Politecnico di Milano
Department of Electronics, Information and
Bio-engineering
Doctoral Programme In 2019

Real Time Trading of Mobile Resources in Beyond-5G Systems

Doctoral Dissertation of:
Özgür Umut Akgül

Supervisors:
Prof. Antonio Capone
Dr. Ilaria Malanchini

Tutor:
Prof. Matteo Cesana

The Chair of the Doctoral Program:
Prof. Barbara Pernici

2019 – 31st Cycle

Acknowledgement

FIRST of all, I would like to express my deepest gratitude to my supervisors, Prof. Antonio Capone and Dr. Ilaria Malanchini, for their continuous support during my PhD study and research, for their patience, enthusiasm, and knowledge. Their guidance has been fundamental during these years.

Beside my supervisors, I would like to thank all the professors and the researchers that have contributed this work with valuable feedback, insightful comments, and suggestions: Dr. Vinay Suryaprakash, Prof. Matteo Cesana, Prof. Nikos Pappas, Prof. Anthony Ephremides, Dr. Markus Gruber, Dr. Qi Liao, Prof. Di Yuan and Dr. Vincenzo Sciancalepore.

My heartfelt thanks go to my friends who provided the necessary encouragement and distraction that made my time more enjoyable. I would like to thank also all the people of the AntLab and the ISPL at Politecnico di Milano for the memorable and enjoyable working environment. I am especially thankful to Davide Sanvito, Sara Mandelli, Edoardo Longo and Andrea Pimpinella for their help during the translation of the thesis.

Last but not least, I am grateful to my family, particularly to my parents, for their continued support of my plans, and their encouragement both before and during my studies.

This work is funded by the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 643002.

List of Figures

List of Tables

Abstract

5G is expected to offer download speeds as high as 1 GBps and latency lower than 1 ms. Although 5G networks are planned to be fully operational by 2020, the unprecedented technical and economic challenges still need to be solved. The biggest problem from a network operator's perspective is the tight profit margins. The lofty expectations from 5G connectivity lead to the need for enormous investments on infrastructure. However, many small operators simply do not possess the necessary revenue in order to deploy the required infrastructure, while the rich operators are unwilling to burden this extreme cost due to the very long return of investment duration. Moreover, 4G technology is reported to reach a fairly close to the Shannon capacity in the available spectrum, and the further improvements on the physical layer are very expensive with respect to the capacity gains. This techno-economic pressure is forcing mobile operators to make pivotal changes in their *modus operandi*. A simple solution is to extent the conventional *infrastructure sharing* agreements to cover the active network components, e.g. radio access network and the spectral resources, and decrease the total costs as well as increasing the spectral efficiency. From a regulatory perspective, the most reasonable scenario is the sharing of the resources brought by a neutral 3rd party, i.e. *infrastructure provider*. However, despite the offered cost efficiency, the conventional sharing approaches rely on well defined service level agreements that cover very long time intervals (e.g. years). However, this static sharing attitude cannot provide the envisioned flexibility and the efficiency in next generation networks.

On the other hand, while the aforesaid techno-economic pressure is forcing the operators to share their networks, the heterogeneity of the service types requires revolutionary changes in the network management. The 5G network is envisioned to host a multitude of services and devices with unique requirements and service priorities. The traditional solution of optimizing the complete network for a particular service type is no longer applicable due to the conflicting requirements posed by different services. A way out is to vertically group network resources, i.e. *slicing the network*, in order to create virtual dedicated networks per service. This way, each resource group (i.e. *slice*) can be customized to serve the respective service in the best possible way. The simplest form of this approach is slicing the network in a static manner, based on some statistical information. However, the conventional network provisioning techniques show that static resource allocation has a tendency towards over-provisioning the network, which causes the inefficient usage of scare spectral resources. Dynamic network slicing can increase efficiency, yet enabling inter-service and inter-tenant priorities in a dynamic negotiation and resource allocation framework is still open in the literature.

In order to address the aforementioned challenges, the main research question in this PhD thesis revolves around how to achieve flexibility and efficiency in a shared mobile network. More specifically, this thesis targets answering the following research questions.

- How can the network resources dynamically and flexibly be shared in a multi-tenant network?
- How can the tenants differentiate their services in a shared infrastructure?
- What are the long and short term implications of anticipatory network sharing and resource trading?

The proposed dynamic negotiation and resource allocation framework proposes a novel service level agreement formulation that allows the operators to renegotiate their shared resources in very short time scales, i.e. in the order of seconds. Moreover, we demonstrate how to exploit anticipatory information regarding the users' achievable rates in order to improve the real time scheduling and resource trading. Lastly, we present a novel self-dimensioning algorithm in order to exploit the short term observations on the traffic demand in order to fine-tune the network capacity. A number of simulations with both synthetic and real data have been performed in order to investigate the characteristic of the proposed framework.

Summary

THE evolution toward 5G and beyond networks brings out novel techno-economic challenges. First and most important of these problems is providing the quality of service expectations with an economically sustainable model. The exponentially increasing broadband demand along with the challenging throughput and delay constraints requires pivotal changes in the network infrastructure. However, the total cost of network upgrade further tightens an already condensed profit margin of the mobile operators and turns the network provisioning into a non-profitable business model. In order to decrease the total costs, a possible way is to extend the infrastructure sharing agreements to include active network components as well as the spectral resources. However, as the number of shared components in the wireless network increases, the risk that individual operators take also increases. Moreover, in order for such a static sharing agreement to be effective, the operators have to have a very good estimation of their current resource needs as well as the evolution of this needs in a relatively long period of time (e.g. a year). Therefore, the static approaches nearly always ends in over-provisioning of the valuable spectral resources. Dynamic infrastructure sharing can increase the resource efficiency as well as further decrease the costs. In order to achieve cost reduction without loss of business potential, all parties in the sharing agreement must be able to renegotiate their resource shares within very short time windows (e.g. hours or days).

Another big challenge in 5G and beyond networks is the heterogeneity of the traffic requirements. Unlike predecessor technologies, 5G is en-

visioned to contain a multitude of industry driven services with specific requirements and priorities. The conventional method to handle these requirements, i.e. optimizing the network resources in line with the quality of service expectations, is no longer an option due to the diverse performance indicators. A way to handle this heterogeneity is to vertically group the network resources (i.e. slicing the network) and optimize the separate groups (i.e. slices) in line with the requirements. Similar to aforementioned problem with the infrastructure sharing, static network slicing requires the mobile operators to have a very good estimation on their needs. However, being an industry driven wireless technology, 5G requires a great level of flexibility in resource allocations and network configurations in order to adapt to the evolution of the traffic conditions and the service needs. A possible way to provide adaptability to the changing conditions (both in terms of the demand and the service type) is dynamically slicing the network resources based on some slice templates. On the other hand, dynamically slicing the network makes the different services to be connected in terms of the resource allocations which can easily lead service level agreement violations. Moreover, in order to provide a robust business ecosystem, a key attribute is to guarantee resource availability. Therefore, the network capacity has to be scaled in line with the evolving needs of the network traffic. On the other hand, the heterogeneous traffic requirements and priority make it harder to compare the urgency of the capacity expansion need among two different regions. Consequently, the conventional capacity management strategies are required to be revisited in order to fit the changing dynamics in the network provisioning.

In this thesis, we have focused on the aforementioned need for a flexible and efficient network management and proposed a novel dynamic negotiation and resource sharing framework for sliced networks. In order to eliminate the over-restrictive structure of the conventional service level agreements, in our model we propose a novel approach where the operators only define their slice types, their utility expectations and their budgets in order to reach these expectations. The rest of the negotiations are automatically handled using a set of parameters that are introduced by the tenants. The minimization of the human-based negotiations allows us to repeat the inter-tenant renegotiations with a very high frequency and dynamically update the operators' resource shares in line with their objectives and the instantaneous traffic conditions. Since the determination of the resource shares per tenant while simultaneously allocating the resources is quite challenging, we have logically separated our model into two sub-problems, where in one of them we determine the resource allocations based on the predefined resource shares and in the other one, we determine the optimal shares per operator based on the observed traffic conditions in the past (i.e. the traffic mixture and the achievable rate per user). Moreover, in order to extend the network resources in line with the demand, we propose a novel pricing strategy that maps the microeconomics' law of supply and demand. The simulation results show that the proposed model increases the cost efficiency while providing an adaptive network management framework.

Next, we have extended the proposed framework to be able to handle heterogeneous traffic requirements. As a first step we have defined a novel piece-wise linear utility function which can be customized according to the service type. Although the proposed utility function is designed to be scaled according to the achieved rate of services, the delay constraint of services are integrated to the proposed scheduler. Through a large set of Monte Carlo simulations, we show how the tenants can differentiate their services. Moreover, we also explain how to control the interconnection between different slices by adapting the slope of the utility functions and how the tenants can exploit this interconnection to maximize their spectral efficiency.

Finally, we focus on the anticipatory network slicing and resource sharing. Nowadays, the research on artificial intelligence provides a large variety of prediction tools with reasonable accuracy levels. As a consequence, the researchers can anticipate the variations in the traffic conditions (i.e. both channel conditions and traffic demand), which gives them the unique opportunity to increase the network efficiency and flexibility by using predicted information to guide the resource allocations. Therefore, we have focused on how to exploit predicted data regarding the upcoming channel conditions. However, since these predictions are required to be done within very short time intervals, the accuracy of the prediction data is low. Consequently, we have proposed a novel filtering mechanism that can exploit the accurate predictions and eliminate the effects of inaccurate information. Our results firstly prove that the proposed filtering mechanism can minimize the impact of inaccurate prediction. Moreover, we show that the efficiency of the network sharing approach can be maximized through anticipating the upcoming channel conditions. Lastly, we have analyzed the long term impacts of anticipatory networking on the evolution of the network infrastructure. Due to the techno-economic constraints, the transition to 5G is envisioned to be distributed over time, making deployments with the rising need. However, the envisioned business ecosystem imposed by multi-tenancy as well as the large variety of traffic needs require a new approach in network capacity planning. Therefore, we proposed a novel slice-aware capacity expansion strategy that can provide efficient deployment decisions. Moreover, our investigations prove that the propose algorithm does not require long term observations on the network status, and can be used in shorter observation windows.

Abstract (Italiano)

SECONDO le previsioni, il 5G offrirà velocità in download fino a 1 GBps e latenza al di sotto di 1ms. Sebbene sia previsto che le reti 5G siano completamente operative entro il 2020, ci sono sfide tecniche ed economiche ancora da risolvere. Il problema più grande dal punto di vista dell'operatore di rete è il ristretto margine di profitto. Le aspettative elevate riguardo la connettività 5G richiedono un enorme investimento nelle infrastrutture. Tuttavia, i piccoli operatori non possiedono le risorse sufficienti a sviluppare le infrastrutture richieste e allo stesso tempo i grandi operatori non vogliono prendersi carico di questi ingenti costi a causa della lunga durata del ritorno sugli investimenti. Inoltre, è previsto che la tecnologia 4G raggiungerà il limite della capacità di Shannon nello spettro, e sviluppare ulteriori migliorie nel livello fisico prevede costi onerosi rispetto ai corrispondenti guadagni in capacità. Questa pressione tecno-economica sta portando gli operatori mobili a cambiamenti cruciali nel loro modus operandi. La soluzione più semplice sarebbe quella di estendere gli accordi convenzionali sulla condivisione dell'infrastruttura al fine di includere i componenti di rete attivi, ad esempio la rete d'accesso e le risorse dello spettro, in modo da ridurre i costi totali e aumentare l'efficienza spettrale. Secondo la regolamentazione, lo scenario più ragionevole sarebbe quello di condividere risorse acquistate da una terza parte neutrale, come un fornitore dell'infrastruttura. Tuttavia, nonostante la potenziale efficienza nei costi, gli approcci convenzionali di condivisione prevedono accordi sul livello del servizio (*Service Level Agreements*, SLA) ben definiti che coprono intervalli temporali molto larghi, nell'ordine degli anni. Tuttavia, una condivisione

statica di questo tipo non è in grado di garantire la flessibilità e l'efficienza richieste dalle reti di nuova generazione.

D'altro canto, mentre la pressione tecno-economica sopracitata sta forzando gli operatori a condividere le loro reti, l'eterogeneità dei servizi richiede cambiamenti rivoluzionari nella gestione stessa della rete. È previsto che il 5G ospiterà una moltitudine di servizi e di dispositivi con requisiti e priorità uniche. L'approccio tradizionale per ottimizzare l'intera rete in base al tipo di servizio non è più utilizzabile a causa dei requisiti contrastanti imposti dai diversi servizi. Una possibile soluzione consiste nel raggruppare verticalmente le risorse di rete, ad esempio adottando il network slicing, per creare reti virtuali dedicate per ciascun servizio. In questo modo, ogni gruppo di risorse (lo *slice*) può essere configurato su misura per servire nel miglior modo possibile il servizio corrispondente. L'approcio più banale è adottare il network slicing di tipo statico, basato su alcune informazioni statistiche. Tuttavia, adottando le tecniche convenzionali di fornitura della rete l'allocazione statica delle risorse tende a richiedere il sovradimensionamento della rete, portando ad un uso inefficiente delle già scarse risorse spettrali. Il dynamic network slicing può aumentare l'efficienza, ma un framework che permetta la negoziazione dinamica di priorità tra i servizi e tra gli utenti e l'allocazione delle risorse è ancora mancante in letteratura.

Per poter affrontare le sfide sopra descritte, il problema principale affrontato in questa tesi di dottorato è quello di ottenere flessibilità ed efficienza in una rete mobile condivisa. Nello specifico, questa tesi risponde alle seguenti domande:

- Come è possibile condividere le risorse di rete in modo dinamico e flessibile in una rete multi-tenant?
- Come possono i diversi tenants differenziare i proprio servizi in un'infrastruttura condivisa?
- Quali sono le implicazioni nel lungo e nel breve termine di una condivisione della rete e dello scambio di risorse basato sulla predizione?

Il framework per la negoziazione e l'allocazione dinamica delle risorse qui presentato propone una nuova formulazione degli accordi sul livello del servizio che permette agli operatori di rinegoziare le loro risorse condivise in un tempo molto ristretto, nell'ordine dei secondi. Inoltre, dimostriamo come sfruttare la disponibilità a priori dell'informazione sulle velocità raggiungibili dagli utenti per migliorare lo scheduling in tempo reale e lo scambio delle risorse. Infine, presentiamo un nuovo algoritmo per sfruttare l'osservazione nel breve periodo delle richieste di traffico al fine di correggere l'assegnamento della capacità di rete. Abbiamo eseguito diverse simulazioni sia con dati sintetici che reali per investigare le caratteristiche del framework proposto.

Summary (Italiano)

L'evoluzione verso il 5G e i suoi futuri sviluppi porta alla luce nuove sfide tecno-economiche. La prima di queste sfide consiste nel soddisfare le aspettative sulla qualità del servizio con un modello economicamente sostenibile. La richiesta di banda larga, che sta crescendo in modo esponenziale, insieme ad impegnativi vincoli su throughput e tempi di ritardo, necessita di cambiamenti cruciali nell'infrastruttura di rete. Tuttavia, il costo totale dell'aggiornamento della rete limita ulteriormente un margine di profitto già ridotto per gli operatori mobili e trasforma la fornitura di rete in un modello di business non redditizio. Al fine di ridurre i costi totali, una strada possibile è estendere gli accordi di condivisione dell'infrastruttura per includere i componenti di rete attivi e le risorse spettrali. Purtroppo, con l'aumentare del numero di componenti condivisi nella rete wireless, aumenta anche il rischio per i singoli operatori. Inoltre, affinché un tale accordo di condivisione statica sia efficace, gli operatori devono avere una stima accurata delle loro esigenze attuali in termini di risorse e dell'evoluzione di tali esigenze in un periodo di tempo relativamente lungo (ad esempio un anno). Pertanto, gli approcci statici finiscono quasi sempre in una sovra-fornitura di preziose risorse spettrali. La condivisione dinamica dell'infrastruttura può aumentare l'efficienza delle risorse e ridurre ulteriormente i costi. Al fine di ottenere una riduzione dei costi senza perdita di potenziale commerciale, tutte le parti nell'accordo di condivisione devono essere in grado di rinegoziare la loro fetta di risorse in tempi brevissimi (ad esempio ore o giorni).

Un'altra grande sfida del 5G e dei suoi futuri sviluppi è l'eterogenei-

tà del traffico. A differenza dei suoi predecessori, il 5G comprende una moltitudine di dispositivi industriali che necessitano requisiti specifici e diverse priorità. Il metodo classico per soddisfare questi requisiti, ad esempio ottimizzando le risorse di rete in linea con le aspettative di qualità del servizio, non è più un'opzione a causa di diversi indicatori di performance. Un modo per gestire questa eterogeneità è raggruppare verticalmente le risorse (*network slicing*) e ottimizzare i gruppi (*slices*) in base ai requisiti. Simile al problema spiegato in precedenza di condivisione delle infrastrutture, lo *static network slicing* richiede agli operatori mobili di avere una stima accurata dei loro bisogni. Ciò nonostante, visto il focus industriale, il 5G richiede un alto livello di flessibilità sia nell'allocazione di risorse che nella configurazione di rete in modo da adattarsi all'evoluzione del traffico e ai bisogni del servizio. Un modo possibile per fornire adattabilità ai cambiamenti (sia nei termini di domanda e tipo di servizio) è sfruttare lo slicing dinamico delle risorse di rete in base a template già noti. Questo può causare violazioni di accordi sul livello di servizio (SLA) a causa dei diversi servizi connessi. Inoltre, per poter fornire un ecosistema di business robusto, è necessario garantire la disponibilità delle risorse. Quindi la capacità di rete deve essere scalata in base alle variazioni del traffico. D'altro canto, l'eterogeneità dei requisiti e della priorità del traffico rendono complicato confrontare in due regioni diverse la capacità di espansione in base all'urgenza. Di conseguenza, le strategie note di capacity management hanno bisogno di essere revisionate in modo da adattarsi ai cambi dinamici del network provisioning.

In questa tesi ci siamo concentrati sui problemi descritti in precedenza per una gestione flessibile e efficiente della rete. Inoltre, viene proposta un'innovativa negoziazione dinamica e un framework di condivisione delle risorse per il network slicing. In modo da eliminare la struttura troppo restrittiva degli accordi sul livello di servizio attuali, proponiamo un approccio dove gli operatori definiscono i tipi di slice, le aspettative e il budget. Il resto della negoziazione viene gestita automaticamente utilizzando una serie di parametri che sono introdotti dai tenant. La minimizzazione del fattore umano nella negoziazione ci permette di ripetere frequentemente la rinegoziazione tra tenant. Inoltre, permette di aggiornare dinamicamente la condivisione delle risorse degli operatori in base ai loro obiettivi e alle condizioni di traffico istantanee.

Visto il bisogno crescente di flessibilità ed efficienza nella gestione della rete, in questo lavoro si propone un modello innovativo di negoziazione dinamica e di condivisione delle risorse per reti *sliced*. Per eliminare la struttura iper-restrittiva dello SLA convenzionale, nel nostro modello gli operatori definiscono solamente il tipo di *slice* desiderato, il modo con cui prevedono che venga utilizzato e le risorse economiche disponibili per soddisfare tale previsione. Successivamente, il resto delle negoziazioni saranno gestite automaticamente tramite un gruppo di parametri introdotto dagli utilizzatori. La minimizzazione dell'interazione umana nel processo di negoziazione delle risorse ne permette la rapida e frequente ridefinizione e di conseguenza garantisce la possibilità di aggiornare dinamicamente la loro associazione rispetto alle reali necessità degli operatori e alle condizioni istantanee del traffico. Dato che la contemporaneità dei processi di determinazione della quantità di risorse per operatore e la loro effettiva allocazione rappresenta una sfida complessa, abbiamo deciso di suddividere il problema in due rami logicamente distinti: nel primo si determina l'allocazione delle risorse sulla base delle richieste predefinite dagli operatori, mentre nel secondo si determina l'allocazione delle risorse ottima rispetto alle condizioni del traffico osservate nel passato (cioè, la tipologia di traffico e il rate medio ottenibile dall'utente). Inoltre, al fine di offrire la giusta quantità di risorse rispetto alla domanda, nel nostro modello proponiamo una strategia di pricing che rispecchia le leggi microeconomiche su domanda e offerta. I risultati delle simulazioni dimostrano che il modello proposto aumenta l'efficienza della rete in termini di costi e allo stesso tempo fornisce uno strumento flessibile ed adattativo per la gestione delle risorse di rete.

Successivamente, abbiamo esteso il modello per poter gestire requisiti relativi a tipologie eterogenee di traffico. IInnanzitutto, abbiamo definito una nuova funzione lineare di utilità che può essere composta in maniera consistente ai tipi di servizio considerati. Sebbene la funzione di utilità proposta sia strutturata in maniera proporzionale al rate raggiungibile dai servizi considerati, i vincoli di ritardo dei servizi stessi sono integrati nello scheduler proposto. Tramite svariate simulazioni di tipo Monte Carlo, abbiamo mostrato come gli utilizzatori possono differenziare la loro offerta di servizi. Inoltre abbiamo presentato, da un lato come sarà possibile per gli operatori controllare la connessione tra le *slice* adattando la pendenza della funzione di utilità, e dall'altro come gli operatori potranno sfruttare tale connessione per massimizzare l'efficienza spettrale.

Infine, lo studio considera il problema dell'anticipazione del network slicing e della condivisione delle risorse. Ad oggi, la ricerca in materia di intelligenza artificiale fornisce un vasto panorama di strumenti di predizione con buoni livelli di accuratezza. È quindi possibile prevedere le variazioni delle condizioni di traffico (sia relativamente al canale sia in termini di richiesta di traffico), opportunità cruciale per aumentare l'efficienza e la flessibilità della rete tramite l'utilizzo dell'informazione predetta, con lo scopo ultimo di guidare il processo di allocazione delle risorse. In questo contesto, abbiamo esplorato le possibilità di sfruttare la predizione delle informazioni circa le condizioni del canale di comunicazione. Tuttavia, visti i vincoli stringenti in termini di tempo di elaborazione delle suddette informazioni affinchè esse siano effettivamente utilizzabili, l'accuratezza delle predizioni è bassa. Di conseguenza, abbiamo proposto un meccanismo innovativo di filtraggio che valorizza le predizioni accurate ed elimina gli effetti delle informazioni inaccurate. In primo luogo, il nostro lavoro dimostra che il meccanismo di filtraggio minimizza l'impatto delle predizioni inaccurate. Inoltre, dimostriamo che allo stesso tempo l'efficienza del processo di condivisione può essere massimizzata predicendo le condizioni del canale. In conclusione, abbiamo analizzato l'impatto a lungo termine delle reti anticipatorie rispetto all'evoluzione dell'infrastruttura di rete. Infatti, a causa dei vincoli tecno-economici, va ricordato che, se da un lato la transizione al 5G è concepita in maniera distribuita nel tempo con il crescere delle effettive necessità, l'ecosistema di business imposto dal mercato concorrenziale e la grande varietà di richieste di traffico richiedono un approccio innovativo in termini di pianificazione di rete. Perciò, proponiamo una nuova strategia di espansione della capacità di rete che sfrutti i vantaggi del network slicing e fornisca strumenti decisionali efficienti relativamente a nuovi deployment. Il nostro studio dimostra che l'algoritmo proposto non richiede osservazioni a lungo termine dello stato della rete, ma può essere utilizzato in finestre di osservazioni a breve termine.

Contents

CHAPTER 1

Introduction

The emerging 5G technology challenges the network operators with a number of economical, technical and regulatory aspects. On one hand, the explosive growth of the mobile network demand requires a high investment from the network operators in order to scale the network capacity in line with the demand. On the other hand, the increase in the network capacity is not always followed with an equivalent increase in the revenues – in some cases, there is even a decrease in the actual revenue of the tenant [66]. For example, due to the regulatory pressure, the network operators are required to provide coverage in some rural areas that are not attractive from a business perspective. According to a recent estimation, 50% of the rural deployment, contributes to less than 10% of the network operator's revenue [57]. This imbalance between the actual revenue and the total expenditure causes an increasing incentive to minimize the total costs and perform accurate investments on infrastructure expansion decisions. The conventional network management strategy relies on worst case scenario solutions that usually end up over-provisioning the resources. This inefficient resource management technique places more strains on the business model of the network operators. Therefore, the network operators' current *modus operandi* has to change in order to suit the shifting economic dynamics of the network provisioning. A possible way out is increasing the flexibility of the business model via sharing the network resources among mobile operators. In this new business approach, the conventional network operators act as mobile virtual network operators (or *tenants*) who buy network resources from *infrastructure providers*. The instantaneous availability of resources allows tenants to scale their networks in line with their needs while focusing on providing the best service to their customers. As a result of sharing the available resources, the total costs, e.g. deployment and operational costs (i.e. CapEx and OpEx, respectively), are split among the tenants. Nevertheless, a major question is how to share the infrastructure resources. The conventional sharing agreements favor static resource sharing where each tenant's resource share as well as their respective costs are determined a priori, usually after very long negotiations. However, it is clear that this approach cannot decrease the need of over-provisioning, as the tenants cannot easily renegotiate their resource share in order to fulfill the dynamic traffic demand. A possible solution, which became available quite recently thanks to the advancements in the virtualization technology, is to dynamically share the resources among tenants. In order for dynamic network sharing to be profitable, the economical implications of the renegotiation as well as the resource allocations, have to be modeled.

The stringent quality of service (QoS) expectations and the heterogeneity of the network services make it harder to serve a multitude of services within the same network infrastructure. The necessity of sharing the infrastructure resources among different devices and services results in the challenge of achieving a harmonious coexistence of a broad range of competing priorities and QoS expectations within the same network. However, supporting these services and achieving the maximum quality of experience (QoE) require the network resources to be customized. The conventional 'one type fits all' strategy simply cannot provide this customization which is a key for harmonious coexistence of all services. The advances in network virtualization technology bring a candidate solution, i.e. *Network Slicing*. Despite the lack of consensus on the definition of network slicing and how it should be handled, a common definition is dividing the available network resources into subgroups that can be separately optimized in order to fully serve to a particular service type [11]. Thus, it proposes dedicated virtual networks that can be customized in line with the needs of each service. A fundamental question, which is also the main focus of this thesis, is how to perform the resource allocations per slice. The simplest and the most straight-forward approach, which naturally comes as a first possible implementation of network slicing, is to slice the network resources statically based on very well-defined service level agreements (SLAs). Despite being the simplest approach, it requires the network operators to have a very good estimation of the traffic mixture and the required resources as well as the evolution of the demand over time. Even with a perfect prediction of the traffic evolution, which can be quite a challenge by itself, such an approach overly restricts the operators as it further increases the time to market of new services. An alternative approach that can provide the flexibility, is to dynamically assign network resources to slices in line with the instantaneous network traffic conditions (i.e. traffic demand, the service mix and the channel conditions) and the long term strategy of the network operator. Such a dynamic approach can provide not only the required resource efficiency but also shorter time-to-market duration for newly defined services. However, it also brings novel challenges in resource allocations among competing services with different priorities and QoS requirements such as achieving SLA guarantees, enabling fairness among services and efficiently assigning resources to the slices.

The joint application of two solutions (i.e. infrastructure sharing and network slicing) brings a new layer of complexity, since providing network slicing in a shared infrastructure, *multi-tenant network*, is much harder than in single operator network due to the new level of priority over the services added by the tenants. In this PhD research, we tackle the challenge of providing such a framework where the tenants can dynamically negotiate over their resource shares and the outcomes of the negotiations are implemented by considering the inter-slice priorities. In line with the "zero-touch network management" concept in the industry, in this thesis, we searched for a way to automatize these two processes, i.e. the negotiations among tenants and the real time resource allocations. We argue that the service level agreements between tenants and the infrastructure provider have to be simplified in a way that it would only provide QoS requirements per tenants and QoE model (namely utility functions) per user as well as unit costs that can be dynamically scaled in line with the actual resource usage. Consequently, this PhD thesis revolves around the question of how to automatize the negotiations between stakeholders (i.e. tenants and the infrastructure provider) and the resource allocation process with the primary goal of maximizing the overall network utility and cost reduction per tenant.

In the following part of this chapter, i.e. Section 1.1, we provide a deeper understanding on the concept of infrastructure sharing. Following the background information, the main research questions focused in this thesis are presented in Section 1.2. Finally, the outline of the following chapters along with the respective contributions are presented in Section 1.3.

1.1 Background

1.1.1 What are the key expectations in 5G?

The conventional network management models are facing revolutionary changes (cf. Fig. 1.1) that are rising from the idea of conneting everthing [79]. Thus, having a good understanding of the expectations of 5G is imperative in order to propose an applicable and durable framework. On one hand despite the huge number of services and devices envisioned to be served in 5G, the customer expectations are pretty much homogeneous and similar to the expectations from the previous technology, namely higher data rates, shorter delays and cheaper services. On the other hand, the major impacts of the 5G is expected to be seen for the network operators. As reported in [46], the key differences between 5G and 4G are on $1)$ achieving higher data rates, $2)$ handling a large volume of traffic, $3)$ providing higher reliability, $4)$ support for higher user speeds, $5)$ achieving lower latency, $6)$ enabling low power communications and $7)$ multi-connectivity. These expectations bring novel challenges that force the network operators to revisit their business and technical models. First of all, a key challenge in the urban network planning is the saturation of the available spectrum resources. The current coding schemes already produce a close approaximization to the achievable spectral efficiency (i.e. up to approximately 80% of the Shannon capacity) [75]. However, from a business perspective, further increasing the spectral efficiency is economically not profitable as the required technology is more expensive than the revenue obtained from the additional spectral resources [1]. A simpler solution to respond the increasing demand on spectral resource is inevitably building new base stations, i.e. densification of the network components [34] [9]. On the other hand, this requires more complicated management mechanisms in order to dynamically manage and synchronize all the base stations.

Another key challenge for the network providers, which is mostly imposed by the Industry 4.0 initiative, is to reduce the required time for network operations, which leads to the minimization of the number of operations that are controlled by human [13]. 5G and beyond networks provide service in a highly dynamic network market. Thus, minimization of the time to market is of uttermost importance. The time to market (TTM) in this context is the time duration between the request for a new service and the provisioning of this service to the end user. It is clear that increasing human involvement in the dynamic resource management decisions increases the TTM [21]. Moreover, the minimization of the human factor in

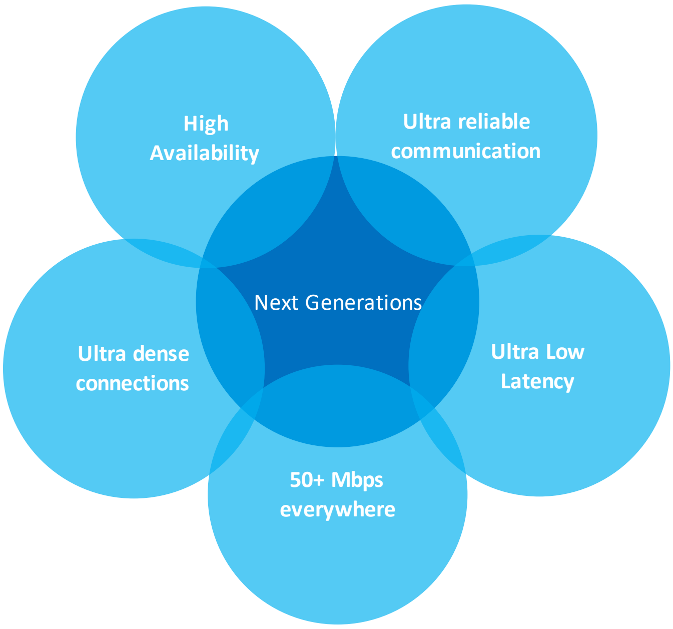


Figure 1.1: *Key challenges in next generation wireless technology.*

the network management is expected to minimize the operational costs in parallel [21]. Last but not least, one of the key attributes of next generation wireless market is the heterogeneity of the business models for the network operators. Since there are multiple services, instead of serving all the services, the mobile operators can specialize in specific portions of the mobile market and provide dedicated services. Therefore, it is crucial for the next generation networks not to pose difficulties to the new entrants as well as supporting specialized tenants [79]. As the mobile operators are increasingly challenged by the economic pressure caused by the technical needs, inter-operator network sharing has emerged as a solution to decrease the costs of network provisioning while providing a flexible and scalable business platform [74].

1.1.2 Why network sharing?

Even though the answer of whether or not to share the network resources is often obvious, the main drivers of network sharing can vary a lot from decreasing the total costs to enhancing existing services or entering a new market [66]. As the main drivers of the network sharing, cost reduction and the capability to scale the network resources in line with the need are first to investigate. From the network operator's point of view, it is essential to dynamically update their infrastructure and business models in line with the changes in the technology. Such updates usually require fundamental changes in the physical topology of the network that lead to enormous CapEx cost. Consequently, as majority of the small mobile operators cannot compensate such big investments with no clear return of investment expectations, the networks are gradually dominated by a small number of big operators. Network sharing guarantees that the network resources are always up to date and the needed technological capacity from an operator can always be met. Consequently network sharing can bring the competition back as the smaller operators can compete against the big ones in terms of their services.

Another key contribution of network sharing is lowering the regulatory pressure on the network operators. In the conventional network provisioning model, the network operators are required to obtain a multitude of regulatory approvals, which causes a serious legal burden. In network sharing, the infrastructure provider handles a multitude of regulatory requirements while the tenants (i.e. mobile virtual network operators) mainly focus on the their service provisioning [97]. Consequently, the network sharing provides predictability and efficiency in the business model of the tenants.

Lastly, enabling flexibility and efficiency in achieving QoS guarantees is another key motivations of network sharing. In the conventional operations, over-provisioning the network resources is the widely applied way to achieve quality of service guarantees. However, over-provisioning decreases the resource efficiency as most of the resources are not actively used other than the peak traffic hours. The immediate availability of network resources gives the tenants a chance to dynamically obtain further resources and allow them to achieve their QoS guarantees with the minimum amount of resources, lower costs and higher spectral efficiency [91].

As an indication of the increasing attention to the idea of network sharing, in 2013, Radio Spectrum Policy Group (RSPG) stated “To meet the growing demand for spectrum, industry and administrations are under pressure to introduce new technologies and regulatory mechanisms to optimize the use of the limited frequency resources” [97]. Indeed, through network sharing the total costs can be decreased, TTM can be shortened and the regulatory pressure can be ceased. Consequently, the key question of whether or not to share the network quickly transforms into how to share which type of resources.

1.1.3 The main sharing approaches in the state of the art

Although various sharing options are shaped based on the long term expectations and short term constraints of the sharing parties, the sharing agreements in the literature can be grouped under three major models [59] [84], i.e.

- *Business Model*: describes the parties involved in the sharing agreement, e.g. network operators or infrastructure providers, and the agreements between these parties
- *Geographic Model*: identifies the physical footprint of individual network operators
- *Technology Model*: describes the technical solutions to enable sharing

From the business perspective the sharing always occurs among network operators with or without an infrastructure provider (not using resources itself) [66]. In a shared network, providing security of confidential business information or customer specific traffic data can be a major challenge. Therefore, among these two scenarios, the neutrality of the outsourcing, e.g. infrastructure provider, can be a better option to preserve sensitive data and bring trust to the sharing process [28]. The involvement of infrastructure provider would be especially effective as the regulatory approval could be obtained more easily.

Secondly, the geographic sharing options can be investigated under four major cases, i.e. full split, unilateral shared region, common shared region and full sharing [33]. The full split sharing option is the case in which the network operators are covering complementary areas and want to involve in sharing to increase their coverage ratios. When one of the operator does not have the necessary infrastructure to perform full coverage, the network operators can involve in a unilateral sharing agreement in which the operators can use others infrastructure to provide full coverage to its customers. In the common shared region model, the network operators that have their own infrastructures make an agreement to install new infrastructure to the region that is not covered by both of them. The key points in this sharing model are that the network operators have their own infrastructure in the region but not in specific areas and they build joint infrastructure in order to provide full coverage. The final geographical sharing model is the full sharing in which the network operators share all the base stations in the coverage area. In this thesis, we focus on the full sharing scenario as it poses the biggest challenge of sharing the available resources among entities with equivalent right to obtain them.

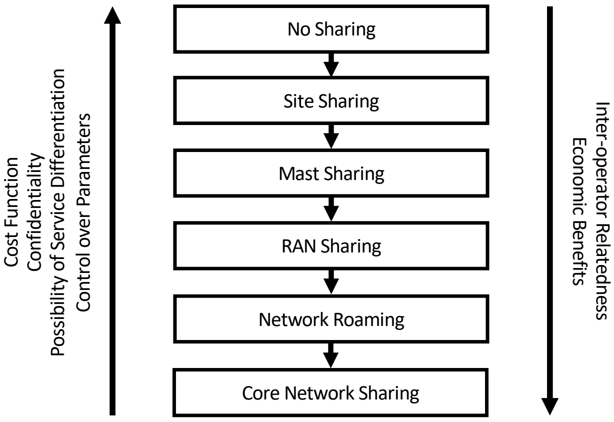


Figure 1.2: *A comparison between different sharing approaches where the direction of the arrow indicates an increase.*

Lastly, from a technological perspective, the sharing agreements can be determined according to the shared entity. As the name suggests, the passive sharing includes the sharing of passive entities in the network such as site or mast sharing. In this type of sharing, the economic impact is usually lower than other types of sharing, as the tenants only share the deployment cost in the region. However, it also gives tenants the maximum control over the resources. In contrast to the passive sharing, active sharing can provide the maximum cost efficiency as the tenants theoretically share all the network components, e.g. RAN sharing or core network sharing. National roaming, where one tenant completely relies on another one's infrastructure for a given geographical location, is also considered as an active sharing agreement. Although, active sharing is considered to be the most cost efficient sharing approach by the OECD report [66] and the references therein, as the depth in sharing increases the tenant's control over the resources decreases. This reverse proportionality is also presented in Fig. 1.2.

Although the envisioned cost reduction and the possibility to have control over the resources are two key aspects to determine the sharing option, another critical variable is the maturity level of the market which is measured by the driving aspect of the wireless market, i.e. either capacity or coverage [33]. In mature markets, where the maintaining coverage phase is completed and the network operators mainly focuses on the network capacity to differentiate their services (in other words capacity-driven networks), the tenants differentiate their services mostly by their capacity and network services, thus for most of the cases, RAN sharing is the best option. On the other hand, in developing markets, where the main competition is mostly on covering as much area as possible, the depth of sharing can be increased since the tenants mostly provide similar services.

The decision of involving in a sharing agreement and the depth of sharing agreement produce the trade-off depicted in Fig. 1.2. As previously mentioned, one of the key challenges in 5G network is achieving an economically efficient network evolution which can be achieved with the maximum sharing depth – i.e. core network sharing. However, especially in a developed wireless market, such a sharing model can result in undifferentiated services among tenants which would bring a huge business risk [53]. A general approach of network operators is to envolve in strong partnership and cooperation agreements for rural areas that usually have least business attraction but needs to be covered due to regulations. It is generally recognized that core network sharing may not be a feasible solution as any service or function that one operator implements can be replicated by the others as they have the same coverage areas and quality of service (QoS) [33]. Based on this trade-off between inter-relatedness of tenant's business applications and cost efficiency, RAN sharing is considered to be the most viable option [17].

1.1.4 Revenue modeling

The applicability of any technological model mainly depends on its economical implications. Therefore, as a first step, the governing economical dynamics are required to be outlined. With this objective in mind, in this subsection, we review the major parts of the cost model. In a classical wireless network topology, due to the well-separated business models of the operators, the per user cost of each tenant does not change with the number of network operators. However, sharing the cost of available network resources decreases the total cost per operator with the total number of operators in the sharing agreement. In a shared infrastructure, in order for the infrastructure provider to set a sustainable platform, the tenants have to compensate the operational expenses of the infrastructure provider as well as a fraction of the infrastructure cost. On the other hand, if the total cost of a tenant is higher than the actual ownership price of a network, it may not be worth to share the network. Therefore the designed pricing mechanism is the most critical aspect for the sharing platform (further insights regarding the impact of pricing on tenant's sharing incentive can be found in [14]).

A detailed cost model is depicted in [84] where the authors also analyze the variations of cost function for different sharing options. The overall cost metrics are modeled under three main categories, i.e. equipment cost, operation cost and transaction cost. In the considered model, the equipment cost parameters are composed of the deployment cost of a network, such as land-lease cost, tower construction cost, equipment cost and spectrum license fee. Being a dynamic metric, the operation cost covers all the direct and indirect costs such as maintenance, planning and R&D etc. Finally, transaction cost is the regulation and connection based costs. If the tenants cover equivalent market portions and have similar economic constraints, we can assume that the tenants equally split the total cost. On the other hand, considering the fact that the tenants can possess different strategies and goals along with different limitations (both in terms of economic and technical), the inaccuracy of this assumption is clear.

In order to provide a sharing agreement that can achieve fairness and efficiency, the total cost per tenant has to be scaled according to the actual resource consumption of this tenant. The simplest way is determining a unit cost per resource and simply multiplying it with the tenant's resource consumption. However, such an approach is rather open to the manipulations of the tenants, which can simply lead to monopolization of the network resources. Therefore, the cost model has to dynamically set the prices of resources while preventing any type of malicious attempt from the tenants.

1.2 Research questions

Despite the economical gains achieved by network sharing, the applicability of it highly depends on providing a sharing platform that can both support inter-tenant service differentiation and maintain fairness and flexibility in resource sharing. Thus, the research activities in this thesis revolve around the design and analysis of a dynamic negotiation and resource sharing platform which would exploit the full potential of infrastructure sharing. The investigations in an overall context include a multitude of aspects: technical and cost performance of the platform, profitability and the sustainability of sharing and service differentiation of the tenants as a result of the business roles that are taken. The analyses performed in the thesis are from a tenant's perspective, namely, the tenants can differentiate their policies and their strategies in accordance with their long term goals. The infrastructure provider is considered to be a zero-profit entity and reinvests all the collected revenues. The research activities in the following sections are focusing on answering three main questions.

> *RQ1. How can the network resources dynamically and flexibly be shared in a multi-tenant network?*

Dynamic sharing of network relies heavily on automatizing the negotiations between tenants and the infrastructure provider. Therefore, economic (such as pricing of the resources, budget per tenant etc.) and technical (e.g. real time resource allocations, expected and achieved QoS etc.) interactions between different stakeholders have to be modeled jointly in the model. Depending on the economical gains and the technical suitability of the framework to the market positions and the long term goals of the operators, the convenience of dynamic sharing can be determined.

> *RQ2. How can the tenants differentiate their services in a shared infrastructure?*

A key aspect of sustainability in a sharing platform is supporting service differentiation among tenants as well as customizing the network resources in order to serve various types of services, e.g. video streaming, calling, text messaging, etc. Therefore, one can naturally ask how the tenants can differentiate their services in order to preserve their business value. The answer to this question will be reached through analysis of sharing parameters that can be used in order to differentiate the services of individual tenants, and the key enabling technology for service differentiation, i.e. network slicing. A key aspect is to model the QoS expectations of different and competing services into a homogeneous and inter-service comparable QoE metric. Modeling different services as well as finding a common objective are the main challenges for this problem.

> *RQ3. What are the long and short term implications of anticipatory network sharing and resource trading?*

This research question focuses on the long and short term effects of network sharing. The research questions revolve around how the short time scheduling and trading decisions can be improved by using some anticipatory information, and how the short term observations can be utilized in order to appropriately scale the network resources in long term.

An investigation of the network evolution in relatively long time instances becomes imperative as well as the possible improvements in short time scale negotiations. The anticipatory information regarding the upcoming time slots as well as the traffic mix and demand can give particular advantages to the tenants in their negotiations as well as the infrastructure provider to increase the spectral efficiency and the network capacity.

1.3 Thesis outline and the contributions

Chapter 2: Review of the state of the art

In Chapter 2, related work on the state of the art is presented in order to accurately position the outcomes of the PhD work in the literature. For the sake of readability, we have separated the related works on infrastructure sharing, network slicing, anticipatory network management and capacity management.

Chapter 3: Dynamic network sharing in a multi-tenant network

In Chapter 3.5 we propose a fundamental sharing platform that can automatize the negotiations between stakeholders in order to be repeated within short time scales and adapt the real-time resource allocations based on the outcome of the negotiations. The demand-supply dynamic in microeconomics is modeled by a novel pricing mechanism which guides the tenants during their resource negotiations in order to find the most convenient way to satisfy their customer's QoS expectations. Finally, we propose a two step implementation of the proposed platform in order to be able to use the algorithm in real-time problems. The performance of the proposed framework is investigated in terms of fairness and efficiency for a traffic type (in other words without considering service heterogeneity). Next the impacts of budgets and the economical incentive for sharing are investigated.

Chapter 4: Enablers of service differentiation in a multi-tenant network

This chapter tackles the problem of how to enable service differentiation and resource customization in a shared infrastructure. We first propose a utility based traffic management model using a generic piece-wise linear function. Using four service types, we then investigate the impact of traffic heterogeneity in our model and demonstrate how the different service priorities can interact with each other.

Next, extending our model into the network slicing approach, we propose a dynamic radio access network slicing approach that can follow the dynamic negotiations and adapt the resource allocations in the slices considering the instantaneous condition of the channel and the traffic mix. Finally, the possibilities of service differentiation among tenants are investigated and the impacts of parameters are analyzed. The performance of the proposed slicing algorithm is evaluated through a large set of simulations.

Chapter 5: Short and long term impacts of anticipatory network sharing and resource trading

This chapter deals with the short and long term effects of infrastructure sharing and tackles the problem of how to enhance the efficiency of the proposed real time sharing and trading platform by exploiting the anticipatory information. Next, we focus on the integration of the anticipatory information into our model, i.e. both the real time resource scheduling and the negotiation processes. The short term implication of the enhanced model is followed by the long term analyzes of the network.

The long term implications are focused on how to utilize the aggregated short term information regarding the traffic demand, the resource usage and the traffic mix while making the investments for expansion. A novel self-dimensioning and capacity management framework is proposed to scale the network's capacity in line with the demand.

Chapter 6: Conclusion

The last chapter includes a summary of the key observations and the potential of this work.

CHAPTER 2

Review of the state of the art

In this chapter, an overview on the state of the art is presented in order to accurately position our contributions with respect to the literature. For organizational purposes, the chapter is separated in four parts, where we have investigated the conventional RAN sharing approaches, network slicing models, anticipatory networking techniques and the capacity dimensioning approaches separately. At the end of the chapter, a summary of the key observations is presented.

Decreasing the cost of network provisioning through network sharing

The aforementioned techno-economic pressure is forcing network operators to perform revolutionary changes in their business models as well as technical structure. Increasing the density of the base stations, which is the conventional solution for similar problems in the past [7], is no longer an option. A possible solution to decrease the total cost of the mobile operators is sharing the active and passive network equipment among multiple operators. Being one of the major drivers, economic models are quite extensively investigated in order to motivate the sharing. There are numerous works similar to [66] and the references therein that investigate the various aspects of sharing the resources. In addition to the economical advantages, pooling all the available resources (i.e. sharing all the available base stations as well as combining the spectrum bands) can maximize the network capacity and minimize the total cost [42].

Research and standardization entities have been exploring how to share the spectral resources, mostly within the context of cognitive radio and heterogeneous networks. A game theoretical analysis of the tenants' tendency towards sharing infrastructure resources, that is given in [18], demonstrates the viability and profitability of the mobile operators' incentive to share the resources. Unlike the former technologies, e.g. cognite networks [96], in the next generation sharing agreements, the spectral resources are shared among equivalent entities, who have equivalent right to access the resources. Therefore, the tenants are obliged to negotiate over the resources in order to obtain the right to access the resources [90]. On the other hand, despite all the advantages of infrastructure sharing, the sharing agreements have rarely covered beyond passive network infrastructure. As aforementioned, this is mostly due to the fear of losing the service differentiation. In a market where the operators compete in terms of their coverage area, resource pooling would arguably homogenize the competition power of the tenants. On the other hand, [53] shows that unlike the common belief in the industry, the tenants can differentiate their services through a set of well defined policies that guide the sharing process. Moreover, according to [97], due to the transition from a coverage driven to capacity driven network, the competition among tenants naturally shifts towards innovation and capturing the industrial needs rather than providing coverage. Therefore, sharing the infrastructure resources is an enabler to unleash the full potential of 5G.

Another key aspect is the pricing of the available resources. The importance of the resource pricing lies within setting the stability of the sharing agreements [14], where sustainability is measured whether or not the tenants are willing to continue the sharing approach or would they prefer a standalone approach. A similar analysis to find the most sustainable sharing model is given in [72], which focuses on spectrum sharing in a multi-tenant network and analyzes how to distribute the economic incentive to the operators in order for them to promote network sharing. This approach relies on a 'proof of sharing' concept which is used to determine if the tenants' sharing levels are satisfactory. On the other hand, from a business perspective, every network is a mixture of high potential and low potential regions where assessing the value of sharing is not obvious. Therefore, it is not clear how the spatially distributed resources can be compared among each other. Also such a framework would only be useful in a capacity driven (i.e. developed) wireless market. Using game theoretic model, [76] determines the Nash equilibrium for the resource prices and the allocation of respective resources. However, a tenant's willingness to pay a price strictly depends on the market dynamics, i.e. their customers' likelihood of accepting a service to a given price, the service types that the tenant is serving, the channel conditions of the users, the long and short term goals of the tenants etc. Moreover, the pricing mechanism has to consider dynamic traffic conditions (e.g. traffic volume, channel condition and traffic mixture) and scale the resource prices accordingly. Such an accurate estimation can only be achieved through a dynamic pricing approach that considers the supply and demand balance. In awareness of this need, [99] focuses on design of an 'optimum' SLA among a set of tenants and an MVNO. However, regardless of how well a sharing agreement is designed, still the pricing mechanism is performed based on statistical information and thus cannot dynamically adjust the prices in line with the evolution need. Another game theoritical framework to determine the pricing is given in [94], where the most profitable price of leasing the resources is derived using Stackelberg game modelling. However, similar to the previous cases, the derived pricing mechanism is static, therefore, it cannot provide an evolution over time, indicating that any possible expansion in the future has to be paid from the profits of the infrastructure provider. With the objective of capturing the market dynamics, [98] develops a pricing system based on the acceptance probability of the users in a shared network. On the other hand, this method cannot capture the microeconomics dynamics to enable the evolution of the network infrastructure over time.

A reasonable approach is that of using variable market-driven prices and allowing the tenants to dynamically trade the resources based on needs and within short time frames. However, it is not possible to understand the relationship between the economic aspects and the technical performance without a well-defined model. Such a model would also enable tenants to exploit the full potential of dynamic sharing. Thus a scheme that is able to automatically define prices and resource allocation based on high level tenant strategies and traffic estimation is of fundamental importance. Even if there are extensive literature on economic aspects (such as [91] and [57]) and technical considerations (such as [71] and [68]) separately, the definition of techno-economic models for resource sharing is still uncovered area.

The changing wireless resource market requires more flexible service level agreements which set the expectations and limitations of the stakeholders (i.e. tenants or infrastructure provider) without proposing any explicit right to access the resources [41]. Therefore, a key issue is to map the tenant based policies as well as the service requirements into a generic model which can be used dynamically during the inter-tenant negotiations and the real time resource allocations.

A level of flexibility in the resource allocations in multi-tenant networks is provided in [55] where the authors propose a maximum deviation parameter from the SLA guaranteed resources. Through this maximum deviation parameter, the tenants can adjust their resource shares while the scheduler can maximize the spectral efficiency. Apart from these clear advantages, the proposed framework in [55] is a pioneering work in the way towards fully automated network management. However, on the down side, the proposed sharing parameters are assumed to be very well defined in the SLA agreements which limits the innovative power of the tenants.

Network slicing

In parallel to enabling cost efficiency, 5G is envisioned to host a multitude of industry driven applications [63]. This high heterogeneity of the service types brings out the novel challenge of service coexistence. More specifically, each service comes with a unique set of requirements and priority level. The conventional solution, i.e. optimizing the available network resources in order to serve best a particular service, is simply not possible due to the number of available services. As a solution, *network slicing* proposes providing vertically grouped dedicated network resources to each service type [74] [88]. Despite the definition of vertically grouped network resources, the specific negotiable attributes of each slice and the tool for service differentiation are still under discussion in academia and standardization bodies. Although, end to end slicing, i.e. grouping the resources from core network to RAN, is usually considered for standalone cases [44], for a multi-tenant network, the key problem is usually RAN slicing [92].

The simplest way of network slicing is ‘static-slicing’ which relies traffic statistics over a very large time interval and fixed service priorities (e.g. [92]) and is not up to change for a predefined (and usually very long) time period. The benefits of static slicing are investigated in [89] [38] [40] considering static SLAs. On this line, due to the business potential of static slicing, [39] provides an auction based pricing framework for the wireless resources and argues that the pricing mechanism can provide the efficient resource allocations.

However, the applications up till now demonstrate the fact that static network provisioning has a tendency towards over-provisioning the network resources. Although such an over-provisioning may be acceptable in the past, the current techno-economic condition of the broadband communication market disfavors the waste of valuable spectral resources. A solution, similar to infrastructure sharing, is dynamically slicing the network resources which would also adjust resource allocations according to the channel conditions, traffic charachteristics and variations, and service heterogenity [73]. Namely, the wireless resources can be dynamically assigned to the different slices in order to meet their needs. The immediate availability of the resources reduces the need for the over-provisioning. Consequently, both the academic and industrial research focus have been directed to the dynamic network slicing frameworks. In [20], the authors propose a dynamic network slicing - resource reallocation framework where the reallocation focuses on providing a QoS threshold for all the services. Therefore, the slice broker, i.e. a logical entity in their framework, observes the QoS of slices and reallocates some of the resources in case a slice cannot reach its QoS expectation while the others are higher than their threshold. On the other hand despite a level of freedom, their proposed framework is incapable of capturing the inter-tenant dynamics. Moreover, they do not consider the inter slice prioritization which makes it even harder to control inter-slice relations.

Similar idea of network slicing has also been implemented in [77], where the separate entities are continuously observing the slice QoS and submitting resource requests accordingly. Despite the focus on inter-slice prioritization in [77], the physical dynamics (e.g. channel conditions) are not considered which can drastically decrease the spectral efficiency. The authors of [50] investigate the network slicing in order to achieve the QoS guarantees. However, the proposed framework does not consider the adaptation to the changes in the traffic conditions (i.e. the channel quality and traffic mix). Also due to its overly complicated structure, the proposed framework is not feasible to be used in a multi-tenant framework. The resource sharing among tenants in a sliced network is also investigated in [100] and [102]. However, built upon well-defined SLA shares, these works are unable to offer the needed flexibility in the next generation wireless networks. Moreover they do not consider the evolution of the infrastructure resources, which requires a dynamic resource pricing in line with the required capacity expansion. A vitualization framework is proposed in [32], where the resources are scaled according to tenant's dynamic needs and fairness is guaranteed not only between tenants, but also between users of different services. The model, however, does not consider adaptation to channel conditions and economic aspects of resource trading. Moreover, it only considers standalone scenario and (similar to majority of the works) relies on well-defined SLA agreements.

Dynamic RAN slicing is commonly assumed to be able to provide the full flexibility to network management. The term dynamic reflects that the resource allocations to the slices vary over time depending to the traffic mixture and the observed channel conditions [58]. A dynamic slicing framework is provided in [69] in order to carter the multiplexing gains. However, despite the level of flexibility they proposed by updating the slice allocations separately, their proposed algorithm does not provide any insight of how the slicing should be handled when there are multiple actors with the same rights to access the resources. An interesting solution is proposed in [48] where the authors define the utility per service using both delay and achieved rate. Their proposed model performs the dynamic resource allocations according to this proposed utility function. However the impact of multi-tenancy on the resource allocations is not covered. Moreover, although using utility based resource allocation is very promising, the integration of delay constraint in the achieved utility can result in over-prioritizing particular services. An interesting approach is presented in [95], where the available resources are also considered to be variable. More specifically the available resources varies according to the harvested energy at time intervals. Their proposed framework uses network slicing in order to distribute resources among competing services, however, it does not consider multi-tenancy either. Finally, [43] proposes a two-step slicing framework in order to address delay guarantees.

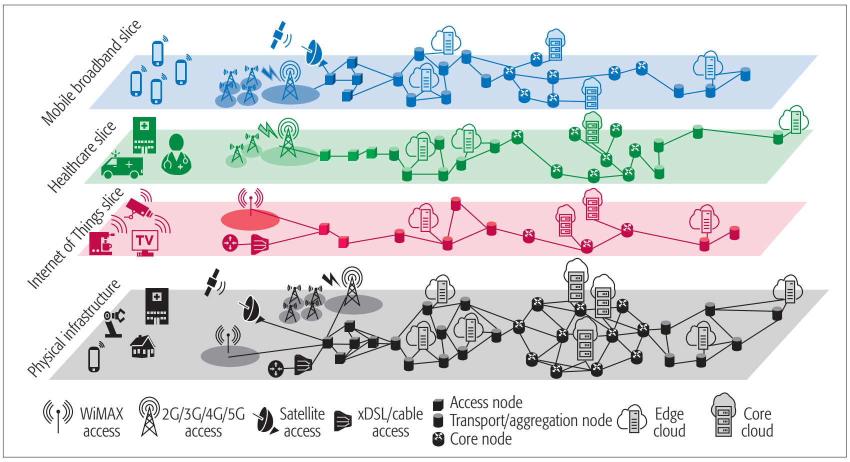


Figure 2.1: *Envisioned multi-ventor and multi-network management of 5G network slices (taken from [67]).*

Dynamic network slicing could be the best possible solution in order to prevent over-provisioning of the network resources. On the other hand, it promotes resource sharing among slices which means that the performance of the separate slices are no longer independent from each other [67]. This inter-slice dependency requires a higher level of control in order to guarantee that the service priorities are still maintained while the maximum spectral efficiency is achieved. Moreover, the multi-tenancy scenario puts a new level of complication onto this problem. Any proposed framework has to be flexible enough to provide high spectral efficiency and low cost while allowing tenants to differentiate their services from their competitors, however at the same time, it should guarantee inter-tenant fairness and inter-service priorities [73].

Anticipatory networking

The next generation wireless networks are envisioned to support delay constraints that are below 1 ms and throughput higher than 10 Gbps [23]. However, reaching this goals is very challenging. The developments in the sensors and the internet of everything create an abundant amount of data. In the recent years, the developments in the artificial intelligence and machine learning give the researcher the necessary tools in order to process this data and produce accurate models. Using these models, the researchers can predict the upcoming changes in the traffic demand and channel. Consequently, exploiting this prediction information in order to govern the decision making is increasingly becoming popular. In contrast to this decentalization approach, [29] focuses on anticipatory resource allocations in order to decrease the delays simply by trying to prepare all the resource available by the time the request is received.

Despite the escalating attention in the literature for the anticipatory networking principle [36], a good portion of the existing works depend on the accuracy of the prediction algorithm. The works such as [52] [49] and [24] can be considered as the representatives of the direct usage of prediction methods with the available data set. Another research on the application of different prediction methods on the network traffic predictions and presents a comparison among them is given in [52]. In [49], the authors tackle the problem of enabling ultra low latency communication within vehicular communications. Therefore, the main focus is on the prediction of the cars' future locations which have been determined from their prior locations. Although their proposed model is particularly interesting, the overall complexity of their algorithm makes the real time implementation of it rather challenging. A similar inter base station coordination problem has been focused in [24], where the authors focus on predicting the achievable data rate of the streaming users. This work focuses on minimizing the need for continuous base station support in a region by accurately buffering the content.

On a different note, unlike the aforementioned data driven anticipatory approaches, the real time scheduling and trading problem has to be able to follow the evolution of the traffic conditions over time. Therefore, the deployed prediction method has to be able to follow the dynamic variations in the time series of the traffic condition. Adopted anticipatory approaches for similar problems that have arisen in different fields, such as energy market ([62]) and stock market ([25], [83]) can shed some light on our problem and provide a beginning point to enable anticipatory network sharing. The non-storability of the electricity requires accurate generation of power in line with the demand and available resources. Built on this need, [62] compares Artificial Neural Network (ANN) and Auto Regressive Integrated Moving Average (ARIMA) methods in predicting the wind speed for energy generation. As an important aspect of their model, the authors propose a hybrid model that can utilize the efficiency of ARIMA in predicting short observation periods while for the long term averages they still rely on ANN. On a similar problem, [25] focuses on forecasting the stock prices using a combination of ARIMA and back propagation neural networks while [83] combines ARIMA with a simple neural network model. The common point of both of these works is their focus on the ARIMA algorithm for shot term predictions while maintaining a ground level of awareness using neural networks.

ARIMA is mostly deployed in the time series prediction as it provides very high prediction accuracy with respect to its complexity. Unlike the well known deep-learning approaches (such as [8] or [65]), ARIMA does not require a very large training set. On the contrary, based on a small set of temporal observations, it models the behavior of the time series and predicts a set of future values for it. The standard form of ARIMA is built upon three control parameters, i.e. the number of auto-regressive terms, the number of seasonal differences and the number of moving average terms [93]. Unlike the deep learning approaches, these terms are derived from a learning window W_L which is usually the last observations of the time series rather than a very long history. Then the ARIMA uses these parameters to predict the upcoming values within the prediction window W_P. Despite being a very simple model, the non-linearity of the time series is quite challenging to model in ARIMA [93].

In order to capture the non-linearity, a favored approach is using ANN or with a more specific name feed forward neural networks (FFNN) [16]. Built upon a training data set, FFNN can learn the correlation of the different time slots by simply updating the weights in the model. Similar to ARIMA, FFNN also requires three major parameters to build its model, i.e. the number of hidden layer nodes, the number of hidden layers and the number of delays. In a most generic sense, the simplest FFNN contains three layers, one input layer, one hidden layer and one output layer. In the execution of the time series prediction, the usually deployed strategy is to feed each observation over time as an input (namely, W_L inputs) and predict the one upcoming time slot value. After that, with a sliding window approach, FFNN iteratively predicts the upcoming time slots one by one with W_L inputs until it reaches the complete prediction horizon W_P. However, this execution of FFNN increases the prediction errors in the single time slot because of the iterative usage of predicted values rather than the actual observations. On the other hand, further improvements in the prediction accuracy can be achieved by simply increasing the complexity of the applied prediction algorithm by changing the iterative prediction algorithm. For instance in [65], the authors deploy a deep learning approach in order to more accurately predict the network traffic. However, such an approach requires the network to be in a steady state so that the training would accurately capture the network dynamics. In contrast to the requirement of steady state, 5G network is expected to be highly dynamic with a multitude of industry driven services. The recurrent neural networks, such as long short term memory can carry the context information over time which immensely increase the prediction accuracy with respect to ARIMA [81]. On the other hand, the computational complexity of such a method makes it impossible to be deployed in model that needs to be run in (close-to) real-time.

Relatively speaking, anticipatory network slicing models have attracted the focus of the research community only in the recent past. Works such as [58] and [12] focus on exploitation of anticipatory information in network slicing. However, these studies rely on relatively long term observations and require large computation duration in order to accurately slice the network resources. In contrast, for our problem the key aspect is to obtain the prediction values within short time scales (in the order of seconds or a minute), even if this causes a fall in the prediction accuracy. More specifically, as shown by [54], the impacts of prediction errors can be compressed while missing the delay constraint is irreversible. The authors of [78] tackle this resource allocation problem using the anticipatory information of the upcoming traffic conditions. However, depending on static SLAs, this work cannot exploit the full flexibility and the efficiency of dynamic network slicing.

Dimensioning the network resources

One of the widespread discussion regarding 5G deployment is whether this deployment is necessary or not considering that the 4G network can still effectively handle the demand. A simple answer to this question is the need to new 5G deployment is driven by the decreased QoE provided by 4G deployment when serving heterogeneous services. Moreover, even the earliest estimations present tougher challenges for the next generation networks (i.e. 6G and beyond) [22]. In 5G, unlike the previous generations, the deployment of new infrastructure can be gradually handled region by region, as the functional and the computational requirements of the newly deployed services evolve over time, and the less demanding cases can be solved with enhanced load balancing approaches, e.g. [47] [36]. The conventional capacity management approach relies on long term observations, i.e. averages, of the traffic demand and the provided QoS metrics and is usually focused on maximizing the minimum QoS of the network [82]. However, it is clear that such an approach results in underutilized network resources out of peak hours [37]. Despite the infrastructure sharing can decrease the problem up to a level, it can only be solved with a concrete capacity expansion strategy that can create timely and accurate expansion. The authors in [82] propose a quality of experience and traffic demand aware capacity scaling strategy that carters the advantages from queuing theory. On the other hand, their estimations still count on statistical approximation rather than estimating the actual needs of the traffic. Moreover, the proposed framework does not consider any priority among different services, which decreases the efficiency of their model in real time cases since the service heterogeneity is one of the key factors to determine the urgency and the length of capacity expansion.

Although the problem of network planning is usually modeled under several steps, the two key aspects are dimensioning and detailed planning [26]. During the dimensioning of the network resources, the generic decisions on resource allocations to different regions are handled without any particular decision on the exact location of the new resources [30] [64]. According to the resource provisioning decisions, detailed planning block determines the physical parameters as well as the exact location of the new deployment based on the radio maps, (e.g. [15] [60]). In this thesis, we focus on the problem of dimensioning the network.

Eventhough the problem of capacity dimensioning is well-investigated in the literature, a majority of these works do not consider the existing infrastructure resources, instead they provide models for complete network deployments [31]. However, unlike the previous technologies, 5G deployment is not envisioned to replace the 4G deployment. Therefore, the network planning in 5G strongly requires awareness of the existing network structure as well as the long and short term variations of the network traffic [70]. Moreover, the unique 5G aspects such as multi-tenancy and the heterogeneous service requirements have to be considered in the planning decisions. In order to decrease the reaction time to the changes in the network and autonomously control the network resources, [61] provides a self-dimensioning algorithm for small cells for multi-tenant networks. However, their model is built upon the average QoS degradation and does not consider how different services perceive this decreased QoS.

Another challenge which is also covered in [80] is the fact that the network operators do not always posses the necessary revenue to upgrade or expand their resources. A detailed cost analysis of network deployments is presented in [87], where the authors consider not only the device cost but also the respective deployment and capacity costs. Following this study, the authors also extend this research to an analysis of the newly deployed base stations' impact on the spectral efficiency in [86].

In parallel to the long term evolution of the wireless infrastructure, relatively short term mechanisms are increasingly becoming popular to handle the short term fluctuations on the total demand [85]. Despite many options in such a short term approach, the majority of the proposed models in the literature have focussed on public safety and disaster management [60] [30]. This limited set of use cases is mostly because of the need for static data. The conventional network planning method requires a statically stable data regarding the region that would allow the determination for the optimum location for a new base station. Therefore, application of mobile base stations require pivotal changes in the network dimensioning approach.

Summary

Although our literature survey has shown a multitude of different approaches and methods to enable 5G, most of these works have common open points which we believe to be crucial for 5G deployment.

- Nearly all the works are built upon the assumption of an existing and very well defined SLA that states the long term shares of individual tenants.

- The conventional fixed sharing cannot provide the needed cost efficiency, while the full sharing models in the literature are incapable of proving the envisioned flexibility and efficiency.
- In order to provide a sustainable business model, a market driven pricing strategy is crucial. Moreover, the pricing strategy has to cover the evolution of the infrastructure resources.
- Using the available anticipatory information is a key aspect in achieving the 5G expectations. However, the well-known tradeoff between time complexity and the prediction accuracy has to be carefully considered in the design of any network management model.
- The conventional network capacity planning strategy that focuses mainly on the QoS degradation has to be revisited because of the changing business ecosystem in 5G. In particular, the inter-tenant dynamics such as fairness and competition and inter-service dynamics, like service priorities and performance indicators have to be considered in order to efficiently manage the capacity scaling.

CHAPTER 3

Dynamic sharing in a multi-tenant network

3.1 Introduction

Infrastructure sharing in a multi-tenant network is among the most viable options to decrease the economic pressure set by the decreasing profitability of the network provisioning. However, the conventional sharing agreements rely on static SLAs to define the resource share of each tenant and the respective prices. On the other hand, the technical dynamics (e.g. channel condition and traffic demand) and business dynamics (i.e. resource scaling and TTM of new services) require a higher level of flexibility in sharing where the tenants can make dynamic decisions on their resource usages in line with the instantaneous conditions of the network and their business strategies. Moreover, the profitability of any market model is strongly related to how well each party understands the resource negotiation and its impact on the total cost and achieved utility. Therefore, a concrete techno-economical model is required to enable real time sharing of the radio resources.

Revisiting all these aspects, in this chapter, we provide a novel dynamic network sharing platform that can provide flexible and efficient resource allocations and negotiations among different parties involved in the sharing agreement. Based on what is observed in Chapter 2, it is assumed that each tenant has an incentive to share the network resources that are provided by an infrastructure provider (or a grand coalition of infrastructure providers that act as a single entity). The key focus of this chapter is to analyze how to negotiate and trade the resources among the key stakeholders in the most cost and spectrum efficient manner. Therefore, in order to preserve the focus solely on the proposed negotiation platform, only elastic traffic is considered in this chapter.

The findings presented in this chapter are published in [4] and [3].

3.1.1 Specific research questions

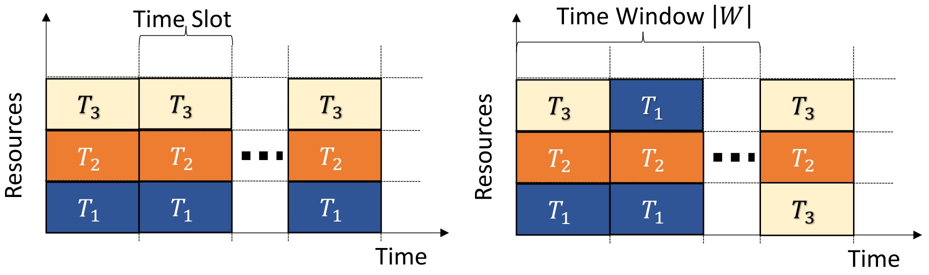
In this chapter the main research question of 'How can the network resources dynamically and flexibly be shared in a multi-tenant network?' is investigated with three main focuses, namely,

- How can the negotiations between tenants and the infrastructure provider be modeled? (Introduced in Section 3.2 and further investigated in Section 3.3)
- How should the resource pricing be handled? (Explored in Section 3.2.5)
- What are the key components in the sharing agreement? (Introduced in Section 3.2 and their impacts are explored in Section 3.4)

In line with these questions, a short time scale dynamic trading model is proposed wherein: $i)$ the cost of resources is market driven, and $ii)$ the tenants trade resources based on their ability to satisfy customer demands as well as meet their respective budget constraints.

3.1.2 Chapter outline

Section 3.2 starts with an in depth analysis of the flexible resource sharing platform that forms the basis of our work. Following this analysis, we present our argument on the simplification of SLA agreements and our novel pricing mechanism. Next the system model and the underlying assumptions as well as the mathematical framework of our work are presented. The modifications on the algorithm in order to run the proposed model in real time are detailed in Section 3.3. Section 3.4 evaluates the performance of the proposed model for a variety of simulation scenarios. Finally, Section 3.5 summarizes the key observations and the findings of this chapter.



(a) *Fixed sharing approach*

(b) *Flexible sharing proposed in [55]*

Figure 3.1: *The level of flexibility in resource scheduling proposed in [55] (on the right) with respect to the conventional fixed sharing approach (on the left).*

3.2 Negotiation platform

3.2.1 Background

In [55], the authors focus on the wireless resource sharing in a multi-tenant network and tackles the problem of increasing the resource efficiency of the network. In order to properly position our contributions in respect of the literature, in this section we outline their proposed framework and key idea behind it.

As previously discussed, the conventional strategy of network sharing relies on fixed resource shares introduced by very well defined service level agreements (i.e. *fixed sharing*). An example scenario is presented in Fig. 3.1a where the network resources are shared among three tenants – T_1, T_2 and T_3. In line with the common application in resource scheduling literature, the time is discritized and divided into time slots, of which the index in generically denoted by n in the rest of the thesis. As depicted in Fig. 3.1a, in the fixed sharing, at each time slot n the tenants always receive a constant fraction of resources, regardless of their actual traffic demand. Thus even if one of the other tenants is facing a resource shortage, i.e. not having sufficient resources to serve all the users, the unused resources from the other tenants cannot be utilized to serve this tenant. However, both from an economical point of view as well as technical point of view, this is quite inefficient.

In order to overcome this inefficiency, the authors of [55] proposed a new level of flexibility on the real time resource allocations of the model. In their proposed approach, during the negotiations between the stakeholders (i.e. tenants and the infrastructure provider), the tenants also decide on a maximum deviation from an agreed sharing ratio which allows the scheduler to deviate from fixed sharing ratios in instantaneous resource allocation. More specifically, the instantaneous deviation of resource allocation cannot exceed this agreed maximum deviation parameter within a time window length, $|W|$. In their considered framework, the time window can be forced by the infrastructure provider or can be an outcome of the negotiations. In case this maximum deviation is set to be zero, as could be the case for the second tenant in Fig. 3.1b, the scheduler provides a fixed sharing approach. On the other hand, if the maximum deviation is greater than zero, cf. tenants 1 and 3 in Fig. 3.1b, the scheduler has the flexibility to assign more or less resources to the tenants to maximize the spectral efficiency given that at the end of time window, the assigned resources in average would not be less or more than the maximum deviation value from the agreed sharing ratio.

In [55], the authors also show how the maximum deviation can increase the utility of multi-tenant resource sharing. Their proposed model can be considered as a preliminary dynamic resource sharing approach where the resource sharing can be adjusted to the instantaneous conditions. However, it is also apperant that this flexibility is only provided in the resource scheduling part. Any update on the sharing parameters would require long negotiations among tenants which increases TTM for the newly defined services. Moreover, once the resource shares are decided, it is hard for tenants to differentiate or adjust their obtained resources. Lastly, despite being a pioneer on flexible sharing, the proposed model is a technical approach that does not possess any economical parts. In our PhD study, taking the flexibility concept defined in [55] as a starting point, we build a novel autonomous negotiation and resource sharing algorithm.

3.2.2 System model and assumptions

In this model, it is assumed that a set of tenants, M, are sharing the downlink of a base station that is provided by an infrastructure provider. A set of users, K, are homogeneously distributed among tenants, and the set of users per tenant is indicated by K_m, where $\cup_{m \in M} K_m = K$. In order to fully focus on the sharing concept, we assumed that the base station's scheduler acts in a standalone mode where the scheduling decisions of this base station would not have any impact or would not be affected by the neighboring sites. Each time slot n is assumed to have a length of 1 transmission time interval (TTI) which can be set according to the technical capability of the base station.

3.2.3 Simplified service level agreements to enable automated negotiations

The service level agreement concept in [55] contains three key parameters that guide the resource sharing mechanism. The first aspect is the guaranteed sharing ratio per tenant m, denoted as $S_m \in [0, 1]$ which defines the fraction of the resources that the tenant expects to receive in average. Secondly, the maximum deviation from the guaranteed sharing ratio, $\Delta_m \in [0, 1]$, which sets the upper and the lower limits of the actual resource allocations in average per tenant in a time window. The last parameter is the time window size, W, that can be given by either the infrastructure provider or a consensus between the tenants.

However, similar to any sharing approach, having nearly fixed parameters to guide the sharing process can result in delays in new service introduction as well as inefficient resource usage. When a tenant wants to introduce a new service, a new negotiation process has to be initiated which causes uncertainty and leads to further delays in TTM and even some lost business opportunities. In the conventional broad band wireless market, such delays can be tolerable due to the relatively low number of services. However, the key attribute in 5G is indeed the heterogeneity of the service mixture which requires the tenants to be flexible not only in the dynamic resource allocations but also in their business models. Therefore, in order to propose a stable and durable sharing platform, we believe the conventional SLA structure has to be revised in a way where the SLAs only provide the high level policies of the tenants regarding their QoS expectations and the budget that they can spare to reach their utility expectations. From the infrastructure provider's point of view, the agreed unit prices for the resources as well as how the pricing will be handled should also be included in the SLA. In summary, we argue that rather than being a complete sharing outline, the SLAs in 5G and beyond networks should define the unit aspects and the generic expectations of the each parties from sharing, and the actual negotiations for the resource allocations should be dynamically handled in order to guarantee adaptability to the temporal changes in the network traffic conditions.

Therefore, in this work, we assume that the tenants declare their budgets per time slot, B_m, their utility thresholds, $U_{\text{th},m}$ and their time windows, W_m (cf. Fig. 3.2). Unlike [55], we let each tenant to determine their individual time windows, W_m, during which the QoS expectations of the service is required to be satisfied. Note that, although we do not consider any explicit analysis of the delay constraints of the services, it is indirectly modeled

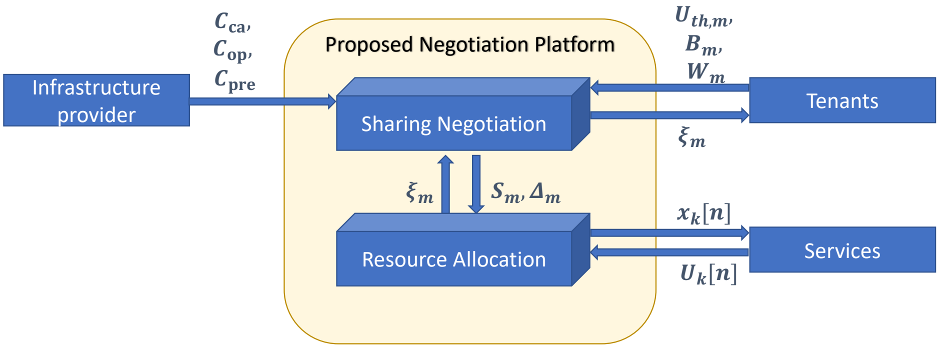


Figure 3.2: *An outline of the inputs and the outputs of different parts in the proposed negotiation and resource scheduling algorithm for multi-tenant networks.*

with W_m. As will be further discussed in the next part, the negotiations of the sharing parameters are handled at the end of every renegotiation interval (*RI*) which is a parameter set by the infrastructure provider. Note that as will be further explained in Chapter 4.5, unlike W_m, the changing values of RI does not have a direct impact on the dynamic resource allocations, and therefore, the least common multiple should be selected (will be further discussed in Section 4.4.5).

3.2.4 Applied notations

The main objective of the tenants is to obtain the necessary downlink resources to fully serve their customers, i.e. *services* in Fig. 3.2. In this thesis, we assume that the main objectives of the infrastructure provider is i) providing a sustainable sharing platform, and ii) maximizing the spectral efficiency while considering the QoE of the services. Therefore, the infrastructure provider is assumed to be a non-profit entity who reinvests all the collected revenue in order to expand the available resources. The achievable rate per spectrum unit [Hz] for each user is given as $r_k[n]$. The assigned resources per user k at time slot n is represented by $x_k[n] \in [0, 1]$. The instantaneous deviation from S_m, i.e. depicted by $\epsilon_m[n]$, is determined based on $x_k[n]\ \forall k \in K_m$. The instantaneous deviation from tenant's guaranteed resource share is limited by the maximum deviation from S_m, i.e. Δ_m. S_m and Δ_m are averaged per time window, W_m, which is considered to be fixed and equal for all the tenants in this chapter, i.e. $W_m = W\ \forall m \in M$. In order to solely focus on the proposed sharing model, we also consider RI to be constant and equal to the time window length, i.e. $RI = W$, but the impacts of choosing different RI is investigated in Chapter 4.5. Finally, the difference between each tenant's utility target, i.e. $U_{\text{th},m}$, and the actually achieved utility is defined as the gap per tenant m and represented as $\xi_m[n]$.

3.2.5 Dynamic pricing of the resources

As outlined in Fig. 3.2, the infrastructure provider sets the unit cost per resource which is modeled by three parts, i.e. capital costs C_{ca}, operational costs C_{op} and the pressure cost $C_{\text{pre},m}$. It is assumed that the tenants commit to pay a CapEx cost which is proportional to their guaranteed resource share, S_m since in case of $\Delta_m = 0$, they receive exactly this amount of resources. However, a level of flexibility in pricing is defined in their OpEx where the tenants only pay for their active resource usage. This flexibility in OpEx payment is designed to motivate the tenants to trade their resources. As an extreme case, if a tenant m chooses *full sharing*, i.e. $S_m = 0$ and $\Delta_m = 1$, then the tenant will pay a lower cost than its counterparts since its resource change is completely opportunistic. However, for this extreme case, the tenants have no guarantees regarding the resource allocations, thus it would be risky to provide inelastic services in full sharing. Consequently, the tenants can customize their guaranteed sharing ratios as well as their flexibility in obtaining these shares in line with their budget. Note that we assume that the tenants define their budgets per time slot n and would not use their budgets for the upcoming time slots to buy resources at any n.

The microeconomic dynamic of demand and supply based pricing is modeled using the pressure cost in our model. In relatively shorter time instances, the pressure cost acts as a regulator for the resource usage, namely the cost of resources are higher when the demand is higher, forcing the tenants to have a more opportunistic resource usage rather than guaranteed resources, thus allowing the infrastructure provider to set the maximum possible spectral efficiency. In case the resources are sufficient to fully satisfy all the tenants, then the pressure cost will be equal to zero. In the long term (e.g. months or years), the collected revenue from the pressure costs allows the infrastructure provider to extend the network's capacity to meet the demands of the tenants. This inversely proportional relationship between the cost and the resource demand is modeled using the gap of each tenant, $\xi_m[n]$. Note that by modeling this inverse relationship with the gap, ξ_m, instead of the available resources $1 - \sum_{k \in K} x_k[n]$, the pressure cost is ensured to be reflecting the capacity expansion need in a cell. Considering all these aspects, the surcharge of the pressure cost is modeled as $\xi_m[n] \times C_{\text{pre},m}$. Thus, in this thesis, the *resource scarcity* is considered to be the case where $\xi_m > 0$ and $\sum_{k \in K} x_k[n] = 1$, whereas, the *resource surplus* is modeled as the scenario where $\xi_m = 0 \, \forall m \in M$ and $\sum_{k \in K} x_k[n] < 1$.

Finally, in order to give the tenants equal opportunity to obtain the resources, the resource prices are scaled proportionally to the tenant's budgets. Thus we guarantee that the resource prices would reflect the actual value for each tenant and it is impossible for a rich tenant to monopolize the market by artificially inflating the costs and pushing the rest of the tenants out of the competition. Therefore, the pressure cost per tenant is defined as

$$C_{\text{pre}^m} = \frac{B_m}{\sum_{m \in M} B_m} \times C_{\text{pre}}. \tag{3.1}$$

3.2.6 Formulation of the proposed framework

Using the notation defined in the previous section, the optimization problem that is run at the scheduler of a base station is given by Equations (3.2.a)-(3.2.i). We define a continuous objective function, composed of two parts. The first part is the minimization of the total gap among all the tenants, which is the difference between the utility goal and the actually achieved utility of each tenant. By focusing on the total gap over all the tenants rather then minimizing the maximum gap, the scheduler is designed to find the resource allocation with the maximum spectral efficiency which would satisfy the techno-economic expectations of all the tenants (i.e. defined by B_m and $U_{\text{th},m}$). Since in this chapter, we only considered elastic services, there is no incentive for tenants to choose $S_m \geq 0$. Consequently, in order to guarantee that the initial deployment cost of the infrastructure provider will be met, we model the second part of the objective function, i.e. the minimization of $S_{\text{max}} = \max(S_m, 1 - S_m)$ which creates a correlation between the guaranteed resource shares of all the tenants. The minimum value of S_{max} in (3.2.b) can be reached when $S_m = \frac{1}{|M|}, \forall m \in M$. Therefore, by minimization of S_{max}, we assure the distribution of the guaranteed resources to be equivalent among the tenants given that all the tenants have sufficient budgets. The deviations in the obtained resources can be done through the selection of $\Delta_m > 0$ which can be interpreted as the trade incentive of the tenants.

In line with our previous definition, (3.2.c) formulates the gap per tenant, $\xi_m[n]$. One can observe from (3.2.c) that even though it is possible to achieve a higher utility than the expectation of the tenant, $U_{\text{th},m}$, in order to reflect the need for required capacity expansion, the gap per tenant, $\xi_m[n]$ cannot receive a value lower than zero. The actual achieved utility, $U_k(x_k[n], r_k[n])$, which will be defined later, is calculated using the achievable rate per user and the assigned resources to this user k at time slot n.

$$\min \sum_{m\in M} \xi_m[n] + S_{\max} \tag{3.2.a}$$

$$\text{s.t.} \qquad S_{\max} \geq \max(S_m,\ 1 - S_m), \quad \forall m \in M, \tag{3.2.b}$$

$$\begin{aligned}\xi_m[n] \geq \max(0, U_{\text{th},m} - \frac{1}{(a+1)|K_m|} \sum_{i=n-a}^{n} \sum_{k\in K_m} U_k(x_k[i],\ r_k[i])),\\ \forall m \in M,\, a \equiv (n-1 \bmod W)\,,\end{aligned} \tag{3.2.c}$$

$$\epsilon_m[n] = \left(\frac{1}{(a+1)} \sum_{i=n-a}^{n} \sum_{k\in K_m} x_k[i]\right) - S_m, \quad \forall m \in M, \tag{3.2.d}$$

$$|\epsilon_m[n]| \leq \Delta_m, \quad \forall m \in M, \quad \forall n \in N, \tag{3.2.e}$$

$$\sum_{i=n-a}^{n} \big(S_m(C_{\text{ca}} + C_{\text{op}}) + \epsilon_m[i]C_{\text{op}} + \min\left(\xi_m[n]C_{\text{pre},m}, B_m\right)\big) \leq B_m(a+1), \ \forall m \in M, \tag{3.2.f}$$

$$0 \leq \Delta_m \leq \max(S_m,\ 1 - S_m), \quad \forall m \in M, \tag{3.2.g}$$

$$\sum_{k\in K} x_k[n] \leq 1, \ \ x_k[n] \geq 0, \ \ \forall k \in K, \tag{3.2.h}$$

$$\sum_{m\in M} S_m \leq 1\ , \ \ S_m \geq 0, \ \ \forall m \in M. \tag{3.2.i}$$

The instantaneous deviation from the guaranteed sharing ratio, $\epsilon_m[n]$, is introduced in (3.2.d). The first term in the right-hand side (RHS) of (3.2.d) calculates the total assigned resources to the tenant in the current renegotiation interval, whereas, S_m is the guaranteed sharing ratio for tenant m. One can observe that $\epsilon_m[n]$ can be both negative and positive, where the former reflects buying additional resources through trading and the latter represents the case where the tenants sell resources through trade. While calculating $\xi_m[n]$ and $\epsilon_m[n]$, the RI is considered to be independent.

Inequality (3.2.e) limits the maximum instantaneous deviation in the assigned resources to m. This constraint defines the depth of sharing by limiting how much flexibility that the scheduler can have while assigning resources to the users. The case where $\Delta_m = 1$ corresponds to the *full sharing scenario*, whereas, the limit case of $\Delta_m = 0$ is the scenario where the tenants do not dynamically share the resources (i.e. *no sharing scenario*). Constraint (3.2.f) integrates the economic aspects into the model. The left-hand side (LHS) of (3.2.f) calculates the total cost of a tenant. As previously defined, the tenants are considered to be paying both CapEx and OpEx proportionally to their guaranteed resource shares, S_m, in order to guarantee that in case the tenants do not involve themselves in the dynamic sharing agreement (i.e. $\Delta_m = 0$), the cost of the resources that they use is still paid. On the other hand, by performing trade activities by selecting $\Delta_m > 0$, the tenants can vary their total OpEx spending. More specifically, if the tenants obtain more resources than their initial estimations (S_m), from (3.2.d), the second term in the LHF of (3.2.f) becomes positive and scales the additional OpEx cost of resources. On the other hand, if the tenants use less resources than what they initially decided, the second term becomes negative and decreases the total cost of the tenant. A key aspect in pricing is the difference between the guaranteed resources and resources obtained through trading. By its definition, guaranteed resources are more expensive than the resources obtained through trade as guaranteeing a fixed portion of the resources decreases the flexibility of the scheduler. Therefore, the model incentivizes the tenants to dynamically share the resources rather than reserve some static portions of the network. The third term in LHS of (3.2.f) is the pressure cost. As previously mentioned, the pressure cost follows "supply and demand" dynamics in microeconomics while collecting the necessary revenue for a future expansion of the network capacity. As (3.2.c) guarantees that $\xi_m[n]$ reflects the unsatisfied traffic demand of tenant m, the pressure cost increases proportionally to the need for the capacity expansion. Further details regarding how to perform the capacity expansion through $\xi_m[n]$ is detailed in Chapter 5.5. The RHS of (3.2.f) sets the available budget of the tenant. The tenant's budget is defined per time slot n, however, tenants can choose to spent their budget in a given time slot or wait for another time slot to spent it. This aspect has been modeled with a scaling factor, i.e. $1 + a$, where $a \equiv (n-1)\ mod\ RI$.

Constraint (3.2.g) ensures that the tenants wold not trade the resources that they do not possess. Constraints (3.2.h) and (3.2.i) set the physical limitations. More specifically, (3.2.h) ensures that the total assigned resources to the users at any time slot n cannot exceed the total amount of resources, whereas, (3.2.i) prevents the total sum of the guaranteed resource shares to be more than the available resources. Note that the non-linearity in the presentation of the proposed model is only for the sake of readability, but it can be linearized with standard techniques.

3.3 Real-time implementation of the proposed model

The model given by Equations (3.2.a)-(3.2.i) is designed to work as a real time scheduler. Thus, the determination of the new sharing parameters, S_m and Δ_m, comes up with a further challenge, i.e. determination of the future achievable rates with a high accuracy. However, in a real-time implementation it can be challenging to obtain high accuracy prediction. Thus in this section, we introduce a two-step algorithm to implement the proposed model, that can iteratively reach the optimum sharing parameters, i.e. the sharing parameters that would create the highest spectral efficiency within the budget limits of the tenants. The proposed model is split into two problems, namely P_1 and P_2, as depicted in Fig. 3.3, where the real time scheduling decisions are governed by P_1, and P_2 is responsible for the determination of the most efficient sharing parameters according to the real time scheduling decisions.

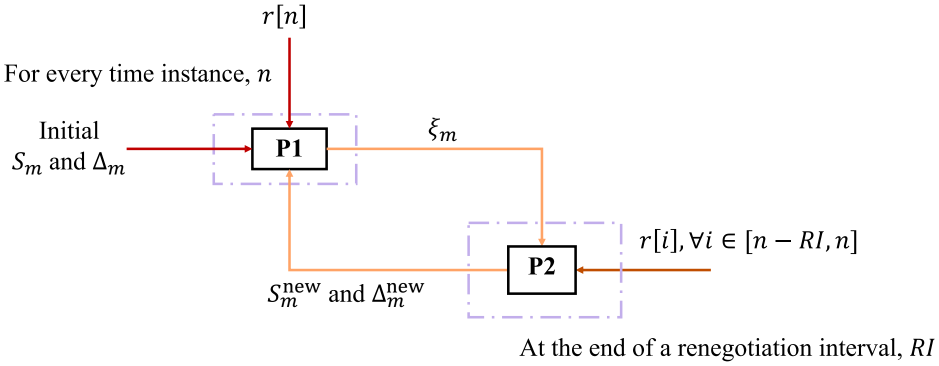

$r[n]$
For every time instance, n
Initial
S_m and Δ_m
P1
ξ_m
P2
$r[i], \forall i \in [n - RI, n]$
S_m^{new} and Δ_m^{new}
At the end of a renegotiation interval, RI

Figure 3.3: *The proposed two step optimization framework.*

In the proposed real time scheduling problem (P_1), the sharing parameters (i.e. S_m and Δ_m) are considered to be constant and the scheduler can minimize the total gap (given in (3.2.a)) only by changing the instantaneous resource allocations. This resource allocation problem is run for ever time slot n within a renegotiation interval, RI according to the instantaneous achievable rates at the respective time slot, n. At the end of every time window, in accordance to the observed achievable rates during the previous time window, the optimizer determines the optimal resource allocation, which would produce the highest spectral efficiency while satisfying individual needs of the tenants (i.e. the resource sharing problem in P_2). Both problems (P_1 and P_2) use the same objective function and the constraints, however, since the variables are different for each problem, the active constraints are different and can be reformulated as follows.

$$P_1 := \begin{cases} \text{(3.2.a)} & \min\limits_{\xi_m,\, x_k,\, \epsilon_m} \sum\limits_{m \in M} \xi_m[n] \\ \text{s.t.} & \text{(3.2.c)(3.2.d)(3.2.e)(3.2.f)(3.2.h)} \end{cases}$$

$$P_2 := \begin{cases} \text{(3.2.a)} & \underset{\substack{\xi_m,\, x_k,\, S_m,\\ \Delta_m,\, \epsilon_m}}{\min} \; \sum\limits_{m\in M} \xi_m[n] + S_{\max} \\ \text{s.t.} & \text{(3.2.b)} - \text{(3.2.i)} \end{cases}$$

The calculated 'optimum' sharing parameters by P_2 only depend on the achievable rates of the previous renegotiation interval. In order to achieve adaptability to the evolution of the network conditions, the new sharing parameters that are given to the P_1 in the upcoming renegotiation interval, i.e. S_m^{new} and Δ_m^{new} are calculated according to

$$S_m^{\text{new}} = (1-\alpha_m)S_m^{\text{old}} + \alpha_m S_m^{\text{opt}}, \tag{3.3}$$

$$\Delta_m^{\text{new}} = (1-\alpha_m)\Delta_m^{\text{old}} + \alpha_m \Delta_m^{\text{opt}}, \tag{3.4}$$

where the optimum sharing parameters (i.e. calculated in P_2) are given by S_m^{opt} and Δ_m^{opt} and the sharing parameters of the previous time window, i.e. S_m^{old} and Δ_m^{old}. α_m present the feature scaling coefficient that we use to calculate the weighted sums. The α_m value becomes critical as it may result in a memoryless resource optimization or a static resource allocation. The efficiency of α_m is measured using the relative distance to optimum (RDO) and calculated as

$$\text{RDO} = \frac{1}{|M|}\sum_{m\in M} \frac{\sum\limits_{i=1}^{N} \xi_m[i] - \xi_m^{\text{opt}}[i]}{\sum\limits_{i=1}^{N} \xi_m[i]}. \tag{3.5}$$

RDO measures how close the performance of the algorithm are with respect to the optimum values. (3.2.c) ensures that for any time slot $\xi_m[n] \geq \xi_m^{\text{opt}}[n]$. Note that for the special case of $\xi_m[n] = \xi_m^{\text{opt}}[n] = 0$, RDO is assumed to be zero. Fig. 3.4 shows the impact of α_m value on the performance of the model for different renegotiation interval lengths. The dynamic scaling coefficient in Fig. 3.4 is calculated in P_2 using

$$\alpha_m = \frac{\left|\sum\limits_{i=n-a}^{n} \xi_m[i] - \sum\limits_{i=n-a}^{n} \xi_m^{\text{opt}}[i]\right|}{\sum\limits_{i=n-a}^{n} \xi_m[i] + \sum\limits_{i=n-a}^{n} \xi_m^{\text{opt}}[i]}, a \equiv (n-1 \mod RI). \tag{3.6}$$

Although our proposed model does not have any limitation in terms of the duration of the renegotiation interval, in order to evaluate the advantages

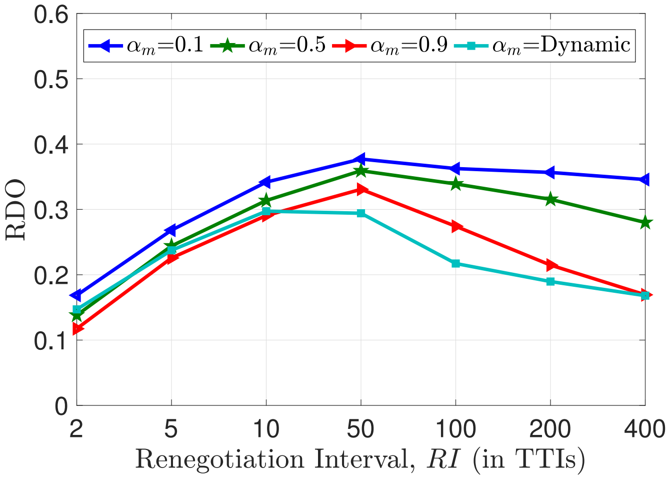


Figure 3.4: *Performance comparison of different scaling coefficients for different renegotiation intervals.*

of dynamic short time scale resource sharing, the RI is set to be between $50 - 200$ TTIs. In Fig. 3.4, where the evaluation of RDO for different α_m and RI values are given, one can see that the dynamic scaling coefficient outperforms the static α_m values between the observed RI region. Therefore, for the reminder of the thesis, dynamic scaling coefficient is applied.

3.4 Evaluation of the proposed negotiation platform

In this section, the proposed dynamic negotiation and trading platform is analyzed with numerous simulations.

3.4.1 Parameters and scenarios studied

The mathematical applicability of our proposed model is investigated for the scenario where the downlink of a base station is shared among three tenants, $|M| = 3$. The set of users K is uniformly distributed throughout the coverage area and it is assumed that at any given time slot n, one of the users of each tenant is active (i.e. $|K_m = 1|$). The simulation horizon is set to be $N = 5000$ TTIs where each time slot, n, is assumed to be 1 TTI. The simulation is run in Matlab while we use Gurobi to solve P_1 and P_2. In a standard commercially available computer, i.e. equipped with i7-4510U CPU and 16 GB RAM, the total run-time of the algorithm (both P_1 and P_2) for one renegotiation interval is 0.998 sec. In order to explore the behavior of the proposed framework, the costs and the budgets are set to have generic values which are in line with their respective real-life counterparts and they are normalized to be between 0 and 100, i.e. $C_{\text{ca}}, C_{\text{op}}, C_{\text{pre}}, B_m \in [0,\ 100], \forall m \in M$. As the real costs and budgets will be similar to these parameters, the behavior of the algorithm is considered to be unchanged.

The communication channel between user k and the base station is assumed to be frequency-flat block fading channel with i.i.d. Rayleigh coefficients. Therefore the channel gains can be modeled as exponentially distributed random values, $|h_k[n]^2|$. Considering Okumura-Hata propagation model, the signal to interferance plus noise ratio of user k at any time instance can be calculated as $\gamma_k[n] = |h_k[n]|^2 \times Pd_k^{-\alpha}/(\sigma^2 + I_0)$ where α represents the the path-loss exponent, σ^2 is the thermal noise, and I_0 is the average interference power. During the calculations, the transmission power is assumed to be P Watts (W), the distance between user and the base station is considered to be d_k meters (m). Using the SINR value, the achievable rate of a user is calculated as $r_k[n] = \log_2(1 + \gamma_k[n])$.

Despite the given utility function in (3.2.c) can be reflecting a set of different aspects, for the sake of simplicity in this chapter we assumed it to be linear to the achieved rate per user namely $U_k(x_k[n],\ r_k[n]) = x_k[n]r_k[n]$. As all the users have unlimited utility functions, i.e. QoE mappings, the users are competing to dominate the resources at any n.

3.4.2 Evaluating the performance of the algorithm

In the first phase of our evaluation, we focused on how close the performance of the proposed heuristic approach is to the performance of the exact model and the fairness of the outcome. In Fig. 3.5 we present the moving harmonic mean of the total gap, i.e. $H(\xi_m, n)$, of our algorithm in comparison to the different feature scaling coefficient α_m. By using the moving harmonic mean, we can capture the peak variations more efficiently with respect to the arithmetic mean. One can observe that the application of dynamic α_m outperforms all the other static approaches, showing that the proposed dynamic scaling coefficient is the better option. Moreover, Fig. 3.5 also contains the results of the optimum approach (cf. continuous pink line in Fig. 3.5), where all the achievable rates are given a priori to the scheduler and then the whole model is optimized at once. Beside outperforming all the other approaches, the optimum algorithm defines the minimum achievable gap with the given infrastructure resources.

A key attribute of any market model is guaranteeing that the resource

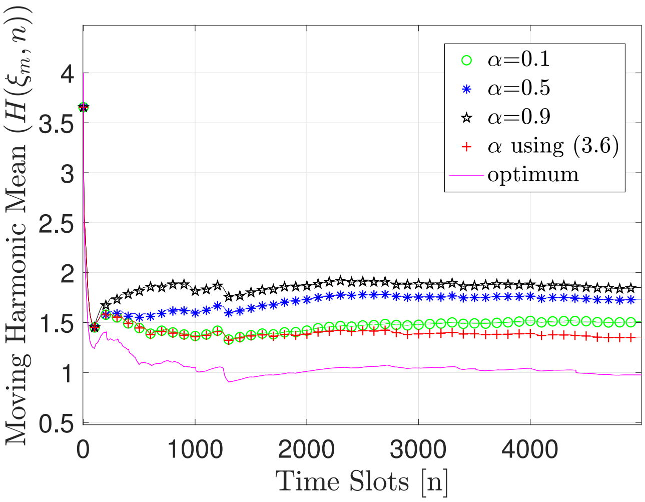


Figure 3.5: *Moving harmonic mean of the total gap over tenants, computed over all the previous time slots up to n where $RI = 100$ TTIs.*

distribution is fair and none of the players is favored. Thus as a second aspect, we are analyzing the inter-tenant fairness of the given model. Fig. 3.6 and Fig. 3.7 outline the variations of the average achieved rates of the tenants, and underline the inter-tenant fairness in terms of achieved rates. More specifically, Fig. 3.6 depicts the cumulative distribution function (CDF) of the achieved rates per tenant. One can see that, the CDF of the achieved rates is almost the same for all the tenants. As a consequence of maximum-rate scheduling and unbounded utility functions, the complete network resources are assigned to only one user with the highest achievable rate $r_k[n]$, which leads to the equal access to the tenants in an overly crowded network. Thus, in total, each tenant obtains resources for $1/|M|$ of the time. Therefore, each tenant can obtain resources approximately 33% of the time.

Lastly, the changes in the moving arithmetic mean of ξ_m, $A\left(\xi_m, n\right)$, over all the time slots till n are given in Fig. 3.7. The key point of Fig. 3.7 is the fact that despite the variations in the achievable rates, after a relatively brief period (i.e. until $n = 2000$ TTI), the functions gain a steady state characteristic. This steady state characteristic shows that our two step algorithm can provide a stable business platform. Note that after the model converges to the steady state, no additional gain can be achieved simply by changing the sharing parameters. The respective investigation is performed in the following chapter.

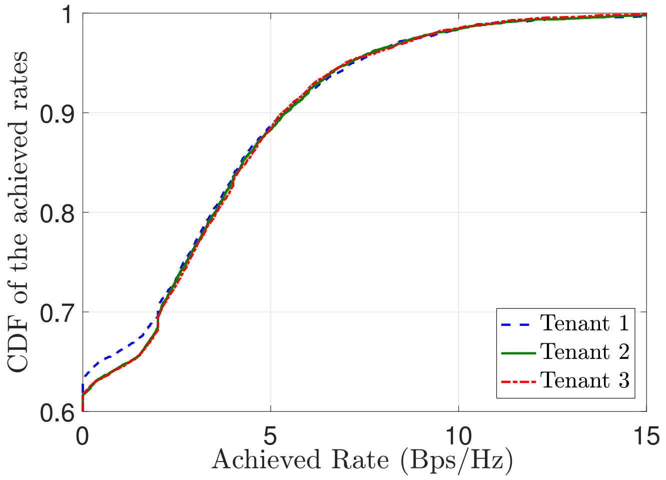


Figure 3.6: *Empirical cumulative distribution function (CDF) of achieved rates.*

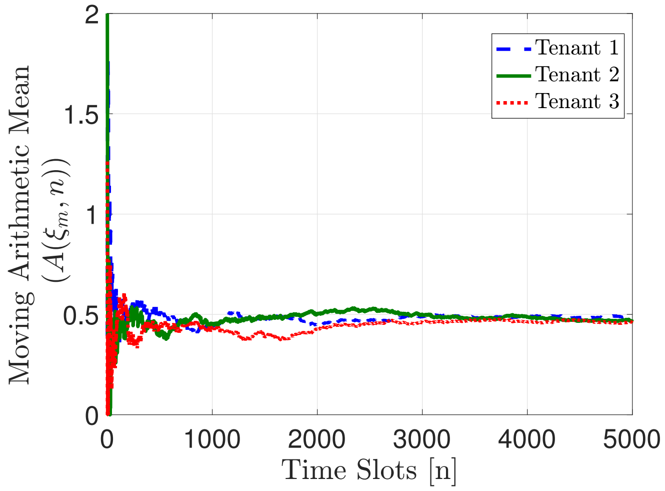
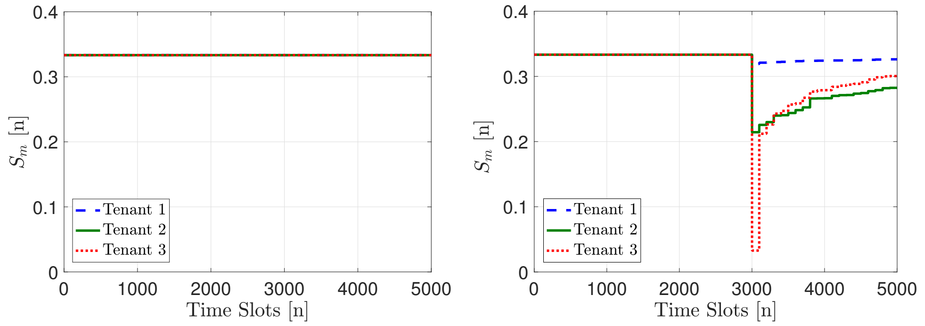


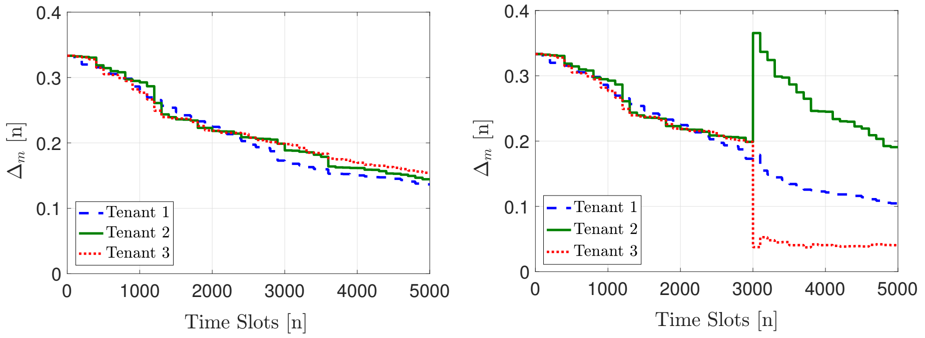
Figure 3.7: *Moving arithmetic mean of the total gap for* $RI = 100$ *TTIs.*



(a) *Sufficient budget scenario*

(b) *Insufficient budget scenario*

Figure 3.8: *Variation in S_m over time under two cost setups.*



(a) *Sufficient budget scenario*

(b) *Insufficient budget scenario*

Figure 3.9: *Variation in Δ_m over time under two cost setups.*

3.4.3 Behavior of our proposed framework under budget insufficiency

The impacts of budget insufficiency on sharing parameters are given in Fig. 3.8 and Fig. 3.9 where the instantaneous values of S_m and Δ_m are observed for two cost setups. In Fig 3.8a, all tenants possess sufficiently large budgets in order to own the resources, therefore, we observe that the guaranteed sharing ratio S_m is shared equivalently among the tenants as a result of (3.2.a). Moreover due to (3.2.f), equivalent distribution of sharing parameters produces same CapEx cost for all the tenants, showing equal contributions from tenants in order to maintain the cooperation. On the other hand, Fig 3.8b shows the case where all the costs are doubled and consequently obtaining the resources becomes more expensive. For this scenario, due to insufficiency of their budgets, the tenants are allowed to decrease their S_m and decrease their expenses on CapEx. Note that this decrease in S_m is only allowed if there is a budget shortfall in order to avoid under utilized resources, but for all the other cases, the S_m distribution is strictly equivalent among tenants.

Fig. 3.9 reports the variation of Δ_m for two different cost value setups. In Fig. 3.9a, the tenants posses the required budget to set their S_m symmetrically and since the time window is high, they can reach their utility target. Therefore, we observe that their Δ_m is also decreasing in an equivalent manner. On the other hand, when the costs are doubled, cf. Fig. 3.9b, the tenants have difficulties in satisfying their utility expectations purely based on S_m. Therefore, when they need to decrease their S_m, we also observe their Δ_m is also changing and they rely on opportunistic access strategy. More specifically, at $n = 3000$ TTI, the insufficient budget is forcing one tenant (Tenant 3) to decrease its S_m and Δ_m. This way, the tenant with the budget insufficiency is decreasing its spending on resources. On the other hand, due to the fairness among tenants, as one tenant is allowed to decrease its S_m, the other two tenants are given the flexibility to decrease their sharing parameters as much as Tenant 2. Consequently, we observe that the other two tenants also decrease their S_m and increase Δ_m in order to minimize their spending on the resources.

3.4.4 Economical impact of sharing

The cost efficiency that can be achieved through sharing with different scenarios along with the total cost of no sharing scenario is given in Fig. 3.10. Here, for the no-sharing scenario, the tenants are considered to own their own infrastructure and they do not involve themselves in the sharing process. Therefore, the total cost of a tenant is the summation of the total CapEx cost of the base station and the OpEx cost of the resources. However, as for the no-sharing scenario the tenants can serve their users according to their needs, there is no observed pressure cost. In the static sharing scenario, the tenants share the resources in a predefined manner, $1/3$ share per tenant, and there is no dynamic negotiations. As they share the same infrastructure equivalently, it is assumed that they bare the cost of any expansion need equally. The last four scenarios in Fig. 3.10 depict the effects of renegotiation interval, RI, on the overall costs of the tenants.

Fig. 3.10 confirms that both the sharing strategy and the flexibility of the scheduler in resource allocations have impact on the overall cost efficiency of the network. The decrease in RI implies that the scheduler has less

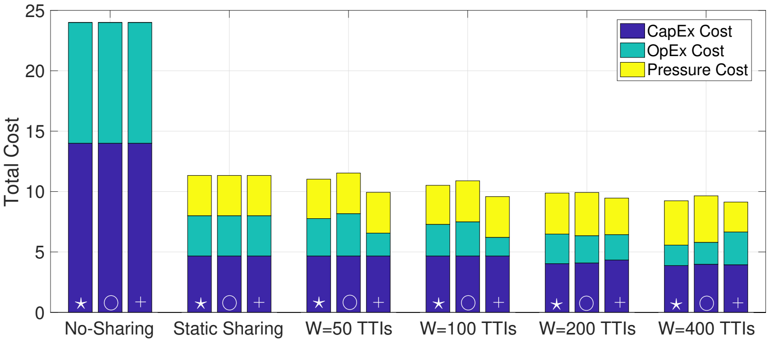


Figure 3.10: *Distribution of the costs over tenants for different sharing options. In the figure, '$\star$' indicates Tenant 1, '$\circ$' indicates Tenant 2 and, '$+$' indicates Tenant 3.*

time to satisfy the sharing parameters. This lack of flexibility forces the scheduler to perform inefficient resource allocation in order to satisfy the negotiated terms. In line parallel to the increasing flexibility, the longer RI produces higher cost efficiency. In the extreme case of full-sharing scenario, i.e. $\Delta_m = 1 \; \forall m \in M$, the highest cost efficiency is reached as the scheduler can perform the highest efficiency. A key aspect is the fact that the proposed model achieves higher efficiency for larger RI than smaller ones. However, for very long RI duration, e.g. more than 200 TTIs, the increased computational complexity of the framework puts the real time implementation of the algorithm into risk. Note that an analysis of the time complexity of the algorithm is given in Chapter 4.4.2

3.5 Summary

In this section we focus on design of a flexible negotiation platform that can exploit the full potential of network sharing without loss of service quality. Consequently, we proposed a novel dynamic pricing and resource sharing model for multi-tenant networks. As a key contribution, the proposed framework brings a new interpretation to the service level agreements and the negotiations among stakeholders, where the resource shares of each tenant is no longer a strict SLA constraint but instead it can be dynamically adjusted based on the traffic conditions and the business model of the tenant. Thus, the proposed model can reach maximum efficiency in resource usage by exploiting the dynamism of the wireless network conditions. The novel pricing model allows tenants to adjust their budget usage in real time and also scales the resource prices in line with the instantaneous resource demand. The proposed pressure cost structure prevents the monopolization of the network resources by penalizing the tenants with large budgets when they try to artificially increase the unit costs by setting unrealistic utility targets. The two-step implementation of the model provides a reactive adaptation to the variations in the network conditions, while the proposed flexibility ensures the instantaneous compensation of the inefficient sharing parameters.

CHAPTER 4

Service differentiation in a sliced multi-tenant network

4.1 Introduction

A broad range of services and devices are envisioned to be sharing the same wireless network in 5G. However, the conflicting priority levels and the different QoS expectations of these services make it challenging for the mobile operators to maintain inter-service fairness while increasing efficiency. Moreover, the ambitious expectations from 5G can only be satisfied by providing dedicated resources that can be customized to serve a particular service. However, the conventional 'one type fits all' approach cannot customize the network resources in order to provide the best possible service. The advances in the virtualization technology make it possible to virtualize the radio network resources and on-demand resource provisioning [19]. Built upon the network virtualization technology, *network slicing* proposes vertically grouping and reserving a set of the network resources[1] in order to be optimized for a single type of service. As each service can have its

[1] As previously detailed in Chapter 4.5, the network slicing concept includes end to end slicing, including both the access network and the core network. In this thesis we focus on the radio access network slicing which is the main bottleneck in the multi-tenant network.

own virtual network that can be shaped in line with its needs, the desired harmonious coexistence among different services in the same network can be achieved [101]. However, similar to any resource sharing problem, the key challenge in network slicing is how to assign resources to different slices in order to achieve the highest efficiency. The easiest way of slicing the network, i.e. statically slicing the network resources, provides dedicated resources to each slice in a permanent manner. However, regardless of how well these slice allocations are determined, the static nature of this allocation scheme requires a perfect estimation of the current and future needs of the mobile operators. This approach almost always ends in overprovisioning of the network resources which is increasingly becoming non-sustainable in the current techno-economic environment. The profitability of the network provisioning as well as achieving maximum QoE among users strictly depend on scaling the resource allocations in the slices in line with the variations in the channel conditions, the fluctuations in the traffic demand and traffic mix [73]. Therefore, performing flexible resource reallocation to the slices, i.e. *dynamic network slicing*, is of crucial importance.

The major issue in dynamic network slicing is determining how to assign the available network resources to the different slices with various traffic conditions and priorities. Although always assigning the resources to the slices with the best channel condition can boost the spectral efficiency, it also hardens satisfying the QoS requirements of the services. A well-known approach to handle this impasse is prioritization of services with strict QoS requirements over the services with more relaxed requirements. Although prioritization works well in cases with highly asymmetric priorities such as coexistence of emergency services and elastic services, the inter-service prioritization can be a further challenging concept in a more heterogeneous context. On one hand, the limited resources force the operators to prioritize some services in order to guarantee that their QoS expectations will be met even in the peak demand time. On the other hand, the level of prioritization can easily end up in very low spectral efficiency and loss of profit. In this PhD work, we argue that regardless of how well the sharing approach is designed, the inter-service fairness and the high resource usage efficiency cannot be simultaneously guaranteed simply by using either QoS requirements or priorities of services. Instead, the QoS expectations and the service priorities need to be mapped into a single QoE model that can let the scheduler dynamically determine the best user to choose at any given time.

The possibility to dynamically slice the networks also allows the network operators to deepen their sharing model and further decrease their costs by sharing a sliced network. However, it is clear that sharing a sliced network among a set of tenants also increases the complexity of the model since the tenants also bring their own priorities and business strategies. Extending our previous dynamic infrastructure sharing model, in this chapter we are proposing a novel automatized dynamic network slicing and slice trading framework for next generation wireless networks. Although the dynamic network slicing problem can be modeled for each layer of the network, we focus on the RAN slicing problem. Note that, however, the proposed model can be applied to the different layers of the network with minor changes. We also demonstrate the impact of specialized tenants, that are envisioned to be the key players in 5G and beyond networks [27], to the sharing agreements.

The materials of this chapter are published in [3] and [5].

4.1.1 Specific research questions

In this chapter, the analysis on the main question of "How can the tenants use the shared resources to serve different services with contradicting priorities?" is performed under three following research questions.

- How can the QoE mappings of different service types be performed? (Explored in Section 4.2)
- How can the tenants differentiate their services to attract more customers from the different portions of the market? (Introduced in Section 4.3 and explored in Section 4.4)
- What are the implications of service differentiation on the sharing agreements? (Evaluated in Section 4.4)

4.1.2 Chapter outline

A novel piece-wise linear QoE mapping of the envisioned services is introduced in Section 4.2. First we have proposed a generic utility function that can be customized in order to model different services, and then we demonstrated the QoE mappings for four different services. Using these services, in Section 4.3, the previously defined network sharing algorithm is extended to a network slicing and slice trading algorithm where the tenants can negotiate in line with their service specifications. The performance of the proposed model is evaluated using a large set of Monte-Carlo simulations. Finally Section 4.5 summarizes the key achievements of this chapter and concludes it.

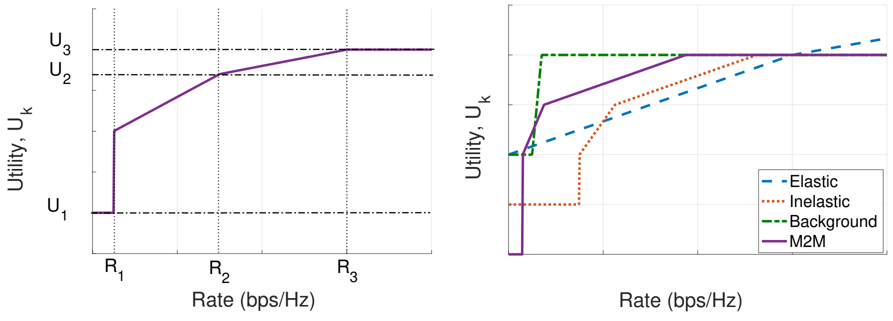
4.2 Envisioned service heterogenity and QoE mapping

Depending on their customer profile and needs, each service is defined by its own QoS expectations. On the other hand, the tenants can differentiate their services by giving different priorities or achieved rate expectations to different services. Therefore, the different QoS expectations as well as the service priorities are required to be integrated into a single form that can be used by the scheduler. In literature, the QoS discussions for the envisioned services for 5G networks are mainly on the rate expectations and delay constraints [45] [56]. In our work we assume that the QoS per service is mainly defined by the achieved rate of the user, while the delay constraint is indirectly defined per tenant. In case tenants need to change their delay constraints, they can do it simply by declaring a different time window (W_m) per service. The service based QoS constraints with the tenant specific policies and priorities are mapped into a QoE mapping. The average QoE per tenant, i.e. average achieved *utility* per tenant, is determined as a summation of their respective user's achieved utilities whereas the achieved utility per user is determined by the user's achieved rate. Assuming that the users are active for the whole time window, the average achieved rate of a user k at any time slot n can be determined as follows.

$$R_k[n] = \big(\frac{1}{a+1}\sum_{i=n-a}^{n} x_k[i] r_k[i]\big), \tag{4.1}$$

where a is determined using the previously defined expanding time window, i.e. $a \equiv n-1 \mod W_m$.

In order to model the heterogeneity of the envisioned services, a generic utility function $U_k(R_k[n])$ is designed (cf. Fig. 4.1a). Although more complicated mappings are possible, for the sake of mathematical tractability of our model, we have used six parameters, namely R_1, R_2, R_3, U_1, U_2, and U_3. By customizing these six key parameters, a variety of services can be modeled. Here, R_1 represents the minimum average achieved rate that is required for a system to be considered active. Below this average achieved rate, the system is considered to be inactive and produces the negative utility of U_1. Once the service is activated, it is considered to be in the standard quality region, where the utility increases rapidly with the achieved rate. R_2 is the transition point between the standard quality and the high quality where the service provides the utility of U_2. The key difference between the standard quality region and the high quality region is the ease of observing the change in the service quality with the increase in the achieved rate. For



(a) *Generic utility function*

(b) *Exemplary utility functions*

Figure 4.1: *Proposed generic utility function and exemplary utility functions.*

further clarification of our design, video streaming can be considered as an example. The video quality between a 144p video and a 480p video is quite clear whereas the difference between a 4K video and 1080p video is not easily noticeable. Therefore, in our model, after achieving R_2 the gradient of the utility function is decreased in order to reflect to the decreased visibility of the quality increase. Finally, R_3 marks the saturation point of the service, where the service has reached its peak utility value (U_3) and after which the utility cannot be increased anymore. Despite the proposed utility function is purely considering the average achieved rate, the latency requirements of particular constraints are indirectly integrated to the proposed model by considering the average achieved rate while calculating the utility. Thus it is possible to argue that the latency constraint per tenant is W_m. In our proposed QoE mapping, W_m, R_1, R_2 and R_3 are used to model the QoS expectations of the services.

Assuming that the base station uses a scheduler that tries to maximize this achieved utility (or average QoE), the gradient of the utility function in Fig. 4.1a determines the priority of the service. Thus, the selection of U_1, U_2 and U_3 sets the priority level of the individual services. This prioritization mechanism and the mapping between achieved rate and the utility provide a new level of flexibility to the scheduler during the resource allocations. More specifically, in most of the cases the scheduler will serve the services with the highest priority first. However, if the channel conditions of a low priority service is much better than the high priority service then it will be served first.

In order to capture the heterogeneity in the 5G networks, we have considered 4 traffic categories, namely: elastic services, inelastic services, background services, and machine to machine (M2M) services. In the next parts, we will detail the customization of the generic utility function in Fig, 4.1b and explain the rationale behind this design.

4.2.1 Elastic services

By definition, the elastic services do not have any strict rate or delay constraint and their service requirement does not have a maximum achieved rate after which the service quality cannot be improved. Therefore, $R_1 = 0$ and $R_3 \to \infty$ for the utility function of the elastic services, as depicted in Fig. 4.1b. Despite they do not have a saturation point after which the utility of the service cannot be increased, we assumed that the utility increase of the service is fairly lax after reaching U_2 in order to maintain a level of fairness among services while providing a level of flexibility to the tenants since they can always increase their total utility by serving the elastic services. Internet browsing or download of a file can be considered as the representative applications of this service type.

4.2.2 Inelastic services

A classic yet very simple example of inelastic service type is video streaming. A relatively high average achieved rate is required in order to assume that the service is active – in other words in order to start the video. Consequently, as visualized in Fig. 4.1b, R_1 is considered to be very high with respect to the rest of the services, during which the service produces the negative utility of U_1. Once the service is activated, i.e. $R_k[n] \geq R_1$, the utility function obtains a quite steep slope until $R_k[n] = R_2$ in order to reflect the visibility of the quality change with the increase of achieved rate (e.g. the perceived difference between 144p and 720p videos). For the rates higher than R_2, the quality change perceived by the user will be small, e.g. the difference between 720p and 1080p, thus the region between R_2 and R_3 is designed to be gradual. The further increase in the average achieved rate after R_3 is considered not to be fruitful since the required highest quality is achieved for this class. Music streaming and virtual reality applications are also some examples of this traffic type.

4.2.3 Machine to machine services

M2M type communication contains a variety of devices and the traffic types. Therefore, in order to model this heterogeneity and the complexity of their traffic type, we aggregate the M2M traffic under three major groups, using which we proposed the generic utility function for M2M service. More specifically we assumed that each M2M service request contains a mix of all these three groups. The first group envisioned in M2M is emergency communications which requires relatively low achieved rates but has strict delay constraint. Therefore, R_1 is set to be relatively small, but the produced utility U_1 is smaller than any other service in order to reflect that not serving these service can result in dangerous situations. Since M2M type has the smallest U_1 value, it is guaranteed that as long as there is any pending critical communication need, the scheduler would always give highest priority to this user. The interval $[R_1,\ R_2]$ is considered to be representing the sensors that transmit relatively small traffic periodically. The prioritization of this traffic is given by the slope of the utility function in this interval which provides the delay constraint while guaranteeing the achieved rate demand. The third and last group that we consider are the sensor aggregation nodes, which are transmitting relatively large traffic without a strict periodicity. Thus achieving a high average rate is considered to be more critical for this type than having a strict periodicity. Therefore, in the modeling of this group, the slope of interval $[R_2,\ R_3]$ is considered to be smaller than the rest of the M2M regions. The visualization of the M2M traffic is given in Fig. 4.1b. Being the broadest category, M2M contains a multitude of use cases, including connected cars, e-health and public security.

4.2.4 Background services

Protocol synchronization or electronic market feed (i.e. high frequency trading messages) can be considered as the use cases for this type of service. It requires relatively low achieved rates, and once this low rate is satisfied, it directly reaches U_3. As there is only standard quality requirement for this type of service, R_2 and R_3 coincide, as showed by the dashed green line in Fig. 4.1b. Considering the non-critical nature of this service type, R_1 is set to be zero, indicating that not serving this type of traffic would not cause a utility penalty. However, due to the very low rate requirement and the ease of reaching U_3, this service has the highest priority.

4.3 Dynamic network slicing and trading in a multi-tenant network

This section illustrates the extension of the mathematical model in (3.2.a)–(3.2.i) to accommodate multiple services.

4.3.1 Mathematical formulation

Our model proposed in the previous section can provide the required flexibility and efficiency in network slicing. During the design of the model, we have assumed that the tenants only have elastic services with unlimited utility functions. Therefore, the fairness among tenants was artificially ensured by introducing S_{max} variable that indirectly sets an upper limit for the maximum guaranteed sharing ratio that a tenant can choose and a tenant can only deviate from this imposed boundary in case of budget insufficiency. However, in this chapter we have considered multiple services with finite QoS expectations and utility outputs (cf. Section 4.2). Therefore, as the first step, the S_{max} variable is omitted from the model along with the respective constraint (3.2.b). Note that in Section 4.4, we compared the two cases where the S_{max} is used and not in order to show why this imposed structure is no longer needed.

Integration of non-elastic services requires a better modeling for the tenants' incentive to trade the resources. In our previous model, the users are assumed to produce only elastic traffic with lax delay constraints, therefore the tenants are eager to trade all the resources in order to maximize the spectral efficiency and minimize the total cost (as formulated in (4.2)).

$$0 \leq \Delta_m \leq \max(S_m, 1 - S_m). \tag{4.2}$$

The main motivation behind this constraint was the fact that in a network with one service type with monotonically increasing utility function, the tenants would try to dominate the resources. Therefore, while the guaranteed sharing ratio is indirectly being limited, the rest of the resources are sold in an opportunistic approach. In this way, the tenants could trade their resources in line with the instantaneous characteristic of their traffic. On the other hand, in this chapter, the tenants can differentiate their guaranteed resource shares in line with their traffic mix and the utility expectations. Consequently their trade incentives depend on the traffic mix they serve. For example, a tenant who mostly serves inelastic traffic would not want to risk the available resources since it could easily end up in QoS violations of the service. Therefore, in this chapter the maximum average deviation from the guaranteed sharing ratio is limited by the total resources that are assigned to the elastic services. The new constraint is given as,

$$0 \leq \Delta_m \leq \frac{1}{a_m + 1} \sum_{i=n-a_m}^{n} \left(\sum_{k \in K_{m,\text{elastic}}} x_k[i] \right), \ \forall m \in M, \tag{4.3}$$

$$\min_{x_k[n], S_m, \Delta_m} \quad \sum_{m \in M} \xi_m[n] \tag{4.4.a}$$

$$\text{s.t. } U_{\text{th}} - \sum_{k \in K_m} U_k(R_k[n]) \leq \xi_m, \; \forall m \in M, \tag{4.4.b}$$

$$\epsilon_m[n] = \left(\frac{1}{(a_m + 1)} \sum_{i=n-a_m}^{n} \sum_{k \in K_m} x_k[i] \right) - S_m, \; \forall m \in M, \tag{4.4.c}$$

$$|\epsilon_m[n]| \leq \Delta_m, \quad \forall m \in M, \quad \forall n \in N, \tag{4.4.d}$$

$$\begin{aligned} \sum_{i=n-a_m}^{n} & (S_m(C_{\text{ca}} + C_{\text{op}}) + \epsilon_m[i]C_{\text{op}} + f_{\text{pre}}(C_{\text{pre},m}, \; \xi_m)) \\ & \leq B_m(a_m + 1), \forall m \in M, \end{aligned} \tag{4.4.e}$$

$$0 \leq \Delta_m \leq \frac{1}{a_m + 1} \sum_{i=n-a_m}^{n} \sum_{k \in K_{m,elastic}} x_k[i], \quad \forall m \in M, \tag{4.4.f}$$

$$\sum_{k \in K} x_k[n] \leq 1, \;\; x_k[n] \geq 0, \;\; \forall k \in K, \tag{4.4.g}$$

$$\sum_{m \in M} S_m \leq 1 \;, \;\; S_m \geq 0, \;\; \forall m \in M, \tag{4.4.h}$$

where $K_{m,\text{elastic}}$ represents the set of users with elastic service demand of tenant m. Note that this change can be considered as a generalized version of the model proposed in the previous chapter. In case the tenant has only elastic traffic, the proposed constraint will be the same as the previous one. The complete model is given in (4.4.a)–(4.4.h).

The optimizer tries to minimize the total gap through the resource allocations, i.e. (4.4.a). By the proposed joint optimization approach, the resource allocation is scaled in accordance to the priority of the services as well as the achievable rate per user. Using the defined utility functions, $U_k(R_k[n])$ in 4.2, (4.4.b) defines the total gap of the tenants. Here the value of $U_{\text{th,m}}$ is set to be $\sum_{k \in K_m} U_{3,k}$, meaning that the main objective of all the tenants is to fully serve their users. A key aspect is the selection of U_3 per user. The tenants can further differentiate their services by selecting asymmetric U_3 parameters for their services, giving higher priorities for some of the traffic while decreasing it for others. However, this case is not covered in this thesis, as it leads to *non-netneutral*[2] network management. Although,

[2] Netneutrality fundamentally argues that all the network traffic is treated fairly, regardless of the content [51]. Namely, the infrastructure provider cannot differentiate the two identical traffic based on the contents or the related user.

the paid prioritization of the particular users is not covered in our framework, the proposed framework can be customized in line with the evolution of the regulations. Note that although the elastic-service users can achieve a utility value greater than U_3, inter-service prioritization, i.e. the gradients of the utility function, forces the scheduler to give higher importance to non-elastic services. Constraint (4.4.d) sets the upper bound of the maximum instantaneous deviation, $\epsilon_m[n]$, that is set in Constraint (4.4.c).

Constraint (4.4.e) sets the total cost-budget balance for the framework. Since the the observed gap depends on the instantaneous conditions of the network, unlike CapEx and OpEx, it is very hard to predict its value in advance. On the other hand, as previously discussed, enabling the price predictability is crucial for the business applications. Therefore, the pressure cost, $f_{\text{pre}}(C_{\text{pre},m}, \xi_m)$, is calculated using the gap calculated in the previous time window. Finally, as detailed in the preivous chaper, (4.4.g) and (4.4.h) define the physical constraints in the framework, indicating that the total assigned resources cannot exceed the available resources and the infrastructure provider cannot sell the resources that do not exist, respectively.

Although the proposed model can capture the techno-economic dynamics that characterize the negotiation process, the real time application of it is rather challenging due to the computational complexity. Thus, as detailed in the previous section, a two step algorithm is designed where the real time resource allocation decisions are separated from the negotiations among tenants. Two different implementations of the proposed framework with regard to the active variables, i.e. P_1 and P_2 are summarized as follows,

$$\begin{aligned} P1 := &\min_{x_k[n]} \xi_m[n] \\ &\text{s.t. } (4.4.\text{b}), (4.4.\text{c}), (4.4.\text{d}), (4.4.\text{e}), (4.4.\text{g}) \\ P2 := &\min_{x_k[n], S_m, \Delta_m} \xi_m[n] \\ &\text{s.t. } (4.4.\text{b}), (4.4.\text{c}), (4.4.\text{d}), (4.4.\text{e}), (4.4.\text{f}), (4.4.\text{g}), (4.4.\text{h}) \end{aligned}$$

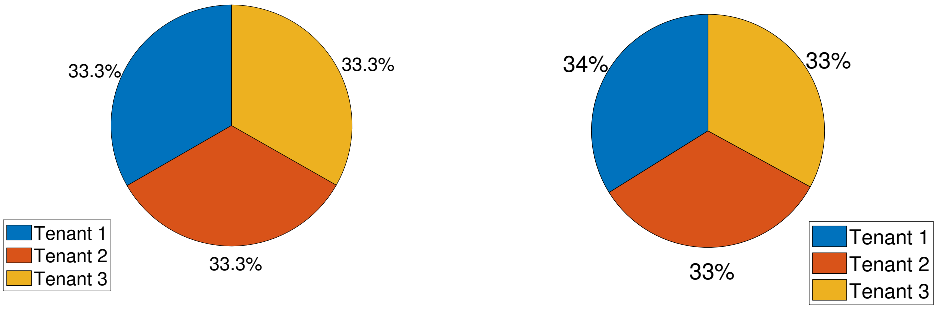
4.4 Numerical evaluation

The service differentiation among tenants becomes a major concern in network sharing which could lead a decrease in tenant's business potential due to the undifferentiated QoS levels. Thus a key aspect in our model is enabling the service differentiation among tenants. Since the key aspect in our framework is dynamic RAN sharing, the service differentiation is investigated in terms of tenant's capability to change the assigned resources by customizing the service parameters. Therefore, we analyze the behavior of the proposed framework in a symmetric scenario, where all the tenants choose the same quality parameters. After the investigation of this symmetric scenario, we investigate the impact of utility function changes, namely one tenant expects higher achieved rates for one of its services. The key aspect of this scenario is the tenant with very high expectations should not impact other tenants. Next the premium service, where a tenant chooses a higher $U_{\text{th,m}}$ than the other tenants is investigated.

4.4.1 Fairness among services

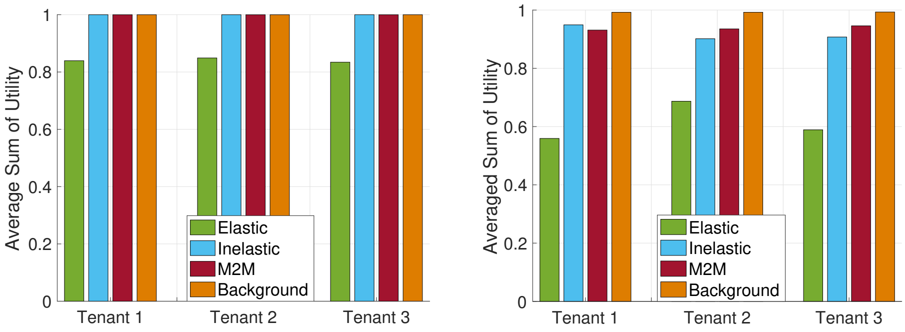
As we previously discussed, providing equal opportunities to the tenants for obtaining resources is one of the key enablers of the inter-tenant fairness. In the previous chapter, due to the traffic elasticity, inter-tenant fairness is achieved by using an artificial parameter, S_{max}, that enables a correlation between guaranteed resource share selections of the tenants. However, in this chapter, the heterogeneity of the defined services and their utility function definition remove the need for an artificial fairness constraint. In order to evaluate whether the market driven fairness model proposed in this chapter can provide tenant based fairness we have compared the resource allocations and the average sum of utilities for two different approaches. Fig. 4.2 shows the resource distribution between the cases where we used S_{max}, cf. Fig. 4.2a, and where we have used market driven S_m determination, cf. Fig. 4.2b. For both of the cases the resources are equally distributed among the tenants, indicating that the proposed QoE driven approach can provide fairness among tenants.

Moreover, in Fig. 4.3, we investigated the inter-service fairness for the two different approaches. Despite the variation of the achieved utilities by tenant between two cases due to the channel conditions, we see a symmetric distributions of achieved utilities among tenants and the average sum of utilities of the same services is similar for all the tenants. Therefore, the proposed QoE based approach can provide fairness among the tenants without any explicit need for a constraint or an artificial parameter. Note that, by removing S_{max}, we provide a further flexibility to our model, which is achieved by the freedom to choose the most appropriate guaranteed sharing parameter for each tenant.



(a) *Equivalent guaranteed resource share* **(b)** *Market driven guaranteed resource share*

Figure 4.2: *Average resource distribution per tenant for two different fairness models.*



(a) *Equivalent guaranteed resource share* **(b)** *Market driven guaranteed resource share*

Figure 4.3: *Average sum of utility per tenant under two different fairness models.*

4.4.2 Time complexity analysis

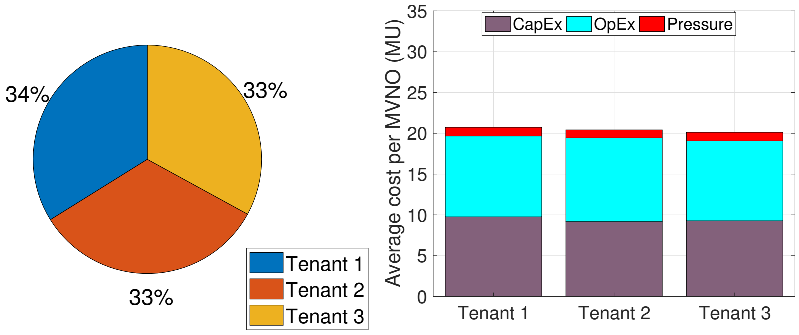
In order to have an estimation of the TTI level, in this section, we have analyzed the time complexity of the proposed framework, depending on the renegotiation interval length, RI. The renegotiation interval length determines the size of the optimization problem in P_2, therefore very large intervals can result in very high time complexity. Note that P_2 is modeled to guide P_1, i.e. the real time resource allocation problem, thus it is run offline.

Table 4.1 depicts the total calculation times of P_1 and P_2 for $|K| = 12$ users and $|M| = 3$ tenants. For this scenario, all the parameters are chosen to be equivalent among tenants (i.e. symmetric traffic scenario). The simulator is run in Matlab, whereas, the optimization problems (i.e. P_1 and P_2) are solved in Gurobi optimizer [35]. Note that the results are obtained from an Intel 2.4 GHz PC with 6 GB of RAM.

Table 4.1: *Effects of renegotiation interval on computation time*

Renegotiation Interval, RI	P_1 duration(sec)	P_2 duration (sec)
5 TTIs	0.0015	0.0431
25 TTIs	0.0012	0.1923
50 TTIs	0.0016	0.5069
80 TTIs	0.0011	1.4832
100 TTIs	0.0015	2.4412

The results report that the computational time of P_2 is strictly depending on the renegotiation interval while no major impact has been measured for the computation time of P_1. Therefore, in order to guarantee a successful implementation of the proposed model as a real time model, P_2 can be run in more powerful machines or a simple heuristic method can be implemented to the framework. Note that since the main problem of dynamic resource sharing and negotiating is divided into two sub-problems (P_1 and P_2), the performance loss due to the application of a heuristic algorithm in P_2 has limited impact on the real time resource allocations (P_1). More specifically, assuming $S_m,\ \Delta_m > 0\ \forall m \in M$, the real time scheduler would have the necessary flexibility to tolerate the performance loss in P_2 up to 20%. However, the algorithm's tolerance also depends on the business model of the tenants, as it may cause higher costs for them.



(a) *Average resource distribution per tenant*

(b) *Average total cost per tenant*

Figure 4.4: *Average resource distribution and average total cost per tenant for* $|K| = 12$.

4.4.3 Symmetric traffic scenarios

This section investigates the behavior of the proposed model when it contains a mix of different services with different QoS expectations and priority levels. In this scenario the base station is shared among $|M| = 3$ tenants with symmetric traffic mix and equivalent budgets. All the tenant specific parameters, e.g. rate expectation, respective utilities for services and $U_{\text{th,m}}$, are assumed to be same for all the tenants. As a result of this symmetry, the tenants receive similar resource shares by paying same costs (cf. Fig. 4.4). The symmetric outcome of the framework for equivalent inputs confirms that our algorithm enables fairness among tenants given that their expectations are similar.

An important aspect to note is that while the proposed model provides inter-tenant fairness, the service prioritization is still being preserved, namely, as reported in Fig. 4.5, the same service from different tenants achieves equivalent utilities. Moreover, the resource scarcity, i.e. the lack of resources to fully satisfy all the services, mostly impacts the elastic services as they have a lower priority in comparison to the rest of the services. Nevertheless, as a direct result of the inter-service prioritization, it is also visible that inelastic and M2M traffic cannot either reach their maximum utility, $U_3 = 1$. More specifically, during the design of services, in order to gain a level of fairness among services, the inelastic and M2M services are designed to have the same priority once they reach a standard quality level, that is indicated by U_2.

Fig. 4.6 presents the steady state behavior of the proposed algorithm in an overloaded network. Despite the same tenant specific parameters are

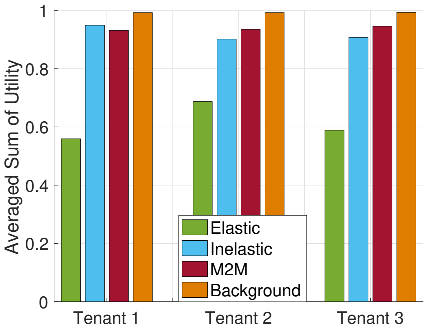
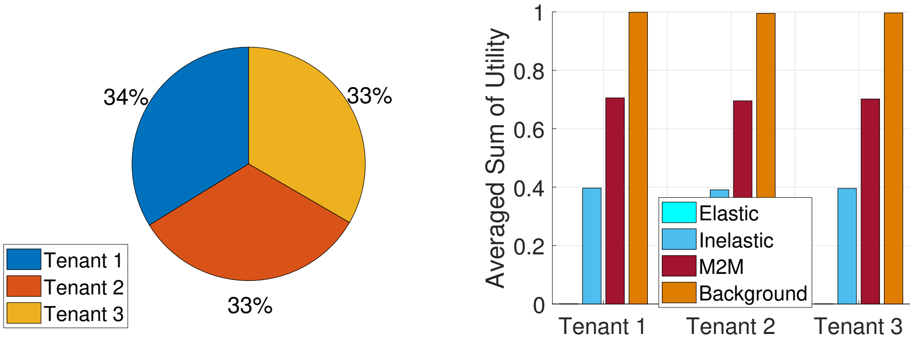


Figure 4.5: *Average achieved utility per service per tenant.*

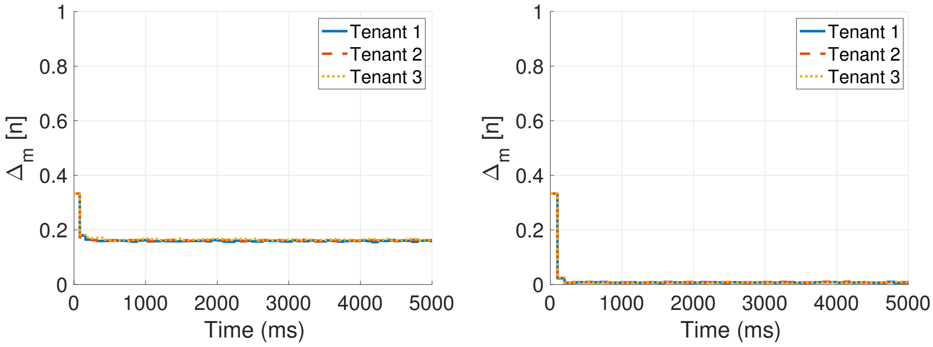
used, i.e. B_m, $U_{\text{th,m}}$, the number of users that each tenant serves is increased to $|K_m| = 16$ (indicating $|K| = 48$ users in total) with the same capacity as described in Fig. 4.4. The decrease in resource availability due to the number of users makes it harder for non-elastic services to be able to reach the standard quality. Despite this strong competition among non-elastic services, Fig. 4.6a shows fairness in the overall resource distribution among tenants. Fig. 4.6b depicts that as a result of the resource scarcity, the elastic services cannot receive any resource, which is a direct result of prioritization. To be precise, unlike Fig. 4.4 where the standard quality for non-elastic services can be achieved, in Fig. 4.6, the inelastic services cannot reach their standard quality expectations (i.e. $U_2 = 0.7$). Thus, the main priority of the scheduler shifts from maintaining fairness among services to achieving the standard quality for non-elastic services.

Fig. 4.7 shows an impact of high load on the tenants sharing parameters. To be precise, the tenant's willingness to trade their resources decreases as their probability to serve their critical services decreases. Therefore, their Δ_m decreases practically down to 0, in order to guarantee that their resources would not be assigned to other tenants. This difference between Fig. 4.7a and Fig. 4.7b, mainly impacts the flexibility possessed by the scheduler. As Δ_m increases, the scheduler can provide resource allocations that can increase the spectral efficiency. On the other hand, for the special case of $\Delta_m = 0$, the scheduler provides the minimum spectral efficiency, indicating that the total costs of the tenants will be higher (since they will require more resources in order to satisfy all the services).



(a) *Average resource distribution per tenant* **(b)** *Average utility per service per tenant*

Figure 4.6: *Average resource distribution and average utility per service per tenant for* $|K| = 48$



(a) Δ_m *variation for* $|K| = 12$ **(b)** Δ_m *variation for* $|K| = 48$

Figure 4.7: *Adaptation of* Δ_m *to the increasing traffic*

4.4.4 Impact of renegotiation interval

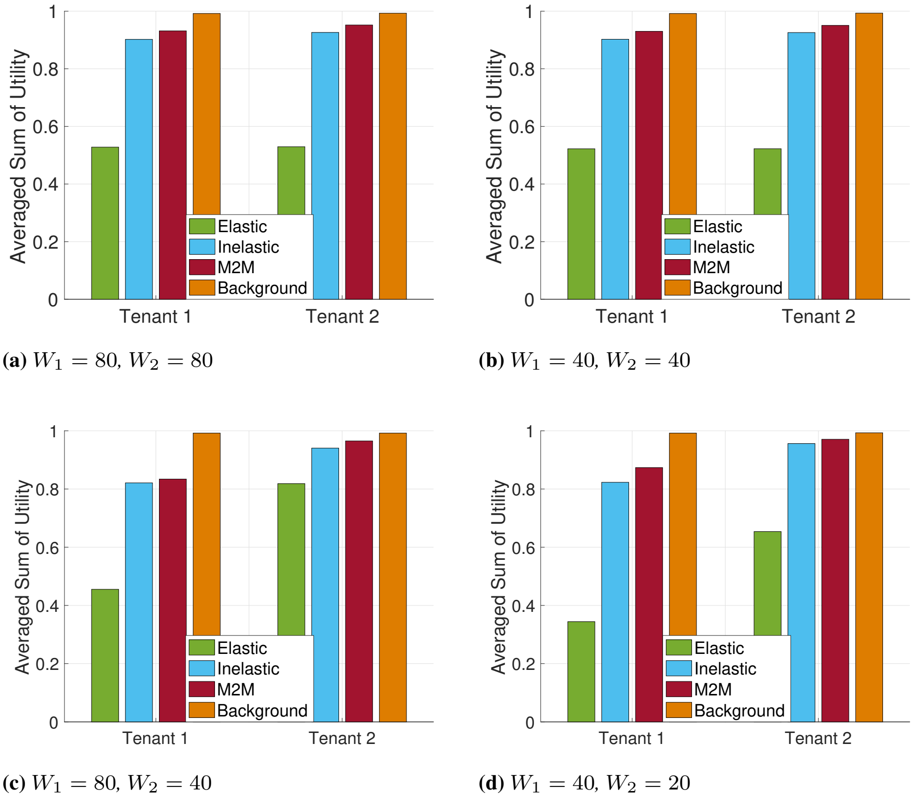
An important question is whether the infrastructure provider's renegotiation interval selection would have any impact on the tenants' resource allocations or not. Fig. 4.8 depicts how the average utilities per tenant is affected by changing the ratio of W_m/RI. In the simulations the renegotiation interval is selected to be $RI = 80$ TTIs. Fig. 4.8a and Fig. 4.8b show the case where both tenants choose the same time window which is different than the renegotiation interval, $W_1 = W_2 \neq RI$. For this case, regardless of their collective time window selection, the achieved utility of each tenant is the same as the other. Moreover, Fig. 4.8c and Fig. 4.8d extend this analysis into the scenario where the tenants choose asymmetric time windows. In this asymmetric selection, the achieved utilities of critical systems are not affected at all, while the elastic service utility of the tenant with the smaller time window is approximately two times higher than the other tenant. The decrease in total achieved utility as the tenants go from a larger time window (i.e. Fig. 4.8c) to a smaller time window (i.e. Fig. 4.8d) is a direct result of decreased spectral efficiency caused by the strict delay constraint.

Consequently, Fig. 4.8 underlines the fact that the renegotiation window selection does not have a direct impact on the resource allocations or the achieved utilities of the tenants. However, based on the results outlined in Section 4.4.2, the smaller renegotiation windows are better from a computational point of view. Therefore, the optimum strategy in RI selection is choosing the least common multiple (LCM) of the time windows of the tenants, i.e. $RI =$ LCM$(W_m)\ \forall m \in M$.

4.4.5 Impact of time window

The impacts of time window on the resource allocation among tenants and the average achieved utility is investigated for $|M| = 2$ scenario, in order to measure the impact of one tenant on the other one, isolated from all other inter-tenant dynamics. Fig. 4.9 and Fig. 4.10 explore the time window differentiation effects on resource sharing under the consideration of resource scarcity and resource surplus, respectively.

In a generic sense, the length of time window determines with which frequency the tenant's expectations are required to be met. Therefore, smaller time windows set more strict delay constraints with respect to the longer ones. Thus, in the scenario where the base station is shared among multiple tenants with different time window lengths, the infrastructure provider has to prioritize the ones with smaller time windows with respect to the rest. The impacts of window length differentiation on resource distribution are



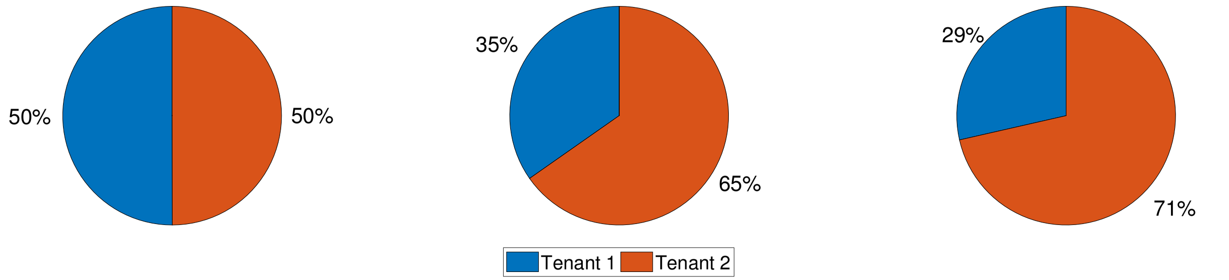
(a) $W_1 = 80$, $W_2 = 80$

(b) $W_1 = 40$, $W_2 = 40$

(c) $W_1 = 80$, $W_2 = 40$

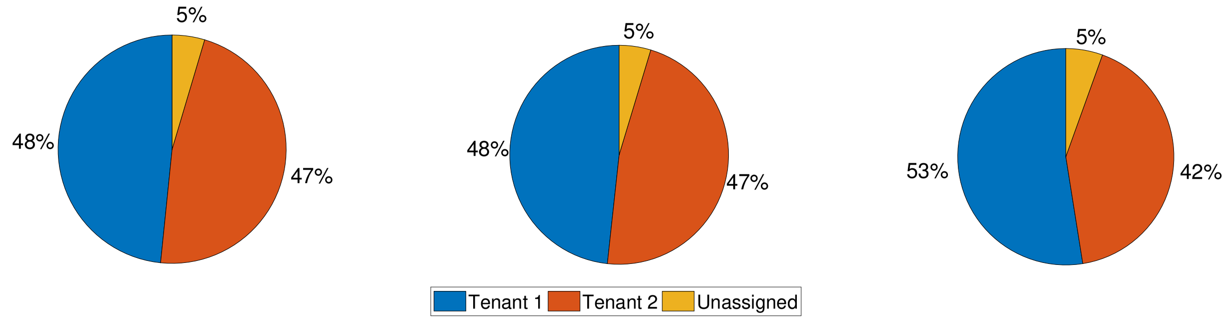
(d) $W_1 = 40$, $W_2 = 20$

Figure 4.8: *The average sum of utility per tenant for different time windows where* $RI = 80$



(a) $W_1 = 80$, $W_2 = 80$ **(b)** $W_1 = 80$, $W_2 = 40$ **(c)** $W_1 = 80$, $W_2 = 20$

Figure 4.9: *Effects of window differentiation on average resource distribution per tenant in the resource scarcity scenario.*



(a) $W_1 = 80$, $W_2 = 80$ **(b)** $W_1 = 80$, $W_2 = 40$ **(c)** $W_1 = 80$, $W_2 = 20$

Figure 4.10: *Effects of window differentiation on average resource distribution per tenant in the resource surplus scenario.*

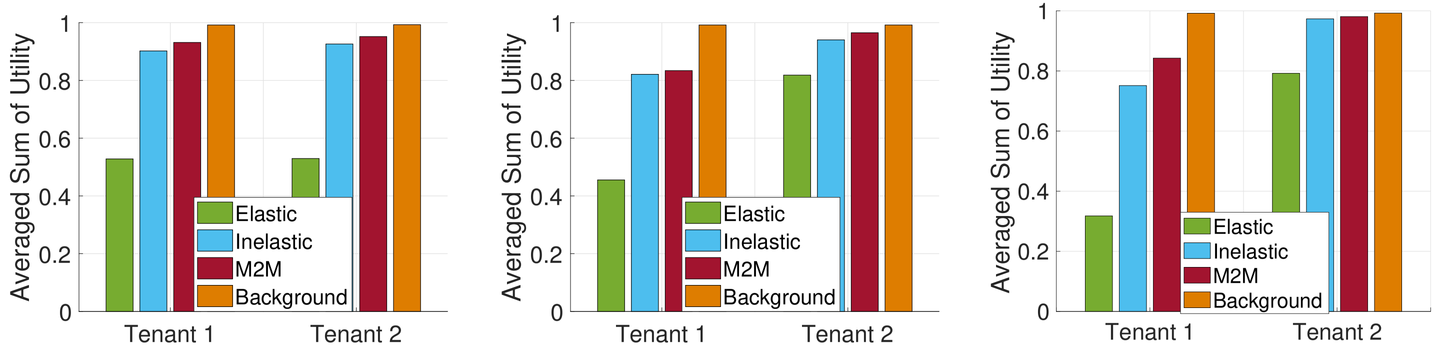
presented in Fig. 4.9 and Fig. 4.10. In case the system posses sufficient resources to satisfy all the tenants, the impact of window differentiation is not visible (cf. Fig. 4.10). On the other hand, in case the resources are not sufficient to satisfy both tenants (i.e. *resource scarcity*) the tenant with a smaller time window, i.e. Tenant 2 in Fig. 4.9, receives higher amount of resources due to the prioritization based on smaller time window selection. As it is possible to exploit this indirect prioritization mechanism, the infrastructure provider or a regulatory body need to monitor the selection of W_m among tenants.

A direct result of prioritization by window length differentiation can be observed in Fig. 4.11, where we compare the average achieved utilities under different window length selections. Due to this priority, the tenant with smaller time window achieves a higher utility with respect to the tenant with longer time window. A key aspect is the fact that the asymmetric prioritization increases as the difference between the selected time windows grows. The major group of services that are afflicted by this asymmetry is the elastic services due to their low priority. All the non-elastic services, due to their inherent priorities, are marginally affected by the prioritization of the competitor tenant. The tenant with the smaller time window achieves higher average utility for all the services, including and particularly in elastic services. However, this way of artificially increasing the priority of services can result in spectral inefficiencies as among the equivalent services from different tenants, the selection is now not only dependent on the achievable rate but also the delay constraint of the tenant.

The economic analysis of window differentiation is given in Fig. 4.12 and Fig. 4.13 where the former one shows the total costs that the tenants pay whereas the latter one presents the actual unit cost per resource that the tenants pay. As the tenant with smaller time window, i.e. tenant 2, obtains more resources the total cost of the tenant also increases. Note that as can be seen in Fig. 4.13, the increase in the total cost is a sole result of higher resource usage as the actual price per unit resource is the same for all three cases.

4.4.6 Impacts of $U_{\mathbf{th},m}$

The service differentiation among tenants can occur in different layers. Unlike the service differentiation through W_m, which is investigated in the previous section, in this section another method, namely service differentiation through buying premium services is investigated. A key parameter that the tenants declare at the beginning of the framework is their utility

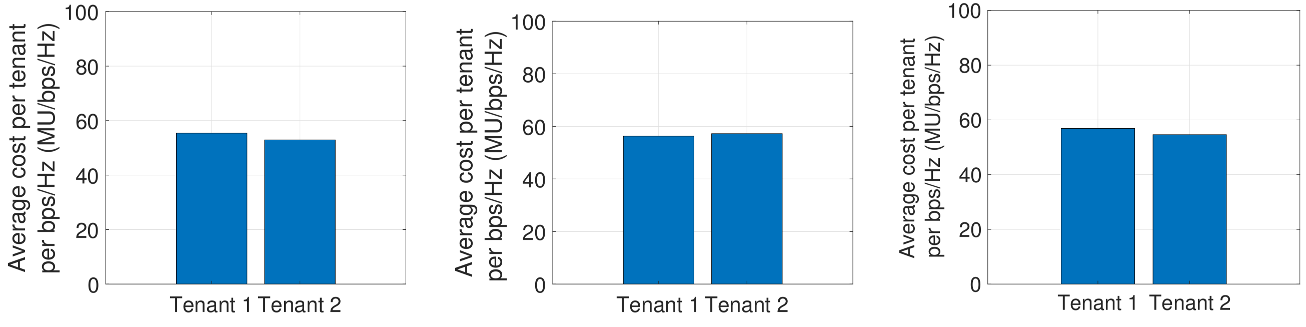


(a) $W_1 = 80$, $W_2 = 80$

(b) $W_1 = 80$, $W_2 = 40$

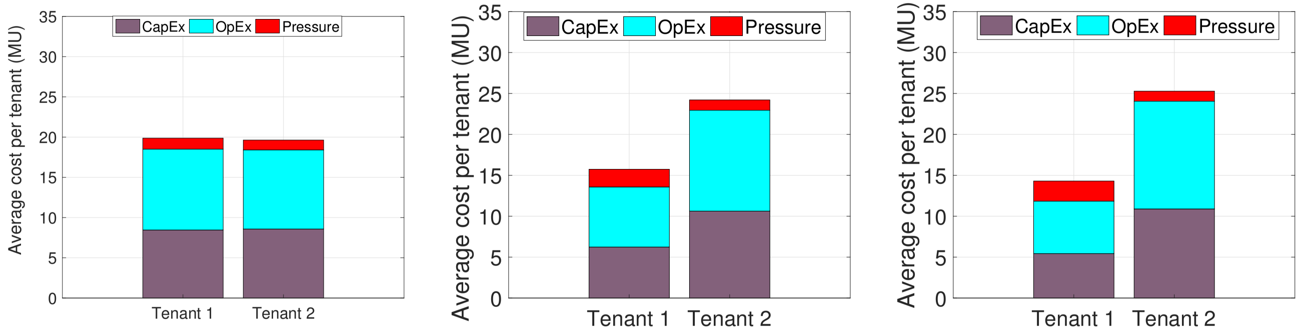
(c) $W_1 = 80$, $W_2 = 20$

Figure 4.11: *Effects of window differentiation on average total cost per tenant in the resource scarcity scenario.*



(a) $W_1 = 80$, $W_2 = 80$ **(b)** $W_1 = 80$, $W_2 = 40$ **(c)** $W_1 = 80$, $W_2 = 20$

Figure 4.12: *Effects of window differentiation on average total cost per tenant in the resource scarcity scenario.*

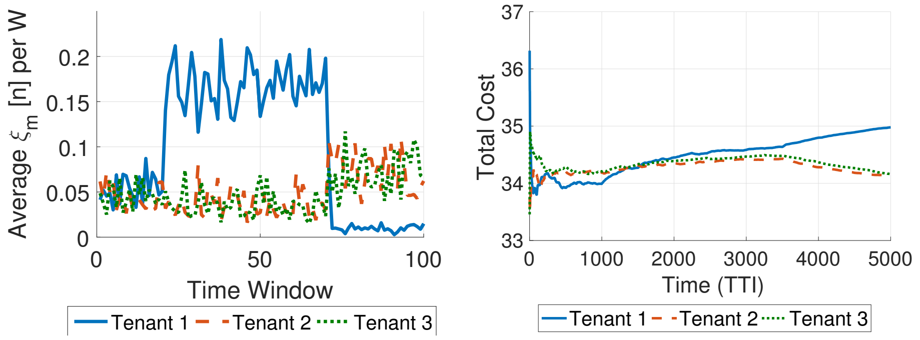


(a) $W_1 = 80$, $W_2 = 80$ **(b)** $W_1 = 80$, $W_2 = 40$ **(c)** $W_1 = 80$, $W_2 = 20$

Figure 4.13: *Effects of window differentiation on average total cost per bps/Hz per tenant in the resource scarcity scenario.*

expectations, $U_{\text{th,m}}$. This expectation (or threshold) is assumed to be same for all the tenants (and similarly same for each service) so far. On the other hand, it can be seen that this parameter can be differentiated among tenants. The proposed gap definition, (4.4.b), shows that higher $U_{\text{th,m}}$ (and indirectly U_3) results in higher priorities. In our scenario, we assume that the tenants can make agreements, i.e. *'premium agreements'*, with the infrastructure provider in order to purchase higher utility thresholds, $U_{\text{th,m}}$, that would reflect as higher priorities to this particular tenant. Note that a key aspect in this utilization is the fact that the tenants gap cannot be lower than zero, so the average achieved utility of tenant is limited by $U_{\text{th,m}}$, such that $\sum_{k \in K} U_{3,k} = U_{\text{th},m}$. Therefore, the tenants can choose to change the maximum achievable utility for particular services ($U_{3,k}$), which would result in much higher priority for these services. In this section we investigate the case where the tenants choose to distribute this premium access right equivalently to all users (K_m).

The impacts of premium service agreements between tenants and the infrastructure provider in order to compensate bad channel conditions is given in Fig. 4.14. More specifically in this scenario, the tenants start the sharing process with equivalent parameters (e.g. $U_{\text{th},m} = 1 \; \forall m \in M$) and similar channel conditions. On the other hand, when the steady state conditions are achieved (i.e. at the 20^{th} time window), the channel conditions of the users from first tenant worsen. Consequently, when the new steady-state is reached, the first tenant faces a higher gap in comparison to the other two tenants, due to the worsened channel conditions. As a response to this new steady-state, at the 75^{th} time window, the first tenant makes an agreement with the infrastructure provider and increases its utility threshold to $U_{\text{th,m}} = 1.2$. This increased threshold is equivalently distributed among all the services, providing a higher priority for these services in comparison to the similar services of other tenants. Thus as can be seen in Fig. 4.14a, in the new steady state, the first tenant decreases most of its gap, while the other two tenants face relatively higher gaps than before. On the other hand Fig. 4.14b depicts the economic impact of this new steady state. One can easily see that in this final steady state the total cost of the first tenant increases enormously in contrast to the other two tenants. This indicates the fact that the tenants can obtain higher utilities than their peers, if they are ready to pay higher costs.



(a) *Average gap of tenants over RI*

(b) *Tenants' total cost per TTI*

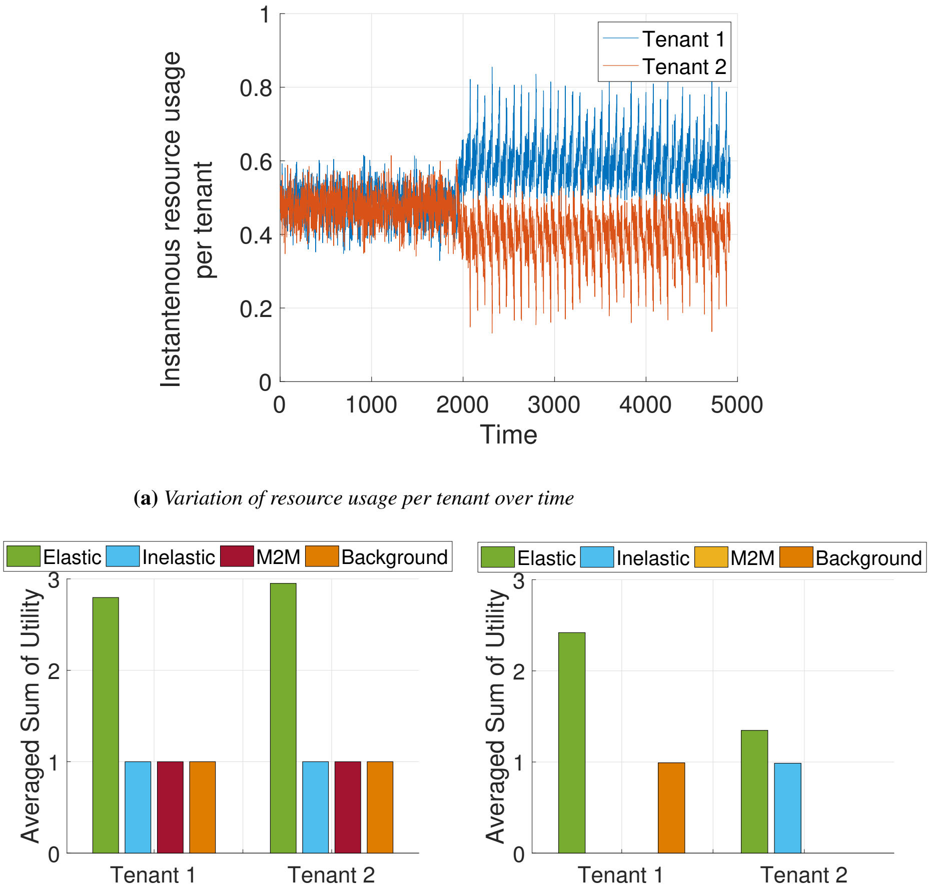
Figure 4.14: *Framework's adaptability to the changes in the channel condition.*

4.4.7 Adaptations to the changes in the traffic mix

Fig. 4.15a investigates the proposed model's capability to adapt to the variations in the traffic mix. For this scenario, we consider that the available resources are sufficient to fully satisfy all the demands, therefore, the services do not need to to compete for the resources. Moreover in the presented scenario, $|K| = 8$ users are equivalently distributed among $|M| = 2$ tenants and the services (i.e. 1 user per service) until $n = 1980$ TTI. At $n = 1980$ the traffic mix of the tenants are changed as follows; the first tenant contains 2 users with elastic traffic and 2 users with background traffic while the second tenant serves 3 users with inelastic traffic and 1 user with elastic traffic. In Fig. 4.15a we can observe the gradual change of the resource allocation with the changing traffic mix. Note that at the time of the change (at $n = 1920$ TTI) the tenants still have the sharing parameters that are calculated for the previous traffic mix. Therefore, the steady state of the resource distribution (and the sharing parameters) is reached after approximately one renegotiation interval later. Fig. 4.15b and Fig. 4.15c show that the average sum of utilities before and after the change in the traffic mix. Note that the decrease in the utility of the elastic users after the traffic mix change is due to the increased number of users with elastic traffic.

4.4.8 Service specialized tenants

In this subsection, we investigate the coexistence of specialized tenants (i.e. they only serve a particular service type) with the conventional tenants that serve a mix of all the services. This analysis focuses on the question if our proposed framework motivates the tenants to be specialized in particular



(a) *Variation of resource usage per tenant over time*

(b) *Average sum of utility before the change of traffic mix*

(c) *Average sum of utility after the change of traffic mix*

Figure 4.15: *Adaptation to the variations in the traffic mix.*

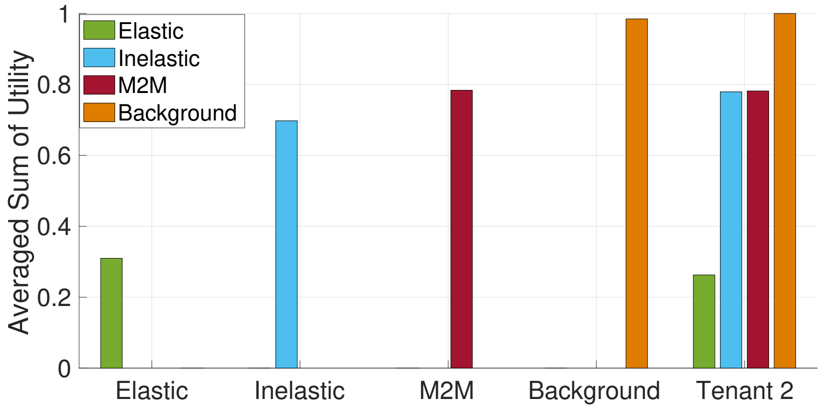


Figure 4.16: *The average utility per service per tenant.*

services or if it is neutral to this choice. In this analysis, we considered $|K| = 16$ users are equivalently shared among two tenants where each service per tenant has 2 users. In this scenario the first tenant prefers to enter to the coalition as four specialized tenants whereas the second tenant demands service as a conventional tenant with all the services.

Fig. 4.16 shows that regardless of how the tenants enter into the market, the proposed model treats the services fairly, giving no advantageous to the specialized tenants. Fig. 4.17 investigates the distiribution of the total cost among the tenants for this particular scenario. Similar to the case with the average utilities, the total costs are equally distributed among the tenants, indicating that the proposed pricing mechanism would not be affected by the specialization of the tenants. Lastly, Fig. 4.18 shows the resource distribution among the tenants where the tenants get equal split of the resources, i.e. half of the resources. Note that Fig. 4.16 and Fig. 4.18 also show that the inter-service resource distribution of Tenant 2 is the same as Tenant 1.

Consequently, one can observe that the proposed framework as well as the pricing mechanism are neutral to the service specialization among tenants. Even in the most extreme scenario, the service level fairness is still preserved and the pricing mechanism is automatically adjusted to maintain fairness among the same services.

4.4.9 Cost and utility in different sharing scenarios

In this section, we have explored the relation between the number of tenants and the sustainability of the sharing platform. More specifically, changes in the total average cost and the total utility with an increasing number of

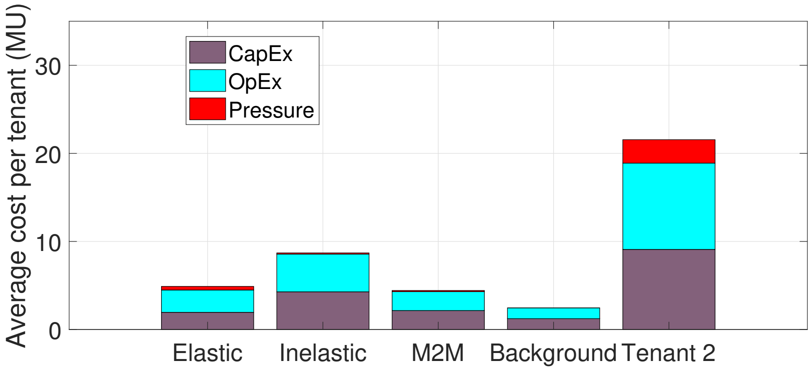


Figure 4.17: *The average total cost per service per tenant.*

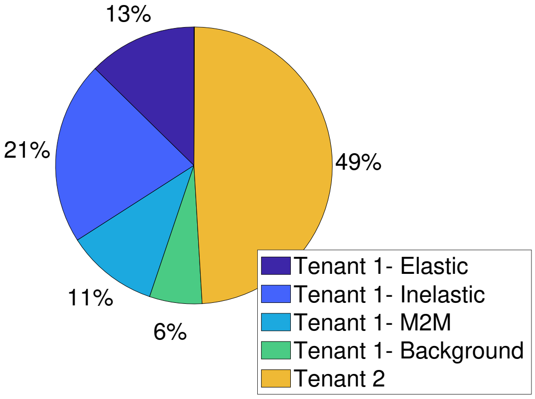
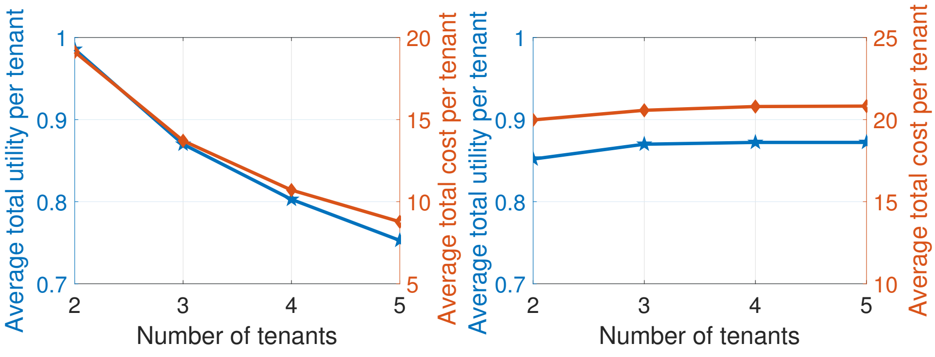


Figure 4.18: *Effects of service specialization on resource distribution.*

tenants are observed within two time scales, i.e. *short term* and *long term*. Short term analysis considers a time interval that is not sufficient for the infrastructure provider to respond to the changes in the traffic demand with an increase in the capacity. Thus it can be considered as the transient state. On the other hand, the infrastructure provider can react to the increased demand with a higher network capacity in the long term, therefore, the long term is the steady state of the network resources.

Fig. 4.19a presents the changes of the total average cost and average utility for different $|M|$ for the short term, whereas, Fig. 4.19b considers long term impacts of the increasing number of tenants. To observe the long term effects, the network capacity is proportionally increased to the traffic demand by increasing the total bandwidth. Our analysis showed that, the increase in $|M|$ can lead to a resource scarcity in short term and can result



(a) *Short-term effects* **(b)** *Long-term effects*

Figure 4.19: *Effects of increasing number of tenants on the average utility (specified with '$\star$') and average costs (specified with '$\diamondsuit$') per tenant.*

in decreased utilities per tenant. On the other hand, as depict in Fig. 4.19a, the average total cost per tenant decreases since the costs are shared among a higher number of tenants. Fig. 4.19b shows that when the traffic demand increase is met with a proportional increase in the capacity, both the utility and the cost per tenant remain the same regardless of the number of tenants. Note that the non-decreasing behavior of the cost function is due to the increase in the unit costs for the excess capacity in the network.

Consequently, it is observable that in the long term, assuming that the infrastructure provider can scale the capacity correspondingly, the increase in the total number of tenants, $|M|$, does not have any implication in terms of the assigned resources per tenant or the unit cost of resources. However, in the short term, since the infrastructure provider cannot scale the capacity, the decrease in the total cost is followed by a decrease in the total achieved utility per tenant. However, Fig. 4.19a is not sufficient to assess whether the decrease in the total cost can compensate the utility decrease or not. In order to have an estimation over the likelihood that a tenant can tolerate the descending utility with the lower cost, we have incorporated the acceptance probability concept from [10]. More specifically, in this work the likelihood of a service to be accepted by user k for a given utility U_k with a price of p is modeled as:

$$A_k(p, U_k) = 1 - exp(-Cp^{-\epsilon}U_k^{\mu}), \tag{4.5}$$

where C is a constant and μ and ϵ are microeconomic parameters. The values of these parameters are directly taken from [10] and constant profits are assumed for the tenants regardless of $|M|$, which means that the variations in the total cost are directly affecting the prices. For the sharing model, it is clear that in order to be profitable in short term, the acceptance probability must be a non-decreasing function of $|M|$. Therefore, assuming $|M_1| \leq |M_2|$, the condition below must be met:

$$A_{k,M_1}(p_{M_1}, U_{k,M_1}) \leq A_{k,M_2}(p_{M_2}, U_{k,M_2}), \tag{4.6}$$

where

$$A_{k,M_1}(p_{M_1}, U_{k,M_1}) = 1 - exp(-Cp_{M_1}^{-\epsilon} U_{k,M_1}^{\mu}),$$
$$A_{k,M_2}(p_{M_2}, U_{k,M_2}) = 1 - exp(-Cp_{M_2}^{-\epsilon} U_{k,M_2}^{\mu}),$$

Assuming that the microeconomic coefficients are independent of $|M|$, (4.6) can be reformulate as follows

$$\left(\frac{U_{k,M_1}}{U_{k,M_2}}\right)^{\mu} \leq \left(\frac{p_{M_1}}{p_{M_2}}\right)^{\epsilon}. \tag{4.7}$$

Consequently, satisfying (4.7) implies that the decrease in the total utility can be compensated by the decrease in the service price, therefore, the tenants would persist in the sharing agreement. The analysis of the scenario in Fig. 4.19a is given in Table 4.2. As can be observed, (4.7) is always satisfied in Table 4.2, indicating that the tenants would always accept the decrease in the utility for the given change in the price. Consequently, this analysis shows that even if the infrastructure provider cannot compensate the increasing demand in short time scales, the proposed model can still provide a sustainable business platform.

Table 4.2: *Variation of average utility and total cost per tenant with respect to the number of tenants in short term*

$\|M_1\| \rightarrow \|M_2\|$	$\left(\frac{U_{k,M_1}}{U_{k,M_2}}\right)^{\mu}$	$\left(\frac{p_{M_1}}{p_{M_2}}\right)^{\epsilon}$
$2 \rightarrow 3$	1,2834	3,7822
$3 \rightarrow 4$	1,1744	2,6893
$4 \rightarrow 5$	1,1372	2,2142

Table 4.3 presents a more detailed analysis of the acceptance probability, namely evaluation of (4.7) per slice type. In this analysis, one can observe that as the resource scarcity further increases (i.e. $|M = 4|$ to $|M = 5|$ for this scenario), the probability of the elastic users' acceptance is decreasing whereas it increases for the rest of the services. Despite being a direct result of inter-service prioritization, this result shows that as the resource scarcity increases some of the elastic service demand may be lost. This risk

can be minimized by the accurate and timely capacity expansion. As it is detailed in the next chapter, the pressure cost concept in our model handles the accuracy of the expansion while collecting the necessary revenue for a timely capacity expansion.

Table 4.3: *Evaluation of the users' acceptance probability for all slice types (We use 'Yes' to indicate that Eq.* (4.7) *holds, 'No' otherwise)*

$\lvert M_1 \rvert \rightarrow \lvert M_2 \rvert$	Elastic	Inelastic	M2M	Background
$2 \rightarrow 3$	Yes	Yes	Yes	Yes
$3 \rightarrow 4$	Yes	Yes	Yes	Yes
$4 \rightarrow 5$	No	Yes	Yes	Yes

4.5 Summary

In this chapter we have extended the dynamic network slicing concept in order to include various service types with conflicting QoS expectations and priorities and showed that dynamic network slicing offers an efficient way to share resources among tenants. As a key advantage, the proposed framework encourages innovation as it decreases time-to-market and guarantees instantaneous availability of the resources. In this new platform the negotiations are influenced by the traffic mix of each tenant as well as the long term business strategies. Our analysis showed that the proposed model provides fairness among tenants while handling the inter service priorities. It is important to note that, although the tenants share a common infrastructure, they can still differentiate their services through customizing a set of parameters. Finally, in this chapter, through numerical analysis we have shown that in order for sharing to be effective for all the services, the network capacity has to be scaled in line with the traffic increase which underlines the importance of having a sound pricing model.

CHAPTER 5

Anticipatory network management

5.1 Introduction

Unlike all the previous technological transitions, the evolution towards 5G technology is accompanied with an excessive amount of data on the user behavior and traffic condition. In parallel, the rapidly improving machine learning techniques have given the researchers the tools to process and build efficient models based on this data. Using these models, the user behaviors and their implications on the physical environment can be better understood, such that the researcher can predict the upcoming changes with a high accuracy level. Similar to all research fields, in dynamic network slicing and resource trading, the anticipation of the upcoming changes can increase the efficiency. In the previous chapters, we focus on reactive network sharing and trading, namely the tenants do not have any information regarding the upcoming traffic demand and the service mix, but instead they update their sharing parameters based on the previous observations. In this chapter, we introduce an anticipatory network slicing and resource trading framework and discuss how the predicted information can be exploited. In particular, we are focusing on the integration of the predicted information regarding the achievable rates of users in the upcoming time slots, and how the impacts of inaccurate predictions can be eliminated or minimized, especially during the resource negotiations. Note that the design of a new prediction algorithm is out of scope of this chapter. Since the proposed framework is envisioned to be run in very short time scales (i.e. in the order of seconds or milliseconds), we assumed that the total user demand, i.e. traffic mixture, does not change in volume or in mixture.

Following our short term analysis, we focus on how the infrastructure provider scales the network resources in long term. In this part, revisiting our previous assumption, we have assumed that the user demand shows large fluctuations over time. From a long term perspective, the critical aspect is to guarantee an accurate and efficient investment of the collected capital resources on the network resources. However, in a scenario with multiple tenants and multiple services, the 'efficiency' of any investment decision can vary from case to case. In this chapter, built on three key metrics, i.e. *1)* measured QoE degradation, *2)* the available revenue and *3)* urgency of expansion, we propose a slice-aware capacity expansion strategy that can provide guidance on the investment decisions in real life scenarios.

The findings demonstrated in this chapter are published in [2], [3] and [6].

5.1.1 Specific research questions

This chapter focuses on the question of "What are the long and short term implications of anticipatory network sharing and resource trading?". This research is complemented with the following specific questions:

- How can the predicted information be integrated to the short time scale negotiations among tenants? (Investigated in Section 5.2)
- What is the value of anticipatory information in a shared multi-tenant network? (Explored in Section 5.2.3)
- How can the aggregated revenue be reinvested efficiently in order to provide a sustainable business platform? (Presented in Section 5.3)

5.1.2 Chapter outline

In Section 5.2, we derive the extension of the model in order to incorporate the anticipatory information, under two parts, namely, proactive resource scheduling and anticipatory resource trading. The proactive approach is detailed in Section 5.2.1 and the respective formulation is presented in Section 5.2.2. Extending the proactive resource scheduling, in Section 5.2.4, an anticipatory resource slicing and trading framework is designed. After the exploration of the short term impacts of anticipatory information, we focus on how to incorporate this information in the long term evolution of the network resources by defining a self-driven capacity expansion model in Section 5.3. The validity of the proposed models are investigated in depth with various simulation scenarios and setups in Section 5.4. Finally, Section 5.5 outlines the key findings of the chapter.

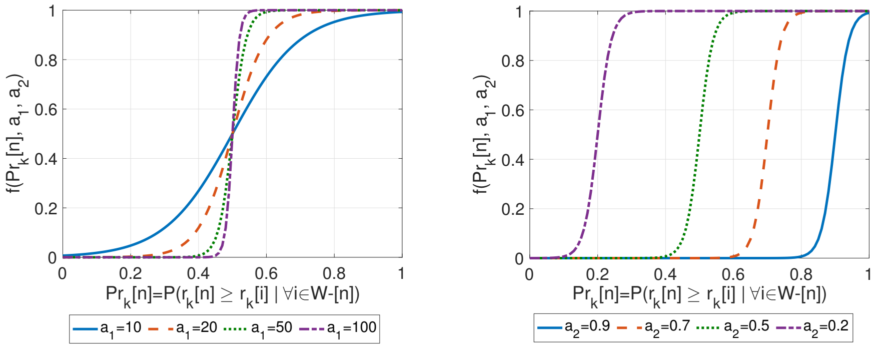
5.2 Anticipatory network slicing

This section investigates the short-term behavior of our model and how it can be improved by exploiting anticipatory information.

5.2.1 Proactive resource scheduling

The real-time scheduling problem myopically focuses on maximizing the achieved utility at a given time slot. On the other hand, this reactive approach cannot exploit the transmission opportunities arising from the instantaneous fluctuations on the achievable rate per user. Consequently, reactive resource scheduling requires a higher amount of resource than what P_2 retrospectively calculates. In order to achieve the full potential of dynamic network slicing, in this part, we evaluate the possible gains of moving towards a proactive approach, by implementing a predictive resource scheduling approach that uses anticipatory information regarding the upcoming time slots. In particular, a channel-aware filter mechanism is integrated to the real time resource scheduler, P_1, in order to evaluate the instantaneous channel rates with respect to the expectations of the upcoming time slots.

As a starting point, we assume that the scheduler possesses an estimation of the channel conditions, namely the probability density function of the achievable rates per user. The probability of the given time slot, n, to be the best time slot to assign resources to the user k is formulated as $Pr_k[n] = P(r_k[n] \geq r_k[i], \forall i \in |W_m|) \in [0,1]$. Entity $Pr_k[n]$ tries to capture the user's time slot with the highest achievable rate in the given time window, W_m. The special case of $Pr_k[n] = 0$ indicates that the instantaneous achievable rate at the current time slot is the lowest within the given time window, thus the scheduler should avoid assigning any resources to this user at that time slot. On the other hand, the case of $Pr_k[n] = 1$ shows that the current time slot n is the best time slot to assign resources to user k as the channel condition is above average conditions.



(a) *Varying a_1, when $a_2 = 0.5$*

(b) *Varying a_2, when $a_1 = 50$*

Figure 5.1: *Variation of the sigmoid function for different a_1 (left) and a_2 (right) values.*

The integration of this channel awareness into our model has to consider inter-user dynamics since purely relying on $Pr_k[n]$ can be misleading. More specifically, although higher $Pr_k[n]$ values demonstrate that n is the best time slot for user k to receive resources, $Pr_k[n] = 1$ does not guarantee that k is the most convenient user to assign the resources at n. Thus, we have integrated this channel information into our model using a two step filter. In the first step of our filter, the statistical values are translated into resources based on a sigmoid function, formulated as

$$f(Pr_k[n], a_1, a_2) = \frac{1}{e^{-a_1(Pr_k[n]-a_2)}}. \tag{5.1}$$

The changes in the characteristic of the sigmoid function for different a_1 and a_2 parameters are shown in Fig. 5.1. As demonstrated in Fig. 5.1a, the linear region of the sigmoid function can be customized using a_1. The length of this linear region determines both the scaling factor between the channel probability and the resource allocations, and the beginning of saturation (i.e. $f(Pr_k[n], a_1, a_2) = 1$) and compression (i.e. $f(Pr_k[n], a_1, a_2) = 0$) regions. The resource efficiency is strongly tied to the value of a_1, as big values can lead to underutilized resources while very small values can end up with low spectral efficiency due to resource allocations in bad channel conditions. The second control parameter, a_2, shifts the sigmoid function (c.f. Fig. 5.1b). Similar to a_1, very high values of a_2 can result in unassigned resources even when the total gap is not zero, while very small values would remove the proactive aspect from the model.

The sigmoid function translates the probability value into resource allocation. However, it is calculated per user, therefore cannot shed sufficient

$$\min_{x_k[n], S_m, \Delta_m} \quad \sum_{m \in M} \xi_m[n] \tag{5.3.a}$$

$$\text{s.t. } U_{\text{th},m} - \sum_{k \in K_m} \beta_k[n] U_k(R_k[n]) \leq \xi_m, \; \forall m \in M, \tag{5.3.b}$$

$$\epsilon_m[n] = \left(\frac{1}{(a_m+1)} \sum_{i=n-a_m}^{n} \sum_{k \in K_m} x_k[i] \right) - S_m, \; \forall m \in M, \tag{5.3.c}$$

$$|\epsilon_m[n]| \leq \Delta_m, \quad \forall m \in M, \quad \forall n \in N, \tag{5.3.d}$$

$$\sum_{i=n-a_m}^{n} \left(S_m(C_{\text{ca}} + C_{\text{op}}) + \epsilon_m[i] C_{\text{op}} + f_{\text{pre}}(C_{\text{pre},m}, \, \xi_m) \right) \leq B_m(a_m+1), \forall m \in M, \tag{5.3.e}$$

$$0 \leq \Delta_m \leq \frac{1}{a_m+1} \sum_{i=n-a_m}^{n} \sum_{k \in K_{m,elastic}} x_k[i], \quad \forall m \in M, \tag{5.3.f}$$

$$\sum_{k \in K} x_k[n] \leq 1, \; x_k[n] \geq 0, \; \forall k \in K, \tag{5.3.g}$$

$$\sum_{m \in M} S_m \leq 1 \; , \; S_m \geq 0, \; \forall m \in M, \tag{5.3.h}$$

light to the inter-user dynamics. With the objective of covering the inter-user dynamics, the output of the given sigmoid function is fed to the second step of the proposed filter which receives $f(Pr_k[n], a_1, a_2)$ and outputs $f(Pr_k[n], a_1, a_2)^p$ where p is a scalar. If the variation among the achievable rate values of users are negligibly small, p can be set to 1, since the output of the sigmoid function is sufficient. On the other hand, as the variation between users' achievable rates increases, p value should be increased as well.

Finally the output of this two-step filter function, *priority coefficient*, is integrated into the gap definition in our framework, i.e.

$$U_{\text{th},m} - \sum_{k \in K_m} \beta_k[n] U_k(R_k[n]) \leq \xi_m, \forall m \in M \tag{5.2}$$

where $\beta_k[n]$ represents the priority coefficient of user k at time slot n.

5.2.2 Mathematical formulation of the model

For the sake of readability, the mathematical formulation of the scheduling problem is presented again in (5.3.a)-(5.3.h). As detailed in Chapter 4.5, the

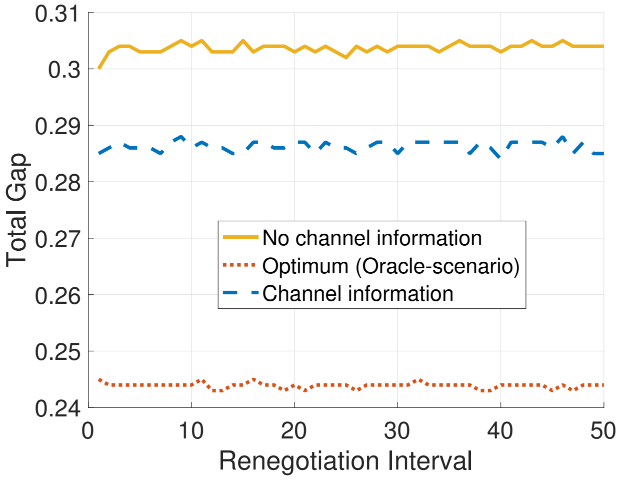


Figure 5.2: *Effects of anticipatory resource scheduling on total gap for* $|M| = 2, \ |K| = 8$.

continuous objective function (5.3.a) minimizes the total gap which is defined in (5.3.b). Constraint (5.3.d) sets the upper and the lower limits for the instantaneous deviation from the guaranteed sharing ratio which is defined in (5.3.c). The economical implications of the technical sharing process is modeled in (5.3.e) while (5.3.f) reflects the tenants' desire to only trade the resources that are intended for the elastic services. Inequalities (5.3.g) and (5.3.h) reflects the physical constraints, i.e. the total assigned resources cannot be greater than the available resources and the infrastructure provider cannot sell resources that he does not possess.

A more detailed examination of the proposed model is given in Section 4.3.

5.2.3 Effects of anticipatory resource scheduling

In this proposed framework, tenants only utilize the statistical observations regarding their traffic conditions (i.e. traffic mix and the achievable rates of the users) without an explicit prediction per time window. We have compared the gains through two-step filter using the probability density functions that are estimated using the Oracle scenario, where the traffic conditions are known and the simulation can be run for the whole RI at once. Fig. 5.2 depicts this comparison between the total gap for three different cases, i.e. with no channel information, using the channel information and the oracle scenario. In the analysis the control parameters are set as follows,

Table 5.1: *Channel information's improvement on the total gap with respect to no-channel information case*

$\lvert K\rvert$	Improvement of total ξ_m
8	33.2%
16	38.5%
24	38.6%

$a_1 = 10,\ a_2 = 0.5,\ p = 3.$

One can observe from Fig. 5.2 that even this simplest form of anticipatory information can create relatively high gains in terms of total gap. Note that the reported gain is achieved through using anticipation only in P_1. However, the negotiations are performed by using the reactive approach, indicating that the performance can be further improved. Moreover, Table 5.1 presents how the achieved gain through anticipation is affected by the increasing number of users. Increasing $|K|$ gives the scheduler a higher flexibility in determining the good channel condition to assign resources, while for lower number of users, the scheduler requires a better accuracy in the prediction process. Further increase in the number of users, i.e. $|K| = 24$ in Table 5.1, cannot increase the efficiency of the model as the possible improvement (that can only be achieved through the flexibility gained by higher number of user) saturates. After this saturation point, further improvements can only be achieved by using a prediction method with an higher accuracy.

5.2.4 Anticipatory resource trading

Following the findings in Section 5.2.1, we move to the exploitation of the predicted achievable rates of the users in the upcoming time slots during the resource negotiations. One way to use this anticipatory information is using the predicted achievable rates in the proposed resource negotiation platform, in (5.3.a) - (5.3.h), and use the predicted assigned resources x_k^{pre} as the actual resource allocations, i.e. $x_k[n],\ n \in RI$. Even though this approach would be the simplest solution, it is highly sensitive to the prediction errors which would lead to big problems in the resource configurations. Thus it can only be used in the perfect prediction scenario. As achieving perfect prediction in real-time algorithm is very challenging, we revisited our heuristic approach of splitting the model into two sub-problems, P_1 and P_2, and redefined it in order to act as a way to minimize the impact of prediction errors while exploiting the anticipatory information in negotiations and resource allocations. Note that unlike the previous chapters, we employ the two step optimization framework with the sole purpose of achieving robustness against prediction errors.

Using the anticipated achievable rates, the P_2 problem is solved using (5.3.a) - (5.3.h) and it outputs the predicted achieved gap ξ_m^{pre}, sharing parameters S_m^{pre}, Δ_m^{pre}, and the resource allocation x_k^{pre}. Note that in case of the oracle scenario, these parameters would be reflecting the optimum values. In order to minimize the impact of prediction error, the sharing parameters for the upcoming renegotiation interval are calculated based on the feature scaling idea defined in the previous section, i.e.,

$$S_m^{\text{new}} = (1-\alpha_m)S_m^{\text{pre}} + \alpha_m S_m^{\text{old}}, \tag{5.4}$$

$$\Delta_m^{\text{new}} = (1-\alpha_m)\Delta_m^{\text{pre}} + \alpha_m \Delta^{\text{old}}, \tag{5.5}$$

where α_m is defined as:

$$\alpha_m = \frac{|\xi_m - \xi_m^{\text{pre}}|}{\xi_m + \xi_m^{\text{pre}}}. \tag{5.6}$$

As an important aspect, the proposed model does not evaluate the success of the prediction method, but considers the impact of prediction errors on the framework's performance. In line with this objective, the scaling coefficient is only considering the impacts of prediction errors on the performance of the proposed model, which is measured by the total gap. In the case of perfect prediction, both algorithms will behave the same, meaning that $\xi_m^{\text{pre}} = \xi_m$ and would set the scaling coefficient to zero. This scenario also allow the direct application of the predicted sharing parameters in the upcoming renegotiation interval. When the prediction accuracy decreases and affects the performance of the algorithm (i.e. $\xi_m^{\text{pre}} >> \xi_m$), α_m value increases up to one, $\alpha_m = 1$. For this case, the inaccurate predicted values are ignored, and the proposed framework works in a reactive manner.

Following P_2, the real time resource scheduling problem, i.e. P_1, receives ξ_m^{pre}, x_k^{pre}, S_m^{pre}, r_k^{pre} and Δ_m^{pre}, and determines the resource allocations using (5.3.a), (5.3.b), (5.3.c), (5.3.d), (5.3.e) and (5.3.g). In order to integrate the predicted resource allocations, the assignable resources to user k is upper limited by the predicted resource allocation, i.e.

$$x_k^{\text{pre}}[n] \geq x_k[n]. \tag{5.7}$$

Consequently, the real time scheduler can also make small adjustments on the predicted resource allocations in line with the dynamic needs of the traffic conditions and prediction errors.

5.2.5 Active filtering

The redefined two-step algorithm can cope with the prediction errors, however, it can also act as a limiter on the algorithm's efficiency. More specifically, if the prediction accuracy is low, (5.7) can decrease the resource efficiency by preventing the users from obtaining resources when their actual channel condition is rather high. In this case, the scheduler is forced to assign resources to the users with lower achievable rates. Thus, in this section built upon the proposed simple filter approach in 5.2.1, we propose an extended filter model that can filter out the impacts of prediction errors while exploiting the advantages of high prediction accuracy.

Using the sigmoid function depicted in Fig. 5.1, the proposed filter is defined as,

$$F(x^{\text{pre}}[n], E_k[n]) = x^{\text{pre}}[n] + \frac{E_k[n]}{1 + e^{-a_{1,k}(E_k[n] - a_{2,k})}} \tag{5.8}$$

where $E_k[n]$ represents the error during the prediction of the achievable rate of user k at time slot n, and x_k^{pre} showes the calculated optimum resource ratio for the predicted achievable rate $r_k^{\text{pre}}[n]$. The value of $E_k[n]$ is calculated using the Euclidean distance of the actual and the prediction values of the achievable rates, i.e. $E_k[n] = |r_k^{\text{pre}}[n] - r_k[n]|$. The implications of low prediction accuracy vary depending on the number of users and how the prediction error is distributed among these users. Therefore, we consider both the positive and negative prediction errors, which is formulated using the absolute value in the error definition. Similar to the definition in Section 5.2.1, $a_{1,k}$ and $a_{2,k}$ are used to control the filter mechanism.

Note that unlike Section 5.2.1, here, the filter's sensitivity to the accuracy of the anticipation technique strictly depends on $a_{1,k}$ and $a_{2,k}$. Therefore, in order to capture the most accurate parameters regardless of the evolution of the prediction errors and changing traffic dynamics, we used an auto-scaling mechanism that is given as,

$$a_{1,k} = \mu_{n \in RI}(E_k), \tag{5.9}$$

$$a_{2,k} = \frac{10}{\sigma_{n \in RI}(E_k)}. \tag{5.10}$$

where the average value and the standard deviation are given as μ and σ, respectively. The updates of these parameters are done at every RI, using the prediction values for the given RI. Based on the calculated $F(x^{\text{pre}}[n], E_k[n])$ value, (5.7) is rewritten as follows.

$$F(x^{\text{pre}}[n], E_k[n]) \geq x_k[n]. \tag{5.11}$$

Note that (5.11) can scale the resource allocation between $[x_k^{\text{pre}}[n], 1]$, depending on the prediction accuracy. Thus it can cover both the oracle scenario and the heuristic scenario.

5.3 Slice-aware capacity expansion strategies in multi-tenant networks

5.3.1 Expansion budget and its implications

Unlike the predecessor technologies, during the deployment of 5G infrastructure, the majority of the base stations from previous technologies are going to be preserved. Therefore, the transition towards 5G is envisioned to be performed gradually with the increasing needs of the network. However, this gradual change requires a novel approach in order to assess the actual needs of the regions and perform the new base station deployments accordingly. The conventional expansion strategy that only relies on the QoE degradation cannot assess the conflicting priorities of different geographical areas that rise from the encountered service mixture. More specifically, even if the users from two different regions experience similar QoE degradation, from a tenants' perspective they do not necessarily have the same economic value. Consequently, from a tenant based perspective, increasing the available capacity in some particular regions, even if they do not face the highest QoE degradation, would be more valuable than the rest of the regions. Moreover, one can see that satisfying the business strategies of tenants in a shared network is very challenging due to the conflicting strategies of tenants and the heterogeneity of the service expectations. Consequently, in this section we propose a novel slice-aware capacity expansion strategy that provides an efficient guideline for capacity expansion. In this part, we assume that the total geographical region is divided into a set of regions, R, where a specific region is indicated as r. Each region r is covered by a set of base stations B_r whose cardinality is shown by $|B_r|$. We assume that the base stations cover the geographical area such that there are no coverage gaps.

The proposed model in (5.3.a) - (5.3.h) performs the real time resource allocations and inter-tenant negotiations and outputs the minimum gap per tenant, ξ_m. In order to differentiate the gaps from different base stations, the tenant specific gap representation in the previous chapters is changed to $\xi_{m,b}[n]$. As previously discussed in detail, each tenant contributes to the aggregated revenue for expansion in line with their gap $\xi_{m,b}$ and the pressure cost unit $C_{\text{pre},m}$. Similar to the previous chapters, time is discretized and divided into time slots which are shown by n. For the sake of simplicity, we assume that the infrastructure provider determines the capacity expansion after a period of time, that we call as '*observation period*' and denote as W_{ex}. During the length of W_{ex}, we assume that the infrastructure provider observes the evolution of the traffic demand and aggregates the revenue. The accumulated revenue over W_{ex} is measured by,

$$R_{acc} = \sum_{n \in W_{ex}} \sum_{r \in R} \sum_{b \in B_r} \sum_{m \in M} \xi_{m,b}[n] C_{\text{pre},m}, \tag{5.12}$$

At the end of W_{ex}, the regions where to expand the resources are selected based on the accumulated revenue (R_{acc}), the traffic mix, and the total gap per region, $\xi_r = \sum_{b \in B_r} \sum_{m \in M} \xi_{m,b}$. In order to determine how to reinvest the collected revenue, first thing to calculate is the number of base stations that can be deployed. In order to calculate the deployment cost of a base station, we are considering the model presented in [87], where the total cost of a new base station deployment is split into the equipment cost, infrastructure cost and the capacity cost. The equipment cost measures the cost of buying the electronic equipment of the base station. The infrastructure cost is the cost required to physically deploy the base station, while the capacity cost is the required cost to provide service. For example, infrastructure cost can be considered as digging the area in order to deploy the cables and all the physical needs. On the other hand, sometimes the existing base station does not require any physical improvements but just an increase in their capabilities or even simply by increasing the available spectral resources. The capacity cost represents the cost of this change. In [87], the authors calculate that the total cost of increasing the base station density of a given area by λ_r will be equal to

$$C_{\text{tot}} = \lambda_r C_1 + \lambda_u A_{01} + \frac{\lambda_u}{\lambda_r \pi} B_{01}, \tag{5.13}$$

where C_1 represents the equipment cost and A_{01} and B_{01} indicate the capacity and the infrastructure costs, respectively. Parameter λ_u is defined as the user density, which can be estimated based on long time observations. For the sake of simplicity, we assume that all the regions have unit coverage area, which makes λ_r and λ_u to be equal to the number of base stations and the average number of users. Note that depending on the type of capacity expansion, the cost structure presented in (5.13) varies. In our research, based on the envisioned requirements of the 5G community, we consider short to medium time scale for the capacity expansion, i.e. time intervals between days to months. However, one can notice that deploying the base station in a given region would require longer time, e.g. a year. Thus, without loss of generality, we have considered that the infrastructure provider considers increasing the available spectral resources as the capacity increase. Therefore, A_{01} and B_{01} are considered to be zero, since the infrastructure provider uses the available expansion revenue only to buy additional spectrum resources. Consequently, one can calculate that the maximum number of new base stations that can be deployed is equal to $\lambda^{\text{max}} = \lfloor R_{acc}/C_{\text{tot}} \rfloor$. Moreover, we assume that the capacity expansion in a region uniformly affects all the base stations in the given region r. Consequently, the capacity increase is homogeneously distributed among all the base stations in r.

5.3.2 The impact of expansion per region

The accumulated economical revenue and the capacity expansion strategy determine the maximum capacity expansion (namely, the maximum number of deployable base stations) in a region. Therefore, the problem of capacity expansion turns into determining where to place the additional resources. Consequently, the objective function of our model is to deploy the maximum network capacity (λ_r) in the regions where it would create the maximum impact, i.e. the minimum total gap (ξ_r),

$$\min \sum_{r \in R} \frac{\partial \xi_r}{\partial \lambda_r} \lambda_r. \tag{5.14}$$

Note that gap is a continuous decreasing function of available capacity. Therefore, the derivative in (5.14) is negative. Consequently, $\sum_{r \in R} \lambda_r = 0$ would not be selected unless the total gap is equal to zero. Since the gap depends on the achieved spectral efficiency, (5.14) can be rewritten as

$$\min \sum_{r \in R} \frac{\partial \xi_r}{\partial R_r} \frac{\partial R_r}{\partial \lambda_r} \lambda_r, \tag{5.15}$$

where ∂R_r represents the unit increase in the achieved spectral efficiency in the region. The first term in (5.15), i.e. $\frac{\partial \xi_r}{\partial R_r}$, estimates the impact of unit change in the achieved spectral efficiency on the measured gap in region r. In Section 4.2, we have detailed the design of customized utility functions per service type and how they interact with each other. The aggregated utility function of a region r, $U_r(R_r, n)$, is defined as

$$U_r(R_r, n) = \sum_{b \in B_r} \sum_{k \in K_b} U_k(R_k, n) \tag{5.16}$$

where K_b indicates the set of active users at base station b at the given time n. Since the total gap is equal to the difference between the desired utility and the achieved utility, i.e.

$$\xi_r[n] = \sum_{m \in M} |B_r| U_{\text{th},m} - U_r(R_r, n), \tag{5.17}$$

(5.15) can be rewritten as follows,

$$\min \sum_{r \in R} \frac{\partial(|B_r| U_{\text{th},m} - U_r(R_r, n))}{\partial R_r} \frac{\partial R_r}{\partial \lambda_r} \lambda_r \equiv \max \sum_{r \in R} \frac{\partial U_r(R_r, n)}{\partial R_r} \frac{\partial R_r}{\partial \lambda_r} \lambda_r. \tag{5.18}$$

The latter term in (5.15) represents the change in the spectral efficiency with increasing capacity. However, the actual change in a region's spectral efficiency requires an in depth analysis of the location of the newly deployed base station and it is very hard to predict. On the other hand, in [86], built upon the homogeneous point processes, the authors argue that the spectral efficiency of a region with a given number of base stations can be calculated using,

$$R_r = \frac{\pi^{5/2}}{2} \sqrt{\frac{\lambda_u \lambda_r P}{\sigma^2}} \text{Erfc}\left[\frac{\pi^2 \lambda_u}{4} \sqrt{\frac{P}{\sigma^2}}\right] exp\left[\frac{\pi^4 \lambda_u^2 P}{16\sigma^2}\right], \tag{5.19}$$

where P indicates the transmission power and σ^2 is the noise power. The authors in [86] prove that if $\lambda_u \sqrt{\frac{P}{\sigma^2}} >> \frac{4}{\pi^2}$ then (5.19) can be simplified to

$$R_r = 2\sqrt{\frac{\lambda_r}{\lambda_u}}. \tag{5.20}$$

Therefore, the second term in (5.15) is equal to

$$\frac{\partial R_r}{\partial \lambda_r} \equiv \frac{1}{\sqrt{\lambda_r \lambda_u}}, \tag{5.21}$$

consequently our objective function becomes,

$$\max \sum_{r \in R} \frac{\partial U_r(R_r, n)}{\partial R_r} \frac{\sqrt{\lambda_r}}{\sqrt{\lambda_u}}. \tag{5.22}$$

However, (5.22) cannot be solved using linear optimization. In order to use linear optimization, we have used a linearized version of (5.22), namely,

$$\max \sum_{r \in R} \frac{\partial U_r(R_r, n)}{\partial R_r} \frac{\lambda_r}{\sqrt{\lambda_u}}. \tag{5.23}$$

Consequently, the expansion model can be written as given in (5.24.a)-(5.24.b)

$$\max_{\lambda_r} \sum_{r \in R} \frac{\partial U_r(R_r, n)}{\partial R_r} \frac{\lambda_r}{\sqrt{\lambda_u}}, \tag{5.24.a}$$

$$\text{s.t.} \sum_{r \in R} \lambda_r \leq \lambda^{\max}. \tag{5.24.b}$$

On the other hand, a direct implementation of this model would not consider the impact of newly deployed base stations. Therefore, the allocations of the new base stations is iteratively performed using an updated version of (5.24.b), i.e.,

$$\sum_{r \in R} \lambda_r \leq 1, \tag{5.25}$$

and by using Algorithm 1. Note that due to the iterative application of the proposed algorithm, the linearization in (5.23) does not cause any loss of generality.

Algorithm 1 Proposed self-expansion algorithm

```
for i=1:$\lambda^{\max}$ & $\sum_{r \in R} \frac{\partial U_r(R_r,n)}{\partial R_r} > 0$ do
    Solve (5.24.a)- (5.25)
    Distribute capacity
    Recalculate $\frac{\partial U_r(R_r,n)}{\partial R_r} \frac{1}{\sqrt{\lambda_u}}$
end for
```

5.4 Performance evaluation

5.4.1 Simulation setup

In this section, we present the long and short term analysis of anticipatory networking. More specifically, we first introduce the anticipatory networking characteristics that revel the economical and the technical improvements. Then, based on this model we investigate the slice-aware capacity expansion strategies.

The short term evolution of the network is governed by the real time resource allocations. Therefore, the anticipatory aspect is integrated by prediction of the achievable rates of the users. Consequently, the main objective of the proposed model is to dynamically prepare and modify the resource allocations in order to fit the upcoming conditions of the network.

On the other hand, the framework is designed to be run per TTI, therefore, immediate availability of the achievable rate predictions for the upcoming time windows is crucial. Thus many efficient yet complex prediction algorithms are not usable in our model. While selecting the prediction methods, we consider the time complexity and the prediction accuracy of the algorithms. As an outcome of our research, we have decided to focus on two well-known prediction approaches, namely, Auto-Regressive Integrated Moving Average (ARIMA) and Feed Forward Neural Networks (FFNN). Note that despite the focus on the two particular prediction methods, the findings of the chapter does not depend on the prediction algorithm selection. In particular, the prediction block can be considered as a black box that can be filled with different prediction tools and our focus is mainly on how to use the output information from this black box.

In the short term analysis, we have considered a simulation horizon of $N = 5000$ TTIs where each time slot is considered to be 1 TTI. Similar to previous chapters, the TTIs are assumed to be scaled in line with the needs and the capabilities of the available technology. The base station is considered to be shared by $|M| = 3$ tenants with $|K| = 12$ users that are homogeneously distributed in the coverage area. The tenants are considered to serve same number of users $|K_m| = 4$ and K_m is assumed to be a mixture of all services. The results presented in the short term analysis are averaged over 50 instances. Despite that the simulations are run in Matlab 2017a, the optimization models are solved in Gurobi solver [35].

Through the simulation the users are assumed to be walking with a speed of $v = 1.5$ m/s on a direct line. The SINR per user is calculated using Shannon-Hartley theorem, i.e.

$$\text{SINR}_k[n] = |h_k[n]|^2 P d_k^{-\alpha} / (\sigma^2 + I_0) \tag{5.26}$$

where d_k is the distance between user and the base station, α is the path loss exponent and I_0 is the inter-cell interference. Values of $h_k[n]$ are the Rayleigh coefficients that are generated from the frequency-flat fading channel between the user and the base station. The maximum Doppler spread is modeled based on

$$F_d = v f_c / c, \tag{5.27}$$

where f_c is the carrier frequency, c represents the speed of light and v is the walking sped. The particular values of these parameters are also presented in Table 5.2.

Table 5.2: *Values of the various parameters in the simulations.*

Parameter	Value
v	5.4 km/hr.
f_c	2 GHz
c	300×10^6 m/s
$\lvert M \rvert$	3
$\lvert K \rvert$	12

5.4.2 Comparison between different prediction methods

As aforementioned, the prediction accuracy and the time complexity are the two main factors in our prediction algorithm selection. A comparison between two different approaches regarding the important aspects to our model is presented in Table 5.3. Note that the values presented here are obtained from a commercially available computer equipped with i7-4510U CPU and 16 GB RAM. The FFNN model is built, trained and tested in Matlab2017a, while for the implementation of ARIMA we have used R. In Table 5.3, the total duration of the algorithm is divided into two parts, namely training time and prediction time. In particular, the training time is the duration of finding the optimum parameters to represent the data set (i.e. training of the model), while prediction time is the time required to generate prediction for the upcoming values of the observed process.

Table 5.3: *Comparison between ARIMA and FFNN in terms of accuracy levels and time complexities*

	ARIMA	FFNN
Time complexity for training process (sec)	0.428	75.03
Time complexity for prediction process (sec)	0.722	0.598
Prediction error for $\lvert W_P \rvert = 10$ms (MAPE)	7.61 %	7.14 %
Prediction error for $\lvert W_P \rvert = 50$ms (MAPE)	160.8 %	216.8 %
Adaptability to varying time conditions	Yes	No

The prediction performance of the model is measured both in terms of the mean average percentage error (MAPE) and the mean square error (MSE), that are formulated as

$$\text{MAPE}(\%) = \frac{100}{N \times |K|} \sum_{k \in K} \sum_{n \in N} \frac{|r_k^{\text{pre}}[n] - r_k[n]|}{r_k[n]}, \tag{5.28}$$

$$\text{MSE} = \frac{1}{N \times |K|} \sum_{k \in K} \sum_{n \in N} (r_k^{\text{pre}}[n] - r_k[n])^2. \tag{5.29}$$

Comparing their computational times, we observe that both of the models' prediction process require less than 1 second. On the other hand, ARIMA outperforms the FFNN in terms of the training duration. Moreover, FFNN requires retraining in order to adapt to the evolving network conditions. In terms of the prediction accuracy, we observe that despite the comparable performances of both methods, ARIMA performs relatively better for larger prediction horizons W_P. Although, there is no major advantage of using one method over the other, due to the time complexity and the adaptability skills of ARIMA, we have decided to proceed with the ARIMA model. Note that it is possible to obtain higher performance with more complicated ANN methods (e.g. RNN or LSTM) but these models bring higher time complexity, therefore, they are not applicable for our problem.

Table 5.4: *Prediction accuracy for different values of W_P and W_L*

Scenario (W_P, W_L)	MAPE (%)	MSE
(10,10)	7.61	0.101
(10,50)	7.34	2.14
(10,90)	12.81	0.69
(25,25)	76.64	1.10
(25,50)	29.86	0.84
(25,75)	28.30	0.77
(50,50)	165.9	3.70

The impacts of having different learning window and prediction window (W_P and W_L) on prediction accuracy are given in Table 5.4. Since the used achievable rates have the correlation window of 100 TTIs, the analysis is limited with the 100 TTIs, namely $W_P, W_L \leq 100$ TTIs. We observe that the selection of W_P and W_L values whose summation is equal to the correlation window (i.e. $W_P + W_L = 100$ TTIs) produces the best prediction accuracy. Moreover, we have also observed that the selection of very high W_L with very small W_P can result in an over-fitting problem which increases the prediction errors. On a different note, our analysis reveals that for some instances the MAPE and MSE values follow an inverse proportional behavior. More specifically, as the MAPE value increases, we can see the decrease in MSE value. In order to better understand the meaning of this behavior, we need to revisit (5.28) and (5.29) . The squared prediction errors in the MSE definition amplify the larger prediction errors (especially larger than 1) and suppress the smaller ones. On the other hand, MAPE gives equal weight to every prediction error. Thus, both MAPE and MSE values are required to have an understanding on the error. For instance comparing the two cases in Table 5.4, (10,50) and (10,90), we can see that (10,90) has a much smaller MSE while its MAPE is approximately two times higher than (10,50). This shows that the prediction errors in (10,90) are homogeneously distributed over the simulation horizon. Moreover, (10,10) and (10,50) have similar MAPE and very different MSE, which indicates that the error is uniformly distributed for the former case while it is concentrated in particular time slots for the later case. Our simulations have showed that between this two scenarios, $|W_L| = 50$ outperforms the other case by a margin of 3%.

5.4.3 Robustness to the prediction errors

Fig. 5.3 depicts the variation of the average total utility for $|M| = 2$ case for different scenarios, i.e. no prediction, prediction without filter (i.e.'no filter') and prediction with filter. When the prediction is not used, the reactive model presented in the previous section is implemented. It is observable that for the small prediction horizons (i.e. $W_p = 10$), performance increase achieved via prediction is very small with respective to the different prediction horizons. Decreasing RI length is forcing the scheduler to perform very close to a real-time scheduler and decreases the visibility of the anticipatory gain. More specifically, in an extreme case of $RI = 1$ TTI, the proposed anticipatory algorithm would perform identically to the no-prediction case due to the per-slot negotiations. As the prediction horizon increases, the performance of the no-filter algorithm decreases sharply due to the prediction errors. However, it is also visible in Fig. 5.3, that using filter always outperforms the rest of the approaches in terms of the achieved utilities. Fig. 5.3 demonstrates that the proposed filter mechanism can prevent the prediction errors from interfering the resource negotiations while providing the benefits of accurate predictions.

The impact of number of tenants $|M|$ (assuming proportionally increasing $|K|$) on average achieved utility is demonstrated in Fig. 5.4 for three different scenarios, namely 'no prediction', 'no filter' and 'with filter'. For this particular scenario, the renegotiation interval is assumed to be same as the prediction window, $RI = W_P = 25$ TTIs, while the learning window is set to be $W_L = 75$ TTIs. As each new tenant m causes a proportional increase in the network congestion, i.e. $K_m = 4$, we can see the decrease in the average utility as $|M|$ increases. More importantly, we can also observe that the advantages of anticipatory network slicing is disappearing as the number of tenants increases. As the number of non-elastic users increases in the network, the infrastructure provider loses its flexibility in assigning

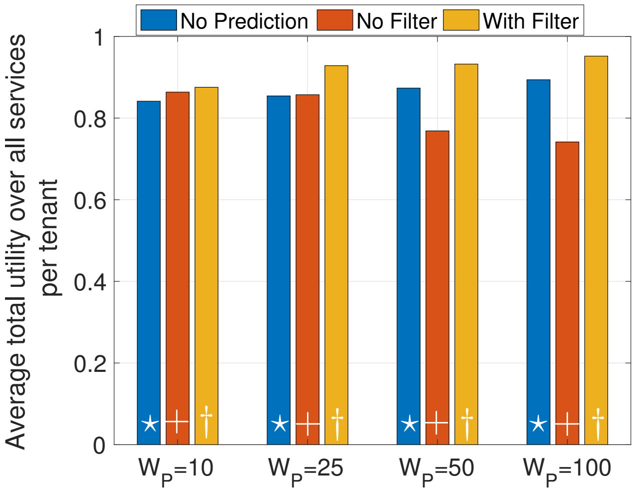


Figure 5.3: *Comparison of different prediction approaches for different prediction horizons and different scenarios, i.e. no prediction (marked with '$\star$'), no filter (specified with '$+$') and with filter (specified with '$\dagger$').*

the resources and the anticipatory gains saturate. Therefore, in order to fully explore the anticipatory advantages, the network capacity is required to be scaled proportionally to the traffic demand.

5.4.4 Business implications of anticipation

The envisioned market model's economic sustainability and the impact of anticipatory networking are investigated using the users' willingness to accept the price p for the given QoS. In order to accurately asses this aspect, we have used the acceptance probability consept that is introduced in [10], i.e.

$$\left(\frac{U_{k,M_1}}{U_{k,M_2}}\right)^{\mu} \leq \left(\frac{p_{k,M_1}}{p_{k,M_2}}\right)^{\sigma}. \tag{5.30}$$

where U_k represents the average achieved utility of the tenants for the given scenario. The service quality is considered to be accepted if (5.30) holds. If the users are satisfied with the received service for the given price, the tenants are assumed to stay in the sharing agreement. Otherwise, the tenants are considered to be leaving the sharing platform as sharing cannot provide the desired QoS for a reasonable cost. The impact of anticipation on sustainability of sharing model is outlined in Table 5.5 and Table 5.6 where we present the cases with and without prediction, when $W_p = 25$ TTIs. The cases where (5.30) holds are indicated with 'YES' and with 'NO' for the

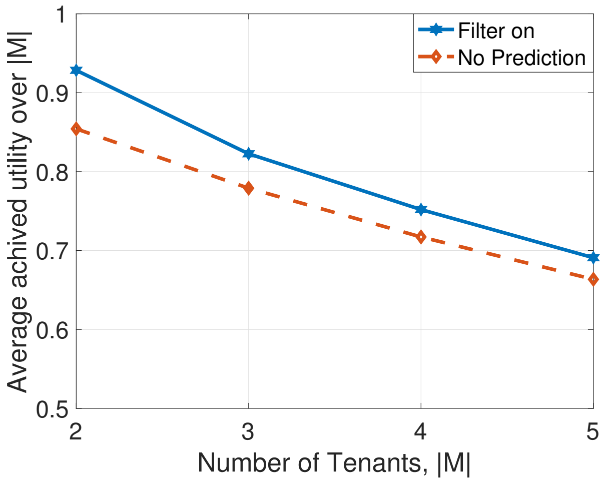


Figure 5.4: *Average achieved utility over $|M|$ for 'no prediction' and 'filter on' scenarios*

cases it is not.

Table 5.5: *Evaluation of Eq.* (5.30) *when increasing the number of tenants when capacity is fixed without prediction*

$\|M_1\| \to \|M_2\|$	$\left(\frac{U_{k,M_1}}{U_{k,M_2}}\right)^{\mu}$	$\left(\frac{p_{M_1}}{p_{M_2}}\right)^{\epsilon}$	Status
$2 \to 3$	1.1598	3.1820	YES
$3 \to 4$	1.1736	1.6810	YES
$4 \to 5$	1.1901	1.1802	NO

Table 5.6: *Evaluation of Eq.* (5.30) *when increasing the number of tenants when capacity is fixed with filter*

$\|M_1\| \to \|M_2\|$	$\left(\frac{U_{k,M_1}}{U_{k,M_2}}\right)^{\mu}$	$\left(\frac{p_{M_1}}{p_{M_2}}\right)^{\epsilon}$	Status
$2 \to 3$	1.2555	2.7378	YES
$3 \to 4$	1.2247	1.6242	YES
$4 \to 5$	1.1815	1.2001	YES

A comparison between two tables demonstrates that the anticipatory network management can increase resource efficiency as well as the cost reduction while providing a comparable average utility. Moreover, the QoE increment achieved by the exploitation of anticipatory information increases the tenants' willingness to accept prices for some cases that would not be accepted in 'no-prediction' scenario. This result shows that the anticipatory network sharing is not only critical for increasing resource effi-

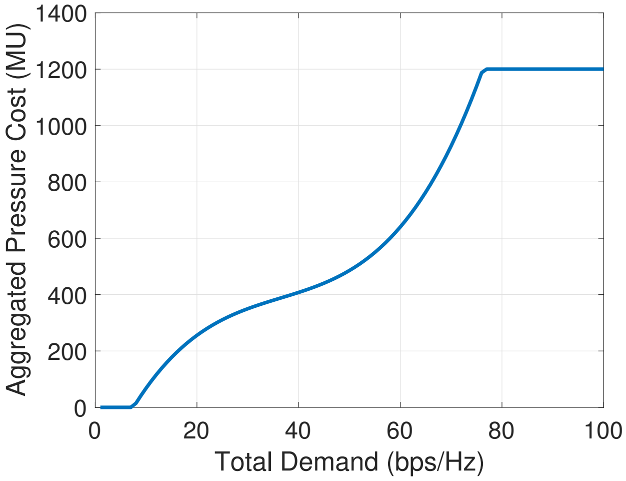


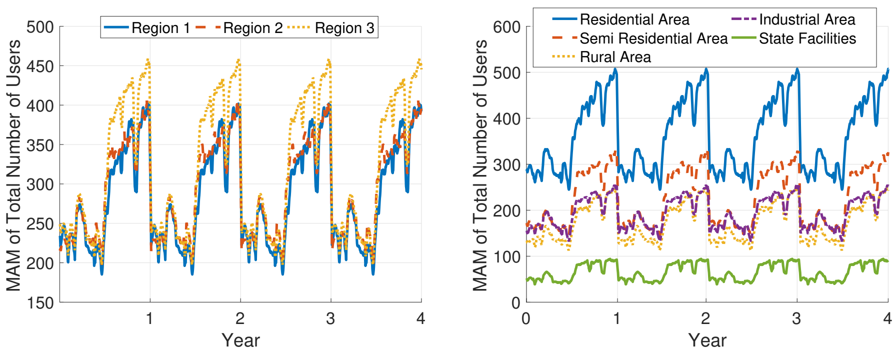
Figure 5.5: *The characterization of the real time resource scheduling and trading framework.*

ciency but also for expanding the wireless market size (i.e. the number of tenants that can be served).

5.4.5 Long term traffic model and the characterization of the real time scheduler

In the following sections, we present the long-term observations (i.e. a few years) of the proposed framework. However, the proposed anticipatory resource scheduling and trading framework is designed to run in real-time, meaning that an exact observation of the model behavior in a year would require a year of simulation. Thus, without loss of generality, we have used the characteristic of our short term model and generated a simple function in order to calculate the aggregated sum of pressure revenue variation by the total number of users. However, in the long term analysis, the number of users and the demanded service types are expected to be varying over time which makes a function that relies on the number of users impractical. More specifically, a base station with twenty elastic services would not be in the same condition as another base station with twenty inelastic services. Therefore, instead of the number of users, we have modeled the daily collected pressure revenue of the base station using the total daily demand in the base station. Based on a large set of simulations with different user locations and number of users, we have empirically derived the function,

$$f(x) = 0.2854x^3 - 57.7510x^2 + 4496.5x + 51663 \tag{5.31}$$



(a) *Symmetric traffic distributions*

(b) *Asymmetric traffic distributions*

Figure 5.6: *Total number of users over time per region.*

whose characteristic is given in Fig. 5.5.

As aforementioned, the long term variation of the network is governed by the traffic demand rather than the users' achievable rates. The performance analysis is performed using actual traffic traces collected from 39 base stations in Bergamo (Italy) between March 2018 and March 2019. We consider two scenarios, one where there are equivalent traffic distributions among regions (cf. Fig. 5.6a), and the one scenario with asymmetric traffic distribution, where the regions are determined according to the traffic mix they produce (cf. Fig. 5.6b). In contrast to the symmetric traffic distribution scenario, where the traffic mix is considered to be the same for all the regions, in asymmetric traffic distribution scenario, the traffic mix is assumed to be as presented in Table 5.7. In order to isolate the effects of the data set, we assume that the cyclic characteristics of the network demand are the same over all the simulation horizon. Therefore, the one year data is extended into four years by copying the available data (cf. Fig. 5.6).

Table 5.7: *The service distributions in percentage (%) per region in scenario 2.*

Region Type	Elastic	Inelastic	M2M	Background
Residential	25	25	25	25
Semi-residential	50	25	0	25
Rural	25	0	0	75
Industrial	25	0	75	0
State-facilities	0	0	100	0

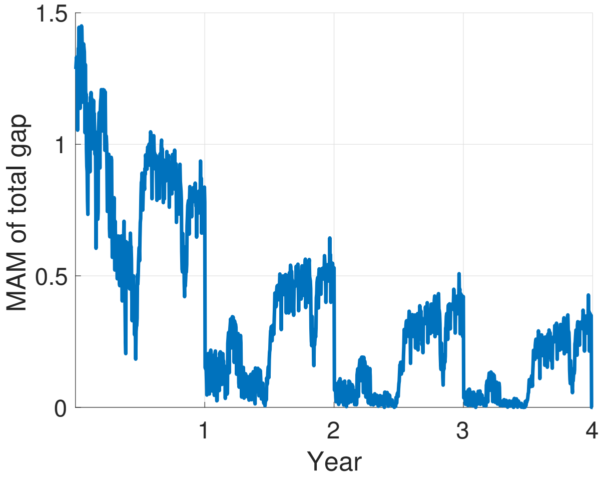


Figure 5.7: *Evolution of total gap over time.*

5.4.6 Capacity evolution for symmetric traffic demand scenario

Fig. 5.7 presents the evolution of the total gap over the complete coverage area. Observation window is chosen to be 1 month, i.e. $W_{ex} = 1$ month. The first year in Fig. 5.7 can be considered to be the transition period for the network capacity. Namely, the network capacity is not shaped in line with the traffic demand, resulting in a very high total gap over all the regions. Consequently, the increase in the available capacity decreases the measured gap quickly, which is especially visible during the first half of the first year. On the second half, however, as a direct consequence of the increasing traffic demand (cf. Fig. 5.6a), a higher total gap is measured.

Over the years, we observe a gradual decrease in the total gap. On the other hand, the decreasing total gap causes a decrease in the total accumulated pressure revenue. Once the traffic profile does not increase over years, also the capacity expansion stabilizes. Consequently, the decrease in the total gap becomes less visible over the years. On the other hand, following the key findings in the previous chapter, this result indicates that the small values of gap only impacts the elastic users, and due to their elasticity, the expansion would require longer time.

Fig. 5.8 shows the variation of the total pressure revenue collected from all the base stations within the total coverage area over time. At the end of each month, the slice-aware capacity expansion algorithm is run to determine the expansion need. In line with Fig. 5.7, the infrastructure provider collects a very high pressure revenue during the first year which is then

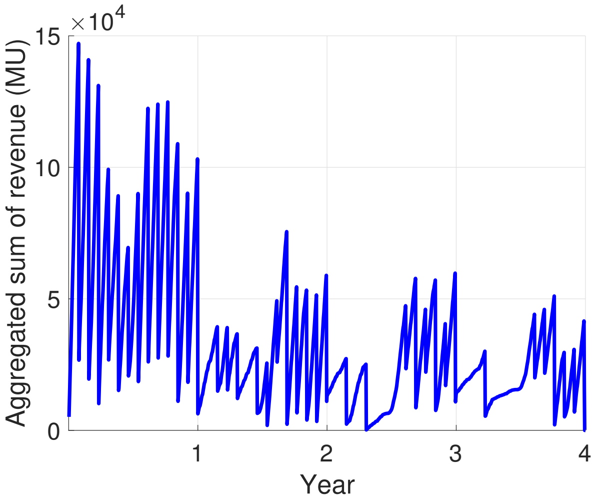
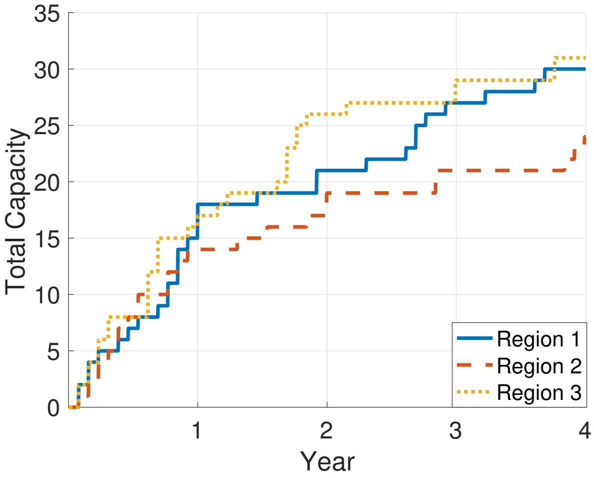


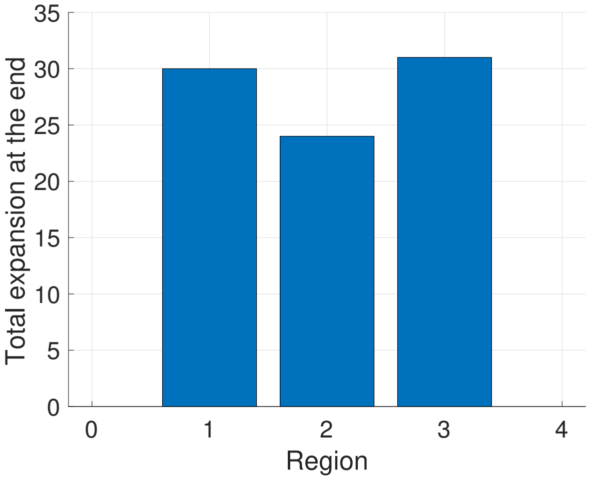
Figure 5.8: *Collected pressure revenue over time.*

used in the capacity expansion. After the first year, we are seeing that the aggregated revenue is decreasing and proportionally the capacity expansion frequency is also decreasing. During the first half of the fourth year, the infrastructure provider only deploys a few base stations. Despite the increase in the aggregated revenue after this new deployments, the necessary revenue could not be collected to trigger additional capacity expansion. This result shows us that the proposed framework is immune to the temporary increases in the network traffic which can be a result of a special event. More specifically, in order for the measured gap to trigger a capacity expansion, it either has to be extremely high - which is not usually possible with an accurately shaped network- or facing the gap over a long time.

Finally, Fig. 5.9 shows the capacity evolution of the network over time (cf. Fig. 5.9a) and place (cf. Fig. 5.9b). During the transition period, i.e. first year, we can see that the capacity deployment is rapidly and equivalently occurring in all three regions, cf. Fig. 5.9a. The proposed framework maintains the fairness among different regions while expanding the capacity. Following the intensive deployment of the first year, we are seeing a rather steep and differentiated curve among different regions. More specifically, after the infrastructure deployments during the first year, the available pressure revenue and proportionally the maximum number of new base stations decrease. Therefore, the slice-aware capacity expansion algorithm considers both the required capacity (i.e. measured in terms of gap) and the total number of users in a region. A very important aspect of the proposed



(a) *Total number of new base station deployments per region over time.*



(b) *The distribution of newly deployed base stations among regions.*

Figure 5.9: *The capacity evolution of the network.*

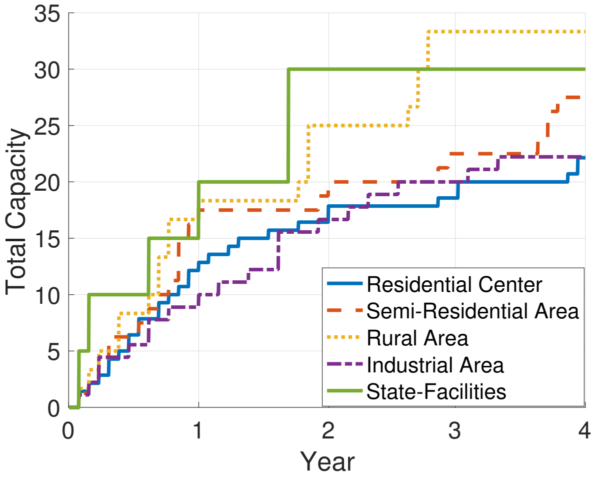


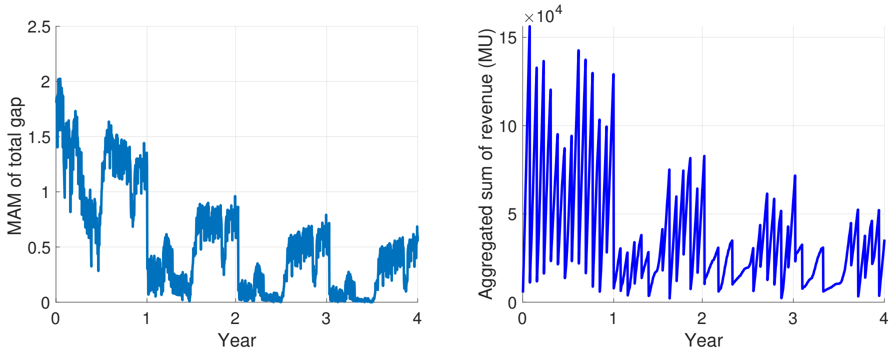
Figure 5.10: *The base station deployment per region for the asymmetric scenario.*

objective function in (5.24.a) is its prioritization to the regions with smaller number of users. Namely, if two regions are facing the same QoS degradation, the scheduler always chooses the region with smaller user density. This idea is especially important as the total number of deployable base stations is fixed. Consequently, the scheduler tries to minimize the observed total gap and the regions with the smallest user density is easier to satisfy with the minimum number of base stations.

Finally, Fig. 5.9b outlines the spatial distribution of the resource allocations. Although the findings in Fig. 5.7 show that the total deployment of the base stations is not finalized (i.e. because $\sum\limits_{r \in R} \xi_r > 0$), we can still observe the fairness among three regions in Fig. 5.9b. Consequently, Fig. 5.9b shows that given that the total demand as well as the traffic mix are comparable with each other, all the regions are equally served.

5.4.7 Capacity evolution for asymmetric traffic mix

The asymmetric traffic distribution scenario is analyzed in this section. As aforementioned, the regions in this scenario are separated according to the type of traffic they produce. The traffic volume per region is also asymmetrically distributed as the number of base stations per region varies (cf. Fig. 5.6b). The evolution of the network capacity per region over time is presented in Fig. 5.10. The highest priority level is given to the state-facilities (green line) as it represents the facilities with the uttermost impor-



(a) *The moving arithmetic mean of the total gap over all the regions* **(b)** *Aggregated pressure revenue*

Figure 5.11: *The evolution of the network performance for the asymmetric scenario.*

tance (e.g. hospitals, police). As a direct result of this high priority, despite its low user density, this region is chosen to be the first to receive additional capacity (cf. green line in Fig. 5.10). Note that the iterative application of our proposed algorithm allows a certain level of investment on the rest of the regions in addition to the state-facilities.

The implications of capacity expansion on network performance is presented in Fig. 5.11, which contains the evolution of total gap (cf. Fig. 5.11a), and the fluctuation of the aggregated revenue over time (cf. Fig. 5.11b). In line with our observations during the symmetric traffic distribution scenario, the deployment of new resources decreases the total gap. The higher observed gap value in Fig. 5.11a with respect to the symmetric traffic distribution scenario (i.e. Fig. 5.7) is due to the higher number of regions (i.e. two new regions are presented in this scenario). Fig. 5.11b also shows that regardless of the asymmetric traffic volume or the traffic mix, the proposed model performs efficient and timely capacity expansion. In particular, the transient conditions such as short term fluctuations in the traffic demand do not impact the expansion decisions as they cannot cause a noticeable impact on the accumulated revenue.

5.4.8 Effects of observation window

The impact of observation period, i.e. W_{ex}, on the total gap over the coverage area is investigated for four different observation window settings, namely, daily, weekly, biweekly and monthly. Fig. 5.13 presents how the network evolves after four years while Fig. 5.12 is focused on the first 6 month of using different observation intervals. One can see in Fig. 5.12

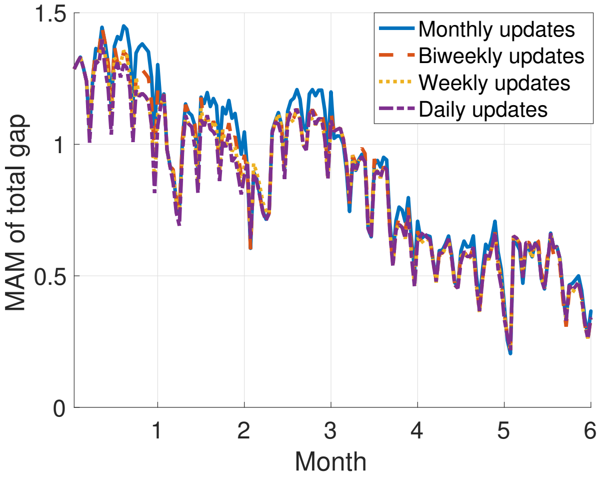


Figure 5.12: *Impact of using different expansion windows on the evolution of the total gap (6 months view)*

that during the month, the daily observation window (i.e. purple dash-dot line) provides less total gap with respect to the monthly observation window (i.e. blue solid line) since the scheduler can deploy new base stations as soon as the necessary revenue is collected for one base station. Although this result is expected in any short observation window, we can also see that at the end of the month, the monthly observation period also achieves the same gap level. This shows that despite the possible variations on the regional base station allocation decisions, the overall evolution of the network, namely the QoE increment with the deployed base station, is equal for both scenarios. More specifically, the designed expansion objective, i.e. (5.15), guarantees that regardless of the observation period, the deployed base station is always placed in the region that it would create the highest impact on the QoE. Therefore, the different placement decisions of the base stations between different observation windows do not have an impact on the long term network performance.

Fig. 5.13 presents the overall network evolution in terms of total gap over four years of time interval. The result also confirms that even though shorter observation windows can be more advantageous within short time intervals, e.g. 1 month, in the long term all the observation periods result in same QoE increases. This result shows that our proposed slice-aware capacity expansion strategy can quickly respond to the increasing demand and can deploy new base stations with the similar QoE gains as the long observation periods.

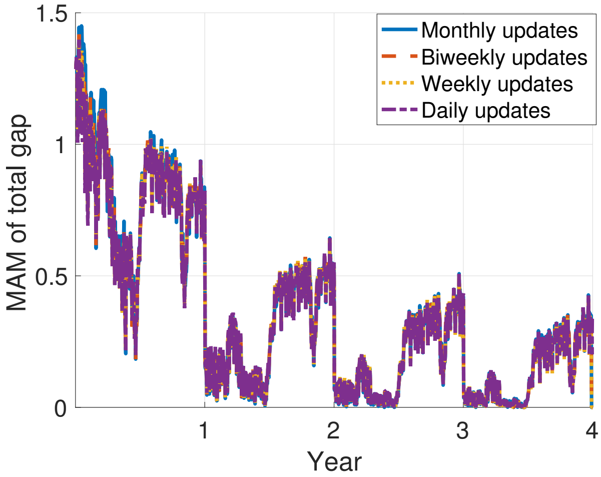


Figure 5.13: *Impact of using different expansion windows on the evolution of the total gap (4 years view)*

5.5 Summary

In this chapter, we have focused on the long and short term implications of anticipatory networking. In the short time scale analysis, we explore the integration of anticipatory information into the short time scale resource scheduling and trading decisions. In order to limit the impact of inaccurate predictions, we have developed a novel filtering mechanism which can filter out the impacts of prediction errors while providing the advantages of anticipatory networking. The proposed filtering mechanism along with the two-step optimization model have shown to provide efficiency and market sustainability. Moreover, our analysis has shown that, in order to take the advantage of anticipatory networking, a sufficient network capacity is required. However, the variety of the services and tenant based policies harden the capacity management decisions for the shared infrastructure resources. In order to maintain the business value and fairness among tenants, the infrastructure provider has to have a clear understanding of the traffic mix, the total required capacity and the urgency of the additional resources. Therefore, on the second part of this chapter, we proposed a novel slice-aware capacity expansion strategy. The proposed framework provides base station deployment decisions based on the QoE degradation among the regions, the available revenue for expansion and the urgency of the expansion. Moreover, the proposed capacity scaling strategy maintains its efficiency even with a smaller observation period (e.g. a day). Therefore, the network evolution can be achieved in proportion to the increment of the traffic demand in an accurate and quick manner.

CHAPTER 6

Conclusions

Starting from the transition towards 3G, the main competition among mobile network operators has been on providing the highest quality for the cheapest price. However, this economic pressure on their business models has created an unsustainable mobile market that is gradually monopolized by a few big telcom operators. This is mostly due to the fact that the sky rocketing network demand is not reflected on the mobile operators' revenues per user and consequently the profit margins of the mobile network operators have been shrinking. The network operators have been responding to the deceasing profits with an overly extensive cost optimization, also including making passive network sharing agreements with competitors, switching off base stations in low demand duration etc. However, the common assumption in all the previous solutions is the understanding of standalone network operator. Namely, the network operators have always been considered to be standalone entities with no long-term cooperation with their competitors. Moreover, in order to guarantee service continuity, the network operators depend on over provisioning the available spectrum resources. This, however, can lead excessive base station deployments in the regions that do not possess any business value. On the excessively demanding techno-economic ecosystem of 5G, this fundamental assumption of standalone network operators has to be revisited.

5G is envisioned to host a multitude of industry driven applications that require not only very high data rate but also low delay constraints. Therefore, in the conventional standalone approach, the network operators are forced to deploy further base stations to meet these expectations. On the other hand, the increased base station density places further strains on the already decreasing profitability of the network provisioning and turns it into a highly unstable business model. Consequently, the research focus has been shifted towards the methods that can increase the available capacity without forcing the network operators to deploy new base stations. On the other hand, the recent estimations state that the predecessor network generation (4G) has already reached a spectral efficiency that is very close to Shannan capacity and further improvements are more expensive than the economic gains from the additional users. As a result, the industrial attention moves to increasing already existing cooperation level between the competing entities in the network and moving the actual competition to the service differentiation rather than capacity provisioning. Although the main competition in developed (i.e. capacity driven) broadband markets has already shifted towards service differentiation, the operators require further investigation on the sharing models and the available flexibility to differentiate their services in a shared infrastructure. In this PhD thesis, a real time network sharing and trading algorithm has been proposed and analyzed in terms of its capability to enable service differentiation, and long and short-term implications on the market evolution.

Our main research question revolves around how the resource sharing process can be automatized and what the long- and short-term implications of this automatized approach are. Three main research questions are identified to guide the research process,

$RQ1$. How can the network resources dynamically and flexibly be shared in a multi-tenant network?

$RQ2$. How can the tenants differentiate their services in a shared infrastructure?

$RQ3$. What are the long- and short-term implications of anticipatory network sharing and resource trading?

Real-time infrastructure sharing

On one hand the sharing has to be guaranteeing the resource flexibility and the inter-tenant fairness while on the other hand it provides enhanced cost efficiency. However, both of these aspects can not be satisfied simultaneously with the long-term service level agreement (SLA) based sharing approaches. Consequently, we start our modeling with a new understanding of the SLAs. In our envisioned system, the SLAs only contain long term aspects, namely available budget of the operators, the utility expectations and the unit price per resource. The rest of the attributes of sharing are dynamically defined according to the instantaneous condition of the network and the tenants' long term expectations. Thus, the envisioned flexibility and efficiency are achieved by increasing the inter-tenant relatedness while the fairness among tenants is induced by a set of constraints that guarantee equal resource distribution for the symmetric scenario. In order to guarantee that the tenants can always pursue their interest and reshape their resource shares accordingly, our framework renegotiates the sharing parameters within a predefined renegotiation interval.

Also, a novel market driven pricing algorithm is proposed to ensure that the infrastructure provider collects the required revenue in order to trigger a capacity expansion. Our proposed pricing mechanism dynamically scales the resource prices proportionally to the total demand. This scaling mechanism (i.e. the pressure cost) regularizes the resource demand while collecting the necessary revenue for a future capacity expansion.

Dynamic network slicing and slice trading

Our second research question addresses how the tenants can differentiate their services in a shared framework. Network slicing is considered to be a key enabler to achieve service coexistence without loss of QoE. On the other hand, the conventional approach relies on static network slicing that is applied based on long term statistics. However, static approaches have nearly always ended up in overprovisioning. Furthermore, in a multi-tenant network, inter tenant dynamics are also required to be considered. Consequently, built upon the proposed market model, we have extended our work using network slicing to serve multiple services with conflicting expectations. As a first step, a novel utility function has been proposed to map the heterogeneous QoS expectations and service priorities into a uniform QoE value. The proposed network scheduler redefines the slice sizes (i.e. the assigned resources per slice) in order to achieve the highest possible QoE. Next, we have concentrated on the service differentiation among tenants. Being a key aspect in competition in the network provisioning, service differentiation lies in the core of our model. We have shown that the tenants can differentiate their services by choosing different parameters and this would be reflected in the resource allocations and total costs of tenants.

However, the differentiated services do not violate the fairness among tenants.

Consequently, built upon the dynamic network sharing, network slicing and resource trading concepts, we have defined a new market model that can support both the main players from the conventional business model (e.g. network operators or infrastructure providers) and the new players such as over the top service providers. The provided framework does not only provide an economical platform but also supports innovation by decreasing the time to market duration and on demand availability of any desired resource.

Long and short term implications

The third and last research focus of this thesis is on the long and short analysis of the proposed market ecosystem. Meeting the expectations of 5G and beyond technology directs the research to mitigating from the reactive network management approach to proactive network management. The advances in the artificial intelligence area provide a large set of tools that can be used to process the available data and provide high accuracy predictions regarding the upcoming changes in the network. A key problem is the well-known tradeoff between prediction accuracy and the time complexity of the available prediction algorithms. In a real time algorithm, such as ours, the predictions regarding the changes in the network conditions have to be available during the scheduling decisions. However, lower time complexity usually comes with low prediction accuracy. Therefore, we developed a novel filtering mechanism in order to filter out the possible implications of the low prediction accuracy. This way, we can filter out the impacts of prediction errors while exploiting the full potential of accurate predictions. Our investigation has shown that, the anticipatory network slicing and trading can increase the spectral efficiency and further decrease the total costs. From a market point of view, it has proven to increase the number of tenants that can be served by the infrastructure provider without a need to expand the network capacity. However, it also underlines the importance of a timely and accurate capacity expansion in order to preserve the advantages of anticipation.

Consequently, on the second part of Chapter 5.5, we have focused on a novel self-dimensioning and network management framework. As aforementioned, our proposed framework dynamically collects the additional revenue in order to trigger the capacity expansion in the long term, i.e. *pressure cost*. The proposed self-dimensioning algorithm uses this revenue in order to scale the network resources in line with the demand. The decisions are performed based on $i.$ needed capacity expansion, $ii.$ urgency of expansion need and $iii.$ available revenue. Our investigations showed that the proposed model can indeed efficiently reinvest the collected revenue and minimize the total gap over time. Moreover, the proposed model can use short term observations in order to perform accurate capacity expansions and, therefore, reduces the need for long term observations. More specifically, the proposed framework does not require long term observations in order to determine the regions that are valuable to expand the capacity. Thus, it is possible to use mobile base stations in order to manage short-term fluctuations in the network demand.

Bibliography

[1] The mobile access network, beyond connectivity.

[2] Ö. U. Akgül, I. Malanchini, and A. Capone. Anticipatory resource allocation and trading in a sliced network. In *2019 IEEE International Conference on Communications (ICC)*, pages 1–7, May 2019.

[3] Ö. U. Akgül, I. Malanchini, and A. Capone. Dynamic resource trading in sliced mobile networks. *IEEE Transactions on Network and Service Management*, 2019.

[4] Ö. U. Akgül, I. Malanchini, V. Suryaprakash, and A. Capone. Dynamic resource allocation and pricing for shared radio access infrastructure. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–7, May 2017.

[5] Ö. U. Akgül, I. Malanchini, V. Suryaprakash, and A. Capone. Service-aware network slice trading in a shared multi-tenant infrastructure. In *IEEE Global Communications Conference (GLOBECOM)*, 2017.

[6] Ö. U. Akgül, I. Malanchini, V. Suryaprakash, and A. Capone. Slice-Aware Capacity Expansion Strategies in Multi-Tenant Networks. Submitted.

[7] Alcatel-Lucent. 5G is coming: Are you prepared? 2015.

[8] S. Anbazhagan and N. Kumarappan. Day-ahead deregulated electricity market price forecasting using recurrent neural network. *IEEE Systems Journal*, 7(4):866–872, Dec 2013.

[9] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta. Are we approaching the fundamental limits of wireless network densification? *IEEE Communications Magazine*, 54(10):184–190, October 2016.

[10] L. Badia, M. Lindstrom, J. Zander, and M. Zorzi. Demand and pricing effects on the radio resource allocation of multimedia communication systems. In *IEEE Global Telecommunications Conference, GLOBECOM '03*, pages 4116–4121, Dec 2003.

[11] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez. Optimising 5G infrastructure markets: The business of network slicing. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, May 2017.

[12] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez. DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning. In *IEEE INFOCOM*, Paris, France, April 2019.

[13] G. Berardinelli, N.H. mahmood, I. Rodriguez, and P. Mogensen. Beyond 5G wireless irt for industry 4.0: Design principles and spectrum aspects. In *GLOBECOM WORKSHOPS 2018 - 2018 IEEE Global Communications Conference Workshops*, Dec 2018.

[14] R. Berry, M. Honig, T. Nguyen, V. Subramanian, H. Zhou, and R. Vohra. On the nature of revenue-sharing contracts to incentivize spectrum-sharing. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2013.

[15] N. Bhushan, J. Li, D. Malladi, R. D. 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*, 52:82–89, 2014.

[16] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer. A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques. *IEEE Communications Surveys Tutorials*, 19(3):1790–1821, 2017.

[17] L. Cano, A. Capone, G. Carello, and M. Cesana. Evaluating the performance of infrastructure sharing in mobile radio networks. In *2015 IEEE International Conference on Communications (ICC)*, pages 3222–3227, June 2015.

[18] L. Cano, A. Capone, G. Carello, M. Cesana, and M. Passacantando. On optimal infrastructure sharing strategies in mobile radio networks. *IEEE Transactions on Wireless Communications*, 16(5):3003–3016, May 2017.

[19] P. K. Chartsias, A. Amiras, I. Plevrakis, I. Samaras, K. Katsaros, D. Kritharidis, E. Trouva, I. Angelopoulos, A. Kourtis, M. S. Siddiqui, A. Vines, and E. Escalona. SDN/NFV-based end to end network slicing for 5G multi-tenant networks. In *2017 European Conference on Networks and Communications (EuCNC)*, pages 1–5, June 2017.

[20] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar. Dynamic network slicing for 5G IoT and eMBB services: A new design with prototype and implementation results. In *2018 3rd Cloudification of the Internet of Things (CIoT)*, pages 1–7, July 2018.

[21] M. Darula and I. Mas. Zero touch networks with cloud-optimized network applications. In *Ericsson White Paper*, 2017.

[22] K. David and H. Berndt. 6G vision and requirements: Is there any need for beyond 5G? *IEEE Vehicular Technology Magazine*, 13(3):72–80, Sep. 2018.

[23] M. Dohler. The future and challenges of communications – Toward a world where 5G enables synchronized reality and an internet of skills. In *Internet Technology Letters*, pages 1–3, 2018.

[24] M. Draxler, J. Blobel, and H. Karl. Anticipatory download scheduling in wireless video streaming with uncertain data rate prediction. In *2015 8th IFIP Wireless and Mobile Networking Conference (WMNC)*, pages 136–143, Oct 2015.

[25] Y. Du. Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network. In *2018 Chinese Control And Decision Conference (CCDC)*, pages 2854–2857, June 2018.

[26] A. Eisenblatter, H. F. Geerdes, and M. Grotschel. Planning UMTS radio networks. *OR/MS Today*, 35:41–46, 2008.

[27] Z. Frias and J. P. Martinez. 5G networks: Will technology and policy collide? *Telecommunications Policy*, 42(8):612 – 621, 2018. The implications of 5G networks: Paving the way for mobile innovation?

[28] T. Frisanco, P. Tafertshofer, P. Lurin, and R. Ang. Infrastructure sharing for mobile network operators; from a deployment and operations view. In *2008 International Conference on Information Networking*, pages 1–5, Jan 2008.

[29] A. T. Gamage, Q. Shen, and X. Shen. Cloud assisted resource management for hyper-dense small cell networks. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec 2015.

[30] A. Georgakopoulos, I. Belikaidis, K. Tsagkaris, V. Stavroulaki, and P. Demestichas. Wireless access infrastructure expansions through opportunistic networks of moving access points. In *2016 European Conference on Networks and Communications (EuCNC)*, pages 163–167, June 2016.

[31] G. D. Gonzalez, H. Hakula, A. Rasila, and J. Hamalainen. Spatial mappings for planning and optimization of cellular networks. *IEEE/ACM Transactions on Networking*, 26:175–188, 2018.

[32] A. Gran, S. C. Lin, and I. F. Akyildiz. Towards wireless infrastructure-as-a-service (WIaaS) for 5G software-defined cellular systems. In *2017 IEEE International Conference on Communications (ICC)*, 2017.

[33] GSMA. Mobile infrasturcture sharing. In *White Paper*, 2012.

[34] J. Guey, P. Liao, Y. Chen, A. Hsu, C. Hwang, and G. Lin. On 5G radio access architecture and technology [industry perspectives]. *IEEE Wireless Communications*, 22(5):2–5, October 2015.

[35] Gurobi Optimization Inc. Gurobi optimizer reference manual, 2015.

[36] S. Han, C. I, G. Li, S. Wang, and Q. Sun. Big data enabled mobile network design for 5G and beyond. *IEEE Communications Magazine*, 55(9):150–157, Sep. 2017.

[37] D. Harutyunyan and R. Riggio. How to migrate from operational LTE/LTE-A networks to C-RAN with minimal investment? *IEEE Transactions on Network and Service Management*, 15(4):1503–1515, Dec 2018.

[38] M. Jiang, M. Condoluci, and T. Mahmoodi. Network slicing management & prioritization in 5G mobile systems. In *22 European Wireless 2016; 22th European Wireless Conference*, pages 1–6. IEEE, May 2016.

[39] M. Jiang, M. Condoluci, and T. Mahmoodi. Network slicing in 5G: An auction-based model. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2017.

[40] M. I. Kamel, L. B. Le, and A. Girard. LTE wireless network virtualization: Dynamic slicing via flexible scheduling. In *IEEE Vehicular Technology Conference (VTC Fall)*, pages 1–5, Sept 2014.

[41] E. Kapassa, M. Touloupou, and D. Kyriazis. SLAs in 5G: A complete framework facilitating VNF- and NS- tailored SLAs management. In *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 469–474, May 2018.

[42] K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini, and T. Braun. Network slices toward 5G communications: Slicing the LTE network. *IEEE Communications Magazine*, 55(8):146–154, Aug 2017.

[43] A. Ksentini, P. A. Frangoudis, A. PC, and N. Nikaein. Providing low latency guarantees for slicing-ready 5G systems via two-level mac scheduling. *IEEE Network*, 32(6):116–123, November 2018.

[44] A. Ksentini and N. Nikaein. Toward enforcing network slicing on RAN: Flexibility and resources abstraction. *IEEE Communications Magazine*, 55(6):102–108, June 2017.

[45] F. Kurtz, C. Bektas, N. Dorsch, and C. Wietfeld. Network slicing for critical communications in shared 5G infrastructures - an empirical evaluation. In *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, pages 393–399, June 2018.

[46] W. Lemstra. Leadership with 5G in Europe: Two contrasting images of the future, with policy and regulatory implications. *Telecommunications Policy*, 2018.

[47] N. Liakopoulos, G. S. Paschos, and T. Spyropoulos. Robust user association for ultra dense networks. *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 2690–2698, 2018.

[48] A. Lieto, I. Malanchini, V. Suryaprakash, and A. Capone. Enabling dynamic resource sharing for slice customization in 5G networks. In *IEEE Global Communications Conference (GLOBECOM)*, 2018.

[49] C. Lin, K. Chen, D. Wickramasuriya, S. Lien, and R. D. Gitlin. Anticipatory mobility management by big data analytics for ultra-low latency mobile networking. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–7, May 2018.

[50] F. Y. Lin, C. Hsiao, Y. Wen, and Y. Wu. Optimization-based resource management strategies for 5G C-RAN slicing capabilities. In *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 346–351, July 2018.

[51] R. T. B. Ma, J. Wang, and D. M. Chiu. Paid prioritization and its impact on net neutrality. *IEEE Journal on Selected Areas in Communications*, 35(2):367–379, Feb 2017.

[52] R. Madan and P. SarathiMangipudi. Predicting computer network traffic: A time series forecasting approach using DWT, ARIMA and RNN. *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–5, 2018.

[53] I. Malanchini and M. Gruber. How operators can differentiate through policies when sharing small cells. In *IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015.

[54] I. Malanchini and V. Suryaprakash. Minimizing the impact of prediction errors during anticipatory resource allocation. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pages 1–6, June 2018.

[55] I. Malanchini, S. Valentin, and O. Aydin. Wireless resource sharing for multiple operators: Generalization, fairness, and the value of prediction. *Computer Networks*, 100:110 – 123, 2016.

[56] K. Mallinson. The path to 5G: as much evolution as revolution. In *3GPP - The Mobile Broadband Standard*, May 2016.

[57] L. Mamushiane, A. A. Lysko, and S. Dlamini. SDN-enabled infrastructure sharing in emerging markets: CapEx/OpEx savings overview and quantification. In *2018 IST-Africa Week Conference (IST-Africa)*, pages Page 1 of 10–Page 10 of 10, May 2018.

[58] 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 *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, MobiCom '18, pages 191–206, New York, NY, USA, 2018. ACM.

[59] D. E. Meddour, T. Rasheed, and Y. Gourhant. On the role of infrastructure sharing for mobile network operators in emerging markets. *Computer Networks*, 55(7):1576–1591, 2011.

[60] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah. Drone small cells in the clouds: Design, deployment and performance analysis. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec 2015.

[61] P. Munoz, O. Sallent, and J. Perez-Romero. Self-dimensioning and planning of small cell capacity in multitenant 5G networks. *IEEE Transactions on Vehicular Technology*, 67(5):4552–4564, May 2018.

[62] K. R. Nair, V. Vanitha, and M. Jisma. Forecasting of wind speed using ANN, ARIMA and hybrid models. In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pages 170–175, July 2017.

[63] NGMN. 5G white paper. 2015.

[64] W. Ni, I. B. Collings, J. Lipman, X. Wang, M. Tao, and M. Abolhasan. Graph theory and its applications to future network planning: software-defined online small cell management. *IEEE Wireless Communications*, 22(1):52–60, February 2015.

[65] L. Nie, D. Jiang, S. Yu, and H. Song. Network traffic prediction based on deep belief network in wireless mesh backbone networks. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–5, March 2017.

[66] OECD. Wireless market structures and network sharing. 2014.

[67] 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*, 55(5):80–87, May 2017.

[68] J. S Panchal, R. Yates, and M. M. Buddhikot. Mobile network resource sharing options: Performance comparisons. *IEEE Transactions on Wireless Communications,*, 12(9):4470–4482, 2013.

[69] E. Pateromichelakis and K. Samdanis. A graph coloring based inter-slice resource management for 5G dynamic TDD RANs. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2018.

[70] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí. Artificial intelligence-based 5G network capacity planning and operation. In *2015 International Symposium on Wireless Communication Systems (ISWCS)*, pages 246–250, Aug 2015.

[71] A. Popovska Avramova and V. B. Iversen. Radio access sharing strategies for multiple operators in cellular networks. In *2015 IEEE International Conference on Communication Workshop (ICCW)*, pages 1113–1118, June 2015.

[72] M. Rahman, S. H. Ahmed, and M. Yuksel. Proof of sharing in inter-operator spectrum sharing markets. In *2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pages 1–5, Oct 2018.

[73] M. Richart, J. Baliosian, J. Serrat, and J. L. Gorricho. Resource slicing in virtual wireless networks: A survey. *IEEE Transactions on Network and Service Management*, 13(3):462 – 476, 2016.

[74] 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*, 55(5):72–79, May 2017.

[75] D. Sahinel, C. Akpolat, M. A. Khan, F. Sivrikaya, and S. Albayrak. Beyond 5G vision for IOLITE community. *IEEE Communications Magazine*, 55(1):41–47, January 2017.

[76] T. Sanguanpuak, S. Guruacharya, E. Hossain, N. Rajatheva, and M. Latva-aho. Infrastructure sharing for mobile network operators: Analysis of trade-offs and market. *IEEE Transactions on Mobile Computing*, 17(12):2804–2817, Dec 2018.

[77] V. Sciancalepore, F. Cirillo, and X. Costa-Perez. Slice as a service (SlaaS) optimal IoT slice resources orchestration. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–7, Dec 2017.

[78] 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 2017 - IEEE Conference on Computer Communications*, pages 1–9, May 2017.

[79] A. Seetharaman, N. Niranjan, V. Tandon, S. Devarajan, M. K. Moorthy, and A. S. Saravanan. What do customers crave in mobile 5G?: A survey spotlights four standout factors. *IEEE Consumer Electronics Magazine*, 6(3):52–66, July 2017.

[80] H. Setiawan, M. F. Rian Dinni, and D. A. Ratna Wati. LTE network planning based on existing base transceiver using a genetic algorithm. In *2016 2nd International Conference on Wireless and Telematics (ICWT)*, pages 106–110, Aug 2016.

[81] S. Siami-Namini, N. Tavakoli, and A. S. Namin. A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1394–1401. IEEE, 2018.

[82] A. Singh, X. Li, I. Abeywickrama, A. Könsgen, C. Görg, P. N. Tran, and A. Timm-Giel. QoE-based access network dimensioning. In *2014 16th International Telecommunications Network Strategy and Planning Symposium (Networks)*, pages 1–6, Sep. 2014.

[83] R. Skopal. Short-term hourly price forward curve prediction using neural network and hybrid ARIMA-NN model. *2015 International Conference on Information and Digital Technologies*, pages 335–338, 2015.

[84] Y. K. Song, H. Zo, and S. Lee. Analyzing the economic effect of mobile network sharing in Korea. *ETRI Journal*, 34(3):308–318, 2012.

[85] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Kténas, N. Cassiau, and C. Dehos. 6G: the next frontier. *arXiv preprint arXiv:1901.03239*, 2019.

[86] V. Suryaprakash, A. Fehske, A. F. dos Santos, and G. P. Fettweis. On the impact of sleep modes and BW variation on the energy consumption of radio access networks. In *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)*, pages 1–5, May 2012.

[87] V. Suryaprakash and G. P. Fettweis. An analysis of backhaul costs of radio access networks using stochastic geometry. *2014 IEEE International Conference on Communications (ICC)*, pages 1035–1041, 2014.

[88] T. Taleb, B. Mada, M. Corici, A. Nakao, and H. Flinck. PERMIT: Network slicing for personalized 5G mobile telecommunications. *IEEE Communications Magazine*, 55(5):88–93, May 2017.

[89] X. Ting, P. Zhiwen, L. Nan, and Y. Xiaohu. Inter-operator resource sharing based on network virtualization. In *International conference on Wireless Communication Signal Processing (WCSP)*, pages 1–6, Oct 2015.

[90] H. Tullberg, P. Popovski, Z. Li, M. A. Uusitalo, A. Hoglund, O. Bulakci, M. Fallgren, and J. F. Monserrat. The metis 5G system concept: Meeting the 5G requirements. *IEEE Communications Magazine*, 54(12):132–139, December 2016.

[91] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso, and C. Verikoukis. Cooperation incentives for multi-operator C-RAN energy efficient sharing. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2017.

[92] P. L. Vo, M. N. H. Nguyen, T. A. Le, and N. H. Tran. Slicing the edge: Resource allocation for RAN network slicing. *IEEE Wireless Communications Letters*, 7(6):970–973, Dec 2018.

[93] H. Wang, K. Wang, and Y. Zhao, H.and Yue. Prediction of user behavior in smart home based on improved ARIMA model. In *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 298–302. IEEE, 2018.

[94] Y. Wang, B. Gu, S.Liu, P. Liu, and X. Zhong. Stackelberg game modeling of pricing for mobile virtual network operators. *EAI Endorsed Transactions on Future Intelligent Educational Environments*, 1(4), 8 2015.

[95] Y. Xiao and M. Krunz. Dynamic network slicing for scalable fog computing systems with energy harvesting. *IEEE Journal on Selected Areas in Communications*, 36(12):2640–2654, Dec 2018.

[96] C. Yang, J. Li, M. Guizani, A. Anpalagan, and M. Elkashlan. Advanced spectrum sharing in 5G cognitive heterogeneous networks. *IEEE Wireless Communications*, 23(2):94–101, 2016.

[97] S. Yrjölä, P. Ahokangas, and M. Matinmikko. Evaluation of recent spectrum sharing concepts from business model scalability point of view. In *2015 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pages 241–250, Sep. 2015.

[98] J. Zausinova, J. Gazda, and T. Maksymyuk. Real-time spectrum secondary markets: Agent-based model of investment activities of heterogeneous operators. In *2018 28th International Conference Radioelektronika (RADIOELEKTRONIKA)*, pages 1–6, April 2018.

[99] D. Zhang, Z. Chang, T. Hamalainen, and W. Gao. A contract-based resource allocation mechanism in wireless virtualized network. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 474–479, April 2018.

[100] D. Zhang, Z. Chang, T. Hamalainen, and F. R. Yu. Double auction based multi-flow transmission in software-defined and virtualized wireless networks. *IEEE Transactions on Wireless Communications*, 16(12):8390–8404, Dec 2017.

[101] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung. Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges. *IEEE Communications Magazine*, 55(8):138–145, Aug 2017.

[102] K. Zhu, Z. Cheng, B. Chen, and R. Wang. Wireless virtualization as a hierarchical combinatorial auction: An illustrative example. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, March 2017.