chi_{tot,u}(b) = \sum_{j=1,j\neq u}^{U} \Pi_{i,j}(b)\boldsymbol{h}_{i,j}(b)\boldsymbol{h}_{i,j}^{\mathrm{H}}(b)$

where $\boldsymbol{h}_{i,u}(b)$ is the array of channel gains from user u to each antenna of the base station, $\boldsymbol{h}_{i,u}^{H}(b)$ is its Hermitian conjugate, N_0 is the thermal

noise power, $\mathbf{I}_M$ is the $M \times M$ identity matrix, and $\chi_{tot,u}(b)$ is the inter-user interference, which depends on the transmission power levels and channel gains of the other users transmitting on the same sub-channel and in the same TTI.

Without loss of generality, it is assumed that CSI feedbacks estimated by the base station are received by the user after τ TTIs from its calculation (i.e., τ represents the CQI reporting delay expressed in TTIs). Therefore, at the beginning of the i-th TTI, the user considers as input parameters of the algorithm the set of values $\Gamma_{i-\tau,u}(b), \forall b \in \mathcal{B}$. Starting from the received CSI feedbacks, the user can calculate the wideband effective SINR, $\Gamma_{i,u}^{eff}$, as:

$$\Gamma_{i,u}^{eff} = -\beta \ln\left(\frac{1}{S}\sum_{b\in\mathcal{S}} \exp\left\{-\frac{\Gamma_{i-\tau,u}(b)}{\beta}\right\}\right) \quad (29)$$

where S and β are the number of sub-channels used to calculate the effective SINR (contained in a subset $\mathcal{S} \subseteq \mathcal{B}$) and a scaling factor that takes into account the different MCSs [72], respectively.

2. *Calculation of the water level for the WF operator*: Differently from [69], $\lambda_{i,u}^{*}$ is obtained as:

$$\lambda_{i,u}^{*} = \max\{min\{\lambda_{i,u}^{a}, \lambda_{i,u}^{b}\}, \lambda_{i,u}^{c}\} \quad (30)$$

where:

- $\lambda_{i,u}^{a}$ represents the solution which maximizes the energy efficiency without taking into account any constraint. The calculation method is very similar to [69].
- $\lambda_{i,u}^{b}$ is evaluated as the water level that minimizes the transmission power for a given bit-rate value (set to the minimum required rate O_u). Unlike [69], where the physical data rate is estimated by the Shannon theorem, in this contribution $R_{i,u}$ is obtained as a function $\mathcal{F}(\cdot)$ of the wideband effective SINR and the number of active sub-channels, as:

$$R_{i,u} = \mathcal{F}\left(\Gamma_{i,u}^{eff}, S\right) \tag{31}$$

In this way, all the details of the underlying technology are carefully taken into account. A detailed description of the function $\mathcal{F}(\cdot)$ can be found in [73].

- $\lambda_{i,u}^{c}$ is calculated as the water level which maximizes the bit-rate using a fixed total power, which is set to the maximum power P_M that mobile terminals are allowed to use.

3. *Evaluation of the optimal power profile*: As a next step of the procedure, the power profile is calculated, although in its preliminary version, through Eq. (26). Moreover, it is subsequently revised in order to reduce the interference levels in highly loaded scenarios, as well as avoid transmitting over sub-channels experiencing peak fading attenuations.

   To this end, not all of the available sub-channels are used for the transmission. Indeed, just a subset of them is selected for this purpose through an iterative procedure. First, all the sub-channels are sorted in decreasing order of SINR. Starting from the sub-channel with the highest SINR value, the corresponding effective SINR and the physical data rate $R_{i,u}$ are obtained using Eqs. (29) and (31), respectively. If the resulting bit-rate is already higher than the target rate, then the iteration stops, and this single channel is used. If $R_{i,u} < O_u$, the calculation is repeated with the best two channels and the bit-rate is checked again. If needed, more channels are added until the required bit-rate O_u is reached. When the required number of channels is found, the remaining channels are excluded and the correspondig allocated power is set to zero.

   The rationale behind the aforementioned sub-channel selection is that a peak fading attenuation experienced in few sub-channels brings

to a low wideband effective SINR value, thus providing a limited physical transmission rate. In other words, the weakest sub-channels are a limiting factor for the throughput, so it is preferable to use as few sub-channels as possible (and choose those with the highest SINR). This issue is not considered in [69], because the sub-channels are assumed to be independent and any additional sub-channel always results in a higher throughput. Also, reducing the number of sub-channels reduces the total power used for transmission and the multi-user interference, allowing more users to transmit simultaneously.

4. *Selection of physical settings*: at the end, the user may finally select the MCS to use for the transmission. This task is done by using the well-known strategy widely accepted for the LTE-A technology [59].

2.4.2 Performance evaluation methodology

The performance of the proposed algorithm is tested using the LTE-Sim open source simulator for cellular networks [59]. To further demonstrate the effectiveness of the designed solution, the comparison with other baseline radio resource management techniques, built on top of the Massive MIMO physical layer, is also provided. In particular, *proportional fair* and *multi-user Massive MIMO* have been considered as reference techniques.

With the proportional fair, the scheduling of uplink resources is fully orchestrated by the base station. In each time interval, and for each user, a scheduling metric is calculated as the ratio between expected physical data rate and the average throughput achieved during past sub-frames. Then, only the user with the highest metric is selected for transmission during that time interval [74]. The inter-user interference within a single cell is completely absent. At the physical layer, mobile terminals leverage a flat power profile because the maximum allowed power is uniformly distributed in the frequency domain. Moreover, Massive MIMO is used to guarantee an improved SINR, due to the combination of a higher number of received signals.

Multi-user Massive MIMO extends the MU-MIMO transmission scheme by

assuming the presence of a higher number of base station antennas. It allows all the users to transmit data at the same time, over the same spectrum. The allocation of uplink resources is not controlled anymore by the packet scheduler. Similar to the proportional fair scheme, multi-user Massive MIMO also uses a flat power profile. The Massive MIMO physical layer is in charge of separating all the multiplexed signals.

Simulations assume a multi-cell environment with 7 cells arranged on a hexagonal grid. Each cell has a radius of 500 m, with an inter-site distance of 866 m. In each cell, a number of users (ranging from 5 to 25) is randomly distributed over the entire simulation area with uniform probability.

A system bandwidth of 5 MHz is used, in the operating band 1920 − 1924.32 MHz (corresponding to LTE band 1), with a frequency reuse factor of 1. This generates strong inter-cell interference and shows how the proposed solution works in extreme conditions. A Massive MIMO physical layer is always considered and the number of antennas at the base station is set to 32.

To take into account the impact of the signal propagation, the 3GPP model for urban macrocells is considered [75]. In addition, other phenomena included are:

- large-scale shadowing, modeled as a log-normal random variable with 0 mean and 8 dB standard deviation [75];
- penetration loss, set to a constant value of 10 dB [75];
- fast fading, introduced as a random variable with Rayleigh distribution, generated using the Jakes' model with 6, 8, 10 or 12 paths [59].

A BLER model based on the Additive White Gaussian Noise (AWGN) channel is taken into account [59]. Each mobile terminal can use a maximum transmit power of 23 dBm, whereas the non-radiative power P_M is set to a constant value of 10 dBm. User mobility is modeled using the random direction model [76] with speed equal to 3 km/h and 120 km/h.

To show the impact of a real protocol implementation, the CSI report delay is set to 2 ms. However, the CSI feedback is assumed to be ideal with regard to

precision and overhead: the user receives a non-quantized channel gain value for each sub-channel, and no occupation of the downlink channel is considered. Quantization, wideband channel status, and downlink occupation can be topics for future research.

To evaluate the behavior of the proposed strategy in different traffic load conditions, the target bit-rates are set to either 300 kb/s or 1000 kb/s per user. When allocating radio resources to a user, the number of bits to be transmitted during one sub-frame is rounded by excess to the nearest TBS allowed in LTE.

Parameter	**Value**
Cell radius	500 m
Number of cells	7
Inter-cell distance	866 m
Receiver type	MMSE
Bandwidth	5 MHz
Number of sub-channels B	25
Operating frequency	1920 − 1924.32 MHz
Number of users per cell	5, 10, 15, 20, 25
Minimum user bit-rate (O_u)	300 or 1000 kbps
Maximum user transmit power	23 dBm
Mobile terminal non-radiative power ($\widetilde{\mathbf{\Pi}}$)	10 dBm
Number of transmit antennas at mobile user	1
User speed	3 or 120 km/h
Mobility model	Random direction
MIMO configuration	Massive MIMO with 32 antennas at the base station side
Path-loss model	$128.1 + 37.6 \log_{10}(d)$ [dB]
Shadowing model	Lognormal with 8 dB std. dev.
Fast fading model	Jakes' model with 6, 8, 10, or 12 paths
Penetration loss	10 dB
Number of runs	25
CSI delay	2 ms

Table 2.6: Simulation parameters

Each simulation lasts 10 seconds and all the simulation results have been averaged over 25 runs. The 95% confidence interval has been also computed by using the t-Student statistic. However, it has not been included in all the graphs because its value is lower than the marker size.

All simulation parameters are summarized in Table 2.6.

2.4.3 Simulation results

The set of metrics measured for the performance evaluation include: average per-user transmission power, average power consumption at the base station, average per-user energy efficiency, average MCS used for the transmission, average number of sub-channels that each user selects for transmitting application data, average application goodput experienced by each user during the simulation, peak data rate, system spectral efficiency, and average outage probability.

2.4.3.1 Power consumptions and energy efficiency

The average power consumption measured by mobile terminals is reported in Table 2.7. With multi-user Massive MIMO, users always consume the maximum power value PM during every time slot. Therefore, the highest power consumption is registered, and that value is not influenced by any simulation parameter. On the contrary, proportional fair guarantees a power consumption that decreases as the network load becomes higher, and remains constant with respect to target data rate and users' speed. In this case, each mobile terminal adopts a power level of PM when it is scheduled, and 0 otherwise. As a result, the number of time slots that each user may spend for the transmission, and therefore the overall power consumption, decreases when the number of users attached to the base station grows. The designed algorithm always registers the lowest power consumption. In particular, it can be observed that the power consumption grows with the required bit-rate, because each user targets a higher SINR to increase its physical data rate. Similarly, to compensate for the additional multi-user interference in highly loaded scenarios, a higher power consumption is observed. Moreover, the impact of user speed on the power consumption is generally very low.

			Number of users				
			5	10	15	20	25
Target rate [kbps]	300	Proportional fair (for both 3 km/h and 120 km/h)	39.91	19.96	13.30	9.98	7.98
		Multi-user Massive MIMO (for both 3 km/h and 120 km/h)	209.53	209.53	209.53	209.53	209.53
		Proposal (for 3 km/h)	10.39	10.45	10.74	10.90	11.03
		Proposal (for 120 km/h)	10.42	10.55	10.76	10.86	10.92
	1000	Proportional fair (for both 3 km/h and 120 km/h)					
		Multi-user Massive MIMO (for both 3 km/h and 120 km/h)	209.53	209.53	209.53	209.53	209.53
		Proposal (for 3 km/h)	11.85	12.49	13.20	13.35	13.73
		Proposal (for 120 km/h)	11.97	12.36	12.84	13.07	13.30

Table 2.7: Average transmission power registered by mobile terminals [mW]. Results demonstrate how the proposed approach always reaches the lowest power consumption, which is properly set by jointly considering QoS requirements, channel quality, and inter-user interference level.

Even if base stations are generally connected to the power grid (and therefore they are not inherently power-limited), they are responsible for about 90% of the amount of energy spent for running a cellular network. Thus, the reduction of their power demands could be very relevant from an environmental and economic perspective. For this reason, the power consumption occurring at the base station side is also calculated, by means of the analytical model presented in [77]. This model has been formulated for the conventional MIMO transmission scheme. It has been extended to Massive MIMO by assuming that the percentage of the power consumption related to the radio-frequency chain and baseband processing is proportional to the number of physical antennas and the number of active MIMO streams, respectively. Results are reported in Table 2.8. As expected, the base station power consumption is not influenced by the users' speed, because the amount of executed operations remains the same. Moreover, it emerges that multi-user Massive MIMO brings to the highest base station power consumption. In fact, it is in charge to address the highest computational load due to the multi-user decoding (note that the base station power consumption significantly increases with the number of users). On the contrary, proportional fair and the proposed approach just register a base station power consumption that slowly increases with the number of users. In this case, the incremental computational load refers to the processing of

downlink communications (that increases with the network load).

	Number of users				
	5	10	15	20	25
Proportional fair (for both 3 km/h and 120 km/h)	2008.52	2061.24	2118	2178	2242.12
Multi-user Massive MIMO (for both 3 km/h and 120 km/h)	2056.88	2277.56	2726.24	3517.28	4762.36
Proposal@300 kbps (for both 3 km/h and 120 km/h)	1980.6	2044.7	2116.4	2185.4	2259.9
Proposal@1000 kbps (for both 3 km/h and 120 km/h)	2002.7	2070.5	2070.5	2227.8	2331.2

Table 2.8: Average power consumption registered by the base station [W].
Multi-user Massive MIMO registers the highest base station power consumption due to the huge computational load introduced by the multi-user decoding. Thanks to its optimal radio resource allocation, the proposed scheme requires power demands for the base station that are comparable to those needed by the baseline proportional fair.

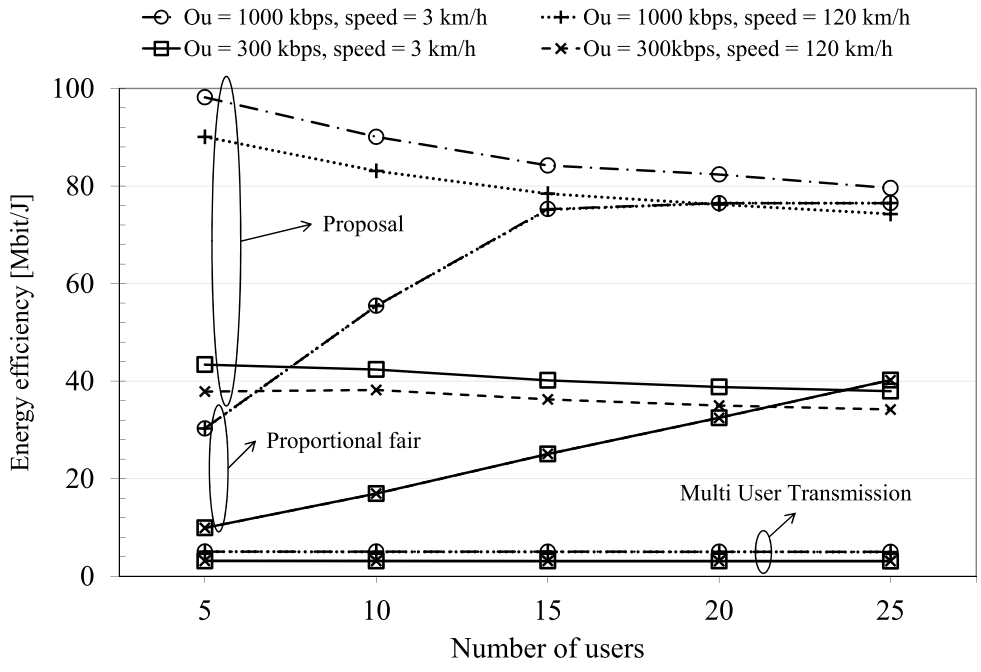


Figure 2.19: Average energy efficiency.
Results clearly confirm that the proposed scheme meets the highest energy efficiency.

Finally, the energy efficiency is a valid metric to demonstrate how energy is effectively used during the transmission. It is obtained as the ratio of the achieved per-user bit-rate and the corresponding transmission power. Figure 2.19 shows that proportional fair ensures an energy efficiency that increases with the network load, because each user transmits for a lower fraction of time and the average power is proportionally lower. On the contrary, multi-user Massive MIMO always registers the lowest energy efficiency, because the maximum power is employed for all the time. With the proposed scheme, instead, the energy efficiency slightly decreases with the number of users, mostly because higher power levels are progressively needed to meet the QoS constraints. Anyway, it is important to remark that the proposed approach guarantees the highest energy efficiency in the most of investigated scenarios. This result well confirms the respect of its design criteria (i.e., the minimization of the energy consumption). Note that proportional fair registers better results only in a scenario with the highest number of users and the target data rate set to 300 kbps. However, other results below will demonstrate that this gain is not important because proportional fair will always register lower user experience data rate, cell throughput, and spectral efficiency.

2.4.3.2 Physical transmission settings and bandwidth usage

			Number of users				
			5	10	15	20	25
Target rate [kbps]	300	Proportional fair (3 km/ h)	25.08	25.23	25.23	25.33	25.34
		Multi-user Massive MIMO (3 km /h)	19.41	18.06	16.37	15.09	13.54
		Proposal (3 km/h)	18.38	18.42	16.43	15.02	14.17
	1000	Proportional fair (3 km /h)	25.08	25.23	25.23	25.33	25.34
		Multi-user Massive MIMO (for 3 km/ h)	19.41	18.06	16.37	15.09	13.54
		Proposal (for 3 km/h)	16.44	14.43	13.50	13.23	12.76
	300	Proportional fair (for 3 km/ h)	24.75	24.81	24.59	24.98	24.90
		Multi-user Massive MIMO (for 3 km /h)	19.10	16.95	14.81	13.02	11.39
		Proposal (for 3 km/h)	17.40	16.62	14.80	14.08	13.67
	1000	Proportional fair (for 3 km/ h)	24.75	24.81	24.59	24.98	24.90
		Multi-user Massive MIMO (for 3 km /h)	19.10	16.95	14.81	13.02	11.39
		Proposal (for 3 km/h)	14.73	13.59	12.72	12.51	12.25

Table 2.9: Average MCS.
When the proposed approach is used, the higher inter-user interference level and the lower power consumption impose to mobile terminals to select lower MCS indexes.

Table 2.9 reports the average MCS index selected by mobile users. As widely recognized, the MCS is selected by the link adaptation process as a function of the quality of the communication link, expressed through the SINR. From one side, the worse the channel condition, the lower the MCS index. From another hand, low transmission power and high interference level generally translate to lower MCS indexes. Now, higher user speeds bring to a worse channel quality. Indeed, users moving at 120 km/h tend to select lower MCS indexes. With proportional fair, users select the highest MCS indexes. In fact, they experience the best channel condition because of the absence of intra-cell interference and the use of the maximum power level. Reported results seem to be independent from the network load, the traffic load and the user speed. Only few statistical variations, due to the variability of the channel quality and the user position, are measured. Intra-cell interference, instead, exists for both multi-user Massive MIMO and the proposed solution. For this reason, they register a lower average MCS index, that also decreases with the network load (because of the increment of the interference level). When multi-user Massive MIMO is used, a constant power profile is always used and the resulting average MCS index does not change with the target user data rate. The proposed solution, however, generally registers the lowest average MCS index. This result can be justified by considering that the formulated algorithm tries to cut down the transmission power, thus generating communication links with lesser quality. At the same time, a higher traffic load produces higher interference levels, which entail, in turns, the selection of lower MCS index.

To provide a further insight on transmission settings, Table 2.10 shows the average number of sub-channels assigned to (or selected by) a user in every time slot. With proportional fair, each user transmits over the entire operative bandwidth only when scheduled and the average number of used sub-channels decreases as the network load increases. Of course, such values do not depend on the target data rate and the users' speed. When multi-user Massive MIMO is used, all users transmit

over the entire bandwidth and the number of sub-channels is always equal to 25.

			Number of users				
			5	10	15	20	25
Target rate [kbps]	300	Proportional fair (for both 3 km/h and 120 km/h)	5	2.5	1.66	1.25	1
		Multi-user Massive MIMO (for both 3 km/h and 120 km/h)	25	25	25	25	25
		Proposal (for 3 km/h)	1.53	1.53	1.84	2.12	2.29
		Proposal (for 120 km/h)	1.57	1.71	2.01	2.18	2.28
	1000	Proportional fair (for both 3 km/h and 120 km/h)	5	2.5	1.66	1.25	1
		Multi-user Massive MIMO (for both 3 km/h and 120 km/h)	25	25	25	25	25
		Proposal (for 3 km/h)	4.84	5.70	6.25	6.37	6.65
		Proposal (for 120 km/h)	5.33	5.85	6.30	6.46	6.62

Table 2.10: Average number of sub-channels per time slot. With the proposed approach, the amount of bandwidth used by mobile terminals is significantly influenced by the network conditions (i.e., network load, traffic load, user speed). Moreover, it is always reduced to the minimum value that guarantees the respect of the QoS constraints.

The proposed algorithm reaches different results: the number of sub-channels selected for the transmission is highly influenced by network conditions. For lower network and traffic loads, a limited amount of bandwidth is required to drain the amount of data generated at the application layer. When the number of users, the target data rate, and the user speed increase, the overall channel conditions decrease (see the comments already provided for the average MCS index) and a higher number of sub-channels is needed to deliver data. It is important to highlight, however, that the proposed scheme tries to reduce the percentage of bandwidth used by each mobile terminal towards the minimum value that guarantees the respect of the QoS constraints.

2.4.3.3 User goodput, spectral efficiency, and peak data rate

The average application goodput is depicted in Figure 2.20. It is evident that proportional fair is not able to reach QoS constraints when the base station is called to serve a higher number of users. When the network load increases, in fact, each user can employ fewer time slots for transmitting application data and too many

users may violate their QoS constraint. Multi-user Massive MIMO always exceeds QoS requirements. When O_u= 300 kbps, for instance, mobile terminals can transmit more data than requested because they can always transmit over the entire available bandwidth, by using the maximum allowed transmission power. This performance gain is reached, however, at the cost of an excessive reduction of the energy efficiency. The proposed algorithm, instead, generally achieves an average user goodput close to the target one, thus demonstrating its ability to adapt transmission settings and reduce energy efficiency, while respecting QoS requirements.

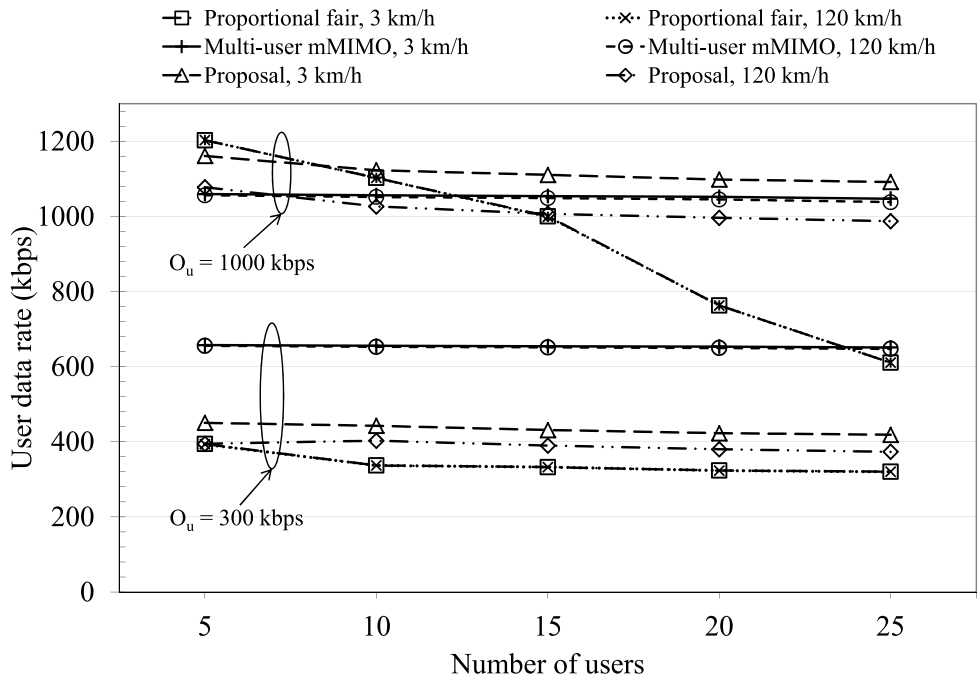


Figure 2.20: Average application goodput.
The proposed approach generally achieves an average user goodput close to the target one.

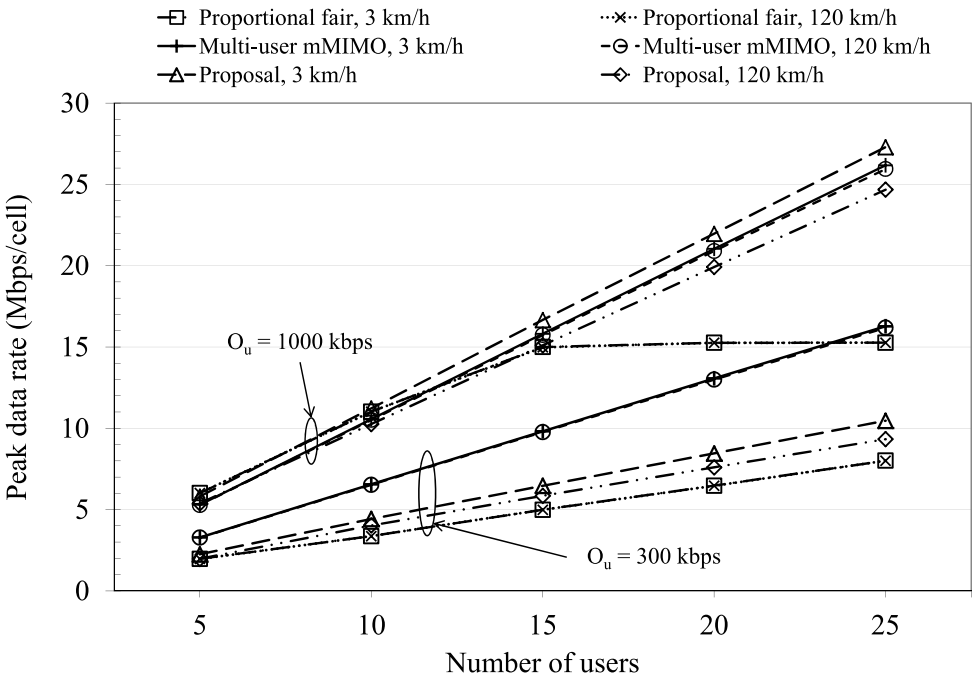


Figure 2.21: Peak data rate.
Results confirm that multi-user Massive MIMO and the proposed approach meet the QoS requirements and achieve a peak data rate that increases with the network load.

Mobile operators are also interested in the peak data rate achieved in a given cell. It is depicted in Figure 2.21. All the results are in line with those just discussed: multi-user Massive MIMO and the proposed solution register a peak data rate that increases with the number of users. proportional fair, however, achieves worse performance (note that when O_u= 1000 kbps, the peak data rate saturates at about 15 Mbps/cell).

Figure 2.22 shows the spectral efficiency, calculated as the ratio of the aggregate transmission bit-rate and the total occupied bandwidth. Also in this case, results confirm the good performance reached by the proposed resource allocation scheme, which always registers higher spectral efficiency when compared with the baseline proportional fair scheme. Similar (or sometime better) results are measured for the multi-user Massive MIMO.

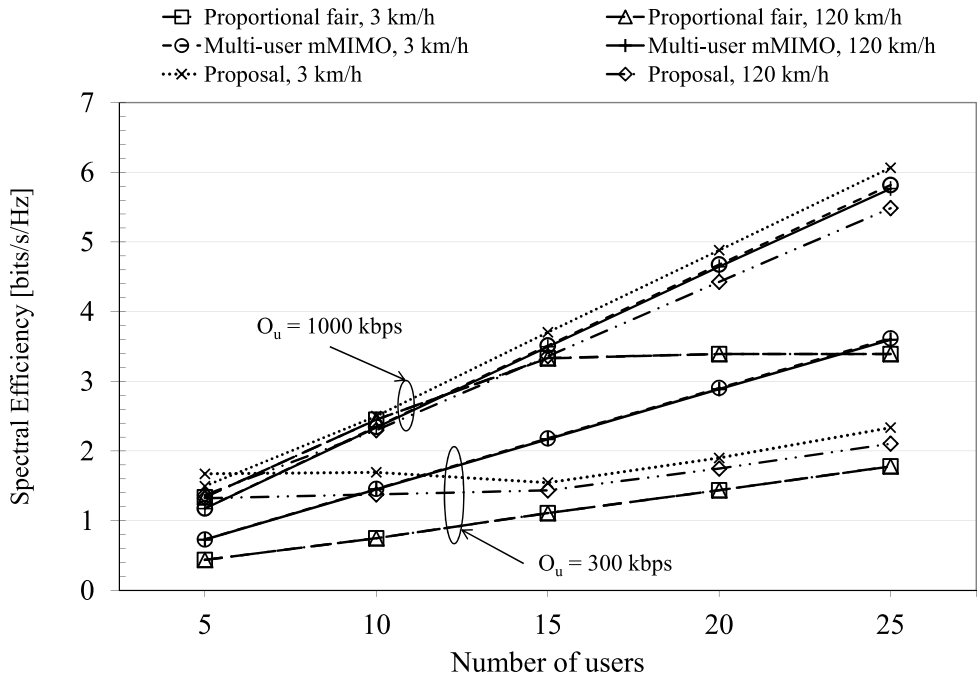


Figure 2.22: Average spectral efficiency.
As expected, the proposed scheme registers a spectral efficiency that increases with the network load. Sometimes, multi-user Massive MIMO reaches better performance because it does not introduce any control on the power consumption.

An interesting behavior is observed for the proposed scheme with O_u= 300 kbps. With a low number of users (up to 15) the spectral efficiency is approximately flat, indicating that a constant number of sub-channels is assigned to each user and they do not overlap. The spectral efficiency is calculated as the ratio of the aggregated goodput and the number of actually allocated sub-channels, so it remains constant as long as there are free sub-channels to employ when the number of users is raised. Instead, with more than 15 users, the spectral efficiency increases because the number of required sub-channels is greater than the available amount. In fact, they begin to overlap in the frequency domain, but the total number of sub-channels used to calculate the spectral efficiency remains constant while the total goodput grows. This confirms that the proposed method guarantees an efficient allocation of

the available sub-channels, and the overlapping of different users is avoided if possible.

2.4.3.4 Outage probability

To conclude, also the outage probability (see Figure 2.23 and Figure 2.24) has been evaluated, by taking the percentage of time in which a user transmits at a bit-rate lower than the target one. This is the most important performance index that effectively highlights the capability of a given access scheme to satisfy users requirements. Obtained results clearly demonstrate that the proposed approach always reaches a low outage probability (better than round-robin and comparable to multi-user Massive MIMO).

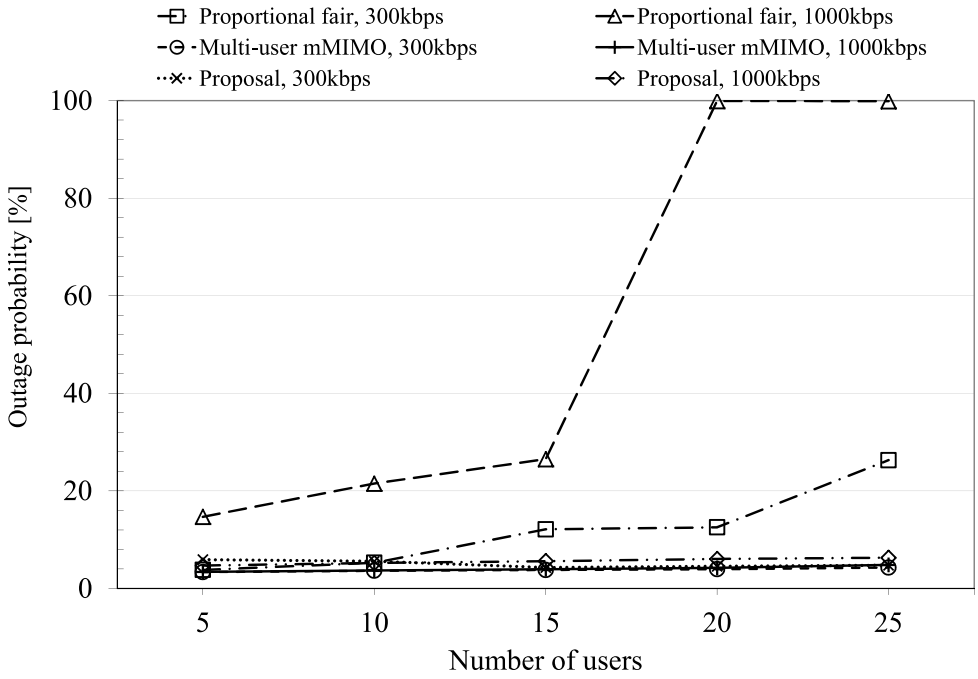


Figure 2.23: Outage probability (3 km/h user speed).
The proposed method can provide a good level of reliability even when the number of users grows.

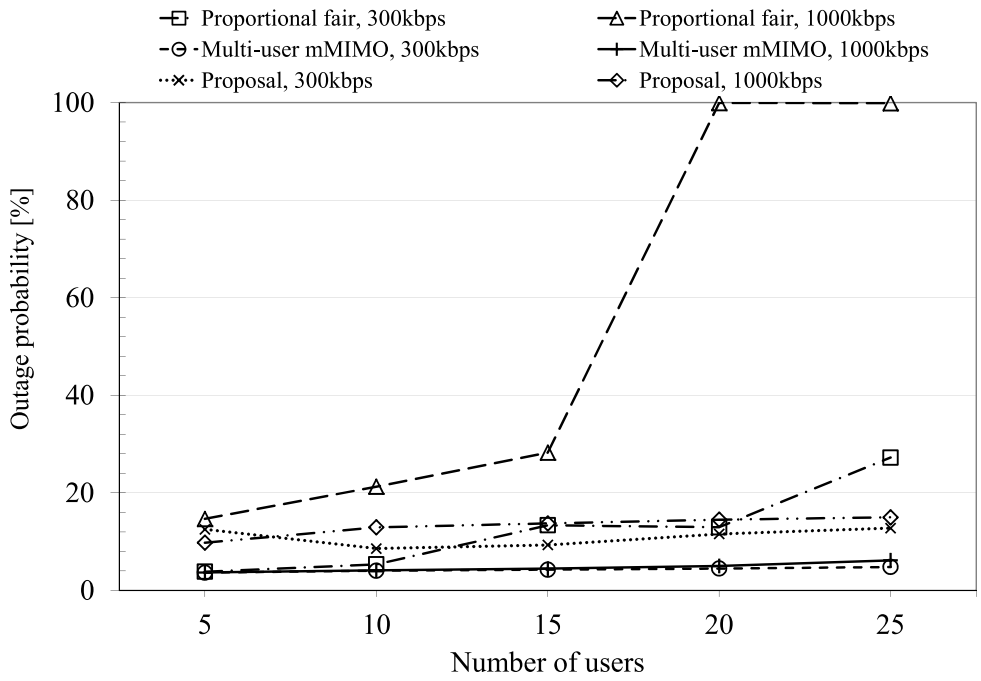


Figure 2.24: Outage probability (120 km/h user speed).
The proposed method can provide a good level of reliability even when the number of users grows.

3 Predictor antennas for high-mobility Internet access

The 5G technology aims at providing broadband-level connectivity to everyone, in every place and in every condition. With this in mind, this chapter deals with a specific scenario where one root cause results in a number of diverse issues, which make the aforementioned target challenging: high-speed trains. As already anticipated in Section 1.3.6, when moving at speeds of a few hundreds of km/h, existing mobile standards are pushed to their design limits and performance degrades substantially [19]. The main effects that come into action are as follows:

- Penetration loss: as train cars are usually made of metal, they introduce a significant attenuation of the radio signals, similarly to indoor users.
- Doppler spread: the combined effect of mobility and multipath propagation leads to spectral broadening of the signal. This is especially problematic for OFDM-based technologies, as the orthogonality between carriers is disrupted and self-interference is introduced as a consequence.
- Frequent handovers: very fast trains can perform handover from one cell to the next one every few seconds. This requires a service interruption time, which can be further increased compared to static users because the high speed degrades the signals. Those interruptions reduce the average bandwidth and introduce intermittent delays.
- Fast channel variations: as the trains move across the land, the radio channel experienced by the users changes very rapidly. Thus, when mobile users make an estimate of the radio channel and send it to the base station, it quickly becomes out-of-date. This is known as the channel aging problem, and it severely hinders MIMO techniques, which are used to enhance data rates in all modern systems but require accurate CSI to work properly.

There are methods that can be used to circumvent or alleviate these issues. For the penetration loss, array of antennas can be placed on top of the train to avoid it, and the train itself would act as a mobile station and relay the traffic to the users via

Wi-Fi hotspots [3]. Along with solving the penetration loss problem, this configuration also allows for more advanced and effective techniques to be implemented on the train, without requiring additional complexity for the end users. The effect of Doppler spread can be reduced by employing waveform or modulation techniques that are more resistant to it, such as the one presented in [78], and handover functionalities can be lightened by avoiding random access procedures, as discussed in [79].

Channel aging is a more complex issue. It exists in almost any condition, but it usually tolerated as most users are static and their performance is not impacted too much. Therefore, few approaches exist to directly face it. A promising one is the Separate Receive and Transmit Antennas (SRTA) techniques, presented in [19]. Section 3.1 describes its details, while Section 3.2 reports a system-level numerical evaluation of its performance, which has been performed in the context of the FANTASTIC-5G project.

3.1 Description of SRTA technique

The family of SRTA techniques tackles the channel aging problem at its roots, by exploiting multiple receiving antennas and the predictable movement of vehicles. The reference scenario considers a moving vehicle that receives a downlink transmission, moving almost in a straight line (which is always true at high speeds). At the base station, a Massive MIMO array is employed to create a beamforming pattern focused on the receiver antenna, based on the latest received CSI. If the environment remains mostly unaltered, then the transmitted signal propagates through it and creates a multipath fading profile, which remains constant over a relatively long observation period [19]. As the vehicle moves along, it sees different points of the multipath profile at different times, thus experiencing a rapidly time-varying radio channel.

The SRTA scheme leverages the concept of Predictor Antenna (PA) for

exploiting the relative time-invariance of the fading profile [80]. All the receiving antennas are placed on the rooftop, regularly spaced along the direction of movement, and the first one of them is designed as the PA. At any given time, it receives pilot signals from the base station and estimates the channel to create a CSI report. Such estimate will quickly be outdated for the PA, but it will later become valid for the other antennas when they occupy the same position where the PA was. In principle, it can be used for beamforming to the other antennas. But for this to work, the TTI should be changed adaptively as a function of the speed, the wavelength, and the antenna spacing, so that the vehicle can receive the transmission with the right antenna at exactly the right time. However, current LTE cellular technology only allows a TTI granularity of 1 ms, and even in 5G it will likely be possible to only use multiples and sub-multiples of that. Without a perfect adjustment of the TTI, there would be a residual mis-pointing and the full potential of the PA can't be realized.

To alleviate this problem, it is possible to store multiple samples of the channel (possibly with multiple PAs), thus using polynomial interpolation to estimate the intermediate positions, without requiring any change to the TTI length. The resulting technique is named SRTA with Polynomial Interpolation (SRTA-PI) [19]. The added complexity is proportional to $N\log(N)$, where N is the number of antennas on the train.

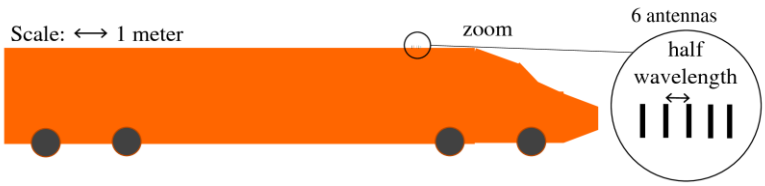


Figure 3.1: Example of the antenna configuration for SRTA-PI.

Figure 3.1 and Figure 3.2 illustrate the application of SRTA-PI to a high-speed train that supports spatial multiplexing of two data streams for speeds up to 500 km/h at 2 GHz, thanks to 6 antennas spaced by half-wavelength, and assuming the latency of an LTE FDD system.

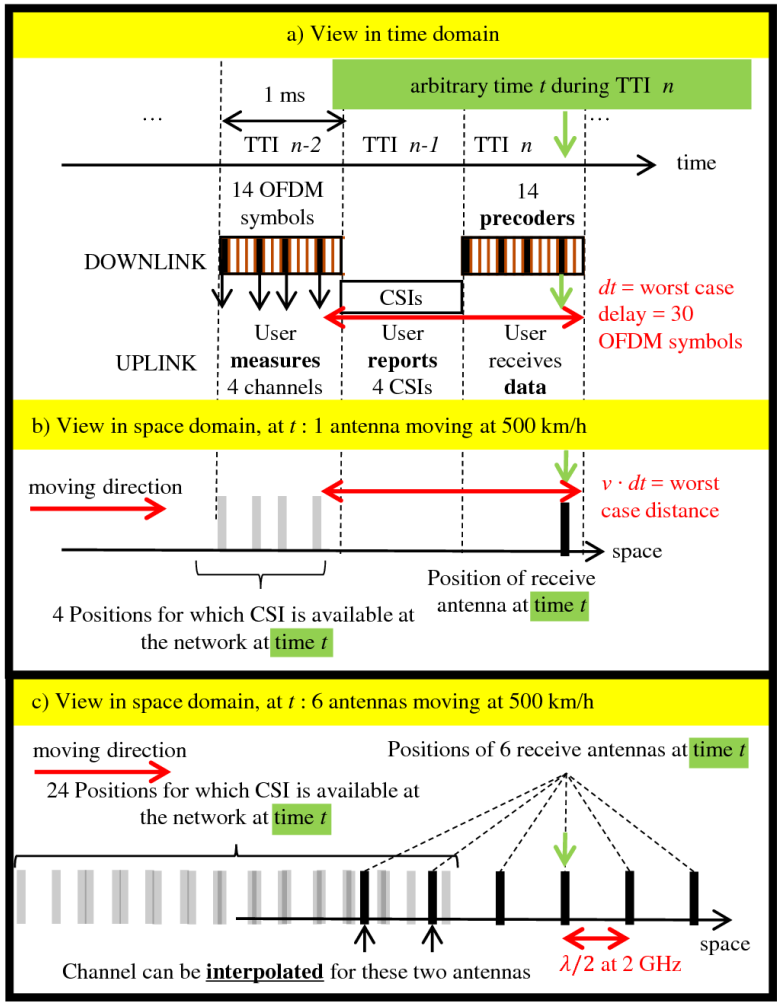


Figure 3.2: Illustration of the SRTA-PI approach.

Figure 3.2-a) illustrates the delays between channel measurement and precoding. The downlink and uplink transmissions are synchronized and using TTIs of 1 ms. During the n-th TTI, 14 OFDM symbols are sent, with 14 different precoders, to track the channel variations within the TTI. During the $(n-2)$-th TTI, the train measures the channel at 4 different times (for which sub-carriers with non-precoded pilots are available) and stores the four corresponding CSIs. Then, assuming the lowest possible latency, the train reports the four CSIs to the network during the $(n-1)$-th TTI. Finally, during the n-th TTI, the network uses the

reported CSIs to build the precoders.

Figure 3.2-b) shows that, when the train has a single antenna, the CSIs cannot help building efficient precoders, as the measured positions are several wavelengths away from the target antenna's position.

Figure 3.2-c) shows that for a train with 6 antennas, $6 \times 4 = 24$ different positions are measured, and that each of the 2 antennas at the back is "surrounded" by measured positions. Using polynomial interpolation over the 24 measurements, the channels of these two antennas can be accurately determined and used for precoding, even though these samples are irregularly spaced and form clusters of very close positions [19].

More generally, N is derived as follows. First of all, closer antennas provide a better spatial sampling of the channel. However, to avoid coupling between antennas, they are spaced by $\lambda/2$ where λ is the wavelength. Then, one must make sure that one antenna is always "surrounded" in the space domain by earlier measurements, even in the worst case in terms of delay dt between channel measurement and data transmission. During dt the train moves by $v \cdot dt$. Hence, the distance between the last antenna and the front antenna, given by $(N-1)\frac{\lambda}{2}$, must verify:

$$v \cdot dt \leq (N-1)\frac{\lambda}{2} \tag{32}$$

If the system needs to send P data streams to P receive antennas instead of 1, then the condition in Eq. (32) becomes:

$$v \cdot dt \leq (N-P)\frac{\lambda}{2} \tag{33}$$

which is equivalent to:

$$2N \geq v \cdot dt + P\lambda \tag{34}$$

With LTE at 2 GHz, $dt = \frac{30}{14}1[\text{ms}] = 2.1429$ ms (as illustrated in Figure 3.2), λ = 15 cm, $N \geq 5.9683$, and $N = 6$.

3.2 System-level simulation and results

This work evaluates the effectiveness of SRTA-PI in scenarios involving a very fast train (up to a speed of 500 km/h) using computer simulations, carried out through an extended version of LTE-Sim [59].

Three different configurations are considered. The first one is the baseline scenario, based on the simple OFDM-based transmission scheme. It has no means to combat the effects of high speed and suffers from both channel aging and Doppler spread. The second one uses SRTA-PI. It removes the channel aging issue, but Doppler spread is still present. Finally, the last scenario models an ideal communication where neither channel aging or Doppler spread are considered. It is used as an ideal solution reaching upper bound performance.

The base stations are configured with Massive MIMO arrays with 256 antennas and configured according to the two-stage JSDM precoding scheme [43]. The handover time is assumed to be negligible by using a solution such as [79]. The 3GPP 3D channel model for suburban environments [60] is taken into account. The bandwidth is set to 100 MHz, centered at 2 GHz. Each base station uses a total transmit power of 53 dBm, and a round-robin scheduler that can serve up to 8 users simultaneously. From the network perspective, the ISD is set to 0.5, 1, and 2 km. During simulations, the train moves on a straight line, with three-sectored base stations placed at both sides of the track in a hexagonal grid, traveling for 1400 m. There are two rows of base stations at each side, where the first one is used for service and the second one for modeling interference from the rest of the network. The speed is set to 30, 120, 250, and 500 km/h, and the typical acceleration of trains is slow enough that the speed can be treated as constant for the purposes of this work. On top of the train there are a number of receiving units, which act as independent user equipments from the point of view of the network. Each one of them has 6 antennas and implements the SRTA-PI technique as described in Section 3.1. Each receiving unit can receive up to two spatially multiplexed streams, and

the received packets are relayed to the users using on-board Wi-Fi hotspots, thus avoiding interference issues with the cellular connection. Results have been averaged over 30 independent simulation runs.

Figure 3.3, Figure 3.4, and Figure 3.5 show the total throughput achieved with 2, 4, or 8 receiving units on the train, respectively. By increasing the number of receiving units, the throughput improves considerably because more data streams can be sent in parallel, but the increment is somewhat less than linear. This is because at least some of the receiving units are co-scheduled on the same base station, and thus they share the same total transmit power.

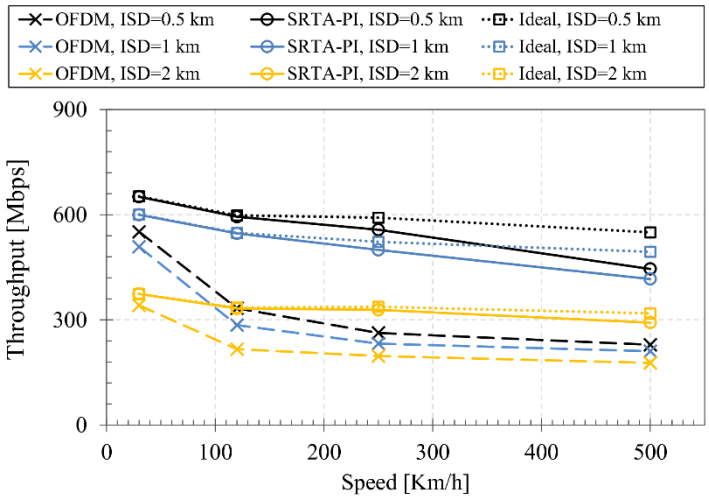


Figure 3.3: Total throughput with 2 receiving units on the train.

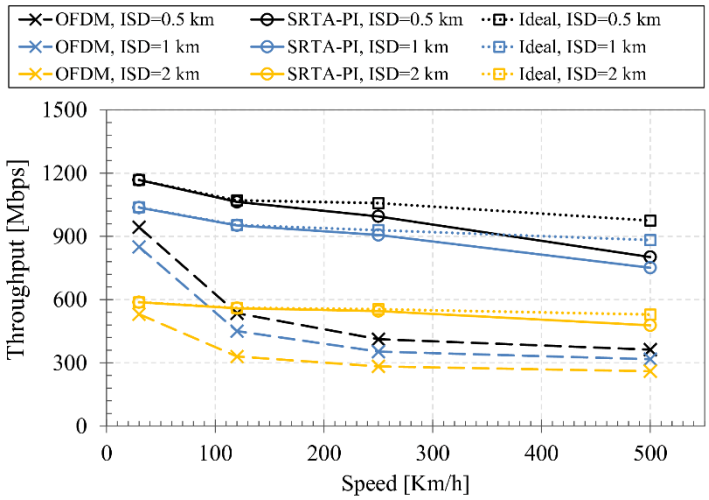


Figure 3.4: Total throughput with 4 receiving units on the train.

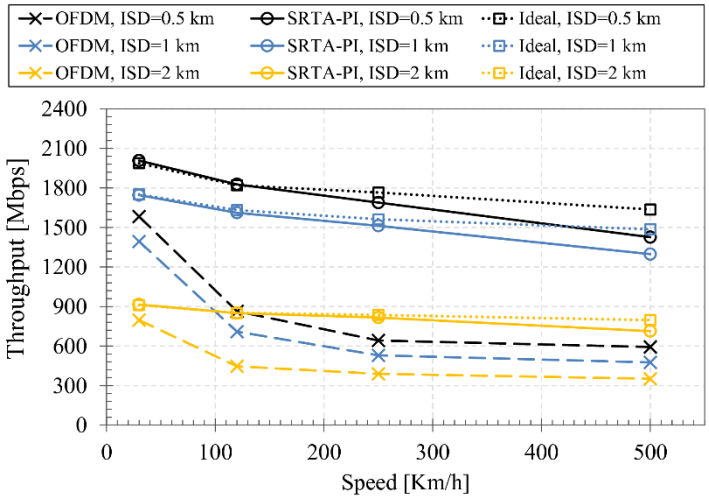


Figure 3.5: Total throughput with 8 receiving units on the train.

In all the cases, the performance of the baseline technology (labeled as OFDM) is the most sensitive to the speed, as its throughput is reduced by around 50% to 65% when the speed increases from 30 km/h to 500 km/h. Instead, SRTA-PI performs much better, as its throughput is already higher by 10-25% at the lowest speed, and the loss at 500 km/h is limited to about 20-30%. This sums up to a throughput gain of 100% or more at the highest speed, in most scenarios. However, there is still a throughput loss of 8-20% compared to the ideal case.

With respect to the ISD, the first increment from 0.5 km to 1 km results in a limited throughput loss, ranging from 6 to 19% across all technologies and speeds. Instead, the loss from 1 km to 2 km is more severe, as it varies from 15 to 47%.

Figure 3.6 shows the distribution of the MCS index selected for transmission, for each speed and each technology. Higher MCS values directly relate to higher SINR, and result in higher spectral efficiency. At 30 km/h, all the rows are quite similar, with most of the values located between 7 and 13. However, as the speed increases, the distribution of OFDM is progressively shifted towards lower values, and MCS indexes above 13 are never used at the higher speeds. On the contrary, for SRTA-PI and the ideal case the most frequent values remain the same, and only

some of the higher MCS indexes are avoided at higher speeds. This confirms that removing the transmission impairments caused by the high speed has a strong effect on the perceived channel quality, and thus on the final throughput.

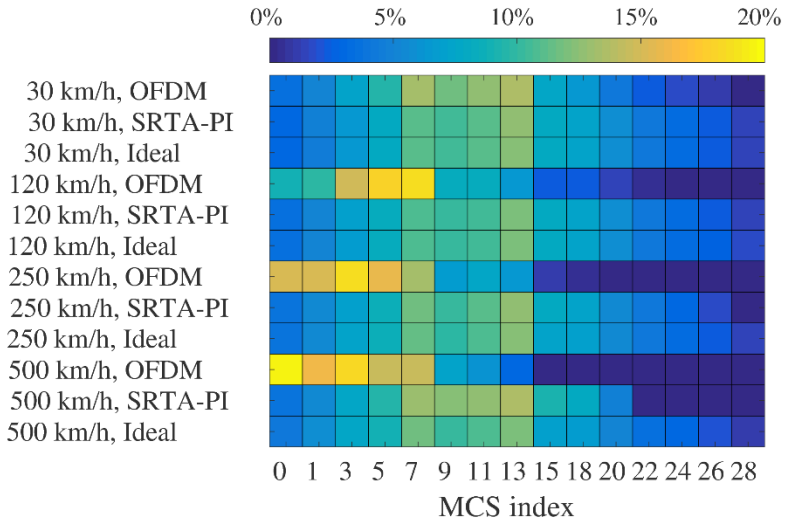


Figure 3.6: Distribution of the MCS index for different speeds and technologies.

4 Enhanced Random Access for massive MTC communications

As hinted in Section 1.3.3, a large number of IoT devices will rely on the cellular infrastructure to attach themselves to the Internet at large. As they typically engage in machine-to-machine interactions, or Machine Type Communication (MTC), they are referred to as MTC devices. This poses some interesting challenges to existing mobile networks as they are now. In fact, they are designed for a relatively small number of high-throughput connections, and that is reflected in different parts of the specifications. On the other hand, MTC devices require very low throughput, but deployments are expected to reach huge numbers and densities, so much that the active IoT devices at any given time could be way more than human users [16].

One of the critical points for this topic is the Random Access Channel (RACH) procedure, that is used by a mobile device to establish a connection when there is none in place. For the very first step, it uses a common channel for all the users with an ALOHA-like access, thus there is a possibility for collisions. The dimensioning is done to ensure a low collision probability with the expected amount of connection attempts. However, MTC devices are expected to come in large numbers and thus put a lot of strain on the RACH, ranging from a higher number of collisions to a total congestion [81].

For this reason, Section 4.1 describes an extension of the LTE RACH procedure to provide greater capacity, and Section 4.2 presents a numerical evaluation to prove its effectiveness.

4.1 Basic and extended RACH procedure

The standardized RACH procedure for LTE is based on a four-message handshake initiated by the mobile terminal (see Figure 4.1) [82]. It allows the mobile terminal to achieve tight synchronization with the base station and to receive an allocation of uplink resources.

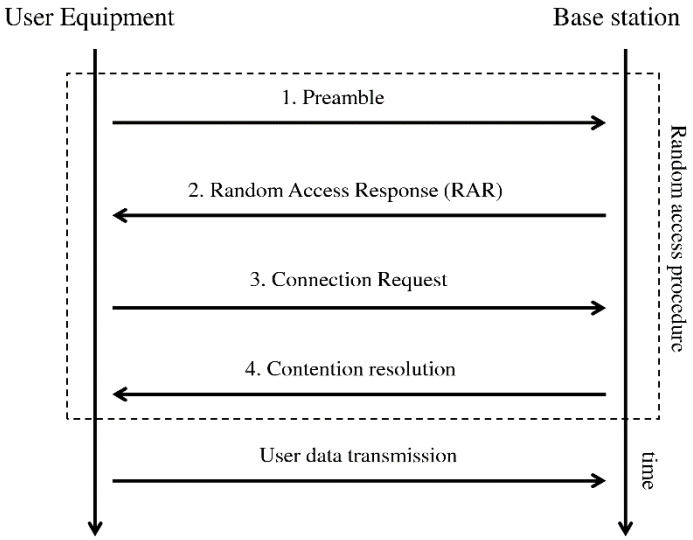


Figure 4.1: Baseline RACH procedure for LTE.

The first message can only be sent during RACH opportunities, periodically scheduled by the base station. It consists of a preamble sequence, randomly chosen from a set of 64 orthogonal sequences. The purpose of the preamble message is to indicate the presence of an access request, and to allow the base station to estimate the distance of the mobile terminal for the Timing Advance procedure (message 1 in Figure 4.1). If two or more devices pick the same preamble during the same RACH opportunity, there is a collision and the procedure will fail immediately or at a later stage. More details on RACH opportunities and preamble sequences are given in [62].

After the preambles are detected, the base station sends back a Random Access Response (RAR) message (message 2 in Figure 4.1). It contains a set of relevant information for each detected preamble, the most important being the allocation of un uplink resource for sending the third message.

If a collision is detected for a specific preamble, then the corresponding information is not sent in the RAR and the devices retry the procedure after a waiting time. On the other hand, if a collision goes unnoticed, then two or more mobile terminals will be assigned the same uplink resource and they will collide again on the third message.

After receiving a resource allocation through the RAR, the mobile terminal can send the Connection Request (message 3 in Figure 4.1). If there was an undetected preamble collision, two or more devices will send this message over the same resource, i.e. they will collide again, and their messages will be lost.

Finally, after a successful delivery of the Connection Request, the base station replies with the last message, that is the Contention Resolution (message 4 in Figure 4.1). A device that receives a Contention Resolution addressed to him assumes that the RACH procedure is completed. Therefore, it can now have a reliable, collision-free communication with the base station. On the other end, if the Contention Resolution is not received at the proper time, the RACH procedure has to be restarted.

In a typical configuration, 54 preambles are dedicated to the contention-based access, while the other 10 are reserved for contention-free access, and the Physical Random Access Channel (PRACH) is scheduled every 5 ms. This gives a theoretical capacity of 10800 preambles per second, which could be sufficient for most MTC scenarios. But, the real capacity is much lower because of the collisions occurring at moderate and high loads, especially in scenarios enabling event-triggered reports.

The procedure described here was extended in [83] through a simple modification. That is, after the base station performs the detection of preambles, it sends back a RAR containing multiple responses for each identified preamble, with different uplink resources assigned. Every mobile terminal which receives the RAR can randomly choose one of the uplink resources reserved for the preamble, selected during the first step. Figure 4.2 shows the modified message sequence chart.

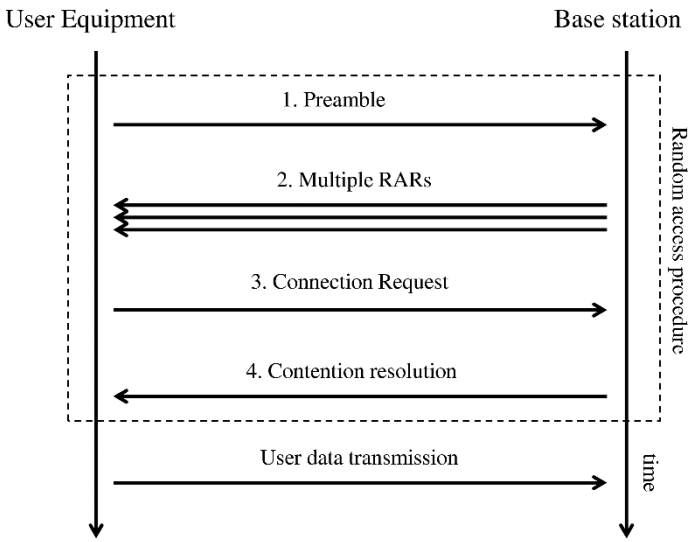


Figure 4.2: Enhanced RACH procedure.

Thanks to this additional randomness, a collision over the selection of the preamble does not necessarily translate to a failure in the access procedure. This way, if two or more mobile terminals use the same preamble, they still have a chance to select a different resource assignment in the RAR and thus avoid the collision at the third step of the protocol. Basically, the multiple RAR responses act as multipliers for the number of preambles.

The downside of this technique is that a correspondingly larger amount of uplink resources must be reserved for the transmission of Connection Request messages, which shrinks the resources available for actual user data.

4.2 Evaluation of the extended RACH procedure

The baseline scheme and the proposed candidate for 5G were implemented in LTE-Sim [59] and extensively tested under high-load conditions via numerical simulations.

The simulated environment consists of multiple cells, managed by base stations having an omnidirectional antenna. Since the RACH protocol works independently in each cell, the results are collected from a specific cell only: the same results can be generated by also considering large-scale simulation environments.

Conducted tests assume a simple single-input single-output communication

scheme, in accordance with the low-cost requirements of MTC devices. The device density varies from 10000 to 1000000 devices/km^2. The inter-site distance is set to 500 m, thus the resulting radius of the cell is equal to 290 m. Users were positioned with a uniform random distribution over the simulation area. The allocated bandwidth is 10 MHz, with a center frequency of 2 GHz.

To reproduce an event-driven transmission burst, the activation time of the devices in the simulations follows a beta distribution with parameters (3,4) over a time interval of 10 seconds [84]. The application payload has a small size (5 bytes), so that it only requires a single LTE RB. Also, it should be delivered within 10 seconds from the generation instant, otherwise it is dropped. Each test was repeated for 50 times with a different seed for the random quantities.

According to the most common configuration in LTE networks, RACH opportunities occur every 5 ms and 54 different preambles are available, while the remaining 10 are reserved for contention-free access [82]. In case of collision, the procedure fails for all the involved devices, and can be repeated for a maximum of 3 times. For the candidate 5G approach, the number of RARs transmitted for each preamble is set to 2 and 4.

We assumed the same modulation and coding as LTE, so as to isolate the effect of the proposed extension of the RACH scheme. This choice ensures that any gain observed is actually due to the new extended procedure rather than difference in the modulation and coding. Moreover, it is expected that 5G will be deployed through incremental steps, and the application services will be quickly deployed while re-using the existing hardware. For this reason, different RACH procedures have been compared, which are implemented on top of the current LTE technology.

The propagation loss is modeled with the urban macro-cell model [85] $L_{db} = 128.1 + 36.7log10(d) + S$, where d is the distance in km and S the large-scale shadowing, with 0 mean and 8 dB standard deviation. The power level employed at the mobile terminals is 23 dB. As frequency and time selectivity are not relevant in this scenario, fast fading was not used. The performance of investigated approaches

was evaluated in terms of success probability of the transmission, average delay, and collision rate of the random access process.

First of all, Figure 4.3 shows the probability that each device completes the transmission of its small data packet within the required delay. It is always equal to 100%, except with the highest density values. At the maximum density of 1000000 devices/km^2, LTE-A performs poorly (6.6%), while the proposed solution shows a clear advantage (21.6% and 31.5% with 2 or 4 RAR responses, respectively). There are two possible causes of failure: (1) a device may not be detected by the base station because of recurring collisions, and (2) transmission can be exceedingly delayed because of the limited amount of physical uplink shared channel (PUSCH) resources. Both of them occur in this evaluation, depending of which technique is adopted, as will be further discussed below.

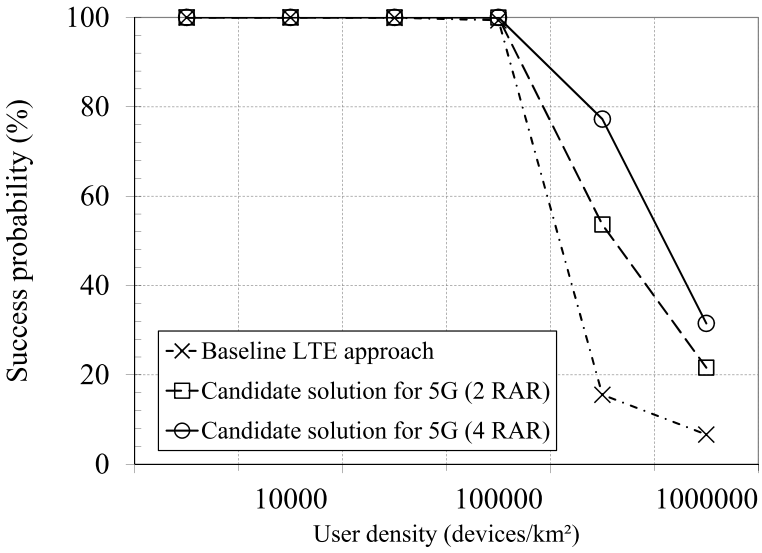


Figure 4.3: Success probability of the complete transmission.

Figure 4.4 shows the average delay of the transmitted packets. This metric refers to application layer data. Therefore, it is calculated by considering both the latency due to the random access procedure and the queuing time at the packet scheduler. Moreover, the delay is only measured when the device completes the RACH procedure and it is scheduled for the transmission of the application packet before it expires. For all the investigated techniques, it increases almost linearly with the number of users, and it progressively lowers when upgrading from LTE-A to the

proposed technique with 2 RARs and finally to 4 RARs. The only exception is with 4 RARs and the highest device density, where the proposed technique grows faster and exceeds the other two cases. This suggests that the PUSCH resources are being saturated for most of the time, due to both the high number of devices handled and the overhead of sending 4 RAR messages.

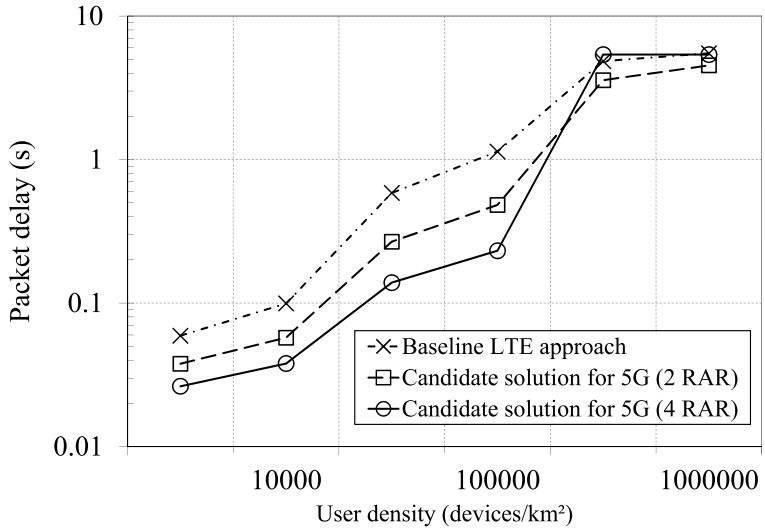


Figure 4.4: Average delay from initial request to successful data transmission.

Finally, Figure 4.5 shows a temporal trace of the collision rate, measured in the highest-loaded scenario with 1000000 devices/km^2. The most significant result is the very high collision rate of LTE from 3 to 13 seconds, which happens during the peak of the arrivals and for some more seconds. During this period, almost no devices can be detected, and they retry connection until they succeed or fail. Therefore, they produce a longer tail than the other curves. Under these conditions LTE-A fails to handle 93.4% of the devices due to collisions, as already shown in Figure 4.3. By contrast, the candidate solution for 5G with 4 RARs experiences only 63% collision rate at most, meaning that eventually all users should be detected thanks to retries and back-off. In this case, the 78.4% of lost transmission attempts is due to PUSCH saturation. The candidate solution for 5G with 2 RARs lies

somewhere in-between, as it experiences a peak of 97% collision rate where part of the users is left out, but recovers quicker than LTE-A. The collision rate for lower user densities is not shown for space reasons, but it is not a concern as it reaches a maximum value of 5%.

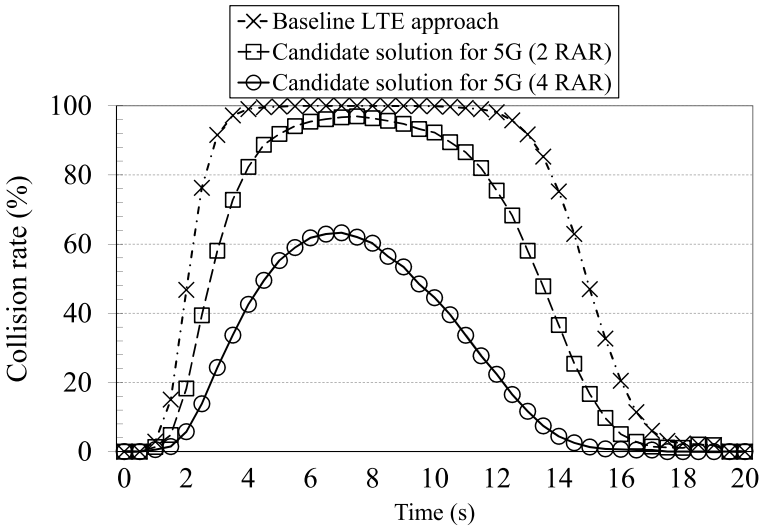


Figure 4.5: Profile of the collision rate occurring during simulations.

5 Advancements for multicast and broadcast

As stated in Section 1.3.7, there are situations where a multicast or broadcast-type transmission is more appropriate than unicast in the context of cellular networks. Whenever there are multiple users interested in the same content at the same time, there is a potential for bandwidth saving: the content is only transmitted once, but it can be received by any interested user.

This base concept was already explored in LTE with the Multicast-Broadcast Single Frequency Network (MBSFN) technique, introduced in Release-9. On the one hand, it takes advantage of the underlying OFDM modulation to work with multiple cells simultaneously without much added complexity. On the other hand, since there are multiple users, the usable MCS levels are limited to consider those users with the lowest channel quality [86].

One important limitation of multicasting and broadcasting in existing mobile networks is that they are strictly downlink-only: the data stream goes from the base station to the users, but there is no reverse channel from the users to the base station that is associated with the multicast/broadcast flow. Therefore, one of the main innovations that is considered for broadcasting in 5G is the inclusion of a return channel. As will be explained later in Section 5.1, that enables different kinds of improvements with respect to the existing state of the art.

Depending on the service area, multicast and broadcast services can be classified as *local*, *regional* or *national*. *Local* services are related to a limited area with a radius of 1 to 20 km. Typical scenarios include local information such as stadium services, advertisements, voucher delivery, outdoor festivals, or congress/convention service. For *regional* services, coverage area could be up to 100 km, to deliver information related to traffic jams or disaster warnings (e.g.

tsunami, earthquakes). Crucially, the feedback channel, if present, can be used to ensure and confirm correct reception of the message from all the recipients. Finally, *national* services could cover the geographical area of one or more nations. They could complement traditional broadcast services (e.g. Radio and TV [87]), and vertical industries can adopt it for firmware distribution and upgrade. Feedback capability can also be used to confirm reception and successful installation, avoid or limiting the need for further intervention.

In this study, the MBSFN technology available in LTE was considered as a baseline. Then, it was extended with an adaptive MCS selection procedure and a Hybrid ARQ (HARQ) technique tailored for the broadcast case. These techniques are described in Section 5.1, and simulation results produced in different scenarios for the FANTASTIC-5G project are reported in Section 5.2.

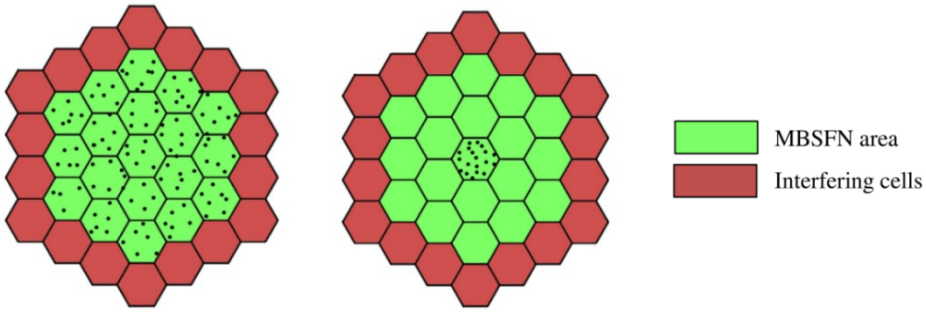
5.1 Baseline and enhanced broadcasting techniques

Multicast and broadcast communication is available in LTE since Release 9, thanks to MBSFN operation. Given an LTE radio frame, which is composed by 10 sub-frames, some of these can be configured as MBSFN sub-frames. Multicast transmission can take place during MBSFN sub-frames, while standard unicast activity takes place during non-MBSFN sub-frames.

It is not possible to use all the sub-frames for MBSFN operation. Some sub-frames in specific positions in the radio frame contain important system information (i.e. sub-frames 0/4/5/9 for FDD mode and 0/1/2/5/6 for TDD mode) and therefore cannot be dedicated to broadcasting [88].

As implied by the name, MBSFN allows the creation of single-frequency-network areas, where all of the involved nodes transmit the same signal at the same time and on the same frequency band. The receiving terminals do not separate the signals coming from each transmitting point; instead they see a single transmission over a time-dispersive channel. The time dispersion effect is dealt with thanks to the cyclic prefix which is usually part of OFDM-based communication [89].

A propagation model specific for the single-frequency-network configuration is used, where the power coming from neighboring cooperating base stations is counted as useful signal rather than interference. In this regard, two deployment types are considered, as shown in Figure 5.1: in type (a), mobile users are simply placed everywhere in the MBSFN area, while in type (b) the users are only created in the center cells. The remaining cells are treated as "assisting rings" and their purpose is to increase the average channel quality by reducing border effects.



(a) without assisting rings *(b) with assisting rings*

Figure 5.1: layout of simulation area.

However, downlink-only broadcast transmissions cannot benefit from link adaptation because channel state information of receivers is not known to the transmitter. Assuming a channel state reporting functionality similar to the unicast case (i.e. CQI feedbacks), a dynamic selection of the MCS has been added to MBSFN. The selection is based on the aggregate CQI data collected during MBSFN sub-frames, which is incompatible with unicast CQI reports. The MCS is then chosen on a worst-case basis to satisfy the BLER target for all the users, or possibly a sub-group of them. As an example, if 95% is used as the threshold, then the 5% of the users with the worst channel quality are excluded and the MCS is tuned to the lowest CQI among the remaining users. This restriction tries to avoid the possibility that few users with bad channel condition can negatively influence the

transmission quality for all the other users.

Finally, HARQ retransmissions are considered because, at the lower layers of the protocol stack (PHY/MAC), broadcast transmission in LTE is unreliable: when a packet is not received correctly, no retransmission is triggered, because there is no return channel to send a positive or negative acknowledgment.

When such return channel is introduced, a hybrid broadcast/unicast scheme can be realized. All data is first transmitted on the broadcast channel, then the base station waits for ACKs or NACKs from all the users registered to the data stream. When NACKs are received, the corresponding data is queued to dedicated unicast channels for each of the users [90].

In this context, using unicast retransmissions seem preferable over sending the negative-acknowledged packets over the multicast channel again. Unicast channels can make use of multi-antenna techniques to improve the channel quality, such as beamforming and MIMO, which are not applicable to multiple users simultaneously on a broadcast channel. Moreover, unicast retransmissions are localized to the cell where the user is located, rather than the entire broadcasting area, therefore saving resources if occurring sporadically.

5.2 Simulation and results

Simulations were carried out with the LTE-Sim simulator, which has been extended with support for MBSFN and with the additional enhancements as described in Section 5.1. The simulation environment consists of 37 cell sites in a hexagonal deployment, with an ISD of 600 m. The 19 innermost cells are part of the MBSFN area and the surrounding ring only contains interfering cells. The bandwidth is 20 MHz, with a center frequency of 2 GHz, and 6 sub-frames are dedicated to MBSFN for every radio frame of 10 sub-frames.

There are two configurations with respect to assisting rings: when the assisting rings are employed, the users are placed only in the central cell, surrounded by two rings of MBSFN cells without any users. Instead, when the assisting rings are not used, users are placed everywhere in the 19 innermost cells. In both cases, their density is 400 users/km^2 and most users are static (65%), while some of them have

moderate speed (30% at 0-120 km/h) and few have high speeds (5% at 0-500 km/h). At the application level, a video stream is transmitted to all the connected users, with a bit rate of 17 Mbps.

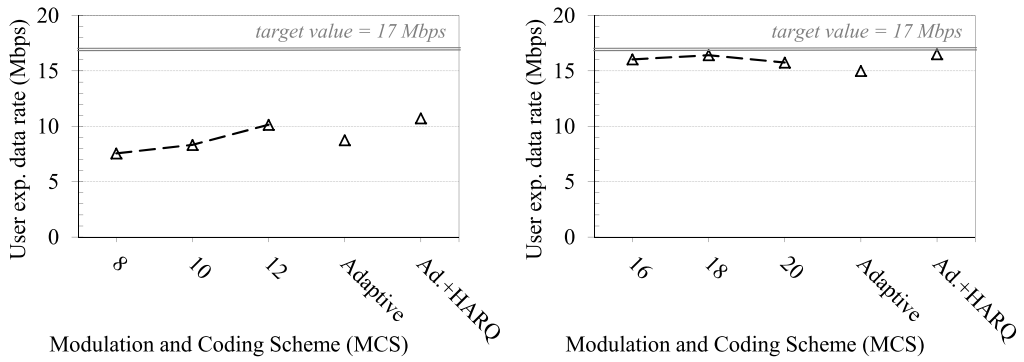
When standard MBSFN is used, without any extension, the MCS is set to a fixed value for the entire simulation. Few different values are evaluated, depending on the configuration, so that they are similar to those chosen by the adaptive selection. Specifically, without assisting rings, the MCS is set to 8, 10, or 12, and with assisting rings, it is set to 16, 18, or 20. When adaptive selection is used, the target CQI value is based on the lowest CQI among the 95% of users with the best channel quality, so that at least 95% coverage is achieved. Finally, when HARQ retransmissions are also employed, every lost packet can be retransmitted for 4 times maximum.

5.2.1 Local broadcast services

The simulation results presented here refer to relatively small MBSFN areas, which relates to the category of "local" broadcast services. This is the most difficult situation, because the extension of the border area (which is subject to strong interference) is non-negligible compared to the total area. Indeed, introducing the assisting rings improves the performance and comes quite close to the targets. The downside is that the spectral efficiency is reduced because of the unused cells.

Figure 5.2 reports the user experienced data rate as a function of the MCS index, with and without assisting rings. This includes the baseline approach (that is, some fixed values of MCS) and the proposed enhancements (i.e., adaptive MCS selection without HARQ and adaptive MCS selection with HARQ). Most results with assisting rings are very close to the target, while results without assisting rings are considerably lower because users at the edge of the area receive a lot of interference. Using the adaptive MCS approach without HARQ seems to make a conservative choice which reduces the throughput, but when coupled with HARQ, a higher

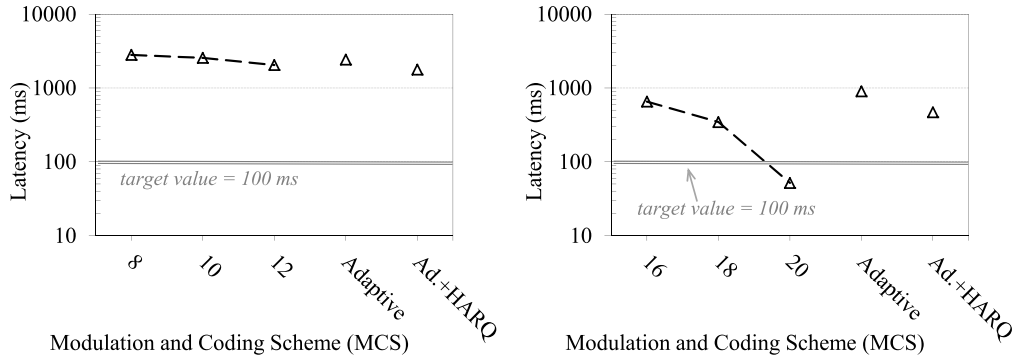
average MCS can be used and there is a notable advantage.



(a) without assisting rings. *(b) with assisting rings.*

Figure 5.2: Results for user experienced data rate.

Figure 5.3 reports the KPI 2 with the same parameters as the previous figure. In this case, the minimum value is achieved with assisting rings, MCS 20 and 6 sub-frames, and it is the only value below the target. Also in this case, adaptive MCS is not very good on its own but it is improved by HARQ. However, the adaptive solutions cannot reach the minimum value because the MCS required to support 95% of the users is lower than 20, thus packets remain queued for longer times. The reason is that MCS 20 does not offer sufficient coverage, despite exhibiting the lowest latency.

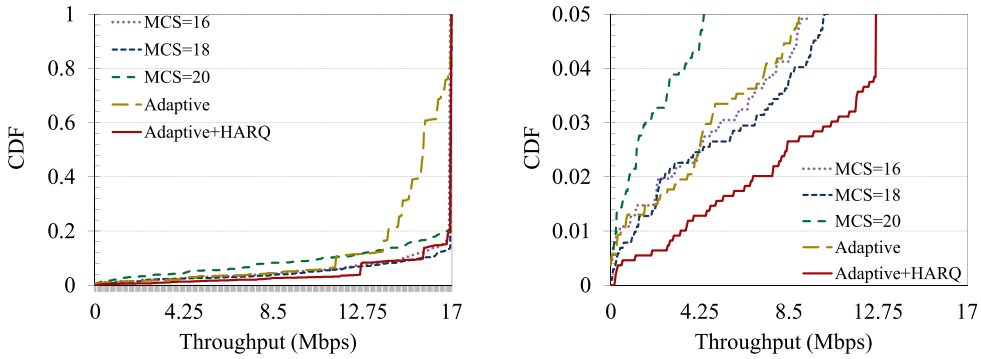


(a) without assisting rings. *(b) with assisting rings.*

Figure 5.3: Results for latency.

This is confirmed by Figure 5.4, which depicts the CDF of the average user throughput. Considering the 5% of users with the lowest throughput, it is easily seen

that most configurations perform similarly in this area. The only exceptions are MCS 20, which is inferior to the other configurations, and the adaptive solution with HARQ, which provides the best result for cell edge users and increases the 5 %-ile throughput from 9-10 Mbps to 12.6 Mbps. Figure 5.5 shows the number of additional RBs required for HARQ. Overall, the resource overhead due to the introduction of HARQ is lower than the reservation of an additional sub-frame.



(a) without assisting rings. *(b) with assisting rings.*

Figure 5.4: CDF of average user throughput.

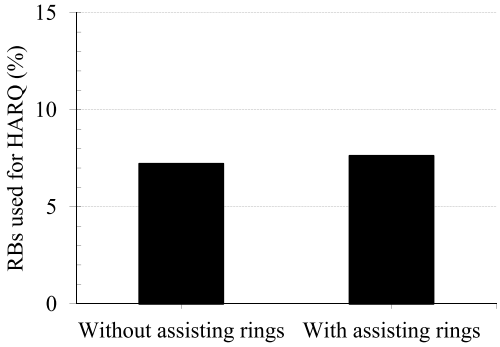
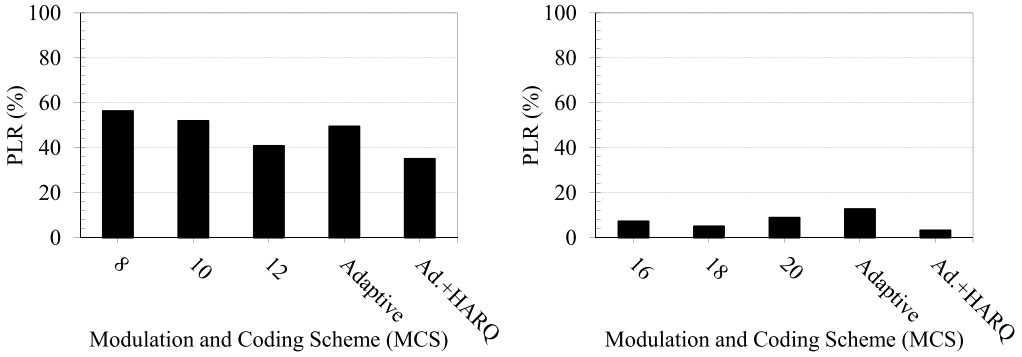


Figure 5.5: Number of additional RBs used for HARQ.

Figure 5.6 shows the PLR obtained with the parameters described above. Packet loss is caused either by data received with errors (if not corrected by HARQ) or data that is still queued at the end of the simulation. Two important highlights emerge: the use of assisting rings always reduces the PLR, and so does the introduction of

HARQ over adaptive MCS alone. Indeed, the combination of adaptive MCS and HARQ (with assisting rings) achieves the lowest PLR value in this simulation study.



(a) without assisting rings. *(b) with assisting rings.*

Figure 5.6: Results for Packet Loss Ratio.

Please note that the evaluated scenario with 19 cells and only 1 active cell is the most pessimistic case, with regard to both interference and spectral efficiency. Deployments with more active cells surrounded by two assisting rings would have lower interference (because interfering cells are farther on average) and higher spectral efficiency (because the ratio of active cells over total cells would be larger). Therefore, this scenario should be considered as a worst case, which is unlikely to be encountered in practice.

5.2.2 Regional and national broadcast services

The evaluated scenario with 1 active cell and 2 assisting rings can also serve as a performance lower bound for larger MBSFN areas. In fact, considering the users of any internal cell, they will see a number of closer surrounding cells which provides a totally constructive signal, while further cells become progressively less constructive and more interfering. This situation is shown in Figure 5.7.

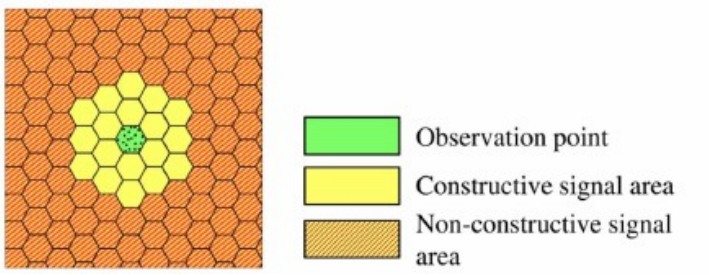


Figure 5.7: breakout of large MBSFN area.

The exact size of the completely constructive area depends on the cyclic prefix and the ISD, and in most cases should be larger than the 2 rings shown in the picture (with the extended cyclic prefix it has a radius of around 5 km). Also, the first level of cells beyond the constructive area (which creates the largest contribution to interference) is not completely interfering, but only partially so [91]. Finally, the overhead due to assisting rings at the border of the area is low when the area is large.

From these observations, it is possible to state that large MBSFN service areas are definitely less challenging than the small areas evaluated in this study, and the performance would at least equal to the case with assisting rings (in most cases they would be significantly better).

6 Conclusions and Future Research Directions

This Ph.D. work pursued the goal of identifying, modeling, and evaluating advanced transmission techniques for the upcoming 5G cellular systems. Since the main goal of 5G is to support a very wide array of use cases, different approaches are required for each specific situation, and this is clearly reflected in this work.

Many contributions are centered on the topic of Massive MIMO, that is explored in Chapter 2. Thanks to the use of very large and fully digitally controlled antenna arrays, this new technology holds promises of great improvements in capacity, spectral efficiency, and energy efficiency. However, this may be hindered by the conflict between the TDD mode assumed in its original formulation, and the FDD mode employed in most deployed networks. The JSDM two-stage precoding technique, complemented with the beam coordination technique for interference reduction, is explored as a way forward to adapt Massive MIMO to FDD mode, with encouraging results. Moreover, the problem of how to suitably detect/create clusters of users for JSDM is explored in detail, by extending the DBSCAN algorithm and providing a useful partitioning in a larger number of scenarios. For the uplink direction, a distributed resource allocation protocol is presented, with fully autonomous user operation, optimized energy efficiency, and respect of QoS constraints.

Chapter 3 describes the issues that arise with providing broadband Internet access on board of very fast trains. Among these, the channel aging effect emerges as an important obstacle, and the Predictor Antenna concept is studied and evaluated as an effective solution.

Chapter 4 deals with the problem of supporting massive deployments of IoT on cellular networks, especially with respect to the initial random access procedure. With the standard parameters considered in the current LTE technology, there is the possibility for collisions, congestion, and total blockage of such procedure. Therefore, a simple and effective extension is investigated and tested via

simulations to provide support for a larger number of devices.

Chapter 5 explains the opportunities offered by multicast and broadcast transmission modes in mobile networks. By improving over the MBSFN architecture already implemented in LTE, an additional return channel is employed to provide dynamic selection of the modulation and coding settings and selective retransmission of damaged packets, resulting in improved performance.

The techniques considered in this work have been demonstrated to be valuable and effective for their respective use cases. Therefore, these are plausible candidates for being included in the modular framework of the 5G air interface. As the 5G ecosystem is still in its infancy, there are plenty of opportunities for further research on it. With respect to the use cases covered here, there are still unsolved problems, and more techniques or combinations thereof require investigation. For example, the JSDM scheme with coordinated beamforming presented in Section 2.2 could be combined with a layer of micro-cells, to be activated only when the corresponding area is not served by the macro-cell, thus avoiding its strong interference. As for the user grouping algorithm of Section 2.3, an open question remains, that is how to create optimal subsets of the generated groups, for the purpose of scheduling them in different time slots. In very high-speed environments the SRTA-PI technique employed in Section 3 can reduce the effects of channel aging, but the interference due to Doppler spread still remains, so additional solutions can be helpful. Finally, there are also whole other scenarios that were not covered, and more will come up in the next years depending on the evolution of society and users' needs.

Appendix

The LTE-Sim simulator

LTE-Sim is a mobile network simulator, developed at the Polytechnic of Bari [59]. It is focused on LTE networks, and more features are being added to support 5G as well. A large part of the research works presented in this thesis were carried out by extending LTE-Sim with the required functionalities, and then using them to conduct simulation campaigns on the topic at hand. This appendix describes some of its main characteristics and recent developments.

LTE-Sim is a system-level simulator. This means that it can model the interactions among many different devices at the same time, including e.g. base stations, mobile users, gateways, and femtocells. In contrast, link-level simulators usually consider only two devices. A key difference is that link-level simulators go into great detail about low-level details of the communication links, such as channel coding and modulations, while these must be abstracted away with simplified models in system-level simulators, otherwise the total complexity would be overwhelming. This difference sets the scope of the two categories: a link-level simulator is useful to test techniques that involve a single communication link, while a system-level simulator deals with more complex interactions, such as scheduling and multi-user interference.

The simulator is written in the C++ language with an object-oriented structure, making use of classes, encapsulation, and inheritance. It is structured as an event-driven application: the *Calendar* class holds a list of events to be executed, with each item containing the required time of execution, the method to execute, the object on which the method should be called, and possibly some parameters. At the beginning of the program's flow, the description of the scenario creates most of the required objects and puts some events in the calendar. Then the *Simulator::Start()* method is called, and the calendar starts executing the registered events in chronological order. Each event can generate new events and put them in the calendar, resulting in a sustained supply of events to process until the end of the

simulation. While a simulation is running, the *FrameManager* class tracks the flow of time and increases the counters related to frames and sub-frames, and in some cases, it marks sub-frames dedicated to different functions (e.g. downlink versus uplink sub-frames in TDD mode).

LTE-Sim runs in a console window, where output is displayed. It consists of a long sequence of structured lines, composed of a number of fields. This simplifies subsequent processing of the generated data, which can be processed line-by-line by extracting the fields of interest. An example is shown in Figure A1. The amount and type of information generated can be customized to a great extent.

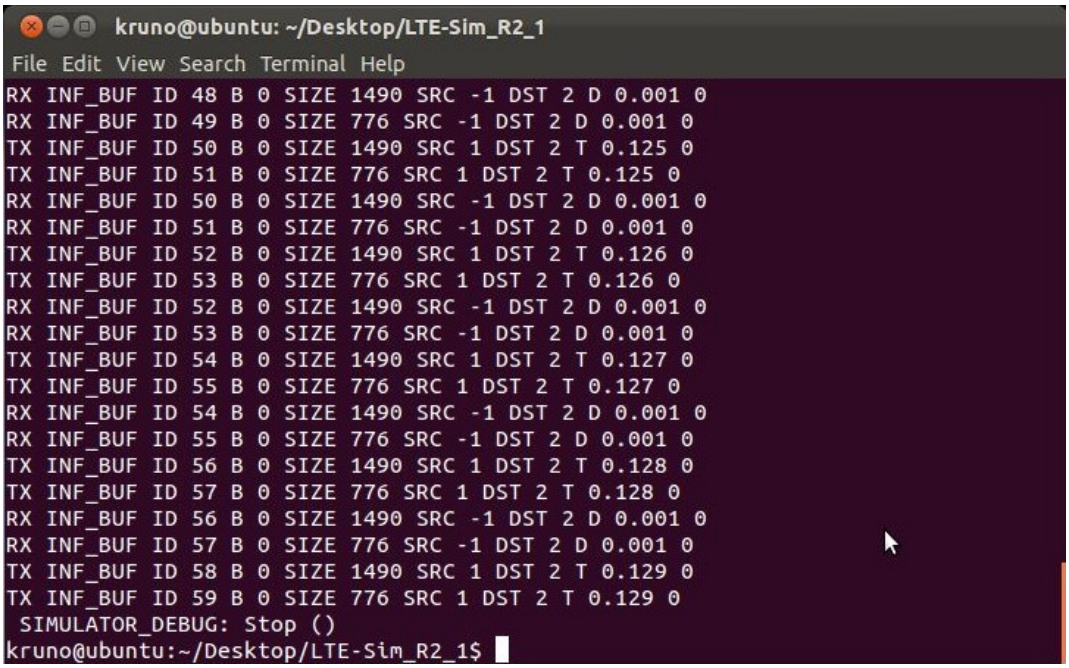


Figure A1: output of LTE-Sim.

Figure A2 shows the interactions among the main blocks of the transmitting/receiving chain. The models for channel, noise, and interference determine how the signal is received at the mobile terminal. Then, the receiver model describes how the signal is processes, and especially what is the effective SINR at the end of the process, considering e.g. multiple sub-channels. This

information is useful for two purposes: first, it is quantized into the CQI values and sent to the base station, where it is used for scheduling and link adaptation; second, it is used in-place at the mobile terminal for the BLER model, i.e. to determine the probability that a data block is received with errors and must then be discarded.

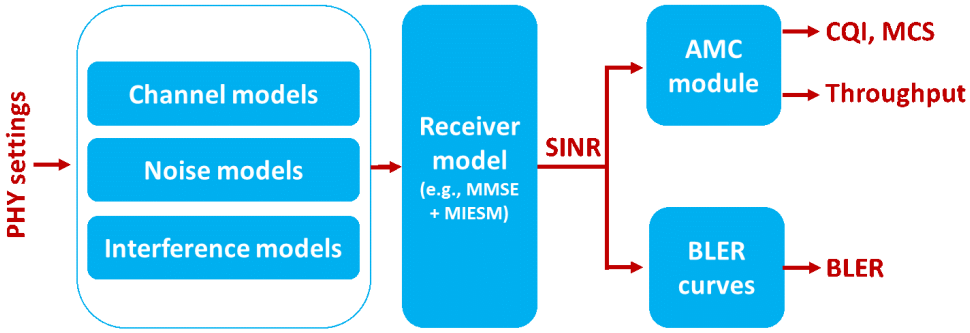


Figure A2: main blocks of the transmission /reception chain.

In Figure A3, the main components recently added to LTE-Sim are shown. Some of them still refer to LTE, while also serving as a foundation for more advanced tools. Other components go beyond the realm of LTE-based technology, so they classify as possible 5G tools. The long-term goal is to implement a full and standard-compliant model of the flexible 5G air interface.

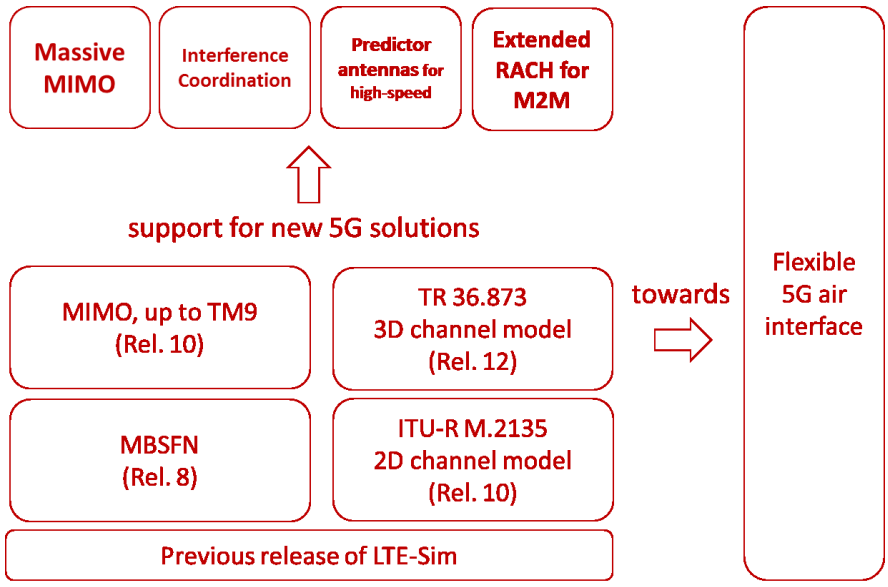


Figure A3: main features developed on top of the previously existing release of LTE-Sim.

The next paragraphs describe the relationship between the new features and the conducted research activities, as well as a general description of the implementation. For the evaluations presented in Section 2.2, LTE-Sim was extended to support MIMO and Massive MIMO transmission modes. This required the entire transmission chain to be extended to use vector/matrix values instead of scalars, including transmitter, channel, and receiver. Before transmission, the base station applies a precoding matrix to the signal, to modify the allocation of each input signal to the different antennas. Such precoding can be based on codebooks or on explicit channel feedback. When the receiver has multiple antennas, it also performs processing to improve or separate the received signals. Moreover, the user equipment is extended to provide additional CSI to the base station, besides regular CQIs, that assists the base station for the precoding operation.

For the work in Chapter 5, multicast/broadcast functionality is introduced in the simulator, specifically in the form of the MBSFN technique available in LTE. This is accomplished by deriving from the *UserEquipment* class a new class *MulticastDestination*, which represents the entire group of users receiving the MBSFN transmission. For the purposes of scheduling, generating application packets, and collecting CQI feedbacks, it acts similarly to a single mobile terminal. However, when the scheduler allocates resources, all the users in the groups are instructed to receive the same data on the same RBs. To model the MBSFN frame structure, the *FrameManager* class is also extended to store a given pattern of MBSFN and non-MBSFN sub-frames. This pattern is repeated for every radio frame and some parts of the simulator behave differently depending on the current sub-frame type. For example, during MBSFN sub-frames the signal coming from cells in the MBSFN area (other than the serving cell) is treated as useful signal, rather than interference. Moreover, when CQIs are calculated during MBSFN sub-frames, they are not sent directly to the base station, but to the *MulticastDestination* object, which calculates the lowest common denominator according to some criteria and then sends the result to the base station.

For the work in Chapter 4, the simulator has been extended with the RACH procedure, which needs to be completed before a user equipment can transmit data in the uplink direction. This is controlled by its active/inactive status: at the beginning of the simulation, users are in the inactive state, and when they generate data to send, the RACH procedure is triggered, involving the exchange of multiple message as described in Chapter 4. Most of this procedure is self-contained in dedicated classes, and only in case of a successful completion the user is switched to the active state. At this point, the user is able to transmit a scheduling request and receive an uplink allocation. Instead, if the procedure fails, it is restarted after a timeout, up to a maximum number of times. Even when the process succeeds, the user will return to the inactive state after a given inactivity period, and the procedure must be performed again for new data. The two versions of the RACH procedure are implemented in different classes derived from common base classes, so that is would be relatively easy to add new schemes for the RACH procedure by writing the corresponding classes.

Most of the new functionalities of LTE-Sim have been released and are available at https://bitbucket.org/telematicslab/5g-simulator. Detailed information on how to compile, debug, and use it are included. More features will be released as soon as they are added and/or refined.

References

[1] D. Astély, E. Dahlman, A. Furuskär, Y. Jading, M. Lindström e S. Parkvall, «LTE: the evolution of mobile broadband,» *IEEE Communications magazine,* 2009.

[2] M. R. Bhalla e A. V. Bhalla, «Generations of mobile wireless technology: A survey,» *International Journal of Computer Applications,* vol. 5, n. 4, 2010.

[3] NGMN Alliance, «5G White Paper,» 2015.

[4] G. Gu e G. Peng, «The survey of GSM wireless communication system,» in *Proc of. International Conference on Computer and Information Application (ICCIA)*, 2010.

[5] C. Bettstetter, H.-J. Vogel e J. Eberspacher, «GSM phase 2+ general packet radio service GPRS: Architecture, protocols, and air interface,» *IEEE communications Surveys,* vol. 2, n. 3, pp. 2-14, 1999.

[6] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjo, M. Persson e C. Roobol, «WCDMA-the radio interface for future mobile multimedia communications,» *IEEE Transactions on vehicular technology,* vol. 47, n. 4, 1998.

[7] D. Martín-Sacristán, J. F. Monserrat, J. Cabrejas-Peñuelas, D. Calabuig, S. Garrigas e N. Cardona, «On the way towards fourth-generation mobile: 3GPP LTE and LTE-advanced,» *EURASIP Journal on Wireless Communications and Networking,* 2009.

[8] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe e T. Thomas, «LTE-advanced: next-generation wireless broadband technology,» *IEEE wireless communications,* vol. 17, n. 3, 2010.

[9] R. W. Heath Jr, M. Honig, S. Nagata, S. Parkvall e A. C. Soong, «LTE-advanced pro: part 1 [Guest Editorial],» *IEEE Communications Magazine,* vol. 54, n. 5, 2016.

[10] S. Chen e J. Zhao, «The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication,» *IEEE communications magazine,* vol. 52, n. 5, 2014.

[11] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka e others, «Scenarios for 5G mobile and wireless communications: the vision of the METIS project,» *IEEE Communications Magazine,* vol. 52, n. 5, 2014.

[12] F. Schaich, B. Sayrac, M. Schubert, H. Lin, K. Pedersen, M. Shaat, G. Wunder e A. Georgakopoulos, «FANTASTIC-5G: 5G-PPP Project on 5G air interface below 6 GHz,» in *Proc. of European Conference on Network and Communications (EUCNC)*, 2015.

[13] 3GPP, «NR; Physical channels and modulation,» 2018.

[14] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfr\"ojd e T. Svensson, «The role of small cells, coordinated multipoint, and massive MIMO in 5G,» *Communications Magazine, IEEE,* vol. 52, n. 5, pp. 44-51, May 2014.

[15] L. Chiaraviglio, N. Blefari-Melazzi, W. Liu, J. A. Gutierrez, J. Van De Beek, R. Birke, L. Chen, F. Idzikowski, D. Kilper, J. P. Monti e others, «5G in rural and low-income areas: Are we ready?,» in *Proc. of ITU Kaleidoscope: ICTs for a Sustainable World*, 2016.

[16] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti e A. Ghosh, «Machine-type communications: current status and future perspectives toward 5G systems,» *IEEE Communications Magazine,* vol. 53, n. 9, 2015.

[17] M. Simsek, A. Aijaz, M. Dohler, J. Sachs e G. Fettweis, «5G-enabled tactile internet,» *IEEE Journal on Selected Areas in Communications,* vol. 34, n. 3, 2016.

[18] X. Cheng, C. Chen, W. Zhang e Y. Yang, «5g-enabled cooperative intelligent vehicular (5genciv) framework: When benz meets marconi,» *IEEE Intelligent Systems,* vol. 32, n. 3, 217.

[19] D.-T. Phan-Huy, M. Sternad e T. Svensson, «Making 5G adaptive antennas work for very fast moving vehicles,» *IEEE Intelligent Transportation Systems Magazine,* vol. 7, n. 2, pp. 71-84, 2015.

[20] M. Vondra, E. Dinc e C. Cavdar, «Coordinated Resource Allocation Scheme for 5G Direct Air-to-Ground Communication,» in *Proc. of 24th European Wireless Conference*, 2018.

[21] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro e A. Iera, «Multicasting over emerging 5G networks: Challenges and perspectives,» *IEEE Network,* vol. 31, n. 2, 2017.

[22] S. Din, A. Paul, A. Ahmad e S. Rho, «Emerging mobile communication technologies for healthcare system in 5G network,» in *Proc. of 14th Intl Conf on Dependable, Autonomic and Secure Computing; 2nd Intl Conf on Pervasive Intelligence and Computing, Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, 2016.

[23] T. L. Marzetta, «Noncooperative cellular wireless with unlimited numbers of base

station antennas,» *IEEE Transactions on Wireless Communications,* vol. 9, n. 11, pp. 3590-3600, 2010.

[24] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin e R. Zhang, «An overview of massive MIMO: Benefits and challenges,» *IEEE Journal of Selected Topics in Signal Processing,* vol. 8, n. 5, pp. 742-758, 2014.

[25] E. G. Larsson, O. Edfors, F. Tufvesson e T. L. Marzetta, «Massive MIMO for next generation wireless systems,» *IEEE Communications Magazine,* vol. 52, n. 2, pp. 186-195, 2014.

[26] J. Choi, J. Mo e R. W. Heath, «Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs,» *IEEE Transactions on Communications,* vol. 64, n. 5, pp. 2005-2018, 2016.

[27] J. Vieira, F. Rusek e F. Tufvesson, «Reciprocity calibration methods for massive MIMO based on antenna coupling,» in *Proc. of IEEE Global Communications Conference (GLOBECOM)*, 2014.

[28] K. Zheng, S. Ou e X. Yin, «Massive MIMO channel models: A survey,» *International Journal of Antennas and Propagation,* vol. 2014, 2014.

[29] J. Jose, A. Ashikhmin, T. L. Marzetta e S. Vishwanath, «Pilot contamination and precoding in multi-cell TDD systems,» *IEEE Transactions on Wireless Communications,* vol. 10, n. 8, pp. 2640-2651, 2011.

[30] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo e S. Z. Iliya, «A comprehensive survey of pilot contamination in massive MIMO-5G system,» *IEEE Communications Surveys \& Tutorials,* vol. 18, n. 2, pp. 905-923, 2016.

[31] A. Ashikhmin e T. Marzetta, «Pilot contamination precoding in multi-cell large scale antenna systems,» in *Proc. of IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2012.

[32] F. Fernandes, A. Ashikhmin e T. L. Marzetta, «Inter-cell interference in noncooperative TDD large scale antenna systems,» *IEEE Journal on Selected Areas in Communications,* vol. 31, n. 2, pp. 192-201, 2013.

[33] S.-C. Lin e H. Narasimhan, «Towards Software-Defined Massive MIMO for 5G\&B Spectral-Efficient Networks,» in *Proc. of IEEE International Conference on Communications (ICC)*, 2018.

[34] E. Björnson, E. G. Larsson e T. L. Marzetta, «Massive MIMO: Ten myths and one

critical question,» *IEEE Communications Magazine,* vol. 54, n. 2, pp. 114-123, 2016.

[35] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone e W. Yu, «Multi-cell MIMO cooperative networks: A new look at interference,» *IEEE Journal on Selected Areas in Communications,* vol. 28, n. 9, pp. 1380-1408, 2010.

[36] Z. Jiang, A. F. Molisch, G. Caire e Z. Niu, «Achievable rates of FDD massive MIMO systems with spatial channel correlation,» *IEEE Transactions on Wireless Communications,* vol. 14, n. 5, pp. 2868-2882, 2015.

[37] X. Rao e V. K. Lau, «Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems,» *IEEE Transactions on Signal Processing,* vol. 62, n. 12, pp. 3261-3271, 2014.

[38] M. S. Sim, J. Park, C.-B. Chae e R. W. Heath, «Compressed channel feedback for correlated massive MIMO systems,» *Journal of Communications and Networks,* vol. 18, n. 1, pp. 95-104, 2016.

[39] J. Choi, D. J. Love e P. Bidigare, «Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory,» *IEEE Journal of Selected Topics in Signal Processing,* vol. 8, n. 5, pp. 802-814, 2014.

[40] J. Choi, Z. Chance, D. J. Love e U. Madhow, «Noncoherent trellis coded quantization: A practical limited feedback technique for massive MIMO systems,» *IEEE Transactions on Communications,* vol. 61, n. 12, pp. 5016-5029, 2013.

[41] W. Zirwas, L. Thiele, M. Kurras e G. Wunder, «Flexible 5G below 6GHz mobile broadband radio air interface,» in *Proc. of IEEE 83rd Vehicular Technology Conference (VTC Spring)*, 2016.

[42] M. Kurras, S. Jaeckel, L. Thiele e V. Braun, «CSI Compression and Feedback for Network MIMO,» in *Proc. of IEEE 81st Vehicular Technology Conference (VTC Spring)*, 2015.

[43] A. Adhikary, J. Nam, J.-Y. Ahn e G. Caire, «Joint Spatial Division and Multiplexing: The Large-Scale Array Regime,» *Information Theory, IEEE Transactions on,* vol. 59, n. 10, pp. 6441-6463, 2013.

[44] C. B. Peel, B. M. Hochwald e A. L. Swindlehurst, «A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization,» *IEEE Transactions on Communications,* vol. 53, n. 1, pp. 195-202, 2005.

[45] W. Zirwas, M. B. Amin e M. Sternad, «Coded CSI Reference Signals for 5G-Exploiting Sparsity of FDD Massive MIMO Radio Channels,» in *Proc. of 20th International ITG Workshop on Smart Antennas (WSA 2016)*, 2016.

[46] M. Kurras, L. Thiele e G. Caire, «Multi-stage beamforming for interference

coordination in massive MIMO networks,» in *Proc. of 49th Asilomar Conference on Signals, Systems and Computers*, 2015.

[47] B. L. Ng, Y. Kim, J. Lee, Y. Li, Y.-H. Nam, J. Zhang e K. Sayana, «Fulfilling the promise of massive MIMO with 2D active antenna array,» in *Proc. of IEEE Globecom Workshops (GC Wkshps)*, 2012.

[48] Y. Kim, H. Ji, H. Lee, J. Lee, B. L. Ng e J. Zhang, «Evolution beyond LTE-advanced with Full Dimension MIMO,» in *Proc. of IEEE International Conference on Communications Workshops (ICC)*, 2013.

[49] 3GPP, «Study on Elevation Beamforming/Full-Dimension (FD) MIMO for LTE,» 2015.

[50] ZTE, «R1-1715253 Calibration results for Phase 2 NR MIMO system level calibration,» in *V3GPP TSG RAN WG1 Meeting #90*, 2017.

[51] X. He, K. Niu, Z. He e J. Lin, «Link layer abstraction in MIMO-OFDM system,» in *Proc. of International Workshop on Cross Layer Design*, 2007.

[52] D. Neumann, M. Joham, L. Weiland e W. Utschick, «Low-complexity computation of LMMSE channel estimates in massive MIMO,» in *Proc. of 19th International ITG Workshop on Smart Antennas (WSA 2015)*, 2015.

[53] R. M. Gray e others, «Toeplitz and circulant matrices: A review,» *Foundations and Trends{\textregistered} in Communications and Information Theory,* vol. 2, n. 3, pp. 155-239, 2006.

[54] J. Nam, A. Adhikary, J.-Y. Ahn e G. Caire, «Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling,» *IEEE Journal of Selected Topics in Signal Processing,* vol. 8, n. 5, pp. 876-890, 2014.

[55] Y.-G. Lim, C.-B. Chae e G. Caire, «Performance analysis of massive MIMO for cell-boundary users,» *IEEE Transactions on Wireless Communications,* vol. 14, n. 12, pp. 6827-6842, 2015.

[56] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel e S. Geirhofer, «Network densification: the dominant theme for wireless evolution into 5G,» *Communications Magazine, IEEE,* vol. 52, n. 2, pp. 82-89, February 2014.

[57] M. Polignano, P. Mogensen, P. Fotiadis, L. Chavarria, I. Viering e P. Zanier, «The Inter-Cell Interference Dilemma in Dense Outdoor Small Cell Deployment,» in *Proc.*

of IEEE 79th Vehicular Technology Conference (VTC Spring), 2014.

[58] M. Kurras, L. Thiele e G. Caire, «Interference Mitigation and Multiuser Multiplexing with Beam-Steering Antennas,» in *Proc. of 19th International ITG Workshop on Smart Antennas (WSA)*, 2015.

[59] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi e P. Camarda, «Simulating LTE cellular systems: An open-source framework,» *IEEE transactions on vehicular technology,* vol. 60, n. 2, pp. 498-513, 2011.

[60] 3GPP, «Study on 3D channel model for LTE,» 2015.

[61] 3GPP, «Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures,» 2016.

[62] 3GPP, «Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation,» 2008.

[63] X. Sun, X. Gao, G. Y. Li e W. Han, «Agglomerative user clustering and downlink group scheduling for FDD massive MIMO systems,» in *Proc. of IEEE International Conference on Communications (ICC)*, 2017.

[64] M. Kurras, S. Fähse e L. Thiele, «Density Based User Clustering for Wireless Massive Connectivity Enabling Internet of Things,» in *Proc. of Globecom Workshops*, 2015.

[65] Y. Sun, S. Lv, S. Liu e Y. Zhang, «Density based user grouping for massive MIMO downlink in FDD system,» in *Proc. of IEEE 9th International Conference on Communication Software and Networks (ICCSN)*, 2017.

[66] M. Ester, H.-P. Kriegel, J. Sander, X. Xu e others, «A density-based algorithm for discovering clusters in large spatial databases with noise.,» in *Proc. of Conference on Knowledge Discovery and Data Mining (KDD)*, 1996.

[67] A. Van Zelst, «A compact representation of spatial correlation in MIMO radio channels,» in *submitted to International Conference on Communications (ICC)*, 2004.

[68] D.-S. Shiu, G. J. Foschini, M. J. Gans e J. M. Kahn, «Fading correlation and its effect on the capacity of multielement antenna systems,» *IEEE Transactions on Communications,* vol. 48, n. 3, pp. 502-513, 2000.

[69] G. Bacci, E. V. Belmega, P. Mertikopoulos e L. Sanguinetti, «Energy-Aware Competitive Power Allocation for Heterogeneous Networks Under QoS Constraints,» *IEEE Trans. Wireless Commun.,* vol. 14, n. 9, pp. 4728-4742, #sep# 2015.

[70] G. Bacci, L. Sanguinetti e M. Luise, «Understanding game theory via wireless power control,» *IEEE Signal Processing Magazine,* vol. 32, n. 4, pp. 132-137, #jul# 2015.

[71] F. Facchinei e C. Kanzow, «Generalized Nash equilibrium problems,» *4OR,* vol. 5, n. 3, pp. 173-210, 2007.

[72] J. Olmos, S. Ruiz, M. García-Lozano e D. Martín-Sacristán, «Link abstraction models based on mutual information for LTE downlink,» in *COST*, 2010.

[73] G. Piro, A. Orsino, C. Campolo, G. Araniti, G. Boggia e A. Molinaro, «D2D in LTE vehicular networking: system model and upper bound performance,» in *Proc. of IEEE International Congress on Ultra Modern Telecommunications and Control Systems*, 2015.

[74] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia e P. Camarda, «Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey,» *IEEE Commun. Surveys and Tutorials,* vol. 15, n. 2, pp. 678-700, Apr. 2013.

[75] 3GPP, «Physical layer aspect for evolved Universal Terrestrial Radio Access (UTRA),» 2006.

[76] T. Camp, J. Boleng e V. Davies, «A survey of mobility models for ad hoc network research,» *Wireless communications and mobile computing,* vol. 2, n. 5, pp. 483-502, 2002.

[77] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. J. Gonzalez e others, «Flexible power modeling of LTE base stations,» in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, 2012.

[78] R. Hadani, S. Rakib, M. Tsatsanis, A. Monk, A. J. Goldsmith, A. F. Molisch e R. Calderbank, «Orthogonal time frequency space modulation,» in *Proc. of Wireless Communications and Networking Conference (WCNC)*, 2017.

[79] S. Barbera, K. I. Pedersen, C. Rosa, P. H. Michaelsen, F. Frederiksen, E. Shah e A. Baumgartner, «Synchronized RACH-less handover solution for LTE heterogeneous networks,» in *Proc. of International Symposium on Wireless Communication Systems (ISWCS)*, 2015.

[80] M. Sternad, M. Grieger, R. Apelfröjd, T. Svensson, D. Aronsson e A. B. Martinez, «Using "predictor antennas" for long-range prediction of fast fading for moving relays,» in *Proc. of Wireless Communications and Networking Conference Workshops (WCNCW)*, 2012.

[81] R.-G. Cheng, C.-H. Wei, S.-L. Tsao e F.-C. Ren, «RACH collision probability for machine-type communications,» in *Proc. of Vehicular Technology Conference (VTC Spring)*, 2012.

[82] A. Laya, L. Alonso e J. Alonso-Zarate, «Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives,» *IEEE Communications Surveys \& Tutorials,* vol. 16, n. 1, pp. 4-16, 2014.

[83] J. Kim, D. Munir, S. Hasan e M. Chung, «Enhancement of LTE RACH through extended random access process,» *Electronics Letters,* vol. 50, n. 19, pp. 1399-1400, 2014.

[84] 3GPP, «RAN Improvements for Machine-type Communications,» 2014.

[85] ITU-R, «Guidelines for evaluation of radio interface technologies for IMT-Advanced,» *Tech. Rep. M.2135,* 2009.

[86] A. Alexiou, C. Bouras, V. Kokkinos, A. Papazois e G. Tsichritzis, «Spectral efficiency performance of MBSFN-enabled LTE networks,» in *Proc. of International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2010.

[87] P. Marques, J. Rodriguez, T. Forde, L. Doyle, K. Won Sung, J. Lauterjung e U. H. Reimers, Towards a Unified 5G Broadcast-Broadband Architecture, 2015, pp. 191-206.

[88] 3GPP, «Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects,» 2010.

[89] D. Lecompte e F. Gabin, «Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and Rel-11 enhancements,» *IEEE Communications Magazine,* vol. 50, n. 11, 2011.

[90] B. Mouhouche e A.-I. Mohammed, «Optimization of delivery time in broadcast with acknowledgement and partial retransmission,» in *Proc. of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2016.

[91] L. Rong, O. B. Haddada e S.-E. Elayoubi, «Analytical analysis of the coverage of a MBSFN OFDMA network,» in *Proc. of IEEE Global Telecommunications Conference*, 2008.