Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

PhD Thesis

Enabling Adaptable Future Networks: Trade-Offs and Resource Allocation Problems

Author:
Conor Sexton

Supervisors:
Prof. Luiz DaSilva
Prof. Nicola Marchetti

26th May 2020

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed:

Conor Sexton

Conor Sexton, 26th May 2020

Summary

The use cases to be supported by future networks, ranging from massive machine type communications to ultra-reliable low latency applications, are too diverse and contrasting to be served efficiently by a single type of network. In addition, the new technologies being suggested for future networks, such as in-band full duplex, new waveforms, Massive Multiple-Input Multiple-Output (Massive-MIMO), and millimetre wave, display clear heterogeneity in their capabilities, proving advantageous in certain scenarios but not in others. As a result, future networks will be more adaptable in nature, requiring the coexistence of contrasting technologies and offering customisable network behaviour on-demand through virtualisation.

In this thesis, we illuminate the various trade-offs arising from the trend towards customisable networks, and propose resource allocation procedures to balance these trade-offs and facilitate the necessary coexistence of Radio Access Technologies (RATs) required to enable adaptable future networks. The problems we address in this thesis can be stated in the form of the following research questions:

- What choices do the range of proposed RATs and system-level techniques afford operators in the context of enabling an adaptable network?
- How can the diverse use cases in future networks be satisfied through the coexistence of multiple waveforms, with each service employing a waveform that is best suited for it?
- What are the implications of a system comprising tailored virtual networks on radio resource allocation and admission control?
- How can the twin goals of slice-tailored performance and increased resource utilisation in network slicing be simultaneously realised in future networks?

To address these questions we apply a range of analytical tools, including optimisation, matching theory, and stochastic analysis. We verify our results using large-scale system-level simulations where possible.

First, we survey the choices and adaptability afforded by some of the RATs being considered for future networks and explore how several system-level techniques, such as Software-Defined Networking (SDN) and Virtualised RAN (VRAN), can be utilised to enable and manage versatile networks. Specifically, we focus on the need for multiple technologies to coexist in future networks to satisfy the diverse range of requirements demanded of the network. In this regard, we identify virtualisation as an enabling technology, permitting service-tailored network slices to be created on-demand. This lays the foundation for the rest of the thesis, underscoring the need for customised network behaviour and highlighting some of the resulting trade-offs that occupy the focus of subsequent chapters.

Then, we concentrate our attention on the coexistence of technologies that offer contrasting performance and potentially interfere with one another. Specifically, we explore the idea that future networks will permit the use of multiple waveforms, with each service employing a waveform that is best suited for it. In particular, we look at a scenario consisting of clustered Device-to-Device (D2D) communication, representative of a future network use case such as a smart factory with inter-communicating machinery. We first demonstrate that inter-D2D interference, resulting from misaligned communications, plays a significant role in clustered D2D topologies. We then demonstrate, through simulation, that this interference can be mitigated through the use of alternative waveforms, without needing to increase the complexity of current resource allocation strategies.

With the need for customisable network behaviour evident, we consider approaches that utilise virtualisation to manage the coexistence of technologies through specialised virtual sub-networks. We begin with an examination of the potential time-frequency resource structure of the Radio Access Network (RAN), focusing on the trade-off between flexibility and the overhead related to ensuring coexistence between contrasting RAN slices. Based on this analysis, we suggest an approach that permits the allocation of resources to a service-type to be performed separately to resource allocation for individual services belonging to that type. The advent of service-tailored networks opens the door for new business models, enabling a marketplace of specialised network operators who create and manage bespoke tailored networks. We investigate a network model in which entities called subscription brokers group service level agreements with multiple Specialised Mobile Network Operators (SMNOs) into a single subscription bundle, and focus on how to perform the matching between users and SMNOs in a bundle, adopting the Gale-Shapley matching algorithm. Overall, the proposed broker-based model performs at least as well as any one SMNO for lower priority users, and outperforms any one SMNO for higher priority users.

Finally, we examine the trade-off in network slicing between the twin goals of providing tailored performance and increasing resource utilisation through increased opportunities for sharing. To balance this trade-off, we propose a system consisting of assured resources, which are available if needed over the lifetime of a slice, and auxiliary resources, which are offered on a probabilistic basis in the short-term based on forecasted resource demand. We employ the practice of overbooking, which is widely used in many industries such as airlines and hotels, to increase resource utilisation when offering auxiliary resources. After deriving probabilistic results relating to the availability of auxiliary resources, we then design an algorithm to determine how much to overbook by. We also propose several ways of distributing auxiliary resource offers among slices, as well as providing an approach for dealing with overbookings. Finally, we highlight the conditions that are conducive to effective overbooking by examining the effects of varying several key system parameters.

Acknowledgements

I am indebted to my supervisors, Luiz DaSilva and Nicola Marchetti, for their support and guidance over the previous four years. I have learnt so much during the PhD from them both, much of which extends far beyond academia. In particular, I wish to express my gratitude to Nicola for initially inducting me into the research world and introducing me to CONNECT.

Without question, the greatest aspect of doing this PhD has been the fantastic people with whom I have shared this journey. Although many people passed through the doors of CONNECT during my four years, and each made an impression in their own way, there are a few who deserve a special mention. Thank you to Merim for being a great friend, for helping me discover my business persona, and for inadvertently improving my grammar and vocabulary. To Joao for his energy and infectious enthusiasm. To Nima for his kindness. To Erika for her thoughtfulness. To Maicon for being a calming presence in the lab. To Alan for always being good craic. To Jernej for the football chats. To Parna for being a good neighbour.

I also owe a huge thank you to Nicholas Kaminski and Johann Marquez-Barja, who helped me find my feet in the first year of my PhD and left a lasting impression. Thank you to Arman Farhang for his unquestionable knowledge (and questionable humour). A special thank you to Quentin Bodinier, whose collaboration was one of the most enjoyable periods of my PhD, and helped me endlessly. Thank you also to all the administrative staff, and in particular Monica, for assisting with all my clueless enquiries. Finally thank you to everyone else for the memories: Fadhil, Ramy, Andrea, Andrei, Boris, Jacek, Eamonn, Andrew, Diarmuid, Stefan, Yi, Elma, Tom, both Lindas, Harleen, Georgios, Marcelo, Marco, Irene, Francisco, Pedro, Harun, Jonathan, and so many more.

I would like to thank my parents, John and Louise, for always being there. This PhD is the product of the never-ending encouragement they have always provided, and would not have been possible without them. Thank you also to my siblings, Mark and Martina, for spurring me on and giving me the belief to accomplish this. I would also like to thank Lucy, Tomás, and Rory for the countless laughs and abundance of joy that they brought me in the final years of my PhD.

And last, but certainly not least, a huge thanks to my loving (and incredibly patient) girlfriend, Katie, who provided me with endless support and believed in me when I did not. I could not have done this without you.

Contents

List of Figures

List of Tables

1 Introduction

Introduction

Unlike previous generations, which were primarily defined by their approach to the air interface and multiple access scheme (i.e. Universal Mobile Telecommunications Service (UMTS)/Wideband Code Division Multiple Access (WCDMA) and Long Term Evolution (LTE)/Orthogonal Frequency Division Multiple Access (OFDMA)), future telecommunication networks are envisioned to be a very different type of network [5]. This difference can be largely attributed to the core ideas of flexibility and adaptability; future networks should present a wireless access platform that offers connectivity to any entity that may need it, designed in a flexible manner to meet the requirements of a diverse range of service requirements.

Two factors which motivate this vision of future networks as different compared to previous generations are:

- Diverse service requirements. Previous generations targeted a limited set of service requirements corresponding to the killer application of the time; for example, voice in the Global System for Mobile Communications (GSM) or data in LTE. In contrast, it is widely accepted that a future network will need to be able to handle 1000x traffic volumes, provide a 100x increase in the edge data rate, have near zero latency, provide ultra-high reliability and availability, while reducing or at least maintaining current energy consumption and costs. Not every service will require each of the above requirements, influencing the selection and allocation of resources.
- New Physical Layer (PHY) technologies. Although every generation has introduced new technologies, the technologies being considered for future networks introduce new levels of heterogeneity in their uses and advantages. Millimetre wave communication, for example, offers unprecedented data rates but also presents new challenges such as extreme susceptibility to blockages, making it less than ideal in scenarios requiring high mobility. As another example, In-Band Full Duplex (IBFD) promises the potential to improve spectral efficiency, however due to the new types of interference introduced into the network, it may not always outperform its Half Duplex (HD) counterpart [6]. Even from such a small sample of examples, it is clear that new technologies do not always equate to better performance in every situation.

The two points above are complementary. On one hand, we have numerous PHY technologies which, due to their various characteristics, are ideally suited to certain services and not others. On the other hand we have a diverse set of service requirements to be supported by future networks, each of which demands different capabilities from the network.

It is evident that a future network will comprise several coexisting, yet potentially conflicting, technologies to provide the multi-service capabilities demanded of it. This is a marked change from previous generations, which could be optimised towards a single goal related to capacity. The trend towards a multi-service system raises several challenges surrounding how best to achieve the flexibility required in the network while maintaining a high level of performance.

One way to achieve this flexibility is through the ability to slice resources logically, where each slice represents a partition of network resources and is tailored to meet the demands of a particular service [7]. Future networks should not therefore be imagined as an incremental upgrade to previous generations, but rather as an all encompassing platform that allows the creation of service-specific slices on-demand.

Hence, we can encapsulate the shift towards customisable networks as follows. Diverse and non-traditional use cases for future networks require a multi-service network capable of satisfying an array of often contrasting Key Performance Indicators (KPIs). At the same time, we have at our disposal numerous technologies that are advantageous in some scenarios and unfavourable in others, necessitating the coexistence of technologies in a single network. This can be achieved by confining specific technologies in logically isolated virtual networks which have been constructed to serve a particular communication demand. Network slicing has been advanced as an enabling technology in this regard.

We focus on several resource allocation problems related to enabling a multi-service system consisting of service-tailored slices. In general, every resource allocation problem is built upon a trade-off between two or more desirable states that cannot be achieved simultaneously. The primary trade-off at the heart of the resource allocation problems in this thesis relates to the cost of adaptability in a network, with the cost referring to increased complexity, increased overhead, and reduced resource utilisation efficiency.

1.1 Scope

We employ an array of analytical and simulation-based techniques to address several problems in the area of resource allocation in adaptable wireless networks, with the overall objective being to build a clear and comprehensive view of the trade-offs, opportunities and considerations associated with service-tailored future networks.

This thesis has a wide ranging scope, from discussing the manner in which networks can increase flexibility by leveraging new technologies, to modelling, assessing and simulating the effects of various new resource allocation procedures. In general, we address the following four research questions:

What choices do the range of proposed physical layer technologies and system-level techniques afford operators in the context of enabling an adaptable network?

The Radio Access Network (RAN) in future networks will require versatile solutions that can be adapted to satisfy many different services and applications. We appraise various Radio Access Technologies (RATs) and architectures in terms of the benefits they can offer in the three following areas:

- enhanced Mobile Broadband (eMBB);
- Massive Machine-Type Communication (mMTC);
- Ultra-Reliable Low Latency Communication (uRLLC).

The core network will also undergo fundamental changes, with increased levels of abstraction allowing for further reconfiguration of the network. As part of this question, we examine the choices afforded by some of the RATs being considered for future networks and explore how several system-level techniques, such as software-defined networking and Virtualised RAN (VRAN), can be utilised to enable and manage versatile future networks. The necessity for multiple technologies to coexist and the potential of virtualisation to enable the creation of service-tailored, logically isolated sub-networks is a key focus and is the foundation upon which the remaining research questions in this thesis depend.

How can the diverse use cases in future networks be satisfied through the coexistence of multiple waveforms, with each service employing a waveform that is best suited for it?

Focusing on the need for multiple technologies to coexist in future networks, we examine the relative merits of several alternative waveforms to Orthogonal Frequency Division Multiplexing (OFDM) for different services. We suggest that future standards may allow for multiple waveforms to be adopted dependent on the use case being served. We look at a Machine-Type Communication (MTC) scenario consisting of clustered user equipment employing Device-to-Device (D2D) communication, such as a smart factory with inter-communicating machinery. The overhead associated with synchronising a large number of machine-type Directly Communicating Users (DUEs) comes at a cost that may render synchronous communication infeasible or undesirable. Based on this motivation, we consider multiple possible combinations of prominent future network waveform candidates for cellular users and DUEs, examining the asynchronous performance of all waveforms under consideration and using the performance of synchronous OFDM as a baseline for comparison. Specifically, we focus on the coexistence of waveforms in which the ordinary cellular users employ OFDM for synchronous communication, as in LTE, and the machine-type DUEs, operating asynchronously, employ a different waveform. The aim is to determine whether permitting multiple waveforms to coexist will provide a performance increase to the targeted services, and to investigate the effects of this set-up on traditional resource allocation schemes.

What are the implications of a system comprising tailored virtual networks on radio resource allocation and user admission?

We investigate the effects that service-tailored slices will have on resource allocation, focusing on two areas:

- the allocation of time-frequency resources to slices;
- the association between users and slices.

In the first case, we consider slices which have been tailored by using different numerologies and examine the key trade-off between the flexibility that this introduces and the associated overhead. It is imperative to the success of RAN slicing that the time-frequency resource grid is designed in such a manner as to manage this trade-off and facilitate the multiple tiers of resource allocation that are required. In the second case, we investigate the issue of how users select slices in a business model consisting of several independently operated, service-tailored slices. We employ a branch of game theory, called matching theory, to examine how the matching between users and slices should be performed. We suggest the introduction of a new entity, called a subscription broker, to assist in the matching process and examine the performance of such a scheme.

How can the twin goals of slice-tailored performance and increased resource utilisation in network slicing be simultaneously realised in future networks?

Network slicing offers two distinct benefits. First, its ability to offer tailored network behaviour allows it to satisfy many different types of requirements and target new industries, verticals, and user communities. Secondly, the statistical multiplexing gains arising from shared network functions and infrastructure can keep costs reasonable. The two benefits are in conflict, however, as permitting flexible sharing hinders a slice's ability to provide a guarantee of performance to its subscribers. Both elements are necessary: network slicing that provides tailored networks but limited sharing essentially reduces to dedicated networks, while aggressive sharing without providing customisability reduces to resource management. We propose an approach to simultaneously realise the twin goals of slice-tailored behaviour and resource efficiency using a combination of long-term and short-term commitments regarding resource availability. The long-term guarantee provides certainty to slices regarding the performance that they can offer to their subscribers, while in the short-term we take advantage of fluctuating demand to increase resource utilisation through the practice of overbooking.

1.2 Outline

We divide the remainder of this thesis into seven chapters. Chapter 2 provides the foundation for the thesis by providing the setting and motivation for the vision of customisable future networks, and we survey opportunities and trends which pave the way towards this vision. Chapter 3 then provides the necessary background into network virtualisation, slicing, and the associated business models. Chapter 4 examines the coexistence of different technologies, namely waveforms. Chapters 5 and 6 focus on the implications of customisable slices on resource allocation, while Chapter 7 examines one of the key trade-offs associated with network slicing. Finally, Chapter 8 summarises our main findings and concludes the thesis.

Chapter 2

In this chapter, we examine the trend in telecommunication networks towards differentiated services, and review the diverse range of requirements demanded of future networks. We then survey some of the RATs and system-level architectures that are critical to achieve the level of versatility required in future networks to satisfy these requirements. Specifically, we explore the choices and opportunities presented by these RATs and system-level techniques to create a more versatile and adaptable network.

Chapter 3

We discuss the importance of network virtualisation as an enabler for customisable future networks. We then survey the state-of-the-art in network slicing, focusing on its ability to create slices with tailored network behaviour, as well as some of the associated proposed business models.

Chapter 4

We first provide an overview of the most prominent alternative waveforms to OFDM, and discuss their relative strengths and weaknesses. We initially consider a single-cell scenario consisting of clustered D2D communication and investigate whether Filter Bank Multi-Carrier with Offset Quadrature Amplitude Modulation (FBMC/OQAM) can reduce the complexity of resource allocation procedures in this scenario. We then expand our work to a multi-cell, multi-cluster network consisting of cells which employ strict Fractional Frequency Reuse (FFR), and consider a wider range of waveforms. Through extensive Monte Carlo simulations, we evaluate the relative asynchronous performance of these waveforms for use in clustered MTC scenarios, compared against a baseline case consisting of synchronous OFDM.

Chapter 5

Considering a future network consisting of service-tailored slices, we begin the chapter with an examination of the benefits of various lower-layer design choices in the RAN for serving different use cases. We then propose four different configurations of the time-frequency resource grid, with each managing the coexistence of RAN slices with approaches ranging from very static to very dynamic. Focusing both on the ability to adapt to traffic changes and how multiple RAN slices can coexist, we examine the relative performance of these different configurations of the resource grid. Finally, based on this analysis, we propose a two-tier time-frequency resource grid that can balance this trade-off between adaptability and coexistence overhead.

Chapter 6

We examine a model consisting of Specialised Mobile Network Operators (SMNOs) which operate independent service-tailored networks, and propose the adoption of a broker-based approach in which users can switch between multiple SMNOs in a subscription bundle based on the type of service that they require. We examine user admission to service-tailored SMNOs in this model, and outline how to perform the matching between users and SMNOs based on their preferences. Finally, using a maximum-utility optimisation approach as a baseline, we evaluate the performance of the proposed broker-based model using a proof-of-concept example case.

Chapter 7

Network slicing is an attractive solution for future networks, offering the potential to provide both customisable behaviour and increased sharing opportunities. The trade-off between these two benefits can be managed by providing both long-term and short-term commitments regarding resource availability. We derive the probabilities of auxiliary resources being available according to a specified guarantee and then we design an algorithm that can be used to determine how many auxiliary resources to offer to slices. The assured resources provide slices with an assurance so that they can provide tailored service to the users, while the specified probability of availability associated with the auxiliary resources facilitates resource overbooking and increases resource sharing. Our analytical results show that the interplay between the controllable and non-controllable inputs into the system influence how effective resource overbooking is, and places limits on the range of guarantees that can be offered.

Chapter 8

Finally, in this chapter, we summarise the main insights of the thesis, and discuss the outlook and future directions for adaptable future networks. We also briefly present another vision for adaptable networks based on cognitive networks instead of virtualisation.

1.3 Contributions

In this section, we present the main research contributions made as part of this thesis. We state our general contribution in this thesis to be one of providing comprehension through the presentation of conceptual insights, as well as analytical and simulation-based results, of the challenges associated with enabling adaptable future networks. We group our contributions based on the chapter in which they appear.

Chapter 2

In this chapter, our main contribution is the provision of a comprehensive survey of the demands of future networks, the tools and techniques available to satisfy those demands, and the opportunities to create an adaptable network. The contributions in this chapter, listed below, relate to providing a vision for future networks and highlighting the research questions that need to be considered to realise it.

- We promote the vision of future networks as a highly versatile and reconfigurable network, capable of adapting to many different service requirements.
- We survey the technologies and tools that will be instrumental in realising this vision of flexible networks.

Chapter 3

Our chief contribution in this chapter lies in highlighting the benefits of virtualisation and network slicing in the context of enabling customisable virtual networks on-demand.

- We provide an overview of virtualisation and network slicing, demonstrating how they may be used to enable flexible future networks. In particular, we focus on RAN slicing, and discuss potential deployment models for network slicing.

Chapter 4

Although waveform research is very mature at a signal processing level, the system-level implications and performance of employing different waveforms in different scenarios is not well studied. Our main contribution, therefore, is to demonstrate the benefits that alternative waveforms to OFDM can provide in future network scenarios such as D2D-enabled MTC. Furthermore, we demonstrate at a system-level that waveforms can coexist in future networks, with different devices potentially adopting different waveforms. We state our contributions as follows:

- We show the effects of inter-D2D interference, taking into account leaked power from sub-bands, for clustered D2D scenarios under various system set-ups.

- We demonstrate that the optimal Resource Allocation (RA) and power-loading schemes can be simplified when D2D pairs use an alternative waveform, as the inter-D2D interference becomes negligible.
- We demonstrate using system-level simulations that it is feasible for cellular networks to serve high rate clustered MTC use cases using D2D communication through the coexistence of alternative waveforms and OFDM, and quantify the benefit of doing so. In particular, we show that DUEs can avail of the benefits of asynchronous communication, without suffering a performance loss, by employing an alternative waveform such as FBMC/OQAM, even if regular Cellular Users (CUEs) continue to use OFDM.
- We characterise the performance of several prominent alternative waveforms across a range of MTC scenarios by varying key system parameters such as cell size, cluster size, DUE transmit power, and maximum possible Timing Offset (TO) and Carrier Frequency Offset (CFO).

Chapter 5

Our contribution in this chapter relates to the allocation of time-frequency resources in an adaptable RAN. While several papers have examined conceptual architectures [8, 9] or proposed radio resource allocation schemes for RAN slicing [10], our contribution is instead on how lower layer RAN concerns pertaining to customisation and flexibility affect the development of RAN slicing, particularly in relation to ensuring isolation between RAN slices. We list our contributions in detail as follows:

- We analyse how four different proposed configurations of the time-frequency resource grid perform with regard to the adaptability/overhead trade-off, providing both a qualitative and quantitative analysis of both sides of the trade-off for each configuration under consideration.
- Informed by this analysis, we distinguish between service-types and individual services through the concept of a RAN profile, which enables resources to be allocated at both different time-scales and granularities based on this distinction.

Chapter 6

In this chapter, we address how to associate users to service-tailored slices, and introduce a new network entity to perform this matching. We list our contributions in detail as follows:

- We show the performance benefits of the proposed broker-based model, allowing users to choose SMNOs according to the needs of the service that they are using. In particular, we demonstrate how to perform the matching between users and SMNOs in a bundle, adopting the Gale-Shapley matching algorithm.
- We outline a framework based on the concept of utility for devising the preference lists of users, while the approach we propose for building the preference lists of SMNOs

can differentiate between different classes of users based on the price they pay for their subscription.
- We evaluate the performance, adopting utility as a metric, of the broker-based model compared to a sum utility maximisation matching approach.

Chapter 7

Our chief contribution in this chapter relates to managing the trade-off between slice-tailored behaviour and resource utilisation by proposing the division of resources into assured resources and auxiliary resources, with the former available if needed over the lifetime of a slice, and the latter offered on a probabilistic basis using short-term forecasts of resource demand. Our contributions are as follows:

- We derive probabilistic results relating to the availability of auxiliary resources.
- We design an algorithm for determining how many auxiliary resources the slice provider can offer to slice tenants, which determines the amount of overbooking that occurs.
- We provide insight into the factors which affect the overbooking of auxiliary resources, and show when conditions dictate that overbooking will be effective.
- We propose several methods for distributing auxiliary resources among slices, including an approach to maximise resource utilisation.

1.4 Dissemination

This section lists the relevant publication works disseminated as part of this PhD project.

Journal Papers

1. C. Sexton, N. Marchetti, and L. A. DaSilva, ‘On Provisioning Slices and Overbooking Resources in Service Tailored Networks of the Future,’ in IEEE/ACM Transactions on Networking, 2019 (under review).
2. C. Sexton, N. Marchetti, and L. A. DaSilva, ‘Customization and Trade-offs in 5G RAN Slicing,’ in IEEE Communications Magazine, vol. 57, pp. 116-122, Apr. 2019.
3. C. Sexton, Q. Bodinier, A. Farhang, N. Marchetti, F. Bader, and L. A. DaSilva, ‘Enabling Asynchronous Machine-Type D2D Communication Using Multiple Waveforms in 5G,’ in IEEE Internet of Things Journal, vol. 5, pp. 1307-1322, Apr. 2018.
4. C. Sexton, N. J. Kaminski, J. M. Marquez-Barja, N. Marchetti, and L. A. DaSilva, ‘5G: Adaptable Networks Enabled by Versatile Radio Access Technologies,’ in IEEE Communications Surveys & Tutorials, vol. 19, pp. 688-720, Second Quarter 2017.

Conference Papers

1. C. Sexton, M. Butt, N. Marchetti, and L. A. DaSilva, ‘On Matching Users to Specialised MNOs in Service Tailored Networks of the Future,’ in IEEE Global Conference on Communications (GLOBECOM) Workshops, 2018.
2. C. Sexton, Q. Bodinier, A. Farhang, N. Marchetti, F. Bader, and L. A. DaSilva, ‘Coexistence of OFDM and FBMC for Underlay D2D Communication in 5G Networks,’ in IEEE Global Conference on Communications (GLOBECOM) Workshops, 2016.

2 The Trend Towards Customisable Networks

The Trend Towards Customisable Networks

This chapter lays the foundation for the rest of the thesis by demonstrating the need for adaptability in future networks. It begins by highlighting that there has been a trend in the area of Resource Allocation (RA) for the last few decades towards increased service differentiation, culminating in the unprecedented adaptability necessary today to satisfy the diverse and heterogeneous requirements demanded of any future network. We then examine how the Radio Access Technologies (RATs) and system-level techniques that will be used to build future networks both offer unprecedented design choices owing to their suitability to some use-cases, but not others.

2.1 Resource Allocation: A Little Bit of History

In the early days of cellular networks, a Mobile Network Operator (MNO) typically built and controlled its own infrastructure. Spectrum, a limited resource, was the root of an MNO's value-chain [11]. As a result, RA has traditionally been primarily concerned with utilising the frequency bands purchased and licensed for use by an MNO. The Global System for Mobile Communications (GSM), the first truly global standard, approached RA through channelisation of the spectrum.

In GSM, the licensed bandwidth owned by an MNO is divided into orthogonal channels using a combination of Time-Division Multiple Access (TDMA) and Frequency-Division Multiple Access (FDMA) with FDMA used to split the spectrum into 124 carrier frequencies of 200 kHz each and TDMA then used to divide each carrier frequency into burst periods, with eight burst periods per TDMA time frame. The authors in [12] present a general formulation of the resource management problem at the time, identifying three key allocation decisions.

- assigning an access port (Base Station (BS));
- assigning a waveform (channel);
- assigning transmit powers.

GSM was designed with an application to voice, using circuit-switched technology to dedicate resources for the duration of a call and hence guarantee a particular Quality of Service (QoS). With the advent of multimedia services at the time, the need for greater support of data services was realised and a new standard for packet data transmission named General Packet Radio Service (GPRS) was introduced to efficiently utilise radio resources. With different QoS now required by circuit-switched voice and packet-switched data, RA became an important research topic to enable the integration of GPRS and GSM without degrading the QoS of voice services [13], [14]. This was achieved through the dynamic allocation of resources, allowing unused voice channels to be allocated to data temporarily, with voice having a pre-emptive priority over data.

GPRS was considered to be an important development step of GSM towards the next generation of mobile communication systems such as Universal Mobile Telecommunications Service (UMTS). Much of the core network in UMTS remained similar to that of GSM/GPRS, with changes mainly focused on the all-important air interface. Wideband Code Division Multiple Access (WCDMA) was chosen as the radio-access technology for UMTS, allowing multiple users to simultaneously utilise a single frequency band by assigning each user a code with good auto and cross-correlation properties.

The problem of allocating resources to meet different QoS requirements, which was introduced with the integration of GPRS and GSM, was further recognised in UMTS with four service classes categorised: conversational, streaming, interactive, and background [15]. Delay sensitivity is the primary distinguishing feature between classes with the conversational class intended for very delay sensitive traffic while the background class is the most delay insensitive and hence has the lowest priority. The research problem of allocating resources in an efficient manner to satisfy various levels of QoS was addressed by many papers at the time [16], [17], [18].

The latest standard for mobile communications, Long Term Evolution (LTE), reflected the dominance of data traffic and targeted simplifying the network architecture from a combined circuit- and packet-switched network to a flat all-IP network. In addition, having realised the limitations of Code Division Multiple Access (CDMA), Orthogonal Frequency Division Multiple Access (OFDMA) was adopted in the downlink and Single Carrier - Frequency Division Multiple Access (SC-FDMA) was used in the uplink to mitigate problems relating to the high Peak-to-Average Power Ratio (PAPR) associated with OFDMA. RA in OFDMA is achieved by allocating resource blocks to users, where each resource block is 180kHz in the frequency domain and 0.5ms in the time domain.

The allocation of resources in LTE to achieve QoS differentiation is handled using the concept of a *bearer*, which is a virtual connection used to transfer data between two endpoints in the network (such as between a User Equipment (UE) and Packet Data Network Gateway (PGW)). Each bearer has an associated QoS Class Identifier (QCI) which classifies the bearer as either Guaranteed Bit Rate (GBR) or non-GBR. GBR bearers have a predefined minimum long term average data rate, as well as upper bounds on both the packet loss rate and the packet delay budget [19]. No guarantee is provided to non-GBR bearers, making them unsuitable for real-time services such as video streaming. The QCI also specifies a priority used to determine the order in which bearers are handled. Table 6.1.7-A in [20] provides a list of example services and how they may be accommodated using the QCI characteristics described above.

2.2 Future Network Requirements: A Brief Overview

Similar to previous generations, future networks will also introduce new services to be supported and new requirements to be fulfilled. However, it is the diversity of these services and requirements that motivate the need for a different kind of network from previous generations. We explore the different requirements that will have to be considered through the scenarios specified in the METIS project for 5G [21]:

1. *Amazingly fast:* This scenario addresses the requirement to provide very high data rates with low latency. The emphasis here is on instantaneous connectivity. Examples include a virtual reality office which requires the transfer of huge amounts of high-resolution 3D imagery.
2. *Great service in a crowd:* This scenario focuses on providing a reasonably good service to users in crowded locations such as stadiums.
3. *Best experience follows you:* This scenario stresses the importance of providing quality service to users on the move, such as a high speed train.
4. *Super real-time and reliable connections:* This scenario deals with providing very low end-to-end latency and high reliability to mission critical applications such as traffic safety.
5. *Ubiquitous things communicating:* The final scenario focuses on providing connectivity to massive numbers of machine-type devices. Such devices need to have extremely low power-consumption to support battery life of up to a decade.

While the scenarios listed above were described in the context of 5G, they remain relevant in the broader discussion of future networks due to their generality. Although it is not practical to list every possible scenario, the five outlined above provide a strong insight into the level of flexibility needed in future networks. The above scenarios demand the following requirements from the network, again provided in the context of 5G:

- **Edge data rate** In order to support a new class of applications such as high definition video or virtual reality, it is generally agreed that a 100x increase in the edge data rate will have to be achieved. Current edge rates in 4G systems are in the range of a few Mbps, placing the 5G target for the minimum rate that can be expected by the majority of users in a cell in the range of several hundred Mbps.
- **Capacity** The general agreement is that 5G networks will have to be able to handle 1000x the current capacity in order to support the scenarios above. The *Ubiquitous things communicating* scenario sees the addition of massive numbers of Machine-Type Communication (MTC) devices to the network (albeit at much lower data rates per device).

- **Reliability** Mission critical applications demanding very high reliability have traditionally been forced to use wired access. 5G aims to provide pseudo-wired connectivity to any application that needs it, such as those in the field of healthcare.
- **Latency** While the mission critical applications mentioned above often require near-zero latency, so too will a wide range of emerging applications such as real time gaming and virtual reality. We expect near-zero round-trip latency to mean in the vicinity of 1ms, an order of magnitude lower than the current 4G times of around 15ms. [22].
- **Energy Efficiency** As networks grow larger and more dense, energy consumption also grows. Energy consumption in networks contributes significantly to operators' costs and also contributes to CO_2 emissions. Hence, a great deal of emphasis is being placed on energy-efficiency in future networks [23]. Given the ultra-dense networks being considered for future networks, increases in energy-efficiency will be necessary just to keep energy consumption at current levels.
- **Cost** In addition to the above demands, it is necessary to reduce Operational Expenditure (OPEX) and Capital Expenditure (CAPEX) for MNOs. MNOs are expected to provide the latest services to users but are finding it increasingly difficult to get a return on their investment in new infrastructure and technologies, particularly with the increase in the number of over-the-top services which do not contribute to operator revenue [11]. New strategies for extracting value from the network and making it affordable must be devised.

2.3 Radio Access Technologies in Future Networks

Future networks have the potential to offer an unprecedented level of flexibility in the RATs it employs. With so many diverse requirements to satisfy, these Physical Layer (PHY) technologies provide the basic building blocks from which to construct versatile networks that can be adapted according to the services to be supported. Future networks will be characterised by both the specific technologies they adopt, and the ability they offer to configure these technologies to suit particular use-cases.

In this section, we focus on three core areas that will form the main ingredients of any future network PHY layer: duplexing, multiple antenna use, and waveforms.

2.3.1 Duplexing

The notion that radios cannot send and receive simultaneously using the same spectral resources is based on the fact that the locally generated transmitted signal can be several orders of magnitude stronger than the signal to be received, essentially drowning it out and resulting in severe cross-talk between the transmitter and receiver. However, given developments in Self-Interference Cancellation (SIC), In-Band Full Duplex (IBFD) is now feasible for low-power, short-range systems such as small cells [24] and Device-to-Device

(D2D) communication [25], which are expected to play an important role in future networks. The main benefit that IBFD offers is the potential of doubled spectral efficiency and capacity.

Increased spectral efficiency is not the only benefit that IBFD can offer. IBFD can be used to reduce control plane latency, since feedback information such as Channel State Information (CSI) and acknowledgements can be received during data transmission. In addition, advances in SIC enable faster collision detection since a transmitting device can simultaneously listen for collisions. This is of particular interest for contention based protocols or Dynamic Spectrum Access (DSA).

IBFD has great potential; however, understanding when and how to use it is critical to its successful adoption. In this section, we survey the many choices and options presented by IBFD, and explore the flexibility that it introduces into the network.

Hybrid Duplexing

IBFD performance depends on numerous factors such as SIC capabilities, pathloss between devices, and transmit power. In addition, IBFD introduces two new types of interference to the cellular network [26], namely BS-BS and UE-UE. In many cases, the new types of interference introduced into the network prevent the promise of potentially doubled capacity from being realised [27]. In some cases, strong interference may even render IBFD less favourable than conventional duplexing[1] techniques. The conditions for IBFD gain in a single cell scenario are derived in [28], which proposes a hybrid scheduler to decide whether to schedule both an uplink and downlink UE in a resource block, or whether to default to traditional HD[2] operation. This leads to the concept of hybrid duplexing, in which the duplexing scheme is chosen depending on current conditions.

With regard to choosing a duplexing mode, four choices reveal themselves:

1. Time-Division Duplexing (TDD);
2. Frequency-Division Duplexing (FDD);
3. In-band Full Duplex (IBFD);
4. Hybrid Duplexing.

Hybrid duplexing involves a controller which, based on a set of parameters of concern, decides when to exploit IBFD communications and defaults to a conventional duplexing technique if conditions are not favourable.

- Hybrid duplexing for cellular access: The use of pure IBFD may not be optimal in every situation due to the effects of interference, motivating the use of a centralised

[1] Conventional duplexing refers to the use of either Half Duplex (HD) techniques, where simultaneous transmission and reception is not possible, or Out-of-Band (OOB) full duplex, where simultaneous transmission and reception is only possible using different frequency bands.

[2] To be precise, HD refers only to techniques that may either transmit or receive in a certain time slot, but not both. However, in keeping with the literature on hybrid duplexing, we use the term HD instead of conventional duplexing to mean either Time-Division Duplexing (TDD) or Frequency-Division Duplexing (FDD).

adaptive scheduler in a scenario consisting of an IBFD BS and HD UEs [6]. The scheduler may decide to schedule either one uplink, one downlink, or a pair of uplink and downlink UEs in a resource block depending on the interference [29, 30].

- Hybrid duplexing for D2D communication: D2D allows nearby devices to establish direct links, negating the need to make a round trip via the BS and hence increasing the overall system throughput. Research into the coexistence of D2D and IBFD primarily focuses on using IBFD communications between device pairs to boost spectral efficiency [31–33], requiring devices to be IBFD capable. This is achieved at the cost of increasingly complicated interference channels to be considered. Hybrid duplexing in an IBFD D2D scenario may take the form of BS assisted hybrid scheduling, or each individual D2D pair may autonomously decide for themselves. The decision between using full or HD may be influenced by a number of factors related to the interference profile of the cell, including self-interference, D2D-to-UE interference, and UE-to-D2D interference.
- Hybrid duplexing for relaying: The concept of hybrid duplexing is again relevant in this scenario, as highlighted by [34], which considers hybrid IBFD/HD relaying with opportunistic mode selection and demonstrates the performance gain offered by such a system over a system confined to a single duplexing scheme. Hybrid transmission mode for relays can achieve better performance than just using IBFD or HD transmission mode alone [35]. The subject of resource allocation in virtualised IBFD relays is discussed in [36, 37], considering spectrum, BSs, and relays as virtual resources.
- Hybrid duplexing for self-backhauling: Self-backhauling refers to a technique whereby a BS uses part of its available spectral resources for wireless backhauling. The authors in [38] highlight the importance of backhaul-aware radio resource management. This is particularly important in an IBFD-capable small-cell that uses spectral resources simultaneously for both access and backhaul. The idea of hybrid duplexing for cellular access is even more prevalent in a scenario involving in-band backhauling. Furthermore, this concept can be extended to the backhaul case as explored in [39], in which the authors demonstrate the usefulness of adaptive IBFD/HD self-backhauling over IBFD self-backhauling alone. In adaptive IBFD/HD self-backhauling, the duplexing scheme is dynamically changed according to the current interference conditions.
- Hybrid duplexing for DSA: DSA has been heralded as a promising technique to deal with the perceived spectrum shortage at microwave frequencies, allowing unlicensed Secondary Users (SUs) to avail of licensed bands according to a strict set of rules. The rules defining how and when an SU can use licensed spectrum are designed with a strong emphasis on protecting the incumbent. Typically, in a cognitive radio, the SU will perform spectrum sensing at the beginning of each time slot and begin transmitting if the received power is below some predefined threshold. Two problems are evident with this approach. Firstly, multiple SUs might opportunistically attempt to access the medium, resulting in secondary collisions. Secondly, the Primary User (PU) may become active at any time and the SU cannot detect this while it is transmit-

ting. SIC has been proposed to enhance the performance of cognitive radios, reducing the number of SU collisions and offering greater protection to the incumbent, as it allows SUs to perform spectrum sensing while simultaneously transmitting [40–42]. The benefits afforded by IBFD may be utilised by cognitive radio in an adaptive transmission-reception-sensing strategy [43, 44], comprising two-modes:

1. simultaneous transmission-and-sensing mode to improve detection probability;
2. simultaneous transmission-and-reception mode to improve throughput.

The adaptive switching strategy results in a spectrum awareness/efficiency trade-off, with a threshold between the two depending on the SU's beliefs about PU activity. If an SU has a strong belief regarding PU idleness in a certain channel, the SU should operate in simultaneous transmission-and-reception mode. If this belief decreases, the SU should switch to simultaneous transmission-and-sensing mode in order to constantly monitor PU activity while transmitting.

SIC Enabling Flexible Use of Spectrum

One of the greatest advantages that SIC introduces in the context of enabling versatile networks is the potential for network operators to make use of their licensed spectrum as they see fit, with some of the possible ways that SIC can be utilised including any-division duplexing and spectrum virtualisation [45].

- *Spectrum virtualisation:* SIC's ability to isolate any pair of transmit and receive frequencies allows it to act as a software controlled duplexer. IBFD relates to the case whereby the uplink and downlink channels are completely overlapped. SIC allows any two channels to be paired, including partially overlapped channels. A software defined duplexer would simplify the effort associated with supporting fragmented spectrum.
- *Any-division duplexing:* SIC can enhance FDD with increased opportunities to be configurable, allowing it to exploit carrier aggregation. For example, similar to the concept of spectrum virtualisation, SIC enables partially overlapping channels to be paired for uplink and downlink in FDD. This is complemented by IBFD, which allows completely overlapping bands to be paired. The different duplexing possibilities are illustrated in Fig. 2.1.

The introduction of IBFD communication, and more generally the concepts of any-division duplexing and software controlled duplexing, will have implications on the manner in which spectrum is auctioned to the highest bidder and assigned. In [11], the authors highlight the inefficiencies in current practices for allocating spectrum to operators and call for the removal of restrictions on spectrum. One of the restrictions highlighted in the paper, and which is most relevant in this case, is the pre-designation of spectrum as either FDD or TDD prior to allocation. Spectrum to be auctioned is stipulated to either be FDD or TDD irrespective of the services that it will be used to support or expected traffic patterns. Clearly this imposes severe difficulties for the introduction of any-division duplexing. The

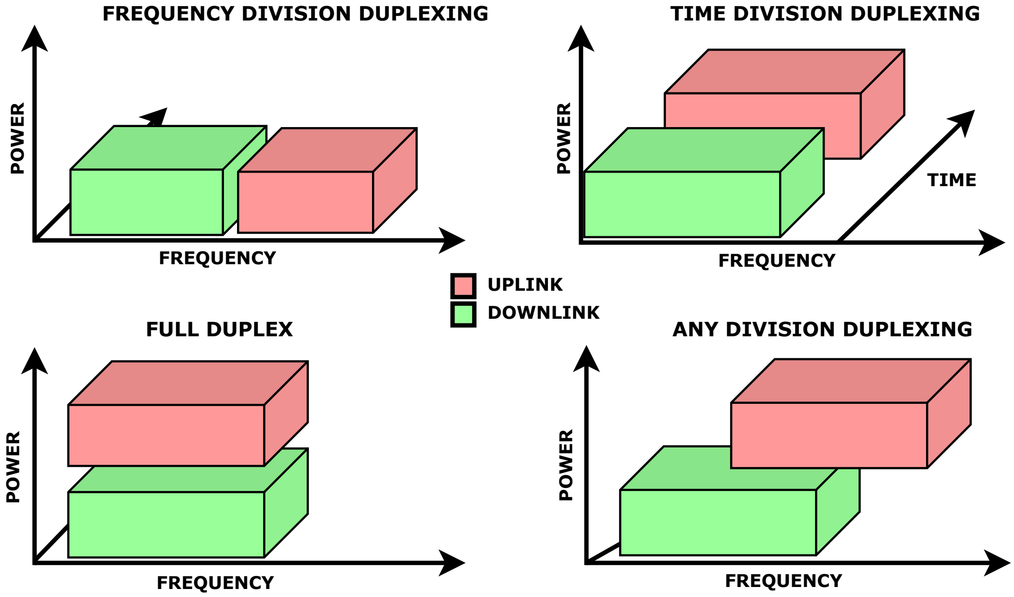


Figure 2.1: Different duplexing options in future networks. Improvements in SIC allows both full and partial overlap of uplink and downlink bands.

advent of new concepts in duplexing such as IBFD and any-division duplexing requires further work in the area of spectrum auctions in order to allow network operators to bid for spectrum irrespective of the duplexing scheme they wish to implement.

2.3.2 Multiple Antenna Use

The ability to utilise multiple antennas comes with inherent choice in how to use them, which directly dictates the resulting benefits. Future network architectures are likely to consist of dense small cell deployments underlaying Massive Multiple-Input Multiple-Output (Massive-MIMO) enabled macro-cells, massively deployed Remote Radio Heads (RRHs) in Virtualised RANs (VRANs), and Distributed Antenna Systems (DASs) (note that these are not mutually exclusive).

The availability of multiple antennas offers many potential advantages to network operators, dependent on how they wish to utilise the antenna resources at their disposal. Capacity, data-rate, and reliability gains are all possible depending on the multiple antenna technique in use. Depending on the service being considered, a network operator may decide to employ techniques including Multiple-Input Multiple-Output (MIMO), spatial modulation, and Coordinated Multi-Point (COMP). Each technique offers varying advantages, as well as different levels of flexibility and customisability. In this subsection, we explore the flexibility and choices associated with multiple antenna use and how it can aid network operators in the creation of versatile networks.

Diversity/Multiplexing Choices

Multiple antenna technologies such as MIMO can be used in two broad formats: diversity for increased reliability, or multiplexing for increased capacity. The decision whether to utilise the multiplexing or diversity gains of MIMO depends on the particular propagation environment, and the priorities of the network operator, who may value reliability over capacity or vice-versa.

1. *Diversity/robustness:* Multiple copies of the signal are received over independently fading channels, increasing the probability that the receiver will be able to detect the transmitted signal without error and, hence, improving reliability.
2. *Spatial multiplexing/throughput:* Spatial multiplexing aims to increase the capacity of a system by sending different signals over the different paths between the transmitter and receiver. Multiplexing is best suited to environments consisting of high multipath in which the various MIMO channels are uncorrelated.

The trade-off between diversity and multiplexing gains offered by MIMO systems is a well researched topic in literature. The authors in [46] demonstrated that both diversity and multiplexing gains could be simultaneously obtained, with a fundamental trade-off between the two. Since then, there has been a wealth of research into the diversity/multiplexing trade-off for MIMO systems. For example, in [47], the authors suggest a framework for devising practical adaptive MIMO architectures, focusing on switching between three MIMO schemes: diversity, hybrid diversity/multiplexing, and multiplexing. In the context of adaptable future networks, it would be beneficial to let the network operator control the diversity/multiplexing gain through adaptive precoding.

The diversity/multiplexing trade-off is already incorporated into LTE, which was designed to natively support the use of multiple antennas in both BSs and user devices, with both diversity and multiplexing configurations permitted. The receiver measures the channel elements and calculates the rank indication, which indicates the number of symbols it can successfully receive. The receiver then feeds back the rank indication to the transmitter. Spatial multiplexing works best in high-scattering environments when the channel elements are highly uncorrelated with each other, allowing the receiver to separate the received symbols from each other [19].

While adaptively switching between diversity and multiplexing may be currently implemented in LTE, its usage is relatively basic. The advent of vastly greater numbers of antennas in future networks, both distributed and co-located, introduces many new challenges and considerations in this area, ensuring that adaptive switching between multiplexing and diversity will remain relevant.

Adaptive Spatial Modulation (ASM)

Spatial modulation (SM) is a MIMO technique which extends traditional digital modulation techniques such as Quadrature Amplitude Modulation (QAM) into the spatial domain. In SM, only one transmit antenna is active at any time, with the index of the transmit antenna used to convey information. Blocks of bits are mapped to both a symbol from the constellation diagram, and a unique transmit antenna number chosen from the set of possible transmit antennas. Spectral efficiency is increased by the base-two logarithm of the number of transmit antennas.

SM takes advantage of the uniqueness and randomness properties of the wireless channel, since each antenna in the possible transmit antenna set will experience different channel conditions. The receiver can then determine the transmit antenna index, which is used in demodulation. Since only one transmit antenna is active at any one time, SM can be considered to be a type of single RF-chain MIMO. This results in a greatly reduced complexity compared to conventional MIMO, which requires an RF-chain per antenna.

Spatial modulation offers yet another way of utilising multiple antennas, representing a new type of modulation and bringing new challenges in this respect. Similar to the idea of Adaptive Modulation and Coding (AMC), which adapts the coding rate and constellation size according to channel conditions, ASM [48, 49] aims to dynamically adapt the modulation order assigned to the transmit antennas according to the channel quality. As illustrated in [50], the fundamental trade-off in adaptive spatial modulation is between constellation size and the number of transmit antennas. In poor channel conditions, a small symbol constellation size is required as the distance between symbols is reduced. However, the poor channel may result in highly uncorrelated antennas, allowing the number of transmit antennas to be increased. Conversely, in good channel conditions, a larger symbol constellation and small number of transmit antennas may be preferable. Therefore, dynamic link adaptation has an important role to play in adaptive networks utilising SM.

Adaptive Precoding

Precoding is a core concept in MIMO systems and refers to maximising the signal quality at the receiver by applying appropriate weightings at each antenna to the multiple data streams being transmitted. Precoding essentially takes advantage of channel state information at the transmitter to perform processing on the signal before transmission. Techniques can be divided into linear and non-linear. Non-linear techniques such as Dirty Paper Coding (DPC) achieve the channel capacity at the cost of high complexity. Linear techniques, such as zero-forcing, block diagonalisation, and maximum ratio transmission (MRT), are less performant but come with reduced signal processing complexity.

In the case where the number of antennas is significantly greater than the number of users, as is the case in Massive-MIMO, simple linear precoders are close to optimal under favourable propagation conditions [51]. However, as demonstrated in [52], this does not hold true when realistic array deployment (taking the physical separation of antennas in account) is considered and there in-fact remains a performance gap between linear and non-linear precoding for dense large scale arrays. This fundamental performance/complexity trade-off naturally leads to the concept of adaptive precoding. In this case, antennas become a fundamental building block for networks, with network operators possessing the power to decide how to use them and what precoding techniques to employ.

One currently existing example of adaptive precoding is the precoding matrix indicator (PMI) in LTE, which is passed from the receiver to the transmitter. The PMI controls the precoding step in the transmitter if diversity is selected, and prevents symbols from cancelling each other out at the receiver by controlling the phase shifts of the transmitted symbols. Adaptive precoding also enables the adaptive switching between diversity and multiplexing techniques. Typically, the UE selects the best precoder from a predefined precoder codebook that maximises the transmission rate for a particular MIMO channel, and feeds this information back to the BS. The precoding choice may also depend on many factors including the number of users to be served, the number of antennas in the array, the signal processing complexity budget, and channel statistics.

Inter-Cell Interference Cancellation (ICIC)

In dense deployments of small cells, inter-cell interference becomes the limiting factor. ICIC techniques such as COMP aim to convert this potential interference into useful signals. COMP refers to a collection of techniques that involve coordination between multiple BSs/antennas during transmission and/or reception to improve the service provided to cell-edge users.

COMP requires coordination between multiple BSs in order to mitigate inter-cell interference and potentially form useful signals. COMP is generally categorised into two main groups.

1. Joint Transmission (JT)/Joint Reception (JR): In the downlink, data is transmitted from each BS in the serving group simultaneously in order to boost the signal strength at the receiver. In the uplink, each BS in the serving group receives the signal from the UE. Signals from each BS are then combined and jointly processed. Data must be shared between each BS, placing increased load on the backhaul between cells.
2. Coordinated scheduling: In the downlink, data is transmitted from only one BS in the serving group to the receiver at any time instant. In the uplink, cooperating cells schedule which BS will receive the data. Scheduling is coordinated among cells in the serving group to mitigate interference and select the BS that can offer the best service to the UE. This reduces the load placed on the backhaul between cells as data does not need to be available in each cell, only channel state information and scheduling decisions are shared among cells.

It is apparent that multiple antenna use involving COMP techniques affords choice to network operators, particularly in small cell architectures and DAS. It is the prerogative of the network operator to decide whether they wish to employ COMP techniques, or not, and if so, choose between JT/JR or coordinated scheduling/beamforming. Taking advantage of the benefits of COMP involves choosing suitable clusters of cooperating BSs. These clusters may be assigned in a static or dynamic manner, possibly requiring the network operator to perform frequent re-selection.

In the case of JR and JT, the question of which entity performs processing is also relevant. Processing may be centralised, maximising the load placed on the backhaul, or it may be distributed among BSs in the cooperating set.

Distributed Antennas

The availability of large numbers of antennas in future networks presents network operators with many options. In dense deployments, the cost of acquiring additional antenna nodes may be cheap and antennas may be considered a resource. Below, we examine three ways in which distributed antennas may be used:

- *DAS:* In a DAS, antenna elements are separated spatially and are connected to a common controller. The principal idea is to extend the coverage of a BS by distributing antennas throughout the environment, retaining the same power budget so that each antenna transmits with reduced power. DAS is popular in indoor environments, in which antennas distributed throughout a building connect to a macro BS (often located on the roof) and serve as repeaters to improve indoor coverage.
- *Distributed MIMO:* MIMO can consist of co-located[3] antennas as part of the same physical array, or distributed throughout the environment. The authors in [54] assert that distributed MIMO systems can achieve higher diversity gain compared to co-located MIMO, as co-located antennas may experience a similar scattering environment. Concepts such as adaptive precoding and the diversity/multiplexing trade-off for MIMO systems, which were discussed previously in the section, remain relevant here.
- *COMP:* In interference limited environments, network operators may use its distributed antennas to avail of the benefits of ICIC techniques such as COMP.

2.3.3 New Waveforms

One of the defining characteristics of each generation change has been the question surrounding the choice of modulation format and Medium Access Control (MAC) strategy. Orthogonal Frequency Division Multiplexing (OFDM) and OFDMA were chosen to be the

[3] The word ‘co-located’, as used in [53], represents antennas on the same array, as opposed to distributed throughout an environment.

modulation format and MAC strategy respectively for LTE due to the advantages they offered over the CDMA systems used in the preceding generation, including higher spectral efficiency and efficient realisation using Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT) blocks.

OFDM's ease of implementation, and its inertia, have ensured that it has maintained its place in the 5G New Radio (NR) standards. However, despite its advantages, OFDM's place in future networks may be challenged by new techniques [55] that aim to deal with some of its shortcomings such as:

1. Large OOB transmissions, resulting in interference issues. This also adversely affects the ability of carrier aggregation to exploit non-contiguous spectrum, a topic that is likely to play an important role in future networks.
2. High sensitivity to synchronisation errors and Doppler shift. The European FP7 research project 5GNOW deems it essential to introduce waveforms that are less sensitive than OFDM to frequency misalignments [56]. In [57], the authors demonstrate that the high sensitivity of OFDM to frequency offsets in a multi-user scenario requires advanced interference cancellation techniques, in turn leading to complex yet low performance systems. Thus, one of OFDM's main advantages in the form of simplicity is lost.
3. Although we listed OFDM's spectral efficiency as an advantage, that was in comparison to previous generations, and there is potential for new techniques to further improve upon this. In particular, the need for a Cyclic Prefix (CP) in OFDM and the large side-lobes at spectrum edges reduce its spectral efficiency.
4. The strict synchronicity demands of OFDM introduces a substantial control overhead in the network. In particular, the emergence of MTC as a major topic introduces new considerations in this area. With the introduction of massive numbers of devices to the network, coordinated access would generate huge signalling overhead, potentially flooding the radio access network. In this regard, a strong case is being made for techniques that facilitate uncoordinated access.

As a result, future networks see a variety of candidate waveforms competing to satisfy the myriad of scenarios and requirements mentioned in Section 2.2. Filter Bank Multi-Carrier (FBMC) schemes aim to achieve higher spectral efficiency than OFDM by suppressing large side-lobes through per-subcarrier filtering, and negating the need for a CP by using narrow channels with flat gain. Universal Filtered Multi-Carrier (UFMC), also known as UF-OFDM, applies filtering to groups of adjacent subcarriers. This idea is based on the observation that asynchronicities tend to occur at block edges, while orthogonality can be maintained within the block itself. Due to the development of equalisers that approach OFDM in complexity, Single Carrier Modulation (SCM) may be, as the authors of [58] suggest, a technique whose time has come again. The main potential for SCM in future networks would be in low latency applications, since delays related to the block processing of data can be avoided [55]. Generalised Frequency Division Multiplexing (GFDM), first introduced in [59], is a multi-carrier modulation scheme with flexible pulse shaping that targets low OOB emissions and frequency agility.

Designing a Flexible Air Interface

This idea of a reconfigurable air interface is explored in [60]. The Software-Defined Air Interface (SDAI), enabled by Software Defined Radio (SDR), consists of an intelligent controller and multiple configurable fundamental building blocks such as the frame structure, waveform, multiple access, modulation and coding, etc. Different services can be supported using different configurations of the fundamental building blocks, which are controlled through software.

In terms of designing an adaptable network, the SDAI concept offers numerous advantages. Through SDR, many aspects of the air interface become configurable, allowing the network to be tailored towards different services. In contrast to an air interface optimised for a single application, we instead have a fluid and adjustable system. Achieving reconfigurability in every facet of the air interface presents several challenges. Current LTE networks already implement a form of adaptability through AMC, in which the coding rate and modulation scheme are chosen according to the link quality. We have already discussed adaptive duplexing and multiple antenna use in the previous subsections. In this subsection, we focus our attention on the multiple access and waveform choice.

Although there are many waveforms being considered for future networks, each presents advantages and disadvantages depending on the scenario under consideration. For example, SCM techniques may lend themselves to low latency applications since they do not incur the delays associated with the block-processing of data. FBMC, on the other hand, may be preferable in an MTC scenario as it facilitates asynchronous access [56]. Hence, an adaptable and flexible solution is required in relation to the choice of waveform and multiple access technique. Below, we explore some of the possible different strategies involving the selection of one or more waveforms for future networks that would lend themselves to the goals of flexibility and versatility.

Single Waveform - Adjusting Parameters of a Configurable Waveform

This option advocates the standardisation of a single configurable waveform, which can be tweaked through tunable parameters. We can begin with a single malleable waveform and mould it according to our needs. This concept is best described as a Software-Defined Waveform (SDW), an idea that resonates with many of the current trends in future networks such as Software-Defined Networking (SDN) and SDR. This option relies heavily on the softwarisation of the Radio Access Network (RAN) in order to be able to present configurable waveform parameters which can be adjusted according to the scenario to be supported. SDR is therefore an enabling technology, and the concept of a configurable waveform fits with the previously described concept of an SDAI. The configurable parameters form a numerology for the waveform, which define it for a particular use case. The content of this numerology, i.e. the parameters themselves, depend on the base waveform in use.

This idea of tweaking the parameters of a waveform according to the use-case is hinted at in [22], in which the authors envision a type of tunable OFDM. In this vision, OFDM would permit configuration through tunable parameters such as subcarrier spacing, CP length, and FFT block size. For example, user specific subcarrier spacing and symbol period is considered in [61], which compares several variants of OFDM which employ either a CP or zero post-fix as a guard interval to prevent Inter-Symbol Interference (ISI), and use either windowing or filtering.

Filtered Orthogonal Frequency Division Multiplexing (f-OFDM) is presented as an enabler for a flexible waveform in [62]. In this vision, with f-OFDM, the assigned bandwidth is divided into several sub-bands. Each sub-band employs OFDM with a numerology tailored to satisfy a particular service. The parameters of such a numerology may include subcarrier spacing, CP length, and Transmission Time Interval (TTI). Asynchronous transmission across sub-bands is supported through sub-band-based filtering.

GFDM is identified as a promising solution for future networks in [63] as a result of its inherent flexibility. In particular, the authors in [63] show how GFDM can be tailored according to several broad scenarios such as Bitpipe, MTC and Tactile Internet, by adjusting a set of GFDM parameters. The need for a flexible PHY layer in future networks and a waveform with many degrees of freedom is highlighted in [64], and a flexible FPGA implementation of GFDM that permits run-time reconfiguration is proposed. Multiple applications can be supported through configuration of several parameters such as filter coefficients, the number of subcarriers in a block, and the number of sub-symbols per subcarrier.

The primary aim here is to use a configurable waveform to expose PHY flexibility to higher layers. The role that techniques at these higher layers perform, and the manner in which they interact with the PHY layer, is critical to the successful implementation and adoption of an SDW vision. Clearly the concept of SDW lends itself to a coupling with techniques such as SDR and SDN, such as the possibility of incorporating SDN and a centralised controller which defines the set of parameters for the waveform to be used for a particular scenario.

Multiple Waveforms - Selecting from a Pool of Waveforms

Future networks may permit the coexistence of multiple waveforms. Given a choice of waveforms, each suited to different use-cases, the waveform itself can be viewed as an addition to the resource pool. Different applications or services may benefit from the use of different waveforms, according to their specific requirements. For example, clustered D2D pairs underlaying an OFDMA macro-cell may use a different waveform such as FBMC in order to reduce the leakage interference between devices. Mission critical applications such as vehicular traffic safety may require ultra low latency, and hence may use a waveform or frame structure capable of supporting short TTIs. We are therefore motivated to investigate how multiple waveforms may impact upon one another and ultimately coexist.

Several works investigate the coexistence of various waveforms by characterising the cross-waveform leakage interference. The authors in [2] consider a scenario consisting of asynchronous D2D communication overlaying an OFDMA macro-cell, and investigate the benefits of D2D pairs adopting alternative waveforms to OFDM. The authors generate interference tables characterising the interference from several new waveforms onto an OFDM receiver. In [65], the authors investigate D2D communication in an OFDMA/SC-FDMA based cellular network, in which D2D pairs may use FBMC to reduce interference. The limitations of using a Power Spectral Density (PSD) based model when evaluating the interference between OFDM/Offset Quadrature Amplitude Modulation (OQAM) and CP-OFDM is considered in [3], with the importance of considering demodulation effects at the receiver emphasised.

There is a shift in emphasis involved in this vision from standardising a particular waveform that all future networks must use, to standardising a set of procedures and protocols that allow network operators to choose a modulation scheme from a set of possible candidates.

Adjusting the Multiple Access Procedure and Level of Synchronicity

This option proposes choosing a single waveform that is suitable for many applications, and using it with different synchronisation procedures and access methods.

At the beginning of this section, when discussing the motivation behind researching alternative waveforms, we discussed how the strict synchronicity demands of OFDM introduce a substantial control overhead in the network when a large number of MTC devices is considered. MTC is characterised by high-volume sporadic traffic consisting of short packet sizes, indicating that it may be best served using contention-based access with relaxed synchronism. Hence, we first explore the possibility of using a frame structure that can support different access procedures, as well as different levels of synchronicity and orthogonality. Classical bit-pipe traffic can be served using scheduled access with strict orthogonality and synchronism. MTC traffic, on the other hand, may use contention-based access and abandon synchronism in order to reduce the signalling overhead.

We examine the findings of the 5GNOW project, which advocated the adoption of a unified frame structure to satisfy the various traffic types [56] to be supported in future networks. The concept of a unified frame structure is also described in [66], in which the authors advocate for the use of UFMC. The unified frame structure aims to be flexible and scalable, incorporating a mix of synchronous/asynchronous and orthogonal/non-orthogonal traffic types. In total, four traffic types are defined, with each targeting a different class of application or service. Each traffic type uses an access procedure, and level of orthogonality and synchronism, appropriate for the traffic that it accommodates.

Three of the traffic types abandon synchronism and hence do not incur the overhead and energy required by a closed-loop synchronisation procedure. Instead, these traffic types could achieve coarse time-alignments by listening to the downlink, in an open-loop synchronisation procedure. The use of Automatic Timing Adjustment (ATA) [67], whereby devices estimate their propagation delay in an open-loop procedure and adjust their transmission timing to compensate, may also be adopted [66].

The concept of a unified frame structure demonstrates how various scenarios can be handled by a single waveform by altering the access procedure (scheduled/contention), and the level of synchronicity (closed-loop/open-loop with ATA). A flexible frame structure is also discussed in [68] and [69], which supports the dynamic adjustment of the TTI according to the service requirements of the link. Given the targeted 1ms latency support for mission critical applications in future networks, TTIs of no more than 0.2-0.25ms are required. Hence, latency critical links may benefit from a small TTI in the flexible frame structure.

Another area in multiple access that has been gaining traction recently is Non-Orthogonal Multiple Access (NOMA). The conventional multiple access schemes used in previous generations, such as TDMA in 2G, CDMA in 3G and OFDMA in 4G, are all orthogonal multiple access schemes, allocating orthogonal resources in either the time, code, or frequency domains. In contrast, NOMA uses non-orthogonal resource allocation to accommodate larger numbers of users. Interestingly, the authors in [70] propose the concept of Software Defined Multiple Access (SoDeMa), which can support diverse services and applications through adaptive configuration of available multiple access schemes. This resonates with the aforementioned idea of a SDW, and highlights the ongoing trend towards softwarisation of the network in response to the need for greater control and versatility.

2.4 System-Level Techniques for Future Networks

Having obtained a clearer idea in Section 2.2 of the scenarios and requirements to be satisfied in future networks, and surveying the array of RATs and their capabilities in Section 2.3, we now take a system-level view of the network. We choose to focus on SDN and VRAN, not only because they represent two of the largest topics in this area, but also because of the potential they possess in the context of enabling adaptable networks. Both techniques aim to achieve a higher level of abstraction in the network, which brings an inherent increase in flexibility and the ability to dynamically control resources. SDN abstracts network control into a logically centralised controller, decoupled from the data plane. VRAN abstracts processing power into separate pools of resources that can be dynamically assigned as needed to RRHs.

2.4.1 SDN

SDN resonates with the trend towards increased softwarisation, and offers the ability to dynamically alter the flow of traffic through a network by decoupling the control plane from the data plane, allowing centralised control over the behaviour of the entire network. The rules for handling data can now be specified in software at the controller, which communicates with the data plane (i.e. switches, routers) through an open interface. As a result, it is possible to alter the entire behaviour of the network from a single logical point without needing to physically touch the hardware.

The essence of SDN is possibly best characterised by four of its core principles [71, 72]:

1. *Decoupling of control and data planes:* This principle is the foundation of the SDN concept, advocating the separation of the control plane into a logically centralised software controller which is capable of managing and altering the routing of data through the network.
2. *Logically centralised controller:* The extracted control plane is logically centralised into a single controller with a network wide view. This logically centralised controller may in fact consist of multiple virtual or physical controllers operating in a distributed manner, depending on the scale of the network.
3. *Open interfaces:* One of the motivating factors behind SDN was to reduce the effort and cost associated with reconfiguring the vendor-specific devices in the network. An open, standardised interface between devices in the control and data planes, known as the southbound Application Programming Interface (API), is therefore a key principle of SDN.
4. *Programmability by external applications:* The controller in SDN allows for programmability by external applications through the so-called northbound API. This naturally lends itself to the concept of adaptability. It allows the network operator to view the myriads of physical hardware under its control as a single programmable entity which it can configure.

In this section, we are primarily concerned with how SDN can be used to increase the versatility of future networks, and create and manage adaptable networks using RATs as building blocks. SDN offers potential in this regard in the following ways:

- Wireless SDN: SDN itself is inherently adaptable, introducing greater abstractions into the network by decoupling the control and data planes. The flow of data through the network can be altered through programmable controllers. We first explore the flexibility that SDN introduces by examining its application in a wireless context, noting that SDN has thus far mainly been researched in the wired domain.
- Slicing: Slicing refers to partitioning resources and isolating the traffic between multiple coexisting virtual networks.[4]

[4] As has become common parlance when discussing virtualisation, we use the term slice to refer to a

- Gathering of statistics: SDN can be used to gather usage statistics and obtain a global view of the network. From a virtualisation point of view, it allows the Mobile Virtual Network Operator (MVNO) to make informed decisions about the management of virtual resources.

Wireless SDN

The application of SDN to the wireless domain in mobile networks is explored in [73], which discusses some of its potential use cases and benefits including virtualisation and Quality of Experience (QoE)-Aware Network Operation. The authors also describe a generic software-defined wireless network architecture using the 3GPP Evolved Packet System as a reference. In the proposed architecture, the southbound interface now connects to three types of entities: user plane entities in the core network, user plane entities in the RAN, and mobile nodes. The authors in [74] provide a useful survey into the primary trends and ideas involving SDN in the context of wireless networks, partitioning the literature into three main target areas: Wireless Local Area Networks (WLANs), cellular, and multi-hop wireless networks. Most research on wireless SDN has so far focused on WLANs, with virtualisation and the ability to slice the network present as recurring themes. Research in the cellular area is divided into both the RAN and the core network.

The authors in [75] make first steps in exploring the application of SDN in cellular networks by proposing extensions to existing controller platforms and switches that enable high-level policies to be enforced based on subscriber attributes. A scalable architecture employing SDN concepts called SoftCell is employed in [76]. SoftCell supports high-level service policies based on subscriber applications through fine-grained packet classification, which is performed at the access edge. In contrast, the authors of [77] focus on the RAN and propose SoftRAN, a software-defined centralised control plane for RANs. SoftRAN introduces the concept of a virtual big-base station which is an abstraction consisting of a central controller and all of the physical BSs in a given geographical area. This permits effective load-balancing and interference management within the encompassing area. An architecture for mobile carrier core networks based on SDN principles is presented in [78], detailing its development and the use of a proof-of-concept prototype. Interestingly, the authors highlight the potential that a software-defined mobile network provides in terms of enabling innovation and permitting the creation of any network type on-demand, which are two focuses of this chapter.

Slicing Networks

In the context of this discussion, we consider OpenFlow due to its current dominance in the SDN landscape. In an OpenFlow architecture, data forwarding devices are considered

virtual network instance. A slice comprises the virtual resources that constitute that particular virtual network.

to be switches and routers. OpenFlow-enabled switches consist of flow tables which are used to match particular data flows, garner statistics on each flow, and specify how they should be handled. The flow table entries are controlled by the SDN controller through a standardised southbound API. SDN, as implemented by OpenFlow, therefore consists of three main components:

1. packet matching for flow-based routing;
2. reporting of flow statistics for global network view;
3. traffic isolation between different virtual networks.

Slicing refers to the task described in the third point above. In OpenFlow, this task is performed by a unit called FlowVisor [79]. FlowVisor sits logically between the SDN controller and the SDN-enabled device, and ensures that the controller can only alter flows belonging to its own virtual network. It therefore helps satisfy the core virtualisation principle of isolation. In order to achieve this, FlowVisor partitions the flow-table, assigning a number of flow-entries to different virtual networks. It also partitions bandwidth resources by setting limits on the data rate of a set of flows for a particular slice. FlowVisor acts as a proxy between OpenFlow enabled hardware and multiple SDN controllers belonging to different virtual networks, using the OpenFlow protocol to communicate with both controllers and hardware. From the controllers' viewpoint, it appears as if they are communicating directly with the hardware.

Gathering Statistics

In addition to managing the forwarding plane, the OpenFlow protocol also permits per-flow counter statistics to be requested from OpenFlow enabled switches. Network monitoring can therefore be achieved through the addition of a monitoring module in the controller which gathers statistics. Controllers periodically query switches for flow statistics, resulting in a trade-off between accuracy and network overhead. The ability to collect per-flow statistics in SDN has been the focus of several works in the literature [80–83]. For example, OpenNetMon [83] is an open-source software implementation that provides monitoring of per-flow metrics such as throughput, delay, and packet loss, and can be used to determine whether end-to-end QoS parameters are actually met. OpenNetMon was written as a module for the OpenFlow controller platform POX, a Python-implemented platform targeting research and education.

From a virtualisation point of view, the availability of statistics allows the MVNO to identify underutilised virtual resources that can be released, and request additional virtual resources in places where the network is over-loaded. For example, in an area where multiple virtual BSs are reporting low usage, the MVNO may decide to release some of its virtual resources. Conversely, in an area where multiple virtual BSs are reporting high usage, the MVNO may decide to acquire more virtual resources in that area.

2.4.2 VRAN

Traditional RAN architectures consisted of a Baseband Unit (BBU) located at the BS, with either a co-located radio unit located beside the BBU or a RRH located a short distance away on the tower itself. In contrast, VRAN seeks to decouple baseband processing power from individual sites, allowing it to be flexibly allocated to parts of the network experiencing the most demand, and also permitting new services to supported in a time and cost efficient manner. The progression from the traditional RAN architecture to the now standardised VRAN architecture comprised many intermediary proposed splits, with the concept of Cloud Radio Access Network (Cloud-RAN) capturing the core split between local radio units at individual BSs and shared centralised processing power.

The Cloud-RAN paradigm proposed splitting the RAN into three components:

1. *BBU pool:* The BBU pool performs baseband and packet processing, separating and migrating this functionality from individual BS sites to a centralised location. One of the motivating factors for Cloud-RAN is the so-called tidal effect, in which the traffic experienced by a particular BS fluctuates both temporally and spatially as users travel to and from work each day. In Cloud-RAN, BBUs can be dynamically assigned to overloaded areas as required, in accordance with the tidal effect.
2. *RRH:* RRHs can be considered dumb compared to current BSs, as processing capabilities have been abstracted away. RRHs simply transmit and receive signals, perform analog-to-digital and digital-to-analog conversion, and send signals to/from the BBU pool for processing.
3. *Fronthaul link:* The fronthaul connects the RRHs and BBUs. Due to the large bandwidth requirements, optical fibre is generally used in the fronthaul. Signals are transmitted as either digitised radio signals over Common Public Radio Interface (CPRI), or analog signals using Radio over Fibre (RoF). Wavelength division multiplexing is used to separate signals. CPRI is more robust than RoF over long distances as it suffers less degradation; however, this advantage comes at the cost of increased bandwidth requirements.

At this point, it is worth distinguishing between a Centralised RAN and Cloud-RAN. In a Centralised RAN, the BBU pool still consists of proprietary hardware and is limited in its flexibility, where as Cloud-RAN supports commercial off-the-shelf hardware which can be virtualised. VRAN further disaggregates this virtualised BBU pool into Distributed Units (DUs) and Centralised Units (CUs) as follows:

1. *DU:* DUs are placed closer to the RRHs so that they can meet the stringent latency and higher bandwidth requirements required by real-time functions in the protocol stack. Each DU supports multiple RRHs. Although many functional splits are possible, low-level physical layer functions are typically confined to the RRH, with higher-level physical layer, MAC and Radio Link Control (RLC) functions performed by the DU. The fronthaul link between the RRH and the DU will adopt Evolved

Common Public Radio Interface (ECPRI) which is an open interface evolution of CPRI which is capable of supporting higher throughput on less fibre.

2. *CU:* CUs support multiple DUs and support non-real-time functions, typically in the upper layers of the protocol stack. CUs and DUs are connected by the F1 interface, defined in [84], known as mid-haul.

One of the most apparent benefits of VRAN is its adaptability to non-uniform traffic. In the traditional network architecture currently employed, BSs are designed to handle peak traffic loads, which can be several times higher than normal usage. VRAN benefits from statistical multiplexing gain by dimensioning the processing capacity of the DU and CU pools to be less than the sum of the capacities of individual BSs. This is motivated by the fact that BSs in different areas experience peak load at different hours of the day. Hence, VRAN can adapt to traffic fluctuations throughput the day by permitting overloaded BSs to use more processing power.

VRAN provides numerous other advantages. By confining radio functions to RRHs and centralising processing across DUs and CUs, the cost of deploying additional radio heads to improve coverage is now reduced - an advantage which will be hugely beneficial in the ultra-dense networks envisioned for future networks. Better energy efficiency can be achieved as processing power can be dynamically allocated and BSs can be turned off when not needed. The VRAN paradigm also facilitates the sharing of information between cooperating BSs, leading to improved spectrum utilisation [85].

The VRAN concept alters the manner in which resource allocation is performed. Processing power is now a resource to be allocated as needed. In addition, cooperation between RRHs can be realised as the CUs and DUs have access to the channel state information and other information supplied by neighbouring RRHs. The ability to treat both RRHs and processing power as resources offers great potential in the pursuit of creating a flexible, adaptable network. In this subsection, we explore how this potential may be realised.

Flexible Functional Splits

One of the important questions in VRAN surrounds the functional split of processing, i.e. which functions should be implemented locally at the radio head site, and which should be handled remotely in the processing pool. With regards to the processing pool, there is a further question regarding which functions should be handled in a DU and which should be handled in a CU.

The various split options have differing requirements for the fronthaul in terms of both bandwidth and latency. Several possible splits of the LTE baseband processing chain are analysed in [86], taking into account bandwidth and latency requirements. The authors in [87], on the other hand, focus on the opportunities provided by a flexible split, detailing the advantages and disadvantages of several options.

In summary, the more lower-layer functions that are moved into the centralised processing pool, the higher the demands are on the fronthaul in terms of latency and capacity. Dividing the processing pool into DUs and CUs helps alleviate this issue, with real-time functions placed in DUs. As a result, CUs show greater potential for virtualisation and resource sharing.

The O-RAN alliance is a consortium of telecommunication companies which is focused on the development and specification of an open, vendor-agnostic, interoperable RAN. Defining functional splits supported by open interfaces so that products from multiple vendors can be utilised in a single network is an essential part of the O-RAN vision.

The possible functional splits between the radio unit, DU and CU are outlined in the Intel white paper [88]. In summary, the desired outcome is to achieve a split that ensures that functions with stringent latency and bandwidth requirements are handled close to the BS, while ensuring that there is still plenty of scope to reap the benefits of virtualisation by using commercial off-the-shelf hardware in a centralised processing pool for higher-layer functions.

Eight possible functional split options were specified by the 3GPP in [89]. The option adopted by the O-RAN alliance in their specifications is known as Split Option 7-2x, in which radio-frequency and lower-level PHY functions are handled in the RRH, and higher-level PHY up to RLC are handled in the DU (incorporating resource element mapping). Functions in layers above RLC are then handled in the CU. One advantage of this split is that it keeps RRH costs low, allowing them to be easily deployed and treated as a resource.

VRAN and PHY Flexibility

The flexibility of the CUs and DUs, enabled by SDR, allow MNOs to customise the low-level details of their network according to their needs. This could allow an MNO to control the choice of duplexing method or waveform through the software in the CUs and DUs. In essence, VRAN isolates the radio head as a fundamental building block, and offers the means to customise PHY operations through the flexibility of the CUs and DUs. We briefly examine the relationship between VRAN and each of the three PHY technologies discussed in Section 2.3.

1. *Waveforms:* In Section 2.3.3, we highlighted the possibility that future networks may permit the coexistence of multiple waveforms. This would allow a choice of waveform depending on the use case being served. This involves a change in thinking, from standardising a single waveform that all future networks must use, to standardising an interface that allows MNOs to choose any modulation scheme as needed. VRAN makes this possible, as CUs and DUs can be configured to use any modulation scheme through the advent of softwarisation.
2. *Duplexing:* We propose to allow the MNO to fully control the choice of duplexing scheme. VRAN again permits this vision, with the duplexing scheme custom-

isable through software in the CUs and DUs. The MNO may be presented with many choices including which duplexing scheme to use, and choosing which bands to pair for uplink and downlink. If the radio head has SIC capabilities, the MNO must also decide how to utilise them; increased spectral efficiency or reduced control plane latency are both possible, as outlined in Section 2.3.1. The advantages of a Cloud-RAN architecture coupled with IBFD communications is outlined in [90], particularly in mitigating the BS-to-BS/downlink-to-uplink interference introduced by IBFD. The centralisation of processing allows the BBU to perform cancellation of the BS-to-BS/downlink-to-uplink interference since the downlink signal of neighbouring RRHs is known by the BBU. These advantages of a Cloud-RAN architecture are also relevant for VRAN.

3. *Multiple Antenna Use:* Of the system-level techniques considered in this chapter, multiple antenna use offers the most apparent links with VRAN. VRAN can enable the flexibility that multiple antenna use affords. VRAN abstracts the actual antenna from the associated processing, allowing both processing power and radio heads to be viewed as resources. RRHs then form the basic building blocks for the network, while the decoupled CUs and DUs allow the MNO to utilise the RRHs whatever way they wish. In the case of distributed RRHs throughout an environment, the MNO may wish to employ antennas for either distributed MIMO or COMP. The centralised aspect of VRAN also permits coordination between selected RRHs, particularly for the JT and JR option in COMP. In effect, VRAN is a direct realisation of COMP.

2.5 Summary

In Section 2.2, we highlighted that the requirements for future networks are extremely diverse, requiring a versatile network capable of adapting to the service demands placed on it. There are a multitude of technologies being considered to meet these demands, each varied in its advantages and disadvantages. In Section 2.3, we surveyed some of the new RATs being considered for future networks in the context of the choices and flexibility they afford. In effect, given the wide range of service requirements, new techniques may only offer advantages in certain scenarios. The role that new system-level techniques have to play in both directly introducing greater flexibility, and managing adaptable networks is very important.

While we may have an idea of the scenarios to be supported and the technologies that may potentially be beneficial, the final constitution of any future network is still unknown. What is clear, however, is that future networks will need to be much more adaptable than previous generations. The technologies and techniques discussed in this chapter are likely to play a role in the future in some shape or form. Based on the current literature, we have extrapolated research trends in order to present a survey of the possible ways that the chosen technologies can facilitate an adaptable, versatile network. In the next chapter, we review a technique that can allow all of these new technologies to be combined to offer tailored network behaviour as-needed.

3 Background on Network Slicing

Background on Network Slicing

In the previous section, we highlighted that telecommunication network operators are seeking to find new value by serving non-traditional use cases in vertical industries such as automotive or healthcare. These new use cases place contrasting demands on the network regarding latency, reliability, capacity and energy efficiency as outlined in Section 2.2. If taken in isolation, the tools and technologies are available to satisfy any one of these demands, as described in Section 2.3. The challenge lies in managing the coexistence of technologies to serve all of the use cases simultaneously, necessitating a network that can exhibit customisable behaviour based on the requirements to be satisfied.

This need for a customisable network is the continuation of a decades long trend towards service differentiation in telecommunication networks. As networks become more adaptable, they can serve more diverse use cases, which in turn drives the need for further adaptability. In effect, the ability to create an adaptable network necessitates such a network.

Creating networks capable of providing customisable behaviour requires the ability to seamlessly divide and aggregate entities across the entire network. Until recently, this was not generally feasible. In recent years, however, network virtualisation has emerged as a means to create virtual instances of network entities that facilitate sharing. We begin this section with an in-depth examination of virtualisation, its relationship to network slicing, and the role both of these have in future networks.

3.1 Network Virtualisation

Virtualisation enables the creation of a virtual representation of a physical resource which is independent from the underlying hardware realisation. Virtual resources are not constrained by the physical constitution of the substrate on which they are constructed, and may be treated as if they were *real* resources. Virtualisation is not a new concept in Information and Communication Technology (ICT), and has been widely applied to storage devices and computer hardware platforms. We are solely interested in the application of virtualisation to networks.

Network virtualisation was first proposed as a solution to deal with the ossification of the internet due to its potential to create a virtual testbed to experiment with new architectures [91]. While architectural purists at the time viewed virtualisation as a tool for simply evaluating new architectures, there was another school of thought, referred to as pluralists, who advocated for virtualisation to be a core part of the architecture itself. Pluralists envisioned virtualisation as a way to permit multiple heterogeneous networks to coexist on the same physical substrate, resulting in a more flexible internet [92].

In the wired domain, network virtualisation has been present for decades in the form of Virtual Private Networks (VPNs) and Virtual Local Area Networks (VLANs). These materialisations of virtualisation are quite limited in the flexibility that they offer. While VPNs have proven to be a popular way of creating a logical network from geographically distributed nodes, they do not offer the ability to independently program the network and are still constrained by the architecture of the underlying network [93].

From as early as the pluralist vision for the internet, we can see that network virtualisation has always been viewed as a tool for allowing specialised virtual networks to coexist. Although virtualisation principles have been applied in the form of overlay networks such as VPNs and VLANs, the limitations of the underlying physical infrastructure still persist, demanding the need for a more comprehensive implementation of network virtualisation that permits true isolation and programmability of virtual network instances [93].

Obtaining a precise definition of network virtualisation is surprisingly difficult. The authors in [94] review several definitions in the literature before offering their own attempt:

Network virtualisation is any form of partitioning or combining a set of network resources, and presenting (abstracting) it to users such that each user, through its set of the partitioned or combined resources has a unique, separate view of the network. Resources can be fundamental (nodes, links) or derived (topologies), and can be virtualised recursively. Node and link virtualisation involve resource partition/combination/abstraction; and topology virtualisation involves new address (another fundamental resource we have identified) spaces.

Although they might exhibit some differences, the many definitions of network virtualisation found in the literature share some common attributes. In general, network virtualisation should exhibit:

1. **Isolation** Virtual network instances should be logically isolated, and provide the illusion of full control of a dedicated network to the Mobile Virtual Network Operator (MVNO).
2. **Programmability** Virtual networks should offer the ability to configure and customise the network according to the requirements of the MVNO.
3. **Recursion** Virtual networks should appear real, and it should be possible to further virtualise them.
4. **Network Abstraction** A virtual network should be able to accommodate a new architecture independent of the underlying physical substrate.

The features of virtualisation listed above provide several advantages that cannot be achieved using dedicated networks or overlay techniques [95]. First, network virtualisation allows new protocols and architectures to be developed much faster as they can be tested using existing infrastructure in an isolated virtual network. Secondly, network virtualisation provides financial benefits to MVNOs, providing cost savings through the benefits of sharing the same physical infrastructure with other MVNOs and the ability to dynamically scale a virtual network to meet current demand. Finally, and most importantly in the context of this thesis, network virtualisation enables the creation of coexisting heterogeneous virtual networks that have been created on-demand to meet the requirements of a particular service or vertical industry.

It is this last advantage of virtualisation that captures our interest. We stated at the beginning of this chapter that creating customisable networks requires the ability to seamlessly divide and aggregate network entities. Virtualisation provides a means to do this by creating virtual representations of resources, which permits virtual networks to be built without concern for the constraints imposed by the granularity of the underlying physical resources.

The authors in [96] present a theory of virtualisation and emphasise that while network resources exist in the physical domain, virtualisation can only be done in the abstract domain. It is this ability to work in the abstract domain, in which the quantity of resources can be altered irrespective of the physical reality, that permits the seamless aggregation and division of a single shared substrate into multiple tailored virtual networks.

The challenge of achieving virtualisation can be reduced to two distinct problems known as the isolation problem and the embedding problem. The isolation problem refers to ensuring that virtual networks can coexist without interfering with one another. Virtual resources should be presented in such a manner as to give the illusion that they are real. The embedding problem refers to the necessary task of mapping virtual resources to real resources and deciding how to allocate them to virtual networks. The challenge in creating a virtual network with custom behaviour on-demand lies in solving the embedding problem while respecting the constraints of the isolation problem.

Network virtualisation is more mature in the domain of wired networks, having been adopted in several test-beds and extensively in data centres. The authors in [97], asking the question why wireless network virtualisation is needed, demonstrate the reduction in Operational Expenditure (OPEX) and Capital Expenditure (CAPEX) that it can bring. For the most part, the motivation for introducing network virtualisation in general hold true for wireless network virtualisation. The flexibility it introduces permits the creation of virtual wireless networks that can be configured to target specific services.

Network virtualisation can be divided into network device virtualisation and link virtualisation [94]. The authors in [96] note that additional challenges arise in wireless network virtualisation from the shared and stochastic nature of the link. In the first instance, the broadcast nature of the wireless medium makes it more difficult to satisfy the isolation problem. In the second instance, the capacity of a virtual link is difficult to predict due to the inherent variation of the wireless channel, which makes the embedding problem more challenging.

The core focus of this thesis resolves around enabling customisable future networks. Network virtualisation, including wireless network virtualisation, is of critical importance in this pursuit. As discussed, by representing resources in the virtual domain, virtualisation removes restrictions relating to the granularity of the physical underlying resources. This permits the aggregation and division of network resources to create virtual networks on-demand in a seamless manner. In this respect, network virtualisation is an enabling technology on which this thesis depends.

3.2 Virtual Network Slices

The authors in [96] make a distinction between slicing and aggregation, with slicing referring to a one-to-many virtualisation scenario in which a single piece of hardware is shared through multiple virtual instances. In contrast, aggregation is the use of virtualisation in a many-to-one scenario to group multiple pieces of hardware together as a single virtual instance.

Creating a virtual network is a process involving both aggregation and slicing using network device virtualisation and link virtualisation. However, when we discuss slicing, we refer to the process of partitioning a physical network of infrastructure, as a whole, into multiple virtual networks. As is common parlance in the literature, we will use the terms slice and virtual network interchangeably. Our interest in slicing is at a macro-scale of creating an entire virtual network, as opposed how the term is applied when discussing the virtualisation of individual links or nodes.

We define the scope of what we mean by network slicing as follows. Network slicing is the process of creating multiple logically isolated virtual networks using a shared physical substrate. Each slice consists of a set of resources (virtual or real), a set of functions operating on top of these resources, and a slice-specific configuration. As virtualisation is a prerequisite for network slicing, the attributes of virtualisation outlined in Section 3.1 also apply to network slices. Network slices should be isolated from one another, capable of possessing a different architecture than the underlying infrastructure, and should be customisable. This last point is of critical importance in the context of this thesis. The set of resources, network functions, and the configuration should all be chosen on a per-slice basis to enhance the slice's ability to satisfy the demands of a specific service.

3.2.1 Benefits of Network Slicing

The community's interest in network slicing stems from three core technical advantages that it offers.

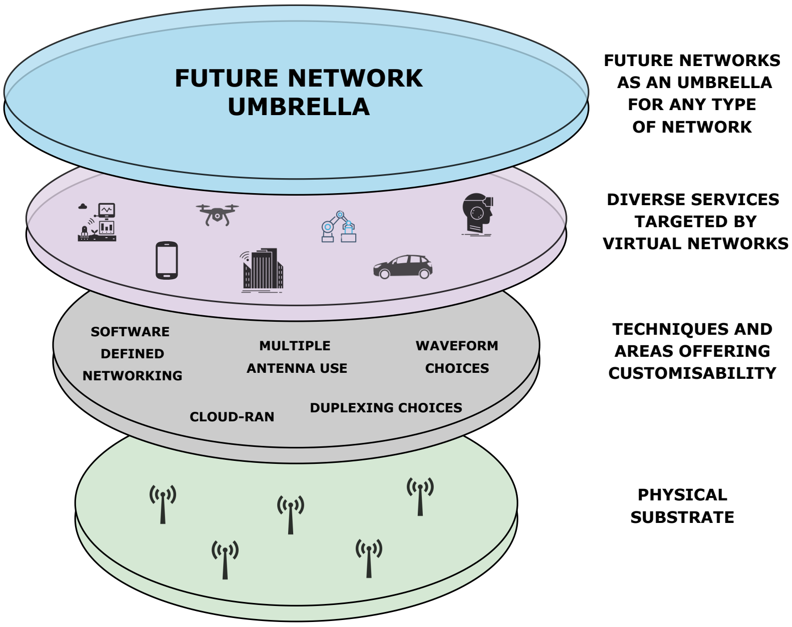

FUTURE NETWORK UMBRELLA
FUTURE NETWORKS AS AN UMBRELLA FOR ANY TYPE OF NETWORK
DIVERSE SERVICES TARGETED BY VIRTUAL NETWORKS
SOFTWARE DEFINED NETWORKING
MULTIPLE ANTENNA USE
WAVEFORM CHOICES
CLOUD-RAN
DUPLEXING CHOICES
TECHNIQUES AND AREAS OFFERING CUSTOMISABILITY
PHYSICAL SUBSTRATE

Figure 3.1: The future network umbrella. Future networks will encompass all types of networks, allowing customised virtual networks which target specific services and use-cases to be instantiated.

1. *Network slicing enables greater resource sharing.* While some resources in the network have always been shared through necessity (e.g. spectrum), slicing promises to broaden the scope for resource sharing with increased opportunities for statistical multiplexing gain. Virtualisation and associated enabling techniques such as Network Function Virtualisation (NFV), Software-Defined Networking (SDN), and Software Defined Radio (SDR) make it possible to dynamically share many types of infrastructure such as switches, radio heads, baseband processing units, and servers running network functions.
2. *Network slices can be tailored for different verticals or services.* Future networks are expected to cater to a vast array of service types, placing a diverse set of often contrasting requirements on the network. Through the judicious selection and configuration of the virtualised resources constituting a slice, network slicing allows the creation of bespoke networks that can target industries not traditionally served by telecommunication networks, such as in automotive and industrial verticals. Future networks, therefore, might not be considered one single type of network but rather an umbrella for a host of customised virtual networks (Fig. 3.1).
3. *Network slicing facilitates more flexible management of the network.* The programmability inherent in network slicing simplifies the management of the virtual network. This is part of a wider trend in telecommunication networks towards softwarisation. This is exemplified by SDN, which facilitates the separation of the user and control planes, and permits the reconfiguration of network switches using a centralised controller. The on-demand nature of network slicing, and the core ability to cus-

tomise the network, permit the scaling of virtual resources to match demand and to reconfigure the network through a centralised controller.

If network slicing is to be adopted, it must prove attractive for network operators. Each of the three technical advantages above translate to financial gains for operators. Network slicing offers a means for multiple MVNOs to share the same infrastructure, which can reduce their OPEX and CAPEX [98]. The ability to offer tailored networks on-demand is critically important for operators; after performing an analysis of the potential new revenue sources, [99] states that increases to operators' revenue streams depends on their ability to provide specialised services to niche user communities. Finally, network slicing permits the management of virtual resources through programmable interfaces, making it easier and cheaper to test, launch and maintain new functionalities, and reducing the time-to-market for new services.

The above advantages have resulted in significant commercial interest in network slicing, with several major industry players championing its cause and influencing its development. In the next section, we examine some of the visions for network slicing in industry.

3.2.2 Industry Visions

We will begin by examining the industry visions of several trade organisations which represent the interests of the telecommunication industry worldwide. We will then follow this up by directly surveying some of the visions of a selection of major worldwide equipment vendors. The organisations and white papers examined are listed in Table 3.1.

Trade Organisations

In 2015, the NGMN Alliance released a white paper [100] with the purpose of building a 5G vision informed by operator requirements, in consultation with NGMN partners. Network slicing was a key part of this vision with the flexibility it introduces identified as a key element to create new business opportunities. This white paper was subsequently followed up by another [101], providing conceptual models for network slicing along with precise definitions.

5G Americas is an industry trade organisation composed of the leading telecommunications service providers in the Americas. Their white paper [102] provides a comprehensive overview of network slicing, ranging from benefits and requirements to architectural discussions and operational aspects. In particular, it describes the full life-cycle of a network slice from the creation of a slice template, to the instantiation of the slice itself, up and down-scaling of resources, ensuring isolation and conducting routine maintenance.

A white paper [103] released by the GSMA, which represents almost 800 operators worldwide, presents the requirements of several vertical industries which can be served by customised network slices, including automotive, healthcare and industry 4.0. The white

Table 3.1: Table of industry white papers that discuss network slicing.

Organisation	White Paper
5G Americas	[102]
NGMN	[100], [101]
GSMA	[103]
FCC	[104]
Huawei, China Mobile, Deutsche Telekom, and Volkswagen	[105]
Ericsson	[106]
Nokia	[107], [108]
Qualcomm	[108]

paper advocates for the adoption of a generic slice template that can be used to describe the type of slice that is required by a particular service.

Finally, the Federal Communications Commission (FCC) released a white paper [104] highlighting the trend from service specific networks to converged networks and views network slicing to be the natural progression of this and a key element of future networks.

Vendors

At the 2017 Mobile World Congress, Huawei, China Mobile, Deutsche Telekom, and Volkswagen jointly released a white paper titled '5G Service-Guaranteed Network Slicing' [105]. The white paper outlines their shared vision for 5G, with network slicing deemed to be an essential part of it. In particular, the authors suggest that 5G has the potential to 'open up the telecom ecosystem to vertical industries', and hence allow the network to respond to the needs of emerging industries.

In a white paper [106] outlining Ericsson's 5G vision, network slicing and the ability to provide services with a customised network behaviour again plays a prominent role. A list of use cases with varying requirements is presented which further exhibits the diversity of the services that 5G must cater to.

A white paper by Nokia [107] highlights the need to shift from a 'network for connectivity model to a network for services model', with network slicing heralded as the way to achieve this. In both Nokia's and Ericsson's white papers discussed, virtualisation is presented as a key enabling technology.

Finally, a joint white paper [108] by Qualcomm and Nokia states that network slicing will be a key enabler for future networks, with the flexibility it introduces and the ability to offer dedicated network slices to customers essential to the creation of new business models and revenue streams.

3.2.3 NFV and SDN

Several of the industry white papers listed in Table 3.1 emphasised the important roles that SDN and NFV have to play in making network slicing a reality [100, 102, 105–107]. SDN and the role it has in enabling adaptable future networks was discussed extensively in Section 2.4.1.

NFV is a network architecture which utilises virtualisation to decouple network functions from the dedicated proprietary hardware on which they are typically implemented. With NFV, network functions can be implemented in software and deployed in virtual machines in commodity hardware in the network.

When NFV was first presented in a white paper in 2012 [109] by a consortium of companies, the motivation was to reduce the difficulties associated with deploying new network services. Standard industry practice for adding new network functions required physical hardware to be manually connected together, requiring significant expertise and CAPEX. The benefits of adopting NFV were stated to be reduced equipment costs, faster time-to-market for new services, and rapid up and down-scaling of resources. It also potentially enabled the sharing of resources across services and customer bases, and tailored service for different customer segments.

The potential that NFV could offer to network slicing was quickly realised, particularly for slice orchestration. Orchestration is the task of arranging and coordinating unconnected elements into an intelligible whole. Our definition of network slicing at the beginning of this section stated that a set of functions was a key element of a network slice. The last two benefits of NFV mentioned in the previous paragraph are beneficial in this regard. Each slice can possess its own instances of the network functions it requires running in virtual machines, allowing easy up and down-scaling to meet the demand of the virtual network.

The seminal white paper [109] noted that NFV and SDN are not dependent but could be mutually beneficial. The authors in [110] remark that there is already a consensus within academia and industry about the form that this mutually beneficial relationship will take. NFV can be used for the management and orchestration of network functions for slices, while SDN can be used as an assistive technology to manage the routing of flows between virtual machines in service chains comprising virtual network functions.

Architectures for network slicing built upon the concepts of NFV and SDN are prevalent in the literature [111–113]. Although NFV and SDN are not the direct focus of this thesis, any discussion on network slicing in the current climate would be incomplete without them due their positions as key enablers of customisable future networks.

3.3 Radio Access Network (RAN) Slicing

The slicing vision outlined in the 5G Americas white paper envisioned core slices, access slices and a selection function to map both of these slices to an end-to-end slice. Network slicing in the core network, enabled by SDN, NFV and cloud computing, is at a much more mature stage of research than RAN slicing [9, 114]. Slicing in the RAN introduces challenges. First, as noted in Section 3.1, virtualising the wireless links in the RAN is difficult due their shared and stochastic nature. Secondly, while capacity in core slicing can be scaled-up by simply adding more hardware to the underlying physical substrate, spectrum provides a physical constraint in RAN slicing [115].

In this section, we first examine whether RAN slicing is actually required, or whether core slices running tailored virtual network functions would provide sufficient customisability to serve diverse vertical industries.

3.3.1 Why is RAN Slicing Needed?

For RAN slicing to be warranted, it must assist in providing a solution to one of the main challenges currently facing the telecommunications industry, which we will call the requirements problem; namely that future networks are expected to cater to a vast array of verticals, placing a diverse set of often contrasting requirements on the network. We explore the potential changes to the RAN through the three main 3GPP target areas under consideration: enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communication (uRLLC), and Massive Machine-Type Communication (mMTC). In addition, we also consider a high mobility case.

eMBB

Mobile broadband was the primary target of 4G wireless communication systems, and eMBB represents a continuation of this. Capacity can be increased through multiple antenna techniques such as Multi-User Multiple-Input Multiple-Output (MU-MIMO) and Massive Multiple-Input Multiple-Output (Massive-MIMO). Densification of the network appears inevitable, with plans to use small cells to boost the capacity of Long Term Evolution (LTE) anchor cells outlined in the 3GPP non-standalone[1] 5G NR specifications. The use of multiple Radio Access Technologies (RATs) may also be employed to increase capacity in the future, with strategic offloading of eMBB traffic to supplementary WiFi hotspots or millimetre wave access points likely to play an important role.

While the above techniques will hopefully combine to provide the necessary capacity increase, they may not be suitable for service types other than eMBB. Massive-MIMO

[1] *Non-standalone* refers to the fact that an LTE anchor cell is required for control plane communications, while 5G NR cells are used to boost capacity.

requires accurate channel estimation and significant training, and may not be suited to high mobility or low latency use cases. The adoption of small cells creates an interference limited network that requires advanced interference coordination schemes, resulting in increased control plane overhead. Offloading to alternative RATs such as millimetre wave may not be suitable for mission critical applications due to the poor non-line-of-sight propagation characteristics at millimetre wave frequencies. Typically, techniques for increasing capacity require carefully coordinated scheduling of resources to avoid interference and optimise throughput. This generally results in significant levels of control signalling, making these techniques unsuitable for serving non-eMBB verticals.

uRLLC

Future networks are expected to provide end-to-end latencies of as low as 1ms to enable vertical industries that rely on tactile functionality and near-zero response times, resulting in a latency budget in the order of several tenths of a millisecond in the RAN. The Transmission Time Interval (TTI) places a lower bound on the latency associated with the air interface, as this is the minimum transmission unit in LTE. It is clear that current TTI values in LTE of 1 ms will not suffice, requiring support for shorter TTIs in 5G NR. Examining a breakdown of the latency across the RAN in the uplink and downlink, provided in Tables 5.2.1-1 and 5.2.1-2 of a 3GPP study on latency reduction techniques in [116], the actual transmission of data is not the dominant factor in determining latency in the RAN. However, data processing time is related to the transport block size, which in turn depends on the TTI, so reducing the TTI should also result in a reduction in processing times.

Slices targeting verticals that require high reliability may modify the adaptive modulation and coding scheme employed in LTE to use smaller constellation sizes and lower coding rates. Reliability can be improved through diversity. However, diversity techniques exploiting the time domain are not conducive to low latency, while frequency diversity can be expensive in terms of resources. Spatial diversity, in which the same information is transmitted over multiple spatial streams, reuses the same resources at the cost of increased control plane overhead. Interface diversity presents another option, but requires optimisation of selected links and increased control plane signalling.

mMTC

Machine-type communication is a key target for future networks, with the term used to encompass many use cases such as the Internet of Things (IoT), Industry 4.0, and vehicular connectivity. According to [117], the FP7 project METIS classified MTC into mMTC and Ultra-Reliable Machine-Type Communication (uMTC). mMTC targets connecting billions of low-rate, low-energy, and low-complexity devices (such as in IoT), while uMTC is concerned with providing low latency and high reliability (such as in V2X) and is similar to uRLLC.

mMTC represents a different type of traffic than what current mobile communication networks are designed for, and is not compatible with a system designed for eMBB communication. Extensive control plane signalling, including pilot signals, channel state feedback and hybrid automatic repeat requests, is required to ensure reliability, security, and enable schedule-based access for eMBB traffic. The performance improvements attributed to such techniques make the control plane overhead worthwhile. However, this is not necessarily the case for mMTC. Small packet sizes make the control overhead unaffordable, with the control signalling required potentially eclipsing the payload. Similarly, the massive number of devices could result in control signalling flooding the network. Finally, it is desirable to minimise the control signalling for low-complexity and battery-constrained devices such as IoT sensors. Notably, downlink traffic tends to dominate in eMBB, while mMTC systems tend to be uplink intensive.

As a result of these differences, contention-based and grant-free multiple access schemes are being considered for mMTC. Orthogonality may also be relaxed to increase capacity to accommodate massive numbers of devices.

High Mobility

In LTE, reference symbols known as pilots are distributed throughout the resource grid according to a fixed pattern. These pilots are used for two purposes:

- to estimate the channel for the purpose of one-tap equalisation in Orthogonal Frequency Division Multiplexing (OFDM);
- to allow the User Equipment (UE) to provide channel state information to the network to facilitate frequency selective scheduling.

In relation to the first point above, equalisation is performed for each OFDM symbol by interpolating the channel estimates obtained at pilot locations. If the channel coherence time reduces, such as in a high mobility scenario, the pilot placements are no longer sufficiently close for interpolation to provide accurate channel estimates for equalisation of OFDM symbols. Similarly, referring to the second point, the feedback provided to the network could be outdated when the next round of scheduling is performed, which reduces the gains from frequency selective scheduling. In these cases, pilot symbols should be placed closer together. This can be achieved by either increasing the density of the pilot symbols, or by reducing the symbol length. However, increasing the density reduces the number of information carrying symbols in a frame. In high mobility scenarios, large Doppler spreads also result in inter-carrier interference. One solution to mitigate this is to increase subcarrier spacing to separate carrier frequencies, which also results in shorter symbols. Hence, shorter symbols may be advantageous in high mobility scenarios to combat low channel coherence times and large Doppler shifts.

The Slicing Solution

It is obvious from the above examination of the main service-types in future networks that techniques and technologies which are required to meet the demands of one service-area can be detrimental to performance in a different service-area. RAN slicing offers a solution to this apparent conflict, allowing logical isolated sub-networks to be created that only employ the technologies and techniques which are optimal for the targeted vertical or service. Core slicing alone is not able to provide the necessary level of adaptability required to make customisable future networks a reality, as evident from the discussion presented in this section. In the following sections, we examine approaches for sharing the RAN among multiple slices, as well as techniques for customising the slices.

3.3.2 RAN Sharing Techniques

Passive sharing constitutes the simplest form of RAN sharing, and has been widely adopted in industry for the past two decades. Passive sharing refers to the sharing of sites and support infrastructure such as masts, power supplies and air-conditioning. Each Mobile Network Operator (MNO) still owns and manages its own radio and backhaul equipment.

Active RAN sharing is more ambitious and concentrates on the sharing of radio access equipment such as antennas, as well as the backhaul. It may also include the sharing of bandwidth and network management systems. 3GPP defines two forms of active RAN sharing [118]:

- **Multi-Operator Core Network (MOCN)** - Each operator possesses its own core network and shares the Base Station (BS); this is achieved by broadcasting the Mobile Country Codes (MCC) and Mobile Network Codes (MNC) for each MNO sharing the BS. MOCN consists of a shared baseband, shared spectrum, shared licences and can support up to six operators.
- **Gateway Core Network (GWCN)** - In addition to the sharing outlined for MOCN, MNOs may also share the Mobility Management Entity (MME).

Although not standardised by 3GPP, Multi-Operator Radio Access Network (MORAN) presents another approach for active RAN sharing which is supported by some vendors [119]. MORAN is distinguished from MOCN in that each MNO sharing a BS uses its own dedicated spectrum, allowing operators to independently select their own radio parameters for each cell.

The road from the RAN sharing techniques above to full multi-tenancy relies on virtualisation. A full realisation of network slicing in the RAN is required to enable the vision of customisable future networks. BS virtualisation can be categorised according to the level of resource isolation, with dedicated and shared resource models [9].

Dedicated Resource Model

In the dedicated resource model, physical resources, such as Physical Resource Blocks (PRBs) or carriers, are exclusively granted to each slice. As noted in [120], this is typically done at the hardware level, with radio components shared and separate BS protocol stacks running as software in virtual machines. Each slice has its own instance of the radio resource control, radio link control, packet data convergence protocol and medium access control layers.

The use of dedicated resources, while providing a high level of certainty to slices regarding their expected performance, reduces the statistical multiplexing gain possible and makes it difficult to up and down-scale slices according to demand.

Shared Resource Model

In the shared resource model, slice traffic is typically separated at the flow-level [120], permitting better multiplexing of resources and enabling MVNOs who do not own spectrum to deploy virtual networks.

Flow-level slicing is differentiated from spectrum-level slicing in [121], which defines spectrum-level slicing as the application of dynamic spectrum access and spectrum sharing to link virtualisation where the focus is on sharing the data bearer. In contrast, flow-level slicing differentiates between sets of flows belonging to different MVNOs. This may be bandwidth-based, in which the resource allocation to a slice is defined in terms of a data-rate, or resource-based, in which a slice's allocation of resources is defined in terms of a fraction of the available resources.

An example of a shared resource model is the Network Virtualisation Substrate (NVS) proposed in [122], which can accommodate slices using both bandwidth-based and resource-based flow-level slicing simultaneously. To achieve this, NVS adopts weighted log-utility functions which permit comparison between the two approaches. Minimum allocation units are defined for slices using each approach in Service Level Agreements (SLAs). After satisfying the minimum allocations, the goal of the slice scheduler is to maximise the overall utility of the BS.

In discussing the feasibility of BS virtualisation and the benefits of NVS in this regard, the authors in [120] highlight the importance of ensuring resource isolation between slices. Isolation was one of the key attributes of virtualisation listed in Section 3.1. RadioVisor [123] aims to accomplish radio resource allocation by dynamically slicing a three-dimensional resource grid consisting of time, frequency and space between virtual operators.

The authors in [10] state that while traffic isolation is easily obtained in a single cell scenario by granting orthogonal resources to slices, inter-cell interference must be considered in multi-cell scenarios. Hence, it is necessary to consider both *traffic isolation* and *radio-electrical isolation*. Four different approaches for achieving RAN slicing are proposed using different radio resource management functions to divide radio resources between MVNOs. The approaches are compared in terms of the level of radio-electrical and traffic isolation they provide, along with the level of customisability and flexibility that they provide. The authors comment that the approaches which perform the best in terms of isolation typically perform the worst in terms of customisability and flexibility.

Often a shared resource model will use a two-tier Medium Access Control (MAC) scheduler [9, 120], with the lower-tier enabling slice or flow-specific scheduling rules, and the upper-tier mapping the scheduling decisions from the lower-tier to a specific set of physical radio resources. While this provides better flexibility compared to the dedicated resource model, it cannot provide the same level of resource isolation and Quality of Service (QoS) [9].

This trade-off between resource isolation and efficient resource utilisation is at the core of RAN slicing. Greater isolation leads to less granularity in resource allocation, reducing flexibility and statistical multiplexing gains. As well as the static (dedicated) and dynamic (shared) techniques already discussed, the authors in [113] also introduce a third category of mixed resource allocation in which approaches attempt to adopt features of both techniques, with a portion of resources dedicated to slices and the remaining resources pooled and distributed for statistical multiplexing gains.

In this thesis, in Chapter 7, we tackle a related trade-off fundamental to network slicing; namely, the trade-off between statistical multiplexing gains arising from increased sharing, and the need for slices to provide assured performance to their subscribers. While the shared model is conducive to the former, dedicated resources facilitate the latter. Instead of using a mixed resource allocation approach, we aim to balance this trade-off through a combination of providing absolute assurances regarding resource availability over the lifetime of a slice, and utilising the practice of overbooking in the short-term to take advantage of fluctuating demand and increase resource utilisation.

3.3.3 Customising RAN Slices

The options for customising RAN slices using the collection of available RATs was extensively discussed in Section 2.3. Furthermore, the advantages that Virtualised RAN (VRAN) can offer in this regard are discussed in Section 2.4.2.

In particular, in Section 2.3.3, we present three options for designing a flexible air interface. One of these options suggested that future networks may permit the use of multiple waveforms, with the waveform employed dependent on the use-case being served. In Chapter 4, we explore the feasibility of multiple coexisting waveforms for a specific future network scenario.

As the standards for 5G New Radio (NR) matured, it became clear that the air interface would be based on Orthogonal Frequency Division Multiple Access (OFDMA) with flexible subcarrier spacings and TTIs. This is similar to the first option advocated for in Section 2.3.3, titled '*Single Waveform - adjusting parameters of a configurable waveform*'.

As per [124], subcarrier spacings of $\Delta f = 2^{\mu} \times 15$ [kHz] for $\mu \in \{0, 1, 2, 3, 4\}$ are supported in 5G NR, with $\mu = 0$ corresponding to the current LTE specification. Two cyclic prefixes are supported, and the number of slots and sub-frames per frame scales according to μ to ensure a constant frame length in the time domain.

To facilitate the coexistence of multiple numerologies in 5G NR, 3GPP is introducing a new concept called a bandwidth part. Defined in [124], a bandwidth part is 'a subset of contiguous common resource blocks defined...for a given numerology μ_i in the bandwidth part i on a given carrier'. Hence, each bandwidth part represents a contiguous portion of the resource grid, employing a particular numerology. Each end-user can be configured with up to four bandwidth parts, with one in use at any given time.

The introduction of flexible subcarrier spacings and TTIs affects the homogeneity of the time-frequency resource grid, as irregular shaped blocks of time-frequency resources granted to a slice must be multiplexed onto the resource grid with other slices, where each slice may potentially be using a different numerology. This problem, referred to as the tiling problem, is highlighted in [115]. In Chapter 5, we perform an examination of the potential time-frequency resource structure of the RAN, focusing on the trade-off between flexibility and the overhead related to ensuring coexistence between contrasting RAN slices. This trade-off is linked to the trade-off between flexibility and isolation described in Section 3.3.2, whereby more granular time-frequency allocations result in more flexibility but incur a greater overhead to ensure isolation between slices.

3.4 Business Models for Future Networks

The challenges currently faced by the telecommunications industry are prompting a change in the deployment scenarios and business models being considered for future networks.

Deployment Strategies

Three different deployment scenarios and the economic benefits of each are considered in [99], and summarised below.

1. Single one-size-fits-all network: The design paradigm adopted in previous generations, a one-size-fits-all style network attempts to satisfy all service requirements as best as it can.
2. Specialised dedicated networks: These specialised networks are built to meet the requirements of a particular vertical or service-type, and include dedicated hardware and software with limited cross-network sharing.

3. Specialised network slices: Using virtualisation techniques, a single network is divided into logically isolated end-to-end sub-networks known as slices. Similar to the specialised dedicated networks, each slice is tailored according to the needs of a particular target market; however, the slicing approach also benefits from a statistical multiplexing gain arising from cross-slice sharing of underlying network resources.

After performing an analysis of the potential new revenue sources, the authors of [99] state that increases to operators' revenue streams depend on their ability to provide specialised services to niche user communities (e.g. smart factories). Unlike specialised dedicated networks and specialised network slices, a one-size-fits-all network does not permit operators to do this. A further analysis of the costs associated with each deployment scenario suggest that the dedicated network approach does not scale well, with the authors concluding that specialised network slices can deliver the highest operating margins. It is notable that the specialised network slicing approach achieves its favourable status due to two characteristics. First, its ability to offer tailored network behaviour allows it to satisfy many different types of requirements and target new industries, verticals, and user communities. Secondly, the statistical multiplexing gains arising from shared network functions and infrastructure can keep costs reasonable.

The quantifiable benefits of network slicing over a single one-size-fits-all network are outlined in [99], with the OPEX costs for maintaining 35 slices only 25% of that of a single network and the CAPEX costs only 20% compared to a single network. Furthermore, assuming network slicing permits 50% of network operations to be automated, the OPEX components can be divided into operational costs, which are 55% of the cost for a single network approach, and components such as real estate and energy connections which are only 20% of the cost associated with a single network. In addition to operational savings, network slicing also provides a 37% increase in Average Revenue Per User (ARPU), if 35 slices are again assumed. This helps illustrate the importance of targeting users with specialised service requirements. Based on a network of 50 slices, a 15% increase in demand for tailored services results in a 46% increase in revenue, reaffirming the significant potential of network slicing for future networks.

Business Models

Traditionally, telecommunication networks have been dominated by a single role, namely the MNO, who owned its own infrastructure and operated its own network. MVNOs, as defined by the business model currently used, do not own any infrastructure and serve customers through wholesale capacity agreements with an MNO.

The advent of network slicing introduces the potential for several new roles and business models. Description of these roles are prevalent in the literature and vary from paper to paper. The authors in [121] describe both a two-level and three-level business model. In the two-level model, MNOs own the infrastructure including the access network, spectrum and backhaul, and offer virtual resources to service providers, who then lease and operate these virtual resources. In the three-level model, the infrastructure provider owns the physical substrate and leases physical resources to MVNOs. The MVNO then creates virtual resources and provides them to service providers, who use these virtual resources to serve their end users.

A similar two-level model is suggested in [112], with an infrastructure provider owning and virtualising a physical substrate, and a tenant leasing these virtual resources and using them to provide customised services to end users. This model is also proposed in [92], which uses the terminology service provider instead of tenant. It also suggests that service providers can become virtual infrastructure providers by partitioning their virtual resources and offering them to other service providers.

A four-level model is presented in [97], which defines an infrastructure provider, a Mobile Virtual Network Provider (MVNP), an MVNO, and finally a service provider. The infrastructure provider owns the physical substrate. The MVNP leases physical resources from the infrastructure provider and creates virtual resources. The MVNO then leases these virtual resources and operates them for service providers, who focus on providing service to end users.

This four-level model is simply an expanded two-level model with more specialised roles. The roles of the infrastructure provider and MVNP in the four-level model is combined into the single role of the infrastructure provider in the two-level model, while the roles of the MVNO and service provider are combined into the single role of service provider/tenant.

We adopt a two-level model, depicted in Fig. 3.2, for this thesis and define the roles and terminology as follows:

1. **Infrastructure provider (slice provider):** The infrastructure provider owns the physical network infrastructure, as well as the spectrum, which it applies virtualisation techniques to and leases virtual resources on-demand to MVNOs.
2. **MVNO (slice tenant):** The MVNO leases virtual resources from the infrastructure provider and operates a virtual network to provide specialised services to its subscribers.

A current realisation of the two-level model described above is that of a neutral host architecture, which is gaining traction in present-day markets. In this model, a third-party referred to as a neutral host [125] provides shared infrastructure on an open access basis to MNOs, who fulfil the role of slice tenants. The neutral host model is typically used to improve coverage in busy locations such as indoor venues in which MNOs may wish to acquire additional capacity on-demand.

The specialised slices approach described in [99] can be deployed using two separate approaches. In the first case, each slice may be a separate virtual network, managed and controlled by an independent MVNO who leases resources from an infrastructure provider. This is in contrast to a multi-service network consisting of slices in which a single MNO

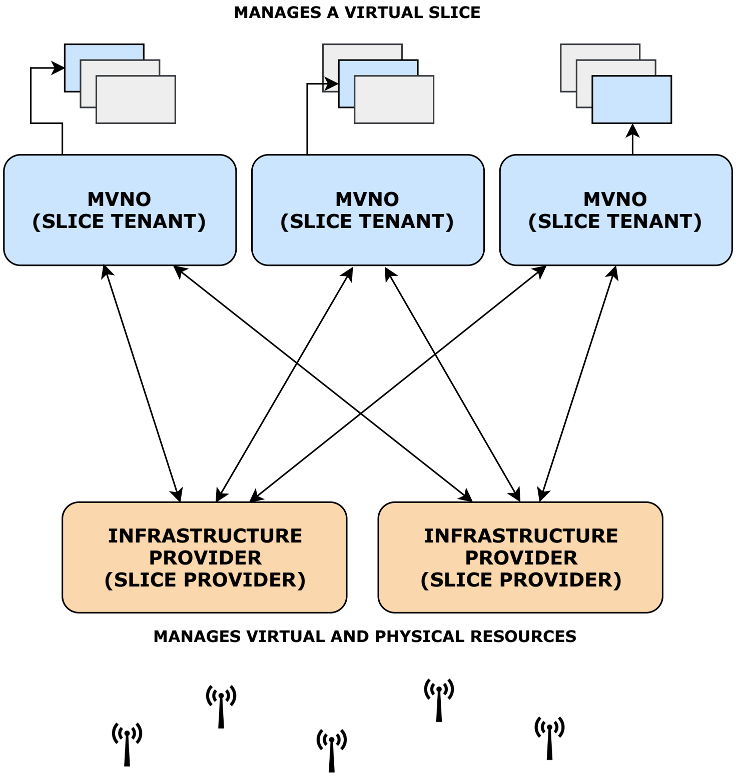


Figure 3.2: Illustration of a two-level business model for network slicing in which slice tenants lease virtual resources from infrastructure (slice) providers and manage tailored virtual networks.

owns and manages the entire network, including all of the slices. Hence, virtualisation need not involve multiple MVNOs; an MNO may virtualise its own resources to conveniently separate the functionality it provides from the infrastructure it owns, and thereby facilitate the creation and management of customisable slices.

In this thesis, we consider both of the above cases, i.e. a single MNO slicing its own network (e.g. Chapter 5) and multiple MVNOs operating slices as independent business ventures (e.g. Chapter 6). In particular, in Chapter 6, we consider a business case consisting of multiple independently operated specialised slices and examine how user admission is performed in this scenario, as a single user may need to avail of the services of more than one slice. We propose a model in which entities called *subscription brokers* group service level agreements with multiple specialised slices into a single subscription bundle, with a fixed data allowance that can be used by the subscriber as needed across any of the slices included in the bundle. Our case study demonstrates how the matching between users and slices may be performed in such a scenario and the benefits that can be obtained.

4 Coexistence of Waveforms to Serve Diverse Use Cases

Coexistence of Waveforms to Serve Diverse Use Cases

The modulation format and multiple access technique for future networks are not yet known, with many contenders under consideration, each proving advantageous in certain scenarios and lacking in others. As discussed in Section 2.3.3, future networks could potentially permit the use of different waveforms for different use cases, each one optimal for a given scenario. In this chapter, we explore this possibility and focus on the coexistence between two future network use cases, broadband services and machine-type Device-to-Device (D2D) communications, which may use different waveforms.

Many low-power, wide-area network solutions, such as Narrowband Internet of Things (NB-IoT), have been developed in response to the low rate, latency-tolerant traffic that is typically associated with the Internet of Things (IoT). However, there exists a contrasting set of inter-machine communication use cases that will possess requirements for low-latency and potentially high data-rate communication resulting from the increased use of robotics, artificial intelligence, and machine learning across multiple sectors such as energy [126], health, industry, and automotive. To enable direct communication between machine-type devices in future networks, D2D communication has been suggested as an enabling technology [127, 128]. Autonomous manufacturing systems, smart factories, and self-organising warehouses are but a few examples of use cases which require direct inter-machine communication in a spatially clustered environment.

The overhead associated with achieving and maintaining synchronous communication for Directly Communicating Users (DUEs) in a Machine-Type Communication (MTC) scenario can be significant, and it may be desirable to reduce the control burden placed on the Base Station (BS) to achieve this. As highlighted in [129], achieving synchronisation in D2D communications is challenging. This is particularly relevant for clustered machine-type D2D communication scenarios, in which the close proximity of the multiple DUE pairs to each other makes them particularly vulnerable to leakage interference arising from synchronisation errors.

Hence, while ordinary Cellular Users (CUEs) are well served using synchronous communication, clustered machine-type DUEs may instead be best served using asynchronous communication. Clustered MTC, based on D2D communication, is sufficiently different from ordinary cellular traffic to warrant investigation into what waveform choice results in the best performance. Furthermore, the types of devices that we are considering, such as machinery in a smart factory, are designed for specific purposes and may not need to support traditional cellular communication through a BS. In this case, it makes sense for them to use a waveform that is better suited for asynchronous MTC.

Orthogonal Frequency Division Multiplexing (OFDM), employed in Long Term Evolution (LTE), performs quite well when synchronisation can be achieved, and some variant of it may continue to be the best choice for cellular communication. However, OFDM's deficiencies in the presence of timing and frequency offsets are well known, with several alternative waveform candidates shown to perform better in asynchronous scenarios [130]. Hence, we are motivated to examine the coexistence of various combinations of waveforms for CUEs and DUEs. We study and quantify how each of the waveforms under consideration performs when employed by DUEs operating in an asynchronous manner, compared to a baseline case consisting of synchronous OFDM. In the remainder of this chapter, we use the term *alternative waveforms* to refer to the multitude of modulation formats that have been proposed in the literature as an alternative to OFDM.

Research Question, Key Contributions and Chapter Organisation

The research question, outlined in Section 1.1, that this chapter addresses is the following:

How can the diverse use cases in future networks be satisfied through the coexistence of multiple waveforms, with each service employing a waveform that is best suited for it?

The use of alternative waveforms to OFDM in D2D communications has been considered previously in several papers, and we provide a brief overview of the main works here. In [131], the authors investigate a D2D video transmission network in which D2D transceivers use Filtered Multi-Tone (FMT) to ensure that no Inter-Carrier Interference (ICI) occurs. In [132], the authors consider power loading for D2D pairs operating in an asynchronous manner, and compare the performance of OFDM and Filter Bank Multi-Carrier with Offset Quadrature Amplitude Modulation (FBMC/OQAM). However, no cellular users are considered in either of the above works and therefore issues regarding the coexistence of waveforms for different types of users do not arise.

The authors in [133] aim to maximise the sum-rate for asynchronous D2D underlay communications and consider the use of both Filter Bank Multi-Carrier (FBMC) and OFDM. Resource sharing between DUEs and CUEs in the downlink band is considered in [134], which suggests the use of Filtered Orthogonal Frequency Division Multiplexing (f-OFDM) for DUE devices to enable them to use parts of the guard band. In [135], the authors study the use of Universal Filtered Multi-Carrier (UFMC) for D2D communication, but focus solely on inter-D2D interference between DUE pairs that are not in the same cell. DUE pairs are assumed to synchronise to CUEs only if they are in the same cell in [136].

Our work differs from the aforementioned papers in many regards. Firstly, we are concerned with asynchronous direct communication whereby each DUE cannot be assumed to be synchronised with any other device in the system. We are also targeting the use of direct communication in spatially clustered MTC applications in which the inter-device leakage interference arising from misaligned communication plays a key role. The key novelty of this chapter lies in its comparison of the performance of multiple waveforms, whereby the waveform used by DUEs and CUEs may be different. Finally, we also consider a multi-cell system employing a frequency reuse technique known as strict Fractional Frequency Reuse (FFR), and consider all possible interference links in the system to obtain realistic results. To the best of our knowledge, no work available in the literature tackles inter-user interference caused by the asynchronous coexistence of multiple use cases employing different waveforms with a similar level of detail and for so many different waveforms.

In this chapter, we first consider a single-cell scenario to demonstrate the effects of inter-D2D interference and the benefits that permitting waveforms to coexist can provide. We then expand our model to a multi-cell scenario and provide an extensive system-level analysis into the performance of coexisting waveforms under varying system parameters. Our contributions in this chapter are as follows:

- We demonstrate that inter-D2D interference can be significant in use cases comprising clustered devices, motivating the use of a waveform with improved spectral containment over OFDM.
- We show that inter-D2D interference becomes negligible if DUE pairs use a suitable alternative waveform, simplifying the optimal Resource Allocation (RA) and power-loading schemes.
- Using system-level simulations, we demonstrate the benefits of serving high rate clustered MTC use cases through asynchronous D2D communication, enabled by the coexistence of waveforms, whereby DUEs employ an alternative waveform such as FBMC/OQAM and regular CUEs continue to use OFDM.
- We provide an analysis of several prominent alternative waveforms across a range of MTC scenarios by varying key system parameters such as cell size, cluster size, DUE transmit power, and maximum possible Timing Offset (TO) and Carrier Frequency Offset (CFO).

The remainder of the chapter is structured as follows. Section 4.1 describes the theory behind several prominent alternative waveforms to OFDM, and discusses their relative strengths and weaknesses. The system model for our single cell scenario is described in Section 4.2, followed by analysis in Section 4.3. Section 4.4 details an expanded multi-cell model, with results provided in Section 4.5. Finally, Section 4.6 concludes the chapter.

4.1 Candidate Waveforms

Below, we briefly present the waveforms that we will consider in this chapter.

1. *OFDM:* Although alternative waveforms are being studied, OFDM may still have an important role to play in future networks. OFDM works quite well in the downlink of cellular systems. In scenarios that do not comprise MTC or delay-intolerant communications, the signalling overhead required to align individual devices becomes affordable. Possibly the biggest advantage OFDM has is inertia. Its wide adoption in 4G systems and Wireless Local Area Network (WLAN) technologies has resulted in a wealth of research in the area, coupled with excellent success in implementation. In particular, issues related to synchronisation, estimation and detection have been extensively studied and solved. However, OFDM suffers from high out-of-band emissions and is known to perform poorly in situations where multiple users transmit adjacently and asynchronously, which is precisely the class of network deployments that interests us in this study.
2. *FBMC:* FBMC waveforms apply an enhanced filtering on a per-subcarrier level to remove the large sidelobes typically associated with OFDM transmission. This filtering makes FBMC subcarriers highly spectrally localised, which reduces the sensitivity to asynchronism. Moreover, FBMC systems generally do not rely on a Cyclic Prefix (CP) to combat Inter-Symbol Interference (ISI). The combination of these two attributes, reduced sidelobes and no CP, results in a time-frequency efficiency that is very close to 1 (and approaching 1 in the ideal case of infinite block lengths). However, to achieve their enhanced spectral localisation, FBMC systems use long prototype filters, which makes them unsuitable for transmission of short packets. The improved spectral containment of FBMC compared to OFDM also proves to be the source of many of its weaknesses, as filter lengths are quite long relative to the length of a single symbol. Consequently, high spectral efficiency is only achievable for long continuous transmissions. Hence, FBMC may not be the optimal choice for MTC which is characterised by short packets sizes and bursty traffic.
   Several FBMC-based schemes have been proposed in the literature. In this chapter, we consider the following ones:
   - FMT [137]: to reduce out-of-band emission, every subcarrier in FMT is filtered by a narrow passband filter, which inevitably results in the loss of the orthogonality between subcarriers according to the Balian-Low theorem. To deal with this, a guard band is added between every subcarrier; however, this reduces the spectral efficiency of the system.
   - FBMC/OQAM [138, 139]: possibly the most well-known alternative to OFDM, FBMC/OQAM achieves maximum spectral efficiency by removing the guard bands used in FMT. ICI and ISI are eliminated by using Offset Quadrature Amplitude Modulation (OQAM) modulation instead of Quadrature Amplitude

Modulation (QAM). However, FBMC/OQAM systems achieve orthogonality only in the real domain, and suffer from pure imaginary interference which can be detrimental for equalisation and makes their application to Multiple-Input Multiple-Output (MIMO) challenging.

- Filter Bank Multi-Carrier - Pulse Amplitude Modulation (FBMC-PAM) (also known as lapped FBMC) [140]: whereas FBMC/OQAM systems double the symbol rate, FBMC/Pulse Amplitude Modulation (PAM) doubles the number of subcarriers. It also uses a short sine filter which achieves a good trade-off between time and frequency localisation.

3. *Generalised Frequency Division Multiplexing (GFDM)* [141]: One of the main drawbacks of the aforementioned FBMC waveforms is the delay incurred by linear convolution with the prototype filter on each subcarrier. To overcome this issue, GFDM applies circular convolution to filter independent data blocks consisting of K subcarriers and M sub-symbols per subcarrier. In addition, only one CP is applied per entire block to reduce the block overhead. Each subcarrier within a block is filtered individually, with pulse shaping applied circularly to remove the filter transient intervals, thereby reducing latency and increasing suitability for MTC. However, circular filtering results in non-orthogonal subcarriers, introducing both ISI and ICI which must be dealt with using interference cancellation techniques, or linear decoders which increase the error rate and complexity of the receiver compared to OFDM. The block structure permits to add a CP which can be used to ease equalisation, relax synchronicity requirements and support uncoordinated access for MTC.
4. *UFMC and f-OFDM:* UFMC [142], also known as UF-OFDM, aims to generalise OFDM and FBMC in order to reap the benefits of both while avoiding their respective limitations. While FBMC filters individual subcarriers, UFMC applies filtering to groups of adjacent subcarriers, which reduces the ramp-up and down delays of the prototype filter and improves performance over FBMC in bursty communications. We note that UFMC is a variant of OFDM and may be described as filtered zero-prefix OFDM [61]. One of the advantages of UFMC lies in the fact that it preserves time orthogonality between subsequent symbols by limiting the filter tails within the guard interval. However, this does not allow for long filters and may therefore make it difficult to achieve satisfactory out-of-band rejection levels when dealing with signals that have a small passband. To overcome this, f-OFDM systems [143] use longer filters, which enable communication devices to achieve sharper frequency localisation at the cost of orthogonality loss between subsequent symbols. Similar to FBMC, both UFMC and f-OFDM offer higher spectral containment than OFDM due to their low out-of-band emissions. However, in contrast to FBMC, which achieves reasonable spectral efficiency only for sufficiently long transmission frames, UFMC and f-OFDM can offer improvements in small, bursty communication situations thanks to their reduced filter tails.

Table 4.1: Parameters of considered waveforms.

Parameter	**OFDM**	**FMT**	**FBMC / OQAM**	**FBMC-PAM**	**GFDM**	**f-OFDM**	**UFMC**
Time-symbol (T)	$\frac{1}{\Delta\mathrm{F}}$	$\frac{1}{\Delta\mathrm{F}-W_{\mathrm{GB}}}$	$\frac{1}{\Delta\mathrm{F}}$				
CP (T_{cp})	$\frac{T}{8}$				$\frac{T}{8}$		
Filter		root raised cosine (RRC), rolloff 0.22, duration $6T$	PHYDYAS, duration $4T$	Sine filter, duration $2T$	RRC, rolloff 1, duration $5T$	Truncated sinc filter [143] with $T_w = \frac{T}{2}$	Chebyshev, -60 dB attenuation, duration T_{cp}
Active subcarriers per RB	12		11		12	11	12
Bandwidth efficiency (Φ_{wf})	8/9	8/9	11/12		$5/(5+1/8)$	$8/9 * 11/12$	8/9

4.1.1 Implementation Parameters Selected for Waveforms under Study

All of the aforementioned waveforms have fostered a wide range of works [144, 145], with a large number of different implementations and parameters considered in the literature. Studying all of the proposed variations of these waveforms would therefore be infeasible. Hence, to ensure fair comparison, we choose parameters, filters, and implementations that are representative of most works in the literature.

As our work focuses primarily on investigating how alternative waveforms can facilitate coexistence in a certain band of the wireless spectrum, we consider that each studied modulation scheme uses the same subcarrier spacing $\Delta\mathrm{F} = 15\mathrm{kHz}$ in accordance with current LTE standards to ensure fairness in our comparisons. Note that, in the particular case of FBMC-PAM, each subcarrier is actually composed of two virtual subcarriers of width $\Delta\mathrm{F}/2$. Other relevant parameters are presented in Table 4.1. T represents the time symbol, T_{cp} is the CP or guard interval duration where applicable, T_w is the duration of the window used by f-OFDM and W_{GB} is the width of the guard band used by FMT. We set W_{GB} so that FMT has the same spectral efficiency as OFDM. Finally, note that the GFDM system that we consider uses blocks of 5 symbols, which is a commonly used value [141].

Due to filtering, leakage interference for FBMC/OQAM, FBMC-PAM and f-OFDM is concentrated in the subcarrier directly adjacent to the active RB. Hence, to take full advantage of the improved spectral properties of these waveforms, we only use 11 subcarriers per RB instead of 12, leaving a guard band of 1 subcarrier between RBs. All other waveforms use the full 12 subcarriers since their sidelobes are larger and leakage interference spans multiple subcarriers. In these cases, adding a single guard subcarrier between RBs would offer little advantage and would just reduce the bandwidth efficiency of the waveforms, thereby reducing the rate achievable with them. Note that FMT is a special case in that it does not require a guard subcarrier between RBs due to its implementation, which places a guard band between every subcarrier.

4.2 Single Cell Scenario: System Model

In the single cell model, we investigate an Orthogonal Frequency Division Multiple Access (OFDMA) based network in which DUE pairs are permitted to reuse the uplink resources of the incumbent cellular users in an underlay fashion, subject to interference constraints. We stay consistent with the literature and consider uplink resource sharing for two reasons. Firstly, in the uplink, all of the interference imposed by the DUE users onto the cellular users is experienced at the BS, enabling this type of interference to be mitigated through BS coordination. Secondly, and most importantly, some of the pilot information broadcast in the downlink is crucial and should not be interfered with. The uplink in an OFDMA cell uses Single Carrier - Frequency Division Multiple Access (SC-FDMA), which can be viewed as OFDM with a pre-coding applied and therefore does not affect our analysis.

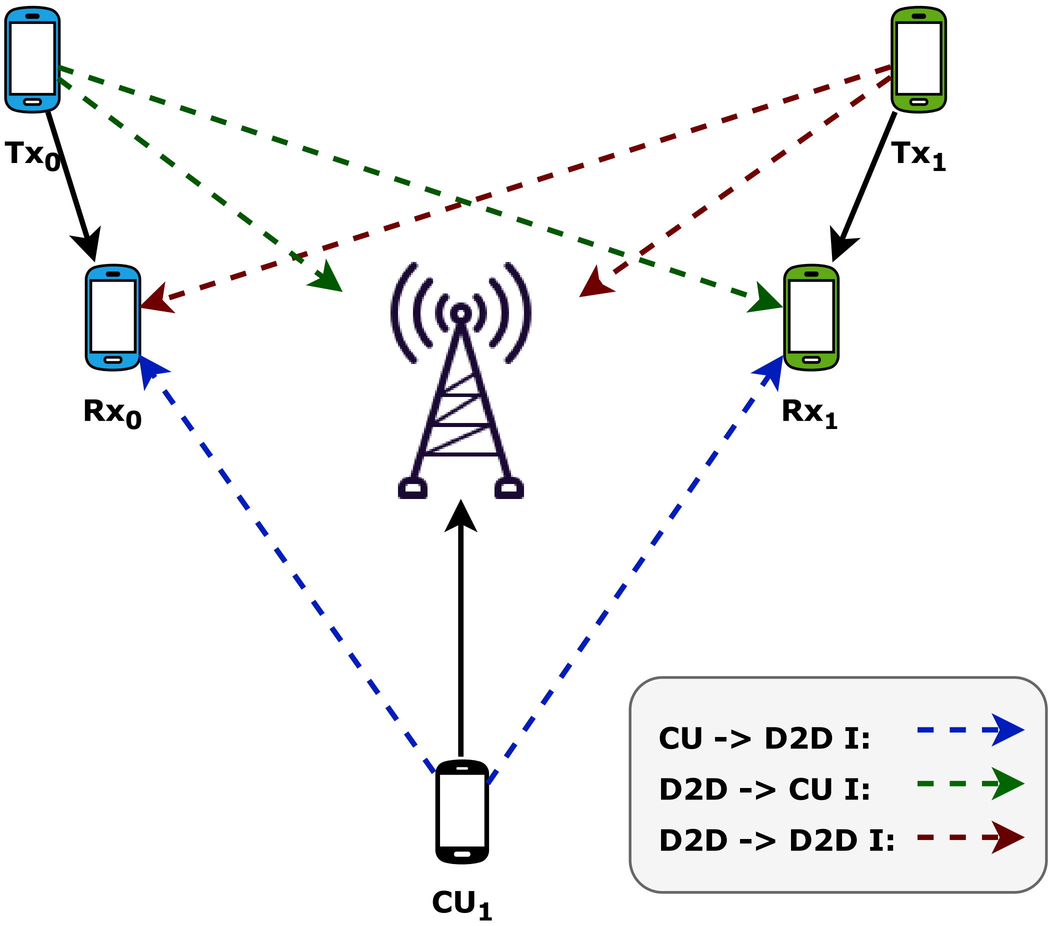


Figure 4.1: Simplified diagram showing two DUE pairs and one cellular user with both interference channels (dashed lines), and useful channels (solid black lines) outlined.

Fig. 4.1 illustrates a simplified scenario in which D2D communication underlays an OFDMA network in the uplink. In a more general scenario, multiple DUE pairs coexist with multiple CUEs, and reuse the uplink spectral resources. C denotes the set of incumbent (cellular) users, D denotes the set of DUE pairs, and R denotes the set of RBs. The useful and interference channels in Fig. 4.1, shown as solid and dashed lines, respectively, are presented in Table 4.2. CUEs do not interfere with each other as we assume they are perfectly synchronised by the BS. Therefore, there are three main interference types requiring consideration:

1. The DUE pairs interfere with the incumbents' transmissions. Since we are investigating uplink resource sharing, this interference is observed at the BS.
2. Conversely, the incumbents interfere with the DUE pairs at DUE receivers.
3. DUE pairs interfere with each other (inter-DUE interference).

In each type, we consider both co-channel and adjacent-channel interference. We model DUE pairs using either FBMC/OQAM or OFDM, while the cellular users are restricted to

Table 4.2: Useful and interference channels for Fig. 4.1.

Useful Channels		
$\mathrm{cu}_1 \rightarrow \mathrm{eNb}$	$\mathrm{Tx}_0 \rightarrow \mathrm{Rx}_0$	$\mathrm{Tx}_1 \rightarrow \mathrm{Rx}_1$
Interference Channels		
$\mathrm{cu}_1 \rightarrow \mathrm{Rx}_0$	$\mathrm{Tx}_0 \rightarrow \mathrm{eNb}$	$\mathrm{Tx}_1 \rightarrow \mathrm{eNb}$
$\mathrm{cu}_1 \rightarrow \mathrm{Rx}_1$	$\mathrm{Tx}_0 \rightarrow \mathrm{Rx}_1$	$\mathrm{Tx}_1 \rightarrow \mathrm{Rx}_0$

using OFDM.

We consider an OFDMA macro-cell with parameters selected based on the 3GPP LTE standard, as outlined in Table 4.3. We assume that the cell is fully loaded with each CUE assigned a single uplink RB. DUE devices underlay the OFDMA cell by reusing a single uplink RB. We also assume that the BS has perfect knowledge of all links in the system, including the interference link between a CUE and a DUE receiver.

We employ the WINNER II channel models [146] to provide us with a distance based pathloss, which also incorporates the probability of line-of-sight. WINNER II provides channel models for many different scenarios, in which the model parameters are based on statistical distributions extracted from channel measurements. Specifically, we use scenario B1 - *urban micro-cell.*

4.2.1 Interference Model

The main measure that we base our analysis upon is the Signal-to-Noise and Interference Ratio (SINR) experienced by incumbent CUEs and DUE pairs. To rate the latter with accuracy, it is necessary to use models of interference that properly estimate the leakage that two asynchronous users inject onto each other. As mentioned earlier, to the best of our knowledge, most studies on D2D underlay operation do not consider leakage between adjacent frequency resource blocks. Papers that do consider leakage, as in [136], rely on the Power Spectral Density (PSD)-based model, the shortcomings of which have been demonstrated in [3].

Fortunately, a number of papers have extensively analysed and modelled the leakage between asynchronous users operating on different parts of the spectrum band, and derived interference tables that we will build our analysis upon [1–3]. More precisely, we draw upon the work of [1] to rate the interference from FBMC/OQAM to FBMC/OQAM users, or OFDM to OFDM users. Additionally, we consider the interference from OFDM to FBMC/OQAM and from FBMC/OQAM to OFDM according to the recent works of [2] and [3]. These works allow us to rate the value of $I\{\mathrm{A} \rightarrow \mathrm{B}\}(l)$, which corresponds to the interference injected by a subcarrier of waveform A to a subcarrier of waveform B at a

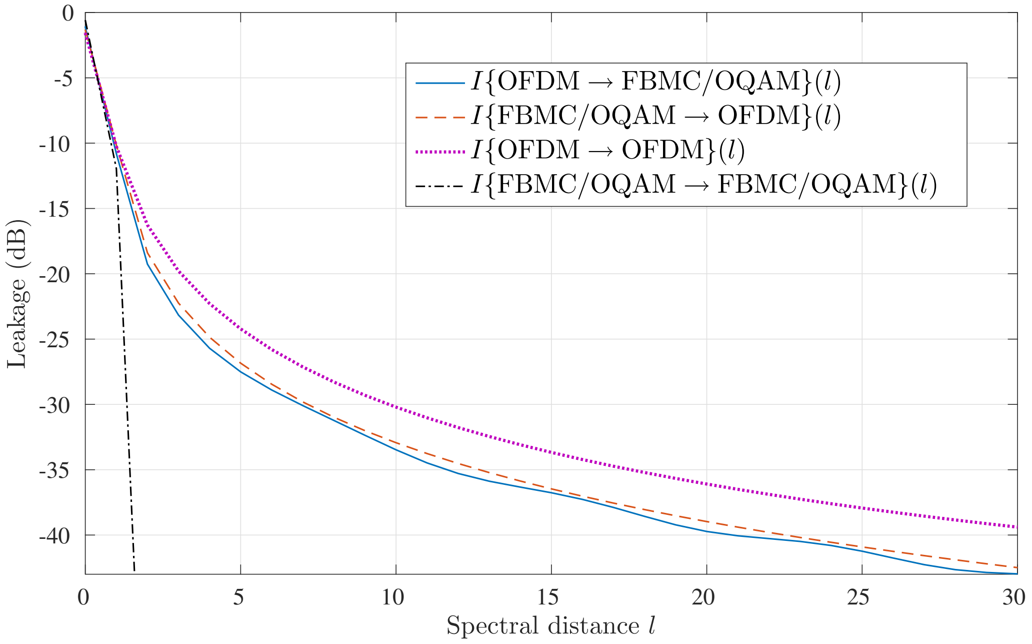


Figure 4.2: Interference tables measuring the value of interference injected between different couples of waveforms according to [1–3].

spectral distance of l subcarriers.

In the single cell model, we use the interference table plotted in Fig. 4.2. This figure shows that the use of FBMC/OQAM for D2D operation will only marginally reduce the interference between cellular and DUE users, as $I\{\text{OFDM} \rightarrow \text{FBMC/OQAM}\}$ is only slightly less than $I\{\text{OFDM} \rightarrow \text{OFDM}\}$. This has been thoroughly explained in [3]. However, the interference between asynchronous DUE users will be drastically reduced if they use FBMC/OQAM instead of OFDM, since $I\{\text{FBMC/OQAM} \rightarrow \text{FBMC/OQAM}\}$ is considerably lower than $I\{\text{OFDM} \rightarrow \text{OFDM}\}$.

4.2.2 Single Cell Scenario: Resource Allocation Formulation

We are interested in comparing the performance of using either OFDM or FBMC/OQAM for D2D communication, rather than proposing a new RA scheme for underlay D2D. Hence, we formulate the RA problem, consisting of both resource block allocation and power-loading, as an optimisation problem that assumes knowledge of all channels in the system. This serves as a platform to allow us to compare the relative performance of both waveforms for varying system parameters.

A DUE pair is allowed to transmit when the interference introduced on the incumbent network does not prevent the incumbent CUEs from satisfying their minimum SINR constraints. The D2D transmissions affect the SINR experienced at the BS and hence the CUEs suffer from adjacent channel interference. The SINR of the CUE indexed by i can therefore be expressed as

$$\gamma_i \; = \; \frac{P_i h_{iB}}{\sigma_\nu^2 + I_{\text{d2d}}^{\text{C}}}, \tag{4.1}$$

where P_i is the transmit power of the i^{th} CUE, h_{iB} is the channel gain between the i^{th} CUE and the BS, and σ_ν^2 is additive white Gaussian noise variance. Finally, $I_{\text{d2d}}^{\text{C}}$ is the interference experienced by the i^{th} CUE from the DUE pairs in the cell, and is given by

$$I_{\text{d2d}}^{\text{C}} \; = \; \sum_{j \in D} \sum_{m \in b_j} \sum_{k \in b_i} h_{jB} P_{jm} I(|k-m|), \tag{4.2}$$

where m indexes the subcarriers in the band b_j used by DUE pair j, P_{jm} is the power of the j^{th} DUE pair on subcarrier m, k indexes the subcarriers in the incumbent band b_i used by user i, and $I(|k-m|)$ is the appropriate interference table $I\{\text{A} \rightarrow \text{B}\}$ in Fig. 4.2, depending on the waveform being used by the DUE pairs.

A DUE receiver will experience two types of interference: i) interference from CUEs, and ii) interference from other DUE pairs. The SINR experienced on subcarrier m at the DUE receiver of pair j is given by

$$\gamma_{jm} \; = \; \frac{P_{jm} h_j}{\sigma_\nu^2 + I_{\text{cu}}^{\text{D}} + I_{\text{d2d}}^{\text{D}}}, \tag{4.3}$$

where I_{cu}^{D} is the interference injected on the m^{th} subcarrier of the j^{th} DUE pair from cellular users using OFDM in the incumbent band, and $I_{\text{d2d}}^{\text{D}}$ is the interference injected on the m^{th} subcarrier of the j^{th} DUE pair from all other DUE users. I_{cu}^{D} is defined as

$$I_{\text{cu}}^{\text{D}} = \sum_{i \in C} \sum_{k \in b_i} h_{ij} P_{ik} I(|k-m|). \tag{4.4}$$

Finally, $I_{\text{d2d}}^{\text{D}}$ is the interference from other D2D links given by

$$I_{\text{d2d}}^{\text{D}} = \sum_{d \in D, d \neq j} \sum_{n \in b_d} h_{jd} P_{dn} I(|n-m|). \tag{4.5}$$

We can now formulate an optimisation problem, using the above SINR expressions, in which the objective is to maximise the sum rate of DUE pairs, subject to a minimum SINR constraint for each CUE.

$$\text{P1} : \max_{P_{jm}, \omega_{jr}} \sum_{j \in D} \sum_{r \in R} \sum_{m \in b_r} \omega_{jr} \log(1 + \gamma_{jm}), \tag{4.6}$$

subject to

$$
\begin{aligned}
&\omega_{jr} \in \{0,1\}, \forall j, r, && \text{(4.6a)}\\
&\sum_{j \in D} \omega_{jr} \leq 1, \forall r \in R, && \text{(4.6b)}\\
&\sum_{r \in R} \omega_{jr} = 1, \forall j \in D, && \text{(4.6c)}\\
&\gamma_i \geq \mathrm{SINR}^{\mathrm{C}}_{\mathrm{min}}, \forall i \in C, && \text{(4.6d)}\\
&P_j = \sum_{m \in b_j} P_{jm} < P^{\mathrm{D}}_{\mathrm{max}}, && \text{(4.6e)}
\end{aligned}
$$

where ω_{jr} is a resource reuse indicator where $\omega_{jr} = 1$ when DUE pair j reuses RB r, and $\omega_{jr} = 0$ otherwise, $\mathrm{SINR}^{\mathrm{C}}_{\mathrm{min}}$ is the minimum acceptable SINR that a CUE must achieve, and $P^{\mathrm{D}}_{\mathrm{max}}$ is the maximum transmit power of a D2D transmitter.

Optimisation problem P1 is a Mixed Integer Non-Linear Programming (MINLP) problem from which it is difficult to obtain the solution directly. Accordingly, we split the optimisation problem into two sub-problems. First, we perform RB assignment, which is a discrete optimisation problem. Once RBs have been assigned, we perform power-loading for the DUE pairs.

Even after splitting P1 into two simpler problems, solving them remains complicated due to the inclusion of inter-DUE interference. The main source of this complexity lies in the fact that $I^{\mathrm{D}}_{\mathrm{d2d}}$ (equation (4.5)) is a function of the power assigned to each subcarrier of each DUE transmitter. Therefore, the different variables over which the optimisation is performed are coupled, as the SINR of each DUE pair affects the SINR of every other pair, complicating equation (4.6). Furthermore, incorporating inter-DUE interference into the RA scheme would assume that every DUE pair is able to obtain information regarding the interference contribution from every other DUE pair. This is an unrealistic assumption, requiring an exchange of information between DUE pairs before any resource is assigned. Hence, in reality, we would like to be able to perform both RB assignment and power-loading without needing to consider inter-DUE interference.

Therefore, we consider a simplification of P1 (equation (4.6)) where the SINR γ_{jm} (equation (4.3)) is reduced to

$$
\gamma'_{jm} = \frac{P_{jm} h_j}{\sigma_\nu^2 + I^{\mathrm{D}}_{\mathrm{cu}}}. \tag{4.7}
$$

The effects of inter-DUE interference are not taken into account in equation (4.7). Instead, we are motivated to develop alternative methods to mitigate inter-DUE interference other than through RA, namely through the use of FBMC/OQAM for DUE pairs. Accordingly, we demonstrate that if DUE pairs use FBMC/OQAM, then there is no significant performance loss incurred by performing RA and power loading without taking into account the inter-DUE interference. This greatly reduces the complexity of the RA schemes and ensures that the power-loading objective function is convex. It is also more realistic as it makes no assumptions regarding the information a DUE pair possesses about every other DUE pair in the cluster.

Given the above simplifications, the two intermediate problems to be solved can be rewritten as follows.

RB Assignment

We assume each cellular user is assigned a single RB and that there are as many CUEs as RBs. Since we only consider pathloss in our channel model, RBs can be randomly assigned to CUEs. We then want to assign one RB to each DUE pair such that the interference experienced by each DUE pair from the CUEs is minimised. The interference from CUEs experienced by DUE pair j on RB r is given by

$$I_{jr} = \sum_{m \in b_r} I^{\mathrm{D}}_{\mathrm{cu}}. \tag{4.8}$$

The assignment problem can be specified as follows

$$\mathrm{P2}: \min_{\omega_{jr}} \sum_{j \in D} \sum_{r \in R} \omega_{jr} I_{jr}, \tag{4.9}$$

subject to

$$\omega_{jr} \in \{0,1\}, \forall j, r, \tag{4.9a}$$

$$\sum_{j \in D} \omega_{jr} \leq 1, \forall r \in R, \tag{4.9b}$$

$$\sum_{r \in R} \omega_{jr} = 1, \forall j \in D. \tag{4.9c}$$

Problem P2 is a combinatorial optimisation problem, made complicated by the fact that multiple DUE users may have the same optimal RB assignment. In line with the literature [147–149], we utilise the well-known Kuhn-Munkres algorithm (commonly known as the Hungarian method), to solve the uplink resource assignment problem for DUE pairs.

Power-Loading

Having assigned an RB to each DUE pair, power-loading can now be performed. The power-loading optimisation problem is similar to P1, with the discrete constraints (equations (4.6a-4.6c)), which relate to RB assignment, removed. Since RB assignment has already been performed, the objective function of optimisation P1, i.e., equation (4.6), can be simplified as follows

$$\max_{P_{jm}} \sum_{j \in D} \sum_{m \in b_j} \log(1 + \gamma'_{jm}). \tag{4.10}$$

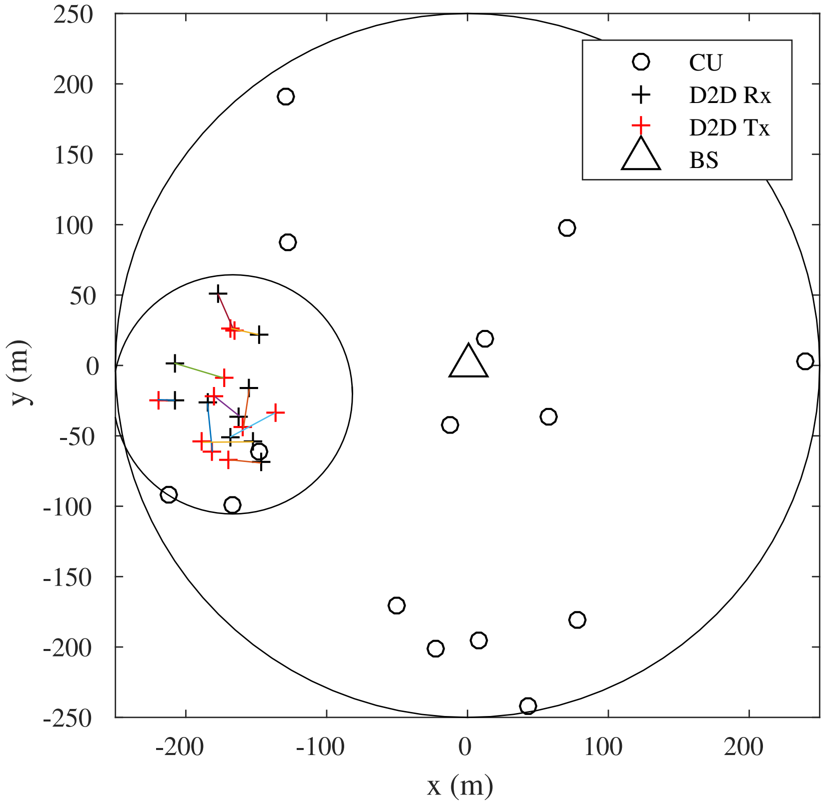


Figure 4.3: Example of clustered scenario consisting of 10 DUE pairs.

The resulting problem is clearly convex and similar to others in the literature, for example [150]. The solution can be readily obtained using an appropriate software package.

4.3 Single Cell Scenario: Results

We perform system-level simulations to investigate the co-existence of FBMC/OQAM and OFDM. Cellular users are uniformly distributed over the coverage area of the encompassing OFDMA cell. In the clustered scenario, the cluster centre is chosen according to a uniform distribution within the macro-cell area and DUE pairs are uniformly distributed within the cluster area. Fig. 4.3 illustrates an example of a clustered scenario with 10 DUE pairs.

Table 4.3 lists the key simulation parameters. After distributing both the CUEs and DUE pairs within the cell, we then perform RB assignment and power-loading as described in Section 4.2.2. The average rate per DUE pair is used as the main output metric from simulations. This metric is calculated for two different cases using the SINR expressions described in Section 4.2.2:

1. Case 1: DUE pairs use OFDM, CUEs use OFDM;
2. Case 2: DUE pairs use FBMC/OQAM, CUEs use OFDM.

Table 4.3: Single cell simulation parameters.

Parameter	Value
Inter-Site Distance (ISD)	500 m
macro-cell radius	250 m
subcarrier spacing	15 kHz
number of RBs	15, 25
number of CUEs	15, 25
scenario type	clustered or non-clustered
maximum cluster radius	ISD/5 m
minimum cluster radius	ISD/10 m
maximum DUE Tx. Rx. distance	(cluster radius) × 2/3
pathloss model	WINNER II scenario B1
CUE minimum SINR	10 dB
noise power per subcarrier [1] (σ_ν^2)	-127 dBm
maximum transmit power	24 dBm
number of iterations	40000

In both cases, we compare the predicted average rate per DUE pair calculated using γ'_{jm} (which does not take into account inter-DUE interference), with the actual average rate per DUE pair calculated using γ_{jm} (which takes into account inter-DUE interference).

4.3.1 Effects of inter-DUE interference for both FBMC/OQAM and OFDM

In the results for the single cell scenario, we show the adverse effects of inter-DUE interference when DUE pairs use OFDM, and how this type of interference may be considered negligible when FBMC/OQAM is instead used. We generate Cumulative Distribution Functions (CDFs) for the average rate per DUE pair for both the clustered (Fig. 4.4) and non-clustered (Fig. 4.5) scenarios in order to demonstrate the effects of inter-DUE interference for both waveforms.

In the clustered scenario in Fig. 4.4, we observe that when OFDM is used, the gap between the actual and predicted values of rate (calculated using γ_{jm} and γ'_{jm}, respectively) is significant. Conversely, when FBMC/OQAM is used, the actual values of achieved average rate per DUE pair are very close to those calculated without taking inter-DUE interference into account. Thus, FBMC/OQAM provides significant improvement over OFDM by virtue of its inherent ability to mitigate inter-DUE interference. In the non-clustered scenario in

[1] Noise power per subcarrier is calculated using the expression $-174\text{dBm/Hz}+10\log_{10}(15\text{kHz})$, where -174dBm/Hz is the background noise and 15kHz is the LTE subcarrier spacing.

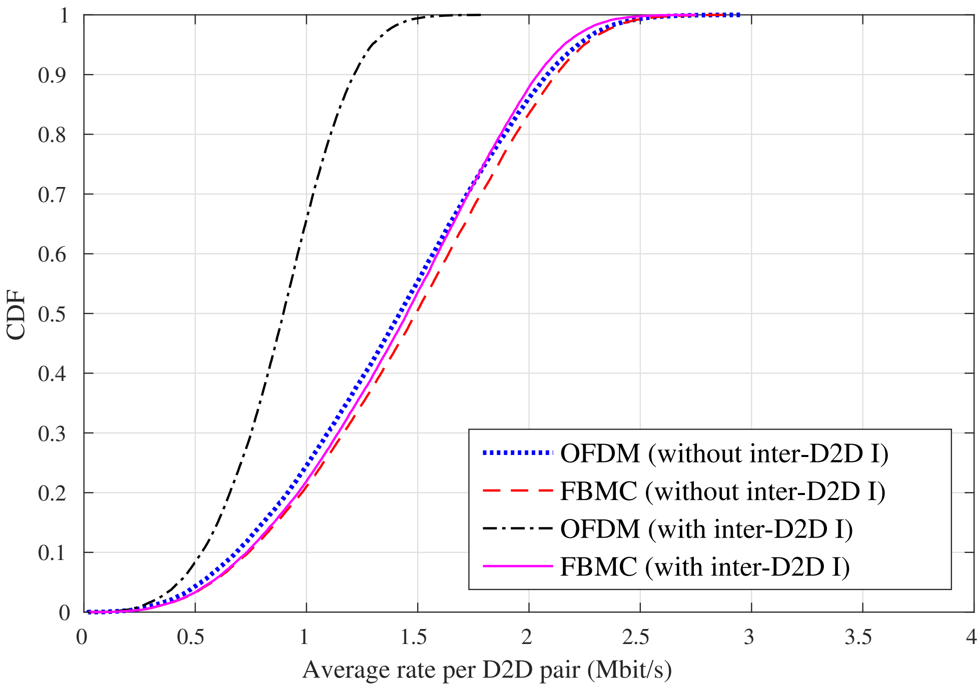


Figure 4.4: Clustered scenario consisting of 10 DUE pairs.

Fig. 4.5, we observe that the advantage of using FBMC/OQAM, even when inter-DUE interference is taken into account, is less than the corresponding clustered scenario. This is intuitive, as DUE pairs are farther apart in the non-clustered scenario and, hence, inter-DUE interference does not play such a significant role.

Consequently, we make two observations. First of all, inter-DUE interference plays a significant role in clustered DUE underlay communication. Second, we observe that while inter-DUE interference is detrimental to performance when OFDM is used, its effect is negligible when FBMC/OQAM is employed. Therefore, Fig. 4.4 and Fig. 4.5 show that permitting DUE users to use FBMC/OQAM can significantly facilitate the RA process in the considered scenarios.

Fig. 4.6 reinforces our observations. We display the CDF of the actual average rate per DUE pair, calculated using γ_{jm} when inter-DUE interference is considered, in a clustered scenario for 5 DUE pairs and 15 DUE pairs. We first observe that the average rate for both OFDM and FBMC/OQAM decreases as the number of DUE pairs increases, since each DUE transmitter must now use a lower transmit power to satisfy the interference constraint specified by equation (4.6d). We also observe that the benefit attributed to using FBMC/OQAM increases as the number of DUE pairs increases, i.e. the gap between the FBMC/OQAM and OFDM curves grows larger as the number of DUE pairs increases. This is due to inter-DUE interference becoming more significant as the number of DUE pairs is increased, despite the fact that pairs must now use less power.

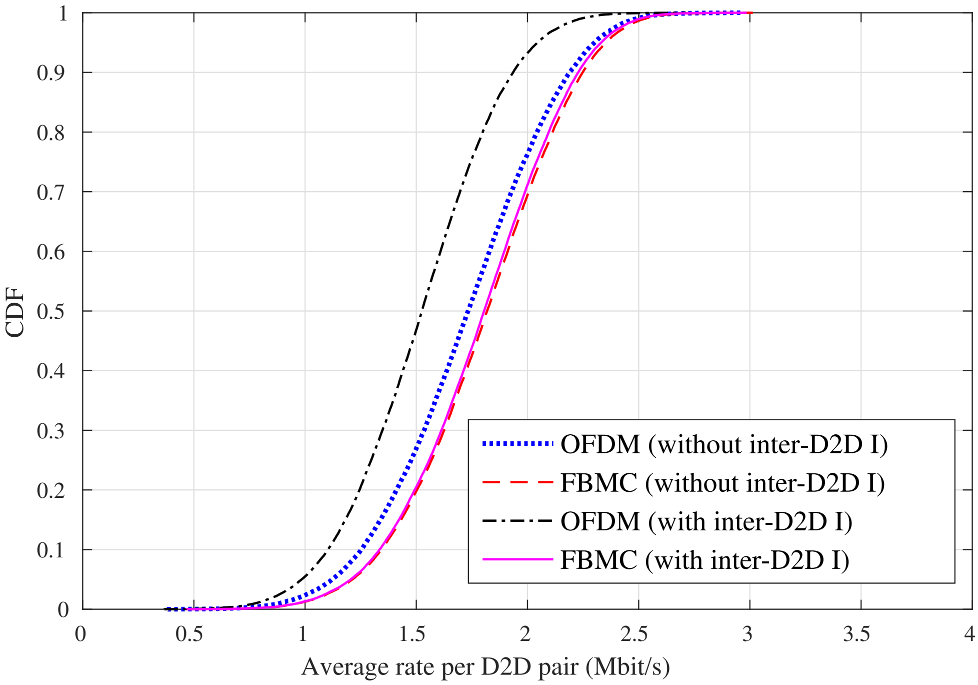


Figure 4.5: Non-clustered scenario consisting of 10 DUE pairs.

4.4 Multi-cell Scenario: System Model

We consider an OFDMA network with parameters selected based on the 3GPP LTE standard, as outlined in Table 4.4 in Section 4.5.1. Cells are modelled as hexagons, with the network consisting of a central cell of interest, an inner ring of direct neighbour cells, and an outer ring of additional cells (totalling nineteen cells). Two rings of cells are commonly used with the hexagonal cellular model, as interference from cells outside of this range can generally be considered negligible. Indeed, we verified this using our simulator, confirming that the addition of a third ring of cells has a negligible influence on the results.

We assume that each cell is fully loaded, with each CUE assigned a single uplink RB. DUEs coexist with the OFDMA cell by reusing a single uplink RB. In reality, LTE networks actually use SC-FDMA in the uplink. However, SC-FDMA is simply OFDMA in which users apply discrete Fourier transform precoding to their transmit signal. In the scope of our study, this precoding is inconsequential and we therefore do not consider it.

CUEs are distributed throughout the entire network according to a Poisson point process. DUEs are employed in high-rate spatially clustered applications such as process control, robotics control, or machine-to-machine communications. To capture this clustering effect in our model, we distribute DUE transmitters in the network using a Matérn point process. For each DUE transmitter, we distribute a receiver at a distance d according to a uniform random variable $U_{[a,b]}$, with a and b representing the minimum and maximum distance, respectively.

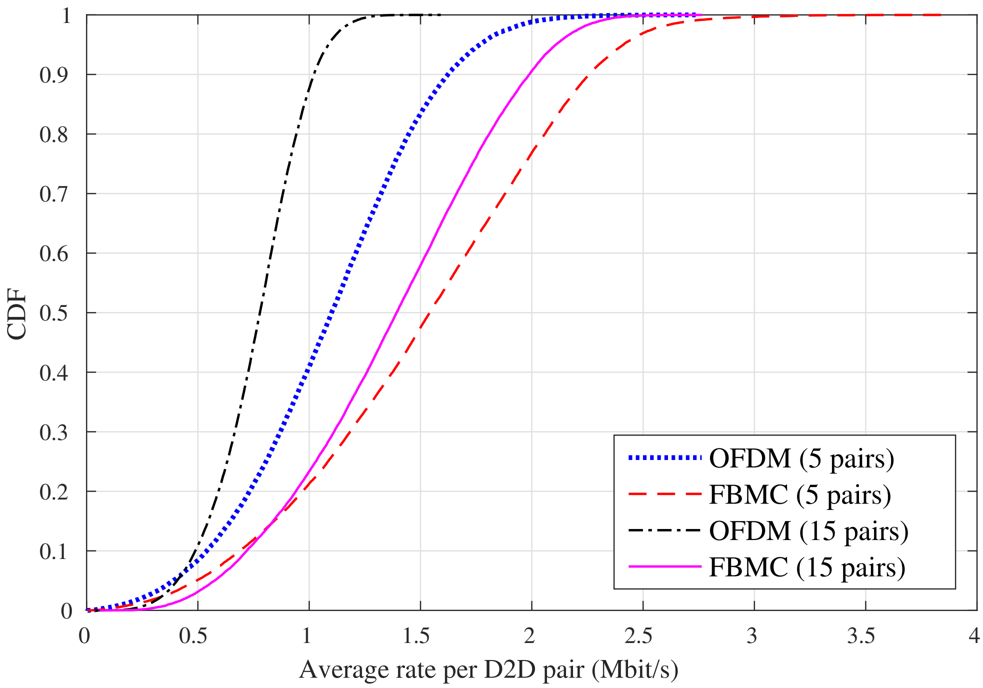


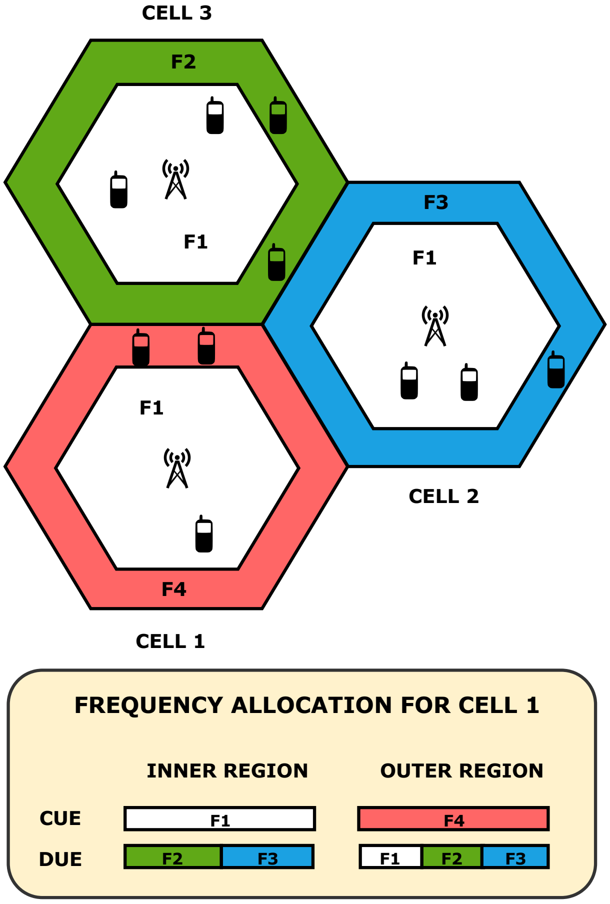
Figure 4.6: Average rate per DUE pair for different numbers of DUE pairs.

We consider the use of strict FFR. Fig. 4.7 illustrates the division of sub-bands between cells. The CUEs in the inner region of each cell are provisioned using a common set of sub-bands. Frequency reuse three is employed for the outer regions of the cells, with cell-edge CUEs in these regions provisioned from one of three sets of sub-bands. Machine-type DUEs are permitted to reuse the spectral resources of cellular users according to the scheme outlined in [4] for D2D communication operating in a network employing strict FFR. Hence, DUEs in the inner region of a cell may reuse the spectral resources assigned to CUEs in the outer regions of neighbouring cells[2]. DUEs in the outer region of a cell may use any spectral resource, except the set assigned to CUEs in the same region.

The scenario under evaluation in this chapter is similar to underlay, since neither D2D devices nor cellular users have exclusive use of the available spectrum across the entire network. However, we also note that due to the manner in which sub-bands are assigned through the use of fractional frequency control, cellular users and D2D devices do not use the same spectral resources within the same region. Hence, the system could be described as overlay at a local level and underlay at a system-wide level, and does not conform to the strict definition of either term.

The ratio of the radius of the inner region (R_{inner}) to the radius of the cell (R_{cell}) is an important parameter in strict FFR systems and influences how sub-bands are divided

[2] Note that DUEs in the inner region of a cell may not reuse the resources of cellular users in the outer region of the same cell, as the DUEs would always be closer to the BS than the CUEs. Since interference in the uplink occurs at the BS, this could result in significant interference from DUEs to CUEs, which is precisely what the reuse scheme is designed to prevent.



FREQUENCY ALLOCATION FOR CELL 1

	INNER REGION	OUTER REGION
CUE	F1	F4
DUE	F2 F3	F1 F2 F3

Figure 4.7: Frequency allocation for multi-cell model. The inner region of each cell uses the same set of sub-bands, while reuse three is employed in the outer regions. CUEs and DUEs are allocated sub-bands in a manner that aims to reduce interference between them, according to the scheme outlined in [4].

between regions. We follow the approach used in [151], and choose the ratio $R_{\text{inner}}/R_{\text{cell}}$ to be 0.65, which was shown in [152] to maximise the average network throughput for uniformly distributed CUEs. Given K_{band} available sub-bands in the system, we can determine the number of resources allocated to each region as follows [152]

$$K_{\text{inner}} = \left[K_{\text{band}} \left(\frac{R_{\text{inner}}}{R_{\text{cell}}} \right)^2 \right], \tag{4.11}$$

$$K_{\text{outer}} = [(K_{\text{band}} - K_{\text{inner}})/3]. \tag{4.12}$$

4.4.1 Interference Model

To model interference between users using each of the aforementioned waveforms considered in this article, we follow the same approach as used in Section 4.2.1 [2]. However, whereas these previous analyses were based on interference tables with a spectral granularity of

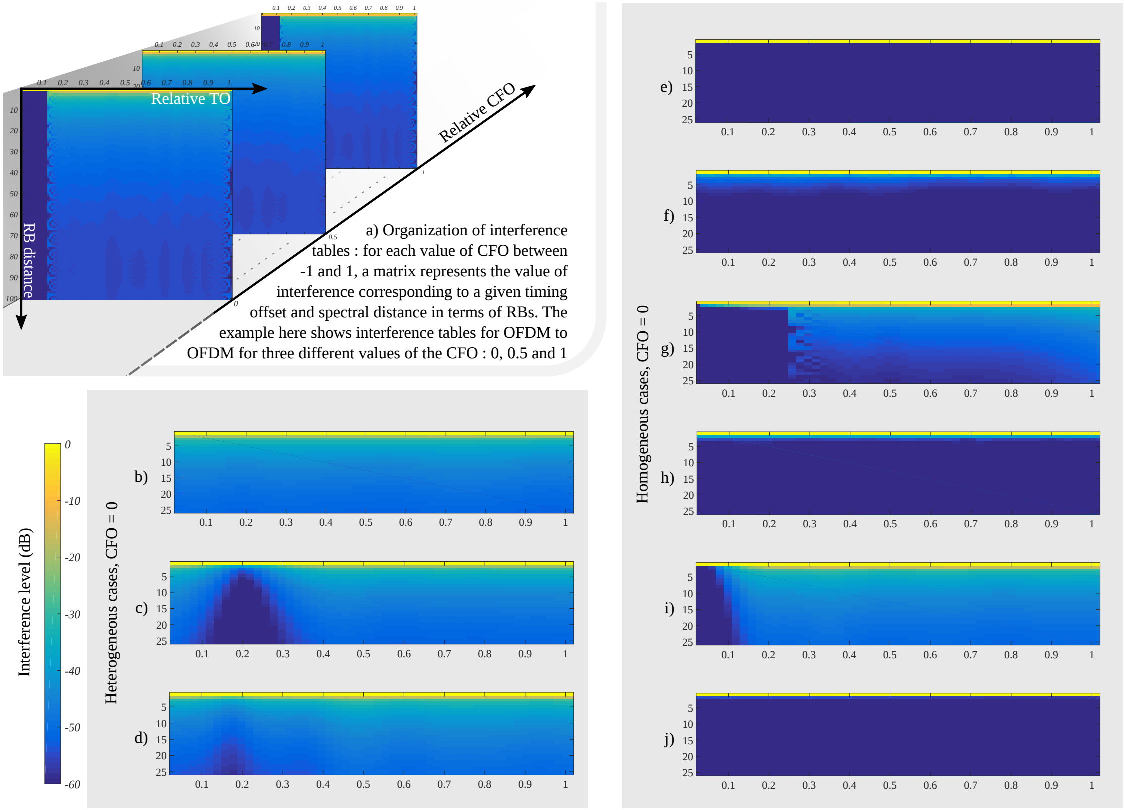

a) Organization of interference tables : for each value of CFO between -1 and 1, a matrix represents the value of interference corresponding to a given timing offset and spectral distance in terms of RBs. The example here shows interference tables for OFDM to OFDM for three different values of the CFO : 0, 0.5 and 1

Figure 4.8: Representative examples of interference tables. TO is expressed in proportion of $T+T_{\text{cp}}$ of a reference OFDM configuration and CFO is expressed relative to ΔF. Values lower than -60 dB appear in dark blue.
a) Organisation of interference tables and OFDM to OFDM example. b) FBMC/OQAM to OFDM. c) f-OFDM to OFDM. d) UFMC to OFDM. e) FBMC/OQAM to FBMC/OQAM. f) FBMC-PAM to FBMC-PAM. g) GFDM to GFDM. h) FMT to FMT. i) UFMC to UFMC. j) f-OFDM to f-OFDM.

one subcarrier spacing, we base the present system-level analysis on interference tables that are defined at the RB level. This is necessary to be able to carry out system-level studies, as RA and other upper layer procedures operate with a minimum granularity of one RB. Besides, whereas most studies on asynchronous networks rely on average values of interference [1, 133], we consider the interference between given pairs of waveforms for specific values of TO δ_{t}. Moreover, we also take CFO into account. We therefore use three-dimensional interference tables which give the interference value at a given RB distance for each possible value of the TO, δ_{t}, and CFO, δ_{f}. We consider that the TO and CFO between users are uniformly distributed in a given interval so that $\delta_{\text{t}} \sim U_{[-\delta_{\text{t}}^{\text{max}},\delta_{\text{t}}^{\text{max}}]}$ and $\delta_{\text{f}} \sim U_{[-\delta_{\text{f}}^{\text{max}},\delta_{\text{f}}^{\text{max}}]}$.

To present our interference model, we display in Fig. 4.8 some of the interference tables that we employ. In particular, we present in Fig. 4.8-a the structure of the interference tables as they are represented in our system-level simulator. For each value of δ_{t} and δ_{f}, the tables provide the corresponding level of interference, up to a maximum spectral distance of 100 RBs. Note that we consider heterogeneous scenarios in which DUEs use an alternative waveform and CUEs employ OFDM, and more advanced homogeneous scenarios in which both CUEs and DUEs use an alternative waveform. To model the interference between different users in these different set-ups, we therefore need to adopt homogeneous interference tables, from a given waveform to the same waveform, and heterogeneous ones, from a given waveform to OFDM and from OFDM to a given waveform.

In Fig. 4.8, for each table *Waveform A* to *Waveform B*, an interfering user using *Waveform A* is active on an RB of index 0, and we show the interference power seen by a victim user using *Waveform B* at a given spectral distance specified in number of RBs, and for given values of the TO δ_{t}. Note that, due to space limitations, we present interference tables only in the case where there is no CFO, i.e. $\delta_{\mathrm{f}} = 0$, and only for spectral distances lower than 25 RBs.

4.4.2 Resource Allocation

RBs are assigned under the condition that an RB may only be assigned to a single CUE, and reused by a single DUE, in a given cell. CUEs transmit on the Physical Uplink Shared Channel (PUSCH) and use a power control procedure [19] that assigns each CUE a power level that results in acceptable signal reception at the BS. In the procedure used, the pathloss between each CUE and the BS is estimated and compensated for to satisfy the power that the BS expects to receive over a single RB $P_{\mathrm{O\_PUSCH}}$. The maximum power at which a CUE may transmit is also capped at P_{cmax}.

Our focus in this chapter is on evaluating the relative performance of the waveforms under consideration for direct communication between devices/equipment in spatially clustered use cases, not on proposing a new RA scheme. Hence, to avoid bias towards any particular scheme, we consider a simple power allocation scheme for machine-type DUEs whereby they are permitted to transmit at maximum power, which is capped by the controlling BS.

Section 4.5 provides detailed insight into the performance of asynchronous communication for various waveforms, compared to a baseline of synchronous communication, allowing informed decisions to be made regarding the choice of waveform for both types of communication. We highlight, however, that the decision of whether to use synchronous or asynchronous communication is multifaceted and there are many reasons why an operator may decide to employ asynchronous communication for MTC scenarios. As mentioned, the overhead associated with achieving and maintaining synchronous communication for MTC may be unattractive. Removing the synchronisation procedure for DUEs could help to reduce the latency experienced by these devices. Asynchronous DUE communication also removes several duties of control from the BS, potentially enabling the network operator to treat RA for high-rate clustered MTC scenarios in a different manner than for CUEs. For example, the network could release spectral resources to a smart factory without actively managing the directly communicating machinery, which may operate autonomously or via a local controller.

4.4.3 Channel Modelling

CUEs in the same cell do not interfere with each other, as we assume they are perfectly synchronised by the BS. Therefore, there are four main interference types requiring consideration:

1. DUE pairs interfere with the CUEs' transmissions. Since we are investigating uplink resource sharing, this interference is observed at BSs.
2. Conversely, the CUEs interfere with the DUE pairs at DUE receivers.
3. DUEs interfere with each other (inter-DUE interference).
4. CUEs in different cells are not synchronised and, hence, interfere with each other (inter-CUE interference).

Owing to their popularity in the existing literature [136, 153, 154], we employ the WINNER II channel models [146] to provide us with a distance based path loss, which also incorporates the probability of line-of-sight. Distinct path loss models are used for the different types of links in the system to represent the network in a realistic manner. Path loss models employed for D2D channels have been modified so that both transceivers in a D2D link are the same height above the ground. The distribution of shadow fading is log-normal, with the standard deviation specified by the Winner II channel models for each scenario.

4.4.4 Performance Measures

Below, we present several metrics that we will use to evaluate the performance of the system. All metrics are evaluated for DUEs and CUEs in the central cell, which represents the cell of interest.

SINR

The SINR of a CUE j in the central cell o using RB k is given by:

$$\gamma_{j_o}^{k} = \frac{P_{j_o}^{k} h_{j_o B}^{k}}{\sigma_\nu^2 + I_{\mathrm{C_N}} + I_{\mathrm{D_N}} + I_{\mathrm{D_S}}}, \tag{4.13}$$

where $P_{j_o}^{k}$ is the transmit power of the CUE, $h_{j_o B}$ is the channel gain between the j^{th} CUE and the BS of the central cell o, and σ_ν^2 is additive white Gaussian noise variance. $I_{\mathrm{C_N}}$ is the interference from CUEs in neighbouring cells and is given by

$$I_{\mathrm{C_N}} = \sum_{n\in N} \sum_{c_n \in C_n} \sum_{r\in R} P_{c_n}^{r} h_{c_n B}^{r} \Omega_{\mathrm{wf}_{c_n} \to \mathrm{wf}_{j_o}}(|r-k|, \delta_{\mathrm{t}}, \delta_{\mathrm{f}}), \tag{4.14}$$

where n indexes the set of neighbouring cells N, c_n indexes the CUEs in the set C_n of CUEs in the n^{th} neighbouring cell, and r indexes the set of RBs R available to the system. $P_{c_n}^{r}$ is the transmit power of the c_n^{th} CUE operating on RB r, $h_{c_n B}^{r}$ is the channel gain between

the c_n^{th} CUE and the BS of the central cell. If the c_n^{th} CUE is not operating on RB r, then $P_{c_n}^r$ is 0. Finally, $\Omega_{\mathrm{wf}_{c_n} \to \mathrm{wf}_{j_o}}(|r-k|, \delta_{\mathrm{t}}, \delta_{\mathrm{f}})$ is the fraction of power injected by CUE c_n using waveform wf_{c_n} and RB r onto CUE j_o using waveform wf_{j_o} and RB k, at a TO of δ_{t} and CFO δ_{f}. For synchronous communication, both δ_{t} and δ_{f} can be set to 0.
$I_{\mathrm{D_N}}$ is the interference from DUEs in the neighbouring cells and is given by

$$I_{\mathrm{D_N}} = \sum_{n \in N} \sum_{d \in D_n} \sum_{r \in R} P_{\mathrm{D}} h_{d_n B}^r \Omega_{\mathrm{wf}_{d_n} \to \mathrm{wf}_{j_o}}(|r-k|, \delta_{\mathrm{t}}, \delta_{\mathrm{f}}), \tag{4.15}$$

which is defined in a similar fashion to equation (4.14), where D_n represents the set of DUEs in the n^{th} neighbouring cell, and P_{D} is the transmit power of DUE devices. Finally, $I_{\mathrm{D_S}}$ represents the interference from DUEs in the same cell, i.e. the central cell, and is formulated in a similar fashion to equation (4.15).

The SINR of a DUE d in the central cell o operating on RB r is given by:

$$\gamma_{d_o}^r = \frac{P_{\mathrm{D}} h_{d_o}^r}{\sigma_\nu^2 + I_{\mathrm{C_N}} + I_{\mathrm{C_S}} + I_{\mathrm{D_S}} + I_{\mathrm{D_N}}}, \tag{4.16}$$

where P_{D} is the transmit power of the DUE devices, $h_{d_o}^r$ is the channel gain between the transmitter and receiver of the d^{th} DUE using RB r, and σ_ν^2 is additive white Gaussian noise variance. $I_{\mathrm{C_S}}$ and $I_{\mathrm{C_N}}$ represent the aggregate interference from CUEs in the same cell and neighbouring cells, respectively. $I_{\mathrm{D_S}}$ and $I_{\mathrm{D_N}}$ represent the aggregate interference from DUEs in the same cell and neighbouring cells, respectively. The expressions for each of the above aggregate interference terms are similar to equations (4.14) and (4.15), with the channel gain h considered between the interfering device and the DUE receiver.

Achieved Rate

We are also interested in the rate achieved by devices, after the bandwidth efficiency of each waveform has been taken into account. The rate of a device using a waveform wf can be calculated as

$$b = \Phi_{\mathrm{wf}} B \log_2(1+\gamma)[b/s], \tag{4.17}$$

where B is the bandwidth of an LTE RB, and Φ_{wf} is the bandwidth efficiency of waveform wf presented in Table 4.1, which is directly computable based on the waveform parameters presented in the same table.

4.5 Multi-cell Scenario: Evaluation of System Performance

4.5.1 Scenario Under Investigation

We first present detailed results for the scenario defined by the parameters listed in Table 4.4. For each set of results, we compare the asynchronous performance of all waveforms

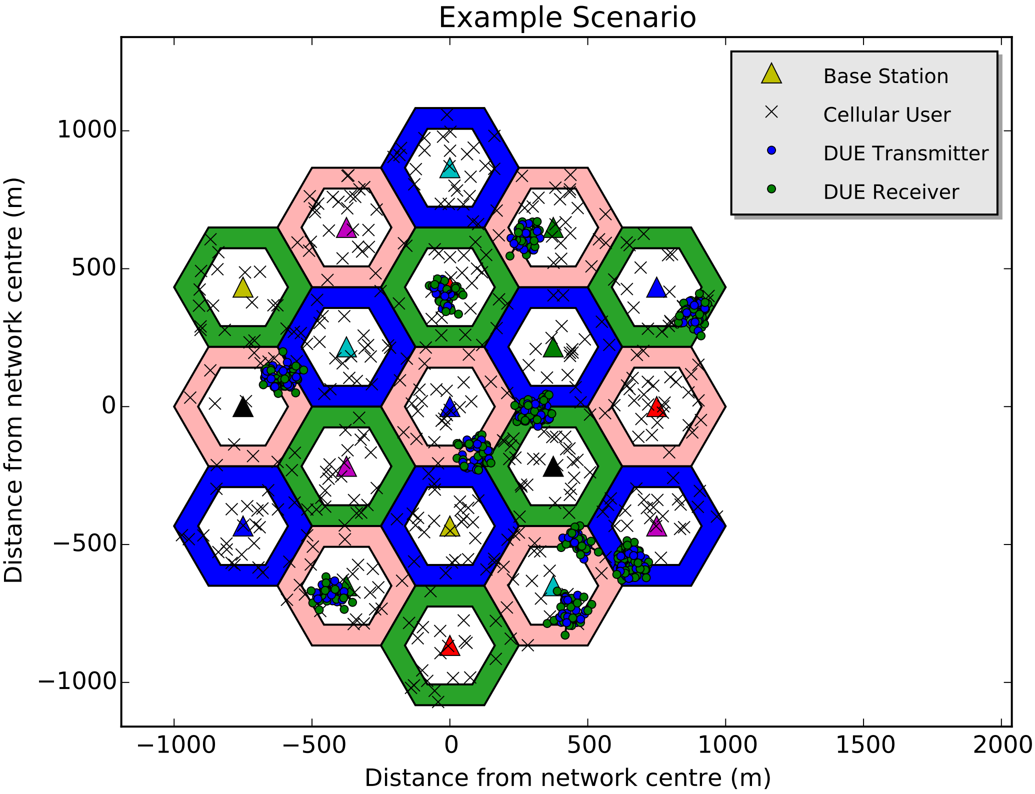


Figure 4.9: Example scenario consisting of 19 cells, with each region coloured according to the spectral resources permitted for use. Each x represents an ordinary cellular user, whereas DUEs involved in direct communication are clustered in groups.

under consideration, and use the performance of synchronous OFDM as an baseline for comparison.

We highlight that the synchronous OFDM case serves as an idealistic baseline for comparison with the asynchronous cases and that, in reality, achieving synchronous communication for the DUEs would be challenging. This is true even if the D2D communication is network assisted [129]. The BS applies a timing advance to ensure all signals reach the BS simultaneously; however, due to the varying distances between DUE pairs, signals will not arrive at DUE receivers simultaneously and hence the DUE pairs will not be fully synchronised with one another. In addition, a DUE pair may span multiple cells, further complicating the issue.

We therefore assume quasi-orthogonality in which all TOs are absorbed by an extended CP of 20%[4] for the synchronous OFDM baseline case. We also do not consider CFO in this case for two reasons. First, the scenario that we are considering typically consists of low mobility, resulting in negligible Doppler shifts and frequency offsets. Secondly, the 3GPP

[3] Noise power per RB is calculated using the expression $-174\text{dBm/Hz}+10\log_{10}(180\text{kHz})$, where -174dBm/Hz is the thermal noise and 180kHz is the bandwidth of an LTE RB.

[4] The value of 20% was chosen as it is similar to the size of the extended CP option in LTE.

Table 4.4: Multi-cell simulation parameters.

Parameter	Value
Cell Radius	250 m
Inner Radius	163 m
Number of Cells	19
CUEs Per Square Km	200
DUEs Per Cluster	30
Clusters Per Square Km	3
Average Cluster Radius	60 m
Average transmitter (Tx)-receiver (Rx) Distance	Uniformly distributed in the range [10, 50] m
Subcarrier Spacing (ΔF)	15 kHz
Noise Per RB [3] (σ_ν^2)	-116 dBm
Number RBs in system	50
$P_{\text{O\_PUSCH}}$	-96 dBm
Max Tx Power CUE	24 dBm
Max Tx Power DUE	-5 dBm
BS Antenna Gain	15 dBi
User Equipment (UE) Antenna Gain	0 dBi
BS Noise Figure	5 dB
UE Noise Figure	9 dB
Max TO	$T + T_{cp}$
Max Local Oscillator (LO) Inaccuracy	2.5 ppm
Waveforms	OFDM, FMT, FBMC/OQAM, FBMC-PAM, GFDM, f-OFDM, UFMC
Number of Iterations	10000

standards specify stringent frequency errors for UEs of less than +/- 0.1 parts per million (ppm)[155] compared to the carrier frequency received from the BS.

In contrast, for asynchronous scenarios, we consider TOs uniformly distributed in the range of 0 to $T + T_{cp}$, where T is the length of an OFDM symbol and T_{cp} is the length of the CP. We also consider less stringent hardware-related frequency error requirements, with LO inaccuracies of +/- 2.5 ppm[5] permitted.

The cell radius value of 250m is based on the 3GPP LTE system scenarios [156], representing an urban macro-cell environment. The antenna gain values, noise figures, and the carrier frequency value are also based on [156]. The values for the maximum CUE transmit power, subcarrier spacing, and number of RBs are based on the LTE standard, with 50 RBs corresponding to a bandwidth of 10MHz. The maximum DUE transmit power of -5 dBm

[5] Generally, strict frequency error requirements require more accurate and expensive clocks. 2.5ppm is the stated frequency accuracy of the NI USRP-292x range of devices.

was chosen as we found through experimentation that it yielded good results. The effects of varying the maximum DUE transmit power will be discussed later in this section.

Fig. 4.9 shows an example of a typical simulation scenario. We explore the case whereby each macro-cell is fully loaded, with all available RBs being utilised, and hence consider a large number of CUEs per square kilometre to ensure this. The parameters relating to the size and frequency of occurrence of clusters are scenario dependant. A cluster of radius 60m, containing 30 inter-communicating devices and with an average of 3 clusters per square kilometre might, for example, represent a factory in an urban area with moderate industrial activities.

Simulating the network for every possible combination of waveform pairs would be impractical and unnecessary. Hence, we only examine the most realistic combinations:

1. Case 1: DUE pairs use an alternative waveform and CUEs continue to use OFDM;
2. Case 2: Both DUE pairs and CUEs use an alternative waveform.

We also examine the effects of the TO on the relative performance of all waveforms, ranging from perfectly synchronised to fully asynchronous communication. An analogous investigation is performed for CFO by varying LO inaccuracies.

4.5.2 System Performance

DUE SINR Performance

Fig. 4.10 presents box plots summarising the SINR distribution for DUEs according to each considered waveform couple. A solid horizontal line in each box represents the median, while the mean is marked with a dashed horizontal line. The ideal baseline OFDM case, assuming no timing or frequency offsets, performs quite well and achieves an average SINR value of approximately 22dB. This, however, reduces to approximately 13dB when asynchronous communication is considered, with UFMC and GFDM exhibiting similar average values. This reduction in performance can be attributed to increased leakage interference between DUEs owing to the large sidelobes exhibited by these waveforms. When both CUEs and DUEs employ an alternative waveform in the set {FBMC/OQAM, FMT, FBMC-PAM, f-OFDM}, performance comparable to the baseline case is achieved even though communication is asynchronous, as the filtering operations substantially reduce the sidelobes of these waveforms. In addition, any leakage remaining after filtering is absorbed by the guard subcarrier for FBMC/OQAM, FMT, and f-OFDM, which explains how two waveforms with different filters, such as f-OFDM and FBMC/OQAM, can present with similar SINR values.

Interestingly, this same set of waveforms performs quite well in the coexistence scenarios in which CUEs use OFDM and DUEs use an alternative waveform, with average values approximately 3dB less than in the baseline case, but up to 6dB greater than asynchronous OFDM. Again, we can explain this by highlighting the increased spectral containment of these waveforms over OFDM, resulting in less inter-DUE interference.

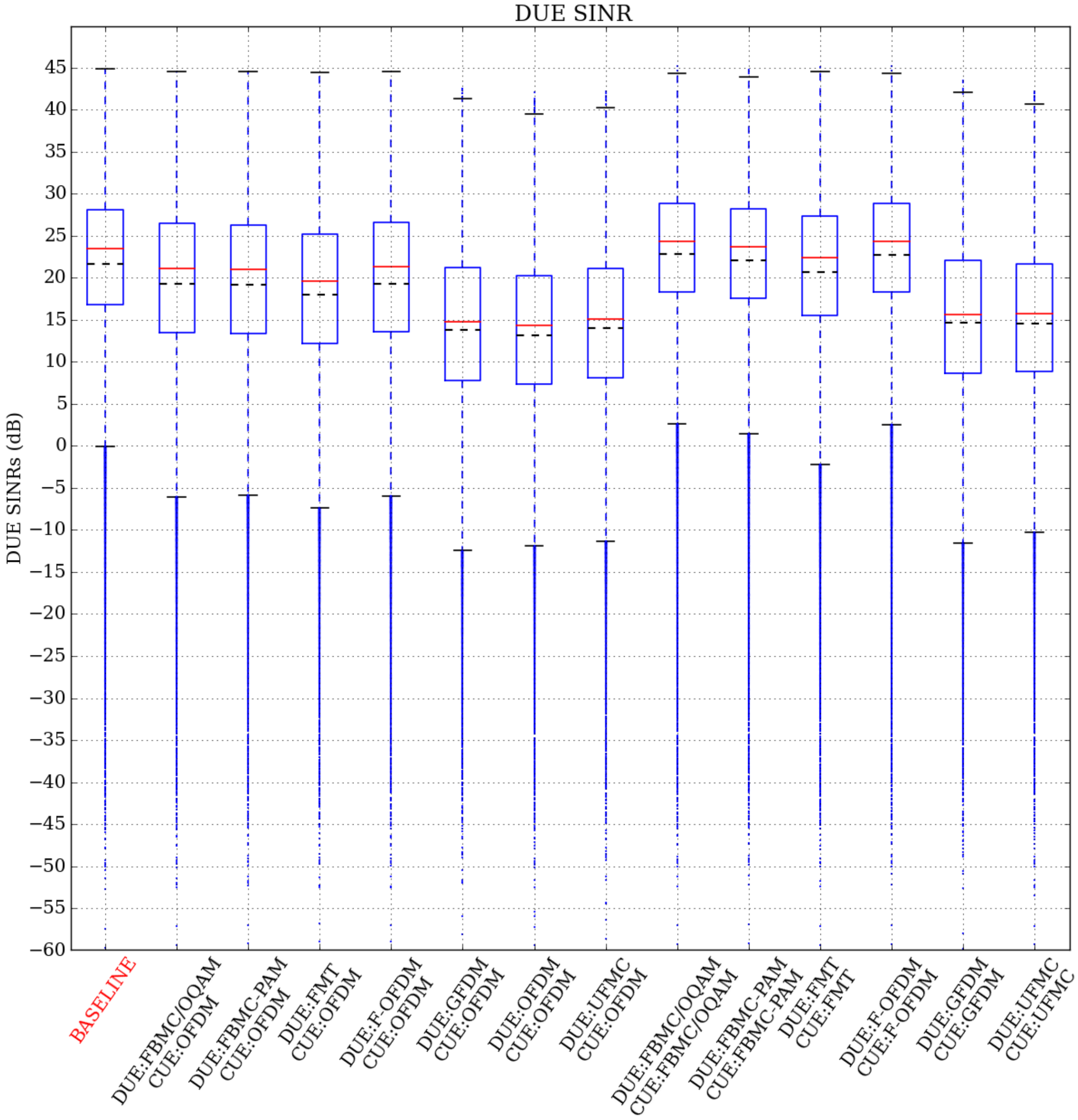


Figure 4.10: Box plots for DUE SINR. A large performance increase can be obtained by choosing an appropriate alternative waveform.

We note that the number of outliers is relatively small (approx. 2.2%) compared to the number of DUEs in the data set. The presence of outliers is not unusual; while the majority of DUEs will experience a similar SINR to one another, especially favourable or unfavourable channel conditions will inevitably result in DUEs with SINRs that are considerably higher or lower than average.

DUE Rate Performance

Fig. 4.11 shows the achieved rate of DUE pairs for each waveform. The greatest performance is achieved when both CUEs and DUEs use FBMC/OQAM, closely followed by

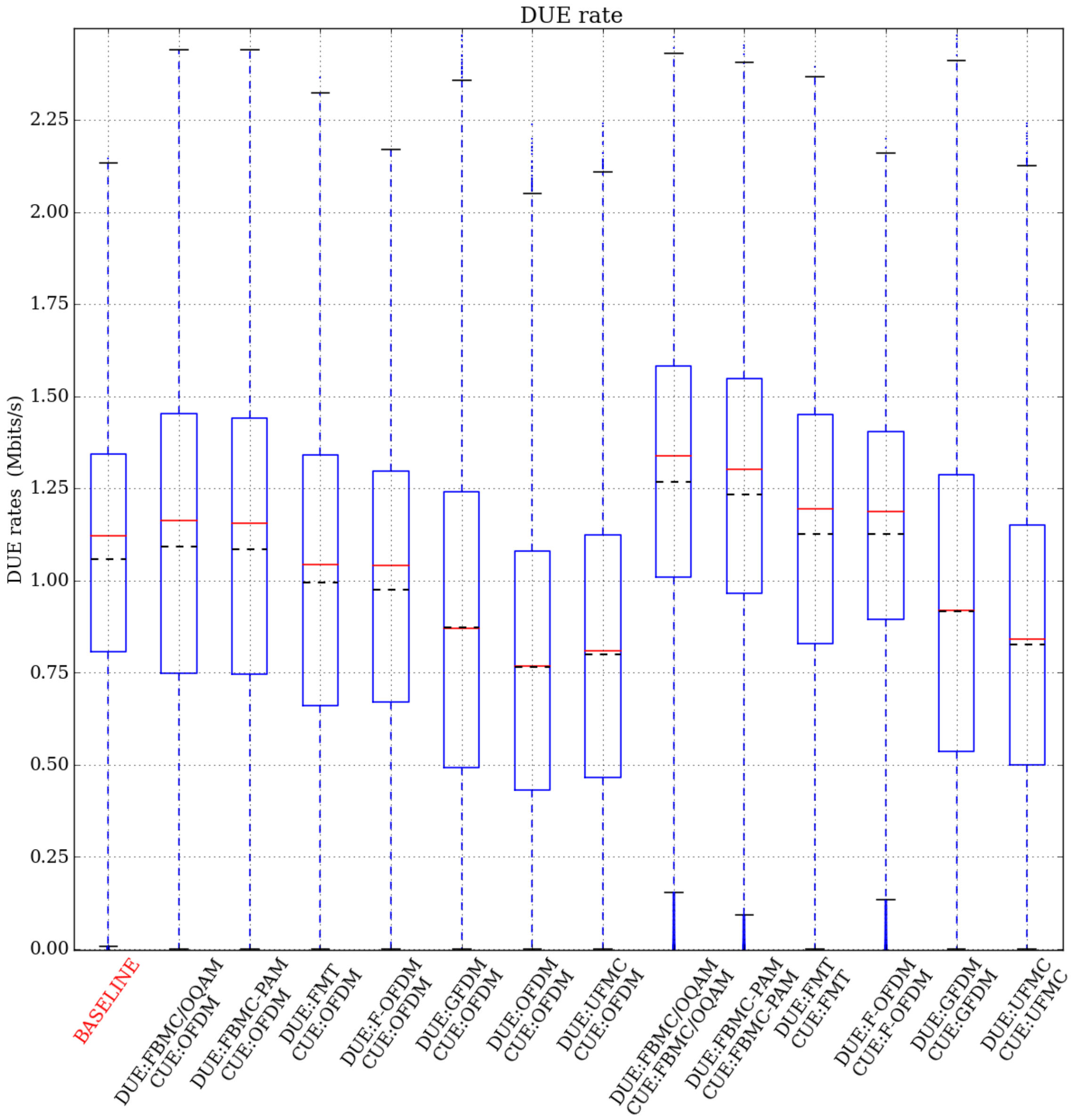


Figure 4.11: Rate performance of DUEs taking into account bandwidth efficiency.

FBMC-PAM. This is understandable, as these two waveforms have the best bandwidth efficiency (see Table 4.1) out of the alternative waveforms considered due to the fact that they do not employ a CP. This explains why FBMC/OQAM outperforms f-OFDM in terms of rate performance, despite both exhibiting similar SINR distributions in Fig. 4.10. Furthermore, recall from Section 4.1.1 that FMT employs a guard band between each subcarrier, with the guard band width set so that FMT has the same spectral efficiency as OFDM.

In the coexistence scenarios, in which CUEs use OFDM and DUEs use an alternative waveform, both FBMC/OQAM and FBMC-PAM again exhibit the best performance. In both of these cases, the achieved rate is marginally greater than for the synchronous OFDM baseline case and approximately 43% greater than for the asynchronous OFDM case. This is an encouraging result, as machine-type DUEs could enjoy the benefits of asynchronous communication without suffering any degradation in performance. It also allows for the possibility that future networks will permit multiple waveforms, whereby different services employ the waveform that is best suited to them. We note that although the average data rate for these two cases is marginally greater than for the synchronous OFDM baseline case, it can be observed in Fig. 4.10 that the baseline case actually achieves a higher SINR. This can again be attributed to the bandwidth efficiencies of the waveforms, with the 20% efficiency loss due to the extended CP in the baseline case causing significant data rate degradation.

CUE Performance

It is imperative that CUE performance not be significantly degraded by the inclusion of MTC in the network. Fig. 4.12 demonstrates that the average DUE to CUE interference is quite low. This can be attributed to two factors: the low transmit power of DUEs (-5dB) and the use of strict FFR.

In the synchronous baseline case, CUEs will still suffer slightly from leakage interference from DUEs, owing to the fact that different cells in LTE are misaligned in time. The use of an appropriate alternative waveform by both sets of users can therefore assist in reducing the interference that CUEs experience from DUEs. Hence, in cases where both CUEs and DUEs use an alternative waveform from the set {FBMC/OQAM, FMT, FBMC-PAM, f-OFDM}, the interference experienced by CUEs is lower than in the baseline case.

In all other cases, the interference is comparable to the noise per RB. However, Fig. 4.12 demonstrates that employing a different waveform for DUEs, while CUEs continue to use OFDM, does little to mitigate DUE to CUE interference; its main benefit is to increase the performance of the DUEs themselves. This is because for CUEs, the interference from DUEs using the same RB will generally be a greater factor than leakage. In contrast, for DUEs, leakage interference from other DUEs in close proximity is the dominant type of interference, and hence they can benefit from adopting a waveform with better spectral localisation than OFDM.

The presence of outliers suggests that while the majority of CUEs suffer little degradation to their performance, a small number of users suffer a large reduction in performance. These users are victims of the specific spatial distribution of transmitting users at that instant, in which strict FFR and the low transmit power of DUEs fail to offer sufficient protection. In these cases, additional protection is needed to ensure that interference to CUEs is kept at an acceptable level and that the minority of users who suffer significant degradation can also obtain adequate performance. This may take the form of intelligent RA schemes that aim to protect vulnerable CUEs by assigning resources in a manner that reduces DUE to CUE interference. As stated previously, such schemes are not within the scope of this chapter as we are solely interested in demonstrating the effect on performance of employing different waveforms.

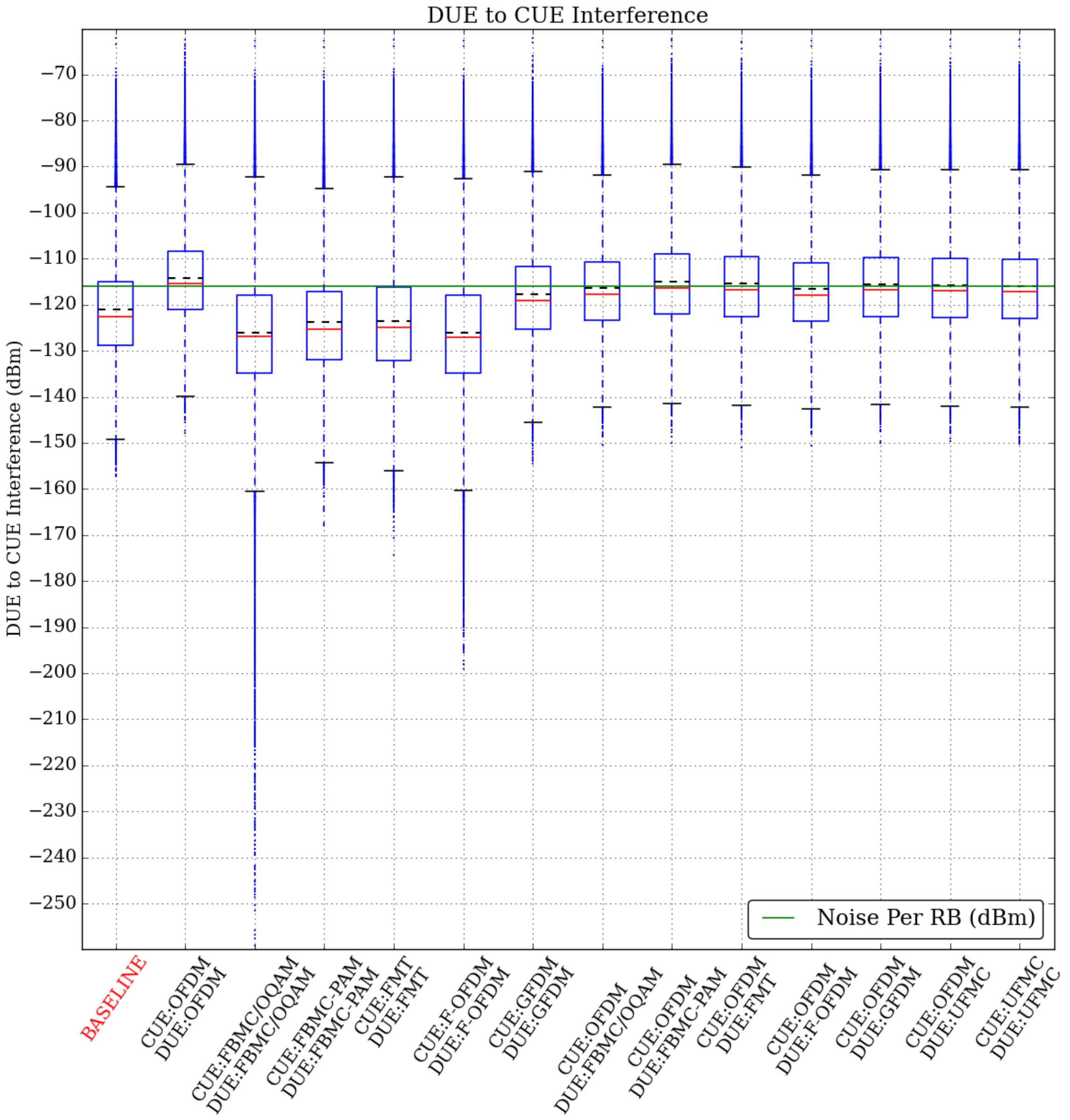


Figure 4.12: Box plots for DUE to CUE interference. The DUE to CUE interference is similar to the value of noise per RB for coexistence cases.

4.5.3 DUE Transmit Power

We investigate the effect that DUE transmit power has on system performance by varying the DUE transmit power from -15dBm to 15dBm in 5dBm increments, while holding all other parameters at the same value as in Table 4.4. We do not include cases in which both CUEs and DUEs use an alternative waveform, as we are more interested in the coexistence cases.

As intuition suggests, Fig. 4.13 shows that increasing the DUE transmit power will increase DUE SINR at the cost of increased interference to CUEs. The case in which DUEs employ OFDM for asynchronous communication exhibits the worst performance, as the large sidelobes of OFDM cause interference with neighbouring users in the resource grid. Leakage interference from DUE pairs in other cells will be present even in the synchronous case as neighbouring cells do not achieve time alignment. While FBMC/OQAM and f-OFDM successfully mitigate this type of leakage interference, it becomes significant at high transmit powers for synchronous OFDM and, hence, the curve representing the baseline case begins to taper as the transmit power is increased. So, while the synchronous baseline case achieves the greatest performance for low transmit powers, it is overtaken by both FBMC/OQAM and f-OFDM at a transmit power of 7.5dBm as leakage from other cells becomes significant.

In particular, we draw the readers' attention to two points.

1. First, for the case in which DUEs do not use an alternative waveform from the set {FBMC/OQAM, FMT, FBMC-PAM, f-OFDM}, successively higher transmit powers provide increasingly diminishing returns since increasing DUE transmit power will also increase the inter-DUE leakage interference. This is evident in Fig. 4.13, in which the set of curves at the bottom of the upper sub-plot gradually begin to level off as the DUE transmit power is increased.
2. Secondly, the benefit to DUEs of using an alternative waveform will be greater at higher values of DUE transmit power since inter-DUE leakage will be more prominent. However, as the DUE transmit power is increased, there is a linear increase in the interference experienced by CUEs.

The main consequence of these observations is that higher DUE transmit powers provide increasingly diminishing returns unless an alternative waveform that adequately mitigates leakage interference is employed. A maximum permissible transmit power should be chosen for DUEs that achieves a balance between adequate average DUE SINR and an acceptable level of interference to CUEs. The value of -5dBm chosen in Table 4.4 reasonably achieves this, with DUEs achieving a SINR close to 20dB when they employ FBMC/OQAM while limiting interference to CUEs to approximately the noise value per RB. We also note that we can trade off some DUE performance for reduced interference to CUEs. We observe, however, that the DUE to CUE interference is at a minimum for the synchronous baseline case, as no leakage interference within the same cell is present.

4.5.4 Cell Radius

In this subsection, we investigate the influence that cell size has on performance by varying the cell radius from 200m to 1000m in 100m increments while holding all other parameters at the same value as in Table 4.4. We display the results in Fig. 4.14. For cell radii under 500m, we consider an urban environment and use the appropriate pathloss models for this scenario, while for cell radii greater than 500m, we consider a suburban environment.

At the smallest cell radius considered (200m), average DUE SINR is at its lowest and average DUE to CUE interference is at its greatest. This is understandable, and readily

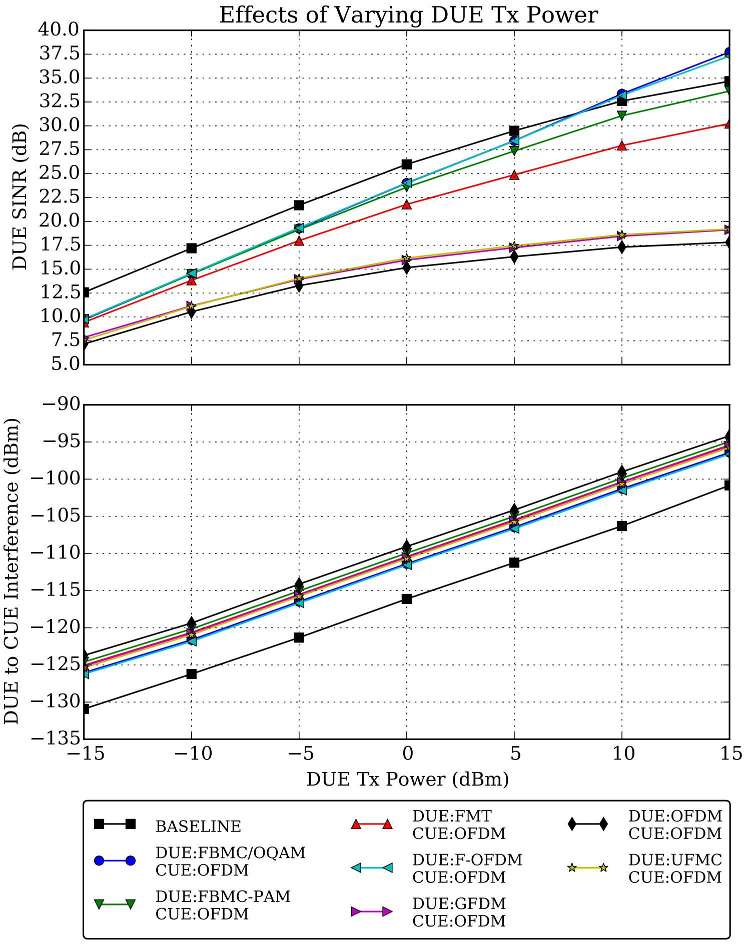


Figure 4.13: Effects of varying DUE transmit power on DUE SINR and DUE to CUE interference. Increasing DUE transmit power results in an increase in DUE SINR at the cost of increased interference to CUEs.

explained as follows. According to the strict FFR scheme employed, DUEs reuse the resources of CUEs in neighbouring reuse regions. At small cell sizes, the average distance between devices in neighbouring reuse regions is reduced. This results in greater CUE to DUE interference and reduces DUE SINR. As the cell radius increases, so too does the distance between reuse regions, and DUE SINR increases. This increase is mainly observed at smaller cell sizes; at large cell sizes, CUE to DUE interference is almost negligible and further increases to cell radius result in little or no increase in DUE SINR.

DUE to CUE interference, on the other hand, occurs at BSs. In small cells, the average distance between clusters and the BSs serving neighbouring reuse regions is shorter, resulting in higher DUE to CUE interference. This is evidenced in the lower sub-plot in Fig. 4.14, in which we observe that the average DUE to CUE interference decreases as the cell radius increases. Therefore, as the cell size increases, average DUE to CUE interference decreases and average DUE SINR increases. Essentially, the greater the cell size the more protection strict FFR offers against the various types of interference, as the reuse regions are further

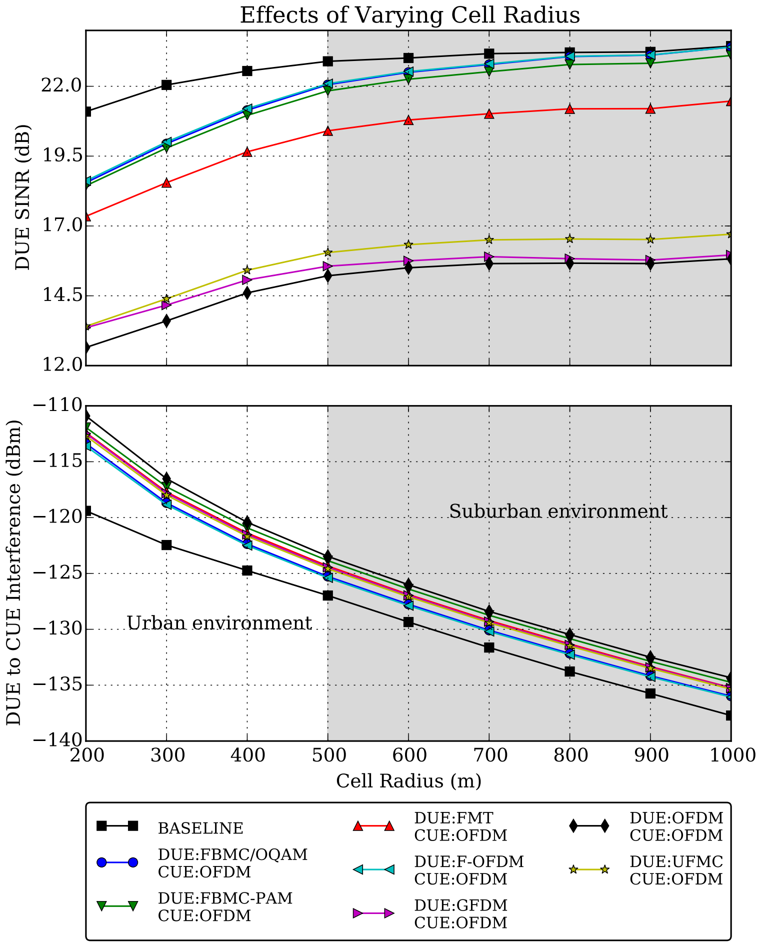


Figure 4.14: Effects of varying cell radius on DUE SINR and DUE to CUE interference. As the cell radius increases, DUE SINR increases and reduction in CUE SINR decreases.

apart.

Over the range of cell radii considered, synchronous OFDM provides the best performance and asynchronous OFDM provides the worst. However, as the cell radius increases, the interference from CUEs in neighbouring reuse regions to DUEs is reduced and the performance of several alternative waveforms approaches that of synchronous OFDM. At large cell sizes, even further gains are achievable as DUEs could transmit at a higher power without affecting CUEs.

4.5.5 Cluster Radius

We investigate the impact that cluster radius has on performance. We present the results in Fig. 4.15, varying the cluster radius from 30m to 100m in 10m increments. Reducing the cluster radius necessitates a corresponding change in the distance between a DUE transmitter and receiver, which we modelled using a uniform random variable. Accordingly, we choose the parameters a and b, representing the minimum and maximum Tx-Rx distances, respectively, of the uniform random variable $U_{[a,b]}$ as follows: $a = 5$m; $b = $ (cluster radius) $- 10$m.

Increasing the cluster radius has two opposing influences on DUE SINR. On the one hand, it results in reduced inter-DUE interference, which should boost the SINR. On the other hand, it also results in reduced received signal power, which should cause the SINR to decrease. In Fig. 4.15, we see that the reduction in received power is more influential and DUE SINR decreases as cluster radius increases. We concede, however, that this is somewhat dependant on how the distance between DUE transmitters and receivers is modelled (such as the parameters a and b), as this affects by how much the received power will decrease. We also note that for small cluster sizes, and in cases where DUEs do not use a waveform in the set {FBMC/OQAM, FMT, FBMC-PAM, f-OFDM}, SINR decreases slowly at first as the reduction in inter-DUE interference is almost significant enough to counter-act the effect of lower received signal powers.

Reducing the cluster radius increases the density of DUEs in the cluster, resulting in greater inter-DUE interference. Hence, employing an appropriate alternative waveform for DUEs yields the greatest benefit in dense clusters in which inter-DUE leakage interference is most significant. The synchronous baseline case again performs the best; however, the performance for cases where DUEs use a waveform in the set {FBMC/OQAM, FBMC-PAM, f-OFDM} approach that of the baseline for small cluster sizes. This can be attributed to reduced leakage interference from CUEs in the same reuse region, as smaller clusters are less likely to encompass CUEs in the same cell.

4.5.6 Amount of Time and Frequency Misalignment between Devices

The final parameters whose influence on performance we investigate are the maximum permitted TO and CFO. Both CFO and TO affect DUE performance similarly, and so it makes sense to isolate them when studying their effects on performance. Hence, when examining the effect of TO on DUE performance, we consider a case involving no CFO. Conversely, when investigating the effects of CFO, we consider devices to be perfectly aligned in time.

Maximum Possible TO

We vary the maximum permissible TO as a fraction of the time spacing between two OFDM symbols from 0 (full synchronism) to 1 (full asynchronism) in 0.1 increments. Limiting the maximum permissible TO corresponds to a case in which coarse alignment has been obtained; for example, 0.2 would correspond to the case in which devices are synchronised to within 20% of an OFDM symbol time.

Fig. 4.16 illustrates the results. The black line representing the case whereby both DUEs and CUEs use OFDM will be our baseline for comparison, and it can be seen that SINR

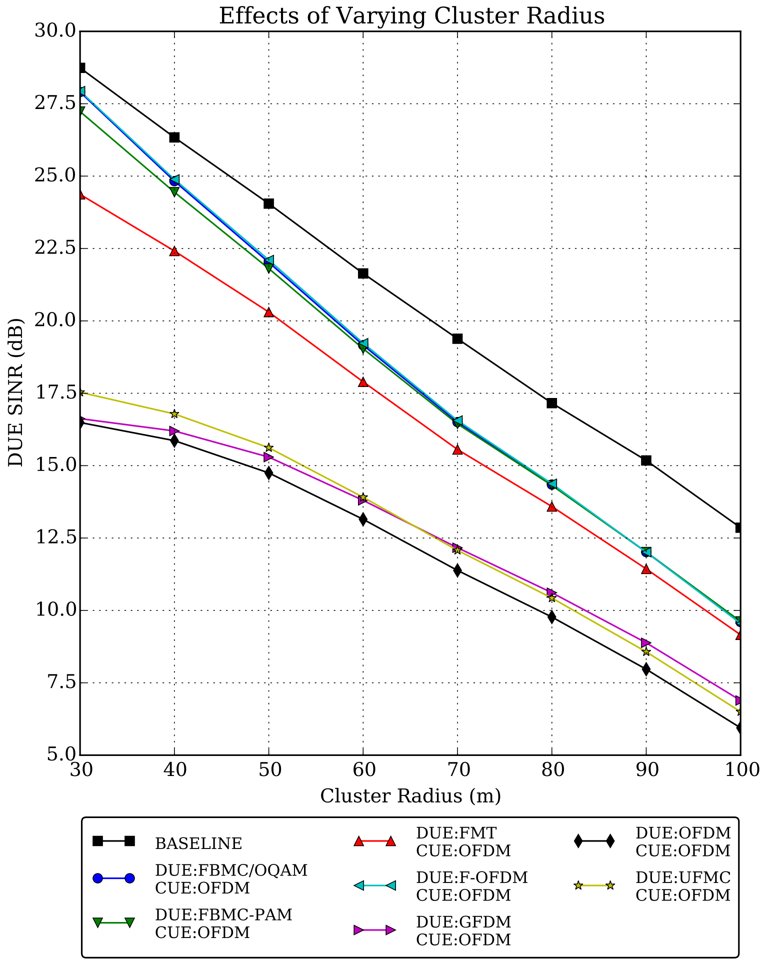


Figure 4.15: Effects of varying cluster radius on DUE SINR and DUE to CUE interference. Employing an appropriate alternative waveform for DUEs yields the greatest benefit in small clusters in which inter-DUE leakage interference is most significant.

drops rapidly when the TO is greater than the CP, as the TOs are no longer fully absorbed by the CP. The CP duration T_{cp} for OFDM is 12.5% of the symbol duration T. We can divide the rest of the graph into two scenarios:

1. *Scenario in which DUEs use a waveform in the set {GFDM, UFMC}, and CUEs use OFDM*: These curves become quite similar as the maximum permissible TO increases, and are out-performed by our baseline OFDM-OFDM case. This seems surprising at first glance, but can be explained. Indeed, we saw in Fig. 4.8 that, with the chosen parameters, UFMC and GFDM still cause a significant amount of interference between coexisting users in homogeneous links in which both users are deploying one of these waveforms; thus, inter-DUE interference is quite important if DUEs use either GFDM or UFMC. Moreover, OFDM based users are orthogonal to one another as long as δ_{t} is contained in the CP duration. However, GFDM or UFMC users never achieve orthogonality with OFDM users, which explains that if

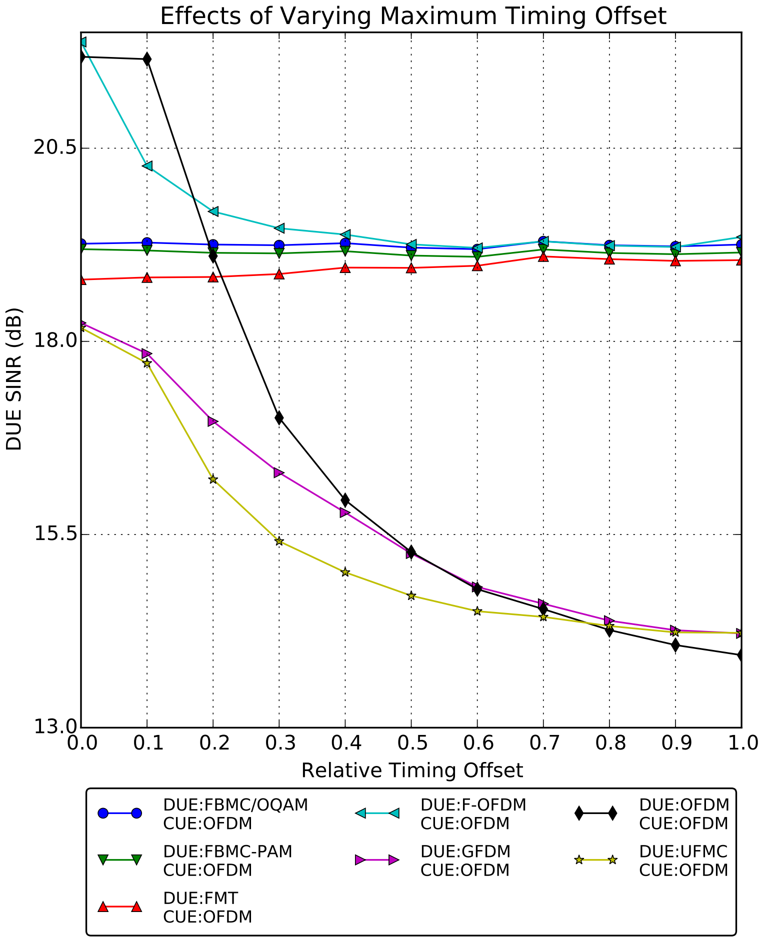


Figure 4.16: DUE SINR performance as the maximum permitted TO is varied.

CUEs use OFDM, CUE to DUE interference is on average more significant if DUEs use UFMC or GFDM than if they also employ OFDM.

2. *Coexistence scenario in which DUEs use a waveform in the set {FBMC/OQAM, FBMC-PAM, f-OFDM, FMT} and CUEs use OFDM*: As the TO increases, the curves exhibit similar performance. At a maximum TO, the benefit to using one of these alternative waveforms for DUEs is considerable, while for very low TOs ($< 20\%$), they are outperformed by the baseline OFDM-OFDM case, since the CP in OFDM absorb much of the TO. With the exception of f-OFDM, the performance of these waveforms varies little according to the TO, as these waveforms all exhibit excellent spectral localisation. In addition, FBMC/OQAM and FBMC-PAM both use a guard subcarrier while FMT is similarly protected by its inbuilt guards. F-OFDM has an interesting behaviour, as it is the only waveform that is affected differently by OFDM according to the value of $\delta_{\mathrm{t}}^{\max}$. This is due to the fact that for small TOs, f-OFDM and OFDM achieve quasi-orthogonality, which is then lost as δ_{t} increases.

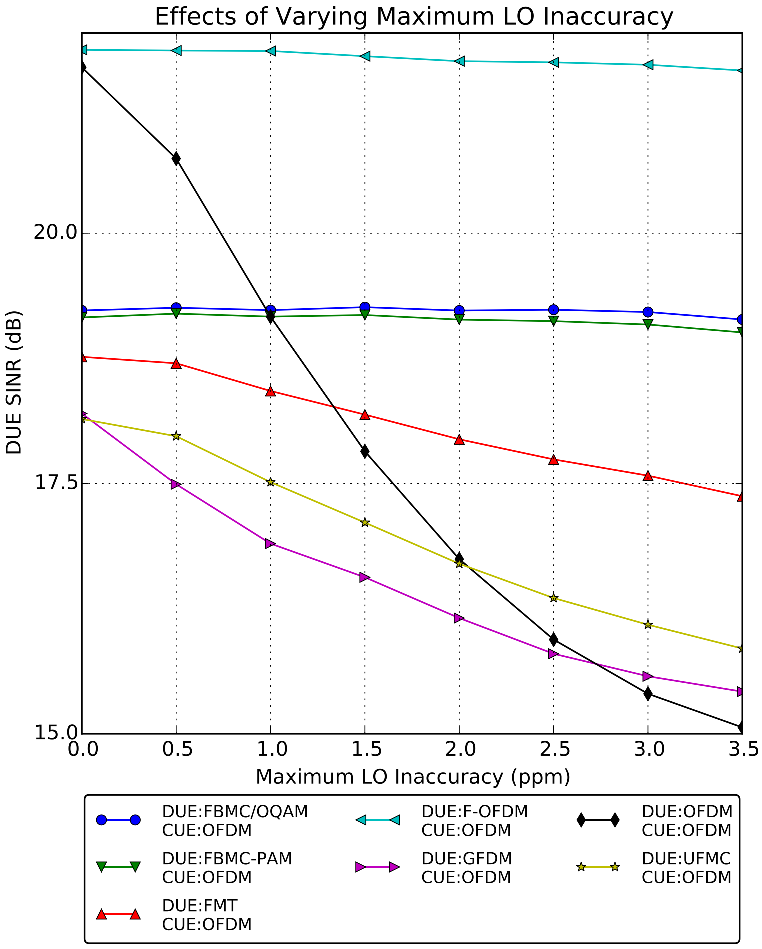


Figure 4.17: DUE SINR performance as the maximum CFO is varied.

Maximum Possible CFO

Having investigated the effect of TO, we now examine the relative performance of the waveforms under various levels of CFO. The LO inaccuracy is varied from 0ppm to 3.5ppm in increments of 0.5, corresponding to frequency offsets of +/- 0kHz to +/- 7kHz in 1kHz increments at a carrier frequency of 2GHz.

In Fig. 4.17, for the case in which OFDM is used by both sets of users, we observe that the average DUE SINR reduces as the frequency offsets become greater. This can be attributed to OFDM's large sidelobes, resulting in significant interference leakage to and from other users. In a similar fashion to the study on the effects of TO, we again take the case in which both sets of users employ OFDM to be our baseline case, and divide the rest of Fig. 4.17 into two scenarios:

1. *Scenario in which DUEs use a waveform in the set {GFDM, UFMC}, and CUEs use OFDM*: When DUEs employ GFDM or UFMC, DUE SINR decreases as the maximum possible LO inaccuracy is increased; however, the decrease occurs at a

lower rate than for OFDM since OFDM possesses the largest sidelobes. For low LO inaccuracies, the baseline OFDM case outperforms the scenarios in which CUEs use OFDM and DUEs use either UFMC or GFDM. This is because OFDM users achieve near orthogonality at low CFOs, while GFDM or UFMC users never achieve orthogonality with OFDM users. However, as the LO inaccuracy is increased, OFDM suffers from increasingly large interference leakage owing to its sidelobes and the waveform choices involving UFMC or GFDM begin to outperform the baseline OFDM case.

2. *Coexistence scenario in which DUEs use a waveform in the set {FBMC/OQAM, FBMC-PAM, f-OFDM, FMT} and CUEs use OFDM*: The waveform choices involving FBMC/OQAM, FBMC-PAM, and f-OFDM are largely unaffected by varying CFO, as evidenced by the horizontal lines in Fig. 4.17. At the LO inaccuracies considered, frequency offsets are contained within +/- half a subcarrier. Given that these schemes use a guard band of half a subcarrier at either side of an RB, and that leakage is confined within a similar range for these alternative waveforms, it is not surprising that very little variation in performance is observed as the CFO is increased. FMT, on the other hand, uses 12 subcarriers per RB. Hence, we observe that the SINR performance of DUEs using FMT reduces as the maximum possible LO inaccuracy is increased. The waveform choices involving FBMC-PAM and FBMC/OQAM only begin to outperform the baseline OFDM case after approximately 1ppm. For DUE users using FMT, improvements in SINR over the baseline case are only observed after a maximum LO inaccuracy of 1.3ppm (based on an interpolated value). Similar to before, this is because OFDM achieves quasi-orthogonality at low CFO, but suffers significant degradation as CFO increases. Out of the waveform couples considered in this scenario, f-OFDM exhibits the best performance and is never outperformed by the baseline OFDM case. Similar to OFDM, f-OFDM achieves quasi-orthogonality at low CFO. However, dissimilar to OFDM, it is protected by its filtering and guard subcarrier as CFO increases and hence does not suffer the performance decrease experienced by OFDM.

4.6 Conclusion

The results presented in this chapter were obtained through simulations, but are built upon theoretical analysis performed at the physical layer which characterises leakage interference between various waveform pairs. In fact, one of the main motivations for the chapter is demonstrating how the well-researched properties of waveforms translate to performance at a system-level in realistic future network scenarios. In this pursuit, simulation is an ideal tool, as it permits us to achieve a high level of realism in our investigations.

When only the SINR metric is considered, the best results are obtained when either synchronous OFDM is used, or both sets of users employ a waveform from the set {FMT, FBMC/OQAM, FBMC-PAM, f-OFDM}. When the achieved rate is instead considered, taking bandwidth efficiency into account, the case in which machine-type DUEs operate asynchronously and both sets of users employ FBMC/OQAM achieves the greatest performance.

As suggested in the previous section, the performance of a waveform in asynchronous communication depends on its sidelobes. Waveforms with very small sidelobes result in less inter-DUE leakage and hence perform the best. We note that the size of a waveform's sidelobes depends largely on the filtering applied, with many filter implementations existing. For example, the performance of FBMC/OQAM could be further improved by using an optimised filter such as the one suggested in [157]. However, as it was not possible to consider every possible filter, and for the sake of fair comparison, we chose filters and parameters that were representative of the most common implementations in the literature.

Promisingly, we also showed that good performance can be obtained when DUEs operate asynchronously and use a different waveform to CUEs, paving the way for the possibility of the coexistence of waveforms in future networks for different use cases, a paradigm shift from previous generations. In particular, when FBMC/OQAM is used by DUEs, the average achieved rate is marginally greater than the synchronous OFDM baseline case, and 43% greater than the asynchronous OFDM case. We also note that these figures are conservative, as they assume perfect synchronisation in the baseline case.

The results indicate that the biggest drawback to using asynchronous communication is the increased interference to cellular users. Unfortunately, employing a different waveform for DUEs does little to reduce this type of interference. We note, however, that interference can typically be kept low through the use of strict FFR and low DUE transmit powers.

To conclude, we have shown that it is feasible for cellular networks to serve clustered MTC use-cases, such as smart factories, using asynchronous direct communication. In particular, we highlighted the benefits to DUEs of using an alternative waveform to reduce leakage interference, and suggested that future networks may permit the coexistence of waveforms. Hence, by employing a waveform with improved spectral localisation compared to OFDM, such as FBMC/OQAM, DUEs can avail of the benefits of asynchronous communication without suffering a performance loss, even if regular CUEs continue to use OFDM.

We highlight, however, that the decision of whether to use synchronous or asynchronous communication is multifaceted and should not be made solely based on performance metrics such as SINR and achieved rate. The computational complexity of the candidate waveforms is also a significant factor in such a decision, and while its consideration is out of scope in this analysis, we refer the reader to Figure 6 in [158] which compares the complex multiplications required in the implementation of each waveform. We note that the complexity of waveforms that rely heavily on filtering, such as UFMC, can be an order of magnitude larger than OFDM, which is the least complex of the waveforms considered. There is a trade-off between low complexity waveforms that suffer greatly from leakage interference in asynchronous communications, such as OFDM, and waveforms that employ filtering to reduce leakage interference, but are more complex as a result.

Hence, the complexity of supporting multiple waveforms in a single user equipment may prove to be challenging for vendors. Instead, the concept of adopting multiple waveforms in future networks is best suited to equipment with specialist communication requirements which need only support a single waveform; for example, machinery in a smart factory.

5 Customisable RAN Slicing: Time and Frequency Resource Allocation

Customisable RAN Slicing: Time and Frequency Resource Allocation

In Section 3.3.1, we demonstrated that slicing the core network alone cannot provide the level of adaptability necessary to satisfy the requirements of future networks. The main categories of services targeted by future networks require disparate Radio Access Technologies (RATs), motivating the need for Radio Access Network (RAN) slicing.

The various RATs and options for customising the future RAN are outlined in Section 2.3. While Chapter 4 investigates the use of multiple coexisting waveforms in the RAN, this chapter focuses on a RAN centred on a single adjustable waveform. This option was discussed in Section 2.3.3, which suggested that the diversity of requirements demanded of the RAN in future networks could be served through physical changes to the manner in which signals are formed and packed together for transmission, such as variable frame structures and subcarrier spacings. We note that this principle is being adopted for 5G New Radio (NR) [124].

For example, a RAN slice serving automotive services in a high mobility scenario may use a wider subcarrier spacing to combat high Doppler shifts, while a RAN slice serving a latency-sensitive service such as real-time gaming may use fewer symbols in each subframe. These lower layer customisations are necessary, but introduce challenges in ensuring isolation between RAN slices.

In particular, the use of different frame structures in a system can result in inter-service-band interference [159]. To ensure that transmissions using different numerologies do not interfere with one-another, it is necessary to adopt a combination of guard bands, enhanced spectral filtering or other advanced signal processing techniques. Hence, although RAN slicing increases the ability of the RAN to serve dissimilar services and industries, this flexibility comes at a cost.

Mapping the resource allocations of multiple slices to a time-frequency resource grid consisting of several numerologies while ensuring isolation between the slices is challenging, and is referred to as the tiling problem [115]. Hence, it is necessary to design a time-frequency resource grid structure which can provide enough granularity in resource allocations to provide flexibility, without compromising the key principle of isolation in slicing.

Research Question, Key Contributions and Chapter Organisation

The research question, outlined in Section 1.1, that this chapter addresses is the following:

What are the implications of a system comprising tailored virtual networks on radio resource allocation and user admission?

We focus on the part of the question relating to radio resource allocation, and deal with user admission in the next chapter.

Specifically, we direct our attention to the trade-off identified in Chapter 3.3.3 between adaptability on the one hand, and the cost/overhead of this adaptability on the other hand. The flexibility arises from being able to serve the needs of multiple vertical industries by tailoring the manner in which the signal is transmitted. The cost/overhead arises from the challenges associated with ensuring the coexistence of multiple RAN slices which have been designed to satisfy diverse use cases.

The main contributions of this chapter are:

- We propose four different configurations of the time-frequency resource grid and analyse, both qualitatively and quantitatively, how each performs in terms of both flexibility and the overhead associated with achieving that flexibility.
- We propose the concept of a RAN profile based on this analysis, which distinguishes between service-types and individual services, and permits the allocation of time-frequency resources to be performed based on this distinction.

The rest of the chapter is structured as follows. Section 5.1 discusses how changes to the numerology can be used to target different use cases. Section 5.2 proposes four different resource grid structures and compares them in terms of the overhead required to ensure isolation, the ability to adapt to traffic changes, and the implications for designing the control plane. Finally, Section 5.3 proposes the concept of a RAN profile to manage the trade-off between flexibility and coexistence overhead.

5.1 Tailoring RAN Slices through Mixed Numerologies

The numerology of a multi-carrier system is the set of design parameters which influence how the system performs, such as the subcarrier spacing, the cyclic prefix length, the symbol length, and the number of symbols in a sub-frame. In previous multi-carrier systems such as Long Term Evolution (LTE) and WiFi, predefined numerologies were chosen that were suited to the primary service and operating environment being targeted. In contrast, future networks may permit each RAN slice to use a separate numerology that is tailored for a particular vertical.

Subcarrier Spacing and Symbol Length

These two design options are related, and cannot be changed independently. In extended coverage scenarios or high multi-path environments, large delay spreads result in inter-symbol interference, and small coherence bandwidths result in frequency selective fading. The solution is to use longer symbols to absorb symbol tails and narrow subcarriers to approximate flat fading channels. In high mobility scenarios, large Doppler spreads result in inter-carrier interference and short coherence times result in poor channel estimation. The solution is to increase subcarrier spacing to separate carrier frequencies, resulting in shorter symbols. The authors in [160] give a good overview of the design options relating to subcarrier spacing (and symbol length). 5G NR will support scalable subcarrier spacings which are an integer multiple of the LTE spacing of 15 kHz, up to a maximum spacing of 240 kHz [161].

Number of Symbols per Transmission Time Interval (TTI)

5G NR provides support for shorter TTI times than the fixed length of 1ms in LTE.

Cyclic Prefix

5G NR supports two cyclic prefix lengths: one for regular use, and one to enable extended coverage (at a subcarrier spacing of 60kHz).

Pilot Placement

Reference symbols known as pilots are used to estimate the channel for the purpose of one-tap equalisation in Orthogonal Frequency Division Multiplexing (OFDM) and to allow the User Equipment (UE) to provide channel state information to the network to facilitate frequency selective scheduling. If the channel coherence time reduces, such as in a high mobility scenario, the pilot placements are no longer sufficiently close for interpolation to provide accurate channel estimates for equalisation of OFDM symbols. Similarly, the feedback provided to the network could be outdated when the next round of scheduling is performed. In these cases, pilot symbols should be placed closer together.

Scheduling Request Period

In the uplink, the scheduling request period can be decreased to reduce the latency. This will increase the control plane overhead and potentially reduce throughput.

Table 5.1 displays a number of channel attributes, their effect on performance, how to mitigate any adverse effects through numerology changes, and the scenario it applies to.

Table 5.1: Numerology considerations for scenarios.

Channel Measure	**Effect**	**Solution**	**Relevant Scenario**
delay spread	large delay spreads result in inter-symbol interference	use longer symbols and/or cyclic prefixes to absorb symbol tails	extended coverage; high multi-path environments
coherence bandwidth	small coherence bandwidths result in frequency selective fading	use narrow subcarriers to approximate flat fading channels	extended coverage; high multi-path environments
Doppler spread	large Doppler spreads result in inter-carrier interference	increase subcarrier spacing to separate carrier frequencies	high mobility
coherence time	short coherence times result in outdated feedback and poor channel estimation	use shorter symbols or redesign pilot placements to increase rate of Channel State Information (CSI) feedback	high mobility

5.2 Time-Frequency Resource Hierarchy of the Future Network RAN

Fig. 5.1 illustrates four potential options for sharing time-frequency resources between RAN slices with different numerologies, which are briefly described below.

1. **Fixed contiguous sub-band approach:** Each RAN slice uses a fixed, predefined contiguous sub-band.
2. **Variable contiguous sub-band approach:** Each RAN slice uses a contiguous sub-band consisting of a fixed region and a variable region. The variable region may be shared between neighbouring RAN slices according to demand.
3. **Sub-band tiling approach:** The resource grid is divided into regular sub-bands of fixed size.
4. **Frame tiling approach:** In this approach, tiling occurs at frame granularity in the time domain, and resource block granularity in the frequency domain.

5.2.1 Qualitative and Quantitative Comparison

We employ Monte Carlo simulation to compare the four possible resource grid designs. In the time domain, each block in Fig. 5.1 represents a frame. As per 5G NR, frames are 10ms in length irrespective of the numerology in use. In the frequency domain, each block

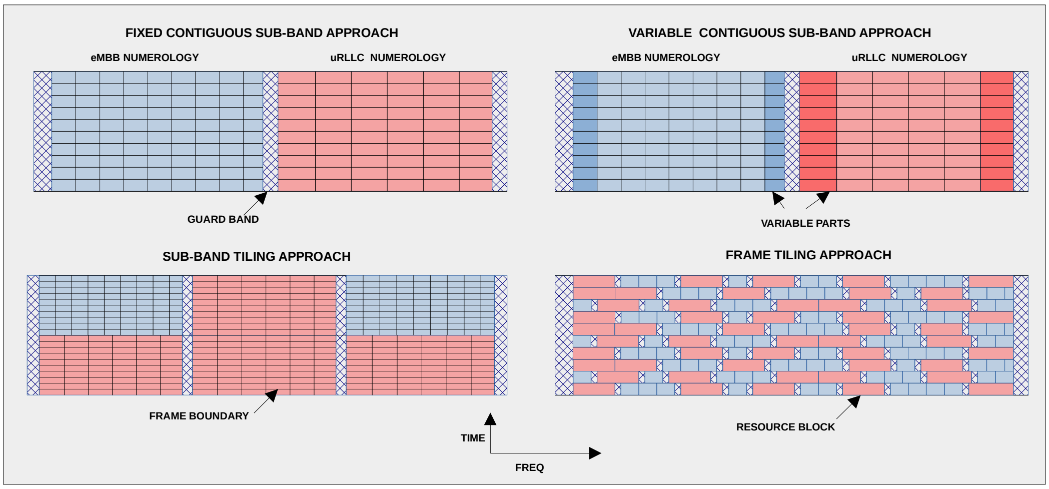


Figure 5.1: Four options for sharing time-frequency resources between two services employing different numerologies.

represents a resource block consisting of 12 subcarriers, which varies in width depending on the numerology.

Simulation Set-Up

We consider three RAN slices, each using a different subcarrier spacing (15kHz, 30kHz, 60kHz), and assume the network is fully loaded with all time-frequency resources in use. In each 1 second interval, sub-bands and resource blocks are randomly assigned according to a uniform distribution to each RAN slice in the sub-band and frame tiling approaches, respectively, until each of the three RAN slices has been granted an equal portion of the time-frequency grid. For the sub-band tiling approach, each sub-band is 10 frames in length and 120kHz wide, while for the variable contiguous sub-band approach, the ratio between the fixed and variable part of each RAN slice is 3:1. For our simulations, we consider a 1 second by 20MHz window of the resource grid. Finally, when considering the adaptability of each configuration, we adapt our simulation model so that, on average, one RAN slice is 50% under-loaded, one RAN slice is fully-loaded, and one RAN slice is 50% over-loaded.

1. Coexistence Overhead

In a multi-service system in which RAN slices utilise different technologies, proactive measures are required to ensure that RAN slices coexist without adversely affecting one another. Of primary concern is that permitting mixed subcarrier spacings and frame structures can result in inter-numerology interference [159], as the side-lobes of the symbols in the frequency domain will no longer overlap in an orthogonal manner. The use of waveforms with enhanced filtering, such as Filtered Orthogonal Frequency Division Multiplexing (f-OFDM), has been considered to overcome this issue [143]. However, even with this enhanced filtering, guard bands may be needed to prevent inter-slice interference. Low complexity inter-service band

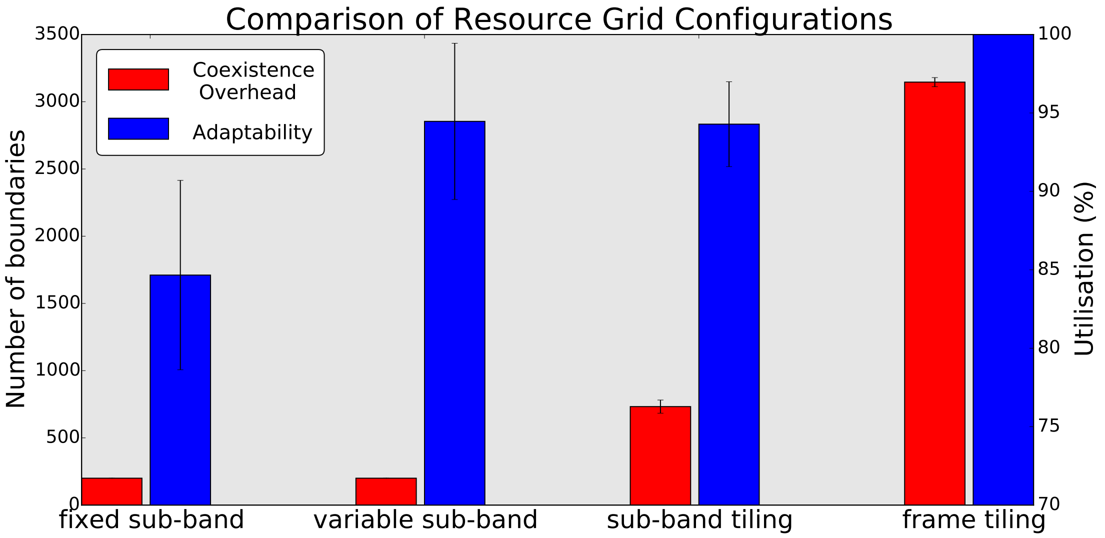


Figure 5.2: Comparison of the overhead and level of adaptability associated with each approach.

interference cancellation algorithms which precode information symbols at the transmitter provide another option [162].

We use the general term *coexistence overhead* to capture the effects of leakage interference between RAN slices, resource loss due to inter-slice guard bands and guard times, and the complexity of advanced interference cancellation techniques. The coexistence overhead is dependent on the number (and length) of boundaries between numerologies over a certain time interval. We consider one boundary to be a single border of length 10ms between two neighbouring radio frames employing different numerologies, and use the previously described simulation set-up to determine the average number of boundaries in a 1 second by 20MHz window of the resource grid.

As can be seen in Fig. 5.2, both the fixed and variable contiguous sub-band approaches minimise the number of boundaries between different RAN slices; hence the coexistence overhead is low and constant. In the sub-band tiling approach, there are more boundaries between different numerologies, which increases the coexistence overhead to approximately 3.5 times that of the contiguous approaches. This is still relatively low compared to the frame tiling approach, which results in a large number of boundaries between numerologies due to the fine granularity at which time-frequency resources can be assigned.

2. Ability to Dynamically Adapt to Traffic

RAN slices must be able to adapt to different traffic loads, requiring the active sharing of time-frequency resources between RAN slices. Future networks will target many different use cases such as smart health-care, automated cars, Industry 4.0, and smart city management, and these will likely present different traffic patterns than classic mobile broadband type, smart-phone driven traffic. As a result, the relative split in network load attributed to each RAN slice can reasonably be expected to vary spatially and temporally, and further investigation on the behaviour of different traffic types based on empirical evidence is required.

We estimate the adaptability of each approach by considering how time-frequency resources can be reassigned among RAN slices when one RAN slice is highly loaded and another is lightly loaded. For each one second time window, we evaluate the ratio between the total number of requests served and the total number of serviceable requests, which provides us with a measure of utilisation of the time-frequency resources. If each request takes a fixed amount of the time-frequency grid to satisfy, the total number of serviceable requests is equal to the lesser value between the total number of requests submitted to all RAN slices and the total number of requests that can be accommodated by the resource grid.

The average utilisation for a one second window is presented in Fig. 5.2. The fixed contiguous sub-band approach proves to be the least flexible, with only 84% average utilisation, due to its inability to reassign unused resources from under-loaded RAN slices to over-loaded RAN slices. The variable contiguous sub-band approach offers more flexibility, providing an improved average utilisation of 94%, due to the fact that RAN slices can be rearranged in the resource grid and can borrow resources from their direct neighbours. Similar performance is provided by the sub-band tiling approach, which divides RAN slices into several smaller sub-bands which are distributed throughout the system bandwidth. Finally, the frame tiling approach is able to provide 100% utilisation of resource blocks, ensuring that no RAN slice has unused resources that could be reassigned to another RAN slice.

3. Control Plane Considerations

The regular resource grid consisting of a single numerology in LTE allows time-frequency resources to be conveniently indexed for the purpose of resource allocation, with the Physical Downlink Control Channel (PDCCH) comprising the first few symbols of each sub-frame and spread across the entire channel bandwidth. The irregularity of the resource grid in a multi-service, mixed numerology system requires more complex indexing schemes for referring to portions of the grid. In response to this, a new concept called a Bandwidth Part (BWP) has been introduced in 5G NR. A BWP is a contiguous set of physical resource blocks with an associated numerology. UEs can be configured with up to four BWPs, with one active at any given time in both the uplink and downlink. The PDCCH for NR is also no longer spread across the entire bandwidth, but is instead localised in a Control-Resource Set (CORESET) within each BWP. Hence, 5G NR allows UEs to be configured to operate in a contiguous bandwidth region using a specified numerology, with the control plane for that region housed internally.

We note that the fixed and variable contiguous sub-band approaches could deploy an LTE-like control plane across the fixed bandwidth allocation of each sub-band. NR could

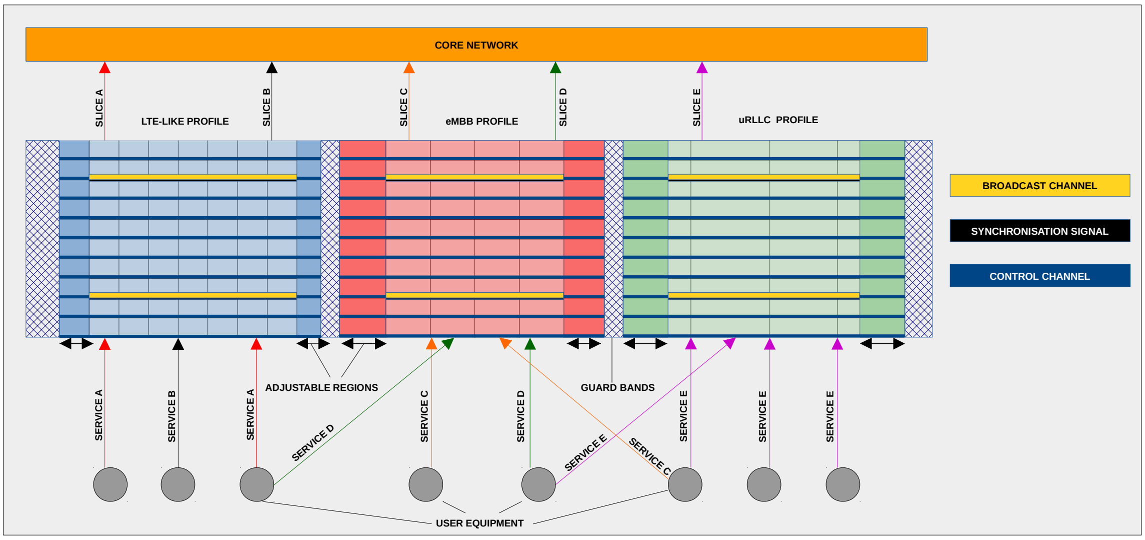


Figure 5.3: Illustration of the RAN profile concept. RAN profiles are based on the variable contiguous sub-band approach, with each profile housing multiple slices.

accommodate both of these options and, additionally, the sub-band tiling approach by configuring sub-bands as BWPs with a CORESET defined within each. An NR-style control plane is not an option for the frame tiling approach, as a minimum frequency width for the CORESET prevents the multiplexing of numerologies as BWPs at a frame/resource block granularity. If multiplexing at this granularity is to be considered in future, alternative control-plane designs must be devised.

5.2.2 Trade-Off: Adaptability vs Coexistence Overhead

From the comparison provided in Fig. 5.2, we can identify a inherent trade-off between adaptability on one hand, and the coexistence overhead on the other hand. At one extreme, the fixed contiguous sub-band approach incurs very little overhead, but is very limited in its adaptability. At the other extreme, the frame tiling approach provides optimal flexibility, but at the cost of a substantial coexistence overhead. The solution appears to be a compromise to both extreme approaches, with both the variable sub-band approach and sub-band tiling approach providing a similar high level of adaptability with a reasonably low overhead.

In short, enabling a highly adaptable system that can adjust to temporally and spatially changing traffic demands for different RAN slices results in an increased number of boundaries between RAN slices, and hence a higher coexistence overhead.

5.3 RAN Profiles

Although RAN slices may be tailored for individual verticals or services, in reality many types of services across different verticals will require similar network behaviour. Hence, we present the idea of RAN profiles: a specific configuration of the air interface that has been selected to meet the demands of a broad category of services. In particular, the RAN profile consists of the more general configuration choices outlined in Section II, relating to lower layer concerns such as subcarrier spacing, TTI length, pilot placement, and scheduling request period. Each of these configuration choices is relevant to an array of similar services, while individual RAN slices within each profile can further configure their behaviour using some of the more specific configuration options outlined in Section 5.1, such as the choice of scheduler, the use of Multiple-Input Multiple-Output (MIMO), and diversity schemes.

We base RAN profiles on the variable contiguous sub-band approach (Fig. 5.1, case 2), due to its relatively low overhead and high adaptability (Fig. 5.2). In the simplest case, the three main 3$^{\text{rd}}$ Generation Partnership Project (3GPP) target areas may be supported in dedicated profiles: enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communication (uRLLC), and Massive Machine-Type Communication (mMTC). Basically, eMBB, uRLLC and mMTC would each have their own version of the air interface, using a frame structure suited to their needs. These air interfaces would each have their own resource grid that could then be further sliced using one of the aforementioned tiling approaches (i.e. sub-band or frame based) since the time-frequency resources are homogeneous and hence do not suffer from inter-numerology interference. This allows for further customisation, which can then be applied to the RAN slices through control over a RAN slice-specific scheduler, modulation order, or other parameters.

This idea of supporting multiple configurations of the RAN in future networks was envisioned in the METIS II project, which outlined the need for Air Interface Variants (AIVs) that are optimised for one or more target scenarios/services [163]. A user-plane design framework for a service-tailored RAN is outlined in [164], incorporating the idea of AIVs.

5.3.1 Coexistence Overhead

Coexistence must be considered at two levels: coexistence of profiles and coexistence of RAN slices. The coexistence of profiles incurs the same overhead as the variable contiguous sub-band approach, which minimises the number of boundaries between different numerologies, as evident in Fig. 5.2. Although the coexistence of RAN slices is based on a tiling approach, the resources within a particular profile are homogeneous in nature, resulting in a reasonably low coexistence overhead.

5.3.2 Control Plane

Each profile could possess a dedicated control plane which can be implemented similarly to LTE, with the first few symbols of each sub-frame carrying control information. In LTE, the central subcarriers house the primary and secondary synchronisation signals and the physical broadcast channel, which is used to broadcast information regarding the configuration of the system. A similar approach could be adopted for a multi-service RAN, and is displayed in Fig. 5.3.

An NR-style control plane is also an option, with each sub-band implemented as a BWP containing the CORESET for that frequency region. The size of the BWP and location of the CORESET within the BWP can be adjusted in response to the variable sizes of the sub-bands.

5.3.3 Ability to Adapt to Traffic Variations

As depicted in Fig. 5.4, the profile-based RAN will be a multi-service, multi-carrier system with a 5-tier frequency resource structure consisting of system bandwidth, profile bandwidth, RAN slice bandwidth, resource block, and subcarrier, listed in descending order of granularity. The system bandwidth is divided into contiguous regions, with each region constituting a different RAN profile. RAN slices are then created dynamically and assigned a number of time-frequency resources from a RAN profile sub-band. The profile bandwidth can be expanded and contracted within defined limits, while RAN slices can be dynamically allocated in a tiling approach within the profile bandwidth.

Hence, the profile approach offers separate control over the resources granted to a particular service area such as eMBB, and the resources allocated to a particular RAN slice in that service area, such as a video streaming service. In this regard, the profile approach proves to be quite flexible. To reduce signalling overhead and complexity, we suggest that profile bandwidths should be updated less frequently than the minimum scheduling unit for users; for example, every 100 TTIs. Once a profile bandwidth has been assigned to a profile, the profile resources can be allocated to RAN slices up until the next profile bandwidth update.

5.4 Conclusion

Fundamentally, there is a trade-off between adaptability on one hand, and low coexistence overhead on the other. Both flexibility and isolation are core traits of network slicing. In this chapter, we focused on the structure of the time-frequency resource-grid, so that the allocation of time-frequency resources can be achieved in a flexible manner, while maintaining isolation between slices. The concept of profiles can be used to balance this trade-off by

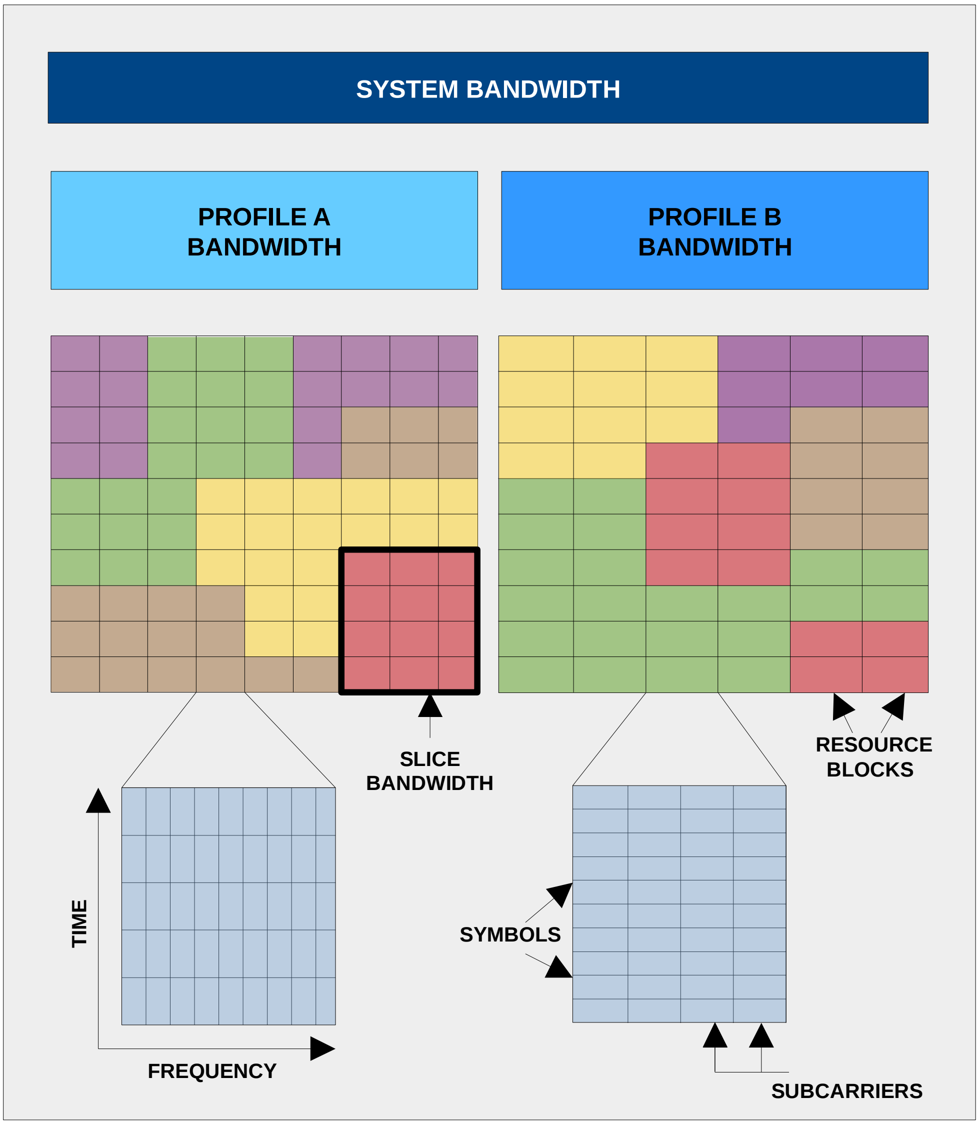


Figure 5.4: A possible time-frequency resource structure for future network RANs. The future network RAN may be a multi-service, multi-carrier system with a 5-tier hierarchy consisting of system bandwidth, profile bandwidth, RAN slice bandwidth, resource block, and subcarrier.

creating a distinction between service-types and individual services, and hence providing an ability to schedule resources at both different time-scales and granularities.

We highlight that the concept of a profile complements the 5G NR standardisation, in particular relation with the concept of a BWP. As discussed in Section 5.3.2, 3GPP are introducing a new concept called a bandwidth part to facilitate the coexistence of multiple numerologies in 5G NR. Defined in [124], a BWP is ‘a subset of contiguous common resource blocks defined...for a given numerology μ_i in the bandwidth part i on a given carrier’. Hence, each BWP represents a contiguous portion of the resource grid, employing a particular numerology. Each end-user can be configured with up to four BWPs, with one in use at any given time. Therefore, a BWP provides the underlying infrastructure from which to implement RAN profiles. For a more detailed overview of BWPs in NR, we refer the reader to [165].

6 Customisable RAN Slicing: User Admission to Slices

Customisable RAN Slicing: User Admission to Slices

We consider the two-tier network slicing model outlined in Section 3.4, in which multiple Mobile Virtual Network Operators (MVNOs) operate service-tailored slices as independent business ventures. For the remainder of this chapter, we will refer to these MVNOs as Specialised Mobile Network Operators (SMNOs).

A user may benefit by subscribing to a bundle comprising services from more than one of these SMNOs. This offers two advantages; first, users do not have to manage multiple separate subscriptions to SMNOs that they may require, and secondly, users can pay a fixed price for a specified data allowance that can be used as needed across a selection of different SMNOs.

Hence, we consider a business model centred on an entity called a subscription broker, which acts as a broker between service-tailored SMNOs and users, selling access to bundles of SMNOs for a fixed price. Each bundle is sold with a fixed data allowance. A similar idea of dynamically switching between multiple Mobile Network Operators (MNOs) to improve performance is used in Google's Project Fi. In our case, however, each MNO has tailored its network to meet the demands of a particular service type, which requires a completely different approach for matching users to MNOs. We adopt the Gale-Shapley *college admission* algorithm [166] to match users to SMNOs and ensure that no matched pair could have achieved a better matching based on their preferences.

Research Question, Key Contributions and Chapter Organisation

The research question, outlined in Section 1.1, that this chapter addresses is the following:

What are the implications of a system comprising tailored virtual networks on radio resource allocation and user admission?

We focus on the part of the question relating to user admission to specialised slices.

Matching theory has been applied to solve resource allocation problems in many areas of wireless communications. In one of the first works to apply matching theory to wireless networks [167], the authors demonstrate that popular schedulers such as maximum throughput or proportional fair do not necessarily result in a stable matching. A comprehensive tutorial on the use of matching theory in wireless networks is provided in [168]. The authors in [169] adopt matching theory to develop a novel user-cell association approach for small cell networks using context information obtained from user devices. User association is also examined in [170], which applies matching theory to pair Base Stations (BSs) and users in a virtualised cellular network. To the best of our knowledge, we are the first to employ matching theory to perform the matching between users and service-tailored slices.

Our focus in this chapter is on how users associate to slices in a marketplace of specialised MVNOs, when a user may require the service of multiple slices. The main contributions of this chapter are:

- We adopt the Gale-Shapley matching algorithm to perform the matching between users and SMNOs in a bundle.
- We propose a method for devising the preference lists of users based on the concept of utility, and the preference lists of SMNOs based on revenue. The revenue-based approach for SMNOs can differentiate between different classes of users based on the price they pay for their subscription.
- We provide a case study to show the performance benefits of allowing users to choose SMNOs based on their requirements in a broker-based approach compared to a dedicated network approach.

The remainder of this chapter is structured as follows. Section 6.1 outlines the system model for the broker-based approach. The methods for constructing the preference lists of both users and slices is presented in Section 6.2. Section 6.3 describes the procedure for performing the matching itself. The cost of achieving stability in terms of utility is discussed in Section 6.4. The parameters of the case study are defined in Section 6.5 and an evaluation of the broker-based model is presented in Section 6.6.

6.1 System Model

We formally define the terms user, SMNO, Service Level Agreement (SLA) and subscription broker:

Definition 6.1.1. A *user* is any entity involved in a process of data transfer that may avail of the services of an SMNO. This may be a smart phone, a connected vehicle, or a sensor.

Definition 6.1.2. An *SMNO* manages a service-tailored network (which may be a virtual network) that is designed to meet the demands of a particular vertical, service, or class of users.

Definition 6.1.3. An *SLA* is a commitment between an SMNO and a user that the SMNO will provide the user with an agreed minimum Quality of Service (QoS), as specified by a set of guarantees advertised by that SMNO.

Definition 6.1.4. A *subscription broker* sells SLAs on behalf of SMNOs to users. The broker groups multiple SLAs into a package, called a bundle, which it sells for a fixed price with a fixed data allowance. This data allowance may be consumed by the user across any of the SMNOs in the bundle as needed. After taking its fee, the remaining revenue from the package is distributed among the included SMNOs according to the percentage of the user's data allowance served by each SMNO.

We note that the concept of a fixed data allowance and fixed price appears, on the surface at least, to contrast the current practice of price-dependent QoS. In this initial investigation of user association in a multi-service network, QoS is not a primary concern. We suggest, however, that the proposed pricing model can be modified in two ways to accommodate price-dependent QoS. In the first approach, we acknowledge that the cost per gigabyte to a service provider of providing a higher QoS increases. Hence, when the subscription broker divides the fixed bundle price among the constituent SMNOs, a higher proportion of the revenue may be allocated to the SMNOs which provide a higher QoS. Secondly, bundles may be structured so that each SMNO in the bundle provides a similar QoS. Premium bundles, which may include the same set of SMNOs but with each SMNO offering a higher QoS, may be offered a higher price-point to the user.

We represent the set of SMNOs as $S = \{s_1, ...s_i, ...s_{N_S}\}, 1 \leq i \leq N_S$ and the set of users as $U = \{u_1, ...u_j, ...u_{N_U}\}, 1 \leq j \leq N_U$. Each SMNO s_i operates a network consisting of N_R^i contiguous Resource Blocks (RBs). The price that a user u_j pays for their bundle with a fixed data allowance D is given by p_j.

Each device estimates its received Signal-to-Noise and Interference Ratio (SINR) using reference signals placed throughout the resource grid and then maps this SINR value to a 4-bit index known as the Channel Quality Indicator (CQI), with each CQI index corresponding to a specific modulation scheme and coding rate as specified in Table 7.2.3-1 in [171].

We adopt higher layer configured sub-band reporting, as per Long Term Evolution (LTE), whereby the band is divided into sub-bands and the User Equipment (UE) reports a single wideband CQI and a CQI for each sub-band. For the purpose of CQI reporting, RBs are grouped into sub-bands consisting of h RBs. Each SMNO's s_i network is therefore divided into $N_B^i = N_R^i/h$ sub-bands. We assume that each user is granted a single sub-band; hence, the number of matches that an SMNO s_i is able to make, known as its quota q^i, is also N_B^i.

Each CQI value corresponds to a particular modulation order and coding rate and hence can be mapped to a measure of efficiency, which can be interpreted as the number of information bits per symbol (see Table 7.2.3-1 in [171]). The spectral efficiency for a user u_j reporting on sub-band b of SMNO s_i is denoted by $\psi_j^{i,b}$. The spectral efficiencies for all users across all SMNOs form the matrix Ψ, which has dimensions $N_U \times N_S \times N_B^i$ (assuming that each SMNO has control over an equal number of sub-bands). Ψ_j^i indexes the vector corresponding to the spectral efficiencies of user u_j for each sub-band b in SMNO s_i.

Resource allocation in a broker-based model is a two step process. First, users must be matched to SMNOs. This is the primary focus in this chapter and we will use matching theory algorithms to accomplish this. However, it should be noted that users are matched to SMNOs and not specific sub-bands. Hence, after this matching has been performed, each SMNO must perform the second step of allocating specific sub-bands to its matched users. SMNOs can use custom schedulers based on well-known approaches such as proportional-fair or maximum throughput to accomplish this.

6.2 Preference Lists

In this section, we outline how to build the preference lists of both users and SMNOs, which are used by the Gale-Shapley algorithm.

6.2.1 User Preference List

We will adopt the concept of utility from the field of economics as a measure of preference that can be used to generate a preference list for users. Utility represents the value (often confined to a real number between 0 and 1) assigned by a user to a service according to the service's performance. A utility function therefore maps a QoS metric such as achievable rate to an abstract unit that captures the utility or value of the service. In this sense, the concept of utility is related to the idea of quality of experience, which examines users' perceptions of a service. Utility-based resource optimisation for wireless networks has received quite a lot of attention for networks with multiple classes of traffic (e.g. [172, 173]).

The user has the following information available to it:

1. QoS guarantees: We assume that an SMNO advertises its QoS guarantees to users.

2. SINR: The user uses reference signals to determine its SINR, and uses this to make an estimate of the rate that it can achieve using a particular SMNO.

Hence, the user has four QoS metrics available to it: average latency, strict latency, packet loss ratio, and achievable rate.

Utility is a service-dependent concept and hence each service that may be employed by a user has an associated utility function. Each utility function takes the four QoS metrics outlined in the previous paragraph as inputs. However, we will first examine how to represent the individual relationships between the QoS metrics and utility. To do this, we will utilise two functions: the normalised logarithmic function and the normalised sigmoid function. In theory, any number of different functions could be used. These two functions are popular in the literature on utility-based optimisation of systems as they allow a range of common relationships to be modelled through correct parametrisation.

The normalised logarithmic function is expressed as:

$$\Omega_{\text{log}}(x) = \frac{\log(1 + kx)}{\log(1 + kx_{\text{max}})}, \tag{6.1}$$

where x is the QoS metric that we are mapping to a utility measure between 0 and 1. x_{max} is the maximum achievable value of the QoS metric (such as maximum rate). k is the rate of increase of the utility measure in relation to the QoS metric. The normalised logarithmic utility function, assuming it is parametrised correctly, can be used to capture any relationship between a QoS metric and service utility measure that is monotonically increasing, given that the performance of the service is relatively elastic in relation to that QoS metric (i.e. not critically reliant on the QoS metric). Monotonically decreasing relationships can be captured using $1 - \Omega_{\text{log}}(x)$. An example of a QoS metric with a monotonically increasing logarithmic relationship with utility would be data rate for best effort communication applications.

The normalised sigmoid function is given by:

$$\Omega_{\text{sig}}(x) = \frac{1}{1 + e^{-a(x-b)}}, \tag{6.2}$$

where a captures the steepness or slope of the curve, while b represents the inflection point of the curve. The normalised sigmoid function can be used to represent monotonically increasing relationships whereby utility has a strict reliance on a particular QoS metric; for example, the utility can be modelled as a sigmoid function of the data rate when a minimum throughput is essential to the performance of the service (such as a video streaming service).

Hence, for each service we will define the relationships between the four available QoS metrics and utility using one of four functions discussed above: $\Omega_{\text{log}}(x)$, $1 - \Omega_{\text{log}}(x)$, $\Omega_{\text{sig}}(x)$ and $1 - \Omega_{\text{sig}}(x)$. Fig. 6.1 shows an example of the shapes of these relationships. The relationship between a QoS metric and utility may be different for different services. For example, the rate relationship may be modelled as $\Omega_{\text{log}}(x)$ for a file sharing service, or

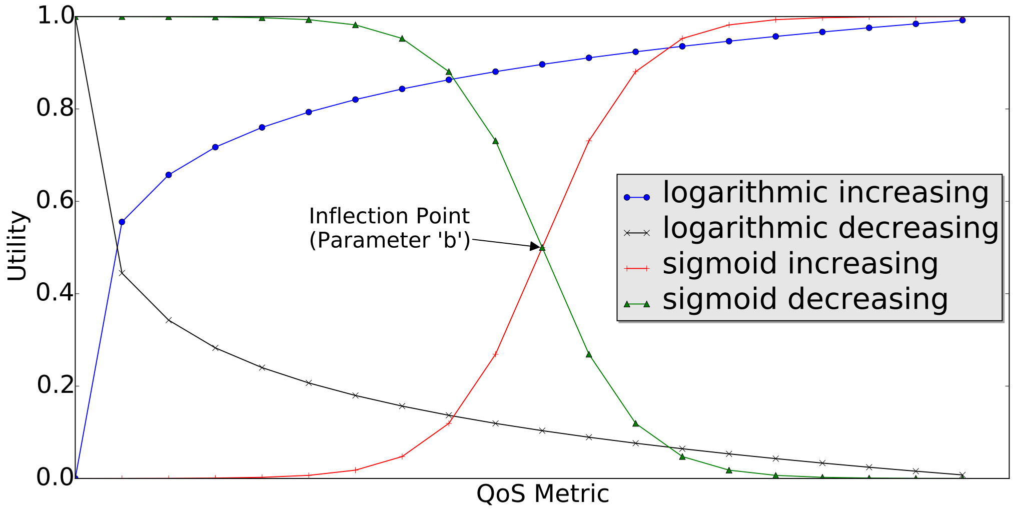


Figure 6.1: The shapes of four functions $\Omega_{\text{log}}(x)$, $1-\Omega_{\text{log}}(x)$, $\Omega_{\text{sig}}(x)$ and $1-\Omega_{\text{sig}}(x)$ which can be used to capture different types of relationships between a QoS metric and utility. For the sigmoid functions, the parameter a determines the steepness of the curve and b determines the inflection point.

$\Omega_{\text{sig}}(x)$ for a video streaming service. The parameters for each relationship are also service-dependent and must be empirically determined.

For each service, having characterised the individual relationships between the QoS metrics and utility, we must combine these into a single utility measure. The choice of method (e.g. multiplicative or additive) again depends on the particular details of a given service [174]. In this chapter, we will adopt a simple multiplicative model in which the four individual utility values are multiplied together to obtain a single overall utility value for the service. Ω_j^i denotes the overall utility, and therefore preference, of user u_j for SMNO s_i.

We comment that users generate a preference measure for an SMNO, although SINR is estimated on a sub-band granularity. Hence, when estimating the achievable rate for an SMNO, the user adopts an optimistic outlook and chooses the sub-band corresponding to the maximum SINR in an attempt to achieve the maximum performance possible.

As a final remark, we note that the proposed model, which consists of a subscription broker in which both SMNOs and users indicate their preference as part of a matching process, has implications regarding the fulfilment of the SLA. Typically, a service provider would suffer a penalty if the terms of an SLA were violated. Since the user now plays a role in the matching process, a case could be made that some of the responsibility for maintaining the terms of the SLA now lies with the user. To clarify this, we highlight that the primary decision that the user is making is to choose a SMNO. Hence, the SMNO should be required to satisfy the SLA when chosen. However, since the SMNO also indicates a preference for individual users, the SLA terms could be softened by specifying that the agreed terms only apply when both a user and SMNO are matched based on a mutual preference, which can be interpreted as an indication that the SMNO is acknowledging that it can provide the required QoS to the user.

6.2.2 SMNO Preference List

For SMNOs, the information available locally which can be used to build a preference list includes:

1. Channel State Information (CSI) for each user;
2. the package that each user is subscribed to.

SMNOs wish to maximise the share of the revenue that they receive from the subscription broker for serving users subscribed to bundles. This can be achieved in two ways: by prioritising users subscribed to more expensive bundles, and by maximising the number of useful information bits that can be transferred using a single resource (i.e. maximising its spectral efficiency). SMNOs will therefore prefer users with good channel conditions subscribed to more expensive bundles. Hence, to maximise its revenue in a given time slot, an SMNO s_i wishes to solve the following maximisation problem:

$$\text{P1} : \max_{\omega_{jr}} \sum_{j=1}^{N_U} \sum_{r=1}^{N_R^i} \omega_{jr} z_i \frac{p_j}{D} \psi_j^{i,o(r)} \tag{6.3}$$

subject to

$$\omega_{jr} \in \{0,1\}, \forall j, r \in \mathbb{N}, j \leq N_U, r \leq N_R^i, \tag{6.3a}$$

$$\sum_{j=1}^{N_U} \omega_{jr} \leq 1, \forall r \in \mathbb{N}, r \leq N_R^i, \tag{6.3b}$$

where ω_{jr} is a binary indicator variable which is equal to 1 if user u_j is scheduled on resource block r, and 0 otherwise. $o(r)$ maps a resource block r to the sub-band containing it, and z_i specifies the number of information carrying symbols (i.e. excluding pilots and symbols belonging to control channels) in a resource block for SMNO s_i. The aim of the maximisation problem is to increase the revenue per resource block through a combination of increasing the price paid per bit and fitting more information bits into a resource block. Constraint (6.3a) ensures that the variable ω_{jr} only takes on binary values, while constraint (6.3b) ensures that a resource block is only allocated to a single user.

Although the solution of P1 would yield the optimal matching of users and resource blocks for a single SMNO, it is not practical as each SMNO is competing for users and cannot simply choose its users. It also does not rank users in terms of their optimality; it just provides the optimal set. Hence, each SMNO applies Algorithm 1, which is based on optimisation problem P1, to generate its preference list. Note that estimating the revenue

Algorithm 1 Generate preference list for SMNO i

```
DECLARE RevenuePerUser: ARRAY[1,$N_U$] of (FLOAT, USER) TUPLES
DECLARE PreferenceList: ARRAY[1,$N_U$] of USERS
for all $u$ in $U$ do
    $\psi_{\max} \leftarrow \max(\Psi_u^i)$
    RevenuePerUser[u] $\leftarrow (z_i \frac{p_u}{D} \psi_{\max}, u)$
end for
SORT RevenuePerUser: descending, sort on first element in tuple
COMMENT: below we extract user IDs using the second element of the tuple in the sorted RevenuePerUser list
$n \leftarrow 1$
for all $j$ in RevenuePerUser do
    PreferenceList[n] $\leftarrow j[1]$
    $n \leftarrow n + 1$
end for
```

that can be obtained from serving a particular user is difficult, as the SMNO will not allocate its resource blocks to users until after the user-SMNO matching has been performed (i.e. it cannot allocate its resource blocks until it knows which users have been matched with it). Hence, each SMNO estimates the potential revenue to be earned from serving a particular user by summing the revenue earned from granting the user each of the available sub-bands.

6.3 Stable Matching using the Gale-Shapley Algorithm

Traditionally, the field of economics has focused on markets in which the price of an article of interest is free to adjust so that supply equals demand. However, in our case, given that a fixed price for access is paid in advance, the price is constrained and other approaches for matching users and slices must be utilised. Such markets, in which price does not play a role, are known as matching markets, and are the subject of a branch of economics known as matching theory.

In addition, when matching users to an SMNO belonging to their subscription bundle, both users and SMNOs must be satisfied that they could not have done better than the match provided to them, otherwise they would not be incentivised to agree to adhere to the broker-based model. To ensure this, we use the Gale-Shapley *college admission* algorithm [166] from the field of matching theory to match users to SMNOs. The Gale-Shapley algorithm guarantees a stable matching between two sets of entities based on the preferences of individual entities; a stable matching is one where there is no pair that would prefer to be matched to each other instead of their current partners.

We will model our matching problem between SMNOs and users as a college admission problem, as named by Gale and Shapley [166] when referring to two-sided, many-to-one matching problems. Our problem is two-sided because it consists of two disjoint sets, and a matching must involve one entity from each set. Our problem is described as many-to-one because each SMNO can form as many matchings with users as its quota permits. The Gale-Shapley college admission algorithm can be used to obtain a stable matching in these kind of problems. Each entity must rank the entities of the other set in order of preference. We assume individual rationality, meaning that each user would prefer to be matched to any SMNO than not matched at all. Hence, each SMNO is dimensioned so that it could provide adequate QoS to any user, although some SMNOs are better suited to some users (depending on the service in use). Individual rationality also implies complete preference lists, i.e. all entities of the opposite set are included in each preference list.

The college admission algorithm works based on a series of proposals from one set to the other. For example, if the users do the proposing, then each user will first propose to its preferred SMNO. The user will then pause if the SMNO provisionally accepts the proposal, and will propose to the next SMNO in its preference list if rejected. SMNOs will provisionally accept a proposal if they have not filled their quota or if the current proposal gives them a better matching than one of its provisionally accepted matchings, and reject the proposal otherwise. The process ends when either all users are provisionally engaged, or no more possible feasible matchings exist.

We now formally define the preference relation for users and SMNOs in Definitions 6.3.1 and 6.3.2, respectively. We use the notation $a \succ_x b$ to define preference, meaning that entity x prefers a to b.

Definition 6.3.1. u_j prefers s_i to $s_{i'}$, if $\Omega_j^i > \Omega_j^{i'}$, denoted by $s_i \succ_{u_j} s_{i'}$, for $u_j \in U$, $s_i, s_{i'} \in S$, $s_i \neq s_{i'}$.

Definition 6.3.2. s_i prefers u_j to $u_{j'}$, if $z_i \times \frac{p_j}{D} \times \max(\Psi_j^i) > z_i \times \frac{p_{j'}}{D} \times \max(\Psi_{j'}^i)$, denoted by $u_j \succ_{s_i} u_{j'}$, for $s_i \in S$, $u_j, u_{j'} \in U$, $u_j \neq u_{j'}$.

We also define the notion of stability in relation to the matching problem outlined in this chapter in Definition 6.3.3.

Definition 6.3.3. A matching M is stable if there exists no user-SMNO pair (u_j, s_i) such that u_j would prefer to be matched to s_i than to its current partner, and s_i would prefer to be matched to u_j than its current partner. That is, there exists no user-SMNO pair (u_j, s_i) such that $s_i \succ_{u_j} M(u_j)$ and $u_j \succ_{s_i} M(s_i)$.

6.4 Cost of Achieving Stability

The Gale-Shapley college admissions algorithm guarantees stability, which is important if both users and SMNOs are to commit to the subscription broker business model. However, we wish to compare the performance of the algorithm with conventional scheduling approaches in order to determine the cost, or performance loss, if any, in achieving stability.

It should be noted, however, that there is no such thing as a conventional scheduler for a multi-SMNO, multi-service system model, as this is still an open research problem.

Although we are free to choose any number of metrics, as system designers, we will state our goal as one of maximising user utility. To simplify the task, we will flatten the problem from a one-to-many mapping between SMNOs and users, to a one-to-one mapping between sub-bands and users. This allows us to express the optimisation problem as one of finding the maximum weight matching in a weighted bipartite graph, which can be solved using the well-known Kuhn-Munkres algorithm.

To simplify the optimisation problem, we assume that allocations are performed at sub-band granularity, and that each user is granted a maximum of one sub-band. Hence, while Ω_j^i provides a user's u_j utility estimate for a SMNO s_i, we instead consider $\Omega_j^{b,i}$, which provides a utility estimate on a sub-band granularity. This allows us to model the problem as a one-to-one matching between users and sub-bands.

The resource allocation problem involving users and sub-bands is outlined in the following optimisation problem:

$$\mathrm{P2} : \max_{\omega_{jib}} \sum_{j=1}^{N_U} \sum_{i=1}^{N_S} \sum_{b=1}^{N_B^i} \omega_{jib} \Omega_j^{b,i}, \tag{6.4}$$

subject to

$$\omega_{jib} \in \{0,1\}, \forall j, i, b \in \mathbb{N}, j \leq N_U, i \leq N_S, b \leq N_B^i, \tag{6.4a}$$

$$\sum_{j=1}^{N_U} \omega_{jib} \leq 1, \forall i, b \in \mathbb{N}, i \leq N_S, b \leq N_B^i, \tag{6.4b}$$

$$\sum_{i=1}^{N_S} \sum_{b=1}^{N_B^i} \omega_{jib} \leq 1, \forall j \in \mathbb{N}, j \leq N_U, \tag{6.4c}$$

where ω_{jib} is a binary indicator variable which is 1 when user j uses sub-band b of SMNO i, and 0 otherwise. Constraint (6.4a) ensures that the variable ω_{jib} only takes on binary values. Constraint (6.4b) ensures that each user is matched with only one sub-band, and constraint (6.4c) ensures that each sub-band is only allocated to a single user.

6.5 Evaluation Parameters

We use Monte Carlo simulations to evaluate the performance of the broker-based model in an example case study. We consider four SMNOs targeting enhanced Mobile Broadband (eMBB) traffic, with each SMNO occupying four sub-bands and 24 resource blocks (six resource blocks per sub-band). The SMNOs advertise their packet-loss ratio, average latency, and strict latency as specified in Table 6.1. SMNOs also advertise a rate multiplier, which is used to scale the rate to account for differences in spatial streams, coding rates, and other rate-affecting factors. A multiplier of one corresponds to the standard LTE values given by Table 7.2.3-1; [171], with only a single spatial stream.

SMNO 1 offers the best data rate, using multiple spatial streams, but has the highest strict and average latencies due to the additional processing and training needed. SMNO 2 offers the lowest average and strict latency by reducing control signalling procedures, resulting in the highest packet loss ratio. SMNO 3 offers the best reliability by employing additional redundancy which results in the lowest data rate advertised. SMNO 4 is an *all-rounder* network, offering decent values for all metrics without excelling in any particular area.

To ensure fair comparison between the stable matching and maximum-utility approaches, we assume that each user gets allocated a single sub-band. The number of users will be varied so that the network is fully loaded (16 users[1], 100% capacity) and overloaded (20 users, 125% capacity).

Each user may employ one of three services, with the relationship between utility and each of the individual QoS metrics specified in Table 6.2. As mentioned in Section 6.2.1, these relationships are service-dependent and should be empirically determined on a service-by-service basis using real-world data. As this is not available to us, we have attempted to choose sensible values for these services. We note, however, that the actual values chosen here are not of great importance. Future networks should be able to cope with any service, including those yet unforeseen.

Service S1 presents the most stringent reliability requirements, moderate data rates, and is relatively latency-tolerant. Service S2 has the lowest latency requirements of all services, but only requires fairly low data rates. Finally, Service S3 has the highest data rate requirements, as well as moderate latency and reliability requirements. As stated in the previous paragraph, we are interested in the system's ability to accommodate any service, rather than specifically the three examples presented.

The subscription bundler sells two subscription packages, A and B, with B costing twice as much as A. Both packages include service level agreements with all four SMNOs, and consist of a fixed data allowance D. After users have been matched to SMNOs, each SMNO allocates sub-bands to its matched users using a maximum throughput scheduler.

6.6 Results

We divide the results presented in Figs. 6.2-6.5 into five key sets of comparisons.

[1] There are 4 SMNOs with 4 sub-bands each, and we assume that each user gets granted a single sub-band.

Table 6.1: Advertised SMNO guarantees.

SMNO ID	Average Latency (ms)	Strict Latency (ms)	Packet Error Rate (per 10^6 packets)	Rate Multiplier
1	10	14	7	2
2	4	8	10	1
3	9	12	2	0.6
4	8	12	8	1

Table 6.2: Utility-QoS relationships for services.

Service	Average Latency (ms)	Strict Latency (ms)	Packet Error Rate (per 10^6 packets)	Data Rate (Mbps)
	1-sigmoid	1-sigmoid	1-sigmoid	logarithmic
S1	a=0.5, b=18	a=1, b=20	a=1, b=10	k=100, $x_{\max}$=2.5
S2	a=0.5, b=12	a=1, b=15	a=0.5, b=15	k=100, $x_{\max}$=1.5
S3	a=0.5, b=16	a=1, b=18	a=0.5, b=15	k=100, $x_{\max}$=5

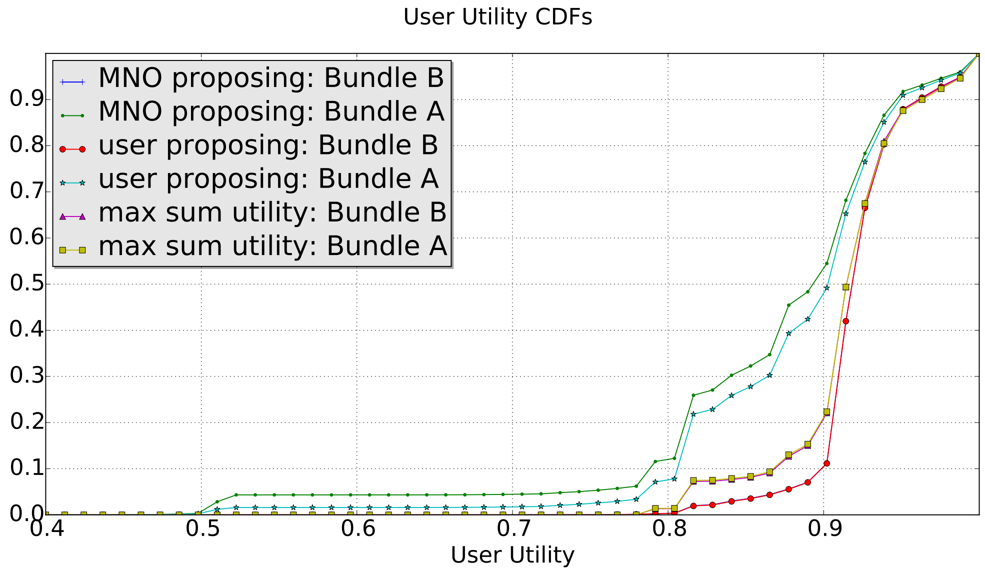


Figure 6.2: Comparison of user-proposing algorithm, SMNO-proposing algorithm, and maximum utility approach for subscribers of both bundles.

Bundle A vs Bundle B

As expected, Fig. 6.2 shows that Bundle B users will achieve greater performance in terms of utility compared to Bundle A users, since Bundle B subscribers are more likely to be at the top of each SMNO's preference list due to the higher revenue that can be earned from them. Note that this only applies when a matching theory algorithm is used, as the utility maximisation approach does not distinguish between Bundle A and Bundle B subscribers. For the most part, the difference in performance between subscribers of the two bundles is

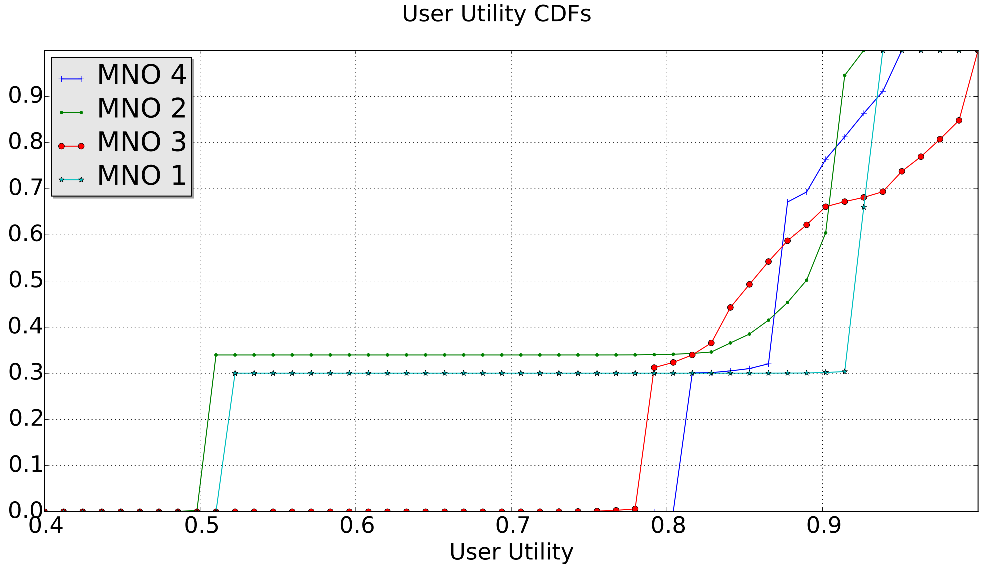
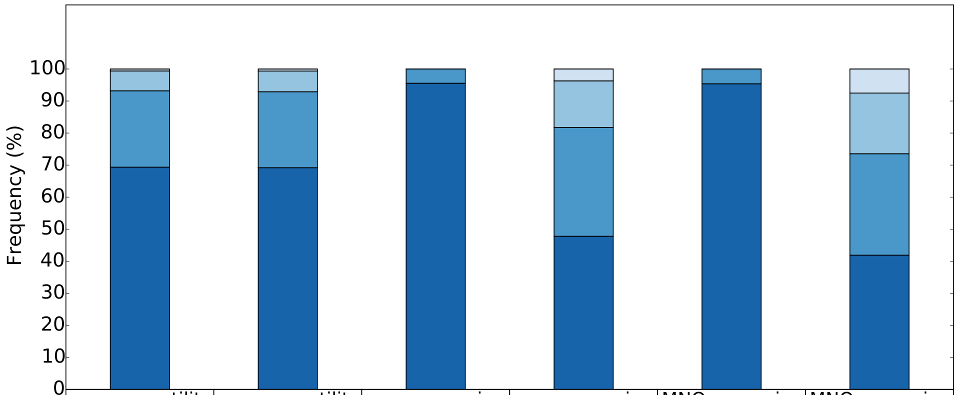


Figure 6.3: User performance without a broker-based network model; no SMNO can provide high performance to all users.

relatively small (less than 0.1 of a utility measure); however, while no Bundle B subscriber experiences a utility of less than 0.8, 10% of Bundle A users experience a utility between 0.5 and 0.8. Hence, subscribing to Bundle B in this case ensures a respectable minimum performance. The matching statistics provided in Fig. 6.4 show that Bundle B users are slightly more than twice as likely to be matched to their first choice than Bundle A users, with no Bundle B subscribers matched with their least preferred choice.

Optimal Sum-Utility vs Gale-Shapley Algorithm

As mentioned in the previous paragraph, Bundle A and Bundle B users achieve the same performance in the sum utility maximisation approach. As seen in Fig. 6.2, when the Gale-Shapley algorithm is adopted, Bundle B subscribers achieve marginally better performance compared to the utility maximisation approach at the expense of Bundle A users. Hence, the performance cost, in terms of utility, of achieving stability is primarily borne by Bundle A subscribers. The combined performance of all users, both A and B subscribers, will be slightly less when the college admissions algorithm is used instead of the sum utility maximisation approach. However, the advantage of using the college admissions algorithm is that stability is achieved. This is important if the broker-based model of packaging service level agreements into bundles is to be adopted. As per definition 6.3.3, stability ensures that there is no user-SMNO pair that would prefer to be matched to each other than to their appointed matches. Without this condition, neither SMNOs nor users would be incentivised to use a broker-based model, as some users would feel that they could gain an advantage by engaging in direct service level agreements with the SMNOs.

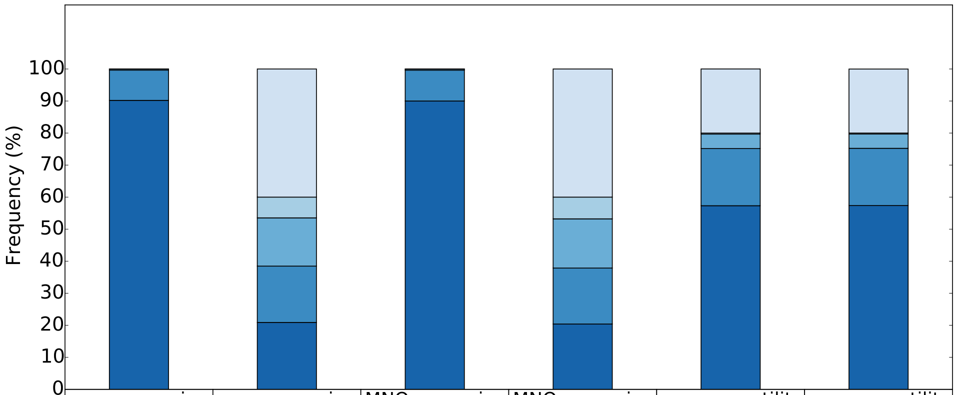


	max sum utility Bundle B	max sum utility Bundle A	user proposing Bundle B	user proposing Bundle A	MNO proposing Bundle B	MNO proposing Bundle A
1st Choice	69.35 %	69.19 %	95.55 %	47.79 %	95.39 %	41.88 %
2nd Choice	23.85 %	23.71 %	4.43 %	33.96 %	4.58 %	31.67 %
3rd Choice	6.19 %	6.46 %	0.02 %	14.55 %	0.03 %	18.93 %
4th Choice	0.61 %	0.64 %	0.0 %	3.71 %	0.0 %	7.52 %
No Match	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %

Figure 6.4: Matching statistics for a fully loaded network; higher priority Bundle B users achieve a higher percentage of first choice matches.

Broker-Based Approach v One-Size-Fits-All Network

Fig. 6.3 shows the performance that would be achieved if users only had an SLA with a single SMNO. None of the SMNOs on their own can provide really high performance to all users, with SMNOs 1 and 2 in particular providing barely adequate performance to about 30% of users. The advantage of the broker-based model is that bundle subscribers can use their data allowance on any SMNO as needed. For example, Fig. 6.2 shows that when using the broker-based model and the Gale-Shapley algorithm, over 90% of Bundle B subscribers achieve a utility of 0.9 or higher, with no subscriber achieving less then 0.8 utility. This is an improvement over any single SMNO in Fig. 6.3. In the broker-based model, using the user-proposing college admission algorithm, Bundle A users also achieve performance that is either better than or comparable to any individual SMNO, with over 90% of users achieving a utility of 0.8 or more, and over 50% of users achieving a utility of 0.9 or more (see Fig. 6.2). SMNO 4, the catch-all network, comes closest to matching this on its own, offering a slight advantage in that no user achieves a utility less than 0.8. Hence, the broker-based model using the college admissions algorithm allows users to select the SMNO best suited to the service in use, resulting in improved performance to Bundle B subscribers, and at least as good performance to Bundle A subscribers when compared to a single one-size-fits-all network approach.



	user proposing Bundle B	user proposing Bundle A	MNO proposing Bundle B	MNO proposing Bundle A	max sum utility Bundle B	max sum utility Bundle A
1st Choice	90.2 %	20.88 %	90.0 %	20.39 %	57.32 %	57.37 %
2nd Choice	9.44 %	17.59 %	9.63 %	17.47 %	17.84 %	17.87 %
3rd Choice	0.35 %	15.06 %	0.36 %	15.36 %	4.51 %	4.45 %
4th Choice	0.01 %	6.47 %	0.01 %	6.78 %	0.32 %	0.31 %
No Match	0.0 %	40.0 %	0.0 %	40.0 %	20.02 %	19.98 %

Figure 6.5: Matching statistics when network is 125% loaded; low priority Bundle A users suffer more than Bundle B users.

User-Proposing vs SMNO-Proposing

As described in Section 6.4, the Gale-Shapley college admissions algorithm can be performed in user-proposing form, or SMNO-proposing form. Since Bundle B subscribers are more profitable to SMNOs, they are generally preferred by SMNOs over Bundle A users. Hence, under the SMNO-proposing variant, Bundle B users are generally proposed to first while under the user-proposing variant, Bundle B users generally have their proposals accepted by SMNOs. Hence, in both cases, Bundle B users achieve a very high percentage of first choice matches and there is no difference in performance. The case is different for Bundle A subscribers. Under the SMNO-proposing variant, an SMNO can propose to as many Bundle B subscribers as it wants before considering Bundle A subscribers. However, under the user-proposing variant, SMNOs will only receive proposals from some of the Bundle B subscribers. As a result, a Bundle A user will compete with fewer Bundle B users under the user-proposing variant than under the SMNO-proposing variant. Hence, Fig. 6.4 shows that a higher percentage of Bundle A subscribers are matched with their first choice under the user-proposing version than under the SMNO-proposing version. This translates to a small performance improvement in achieved utility, as can be observed in Fig. 6.2.

Fully-Loaded vs Over-Loaded

Fig. 6.5 shows the matching statistics when the network is at 125% capacity. Due to the mismatch in cardinality between the set of sub-bands and the set of users, some users will be unmatched after the algorithm is performed. We observe that when the sum utility maximisation algorithm is used, the unmatched users are distributed equally between Bundle A and Bundle B. In the case a matching theory algorithm is used instead, all unmatched users are Bundle A subscribers, with Bundle B users largely unaffected. Hence, while Bundle B subscribers only have a small advantage over Bundle A subscribers in the fully-loaded case (Fig. 6.2), they obtain a much greater advantage when the network is over-loaded, as they are never left unmatched.

6.7 Conclusion

The evaluation of our case study demonstrates the improvement in performance that can be obtained if the proposed broker-based network model is adopted compared to a one-size-fits-all network. The benefits of the broker-based model, which is based on bundles of service level agreements with multiple SMNOs, lie in its ability to match users to suitable SMNOs according to the requirements of the service that the users are employing at that moment in time. However, for the broker-based model to be successfully adopted, both users and SMNOs must be satisfied that they could not have achieved a better match. We adopt the Gale-Shapley college admissions algorithm, which provides a stable matching, to guarantee that this condition is met. We investigated a system comprising two classes of users, and proposed an approach for building the preference lists of SMNOs that relies on the self-interest of operators to maximise their own revenue, and is able to inherently differentiate between different classes of users. We show that compared to a utility maximisation approach the performance cost, in terms of utility, of achieving stability is carried by the lower priority users. Overall, the broker-based model outperforms any one SMNO for the higher priority users, while maintaining a level of performance that is at least as good as any one SMNO for lower priority users.

7 On Provisioning Slices and Overbooking Resources

On Provisioning Slices and Overbooking Resources

The motivation for adopting the network slicing paradigm in telecommunication networks stems from two main benefits that it can potentially provide to operators and end users of the network, which are described in Section 3.2.1:

1. *Network slicing enables greater resource sharing;*
2. *Network slices can be tailored for different verticals.*

The two points above are in conflict. If providing bespoke performance is the crux of network slicing, then slices need to provide a set of performance guarantees (such as rate, latency, etc.) to its subscribers. To satisfy these guarantees, each slice must have a minimum quantity of resources (e.g. resource blocks, processing power) available to it for every user that it must serve. Reserving dedicated resources for every slice reduces the statistical multiplexing gain achievable. Conversely, permitting highly dynamic sharing hinders a slice's ability to provide a guarantee of performance to its subscribers. The contrasting goals in network slicing of providing guaranteed customised services while simultaneously increasing resource utilisation is highlighted in [175], demonstrating the potential efficiency gap resulting from this trade-off.

We consider a slicing business model, described in Section 3.4, comprising a slice provider, slice tenants, and subscribers to each slice tenant. A slice tenant requires assurances over the course of its lifetime regarding the availability of resources from the slice provider so that it can provide a high degree of consistency in the service it offers to its subscriber base. The slice provider wishes to maximise resource utilisation and profit by taking advantage of the short-term fluctuations in the demand across its tenants.

We propose to control the balance in this trade-off between tailored slice behaviour and resource efficiency using a combination of long-term and short-term commitments regarding resource availability. In the long-term, each slice tenant engages in a service contract with the slice provider in which it requests a number of assured resources. These resources are always available to the slice tenant, if needed, for the duration of the contract, affording the slice tenants confidence in the performance that they can offer to their own subscribers.

In the short-term, in addition to its assured resources, a slice tenant may obtain auxiliary resources. Providing auxiliary resources on a best-effort basis is not conducive to maintaining a predictable level of service to slice tenants' subscribers. Instead, the slice provider forecasts the resource demand of the slice tenants in the near future and informs each slice tenant of the predicted quantity of auxiliary resources that will be available to it with a stated probability. The auxiliary resources allow the slice provider to obtain statistical multiplexing gains in the short-term, while the information regarding their availability allows slices to make informed scheduling decisions to ensure that they maintain their desired service quality.

For example, if the slice provider predicts that slice A is unlikely to require all of its assured resources in an upcoming time-interval, it might offer more auxiliary resources to slice B. A slice provider is not obligated to offer any auxiliary resources to a slice; however, if it does offer any resources, it must ensure that the resources are available with a pre-defined confidence level.

To efficiently provide auxiliary resources with a specified probability of availability, the slice provider may rely on overbooking. The practice of overbooking is widely used in many industries, such as airlines [176, 177], to ensure maximum resource utilisation by allocating more resources than actually available to resource consumers.

Research Question, Key Contributions and Chapter Organisation

The research question, outlined in Section 1.1, that this chapter addresses is the following:

How can the twin goals of slice-tailored performance and increased resource utilisation in network slicing be simultaneously realised in future networks?

Accurate prediction of resource demand is critically important when performing overbooking, and the practice is examined in the context of network slicing in [178], in which traffic analysis and prediction are used to inform admission control for slice requests.

The authors in [179] employ overbooking techniques for provisioning resources in shared hosting platforms, with a probabilistic guarantee of resource availability provided to resource consumers. The authors argue that the use of overbooking incentivises the deployment of viable hosting platforms. For the resource provider, overbooking increases resource utilisation and profit. For the resource consumer, the associated guarantee of availability is more valuable than best-effort approaches. Adopting the same motivation, we apply overbooking in the context of network slicing, and provide a combination of both long and short-term guarantees on resource availability to slice tenants.

The substantial revenue gains that slice overbooking can provide in telecommunications in realistic scenarios is demonstrated in [180], with experimental proofs-of-concept outlined in [181] and [182]. In the aforementioned works, a slice provider is periodically presented with a number of slice requests and employs forecasting and overbooking to decide how many requests to facilitate. In this chapter, we instead propose that the slice provider applies overbooking to increase the efficiency in the use of its resources, and avoid over-provisioning, when offering auxiliary resources to already admitted slices.

Our primary contribution in this chapter centres on a solution to manage the trade-off between slice-tailored network behaviour and increased resource sharing between slices. Our contributions are as follows:

- We propose a system consisting of assured resources, which are guaranteed over the lifetime of a slice tenant, and auxiliary resources, which are offered on a short term basis with an associated probabilistic guarantee of availability.
- We devise an algorithm for determining how many auxiliary resources can be offered according to the probabilistic guarantee, taking current resource demand into account.
- We detail how to distribute auxiliary resources among slice tenants in such a manner to maximise the revenue of the slice provider.
- We specify how the slice provider can limit the damage to any individual slice tenant when the system is over-subscribed.
- We provide insight into the factors which affect the overbooking of auxiliary resources, and show when conditions dictate that overbooking will be effective using both analytical and numerical results.

The rest of the chapter is structured as follows. Section 7.1 describes the system model. Section 7.2 consists of three parts. We first derive various probabilities associated with the availability of auxiliary resources and use these probabilities to design an algorithm to determine how much to overbook by. We then propose several ways of distributing auxiliary offers among slices, before finally providing an approach for distributing resources when the system is over-subscribed. In Section 7.3, we provide analytical results for a case of two slices with uniform resource demand to demonstrate how the inputs into the system affect the performance of resource overbooking. We expand this analysis to a more complex model consisting of multiple slices with heterogeneous resource demand profiles in Section 7.4, and provide numerical results to substantiate the findings from our theoretical model. Finally, Section 7.5 concludes the chapter by providing insights into the favourable conditions for overbooking.

7.1 System Model

Slice tenants lease an arbitrary virtual resource from the slice provider and operate slices $s \in S$ as independent business ventures. We assume that the resources (e.g., spectrum sub-bands or time-frequency resource blocks) come from a countable set with cardinality K.

Each slice tenant, denoted by the slice s that it operates, requests a quantity a_s of assured resources from the slice provider in the service contract. To ensure the availability of assured resources over the lifetime of the contract with the slice tenant, the admission

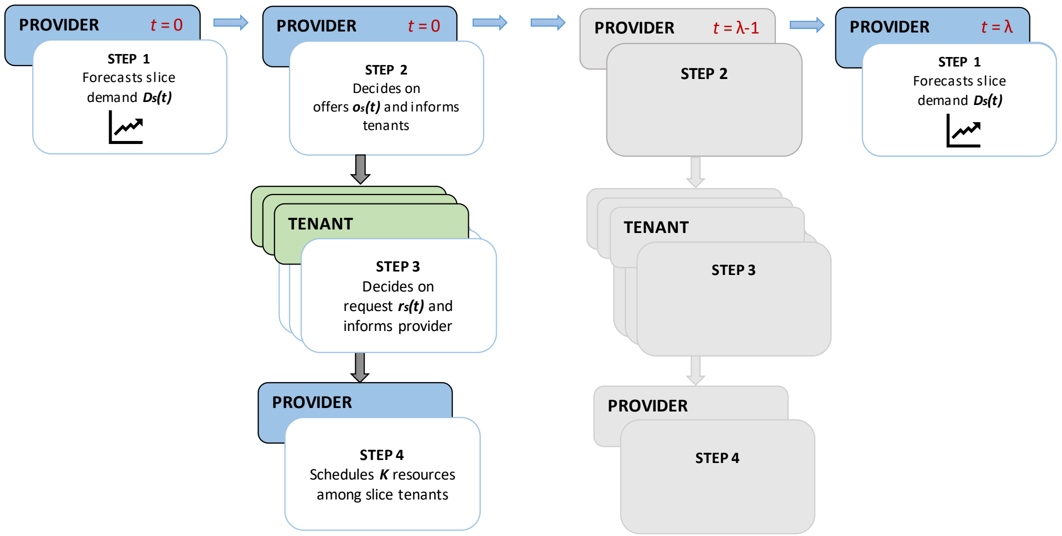

PROVIDER t = 0
STEP 1
Forecasts slice demand Ds(t)
PROVIDER t = 0
STEP 2
Decides on offers os(t) and informs tenants
TENANT
STEP 3
Decides on request rs(t) and informs provider
PROVIDER
STEP 4
Schedules K resources among slice tenants
PROVIDER t = λ-1
STEP 2
TENANT
STEP 3
PROVIDER
STEP 4
PROVIDER t = λ
STEP 1
Forecasts slice demand Ds(t)

Figure 7.1: Illustration of the four-step process of scheduling resources among tenants. The slice provider predicts the resource demand of each slice every λ scheduling windows and issues offers of auxiliary resources for each scheduling window.

control process ensures that the following condition is met at all times:

$$\sum_{s \in S} a_s \leq K. \tag{7.1}$$

The fundamental unit of time is the scheduling window, indexed t, during which the slice provider schedules its K resources among the $|S| = N$ slice tenants, who in turn serve their own subscribers. This process of scheduling resources among tenants consists of four steps, as illustrated in Fig. 7.1.

1. Demand forecast

Every λ scheduling windows, the slice provider forecasts the underlying resource demand for each of its slice tenants for the next λ scheduling windows. The underlying resource demand for a slice is a random process, with each time instance t representing a random variable $D_s(t)$. To provide tenants with information regarding the availability of auxiliary resources at a time t, the slice provider must have knowledge of the distribution of the random variable at time t. The slice provider's estimate of $D_s(t)$ is denoted $\widehat{D}_s(t)$.

Demand forecasting is common practice in communication networks, and there is a wealth of research in this area. The authors in [183] note the importance of predicting traffic in the Radio Access Network (RAN) and analyse the practice of traffic-aware networking. While time-series forecasting and statistical methods have traditionally been employed, advances in machine learning offer significant potential in this area [184, 185]. Demand prediction has even recently been considered for the purpose of resource maximisation in

RAN slicing [178].

Demand prediction techniques are not within the scope of this thesis. Instead, we assume that the slice provider, at any time t, can estimate the joint Probability Density Function (PDF) $f_{\widehat{D}_1(t)\ldots\widehat{D}_N(t)}(\widehat{d}_1(t)\ldots\widehat{d}_N(t))$ of the resource demand of all the slice tenants. The joint PDF can be used to obtain the marginal PDFs for the demand of individual slices (characterised by expected value $\bar{d}_s(t)$ and variance $\widehat{\sigma}_s^2(t)$).

2. Slice provider offers auxiliary resources to slice tenants

The slice provider then uses this forecast to determine how many auxiliary resources $o_s(t)$ it can offer to each slice tenant in each scheduling window; this offer comes with probabilistic guarantees, namely that the offered resources will be available for at least a fraction g of the time slots.

3. Slice tenants request resources from slice provider

Each slice tenant becomes aware of its actual resource requirement $d_s(t)$ for each scheduling window t as it approaches. Based on the quantity of auxiliary resources offered $o_s(t)$ to it in that scheduling window and the probability that those resources will be available, each slice tenant then decides how many resources $r_s(t)$ it will request from the slice provider. Since a slice tenant is only interested in resources which have accompanying availability information, the number of resources $r_s(t)$ that it requests can be assumed to be bound by

$$0 \leq r_s(t) \leq a_s + o_s(t),\ \forall\ t,\ \ s \in S. \tag{7.2}$$

Hence, the resource request $r_s(t)$ from a slice tenant to a slice provider is a function of $d_s(t)$, $o_s(t)$ and g.

4. Slice provider schedules resources among tenants

After receiving the resource requests $r_s(t)$ from all slices in a given scheduling window t, the slice provider then schedules the K available resources among the N slices, with each slice tenant s receiving $u_s(t)$ resources. The quantity of resources scheduled for each slice tenant is bound by

$$0 \leq u_s(t) \leq r_s(t),\ \forall\ t,\ \ s \in S. \tag{7.3}$$

Due to the overbooking of auxiliary resources by the slice provider, in any given time slot the slice tenant may not get the full quota of resources that it requested, as the slice provider is bound by the condition

$$\sum_{s \in S} u_s(t) \leq K,\ \forall\ t. \tag{7.4}$$

Hence, the actual quantity of resources scheduled for a slice tenant is bound by the minimum between the resources that it requested $r_s(t)$ and the number of resources that the slice provider can schedule for it, given the constraints imposed by the probability of availability that it must satisfy for all slices.

While the slice provider is not obliged to offer any auxiliary resources, it must satisfy the accompanying probability g for any offers it does make. The guarantee for each slice s is evaluated using a sliding observation window W_s consisting of the p previous scheduling windows in which $o_s > 0$. The fraction g_s of time slots for which the resource request of a slice s was fully satisfied is given by

$$g_s = \frac{1}{p} \sum_{w \in W_s} \mathbf{1}\Big(u_s(w) = r_s(w)\Big), \tag{7.5}$$

where $\mathbf{1}(x)$ denotes an indicator function, which equals 1 when x is true, and zero otherwise. The slice provider must ensure that

$$g \leq g_s,\ \forall\ s \in S. \tag{7.6}$$

The guarantee in equation (7.5) only applies to auxiliary resources, and hence only scheduling windows in which a slice received a non-zero offer of auxiliary resources are included in the sliding observation window W_s. Hence, equation (7.5) implies that the scheduled resources u_s for a slice should not be less than the number of resources r_s requested by the slice in a fraction g of the time that auxiliary resources have been offered ($o_s > 0$).

The service contract applies to every region over which the resource is reusable. For spectrum, this may be at a cell-level, while for processing power in a baseband unit this may be a city. It is possible to alter the amount of assured or auxiliary resources for a slice on a per-region basis. In this chapter, we will consider a single resource and a single region, as the procedures outlined can be repeated for every additional region and resource of concern.

We propose a pricing model whereby each slice tenant pays a fixed price to the slice provider in proportion to the number of assured resources included in its service contract, and then pays on a per-resource basis for the scheduled resources $u_s(t)$ that it actually received which, according to equation (7.3), is always less than or equal to the number of resources that it requested. The fixed price to reserve the assured resources can be thought of as a non-refundable deposit to ensure their availability if needed. While equation (7.1) ensures that assured resources are not overbooked, assured resources that are predicted to be unused in a given scheduling window will be taken into account when the slice provider applies overbooking to its offers of auxiliary resources.

7.2 Offering Auxiliary Resources

The process of offering auxiliary resources to slice tenants consists of 3 steps: first, the slice provider must determine how many resources are available to issue as auxiliary resources; second, these free resources must be distributed as auxiliary resources among the slice tenants as offers; third and finally, the slice provider must decide how to handle overbookings when $\sum_{s \in S} r_s > K$. Note that for the remainder of this section we drop the time index t, as this procedure repeats for each scheduling window.

7.2.1 Calculating the Number of Auxiliary Resources

Equation (7.5) expresses the guarantee provided on auxiliary resources, stating that a slice tenant should be scheduled a quantity of resources u_s not less than its request r_s for at least a fraction g of the time that auxiliary resources have been offered to it. Hence, given a slice's predicted underlying demand for resources $\widehat{D}_s$, we are interested in its predicted request for resources, denoted $\widehat{R}_s$. Applying the relation expressed in equation (7.2), we can express $\widehat{R}_s$ in terms of $\widehat{D}_s$ as

$$\widehat{R}_s = \min(a_s + o_s, \widehat{D}_s). \tag{7.7}$$

While the random variable $\widehat{D}_s$ captures the underlying resource demand forecasted for a slice s, $\widehat{R}_s$ maps to the predicted demand for assured and auxiliary resources requested by a slice. The PDF of $\widehat{R}_s$ is a piece-wise function given by

$$f_{\widehat{R}_s}(\widehat{r}_s) = \begin{cases} f_{\widehat{D}_s}(\widehat{r}_s), & \widehat{r}_s < a_s + o_s \\ \delta(\widehat{r}_s) \int_{a_s+o_s}^{\infty} f_{\widehat{D}_s}(\widehat{r}_s) \, d\widehat{r}_s, & \widehat{r}_s = a_s + o_s \\ 0. & \widehat{r}_s > a_s + o_s \end{cases} \tag{7.8}$$

where $\delta(x)$ represents the Dirac delta function. Since r_s is upper bounded by $a_s + o_s$, the PDF for the predicted demand for resources requested by a slice tenant is 0 above this number. If the slice's underlying demand exceeds this limit, then it will require its full quota of assured and auxiliary resources, hence the spike at exactly $a_s + o_s$.

The joint PDF $f_{\widehat{R}_1 \ldots \widehat{R}_N}(\widehat{r}_1 \ldots \widehat{r}_N)$ for the forecasted number of resources requested by all slices is also a piece-wise defined function. In each dimension, representing a slice, the probability of requesting more resources than the sum of assured and auxiliary resources being offered is 0. The PDF will also contain spikes at the limit $a_s + o_s$ for each slice in a similar manner to equation (7.8), calculated by integrating over the relevant axes from $a_s + o_s$ to infinity.

A sufficient, but not necessary, condition for the slice provider to ensure that the guarantee g is satisfied for all slices is to ensure that the system is not over-subscribed in a

Algorithm 2 Calculate the number of auxiliary resources that are free a fraction g of the time.

```
**GIVEN** $f_{\widehat{D}_1 \ldots \widehat{D}_N}(\widehat{d}_1 \ldots \widehat{d}_N), [a_1, a_2, \ldots, a_N]$
**SET** $K_{\text{aux}} = 0$
**DISTRIBUTE** $K_{\text{aux}}$ among slices to determine $[o_1, o_2, \ldots, o_N]$
**CALCULATE** $f_{\widehat{R}_1 \ldots \widehat{R}_N}(\widehat{r}_1 \ldots \widehat{r}_N)$ **using** $f_{\widehat{D}_1 \ldots \widehat{D}_N}(\widehat{d}_1 \ldots \widehat{d}_N), [a_1, \ldots, a_N], [o_1, \ldots, o_N]$
**CALCULATE** $f_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}})$ **from** $f_{\widehat{R}_1 \ldots \widehat{R}_N}(\widehat{r}_1 \ldots \widehat{r}_N)$
**WHILE** $\int_0^K f_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}}) d\widehat{r}_{\text{sum}} \geq g :$
    **SET** $K_{\text{aux}} = K_{\text{aux}} + 1$
    **CALCULATE** $[o_1, o_2, \ldots, o_N]$
    **CALCULATE** $f_{\widehat{R}_1 \ldots \widehat{R}_N}(\widehat{r}_1 \ldots \widehat{r}_N)$
    **CALCULATE** $f_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}})$
**SET** $K_{\text{aux}} = K_{\text{aux}} - 1$
**CALCULATE** $[o_1, o_2, \ldots, o_N]$
```

fraction g of all scheduling windows. We can express this condition as

$$\Pr\left(\sum_{s \in S} \widehat{R}_s \leq K\right) \geq g, \tag{7.9}$$

where $\Pr\left(\sum_{s \in S} \widehat{R}_s \leq K\right)$ is the probability that the system is not expected to be oversubscribed. To express this in other words, the predicted sum demand for requested resources should be less than the total number of resources available for at least a fraction g of the time.

Let $\widehat{R}_{\text{sum}} = \sum_{s \in S} \widehat{R}_s$ denote the random variable expressing the predicted sum total of resources requested by all slices. The PDF of $\widehat{R}_{\text{sum}}$ can be obtained by integrating over the joint PDF $f_{\widehat{R}_1 \ldots \widehat{R}_N}(\widehat{r}_1 \ldots \widehat{r}_N)$. We can now rewrite equation (7.9) in terms of $f_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}})$ as

$$\int_0^K f_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}}) d\widehat{r}_{\text{sum}} \geq g. \tag{7.10}$$

The shape of the PDF of $\widehat{R}_{\text{sum}}$ is derived from the joint distribution $f_{\widehat{R}_1 \ldots \widehat{R}_N}(\widehat{r}_1 \ldots \widehat{r}_N)$, which is a piece-wise defined function dependent on the value of o_s for each slice s. Hence, equation (7.10) is dependent on both the number of auxiliary resources being offered, and the specific way that they are distributed among slices. We can numerically evaluate equation (7.10) for a specific set of values $(o_1, o_2, \ldots, o_N)$ and determine whether it satisfies the guarantee g. Starting with zero, the number of auxiliary resources can be incrementally increased until the guarantee can no longer be satisfied according to equation (7.10). Algorithm 2 outlines the procedure for determining how many auxiliary resources can be offered with a guarantee g.

Algorithm 2 purposely does not outline how to distribute auxiliary resources among slices; in the next section we will discuss how to approach this. Any distribution technique can be adopted and the algorithm will determine the number of auxiliary resources that can be offered within the terms of the guarantee g.

7.2.2 Distributing the Auxiliary Resource Offers

After determining the total number of auxiliary resources K_{aux} that can be offered to slices, the slice provider must then decide how to distribute them among slices. We consider three different approaches:

Equal Share

Dividing the auxiliary resources equally among slices is one of the simplest approaches, but may result in auxiliary resources being offered to slices which do not need them. Individual resources are indivisible; any remainder ($K_{\text{remain}} = K_{\text{aux}} \mod N$) is distributed to K_{remain} slices chosen at random.

Proportional Share

Offering a slice a quantity of auxiliary resources proportional to its predicted resource demand ensures that slices with high demand get more auxiliary resources. This approach is indiscriminate in terms of how many assured resources each slice has, and so auxiliary resources may still be scheduled to slices which do not require them. The number of auxiliary resources offered to each slice is given by

$$o_s = \left\lfloor \frac{\bar{d}_s}{\sum_{s \in S} \bar{d}_s} K_{\text{aux}} \right\rfloor . \tag{7.11}$$

Any remainder ($K_{\text{remain}} = K_{\text{aux}} - \sum_{s \in S} o_s$) resulting from the flooring operation is distributed to K_{remain} slices chosen at random.

Revenue Maximisation

The slice provider's aim is to increase its revenue. Since it gets paid for each resource u_s scheduled to a slice, and u_s is upper bounded by r_s according to equation (7.3), the slice provider is motivated to distribute the resource offers in such a way as to maximise the probability that the auxiliary resources will be requested.

We can express the distribution of auxiliary resources with the aim to maximise the slice provider's revenue as an optimisation problem:

$$\text{OP1}: \max_{[o_1, o_2, \ldots, o_N]} \mathbb{E}[\widehat{R}_{\text{sum}}], \tag{7.12}$$

subject to

$$\sum_{s \in S} o_s \leq K_{\text{aux}}, \tag{7.12a}$$

$$o_s \geq 0, \quad \forall s \in S. \tag{7.12b}$$

The expected value of $\widehat{R}_{\text{sum}}$ can be expanded as

$$\mathbb{E}[\widehat{R}_{\text{sum}}] = \sum_{s \in S} \mathbb{E}[\widehat{R}_s].$$

Referring to equation (7.8), for large enough values of $a_s + o_s$, the maximum $\mathbb{E}[\widehat{R}_{\text{sum}}]$ occurs when $\mathbb{E}[\widehat{D}_s]$ is maximised. This corresponds to a scenario when $a_s + o_s$ is greater than the upper bound on the support of the distribution of the estimated demand from slice s, i.e. when $\int_0^{o_s + a_s} f_{\widehat{D}_s}(\widehat{d}_s) d\widehat{d}_s = 1$.

Hence, to maximise revenue, K_{aux} should be distributed among $[o_1, o_2, \ldots, o_N]$ in such a way that $\int_0^{o_1+a_1} \ldots \int_0^{o_N+a_N} f_{\widehat{D}_1 \ldots \widehat{D}_N}(\widehat{d}_1 \ldots \widehat{d}_N) d\widehat{d}_1 \ldots d\widehat{d}_N$ is as close to 1 as possible. We can rewrite optimisation problem OP1 in terms the random variables representing the underlying estimated demand for each slice using the joint Cumulative Distribution Function (CDF) $F_{\widehat{D}_1 \ldots \widehat{D}_N}(\widehat{d}_1 \ldots \widehat{d}_N)$ as

$$\text{OP2}: \max_{[o_1, o_2, \ldots, o_N]} F_{\widehat{D}_1 \ldots \widehat{D}_N}(o_1 + a_1, \ldots, o_N + a_N), \tag{7.13}$$

subject to

$$\sum_{s \in S} o_s \leq K_{\text{aux}}, \tag{7.13a}$$

$$o_s \geq 0, \quad \forall s \in S. \tag{7.13b}$$

$F_{\widehat{D}_1 \ldots \widehat{D}_N}(o_1 + a_1, \ldots, o_N + a_N)$ is a monotonically increasing function, and incrementing the offer made to any slice at any stage in the algorithm cannot decrease the objective function in OP2 (intuitively, it is not possible to reduce profit by offering any additional auxiliary resources to a slice). Since it cannot be guaranteed that this joint CDF will be a convex function, we adopt steepest ascent hill climbing, in which incrementing o_s for all $s \in S$ is considered and the slice which results in the largest increase in $F_{\widehat{D}_1 \ldots \widehat{D}_N}(o_1 + a_1, \ldots, o_N + a_N)$ is chosen. The goal is to climb as high as possible in K_{aux} steps.

Hill-climbing cannot guarantee an optimal solution, but modifications such as stochastic hill-climbing can be adopted to increase the probability of obtaining an optimal solution. We note that random-restart hill climbing, where a different starting point is chosen each time, is not applicable as the starting point is fixed at the point $(a_1, \ldots, a_N)$.

7.2.3 Managing Overbookings

After making its offers $[o_1, ..., o_N]$ of auxiliary resources to slices, the slice provider receives the requests for resources $[r_1, ..., r_N]$ from the slices. If its resources are over-subscribed (i.e. $\sum_{s \in S} r_s > K$), the slice provider aims to minimise the negative effect on any individual slice by scheduling the available additional resources $\left(K - \sum_{s \in S} \min(a_s, r_s)\right)$ across all slices that have requested auxiliary resources, in proportion to their requests. Any slice tenant who requested $r_s \leq a_s$ will be scheduled its full request $u_s = r_s$.

The slice provider first identifies the set S_{R} of slices that have requested auxiliary resources:

$$S_{\mathrm{R}} = \{s \in S \mid r_s > a_s\}. \tag{7.14}$$

The scheduled resources $u_{s'}$ for each slice $s' \in S_{\mathrm{R}}$ is calculated as a fraction of the auxiliary resources requested by that slice:

$$u_{s'} = \left[\frac{r_{s'} - a_{s'}}{\sum_{s \in S_{\mathrm{R}}}(r_s - a_s)}\right] \Big(K - \sum_{s \in S} \min(a_s, r_s)\Big) + a_{s'},\ \forall\ s' \in S_{\mathrm{R}}. \tag{7.15}$$

To ensure that it is satisfying the guarantee g on auxiliary resources, the slice provider must track the current status of its guarantee to each slice. To achieve this, it evaluates the guarantee g_s for each slice s over a sliding observation window W_s, as defined in equation (7.5). If $g_s < g$ for a slice s consistently, it indicates that the slice provider's forecast $\widehat{D}_s$ is not accurate for that slice. In this case, the slice provider should attempt to improve its demand forecasting technique (or, alternatively, modify the guarantee g that it will offer in the future).

7.3 Analytical Results

For concreteness of results, in this section we analyse the case of two slices (indexed 1 and 2) with independent, identically distributed predicted resource demand ($\widehat{D}_1$ and $\widehat{D}_2$), characterised by uniform distributions with upper and lower supports given as b_U and b_L, respectively:

$$f_{\widehat{D}_1}(\widehat{d}_1) = f_{\widehat{D}_2}(\widehat{d}_2) = \begin{cases} \frac{1}{(b_U - b_L)}, & b_L \leq \widehat{d} \leq b_U \\ 0, & \text{otherwise} \end{cases}. \tag{7.16}$$

The assured resources for the two slices are denoted as a_1 and a_2, where $b_L \leq a_2 \leq a_1 \leq b_U$. The middle part of the inequality is without loss of generality, and the provider can safely assume that the assured resources requested by the slices lie within the support of their demand distribution. As in the previous section, the auxiliary resources offered to the slices are denoted by $o_s \geq 0$. We denote the total number of assured and auxiliary resources offered to a slice as $c_s = a_s + o_s$.

Applying equation (7.7), we obtain the predicted resource request for each slice ($\widehat{R}_1$ and $\widehat{R}_2$), with PDF characterised according to equation (7.8). The PDF of $\widehat{R}_1$ is given as

$$f_{\widehat{R}_1}(\widehat{r}_1) = \begin{cases} \frac{1}{(b_U - b_L)}, & b_L \leq \widehat{r}_1 \leq c_1 \\ \delta(c_1)\frac{b_U - c_1}{(b_U - b_L)}, & \widehat{r}_1 = c_1 \\ 0, & \text{otherwise} \end{cases}, \tag{7.17}$$

with $f_{\widehat{R}_2}$ defined similarly.

Let $\widehat{R}_{\text{sum}} = \widehat{R}_1 + \widehat{R}_2$ denote the random variable expressing the predicted sum total of resources requested by both slices. Since the two slices have independent predicted resource demand, the PDF of $\widehat{R}_{\text{sum}}$ can be obtained through the convolution of the PDFs of $\widehat{R}_1$ and $\widehat{R}_2$, given as $f_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}}) = f_{\widehat{R}_1}(\widehat{r}_1) \circledast f_{\widehat{R}_2}(\widehat{r}_2)$, which expands to

$$f_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}}) = \begin{cases} \widehat{\alpha}_f(\widehat{r}_{\text{sum}}), & 2b_L \leq \widehat{r}_{\text{sum}} < b_L + c_2 \\ \widehat{\beta}_f(\widehat{r}_{\text{sum}}), & b_L + c_2 \leq \widehat{r}_{\text{sum}} < b_L + c_1 \\ \widehat{\gamma}_f(\widehat{r}_{\text{sum}}), & b_L + c_1 \leq \widehat{r}_{\text{sum}} < c_1 + c_2 \\ \widehat{\mathit{\Gamma}}_f \delta(c_1 + c_2), & \widehat{r}_{\text{sum}} = c_1 + c_2 \end{cases}, \tag{7.18}$$

where $\widehat{\alpha}_f(\widehat{r}_{\text{sum}}), \widehat{\beta}_f(\widehat{r}_{\text{sum}}), \widehat{\gamma}_f(\widehat{r}_{\text{sum}}), \widehat{\mathit{\Gamma}}_f$ are defined below.

$$\widehat{\alpha}_f(\widehat{r}_{\text{sum}}) = \int_{2b_L}^{\widehat{r}_{\text{sum}}} \frac{1}{(b_U - b_L)^2} dy = \frac{\widehat{r}_{\text{sum}} - 2b_L}{(b_U - b_L)^2}.$$

$$\widehat{\beta}_f(\widehat{r}_{\text{sum}}) = \int_{b_L}^{c_2} \frac{1}{(b_U - b_L)^2} dy + \frac{(b_U - c_2)}{(b_U - b_L)^2} = \frac{1}{(b_U - b_L)}.$$

$$\widehat{\gamma}_f(\widehat{r}_{\text{sum}}) = \int_{\widehat{r}_{\text{sum}}}^{c_1 + c_2} \frac{1}{(b_U - b_L)^2} dy + \frac{(b_U - c_1)}{(b_U - b_L)^2} + \frac{(b_U - c_2)}{(b_U - b_L)^2} = \frac{(2b_U - \widehat{r}_{\text{sum}})}{(b_U - b_L)^2}.$$

$$\widehat{\mathit{\Gamma}}_f = \frac{(b_U - c_1)(b_U - c_2)}{(b_U - b_L)^2}.$$

The CDF of the predicted sum total of resources requested by both slices is given by

$$F_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}}) = \begin{cases} 0, & \widehat{r}_{\text{sum}} < 2b_L \\ \widehat{\alpha}_F(\widehat{r}_{\text{sum}}), & 2b_L \leq \widehat{r}_{\text{sum}} < b_L + c_2 \\ \widehat{\beta}_F(\widehat{r}_{\text{sum}}), & b_L + c_2 \leq \widehat{r}_{\text{sum}} < b_L + c_1 \\ \widehat{\gamma}_F(\widehat{r}_{\text{sum}}), & b_L + c_1 \leq \widehat{r}_{\text{sum}} < c_1 + c_2 \\ 1, & \widehat{r}_{\text{sum}} \geq c_1 + c_2 \end{cases}, \tag{7.19}$$

where $\widehat{\alpha}_F(\widehat{r}_{\text{sum}}), \widehat{\beta}_F(\widehat{r}_{\text{sum}}), \widehat{\gamma}_F(\widehat{r}_{\text{sum}})$ are defined below.

$$\widehat{\alpha}_F(\widehat{r}_{\text{sum}}) = \int_{2b_L}^{\widehat{r}_{\text{sum}}} \widehat{\alpha}_f \; d\widehat{r}_{\text{sum}} = \frac{(\widehat{r}_{\text{sum}} - 2b_L)^2}{2(b_U - b_L)^2}.$$

$$\widehat{\beta}_F(\widehat{r}_{\text{sum}}) = \int_{b_L + c_2}^{\widehat{r}_{\text{sum}}} \widehat{\beta}_f \; d\widehat{r}_{\text{sum}} + \widehat{\alpha}_F = \frac{(\widehat{r}_{\text{sum}} - b_L - c_2)(b_U - b_L)}{(b_U - b_L)^2} + \widehat{\alpha}_F(b_L + c_2).$$

$$\widehat{\gamma}_F(\widehat{r}_{\text{sum}}) = \int_{b_L + c_1}^{\widehat{r}_{\text{sum}}} \widehat{\gamma}_f \; d\widehat{r}_{\text{sum}} + \widehat{\beta}_F = \frac{2b_U\widehat{r}_{\text{sum}} - 0.5\widehat{r}_{\text{sum}}^2 - (2b_U)(b_L + c_1) + 0.5(b_L + c_1)^2}{(b_U - b_L)^2} + \widehat{\beta}_F(b_L + c_1).$$

Algorithm 2 incrementally increases the number of auxiliary resources offered across all slices while ensuring that the condition specified in equation (7.10) is satisfied. We can rewrite equation (7.10) in terms of equation (7.19) as

$$F_{\widehat{R}_{\text{sum}}}(K) \geq g. \tag{7.20}$$

To ensure that our analysis is performed within a valid range specified in equation (7.19) at all times, we assume that the total number of available resources K is at least equal to the expected value of the sum of the predicted resource demand for both slices, i.e. $K \geq \bar{d}_1 + \bar{d}_2$ where $\bar{d}_1 + \bar{d}_2 = b_U + b_L$ (since the expected value for a uniform distribution is given as $\frac{b_U + b_L}{2}$). This ensures that K is always within the range of the part of $F_{\widehat{R}_{\text{sum}}}(\widehat{r}_{\text{sum}})$ expressed by $\widehat{\gamma}_F(\widehat{r}_{\text{sum}})$, as the lower limit of the range for $\widehat{\gamma}_F(\widehat{r}_{\text{sum}})$ never exceeds $b_U + b_L$, since the maximum value that c_1 can attain is b_U. The upper limit of the range of $\widehat{\gamma}_F(\widehat{r}_{\text{sum}})$ is exceeded only if $g = 1$, and it can be reasonably assumed that the slice provider will want to provide a guarantee less than 1.

We wish to calculate the overbooking fraction O given as

$$O = \frac{c_1 + c_2}{K}. \tag{7.21}$$

To achieve this, we first set $c_2 = c_1 - \Delta$ and solve the following equation for c_1:

$$\widehat{\gamma}_F(K) = g. \tag{7.22}$$

The total number of assured and auxiliary resources $(c_1 + c_2)$ can then be calculated as $2c_1 - \Delta$, giving

$$c_{\text{total}} = 2b_U - \sqrt{4(g-1)(b_U - b_L)^2 + 2(2b_U - K)^2 - \Delta^2}, \tag{7.23}$$

where $c_{\text{total}} = c_1 + c_2 = 2c_1 - \Delta$. The overbooking fraction is then given as

$$O = \frac{c_{\text{total}}}{K}. \tag{7.24}$$

From equation (7.23) above, we observe that the total number of assured and auxiliary resources will not exceed the maximum number of resources $(2b_U)$ required by the slices. We can extend this by noting that c_{total} is reduced if a higher guarantee is provided. We also note that c_{total} is reduced if the maximum number of resources that might be required by the slices is considerably greater than the total number of resources available (i.e. c_{total} decreases as $2b_U - K$ increases). Finally, and less intuitively, we note that c_{total} increases as the difference between c_1 and c_2 increases. This implies that the fairest option of offering equal numbers of auxiliary resources to both slices is not necessarily conducive to maximising the total number of auxiliary resources.

In the following subsections, we examine each of these relationships in more detail.

7.3.1 Guarantee

Intuitively, we would expect that a stricter guarantee g would correspond to fewer auxiliary resources. We can prove this by examining the partial derivative of the overbooking fraction O in equation (7.24) with respect to g and demonstrating that it is always negative. The partial derivative of O with respect to g is

$$\frac{\partial O}{\partial g} = \frac{-2(b_U - b_L)^2}{k\sqrt{4(g-1)(b_U - b_L)^2 + 2(2b_U - K)^2 - \Delta^2}}. \tag{7.25}$$

It is easy to see that $\frac{\partial O}{\partial g} \leq 0$ and that increasing the guarantee reduces the overbooking fraction.

To illustrate this graphically, we consider what happens when the number of assured and auxiliary resources offered to each slice is the same (i.e. $c_1 = c_2$), so that Δ is 0. We also set the total number of resources equal to the expected value of the sum of the predicted demand of both slices, such that $K = b_U + b_L$. Finally, we assume that $b_L = 0$. Making

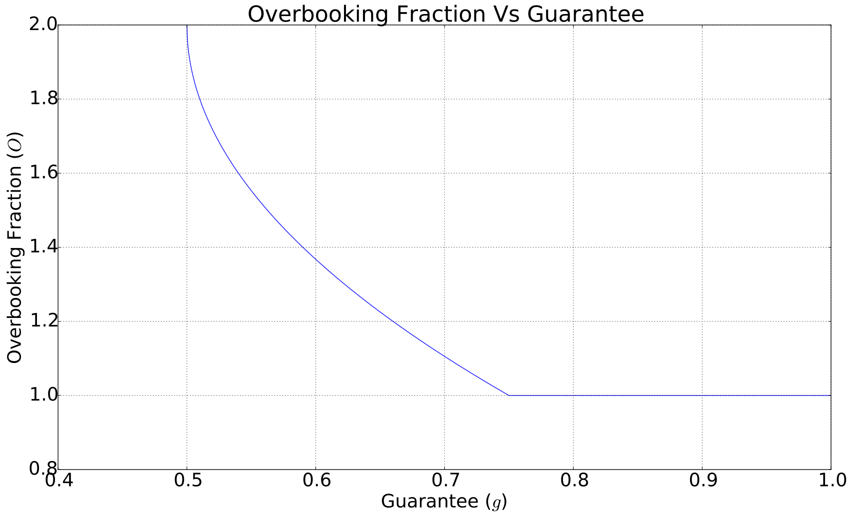


Figure 7.2: Overbooking fraction versus the guarantee provided on auxiliary resources. The overbooking fraction decreases as the guarantee becomes more strict. No overbooking occurs above a guarantee of 0.75.

these substitutions gives

$$c_{\text{total}} = 2b_U - \sqrt{2(2g-1)}b_U.$$

Fig. 7.2 shows the relationship between g and O as g is varied, and confirms that O is reduced as the guarantee is made more stringent. There are two bounds on the guarantee to note.

First, we observe that overbooking only occurs for guarantees less than an upper bound of 0.75. At least K resources can always be scheduled and the overbooking fraction therefore never falls below 1, as represented by the horizontal line in Fig. 7.2. The upper limit of $\widehat{\gamma}_F(\widehat{r}_{\text{sum}})$ in equation (7.19) is no longer satisfied at a guarantee above 0.75, as $c_1 + c_2 < K$. Hence, there is a discontinuity between $g = 0.75$ and $g = 1$ represented by $\widehat{\Gamma}_f$ in equation (7.18). Given that $c_1 = c_2$, $c_1 + c_2 = K$, and $K = b_U + b_L$, then we can confirm that $\widehat{\Gamma}_f$ is in fact 0.25.

At a guarantee of 0.5, then $c_1 = c_2 = b_U$, and the lower limit of $\widehat{\gamma}_F$ in equation (7.19) is reached. We note that the model does not permit $c_1 > b_U$ because providing a guarantee on resources in excess of the maximum possible resource demand for a slice is meaningless. Hence, in this specific case when $K = b_U$, the maximum overbooking fraction is 2, corresponding to $2b_U$ assured and auxiliary resources, which is the maximum predicted sum demand of the two slices.

The guarantee below which overbooking occurs depends on the relationship between K, b_U and b_L, and will be explored in more detail in the next subsection.

7.3.2 Total Available Resources

In Section 7.3.1, we observed that there is an upper limit to the guarantee above which no overbooking can occur, and a lower limit at which overbooking is maximised. The number of available resources K, and its relationship to b_L and b_U, influences these limits.

The upper limit g_u above which overbooking does not occur is given by $1-\widehat{\mathit{\Gamma}}_f$, calculated before any auxiliary resources have been granted to slices (i.e. when $c_s = a_s$ and $o_s = 0$). Again assuming that $\Delta = 0$, we let the sum of assured resources $a_1 + a_2$ equal K, giving

$$g_u = 1 - \frac{(b_U - \frac{K}{2})^2}{(b_U - b_L)^2}. \tag{7.26}$$

The partial derivative of g_u with respect to K is

$$\frac{\partial g_u}{\partial K} = \frac{b_U - 0.5K}{(b_U - b_L)^2}.$$

Since $a_s \leq b_U$ and $K = 2a_s$, $\frac{\partial g_u}{\partial K}$ is always positive and increasing K increases the maximum limit on the guarantee.

The lower limit g_l occurs when the square root in equation (7.23) is equal to zero:

$$4(g-1)(b_U - b_L)^2 + 2(2b_U - K)^2 - \Delta^2 = 0.$$

As before, we assume that $\Delta = 0$ and $a_1 + a_2$ equals K, giving

$$g_l = 1 - 2\frac{(b_U - \frac{K}{2})^2}{(b_U - b_L)^2}. \tag{7.27}$$

Using an approach similar to that used for g_u, it can be shown that increasing K increases the lower limit on the guarantee.

We demonstrate these results graphically by varying a_s from $\frac{b_U+b_L}{2}$ to b_U by setting K equal to $2(b_L + x(b_U - b_L))$ where x is a value in the range $[0.5, 1]$. Making this substitution, we obtain

$$g_u = x(2-x),$$

and

$$g_l = 1 - 2(1-x)^2.$$

Fig. 7.3 shows the upper and lower limits on the guarantee as K is varied from $b_L + b_U$ to $2b_U$. The shaded region represents the range of guarantees that can be provided. No overbooking is possible above this range, while the overbooking fraction cannot be improved by offering guarantees below this range.

As K approaches $2b_U$, the number of actual resources available is similar to the maximum predicted sum demand of the slices and it is possible to offer strict guarantees. The

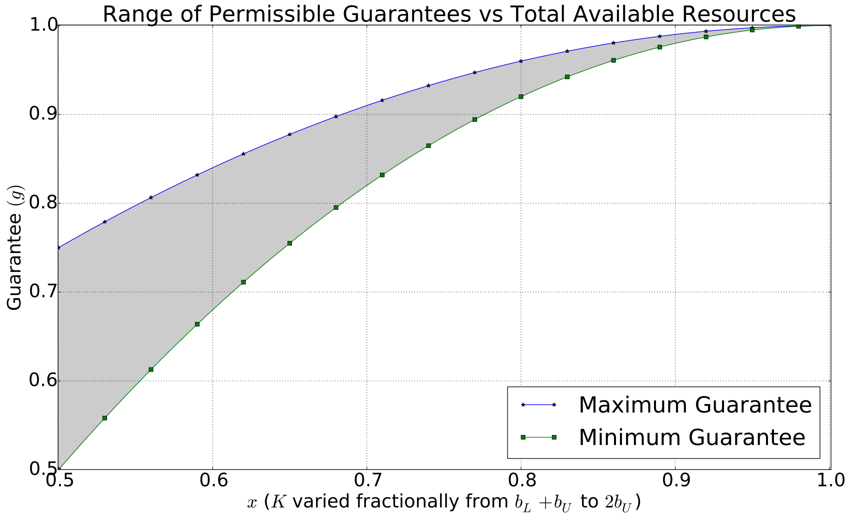


Figure 7.3: Range of permissible guarantees versus the total number of available resources. The range of guarantees that can be offered varies in accordance with the number of available resources.

range is quite small, with little additional benefit obtained from offering significantly lower guarantees.

7.3.3 Uncertainty in Demand Prediction

We use the standard deviation of $\widehat{D}_1$ and $\widehat{D}_2$ as our measure of uncertainty, which is defined as $\sqrt{\frac{1}{12}(b_U - b_L)^2}$. Hence, for example, a standard deviation of 1 implies $b_U - b_L = \sqrt{12}$.

In the previous section, we observed that when $b_U - b_L$ is constant and K is varied, the upper and lower limits on the guarantee vary. It is reasonable to expect that when K is fixed and $b_U - b_L$ is varied, the limits placed on the guarantee will also be affected. To isolate the effects of $b_U - b_L$, we must ensure that the expected value of the predicted demand $\bar{d}_s = 0.5(b_U + b_L)$ remains constant by varying b_U and b_L equally on either side of $\bar{d}_s$. We also assume that $K \leq 2b_U$ (otherwise overbooking is not required).

Taking the expression for the upper limit g_u defined in equation (7.26), we wish to find the partial derivative of g_u with respect to $b_U - b_L$. To achieve this, we make the substitution $b_U = \frac{1}{2}((b_U + b_L) + (b_U - b_L))$, noting that $b_U + b_L$ is in fact a constant. This gives

$$\frac{\partial g_u}{\partial(b_U - b_L)} = \frac{(b_U b_L - 0.5 b_L K + b_U^2 - 1.5 b_U K + 0.5K^2)}{(b_U - b_L)^2}.$$

Given that $b_U \geq b_L$ and $(b_U + b_L) \leq K \leq 2b_U$, we obtain

$$\frac{\partial g_u}{\partial(b_U - b_L)} < 0.$$

Hence, if the total number of resources K is greater than the sum of the expected value of predicted demand ($\bar{d}_1 + \bar{d}_2 = b_U + b_L$), the upper limit on the guarantee g_u will decrease

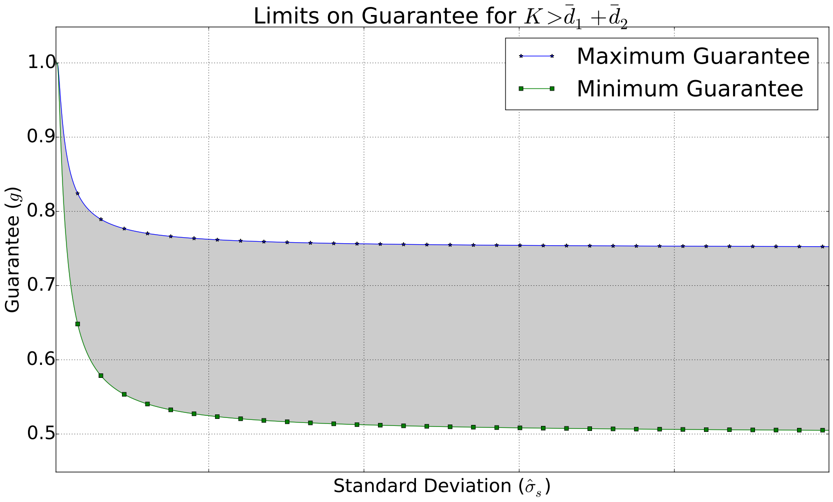


Figure 7.4: Limits on the range of guarantees that can be provided versus the standard deviation of predicted resource demand for slices. When $K > b_U + b_L$, the upper and lower limits of g decrease with respect to increasing $\widehat{\sigma}_s$.

with $b_U - b_L$.

The approach above can be repeated with equation (7.27) to prove that the same holds true for g_l. Fig. 7.4 gives an indication of the shape of the curves for g_u and g_l.

We now examine what happens to g_l and g_u as $b_U - b_L$ grows very large. Again, using the expression for the upper limit g_u defined in equation (7.26), we make the substitution $b_U = \frac{1}{2}((\bar{d}_1 + \bar{d}_2) + (b_U - b_L))$ giving

$$g_u = 1 - \frac{1}{4}\frac{\Big((\bar{d}_1 + \bar{d}_2) + (b_U - b_L) - K\Big)^2}{(b_U - b_L)^2}.$$

The upper limit on the guarantee g_u converges on a fixed value as $b_U - b_L$ grows large, given by

$$\lim_{(b_U - b_L)\to\infty} g_u = 0.75.$$

It can be shown in similar fashion that for g_l, the limit is given as

$$\lim_{(b_U - b_L)\to\infty} g_l = 0.5.$$

Hence, both g_u and g_l converge to $0.75,\ 0.5$, respectively, as $b_U - b_L$ grows large, irrespective of the value of K. This behaviour is evident in Fig. 7.4.

Having examined how the uncertainty in the forecast affects the maximum and minimum bounds on the guarantee that can be provided, we turn our attention to its influence on the overbooking fraction O. Equation (7.23) depends on several variables, both controllable (g and K) and uncontrollable (b_U and b_L). Even though g represents a design choice, the limits

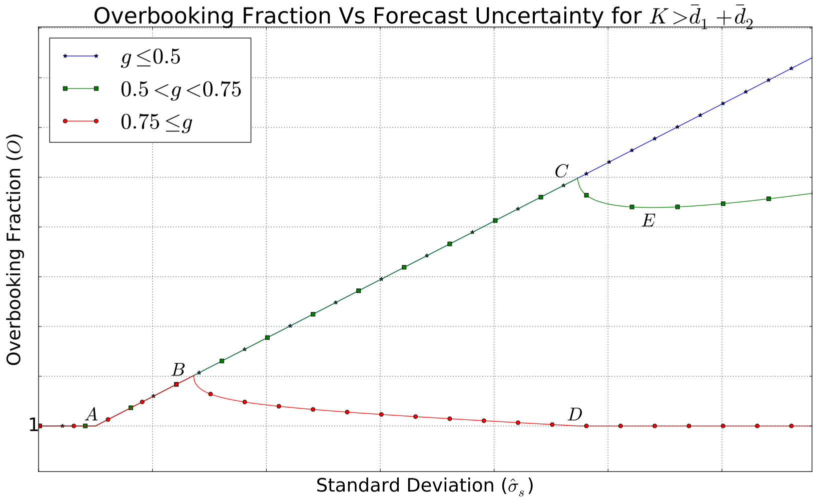


Figure 7.5: Overbooking fraction versus resource demand forecast uncertainty. The relationship between the overbooking fraction and $\widehat{\sigma}_s$ varies according to the value of g when $K > b_U + b_L$.

of the range of possible values is dependent on all other variables as evident in equations (7.26) and (7.27). Furthermore, K is bound by b_U and b_L such that $b_L + b_U \leq K \leq 2b_U$. Finally, to isolate the effects of $b_U - b_L$ alone, we must ensure that the expected value of the predicted demand $\bar{d}_s = 0.5(b_U + b_L)$ remains constant.

Hence, the influence of $b_U - b_L$ on O is complicated, and depends on the interaction between the four variables listed above. In total, there are 3 cases to consider.

Case 1: $K > b_L + b_U, \ g \leq 0.5$

When $K > b_L + b_U$, the lower limit on the guarantee decreases as $b_U - b_L$ increases and converges to a minimum value of 0.5. Hence, it is not possible to give a guarantee less than 0.5 when $K > b_L + b_U$. Instead, the minimum value of the range $[g_u, g_l]$ can be provided. In this case, the square root in equation (7.24) is zero and the overbooking fraction O increases linearly with $b_U - b_L$ according to $O = 2b_U/K$. This behaviour is illustrated by the blue line in Fig. 7.5. Note that no overbooking occurs until $K < 2b_U$, which is marked as point A in Fig. 7.5.

Case 2: $K > b_L + b_U, \ 0.5 < g < 0.75$

Initially in this case, shown as the green line in Fig. 7.5, g will be less than g_l and O will increase according to $O = 2b_U/K$. The behaviour then changes when $g = g_l$ which occurs at

$$b_U - b_L = \frac{\Big(K - (b_U + b_L)\Big)\Big(1 + \sqrt{2(1-g)}\Big)}{2g - 1}, \tag{7.28}$$

which is labelled as C in Fig. 7.5. Above this point, the nature of the relationship is dictated by equation (7.24), as g will never exceed the minimum value of the upper guarantee of 0.75. By examining the partial derivative of equation (7.24) with respect to $b_U - b_L$, we observe that O decreases between the value of $b_U - b_L$ given in equation (7.28) and the value of $b_U - b_L$ given by

$$b_U - b_L = \frac{\Big(K - (b_U + b_L)\Big)\left(1 + \big(2\sqrt{2} - \sqrt{2}\big)\Big(\sqrt{\tfrac{g-1}{(2g-1)(4g-3)}}\Big)\right)}{2g - 1}, \quad (7.29)$$

which is labelled as E in Fig. 7.5. Above this value of $b_U - b_L$, O begins to increase again.

Case 3: $K > b_L + b_U,\ g \geq 0.75$

When $K > b_L + b_U$, the upper limit on the guarantee decreases as $b_U - b_L$ increases and converges on the value 0.75. Initially, as per the previous case, O will increase according to $O = 2b_U/K$ until $g = g_l$. This occurs at the value of $b_U - b_L$ given by equation (7.28), and is marked as point B in Fig. 7.5. After this point, the value of O will decrease until $g = g_U$ which occurs at

$$b_U - b_L = \frac{\Big(K - (b_U + b_L)\Big)\Big(1 + 2\sqrt{(1-g)}\Big)}{4g - 3}, \quad (7.30)$$

which is labelled as D in Fig. 7.5. Above this value of $b_U - b_L$, g is above the maximum value and no overbooking occurs ($O = 1$).

7.3.4 Difference between c_1 and c_2

Examining equation (7.23), we note that c_{total} increases as the difference between c_1 and c_2, denoted Δ, increases. This implies that offering equal numbers of auxiliary resources to both slices is not the best approach to maximise the total number of auxiliary resources.

Taking equation (7.23), we isolate the effect of Δ on c_{total} by setting the total number of resources equal to the expected value of the predicted demand of both slices such that $K = b_U + b_L$. We also assume that $b_L = 0$. We express Δ in terms of b_U as $\Delta = xb_U$, where x is a value in the range $[0, 1]$. Making these substitutions gives

$$c_{\text{total}} = b_U\left(2 - \sqrt{2(2g-1) - x^2}\right). \quad (7.31)$$

Choosing a value for g is more difficult, as the upper and lower limits for g depend on the value of x. The lower limit is calculated as $2(2g_l - 1) - x^2 = 0$, giving

$$g_l = \frac{2 + x^2}{4}.$$

Depending on the value of x, g_l takes on a value in the range $[0.5, 0.75]$.

Similar to Section 7.3.2, the upper limit on the guarantee g_u is calculated as $(1 - \widehat{\Gamma}_f)$ before any auxiliary resources have been granted to slices (i.e. when $c_s = a_s$ and $o_s = 0$). We also simplify the expression by assuming $b_L = 0$, giving

$$\widehat{\Gamma}_f = \frac{(b_U - a_1)(b_U - a_2)}{(b_U)^2}. \tag{7.32}$$

As before, we make a series of simplifications to isolate the effect of Δ on g_u. Again, we set $K = b_U + b_L = b_U$, and $\Delta = xb_U$. Noting that $a_1 + a_2 = K$, we can express a_1 and a_2 as

$$\begin{aligned} a_1 &= \frac{b_U(1-x)}{2}, \\ a_2 &= \frac{b_U(1+x)}{2}. \end{aligned}$$

Substituting this into equation (7.32) gives

$$g_u = \frac{(3 + x^2)}{4}.$$

Depending on the value of x, g_u takes on a value in the range $[0.75, 1]$.

We observe that 0.75 is the only guarantee that can be assured to be valid regardless of the value of x. Setting $g = 0.75$ in equation (7.31), we obtain

$$c_{\text{total, 1}} = b_U(2 - \sqrt{1 - x^2}). \tag{7.33}$$

There is also an upper limit imposed on c_{total} due to the fact that $c_s < b_U$, given as

$$c_{\text{total, 2}} = b_U(2 - x). \tag{7.34}$$

Hence, the total number of assured and auxiliary resources is

$$c_{\text{total}} = \min(c_{\text{total, 1}}, c_{\text{total, 2}}). \tag{7.35}$$

Fig. 7.6 shows the effect of varying the difference Δ between c_1 and c_2 as a fraction of $b_U - b_L$ on the overbooking fraction.

At $\Delta = 0$, no overbooking occurs which is consistent with Fig. 7.2. Increasing Δ reduces the number of auxiliary resources offered to one of the slices. By reducing the number of auxiliary resources offered to one slice, a higher number of resources can be promised to the other slice at the same level of guarantee. Hence, c_{total} increases with Δ until it hits a limit imposed by the constraint that $c_s < b_U$. At this point, c_{total} begins to decreases as the difference between c_1 and c_2 increases. At $\Delta = 1$, no overbooking occurs as c_1 must equal 0 and c_2 must equal 1.

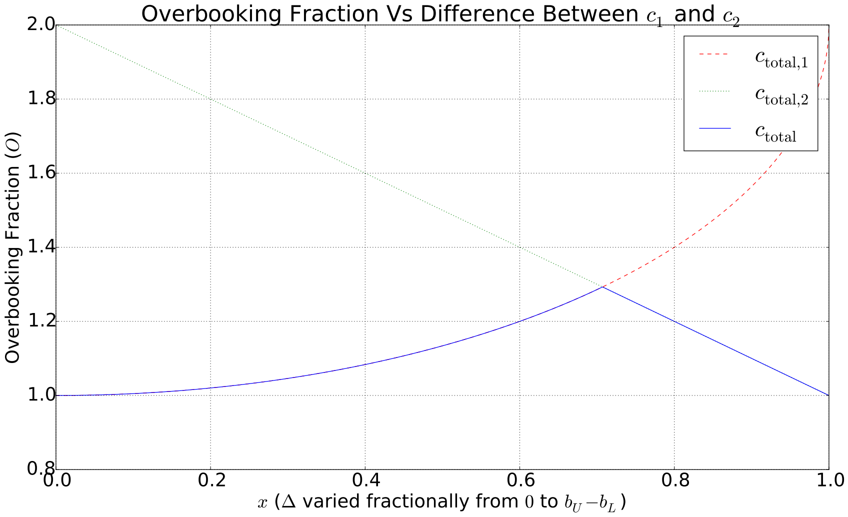


Figure 7.6: Overbooking fraction versus the difference between the quantity of resources scheduled to slices. The overbooking fraction increases at first as Δ increases before hitting a limit imposed by the constraint that $c_s < b_U$.

It is not immediately obvious why increasing Δ would increase the total number of auxiliary resources. On an intuitive level, to maximise the total number of auxiliary resources while satisfying the guarantee, the slice provider should offer each additional auxiliary resource to the slice which is least likely to use it. The result of this policy is that one slice is offered more resources than it will probably need, and the other slice is offered very few resources as a consequence, thereby increasing Δ.

Hence, while increasing the difference between c_1 and c_2 may be used to increase the total number of auxiliary resources, it does not necessarily maximise resource usage. The extra auxiliary resources gained may be less likely to be needed than if fewer were evenly spread across both slices. Section 7.2.2 outlines an approach to distribute auxiliary resources in such a manner as to maximise the revenue of the slice provider, which also results in resource utilisation being maximised.

7.4 Numerical Results

In Section 7.3, we adopted a simple model, consisting of two slices with identical uniformly distributed resource demand, to ensure that we could obtain tractable analytical expressions. This allowed us to demonstrate how the overbooking fraction depended on several key inputs to the system such as the guarantee and the uncertainty in the forecast.

We expand this analysis in this section to a more complex model consisting of four slices with heterogeneous resource demand based on normal distributions. In adopting this more complex model, we sacrifice analytical tractability and instead rely on numerical methods to evaluate performance of the system. Our focus remains on investigating how the overbooking

Table 7.1: Slice Parameters for Numerical Analysis

	Mean Demand	**Standard Deviation**	**Assured Resources**
Slice A	60 (high demand)	15 (high uncertainty)	72 (surplus)
Slice B	30 (low demand)	5 (low uncertainty)	24 (deficit)
Slice C	60 (high demand)	5 (low uncertainty)	48 (deficit)
Slice D	30 (low demand)	15 (high uncertainty)	36 (surplus)

fraction depends on the system inputs, and whether the conclusions drawn in Section 7.3 remain valid.

We consider four slices with resource demand characterised by Gaussian distributions. The base parameters for the slices are specified in Table 7.1. We consider a total of 200 resources in the system, with a targeted guarantee of 90%. Slice A has high resource demand, high uncertainty in the forecast, and is conservative in its policy of reserving assured resources by reserving more assured resources than its expected demand. Slice B, in contrast, has low demand and low uncertainty in the forecast, and reserves fewer assured resources than its expected demand. Slice C has high demand but low uncertainty, and again has fewer assured resources than its expected resource demand. Finally, Slice D has low demand but high uncertainty in the forecast, and has more assured resources than its expected demand.

The four slices chosen provide the basis to investigate how the proposed system of overbooking performs when slices have heterogeneous demand profiles. The base parameters set the initial conditions for each slice, from which we will isolate individual parameters and vary them to investigate their effect on the overbooking fraction.

In the context of the four slices which constitute the model for the numerical analysis, we define the overbooking fraction as the sum of the assured and offered resources for all four slices, divided by the total available resources in the system.

7.4.1 Guarantee and Uncertainty in the Forecast

We examine how the level of guarantee offered affects the overbooking fraction by varying the guarantee level from 80% to 100%. For each guarantee level, we also examine the influence of the uncertainty in the prediction on the overbooking fraction. To achieve this, we increment the standard deviation of each slice's demand from the base levels specified in Table 7.1.

Figure 7.7 illustrates that the overbooking fraction is maximised when the uncertainty is minimised and the guarantee is reduced. At a guarantee of 80% and no increase in standard deviation from the base levels specified in Table 7.1, the overbooking margin is 42%. This decreases sharply to 13.5% as the standard deviation is incremented by 4. Similarly, raising the guarantee to nearly 100% prevents any overbooking from occurring.

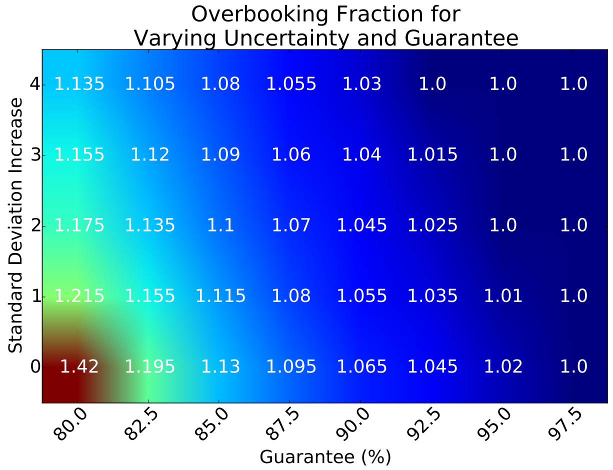


Figure 7.7: The heatmap shows that the overbooking fraction increases as the uncertainty in the forecast decreases and as the guarantee is lowered.

These results confirm the findings in Section 7.3.1 for the guarantee level. This is intuitive as more stringent guarantees require a more conservative approach to overbooking. The relationship between the overbooking fraction and uncertainty is more complex, with Section 7.3.3 demonstrating that the maximum and minimum guarantees that can be offered decrease as the standard deviation increases, while the influence of standard deviation on the overbooking fraction depended on the specific ranges of the two-slice uniform model. Figure 7.7 shows that the relationship between the overbooking fraction and uncertainty simplifies in the more complex model, with the overbooking fraction decreasing as the uncertainty increases.

The ranges on the level of guarantee that can be provided for the two-slice model, as illustrated in Figures 7.4 and 7.5, are quite pronounced due to the hard limits in the uniform distribution. Figure 7.7 shows that the upper limit on the guarantee beyond which overbooking no longer occurs is present in the more complex model, and that this limit is dependent on the uncertainty. This is an important point to note when deciding a guarantee to provide; greater uncertainty necessitates lower guarantees if overbooking is intended.

7.4.2 Total Available Resources Relative to Total Expected Demand

We investigate how the ratio between the total available resources relative to the total expected demand affects the overbooking fraction by varying the total available resources around the base value of 200. Figure 7.8 demonstrates that the overbooking fraction increases as the total available resources increase relative to total expected demand. Little or no overbooking occurs when the total available resources are less than the total expected demand.

Figure 7.8 confirms the trend highlighted in Section 7.3.2. If the expected sum demand is significantly higher than the total available resources, then resource utilisation will be

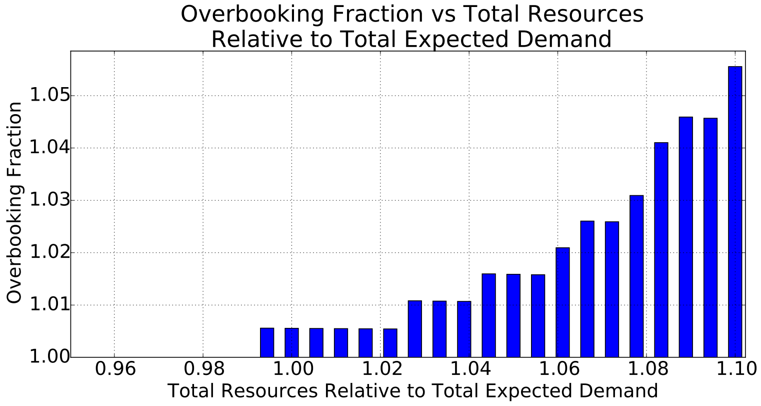


Figure 7.8: The overbooking fraction increases as the total available resources increases relative to total expected demand.

100% with high probability, rendering overbooking unnecessary. This is not recommended, however, as slices will also be under-provisioned with high probability due to a lack of resources, resulting in disgruntled customers (slices) for the slice provider.

7.5 Insights

In Section 7.3, we demonstrated that the effectiveness of overbooking depends on the interactions of several inputs, both controllable and uncontrollable. These results were obtained for an example scenario consisting of two slices with independent forecasted resource demand characterised by uniform distributions. In general, it is challenging to predict when conditions are favourable for overbooking as the demand distributions for the slices may take on different forms, and closed-form expressions for the overbooking fraction typically do not exist.

Hence, it is important to be aware of the influence that the inputs into the system, such as the guarantee and the uncertainty in the demand prediction, have on the overbooking factor. Below, we distil some general insights from the results in Sections 7.3 and 7.4.

7.5.1 Guarantee

As expected, Section 7.3.1 and 7.4.1 confirmed that higher guarantees result in less overbooking. This is an intuitive result; the higher the guarantee, the more conservative the slice provider must be when overbooking. Less obvious, perhaps, is that the slice provider is limited in the choice of guarantee that it can offer, with no overbooking possible above an upper bound, and no further improvement possible below a lower bound.

7.5.2 Total Number of Available Resources

The total number of available resources, relative to the expected value of the sum demand for all slices, influences the bounds on the guarantee. As observed in Section 7.3.2, higher guarantees can be provided as the total number of resources available increases. If the total number of resources is significantly greater than the expected value of the sum demand for all slices, then the slice provider will have enough resources to satisfy the demand of the slices in the majority of the time slots. Overbooking can be then used, with a high guarantee placed on auxiliary resources, to increase resource utilisation. As demonstrated in Section 7.4.2, overbooking is not effective when there are fewer resources available than the total expected demand.

7.5.3 Uncertainty in Resource Demand Forecast

In Section 7.3.3, we observed that the upper limit on the guarantee decreases as the *uncertainty* associated with the forecasted resource demand increases. Hence, the slice provider must reduce the guarantee it offers to slice tenants if the variation in the forecasted demand of the slices increases. Again, this is intuitive; greater *uncertainty* in the forecast requires the slice provider to be more conservative in the guarantee it offers. This result is confirmed in the numerical analysis of the more complex case in Figure 7.7.

Hence, it may then seem counter-intuitive that, for a fixed guarantee in the analytical study in Section 7.3.3, the overbooking fraction increases with the standard deviation at times in Figure 7.5, as it appears to imply that less accurate predictions are more favourable towards overbooking. The key lies in understanding that the variance of $\widehat{D}$ does not represent an error in the prediction model, but rather accounts for the inherent stochastic nature of the underlying random process representing the demand of a slice.

Hence, the results indicate that overbooking is better suited when demand is sufficiently unpredictable. This can be reasoned by considering the extreme case of perfect prediction when $\sqrt{\frac{1}{12}(b_U - b_L)^2} = 0$. In this case, we know exactly how many resources each slice needs and offering resources in excess of this with a guarantee of availability is meaningless. In the presence of a small amount of uncertainty, e.g. one standard deviation, only a small number of auxiliary resources are required.

Conversely, we saw in the numerical analysis of the more complex case in Section 7.4.1 that the overbooking fraction decreased with increasing uncertainty. Hence, the relationship between uncertainty in the forecast and overbooking is dependent on the demand profiles of the slices. We can conclude that while some uncertainty is required to make overbooking possible, too much uncertainty reduces the ability to provide auxiliary resources at a given guarantee level.

We note that any error in the prediction model itself is different from the *uncertainty* relating to the stochastic nature of a slice's demand, and would affect the effectiveness of overbooking. For example, a bias error would shift the distribution of $\widehat{D}_{\text{sum}}$ either right or left and result in ineffective overbooking that would fail to satisfy its guarantees.

7.5.4 Correlation between Slices' Resource Demands

The correlation between slice tenants' demand for resources is also an important factor. Intuitively, we expect that negatively correlated slices should favour overbooking, as higher than expected demand in one slice should be countered by lower than expected demand in another slice. We investigated this further using a normal distribution for the sum forecasted resource demand for all slices and observed that negatively correlated slices result in more overbooking when $g > 0.5$. This is an encouraging result, as it is reasonable to assume that slice providers will want to provide guarantees greater than 0.5.

While it is not possible to make any definitive statements regarding resource demand correlation in future networks without real data, it is easy to imagine that the diverse services being targeted by future networks will be less positively correlated than the set of data services targeted by Long Term Evolution (LTE). Hence, overbooking appears, on paper at least, to be well suited to the diverse requirements and use cases being proposed for future networks.

7.6 Conclusion

The advantage of the approach outlined in Section 7.2 is that it can handle any set of demand distributions. Given the joint distribution of the forecasted demand for all slices, the sum distribution can be calculated numerically. Algorithm 2 then iteratively increases the auxiliary resource offer for each slice until the guarantee is no longer satisfied. If a guarantee is outside of the permitted range, the algorithm will successfully arrive at the correct amount of overbooking specified by the closest limit. Hence, the method proposed can be applied without concern regarding the current conditions of the system. The only assumption is that the joint demand distribution for all slices can be predicted. Forecasting this is challenging, and proposing methods for doing so are outside of the scope of this thesis.

Despite this, it is still important for a slice provider to be aware of the trends and insights presented above, particularly when determining what level of guarantee to offer. Intuition regarding how the system responds to various inputs is imperative when monitoring the system, especially if the offered guarantee is not being satisfied. Hence, the contribution in this chapter is two-fold. First, we provide an algorithm to calculate how many auxiliary resources to offer, which determines the amount of overbooking that occurs. We then analyse how the performance of the system depends on key inputs, providing valuable intuition to the slice provider.

8 Conclusion

Conclusion

There has been a trend since the first generation of telecommunication networks towards differentiated services, as discussed in Section 2.1. These services benefit Mobile Network Operators (MNOs) through new revenue streams, and can foster innovation across service providers. For example, consider the advent of data services in the past decade and the transformative effect it has had on society, with every aspect of a person's life now touched by the ability to both report and access large amounts of information from any habitable location.

The trend mentioned above appears set to peak in the next generation of future networks. As MNOs set their sights on non-traditional industries and user groups, the set of requirements and use cases that they must cater for grows in diversity. This diversity is the crux of the problem currently faced by MNOs, namely how to simultaneously satisfy multiple requirements which are in conflict with one another. These conflicts, the associated trade-offs, and the set of technologies available to fulfil the diverse range of requirements placed on the network are extensively discussed in Chapter 2, which highlights progression from differentiated services to a customisable network.

A customisable network provides tailored network behaviour on-demand, judiciously configuring the network using the techniques and technologies best suited to the targeted use case. Hence, future networks will experience a paradigm shift towards a network for services model, in which the network is defined by its flexibility rather than any individual constituent technology that it employs. As noted in Chapter 3, this may be a self-fulfilling prophecy. As networks become more adaptable, a wider range of industries will innovate and use network services, which in turn requires further capabilities from networks, and so on.

We have shown in this thesis that the advent of customisable networks presents new challenges, requiring multiple contrasting technologies to coexist in a single network. In particular, we identified virtualisation as an enabling technology in this regard, providing the capability to serve diverse use cases through tailored virtual network slices. Throughout this thesis, we examined the implications of introducing increased flexibility into future networks, and the challenges associated with achieving it.

We first demonstrated that future networks may permit multiple waveforms to coexist, with each scenario employing a waveform best suited to it. We then focused on resource allocation in future networks consisting of service-tailored slices, and examined the trade-off between flexibility and the cost of achieving that flexibility. We also examined a second, related trade-off between resource utilisation and slice-tailored performance. Finally, we examined a business model for future networks in which users subscribe to a bundle of service-tailored slices, and focused on how to match users to the appropriate slice in the bundle.

8.1 Summary of the Findings

In this section, we summarise the findings of the thesis.

8.1.1 Coexistence of Technologies

Having identified in Chapter 2 that the trend towards customisable future networks will necessitate the coexistence of contrasting technologies, we explored this concept in detail in Chapter 4. In particular, we focused on the coexistence of waveforms in a single network. The choice of waveform to employ in a network is the ideal scenario through which to explore the coexistence of disparate technologies in future networks for two reasons. Firstly, each waveform exhibits different strengths and weaknesses, proving advantageous in some scenarios and not in others. Secondly, different waveforms operating in adjacent frequency bands can result in interference, making coexistence difficult.

Hence, while the choice of waveform may benefit one use case, it may negatively affect performance in another. We considered a scenario consisting of two coexisting use cases: regular cellular traffic and clustered asynchronous Device-to-Device (D2D) communication. The clustered D2D communication is representative of a future network scenario such as a smart factory with Directly Communicating Users (DUEs) which do not synchronise through a Base Station (BS), while the cellular traffic is representative of a LTE-type scenario with Cellular Users (CUEs) availing of data services and communicating through the network via a BS.

We first demonstrated that inter-D2D interference is significant in the clustered scenario considered and must be accounted for when assigning spectral resources and assigning a power level to DUEs. This greatly complicates the optimal resource allocation problem formulation. We then demonstrated that the inter-D2D interference is rendered insignificant if DUEs use a waveform with enhanced spectral localisation such as Filter Bank Multi-Carrier with Offset Quadrature Amplitude Modulation (FBMC/OQAM), even if CUEs continue to use Orthogonal Frequency Division Multiplexing (OFDM). Hence, the coexistence of waveforms can permit the continued use of relatively simple current resource allocation procedures.

We then performed extensive system-level simulations for a multi-cell network to evaluate the relative performance of adopting a range of alternative waveforms for the clustered

Machine-Type Communication (MTC) scenario under study. We concluded that waveforms with small sidelobes perform best, with DUEs experiencing a higher Signal-to-Noise and Interference Ratio (SINR) due to the reduced inter-D2D interference. This benefit is confined to DUEs; there is no benefit to CUEs if DUEs use a waveform with enhanced filtering.

In summary, our simulations demonstrated that multiple waveforms can coexist in the network, and prove advantageous to the use case that they are targeting without adversely affecting other use cases. We highlight, however, that the decision of whether to use multiple waveforms is multifaceted and should not be made solely based on performance metrics such as SINR and achieved rate. The complexity and financial cost of supporting multiple waveforms in a single device may be undesirable. If the concept is to prove commercially viable, it may be best suited to bespoke devices such as machinery in a smart factory, which are typically designed for a single purpose and need only support a single waveform.

8.1.2 Service-Tailored Slices

While Chapter 4 demonstrated that multiple waveforms can coexist in a single network to serve two diverse use cases, it is difficult to extend this concept to any arbitrary set of use cases and any arbitrary set of technologies. A general framework for providing customisable network behaviour for a specific use case in an isolated manner, without impacting the service provided to other use cases, is required. Virtualisation provides us with this framework, enabling logically isolated service-tailored slices to be instantiated on-demand. The merits of virtualisation and network slicing in the context of creating adaptable future networks is provided in Chapter 3.

As stated in Chapter 1, every resource allocation problem can be expressed as a trade-off between two or more desirable states that cannot be achieved simultaneously. In Chapters 5 and 7, we focus on two of the trade-offs that will be present in any future network consisting of service-tailored slices, and examine the associated implications on resource allocation.

Trade-off 1: Adaptability Vs Cost of Achieving Adaptability

In Chapter 5, we identified that there is a trade-off between increasing the ability of the network to respond to traffic changes, and the overhead required to ensure isolation between network slices. While we considered the coexistence of multiple waveforms in Chapter 4, the flexibility in this case comes from the ability to tweak the parameters of a single waveform according to the use case being targeted.

We examined the design of the time-frequency resource grid, considering four different approaches ranging from static to very dynamic. Slices might employ different numerologies depending on the service they are targeting, resulting in inter-numerology interference. The dynamic approaches permit more granular allocations of resources; this increases the adaptability of the system, but also increases the number of boundaries between slices. As a result, inter-numerology interference is increased, which makes it more difficult to maintain isolation between slices. On the other hand, static approaches make it easier to ensure isolation, but are very limited in their ability to adapt to traffic changes.

Based on this analysis, we proposed the concept of a RAN profile, which distinguishes between the concept of a service and a service-type. Hence, time-frequency resource allocations can be performed in a two-tier manner based on this distinction. Time-frequency resources, consisting of a single numerology, are first assigned to a service-type in a semi-static manner. This assignment is known as a RAN profile and consists of a contiguous portion of the resource grid. Time-frequency resources within these RAN profiles are then allocated to individual slices on a more granular basis. Since each RAN profile consists of a homogeneous numerology internally, inter-numerology interference is only a concern at profile boundaries, which are minimised by the coarse granularity of the profile assignments.

Trade-off 2: Guaranteeing Service-Tailored Behaviour Vs Resource Utilisation

We then focused in Chapter 7 on the twin goals in slicing of service-tailored performance and increased resource utilisation. These two goals are in conflict. The commercial viability of network slicing centres on its ability to provide tailored behaviour on-demand [99], yet guaranteeing the performance of a slice limits the sharing gains that can be obtained. We considered a model consisting of a slice provider who leases virtual resources to multiple slice tenants, with each tenant targeting a specific vertical.

Issuing dedicated resources to slice tenants grants them confidence regarding the level of service that they can provide, but can lead to an inefficient use of resources. On the other hand, providing resources without any guarantee of availability, such as best-effort resources, results in high resource utilisation but low confidence regarding the quality of service that a slice tenant can provide to its subscribers.

To balance this trade-off, we proposed an approach that provided a mixture of both long and short-term guarantees surrounding resource availability. In the long-term, assured resources are made available to slices with a guarantee of being available if needed. In the short-term, auxiliary resources are offered to slices with an associated probability of availability. The guarantees of availability provide confidence to slice tenants regarding the performance they can offer subscribers, while the short-term offers of auxiliary resources allows the slice provider to take advantage of current traffic demand and maximise resource utilisation.

We provided an algorithm to specify how many auxiliary resources to offer based on current demand while still satisfying the probability of availability. Resource utilisation can be improved by offering more resources than are actually available by using a technique known as overbooking. We demonstrated that the effectiveness of overbooking is dependent on many inputs into the system such as the guarantee provided, the number of available resources, and the uncertainty associated with the demand forecast of slices. There are also upper and lower limits to the range of guarantees that the slice provider can offer. The influence of these factors is difficult to predict, as the demand distribution of slices may take on many different forms. While it is important to understand how each of these inputs affect how the system operates, the advantage of the algorithm provided is that it can handle any set of demand distributions.

8.1.3 New Business Models for Slicing

Providing tailored network behaviour to niche user communities and new verticals is critically important for MNOs to increase their revenue streams [99]. The advent of service-tailored slices in future networks demands new business models. In Chapter 6, we suggested that users may wish to avail of the services of multiple service-tailored slices, and proposed a business model to achieve this.

A new entity, called a subscription broker, groups multiple slices into a subscription bundle, and sells this bundle to subscribers with a fixed data allowance that can be used across all slices in the bundle. We focused on how subscribers are matched to slices within these bundles. We asserted that, for each party to be incentivised to adopt the proposed business model, matchings should exhibit the property of stability. That is, both slices and subscribers should be satisfied that they could not have achieved a better matching using a different business model.

To achieve this property, we employed a branch of game theory known as matching theory, and adopted the Gale-Shapley algorithm to match users and slices based on their preferences. We also specified how to build the preference lists of both users and slices. Our case study demonstrated that the proposed model in which users subscribe to a bundle of slices outperforms the traditional case where users only subscribe to a single slice.

8.2 Alternative Vision

In this thesis, we adopted virtualisation as the enabling technology to achieve adaptable future networks which can provide customisable behaviour on-demand. We chose virtualisation because, as outlined in Chapter 3, it is well suited to this task and there is a wealth of research into network slicing. We note, however, that virtualisation is not the only option in the pursuit of customisable future networks. In this section, we briefly present an alternative vision based on the concept of a cognitive network.

We focus on the concept of a cognitive network as first defined in [186]: *A cognitive network is a network with a cognitive process that can perceive current network conditions, and then plan, decide, and act on those conditions. The network can learn from these adaptations and use them to make future decisions, all while taking into account end-to-end goals.*

We explicitly distinguish the cognitive network concept from that of a cognitive radio. A cognitive network possesses end-to-end goals, giving it a network-wide scope. In contrast, a cognitive radio possesses user-centric goals giving it local scope. The two concepts share common traits, however. Both concepts share similar models of cognition, learning from past experiences which influence decisions made in the future. Cognitive radio implements actions based on its observations through tunable parameters supplied by Software Defined Radio (SDR). Cognitive networks on the other hand, dealing on a network-wide scope, require tunable parameters in the form of a Software Adjustable Network (SAN) [186, 187].

The cognitive network definition has similarities to an earlier concept called the Knowledge Plane, described in [188]. The Knowledge Plane construct is described as '*a distributed cognitive system that permeates the network*', with the stated objective of creating a new kind of network that is capable of assembling itself based on a high-level description, detecting faults, and repairing itself. While much of the Internet's remarkable success has resulted from its core design principle of transporting data through the core without concern for what purpose the data serves, this has also resulted in severe limitations in terms of management, configuration, and fault diagnosis, each of which requires manual attention. The Knowledge Plane concept aims to construct a network based on cognitive systems that is able to make low-level decisions on its own, based on current network conditions and high-level descriptions of its design goals.

The concept of a cognitive network is elaborated upon in [187], which emphasises the importance of end-to-end goals. In effect, all elements in the network involved in data flow are part of the cognitive process, capable of providing information about the network and offering adaptability. The network should not be reactive, but should instead be able to make decisions based on predictive models constructed using past observations. In summary, the cognitive network inputs observations of network performance, uses these observations in a decision making process, and implements actions based on these decisions through adjustable network elements.

In order to be effective, the cognitive network requires extensive knowledge of network state for the decision-making process. Focusing on obtaining network state information, the cognitive process must have access to state across the entire network. Knowing the state of the entire network is somewhat unrealistic and, as a result, the cognitive process should be able to deal with incomplete information. Often the process will only require a subset of state information, obtaining the relevant pieces through filtering. The layered nature of networks provides a blockage in terms of the flow of state information in the network. Often a layer may be able to provide information that could potentially influence an adaptation at a different layer. Hence, cognitive networks must operate cross-layer.

Wireless cognitive networks are also the focus of [189], which emphasises the business and management aspects. Interestingly, the authors identify that a complementary idea to the cognitive networking idea is to simply have cooperating networks with different Radio Access Technologies (RATs), from which a network operator can choose the one that best suits their needs. This is similar to the idea of customised virtual network slices presented in this thesis. In effect, a virtual slice gives an operator a customised network which has been tailored to their needs, whereas cognitive networking gives an operator a network that is able to adapt itself according to the demands placed on it.

A cognitive network requires adjustable network elements that allow it to implement a set of actions based on the decisions it makes. In this regard, a cognitive network is limited by the flexibility of the network itself. If the cognitive process is unable to adjust the network based on the decisions it makes and in accordance with its end-to-end goals, then the application of the cognitive network is fruitless. Instead, a SAN is needed which presents tunable or modifiable components, allowing the cognitive process to adjust one or more layers in the network stack belonging to various network elements such as switches or Baseband Units (BBUs).

Cognitive networks offer obvious potential in the context of adjustable networks. The cognitive network removes the need for an operator to tune the network, and is instead capable of autonomously adapting itself to the various service requirements as required. In addition, the RATs presented in Section 2.3 offer the adaptability required by a cognitive network to be effective. Each RAT, such as duplexing or multiple antenna use, offers choices and modifiable elements that the cognitive process can utilise to adapt the network accordingly. Emerging system-level techniques such as Virtualised RAN (VRAN) and Software-Defined Networking (SDN) also offer adaptability that can be used to alter the operation of the network. State information obtained at the radio access level may influence adaptations at the system-level, and vice versa. In this regard, the cognitive network concept unifies the RATs and emerging system-level techniques. In essence, the diverse service requirements and flexible technologies make future networks a potentially excellent fit for integration with the cognitive network concept.

The concept of a cognitive network is a broad topic with many different techniques fitting the description, yet the realisation of a truly cognitive network remains unseen. In [188], published in 2003, the need for an adaptable network designed using artificial intelligence and cognitive techniques was identified. Almost two decades later, our networks are arguably more adaptive, but this adaptivity is confined to certain parts of the network and arises from the use of algorithmic techniques applied in these areas, rather than an inherent intelligence permeating the entire network. The lack of a true SAN has restricted the development of the cognitive network concept; however, it may be on the cusp of experiencing its coming of age moment. Similar to the way in which advances in SDR preceded and enabled a plethora of research in the area of cognitive radio, the current movement towards a software defined RAN, coupled with SDN techniques, may herald a renewed interest into extending the cognitive radio concept to the entire network.

8.3 Outlook and Future Directions

Caution should be exercised when extrapolating any trend, but at this moment in time, it certainly appears that future networks will be more adaptable. It remains to be seen whether virtualisation and network slicing will indeed play a prominent role in future networks, and whether the vision of on-demand tailored network behaviour will be realised. Reviewing the current state-of-the-art in Chapters 2 and 3, steady steps along the path towards this vision have been taken. There is certainly an appetite for this vision in industry, as demonstrated in Section 3.2.2.

With respect to the coexistence of waveforms, the choice of modulation format has been a central debate in each generation, and this is likely to continue into the future. Unlike previous generations, however, there may not be a single choice, with requirements too diverse to be satisfied by a single implementation of a waveform. While we demonstrated that multiple waveforms can coexist and provide performance benefits in certain scenarios, the complexity of supporting multiple waveforms in a single device may not appeal to vendors. Instead, initially at least, the preferred approach is likely to be a single waveform with adjustable parameters; this is the approach that has been adopted for 5G New Radio (NR). Nonetheless, the coexistence of waveforms may be an attractive approach for niche equipment which only needs to support a single waveform, such as in the MTC-type scenario considered in Chapter 4. The viability of coexisting waveforms needs to be considered on a case-by-case basis for other scenarios which have specialised communication requirements.

In the context of a single waveform with adjustable parameters, designing the time-frequency resource grid in such a manner as to manage the isolation and adaptability of slices is important. Further experimental work is required to characterise the inter-numerology interference between slices, which can then be used to precisely optimise the granularity of resource allocations, and the size of the guard bands required. Non-orthogonal multiple access techniques may also have a role to play in future networks, and further work is required to examine how this would affect the adaptability/cost trade-off discussed in Chapter 5.

Focusing on the related trade-off between resource utilisation and providing tailored service with high probability, we demonstrated the advantages of resource overbooking in this regard in Chapter 7. We also examined how the effectiveness of overbooking is affected by the inputs into the system. The next step would be to examine the demand distributions of future network services using real-world data and determine whether they are conducive to overbooking. This was conducted by the authors in [180]; however, large scale data is not yet readily available for many future network services, such as virtual reality and autonomous cars, and further studies should be conducted as these use cases become more prominent.

Finally, further work is required to determine the business model for a future network consisting of service-tailored slices. Traditionally, spectrum was at the root of an MNO's value chain [11]. Network slicing alters this value chain, placing more importance on the provision of specialised services to niche user groups. We demonstrated through our broker-based model in Chapter 6 that the business model not only has implications for the revenue of the slice operator, but can also affect the performance experienced by end users. In this regard, more economic studies into the business models of network slicing are required, and these studies should jointly consider the ramifications of these business models on the service quality experienced by end users.

Acronyms

3GPP	3$^{\text{rd}}$ Generation Partnership Project
AIV	Air Interface Variant
AMC	Adaptive Modulation and Coding
API	Application Programming Interface
ARPU	Average Revenue Per User
ASM	Adaptive Spatial Modulation
ATA	Automatic Timing Adjustment
BBU	Baseband Unit
BS	Base Station
BWP	Bandwidth Part
CAPEX	Capital Expenditure
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CFO	Carrier Frequency Offset
COMP	Coordinated Multi-Point
CORESET	Control-Resource Set
CP	Cyclic Prefix
CPRI	Common Public Radio Interface
CQI	Channel Quality Indicator
Cloud-RAN	Cloud Radio Access Network
CSI	Channel State Information

CU	Centralised Unit
CUE	Cellular User
D2D	Device-to-Device
DAS	Distributed Antenna System
DSA	Dynamic Spectrum Access
DU	Distributed Unit
DUE	Directly Communicating User
eMBB	enhanced Mobile Broadband
ECPRI	Evolved Common Public Radio Interface
FBMC	Filter Bank Multi-Carrier
FBMC-PAM	Filter Bank Multi-Carrier - Pulse Amplitude Modulation
FBMC/OQAM	Filter Bank Multi-Carrier with Offset Quadrature Amplitude Modulation
FCC	Federal Communications Commission
FDD	Frequency-Division Duplexing
FDMA	Frequency-Division Multiple Access
FFR	Fractional Frequency Reuse
FFT	Fast Fourier Transform
FMT	Filtered Multi-Tone
f-OFDM	Filtered Orthogonal Frequency Division Multiplexing
GBR	Guaranteed Bit Rate
GFDM	Generalised Frequency Division Multiplexing
GFDM	Generalised Frequency Division Multiplexing
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GWCN	Gateway Core Network
HD	Half Duplex

IBFD	In-Band Full Duplex
ICI	Inter-Carrier Interference
ICIC	Inter-Cell Interference Cancellation
ICT	Information and Communication Technology
IFFT	Inverse Fast Fourier Transform
IoT	Internet of Things
ISD	Inter-Site Distance
ISI	Inter-Symbol Interference
JR	Joint Reception
JT	Joint Transmission
KPI	Key Performance Indicator
LO	Local Oscillator
LTE	Long Term Evolution
MAC	Medium Access Control
MCC	Mobile Country Codes
MIMO	Multiple-Input Multiple-Output
MINLP	Mixed Integer Non-Linear Programming
MME	Mobility Management Entity
Massive-MIMO	Massive Multiple-Input Multiple-Output
mMTC	Massive Machine-Type Communication
MNC	Mobile Network Codes
MNO	Mobile Network Operator
MOCN	Multi-Operator Core Network
MORAN	Multi-Operator Radio Access Network
MTC	Machine-Type Communication
MU-MIMO	Multi-User Multiple-Input Multiple-Output
MVNO	Mobile Virtual Network Operator

MVNP	Mobile Virtual Network Provider
NB-IoT	Narrowband Internet of Things
NFV	Network Function Virtualisation
NOMA	Non-Orthogonal Multiple Access
NR	New Radio
NVS	Network Virtualisation Substrate
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OOB	Out-of-Band
OPEX	Operational Expenditure
OQAM	Offset Quadrature Amplitude Modulation
PAM	Pulse Amplitude Modulation
PAPR	Peak-to-Average Power Ratio
PDCCH	Physical Downlink Control Channel
PDF	Probability Density Function
PGW	Packet Data Network Gateway
PHY	Physical Layer
PRB	Physical Resource Block
PSD	Power Spectral Density
PU	Primary User
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Identifier
QoE	Quality of Experience
QoS	Quality of Service
RA	Resource Allocation
RAN	Radio Access Network

RAT	Radio Access Technology
RB	Resource Block
RLC	Radio Link Control
RRH	Remote Radio Head
RoF	Radio over Fibre
Rx	receiver
SAN	Software Adjustable Network
SC-FDMA	Single Carrier - Frequency Division Multiple Access
SCM	Single Carrier Modulation
SDAI	Software-Defined Air Interface
SDN	Software-Defined Networking
SDR	Software Defined Radio
SDW	Software-Defined Waveform
SIC	Self-Interference Cancellation
SINR	Signal-to-Noise and Interference Ratio
SLA	Service Level Agreement
SM	Spatial modulation
SMNO	Specialised Mobile Network Operator
SoDeMa	Software Defined Multiple Access
SU	Secondary User
TDD	Time-Division Duplexing
TDMA	Time-Division Multiple Access
TO	Timing Offset
TTI	Transmission Time Interval
Tx	transmitter
UE	User Equipment
UFMC	Universal Filtered Multi-Carrier

uMTC	Ultra-Reliable Machine-Type Communication
UMTS	Universal Mobile Telecommunications Service
uRLLC	Ultra-Reliable Low Latency Communication
VLAN	Virtual Local Area Network
VPN	Virtual Private Network
VRAN	Virtualised RAN
WCDMA	Wideband Code Division Multiple Access
WLAN	Wireless Local Area Network

Bibliography

[1] Y. Medjahdi, M. Terré, D. L. Ruyet, and D. Roviras, "Interference tables: a useful model for interference analysis in asynchronous multicarrier transmission," *EURASIP Journal on Advances in Signal Processing*, no. 54, pp. 1–17, Apr. 2014.

[2] Q. Bodinier, A. Farhang, F. Bader, H. Ahmadi, J. Palicot, and L. A. DaSilva, "5G waveforms for overlay D2D communications: Effects of time-frequency misalignment," in *IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–7.

[3] Q. Bodinier, F. Bader, and J. Palicot, "Modeling Interference Between OFDM/OQAM and CP-OFDM : Limitations of the PSD-Based Model," in *International Conference on Telecommunications (ICT)*, May 2016, pp. 1–7.

[4] H. S. Chae, J. Gu, B. G. Choi, and M. Y. Chung, "Radio resource allocation scheme for device-to-device communication in cellular networks using fractional frequency reuse," in *Asia Pacific Conference on Communications (APCC)*, Oct. 2011, pp. 58–62.

[5] Qualcomm, "5G - Vision for the next generation of connectivity," Mar. 2015.

[6] S. Goyal, P. Liu, S. S. Panwar, R. A. Difazio, R. Yang, and E. Bala, "Full duplex cellular systems: will doubling interference prevent doubling capacity?" *IEEE Communications Magazine*, vol. 53, no. 5, pp. 121–127, May 2015.

[7] Ericsson, "5G systems - enabling industry and society transformation," Jan. 2015.

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

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

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

[11] L. Doyle, J. Kibilda, T. K. Forde, and L. DaSilva, "Spectrum Without Bounds, Networks Without Borders," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 351–365, Mar. 2014.

[12] J. Zander, "Radio resource management in future wireless networks: requirements and limitations," *IEEE Communications Magazine*, vol. 35, no. 8, pp. 30–36, Aug. 1997.

[13] S. Ni and S.-G. Häggman, “GPRS performance estimation in GSM circuit switched services and GPRS shared resource systems,” in *Wireless Communications and Networking Conference (WCNC)*, Sep. 1999, pp. 1417–1421.

[14] M. Ermel, K. Begain, T. Müller, J. Schüler, and M. Schweigel, “Analytical Comparison of Different GPRS Introduction Strategies,” in *ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Aug. 2000, pp. 3–10.

[15] 3GPP, “Universal Mobile Telecommunications System; Quality of Service concept and architecture (TS 23.107 V3.9.0),” 1999.

[16] H. Kröner, “Radio Resource Allocation for Data Services in UMTS Networks,” *AEU - International Journal of Electronics and Communications*, vol. 55, pp. 55–62, Jan. 2001.

[17] S.-E. Kim, H. Kim, and J. A. Copeland, “Dynamic radio resource allocation considering QoS in UMTS network,” in *Workshop on Mobile and Wireless Communication Networks*, Sep. 2002, pp. 636–640.

[18] D. Staehle, K. Leibnitz, K. Heck, P. Tran-Gia, B. Schröder, and A. Weller, “Analytic approximation of the effective bandwidth for best-effort services in UMTS networks,” in *IEEE Vehicular Technology Conference (VTC Spring)*, vol. 2, Apr. 2003, pp. 1153–1157.

[19] C. Cox, *An introduction to LTE: LTE, LTE-advanced, SAE and 4G mobile communications.* John Wiley & Sons, 2012.

[20] 3GPP, “Technical Specification Group Services and System Aspects; Policy and charging control architecture (TS 23.303 V16.1),” Jun. 2017.

[21] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, “Scenarios for 5G mobile and wireless communications: the vision of the METIS project,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014.

[22] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What Will 5G Be?” *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[23] R. L. G. Cavalcante, S. Stanczak, M. Schubert, A. Eisenblaetter, and U. Tuerke, “Toward Energy-Efficient 5G Wireless Communications Technologies: Tools for decoupling the scaling of networks from the growth of operating power,” *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 24–34, Nov. 2014.

[24] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, “Network densification: the dominant theme for wireless evolution into 5G,” *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[25] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, “Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 86–92, May 2014.

[26] Y.-S. Choi and H. Shirani-Mehr, “Simultaneous Transmission and Reception: Algorithm, Design and System Level Performance,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 12, pp. 5992–6010, Dec. 2013.

[27] S. Goyal, P. Liu, S. Hua, and S. Panwar, “Analyzing a full-duplex cellular system,” in *Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2013, pp. 1–6.

[28] S. Goyal, P. Liu, S. Panwar, R. A. Difazio, R. Yang, J. Li, and E. Bala, “Improving small cell capacity with common-carrier full duplex radios,” in *IEEE International Conference on Communications (ICC)*, Jun. 2014, pp. 4987–4993.

[29] S. Goyal, P. Liu, and S. S. Panwar, “User Selection and Power Allocation in Full Duplex Multi-Cell Networks,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2408–2422, 2016.

[30] A. C. Cirik, K. Rikkinen, and M. Latva-aho, “Joint Subcarrier and Power Allocation for Sum-Rate Maximization in OFDMA Full-Duplex Systems,” in *IEEE Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[31] W. Cheng, X. Zhang, and H. Zhang, “Optimal power allocation for full-duplex D2D communications over wireless cellular networks,” in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2014, pp. 4764–4769.

[32] L. Wang, F. Tian, T. Svensson, D. Feng, M. Song, and S. Li, “Exploiting full duplex for device-to-device communications in heterogeneous networks,” *IEEE Communications Magazine*, vol. 53, no. 5, pp. 146–152, May 2015.

[33] S. Kim and W. Stark, “Full duplex device to device communication in cellular networks,” in *International Conference on Computing, Networking and Communications (ICNC)*, Feb. 2014, pp. 721–725.

[34] T. Riihonen, S. Werner, and R. Wichman, “Hybrid Full-Duplex/Half-Duplex Relaying with Transmit Power Adaptation,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 9, pp. 3074–3085, Sep. 2011.

[35] W. Cheng, X. Zhang, and H. Zhang, “Full/half duplex based resource allocations for statistical quality of service provisioning in wireless relay networks,” in *IEEE INFOCOM*, Mar. 2012, pp. 864–872.

[36] G. Liu, F. R. Yu, H. Ji, and V. C. M. Leung, “Distributed resource allocation in full-duplex relaying networks with wireless virtualization,” in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2014, pp. 4959–4964.

[37] G. Liu, F. R. Yu, H. Ji, V. C. M. Leung, and X. Li, “In-band full-duplex relaying for 5G cellular networks with wireless virtualization,” *IEEE Network*, vol. 29, no. 6, pp. 54–61, Nov. 2015.

[38] N. Wang, E. Hossain, and V. K. Bhargava, “Backhauling 5G small cells: A radio resource management perspective,” *IEEE Wireless Communications*, vol. 22, no. 5, pp. 41–49, Oct. 2015.

[39] U. Siddique, H. Tabassum, and E. Hossain, “Adaptive in-band self-backhauling for full-duplex small cells,” in *IEEE International Conference on Communication (ICC) Workshops*, Jun. 2015, pp. 44–49.

[40] Y. Liao, L. Song, Z. Han, and Y. Li, “Full duplex cognitive radio: a new design paradigm for enhancing spectrum usage,” *IEEE Communications Magazine*, vol. 53, no. 5, pp. 138–145, May 2015.

[41] W. Cheng, X. Zhang, and H. Zhang, “Full duplex spectrum sensing in non-time-slotted cognitive radio networks,” in *Military Communications Conference (MILCOM)*, Nov. 2011, pp. 1029–1034.

[42] E. Ahmed, A. Eltawil, and A. Sabharwal, “Simultaneous transmit and sense for cognitive radios using full-duplex: A first study,” in *IEEE Antennas and Propagation Society International Symposium (APSURSI)*, Jul. 2012, pp. 1–2.

[43] W. Afifi and M. Krunz, “Adaptive transmission-reception-sensing strategy for cognitive radios with full-duplex capabilities,” in *IEEE International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, Apr. 2014, pp. 149–160.

[44] ——, “Exploiting self-interference suppression for improved spectrum awareness/efficiency in cognitive radio systems,” in *IEEE INFOCOM*, Apr. 2013, pp. 1258–1266.

[45] S.-K. Hong, J. Brand, J. Choi, M. Jain, J. Mehlman, S. Katti, and P. Levis, “Applications of self-interference cancellation in 5G and beyond,” *IEEE Communications Magazine*, vol. 52, no. 2, pp. 114–121, Feb. 2014.

[46] L. Zheng and D. N. C. Tse, “Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[47] C.-B. Chae, A. Forenza, R. W. Heath, M. R. McKay, and I. B. Collings, “Adaptive MIMO transmission techniques for broadband wireless communication systems [Topics in Wireless Communications],” *IEEE Communications Magazine*, vol. 48, no. 5, pp. 112–118, May 2010.

[48] P. Yang, Y. Xiao, Y. Yu, and S. Li, “Adaptive Spatial Modulation for Wireless MIMO Transmission Systems,” *IEEE Communications Letters*, vol. 15, no. 6, pp. 602–604, Jun. 2011.

[49] P. Yang, Y. Xiao, Y. Yu, L. Li, Q. Tang, and S. Li, “Simplified Adaptive Spatial Modulation for Limited-Feedback MIMO Systems,” *IEEE Transactions on Vehicular Technology*, vol. 62, no. 6, pp. 2656–2666, Jul. 2013.

[50] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, “Spatial Modulation,” *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2228–2241, Jul. 2008.

[51] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays,” *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[52] C. Masouros, M. Sellahurai, and T. Ratnarajah, “Bridging the gap between linear and non-linear precoding in small- and large-scale MIMO downlinks,” in *IEEE International Conference on Communications (ICC)*, Jun. 2014, pp. 4483–4487.

[53] K. A. Alnajjar, P. J. Smith, and G. K. Woodward, “Co-located and distributed antenna systems: deployment options for massive multiple-input multiple-output,” *IET Microwaves, Antennas & Propagation*, vol. 9, no. 13, pp. 1418–1424, Oct. 2015.

[54] S. Ma, Y. L. Yang, and H. Sharif, "Distributed MIMO technologies in cooperative wireless networks," *IEEE Communications Magazine*, vol. 49, no. 5, pp. 78–82, May 2011.

[55] P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek, and A. Ugolini, "Modulation Formats and Waveforms for 5G Networks: Who Will Be the Heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 80–93, Nov. 2014.

[56] G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, and F. Wiedmann, "5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 97–105, Feb. 2014.

[57] H. Saeedi-Sourck, Y. Wu, J. W. M. Bergmans, S. Sadri, and B. Farhang-Boroujeny, "Complexity and Performance Comparison of Filter Bank Multicarrier and OFDM in Uplink of Multicarrier Multiple Access Networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1907–1912, Apr. 2011.

[58] N. Benvenuto, R. Dinis, D. Falconer, and S. Tomasin, "Single Carrier Modulation With Nonlinear Frequency Domain Equalization: An Idea Whose Time Has Come Again," *Proceedings of the IEEE*, vol. 98, no. 1, pp. 69–96, Jan. 2010.

[59] G. Fettweis, M. Krondorf, and S. Bittner, "GFDM - Generalized Frequency Division Multiplexing," in *IEEE Vehicular Technology Conference (VTC Spring)*, Apr. 2009, pp. 1–4.

[60] Q. Sun, I. Chin-Lin, S. Han, Z. Xu, and Z. Pan, "Software defined air interface: a framework of 5G air interface," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Mar. 2015, pp. 6–11.

[61] S. Venkatesan and R. A. Valenzuela, "OFDM for 5G: Cyclic prefix versus zero postfix, and filtering versus windowing," in *IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–5.

[62] X. Zhang, M. Jia, L. Chen, J. Ma, and J. Qiu, "Filtered-OFDM - Enabler for Flexible Waveform in the 5th Generation Cellular Networks," in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2015, pp. 1–6.

[63] N. Michailow, M. Matthe, I. S. Gaspar, A. N. Caldevilla, L. L. Mendes, A. Festag, and G. Fettweis, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks," *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3045–3061, Sep. 2014.

[64] M. Danneberg, N. Michailow, I. Gaspar, D. Zhang, and G. Fettweis, "Flexible GFDM Implementation in FPGA with Support to Run-Time Reconfiguration," in *IEEE Vehicular Technology Conference (VTC Fall)*, Sep. 2015, pp. 1–2.

[65] H. Xing and M. Renfors, "Investigation of filter bank based device-to-device communication integrated into OFDMA cellular system," in *International Symposium on Wireless Communications Systems (ISWCS)*, Aug. 2014, pp. 513–518.

[66] T. Wild, F. Schaich, and Y. Chen, "5G air interface design based on Universal Filtered (UF-)OFDM," in *International Conference on Digital Signal Processing (DSP)*, Aug. 2014, pp. 699–704.

[67] F. Schaich and T. Wild, "Relaxed synchronization support of universal filtered multicarrier including autonomous timing advance," in *International Symposium on Wireless Communications Systems (ISWCS)*, Aug. 2014, pp. 203–208.

[68] K. Pedersen, F. Frederiksen, G. Berardinelli, and P. Mogensen, "A Flexible Frame Structure for 5G Wide Area," in *IEEE Vehicular Technology Conference (VTC Fall)*, Sep. 2015, pp. 1–5.

[69] K. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53–59, Mar. 2016.

[70] L. Dai, B. Wang, Y. Yuan, S. Han, C. l. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[71] M. Jarschel, T. Zinner, T. Hossfeld, P. Tran-Gia, and W. Kellerer, "Interfaces, attributes, and use cases: A compass for SDN," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 210–217, Jun. 2014.

[72] S. Sezer, S. Scott-Hayward, P. K. Chouhan, B. Fraser, D. Lake, J. Finnegan, N. Viljoen, M. Miller, and N. Rao, "Are we ready for SDN? Implementation challenges for software-defined networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 36–43, Jul. 2013.

[73] C. Bernardos, A. La Oliva, P. Serrano, A. Banchs, L. M. Contreras, H. Jin, and J. C. Zuniga, "An architecture for software defined wireless networking," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 52–61, Jun. 2014.

[74] N. A. Jagadeesan and B. Krishnamachari, "Software-Defined Networking Paradigms in Wireless Networks: A Survey," *ACM Computing Surveys*, vol. 47, no. 2, Nov. 2014.

[75] L. E. Li, Z. M. Mao, and J. Rexford, "Toward Software-Defined Cellular Networks," in *European Workshop on Software Defined Networking (EWSDN)*, Oct. 2012, pp. 7–12.

[76] X. Jin, L. E. Li, L. Vanbever, and J. Rexford, "SoftCell: Scalable and Flexible Cellular Core Network Architecture," in *ACM Conference on Emerging Networking Experiments and Technologies*, Dec. 2013, pp. 163–174.

[77] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software Defined Radio Access Network," in *ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, Aug. 2013, pp. 25–30.

[78] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward software-defined mobile networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 44–53, Jul. 2013.

[79] R. Sherwood, G. Gibb, K.-K. Yap, G. Appenzeller, M. Casado, N. McKeown, and G. Parulkar, "Flowvisor: A network virtualization layer," *OpenFlow Switch Consortium, Tech. Rep*, Oct. 2009.

[80] A. Tootoonchian, M. Ghobadi, and Y. Ganjali, “OpenTM: Traffic Matrix Estimator for OpenFlow Networks,” in *International Conference on Passive and Active Measurement*, Apr. 2010, pp. 201–210.

[81] S. R. Chowdhury, M. F. Bari, R. Ahmed, and R. Boutaba, “PayLess: A low cost network monitoring framework for Software Defined Networks,” in *IEEE Network Operations and Management Symposium (NOMS)*, May 2014, pp. 1–9.

[82] Z. Su, T. Wang, Y. Xia, and M. Hamdi, “FlowCover: Low-cost flow monitoring scheme in software defined networks,” in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2014, pp. 1956–1961.

[83] N. L. M. van Adrichem, C. Doerr, and F. A. Kuipers, “OpenNetMon: Network monitoring in OpenFlow Software-Defined Networks,” in *IEEE Network Operations and Management Symposium (NOMS)*, May 2014, pp. 1–8.

[84] 3GPP, “NG-RAN; F1 general aspects and principles (TS 38.470 V15.7.0),” 2019.

[85] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, “Cloud radio access network (C-RAN): a primer,” *IEEE Network*, vol. 29, no. 1, pp. 35–41, Jan. 2015.

[86] U. Dötsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier, “Quantitative analysis of split base station processing and determination of advantageous architectures for LTE,” *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, Jun. 2013.

[87] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, “Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN,” *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[88] Intel, “Exploring 5G Fronthaul Network Architecture Intelligence Splits and Connectivity,” Dec. 2019.

[89] 3GPP, “Study on new radio access technology: Radio access architecture and interfaces (TR 38.801 V14.0.0),” 2017.

[90] O. Simeone, E. Erkip, and S. Shamai, “Full-Duplex Cloud Radio Access Networks: An Information-Theoretic Viewpoint,” *IEEE Wireless Communications Letters*, vol. 3, no. 4, pp. 413–416, Aug. 2014.

[91] T. Anderson, L. Peterson, S. Shenker, and J. Turner, “Overcoming the Internet Impasse through Virtualization,” *IEEE Computer*, vol. 38, no. 4, pp. 34–41, Apr. 2005.

[92] N. M. Mosharaf Kabir Chowdhury and R. Boutaba, “Network Virtualization: State of the Art and Research Challenges,” *IEEE Communications Magazine*, vol. 47, no. 7, pp. 20–26, Jul. 2009.

[93] J. Carapinha and J. Jiménez, “Network virtualization: a view from the bottom,” in *ACM Workshop on Virtualized Infrastructure Systems and Architectures*, Aug. 2009, pp. 73–80.

[94] A. Wang, M. Iyer, R. Dutta, G. N. Rouskas, and I. Baldine, “Network Virtualization: Technologies, Perspectives, and Frontiers,” *Journal of Lightwave Technology*, vol. 31, no. 4, pp. 523–537, Feb. 2013.

[95] A. Nakao, "Network Virtualization as Foundation for Enabling New Network Architectures and Applications," *IEICE Transactions on Communications*, vol. 93, no. 3, pp. 454–457, Mar. 2010.

[96] J. Van De Belt, H. Ahmadi, and L. E. Doyle, "Defining and Surveying Wireless Link Virtualization and Wireless Network Virtualization," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1603–1627, Third Quarter 2017.

[97] C. Liang and F. R. Yu, "Wireless Network Virtualization: A Survey, Some Research Issues and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 358–380, First Quarter 2015.

[98] D.-E. Meddour, T. Rasheed, and Y. Gourhant, "On the role of infrastructure sharing for mobile network operators in emerging markets," *Computer Networks*, vol. 55, no. 7, pp. 1576 – 1591, May 2011.

[99] Nokia Bell Labs, "Unleasing the Economic Potential of Network Slicing," Apr. 2018.

[100] NGMN Alliance, "NGMN 5G White Paper," Feb. 2015.

[101] ——, "Description of Network Slicing Concept," Sep. 2016.

[102] 5G Americas, "Network Slicing for 5G Networks and Services," Nov. 2016.

[103] GSMA, "An Introduction to Network Slicing," Nov. 2017.

[104] FCC Technological Advisory Council – 5G IoT Working Group, "5G Network Slicing White Paper," 2018.

[105] Deutsche Telekom. China Mobile, Huawei and Volkswagen, "5G Service – Guaranteed Network Slicing," Feb. 2017.

[106] Ericsson, "5G systems: Enabling the Transformation of Industry and Society," Jan. 2017.

[107] Nokia, "Dynamic end-to-end network slicing for 5G," Jun. 2016.

[108] Qualcomm and Nokia, "Making 5G a reality: Addressing the strong mobile broadband demand in 2019 & beyond," Sep. 2017.

[109] AT&T, BT, CenturyLink, China Mobile, Colt, Deutsche Telekom, KDDI, NTT, Orange, Telecom Italia, Telefonica, Telstra, and Verizon, "Network Functions Virtualisation: An Introduction, Benefits, Enablers, Challenges & Call for Action," Oct. 2012.

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

[111] J. Matias, J. Garay, N. Toledo, J. Unzilla, and E. Jacob, "Toward an SDN-enabled NFV architecture," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 187–193, Apr. 2015.

[112] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, May 2017.

[113] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, “Mobile network architecture evolution toward 5G,” *IEEE Communications Magazine*, vol. 54, no. 5, pp. 84–91, May 2016.

[114] I. DaSilva, G. Mildh, A. Kaloxylos, P. Spapis, E. Buracchini, A. Trogolo, G. Zimmermann, and N. Bayer, “Impact of network slicing on 5G Radio Access Networks,” in *European Conference on Networks and Communications (EuCNC)*, Jun. 2016, pp. 153–157.

[115] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, “Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks,” *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, May 2017.

[116] 3GPP, “Technical Specification Group Radio Access Network; Study on latency reduction techniques for LTE (TR 36.881 V1.0.0),” Jun. 2016.

[117] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, “Massive machine-type communications in 5G: physical and MAC-layer solutions,” *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[118] 3GPP, “Network Sharing; Architecture and functional description (TS 23.251 V15.1.0),” Sep. 2018.

[119] Nokia, “Network Sharing: Delivering mobile broadband more efficiently and at lower cost,” 2018.

[120] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, “Radio access network virtualization for future mobile carrier networks,” *IEEE Communications Magazine*, vol. 51, no. 7, pp. 27–35, Jul. 2013.

[121] C. Liang and F. R. Yu, “Wireless virtualization for next generation mobile cellular networks,” *IEEE Wireless Communications*, vol. 22, no. 1, pp. 61–69, Feb. 2015.

[122] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, “NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks,” *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1333–1346, Oct. 2012.

[123] S. Katti and L. E. Li, “RadioVisor: A Slicing Plane for Radio Access Networks,” in *Open Networking Summit (ONS)*, Aug. 2014.

[124] 3GPP, “NR; NR and NG-RAN Overall Description; Stage 2 (TS 38.300 V15.2.0),” Dec. 2018.

[125] M. G. Kibria, G. P. Villardi, K. Nguyen, W. Liao, K. Ishizu, and F. Kojima, “Shared Spectrum Access Communications: A Neutral Host Micro Operator Approach,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 8, pp. 1741–1753, Aug. 2017.

[126] Y. Cao, T. Jiang, M. He, and J. Zhang, “Device-to-Device Communications for Energy Management: A Smart Grid Case,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 190–201, Jan. 2016.

[127] M. Bagaa, A. Ksentini, T. Taleb, R. Jantti, A. Chelli, and I. Balasingham, "An efficient D2D-based strategies for machine type communications in 5G mobile systems," in *IEEE Wireless Communications and Networking Conference*, Apr. 2016, pp. 1–6.

[128] J. Zhao, K. K. Chai, Y. Chen, J. Schormans, and J. Alonso-Zarate, "Joint mode selection and resource allocation for machine-type D2D links," *Transactions on Emerging Telecommunications Technologies*, vol. 28, no. 2, p. e3000, Nov. 2017.

[129] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, Apr. 2014.

[130] A. Aminjavaheri, A. Farhang, A. RezazadehReyhani, and B. Farhang-Boroujeny, "Impact of timing and frequency offsets on multicarrier waveform candidates for 5G," in *IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*, Aug. 2015, pp. 178–183.

[131] P. Wu, P. C. Cosman, and L. B. Milstein, "Resource Allocation for Multicarrier Device-to-Device Video Transmission: Symbol Error Rate Analysis and Algorithm Design," *IEEE Transactions on Communications*, vol. 65, no. 10, pp. 4446–4462, Oct. 2017.

[132] M. Pischella, R. Zakaria, and D. L. Ruyet, "Resource Block level power allocation in asynchronous multi-carrier D2D communications," *IEEE Communications Letters*, vol. 21, no. 4, pp. 813–816, Apr. 2017.

[133] ——, "Weighted sum rate maximization with filtered multi-carrier modulations for D2D underlay communications," in *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2016, pp. 1–6.

[134] Y. Li, X. Sha, and L. Ye, "Downlink Resource Sharing for D2D Communications in a Filtered OFDM System," in *IEEE Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–6.

[135] M. Mukherjee, L. Shu, Y. Zhang, Z. Zhou, and K. Wang, "Joint Power and Reduced Spectral Leakage-Based Resource Allocation for D2D Communications in 5G," in *International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP)*, Nov. 2015, pp. 244–258.

[136] H. Xing and M. Renfors, "Investigation of filter bank based device-to-device communication integrated into OFDMA cellular system," in *International Symposium on Wireless Communications Systems (ISWCS)*, Aug. 2014, pp. 513–518.

[137] T. Wang, J. G. Proakis, and J. R. Zeidler, "Interference Analysis of Filtered Multitone Modulation Over Time-Varying Frequency - Selective Fading Channels," *IEEE Transactions on Communications*, vol. 55, no. 4, pp. 717–727, Apr. 2007.

[138] B. Farhang-Boroujeny, "OFDM Versus Filter Bank Multicarrier," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 92–112, May 2011.

[139] P. Siohan, C. Siclet, and N. Lacaille, "Analysis and design of OFDM/OQAM systems based on filterbank theory," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1170–1183, May 2002.

[140] M. Bellanger, D. Mattera, and M. Tanda, "A filter bank multicarrier scheme running at symbol rate for future wireless systems," in *Wireless Telecommunications Symposium (WTS)*, Apr. 2015, pp. 1–5.

[141] N. Michailow, M. Matthé, I. S. Gaspar, A. N. Caldevilla, L. L. Mendes, A. Festag, and G. Fettweis, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks," *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3045–3061, Sep. 2014.

[142] V. Vakilian, T. Wild, F. Schaich, S. ten Brink, and J.-F. Frigon, "Universal-filtered multi-carrier technique for wireless systems beyond LTE," in *IEEE Global Communications Conference (GLOBECOM) Workshops*, Dec. 2013, pp. 223–228.

[143] J. Abdoli, M. Jia, and J. Ma, "Filtered OFDM: A new waveform for future wireless systems," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Jun. 2015, pp. 66–70.

[144] P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek, and A. Ugolini, "Modulation Formats and Waveforms for 5G Networks: Who Will Be the Heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 80–93, Nov. 2014.

[145] R. Gerzaguet, N. Bartzoudis, L. G. Baltar, V. Berg, J.-B. Doré, D. Kténas, O. Font-Bach, X. Mestre, M. Payaró, M. Färber, and K. Roth, "The 5G candidate waveform race: a comparison of complexity and performance," *EURASIP Journal on Wireless Communications and Networks*, no. 1, p. 13, Jan. 2017.

[146] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzic, M. Milojevic, A. Hong, J. Ylitalo *et al.*, "WINNER II channel models," *WINNER II Public Deliverable*, pp. 42–44, 2007.

[147] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, G. Feng, and S. Li, "Device-to-Device Communications Underlaying Cellular Networks," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3541–3551, Aug. 2013.

[148] J. Han, Q. Cui, C. Yang, and X. Tao, "Bipartite matching approach to optimal resource allocation in device to device underlaying cellular network," *Electronics Letters*, vol. 50, no. 3, pp. 212–214, Jan. 2014.

[149] L. Wang and H. Wu, "Fast Pairing of Device-to-Device Link Underlay for Spectrum Sharing With Cellular Users," *IEEE Communications Letters*, vol. 18, no. 10, pp. 1803–1806, Oct. 2014.

[150] M. Shaat and F. Bader, "A Two-Step Resource Allocation Algorithm in Multicarrier Based Cognitive Radio Systems," in *Wireless Communications and Networking Conference (WCNC)*, Apr. 2010, pp. 1–6.

[151] N. Saquib, E. Hossain, and D. Kim, "Fractional frequency reuse for interference management in LTE-advanced hetnets," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 113–122, Apr. 2013.

[152] T. Novlan, J. G. Andrews, I. Sohn, R. K. Ganti, and A. Ghosh, "Comparison of Fractional Frequency Reuse Approaches in the OFDMA Cellular Downlink," in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2010, pp. 1–5.

[153] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, “Device-to-device communication as an underlay to LTE-advanced networks,” *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, Dec. 2009.

[154] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, “Resource Sharing Optimization for Device-to-Device Communication Underlaying Cellular Networks,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[155] 3GPP, “Universal Mobile Telecommunications System; User Equipment radio transmission and reception (TS 25.101 V14.0.0),” May 2017.

[156] ——, “LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios (TR 36.942 V14.0.0),” Apr. 2017.

[157] D. Chen, D. Qu, T. Jiang, and Y. He, “Prototype Filter Optimization to Minimize Stopband Energy With NPR Constraint for Filter Bank Multicarrier Modulation Systems,” *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 159–169, Jan. 2013.

[158] K. Cain, V. Vakilian, and R. Abdolee, “Low-Complexity Universal-Filtered Multi-Carrier for Beyond 5G Wireless Systems,” in *International Conference on Computing, Networking and Communications (ICNC)*, Mar. 2018, pp. 254–258.

[159] L. Zhang, A. Ijaz, P. Xiao, and R. Tafazolli, “Multi-Service System: An Enabler of Flexible 5G Air Interface,” *IEEE Communications Magazine*, vol. 55, no. 10, pp. 152–159, Oct. 2017.

[160] F. Schaich, T. Wild, and R. Ahmed, “Subcarrier Spacing - How to Make Use of This Degree of Freedom,” in *IEEE Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–6.

[161] 3GPP, “NR; Physical channels and modulation (TS 38.211 V15.6.0),” Jun. 2019.

[162] L. Zhang, A. Ijaz, P. Xiao, A. Quddus, and R. Tafazolli, “Subband Filtered Multi-Carrier Systems for Multi-Service Wireless Communications,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1893–1907, Mar. 2017.

[163] METIS-II, “METIS-II final project report,” Jun. 2017.

[164] E. Pateromichelakis, J. Gebert, T. Mach, J. Belschner, W. Guo, N. P. Kuruvatti, V. Venkatasubramanian, and C. Kilinc, “Service-Tailored User-Plane Design Framework and Architecture Considerations in 5G Radio Access Networks,” *IEEE Access*, vol. 5, pp. 17 089–17 105, Aug. 2017.

[165] J. Jeon, “NR Wide Bandwidth Operations,” *IEEE Communications Magazine*, vol. 56, no. 3, pp. 42–46, Mar. 2018.

[166] D. Gale and L. Shapley, “College admissions and the stability of marriage,” *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

[167] E. A. Jorswieck, “Stable matchings for resource allocation in wireless networks,” in *International Conference on Digital Signal Processing (DSP)*, Jul. 2011, pp. 1–8.

[168] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, “Matching theory for future wireless networks: fundamentals and applications,” *IEEE Communications Magazine*, vol. 53, no. 5, pp. 52–59, May 2015.

[169] F. Pantisano, M. Bennis, W. Saad, S. Valentin, and M. Debbah, "Matching with externalities for context-aware user-cell association in small cell networks," in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2013, pp. 4483–4488.

[170] T. H. T. Le, N. H. Tran, T. LeAnh, and C. S. Hong, "User matching game in virtualized 5G cellular networks," in *Asia-Pacific Network Operations and Management Symposium (APNOMS)*, Oct. 2016, pp. 1–4.

[171] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (TR 36.213 V14.5.0)," Jan. 2018.

[172] L. Chen, B. Wang, X. Chen, X. Zhang, and D. Yang, "Utility-based resource allocation for mixed traffic in wireless networks," in *IEEE INFOCOM Workshops*, Apr. 2011, pp. 91–96.

[173] A. Abdel-Hadi and C. Clancy, "A utility proportional fairness approach for resource allocation in 4G-LTE," in *International Conference on Computing, Networking and Communications (ICNC)*, Feb. 2014, pp. 1034–1040.

[174] T. Hossfelt, L. Skorin-Kapov, P. E. Heegaard, M. Varela, and K.-T. Chen, "On additive and multiplicative QOS-QOE models for multiple qos parameters," in *ISCA/DEGA Workshop on Perceptual Quality of Systems*, Aug. 2016.

[175] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How Should I Slice My Network?: A Multi-Service Empirical Evaluation of Resource Sharing Efficiency," in *ACM Annual International Conference on Mobile Computing and Networking*, Oct. 2018, pp. 191–206.

[176] B. C. Smith, J. F. Leimkuhler, and R. M. Darrow, "Yield Management at American Airlines," *INFORMS Journal on Applied Analytics*, vol. 22, no. 1, pp. 8–31, 1992.

[177] M. Rothstein, "An Airline Overbooking Model," *Transportation Science*, vol. 5, no. 2, pp. 180–192, 1971.

[178] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization," in *IEEE INFOCOM*, May 2017, pp. 1–9.

[179] B. Urgaonkar, P. Shenoy, and T. Roscoe, "Resource Overbooking and Application Profiling in Shared Hosting Platforms," *SIGOPS Operating Systems Review*, vol. 36, pp. 239–254, Dec. 2002.

[180] J. X. Salvat, L. Zanzi, A. Garcia-Saavedra, V. Sciancalepore, and X. Costa-Perez, "Overbooking Network Slices Through Yield-driven End-to-end Orchestration," in *ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, Dec. 2018, pp. 353–365.

[181] L. Zanzi, V. Sciancalepore, A. Garcia-Saavedra, and X. Costa-Perez, "OVNES: Demonstrating 5G Network Slicing Overbooking on Real Deployments," in *IEEE INFOCOM*, Apr. 2018, pp. 1–2.

[182] L. Zanzi, J. X. Salvat, V. Sciancalepore, A. G. Saavedra, and X. Costa-Perez, "Overbooking Network Slices End-to-End: Implementation and Demonstration," in *ACM SIGCOMM Conference on Posters and Demos*, Aug. 2018, pp. 144–146.

[183] R. Li, Z. Zhao, X. Zhou, J. Palicot, and H. Zhang, "The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 234–240, Jun. 2014.

[184] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, "Citywide Cellular Traffic Prediction Based on Densely Connected Convolutional Neural Networks," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1656–1659, Aug. 2018.

[185] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, "The Learning and Prediction of Application-Level Traffic Data in Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3899–3912, Jun. 2017.

[186] R. W. Thomas, L. A. DaSilva, and A. B. MacKenzie, "Cognitive networks," in *IEEE International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, Nov. 2005, pp. 352–360.

[187] R. W. Thomas, D. H. Friend, L. A. DaSilva, and A. B. Mackenzie, "Cognitive networks: adaptation and learning to achieve end-to-end performance objectives," *IEEE Communications Magazine*, vol. 44, no. 12, pp. 51–57, Dec. 2006.

[188] D. D. Clark, C. Partridge, J. C. Ramming, and J. T. Wroclawski, "A Knowledge Plane for the Internet," in *Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, Aug. 2003, pp. 3–10.

[189] P. Demestichas, G. Dimitrakopoulos, J. Strassner, and D. Bourse, "Introducing reconfigurability and cognitive networks concepts in the wireless world," *IEEE Vehicular Technology Magazine*, vol. 1, no. 2, pp. 32–39, 2006.