Universitat Politècnica de Catalunya

Doctoral Thesis

Network Virtualization in Next Generation Cellular Networks

Author:
Georgia Tseliou

Supervisors:
Dr. Ferran Adelantado
Dr. Christos Verikoukis
Tutor:
Dr. Ramon Ferrús

Signal Theory and Communications Department

May 2019

Abstract

The complexity of operation and management of emerging cellular networks significantly increases, as they evolve to correspond to increasing QoS needs, data rates and diversity of offered services. Thus critical challenges appear regarding their performance. At the same time, network sustainability pushes toward the utilization of sharing Radio Access Network (RAN) infrastructure between Mobile Network Operators (MNOs). This requires advanced network management techniques which have to be developed based on characteristics of these networks and traffic demands. Therefore it is necessary to provide solutions enabling the creation of logically isolated network partitions over shared physical network infrastructure. Multiple heterogeneous virtual networks should simultaneously coexist and support resource aggregation so as to appear as a single resource to serve different traffic types on demand.

Hence in this thesis, we study RAN virtualization and slicing solutions destined to tackle these challenges. In the first part, we present our approach to map virtual network elements onto radio resources of the substrate physical network, in a dense multi-tier LTE-A scenario owned by a MNO. We propose a virtualization solution at BS level, where baseband modules of distributed BSs, interconnected via logical point-to-point X2 interface, cooperate to reallocate radio resources on a traffic need basis. Our proposal enhances system performance by achieving 53% throughput gain compared with benchmark schemes without substantial signaling overhead. In the second part of the thesis, we concentrate on facilitating resource provisioning between multiple Virtual MNOs (MVNOs), by integrating the capacity broker in the 3GPP network management architecture with minimum set of enhancements. A MNO owns the network and provides RAN access on demand to several MVNOs. Furthermore we propose an algorithm for on-demand resource allocation considering two types of traffic. Our proposal achieves 50% more admitted requests without Service Level Agreement (SLA) violation compared with benchmark schemes. In the third part, we devise and study a solution for BS agnostic network slicing leveraging BS virtualization in a multi-tenant scenario. This scenario is composed of different traffic types (e.g., tight latency requirements and high data rate demands) along with BSs characterized by different access and transport capabilities (i.e., Remote Radio Heads, RRHs, Small Cells, SCs and future 5G NodeBs, gNBs with various functional splits having ideal and non-ideal transport network). Our solution achieves 67% average spectrum usage gain and 16.6% Baseband Unit processing load reduction compared with baseline scenarios. Finally, we conclude the thesis by providing insightful research challenges for future works.

Resumen

La complejidad de la operación y la gestion de las emergentes redes celulares aumenta a medida que evolucionan para hacer frente a las crecientes necesidades de calidad de servicio (QoS), las tasas de datos y la diversidad de los servicios ofrecidos. De esta forma aparecen desafíos críticos con respecto a su rendimiento. Al mismo tiempo, la sostenibilidad de la red empuja hacia la utilización de la infraestructura de red de acceso radio (RAN) compartida entre operadores de redes móviles (MNO). Esto requiere técnicas avanzadas de gestión de redes que deben desarrollarse en función de las características especiales de estas redes y las demandas de tráfico. Por lo tanto, es necesario proporcionar soluciones que permitan la creación de particiones de red aisladas lógicamente sobre la infraestructura de red física compartida.

Para ello, en esta tesis, estudiamos soluciones de virtualización de la RAN destinadas a abordar estos desafíos. En la primera parte, nos centramos en mapear elementos de red virtual en recursos de radio de la red física, en un escenario LTE-A que es propiedad de un solo MNO. Proponemos una solución de virtualización a nivel de estación base (BS), donde los módulos de banda base de BSs distribuidas, interconectadas a través de la interfaz lógica X2, cooperan para reasignar los recursos radio en función de las necesidades de tráfico. Nuestra propuesta mejora el rendimiento del sistema al obtener un rendimiento 53% en comparación con esquemas de referencia. En la segunda parte nos concentramos en facilitar el aprovisionamiento de recursos entre muchos operadores de redes virtuales móviles (MVNO), al integrar el *capacity broker* en la arquitectura de administración de red 3GPP. En este escenario, un MNO es el propietario de la red y proporciona acceso bajo demanda (en inglés *on-demand*) a varios MVNOs. Además proponemos un algoritmo para la asignación de recursos bajo demanda, considerando dos tipos de tráfico. Nuestra propuesta alcanza 50 % más de solicitudes admitidas sin violación del Acuerdo de Nivel de Servicio (SLA) en comparación con otros esquemas. En la tercera parte, estudiamos una solución para el *slicing* de red independiente del tipo de BS, considerando la virtualización de BS en un escenario de múltiples MVNOs (*multi-tenants)*). Este escenario se compone de diferentes tipos de tráfico junto con BSs caracterizadas por diferentes capacidades de acceso y transporte (por ejemplo, Remote Radio Heads, RRHs, Small cells, SC y 5G NodeBs, gNBs con varias divisiones funcionales que tienen una red de transporte ideal y no ideal). Nuestra solución logra una ganancia promedio de uso de espectro de 67% y una reducción de la carga de procesamiento de la banda base de 16.6 % en comparación con escenarios de referencia. Finalmente concluimos la tesis al proporcionando los retos de investigación para trabajos futuros.

Acknowledgments

I would like to sincerely thank the people who have helped, inspired and motivated me during the course of my study.

My research would have been impossible without the aid and support of my supervisors Dr. Ferran Adelantado i Freixer and Dr. Christos Verikoukis. Heartfelt thanks for teaching me how to think as a researcher and always be curious on how to solve applied problems in the field of wireless telecommunications. The guidance, support and motivation throughout my PhD studies has been valuable.

Furthermore I would like to thank my friend Maria for sharing many good and challenging moments during the years of our PhD studies.

Last but not the least, I would like to thank my family: my parents, Maria and Kostas, and my sister, Katerina, for always supporting me in any possible way and believing in my capabilities. They are the most important people in my world and I dedicate this thesis to them.

Contents

List of Figures

List of Tables

Abbreviations

2G	Second Generation access network
3G	Third Generation access network
3GPP	Third-Generation Partnership Project
5G	Fifth Generation Networks
AAS	Antenna Array System
AI	Artificial Intelligence
AMC	Adaptive Modulation and Coding
AWGN	Additive White Gaussian Noise
AxC	Antenna Carrier
BBU	Base Band Unit
BE	Best-Effort
BH	BackHaul
BS	Base Station
C-RAN	Centralized RAN
CA	Carrier Aggregation
CAC	Call Admission Control
CAPEX	Capital Expenditure
CB	Coordinated Beamforming
CBR	Constant Bit Rate
CD	Confidence Degree
CDF	Cumulative Distribution Function
CDMA	Code division multiple access
CN	Core Network
CO	Central Office
CoMP	Coordinated Multi-Point

CPRI	Common Public Radio Interface
CS	Coordinated Scheduling
CSI	Channel Status Information
CSIR	Channel State Information at the Receiver
CSIT	Channel State Information at the Transmitter
CU	Centralized / Central Unit
D-RAN	Distributed RAN
DeNB	Donor eNB
DU	Distributed Unit
E-UTRAN	Evolved Universal Terrestrial Radio Access
eICIC	Enhanced ICIC
eICIC	enhanced Inter-Cell Interference Coordination
eMBB	enhanced Mobile Broadband
eNB	E-UTRAN macro NodeB
EP	Elementary Procedure
EPC	Evolved Packet Core
EvDO	Evolution-Data Only or Evolution-Data Optimized
FFT	Fast Fourier Transform
FH	FrontHaul
FTP	File Transfer Protocol
gNB	5G New Radio (NR) NodeB
GPP	General-Purpose Processor
GWCN	GateWay Core Network
HeNB	GW Home eNB Gateway
HeNB	Home Evolved Universal Terrestrial Radio Access (E-UTRA) NodeB
HetNet	Heterogeneous network
HPLMN	Home Public Land Mobile Network
IA	Interference Alignment
ICIC	Inter-Cell Interference Coordination
IE	Information Element
IFFT	Inverse Fast Fourier Transform
IPsec	Internet Protocol Security
IQ	In-phase and Quadrature

ISD	Inter-site Distance
ISP	Internet Service Provider
ISV	Independent Software Vendor
Itf-B	Type 1 interface
Itf-N	Type 2 interface
ITU	International Telecommunications Union
KDE	Kernel Density Estimation Technique
LTE-A	Long-Term Evolution Advanced
MAC	Medium Access Control
MCS	Modulation and Coding Scheme
MIN-RATE-S	Minimum Rate Slicing
MME	Mobility Management Entity
mMTC	massive Machine Type Communications
MO-NM	Master Operator-Network Manager
MO-SR-DM	Master Operator-Shared RAN-Domain Manager
MOCN	Multi-Operator Core Network
MORAN	Multi-Operator RAN
MuSli	Multi-tenant Slicing
MVNO	Mobile Virtual Network Operator
NetSliC	BS agnostic framework for Network SliCing
NFV	Network Functions Virtualization
NGC	Next Generation Core
NGFI	Next Generation FH Interface
NVS	Network Virtualization Substrate
OFDM	Orthogonal Frequency Division
OPEX	Operational Expenditure
OTT	Over-the-Top Provider
PDCP	Packet Data Convergence Protocol
PDF	Probability Distribution Function
PDSCH	Physical Downlink Shared Channel
PER	Probability Error Rate
PGW	Packet Delivery Network Gateway
PHY	Physical

PLMN	Public Land Mobile Network
PON	Passive Optical Network
PoP	Point of Presence
PRB	Physical Resource Block
PRR	Partial Resource Reservation
QoE	Quality of Experience
RAC	Radio Admission Control
RAN	Radio Access Network
RB	Resource Block
RBC	Radio Bearer Control
RENEV	Resources nEgotiation for NEtwork Virtualization
RF	Radio Frequency
RLC	Radio Link Control
RMSE	Root Mean Square Error
RN	Relay Node
RO	Regional Office
RRC	Radio Resource Control
RRH	Radio Remote Radio Head
RRM	Radio Resource Management
SC	Small Cell
SCTP	Stream Control Transmission Protocol
SDN	Software Defined Network
SDR	Software Defined Radio
SGW	Serving Gateway
SIB	Subscriber Information Base
SINR-S	SINR-based Slicing
SISO	Single Input Single Output
SLA	Service Level Agreement
SLN	Service with Leased Network
SNR	Signal-to-Noise-Ratio
SO-NM	Sharing Operator-Network Manager
SS	Spectrum Sharing
SSL	Secure Sockets Layer VPN

TAC	Tracking Area Code
TCP	Transmission Control Protocol
TTI	Transmission Time Interval
UE	User Equipment
UMi	Urban Micro
URLLC	Ultra-Reliable Low Latency Communications
V-BS	Virtualized BS
V-CRAN	Virtualized cloud radio access network
VPLMN	Visited Public Land Mobile Network
VPON	Virtualized Passive Optical Network
WiMAX	Worldwide Interoperability for Microwave Access
xDSL	x Digital Subscriber Line
ZF	Zero Forcing beamforming

To my sister Katerina and
my parents Maria and Kostas

Chapter 1

Introduction

1.1 Motivation

Cellular communications are evolving to facilitate the current and expected increasing needs of Quality of Service (QoS), high data rates, and diversity of offered services. Mobile networks accommodate not only conventional terminals (i.e., feature phones and smartphones), but also a number of terminals assumed to be embedded in devices are emerging, which in turn create a variety of service demands. As a result, mobile traffic patterns present high temporal and spatial variations as shown in Fig. 1.1. In addition several Mobile Network Operators (MNOs) are expected to operate and manage mobile Radio Access Network (RAN) and Core Network (CN) infrastructure. Thus, new challenges are generated by diversification of terminal requirements, traffic patterns and the plurality of MNOs. Moreover as users are typically shifting between different areas (e.g., residential and office) during the day, peak transmission bandwidth requirements may be as much as 10 times higher than during off-peak hours. The traditional network provisioning approach of considering only busy hours is no longer acceptable. It leads to an inefficient resource usage with high Capital (CAPEX) and Operational Expenditures (OPEX). Thus, operators have to find practical, flexible, and cost-efficient solutions for their networks taking into account the scarce amount of radio resources. Furthermore the appearance of Mobile Virtual Network Operators (MVNOs), which aim to provide specific services without owning infrastructure, is seen as a definitive trend that will modify the mobile infrastructure ownership landscape [1].

In addition Fifth Generation (5G) Networks will become a unified service platform to serve services covering enhanced Mobile Broadband (eMBB), massive Machine Type Communications (mMTC) and Ultra-Reliable Low Latency Communications (URLLC) [2–4] as defined by ITU-R. Certainly new network architecture and technologies are

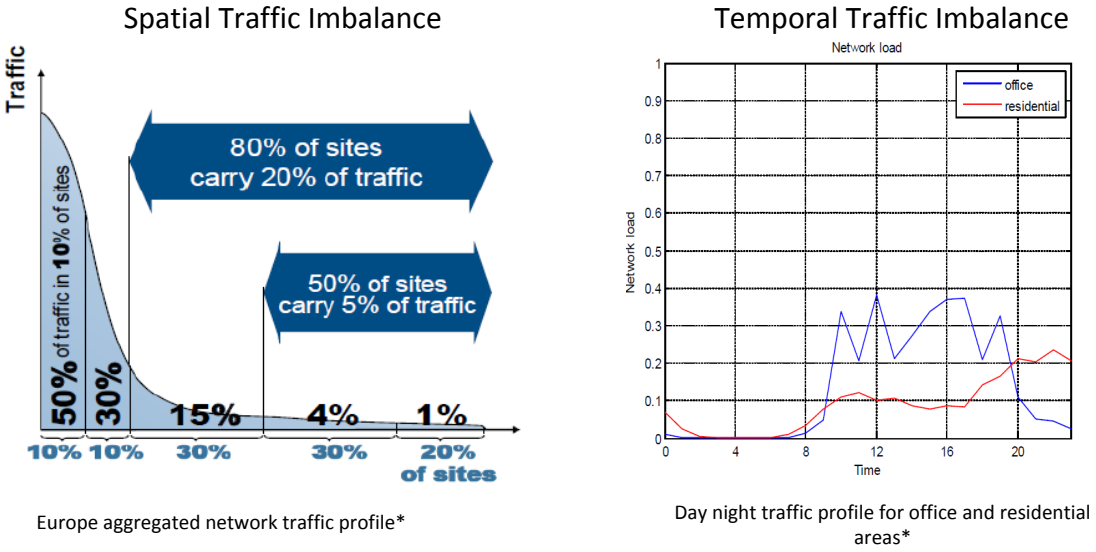


FIGURE 1.1: Traffic Imbalance in spatial and temporal dimensions [1].

required to fuel such a wide spectrum of services. The increasing complexity in 5G RANs [5], which constitute an emerging paradigm, and their coexistence with legacy infrastructure calls for a new network design. Beyond the requirements described by the three ITU categories, the pace and success of the 5G technology roll-out will rely on the ability of the network to accommodate service needs of vertical industries. Good examples of that are Virtual Reality / Augmented Reality (VR/AR) or 3D video as eMBB, smart cities or logistic applications as mMTC, and industrial automation, remote surgery or Vehicle-to-everything (V2X) as URLLC.

In that sense, and aiming to meet such a diverse set of QoS requirements, flexibility and adaptability are the main premises on which 5G has been designed. With an architecture based on Software Defined Networking (SDN), 5G decouples control plane from forwarding hardware, thereby improving the network programmability, enabling the creation of tailored virtual networks on top of the physical network and facilitating the upgrade and introduction of new services, as proposed in [5, 6]. Such programmability and flexibility are key enablers for the Radio Access Network (RAN) slicing, a new concept that allows the differentiated traffic handling depending on the service type, by dynamically allocating software and hardware resources in a customized manner.

However, the traffic diversity is not the unique source of complexity 5G has to face. Unlike the previous LTE standard, 5G New Radio (NR) has been standardized as a non-standalone network at a first stage, and as a standalone network at a subsequent stage.

Indeed 5G introduces the concept of network slicing, which is based on virtualization and softwarization. In [7] a full set of 5G RAN architecture options are described and discussed, ranging from the complete standalone 5G RAN option -NR node (gNB) directly connected to the Next Generation Core (NGC)- to the non-standalone option -gNB connected to the Evolved Packet Core (EPC) through an Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Node B (eNB).

This range of architecture options means that coexistence and inter-working with previous standards are design objectives of the 5G RAN to allow a gradual migration from 2G/3G/4G to 5G, thereby reducing time to market and minimizing initial MNOs' investment.Thus, initial 5G deployments will be multi-standard networks, resulting in a composite of different 3GPP -and eventually non-3GPP- network architectures jointly operated.

The scarcity of available spectrum to fuel the operation of 5G has also become a key element in rolling out the new standard [8, 9]. On the one hand, MNOs aim to achieve full coverage, which makes imperative the allocation of low bands -below 1 GHz- and mid bands -below 6 GHz. On the other hand, eMBB or URLLC call for the deployment of a *capacity tier*, which requires the allocation of high spectrum bands -e.g. 24-28 GHz, 37-40 GHz or 64-71 GHz. Hence, MNOs will have to develop a spectrum strategy to define the transition from the current allocation of spectrum for 2G/3G/4G towards a new spectrum allocation.

Therefore, the management of 5G networks face complexity in three dimensions:

- Traffic dimension: Vertical industries will define a wide range of services with extremely differentiated QoS requirements.
- Technology dimension: MNOs will manage multi-standard networks.
- Spectrum dimension: 5G will operate on a wide range of spectrum bands.

RAN virtualization and network slicing are the key solutions destined to tackle the challenges that the distinct dimensions create. Generally as a concept, network virtualization refers to the architecture and related enabling solutions allowing the deployment of multiple virtual networks on top of a common infrastructure. In essence, it aims at decoupling and isolating virtual networks from the underlying physical substrate infrastructure. However the general definition assumes different nuances as a function of each specific context of the current application. The main objective of network virtualization is to reduce the total OPEX and CAPEX by sharing network resources while still maintaining isolation among them. Regarding network slicing, physical resources that

are pooled together are sliced when and as needed per traffic type, to deliver a service, leading to higher resource utilization and further expenditure reduction. The concept of network slicing has been proposed to address the diversified service requirements in 5G under the background of the aforementioned technologies [10]. From the RAN perspective, the virtualization and network slicing facilitate the orchestration of radio resources as well as the efficient management and sharing of spectrum among different tenants.

Within the context of network infrastructure sharing and multi-tenancy, network virtualization is conceived as an enabler allowing different virtual radio networks (i.e., MVNOs) to operate on a common shared infrastructure. Key element in network virtualization research is the provision of resources for co-existing MNOs to provide isolation between them. Isolation can be provided by a fixed division of the resources but this can be highly inefficient, since efficient utilization is sacrificed [11]. The management of the increase of co-channel interference due to cell densification and the sharing mode of the different wireless resources are basic challenges that must be confronted.

Network virtualization also enables completely new value chains. Small players can come into the market and provide new services to their customers using a virtual network. This also allows completely new future networks, e.g., isolating one virtual network (like a banking network) from a best effort Internet access network [12]. Hence, the design of solutions both in terms of frameworks and architecture that implement RAN virtualization is imperative.

The combination of these challenges lead to the motivation of the current thesis that addresses the need of creating RAN virtualization solutions at the Base Station (BS)[1] level and network slicing solutions for efficient management of varying radio resources shared among operators and guaranteeing agreed Service Level Agreements (SLAs) between distinct types of services. The design of future architectures favors multiple coexisting solutions, designed and customized to satisfy specific network requirements, rather than trying to achieve a global architecture that fits all.

To that end, in this thesis, we attempt to fill the gap in literature regarding the combined study of RAN virtualization solutions in multi-tenant scenarios in terms of distinct services and operators. First we design a scalable virtualization solution at the BS level for a two tier deployment with a macro cell overlaid with numerous small cells within a Long-Term Evolution Advanced (LTE-A) architecture. The requirements imposed by this scenario are the geographical variations of introduced traffic and cell densification. These issues create a number of challenges to be faced when applying resource sharing and isolation among interconnected BSs. Then we propose a capacity broker framework

[1]In this thesis with the general term BS we refer to any RAN node offering service capability to a user, such as macro BS, small cell, RRH, gNB, unless stated otherwise.

and architecture for a scenario where a MNO owns the RAN infrastructure and many MVNOs act as resellers of their host network's capacity under their own brands, to their own customers. Finally we devise and test a framework that creates wireless slices independently of the type of BS (i.e., BS agnostic) and on the same time achieves desirable service capability across the various traffic profiles. This framework is applied in an architecture combining both legacy and future RAN nodes that leverage the concept of functional splits. Our holistic vision for RAN virtualization is depicted in Fig. 1.2. In particular we consider that RAN resources and nodes can be shared in isolation among participating tenants (i.e., either these are BSs, MNOs or traffic types) based on their particular service demands.

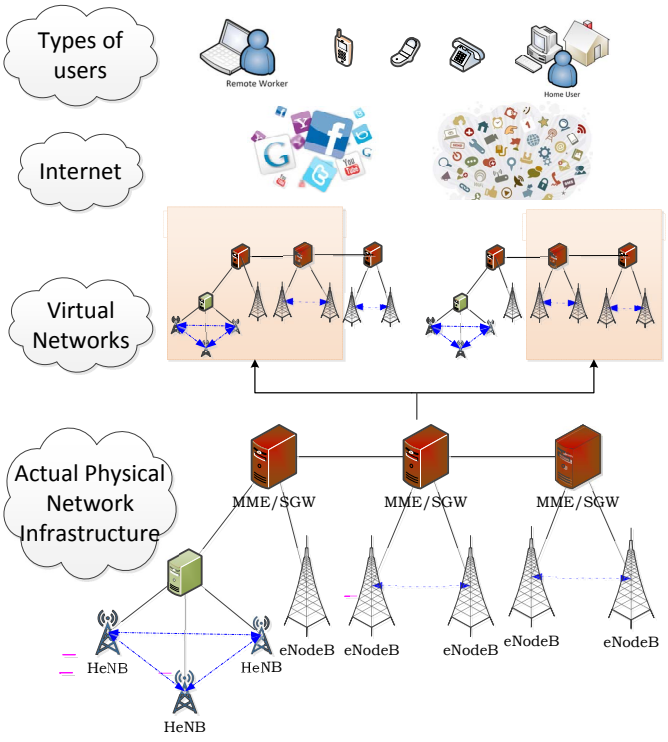


FIGURE 1.2: Vision of this thesis for RAN virtualization.

The structure of the thesis and the main contributions of this work will be discussed in detail in the following section.

1.2 Structure of the Thesis and Contributions

The remaining part of the thesis consists of six chapters.

Chapter 2 provides some necessary background information on the reference legacy and future architectures that we use within our study. Then we define RAN virtualization at the BS level by discussing about functional splits and management of virtualized

wireless resources. Furthermore we present use cases and business requirements about RAN sharing. In addition we present network slicing principles along with future 5G slicing and the basics of FrontHaul (FH) / BackHaul (BH) integration. Finally we discuss about several tools and architectures in the literature to implement network virtualization.

The innovative contributions of the thesis are organized into three chapters: i) in chapter 3 we study radio resource management principles and propose Resources nEgotiation for NEtwork Virtualization (RENEV), a solution that achieves slicing and on-demand delivery of resources at the BS level, ii) chapter 4 wherein we propose the Multi-tenant Slicing (MuSli) of capacity algorithm, to allocate resources towards MVNOs in coarse time scales and iii) in chapter 5 we propose the BS agnostic framework for Network SliCing (NetSliC) to be adopted by the MNOs for creating a virtualization layer in a scenario wherein future and current cellular RAN nodes co-exist. In the following, the main contributions of the thesis will be outlined in more detail.

In our first approach to create a mapping of virtual network elements onto radio resources of the existing physical network, we propose the RENEV algorithm, which is suitable for application in heterogeneous networks in LTE-A environments, consisting of a macro evolved Node B overlaid with Small Cells (SCs). This work is described in chapter 3. By exploiting radio resource management principles, RENEV achieves slicing and on-demand delivery of resources at BS level. Leveraging the multi-tenancy approach, radio resources are transferred in terms of physical radio resource blocks among multiple heterogeneous BSs (i.e., macro cell and SCs), which are interconnected via the X2 interface. The main target is to deal with traffic variations in geographical dimension. All signaling design considerations under the Third-Generation Partnership Project (3GPP) LTE-A architecture are also investigated. Analytical studies and simulation experiments are conducted to evaluate RENEV in terms of network's throughput and additional signaling overhead. Moreover we show that RENEV can be applied independently on top of already proposed schemes for RAN virtualization to improve their performance. The results indicate that significant advantages are achieved both from network's and users' perspective and that it is a scalable solution for different numbers of SCs.

In chapter 4 we study network slicing in the scenario where a MNO owns the RAN and many MVNOs act as resellers of their host network's capacity under their own brands, to their own customers. Resource allocation in multi-operator scenarios requires an estimate of the tenants' traffic needs (i.e., MVNOs with different service requirements). In such scenarios, the forecasted MVNO traffic is the basis for providing resources suitable with the corresponding MVNOs' demand. To that end, the dynamic provision of resources among MVNOs should be performed in flexible, short-term time scales. In

chapter 4, we effectively address this issue by integrating the capacity broker entity into the 3GPP network management architecture using the minimum set of enhancements. In addition, to fully exploit its capabilities, we propose the Multi-tenant Slicing (MuSli) of capacity algorithm, to allocate resources towards MVNOs in coarse time scales. MuSli considers the estimated capacity and the impact of the traffic type (i.e., guaranteed QoS and Best-Effort) in each MVNO, to provide better utilization of the host network's capacity. Our results highlight the gains in the number of served requests without compromising their service quality.

In chapter 5 we propose the BS Agnostic Framework for Network SliCing, NetSliC, that creates network slices for the RAN based on the distinct service requirements. In this chapter we create a scenario wherein the RAN is a heterogeneous complex architecture, comprising legacy distributed SCs, Radio Remote Radio Heads (RRHs) connected to a centralized Base Band Unit (BBU), and future 5G BSs (gNBs) leveraging virtualization with different functional splits, serving distinct traffic profiles. Thus, NetSliC is a common language that manages and controls this heterogeneous infrastructure. We consider several criteria for creating network slices in this particularly interesting future scenario that are applicable to all different types of access and transport interfaces of the distinct BSs. This chapter concludes the vision of this thesis to propose a virtualization layer running on top of the substrate network; our proposal, NetSliC, constitutes a framework that creates wireless slices independently of the type of BS (i.e., BS agnostic) and on the same time achieves desirable service capability across the various traffic profiles.

Finally, chapter 6 discusses the conclusions of the presented work and identifies potential lines for future investigation.

1.3 Research Contributions

All the work presented in this thesis, has been published in three journals and two international conferences. The list of publications follows:

[*J*3] A Base Station Agnostic Network Slicing Framework for 5G, **G. Tseliou**, F. Adelantado and C. Verikoukis, *IEEE Network Magazine*, Accepted as a paper on 22 March 2019, to appear.

[*J*2] NetSliC: Base Station Agnostic Framework for Network SliCing, **G. Tseliou**, F. Adelantado and C. Verikoukis, *IEEE Transactions on Vehicular Technology*, Vol. 68, no. 4, pp. 3820-3832, Apr. 2019 (Published online: 28 February 2019).
URL: https://ieeexplore.ieee.org/abstract/document/8654697

[*J*1] Scalable RAN Virtualization in Multi-Tenant LTE-A Heterogeneous Networks, **G. Tseliou**, F. Adelantado and C. Verikoukis, *IEEE Transactions on Vehicular Technology*, Vol. 65, Issue: 8, Aug. 2016 (Published online: 1 September 2015).
URL: https://ieeexplore.ieee.org/abstract/document/7234933

[*C*2] A Capacity Broker Architecture and Framework for Multi-tenant support in LTE-A Networks, **G. Tseliou**, K. Samdanis, F. Adelantado, X. Costa Pérez and C. Verikoukis, In: *IEEE International Conference on Communications (IEEE ICC 2016), Kuala-Lumpur, Malaysia*, May 23-27 2016.
URL: https://ieeexplore.ieee.org/abstract/document/7511042

[*C*1] Resources Negotiation for Network Virtualization in LTE-A Networks, **G. Tseliou**, F. Adelantado and C. Verikoukis, In: *IEEE International Conference on Communications (IEEE ICC 2014), Sydney, Australia*, June 10-14, 2014.
URL: https://ieeexplore.ieee.org/abstract/document/6883804

1.4 Other Research Contributions

Although the aforementioned research contributions compose the main body of the thesis, there is a publication and other activities that took place during the course of this PhD. As these works are not fully aligned with this thesis, they were not included therein. However, we believe that the following publication should be mentioned in this section:

[*C*3] An SDN QoE-service for dynamically enhancing the performance of OTT applications, E. Liotou, **G. Tseliou**, K. Samdanis, D. Tsolkas, F. Adelantado and C. Verikoukis In: *Seventh International Workshop on Quality of Multimedia Experience (QoMEX), Costa Navarino, Greece*, May 26-29 2015.
URL: https://ieeexplore.ieee.org/abstract/document/7148106

Chapter 2

Background

2.1 Introduction

The main objective of this thesis is to design and evaluate solutions that leverage the concept of RAN virtualization and network slicing, in current and future cellular architectures wherein service types pose distinct requirements. To that end, in this chapter, we will provide the background behind our proposals that will facilitate the understanding of the contributions of this thesis.

Hence, Section 2.2 discusses the reference architectures of Distributed RAN (D-RAN), Centralized RAN (C-RAN) as well as Hybrid architectures combining both distributed and centralized characteristics. Section 2.3 presents the concept of RAN virtualization both at Base Station (BS) and radio resource level. Section 2.4 shows different types of sharing configurations as well as use cases and business requirements whereas 2.5 discusses the role of network slicing in the RAN domain. Finally, section 2.6 presents different tools for implementing RAN virtualization.

2.2 Reference Architectures

2.2.1 Distributed RAN: LTE-A Cellular Networks

Third Generation Partnership (3GPP) defines the standards for LTE and 5G. In particular LTE-Advanced includes all work from 3GPP Release 10 till 3GPP Release 14 [13–15] and it is the leading architecture that meets International Telecommunications Union's (ITU) IMT-Advanced requirements. Consequently, it is of critical importance to

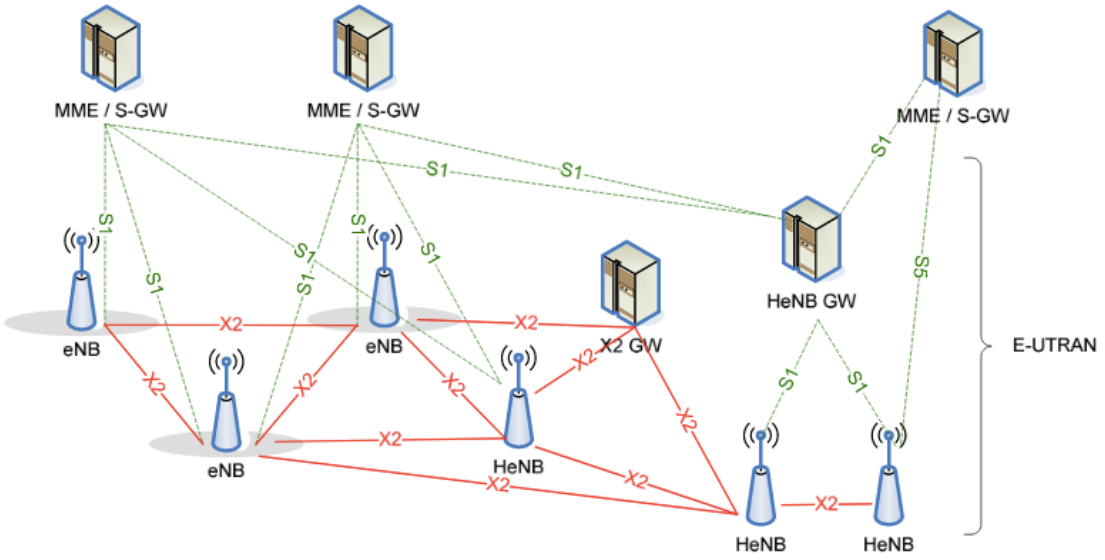


FIGURE 2.1: Overall E-UTRAN Architecture with deployed SCs (i.e., HeNBs) [14-16].

investigate how to incorporate network virtualization in LTE-A access nodes and radio resources, which constitutes our work presented in chapter 3.

Fig. 2.1 presents 3GPP Release 13 ([14–16]). We study this architecture in chapter 3 and we further refer to as legacy / traditional D-RAN architecture. Fig. 2.1 defines a heterogeneous environment consisting of E-UTRAN macro NodeBs (eNBs) and Small Cells (SCs) such as Pico eNBs, Relay Nodes (RNs) and Home eNBs (HeNBs). Evolved Packet Core (EPC) and the Radio Access Network (RAN), called Evolved Universal Terrestrial Radio Access (E-UTRAN) are shown in this figure. 3GPP Release 13 as described in [16] and as shown in Fig. 2.2 does not offer the flexibility of providing RAN sharing on demand to support diverse and bandwidth hungry services. Except from this, signaling is also caused due to abrupt mobility of the involved devices.

The SC densification leads to frequent handovers within the participant tiers. Thus the signaling load is increased due to user mobility and the perceived service quality is decreased due to the degradation of application throughput. The efficient management of a heterogeneous environment, where operators are competing at the service layer, creates the need of separate virtual networks on the same physical infrastructure. Network virtualization lends itself to spectrum and cost savings, efficiency and flexibility. In addition to that, resource virtualization avoids over-provisioning by assigning resources intelligently and elastically per operator and service type based on the actual need. Therefore there is imminent need to implement it in the network. Below we describe the main elements of the architecture to identify where our proposed solutions are applied. It is pointed out that this thesis studies RAN virtualization. For the sake of completeness

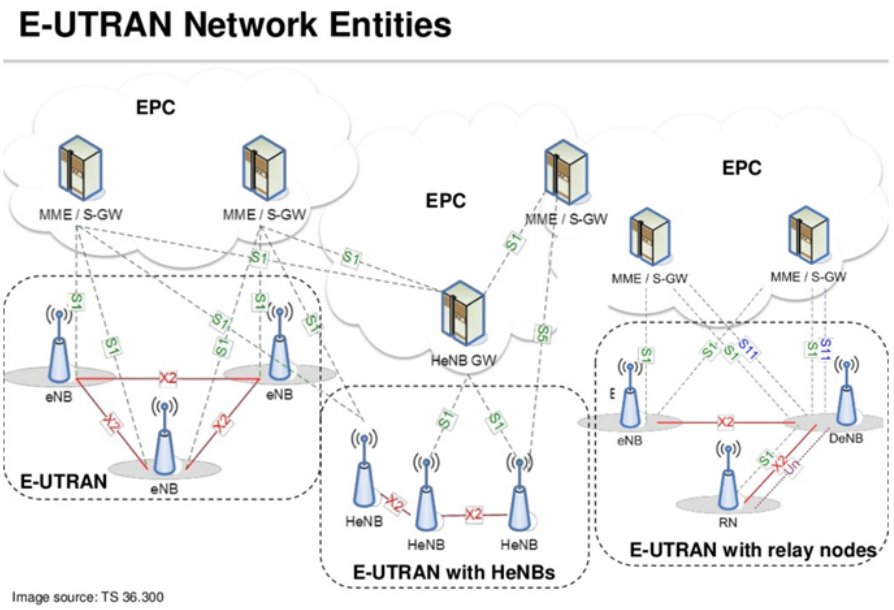


FIGURE 2.2: E-UTRAN Network Entities of Release 13 3GPP [14-16].

we provide a holistic view of the architectural elements in this chapter but more detail is provided for the RAN as shown in Fig. 2.2.

To start off with, the Core Network (CN) consists of the following nodes: Mobility Management Entity (MME), Service Gateway (SGW) and Packet Delivery Network Gateway (PGW). Generally, MME is responsible for the control plane of the LTE-A architecture and the key control-node for LTE-A RAN. In more detail, it is responsible for the idle mode User Equipment (UE) tracking, the paging procedure when retransmissions are required as well as activation and deactivation of the bearer process. Moreover, within its responsibilities lies the user authentication and the provision of control plane function for mobility between LTE and Second / Third Generation (2G/3G) access networks.

Next, SGW node routes and forwards user data packets by being the mobility anchor for the user plane during inter-eNB handovers. It also acts as the mobility anchor for the user plane during inter-eNB handovers and as the anchor for mobility between LTE and other 3GPP technologies. Finally it performs replication of user traffic in case of a potential lawful interception. The last node of EPC, PGW, provides connectivity from the UE to external packet data networks by being the point of exit and entry of traffic for the UE. It also performs policy enforcement, packet filtering for each user, charging support, lawful interception and packet screening. It is the anchor for mobility between 3GPP and non-3GPP technologies such as Worldwide Interoperability for Microwave Access (WiMAX) standard and 3GPP2 (Code division multiple access (CDMA) and Evolution-Data Only or Evolution-Data Optimized (EvDO)).

Furthermore Fig. 2.2 presents the different E-UTRAN (or for the sake of simplicity RAN) network entities included in the release of LTE-A [14] under study. E-UTRAN consists of eNBs, HeNBs and RNs. These nodes have been introduced in LTE-A for efficient heterogeneous network planning. The RNs are low power eNBs that provide enhanced coverage and capacity at cell edges. One of the main benefits of relaying is to provide extended LTE coverage in targeted areas at low cost. The RN is connected to the Donor eNB (DeNB) via radio interface, Un, a modified version of E-UTRAN air interface Uu. DeNB also serves its own UE as usual, in addition to sharing its radio resources for RNs [16]. eNBs and HeNBs are interconnected with each other by means of the X2 interface. They are also connected by means of the S1 interface to the EPC, more specifically to the MME by means of the S1-MME interface and to the Serving Gateway (SGW) by means of the S1-U interface. S1 interface supports many-to-many relations between MMEs / SGWs and eNBs / HeNBs. It is very interesting to note that in real deployments all these nodes are located in different geographical areas. For instance it is very common to group several eNBs of each region to a same Central Office (CO). Every CO is connected to the Regional Office (RO), where network elements such as the core nodes (i.e. MMEs / SGWs) are located.

The functions supported by the SCs (i.e., HeNBs) are the same as those supported by the macro BS (i.e., eNB). The same holds for the procedures run between a SC and the corresponding EPC. In any case both types of BSs [1] (i.e., macro BS or SC) comply to the same set of physical standards but their configuration and parametrization may be different due to the different transmission power. The different coverage scale of each BS has an impact on their role in resources negotiation. The main difference of macro and SC in this case lies in the different transmission bandwidth that each one accommodates, since the one allocated to the SC is quite restricted.

LTE-A is an Orthogonal Frequency Division (OFDM) based system where the system bandwidth is available for a BS. However, not all subcarriers are used simultaneously in a specific set of cells, i.e., according to Inter-Cell Interference Coordination (ICIC) techniques each subcarrier is not usually allocated to more than one BS simultaneously. Actually, ICIC techniques are aimed to reduce the interference level, particularly in cell edge. ICIC actually tries to reuse resources only if interference is low enough (i.e., the interfering source is at a minimum distance), but resources are nevertheless reused. In LTE-A, Enhanced ICIC (eICIC) is further defined, which is an adjusted version of ICIC for HetNet, and Coordinated Multi-Point (CoMP) which uses Channel Status Information (CSI) reported by UE. The basic scenarios where this thesis will focus on, are deployments consisting of numerous SCs with and without the inclusion of an

[1] Throughout the rest of the thesis, the exact definition of the term BS is given for all the particular scenarios when required.

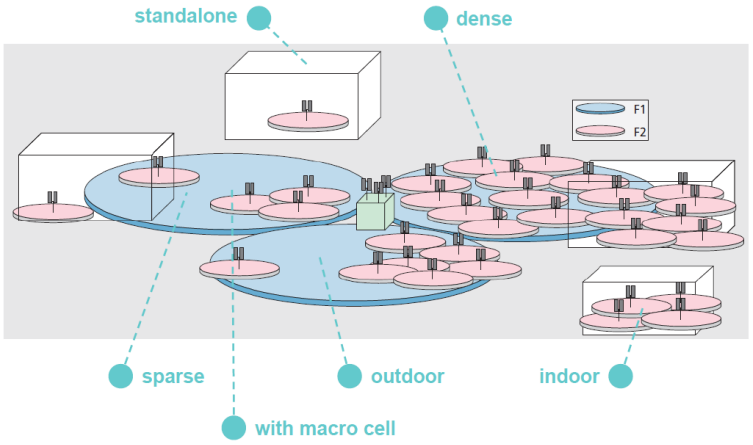


FIGURE 2.3: General deployment of two tier scenario with macro eNB and SCs in different frequencies [17, 18].

eNB. Figure 2.3 gives an overview of a Heterogeneous Two Tier deployment with the coexistence of macro BS and different categories of SCs [17, 18].

2.2.2 Centralized Radio Access Network (C-RAN) Architecture

Traditional D-RAN supports the dense deployment of standalone SCs to increase area spectral efficiency. However, it has been quickly observed that non-ideal BackHaul (BH), such as X2 interface, limits the coordination among SCs. Incremental advancements on traditional D-RAN architecture are not able to satisfy high QoS requirements [19].

In order to cope with the increasing capacity demand and new service requirements, the Cloud / Centralized (C-RAN) paradigm has been introduced to increase the degree of cooperation between cells [20–22]. In this model RAN functions are split between the Base Band Unit (BBU), hosted in the cloud, and Remote Radio Heads (RRHs) / units (RRUs) that provide antenna equipment and radio access, as presented in Fig. 2.4.

The main difference of this architecture in comparison to the LTE-A architecture, shown in chapter 2.2.1, is that it leverages principles of cloud computing. Its goal is to move functionality, especially BBU processing, from the antenna site to a central location [23]. At this central location, which can be several kilometers away from the antenna site, the baseband system modules of several antenna sites are pooled. Thus, the antenna consists only of a relatively simple Radio Frequency (RF) unit which is connected to the central location via optical fibre. Cloud concept and its potential use in wireless networks is also investigated in [24].

The main concept of C-RAN is de-constructing traditional BS to leave a low power unit at the cell site, integrating antenna and radio, while centralizing all the BBU activity and

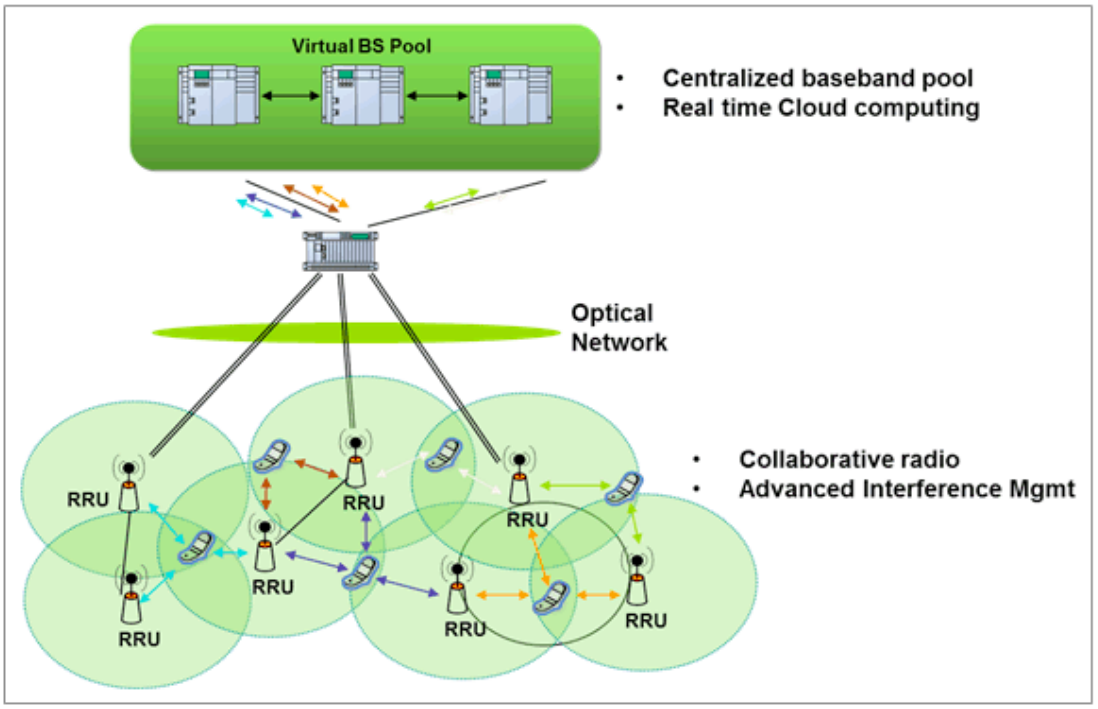


FIGURE 2.4: Centralized RAN architecture by NGMN [20-22].

supporting a plurality of sites. In [20] the authors present a system consisting of a BBU which is a digital unit that implements the Medium Access Control (MAC), Physical (PHY) and Antenna Array System (AAS) functionality and the RRH that contains the base station's RF circuitry plus analog-to-digital or digital-to-analog converters and up-/down converters. C-RAN brings to RAN the advantages of the cloud: resource-sharing, elasticity, on-demand and pay-as-you-go. Based on real-time virtualization technology, C-RAN minimizes CAPEX and OPEX costs, by aggregating multiple BBUs per central office. It enables the fast, flexible and optimized deployment and upgrade of RANs, supporting pay-per-use models. It also eases the flexible and on-demand adaptation of resources to non-uniform traffic. Besides this, the centralized processing of a large cluster of RRUs also enables the efficient operation of inter-cell interference reduction and CoMP transmission and reception mechanisms, and eases mobility between RRHs / RRUs.

However, the centralization of BBU into a common shared BBU pool poses strict capacity requirements to the FrontHaul (FH) connection (i.e., interface between RRH / RRU and BBU). Furthermore the availability of high speed fiber links (i.e., ideal FH), especially in urban deployments, is controversial due to high implementation cost. This is due to the stringent requirements on the FH, which can only be met in practice by costly fiber point to point links.

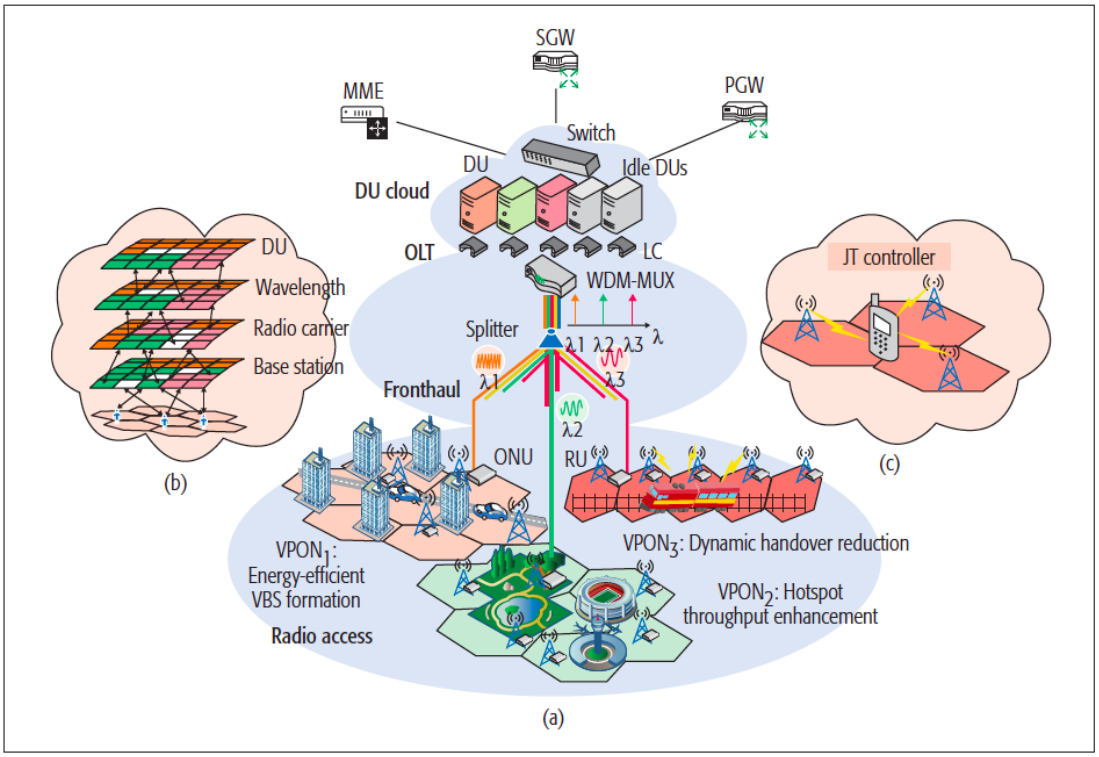


FIGURE 2.5: a) Virtualized-CRAN architecture; b) illustration of joint optimization of resources. c) illustration of joint transmission [19].

2.2.3 Hybrid D-RAN and C-RAN Future Architectures

Both D-RAN and C-RAN architectures face challenges with regard to the implementation of virtualization and joint control / management of transport, computing, and radio resources. First, network resources must be virtualized and provisioned dynamically, so virtualization techniques used in the IT industry must be tailored to satisfy time-sensitive wireless tasks. Second, there is a trade-off between virtualization gain and implementation complexity, for example, whether to allocate resources on a per-user or per-cell basis.

To that end several intermediate architectures have emerged and proposed for 5G (i.e., all the 3GPP specifications from Release 15 onwards). For instance the authors in [19] present a new 5G architecture, called virtualized cloud radio access network (V-CRAN), moving toward a cell-less network architecture as depicted in Fig. 2.5. They leverage the concept of a virtualized BS (V-BS) that can be optimally formed by exploiting several enabling technologies such as software defined radio (SDR) and CoMP transmission /reception. A V-BS can be formed on a per-cell basis or per-user basis by allocating virtualized resources on demand. For the FH solution, the authors exploit the Passive Optical Network (PON), where a wavelength can be dynamically assigned and shared to form a Virtualized Passive Optical Network (VPON).

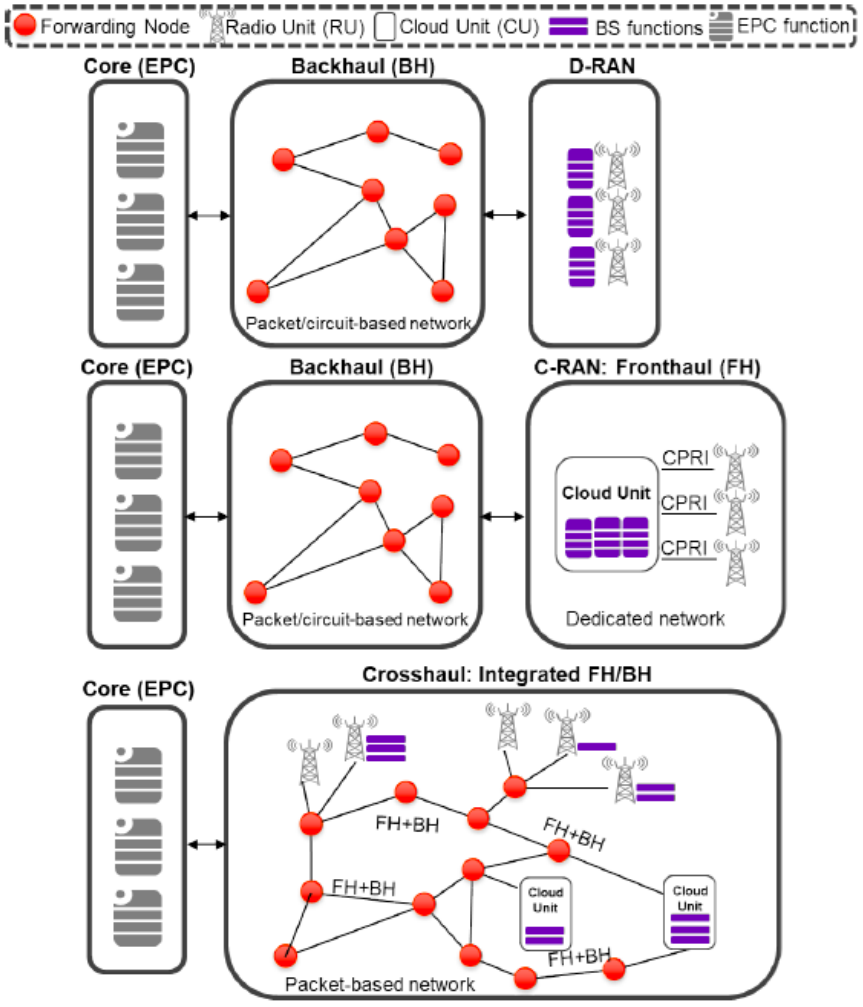


FIGURE 2.6: Traditional Distributed RAN (D-RAN) vs Cloud RAN (C-RAN) vs Next-Generation Fronthaul (Crosshaul) [25].

Another architecture that achieves flexible centralization is introduced in [25] and depicted in Fig. 2.6. In this architecture FH and BH coexist in a common packet-based network, creating a new interface called Next Generation FH Interface (NGFI). Each access node, i.e., BS, adopts flexible splits of RAN functionality. More detailed information about the functional splits can be found in section 2.3.1. The idea is to divide a classic BS into a set of functions that can either be processed at the Distributed Unit (DU) or offloaded into a Centralized Unit (CU), depending on the transport requirements and centralization needs. In this way we can better balance cost / performance (i.e., the more aggressive the offloading, the higher the gains) and requirements (i.e., the softer the offloading, the more relaxed the network constraints). In [25] the authors propose optimization mechanisms that maximize the degree of centralization while meeting the transport constraints of the BSs. The proposed frameworks jointly perform routing and select dynamically the optimal functional split for each BS.

2.3 Virtualization of Radio Access Network

Network virtualization can be applied in several parts of the network. Also in the case of cellular networks two options for its application arise: CN and RAN. This thesis is going to focus on the RAN part that makes possible the easy creation and management of virtual networks, opening up a range of new business models. RAN offers a wide field for potential solutions based on network virtualization. The partitioning and/or pooling of radio physical resources can be introduced by enabling more efficient management of RAN capacity. In this context, the imposed challenges require the development of new concepts such as the design of RRM algorithms that take into advantage the dense heterogeneous architecture of legacy D-RAN and future C-RAN and Hybrid cellular deployments as well as different traffic characterization and geographical distribution. Moreover, virtualization of radio spectrum and BS resources makes possible the creation and management of virtual networks on demand, opening up a range of new business models through which network owners can increase the revenue from their networks. The following two sections define two categories for interpreting RAN virtualization in cellular networks: section 2.3.1 introduces the flexible BS virtualization and the notion of functional splits and section 2.3.2 discusses about the management of virtual resources.

2.3.1 Flexible BS Virtualization and Functional Splits

RAN virtualization is based on the notion of BS softwarization, which allows certain RAN functions to run at remote cloud platforms [7]. The FH link is the foundation enabler for the virtualized use cases to be viable. While it is accepted that dedicated fiber links permit extremely low latency and high bandwidth connections, it is considered cost prohibitive for the volumes of SCs projected and is therefore a barrier to scale. A summary of the split architectures for the use cases is introduced in [26] and depicted in Fig 2.7. Functions to the left of the split are virtualized in the CU, while functions to the right reside in the remote DU. The use cases are presented from left to right as gradually more of the SC functionality is virtualized. For the MAC and PHY use cases exist where the functionality is divided between the CU and DUs, to represent this they are divided into upper and lower components, in these split use cases the upper portion resides in the CU and the lower part in the remote DU.

The most common functional splits are detailed further:

- PHY-layer option provides the highest centralization and can be realized only with an ideal FH, i.e., a high data rate and low latency optical fiber.

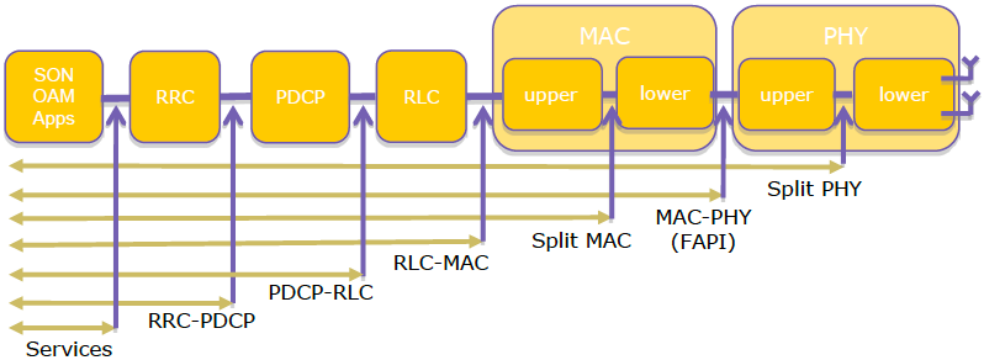


FIGURE 2.7: A high-level overview of the different functional split options [26].

- MAC-layer option where the MAC layer and the layers above it are virtualized and run on a BBU with real time scheduling performed aggregately for multiple RRHs. This option leverages the benefits of connecting distributed RRH physical layers to a common MAC, which allows coordinated scheduling and dynamic point selection, i.e., CoMP. However, this option requires a low latency FH as some of the MAC procedures are time critical (e.g., UE scheduling) and need to generate a configuration at the Transmission Time Interval (TTI) level.

- RLC-layer option where the RLC layer and other layers above it are virtualized at the BBU allowing multiple MAC entities to be associated with a common RLC entity. This option reduces the FH latency constraints as real time scheduling is performed locally in the RRH.

- PDCP-layer option is non-time critical. It runs the PDCP functions at the BBU and may use any type of FH network. The main advantage of this option is the possibility to have an aggregation of different RRH technologies (e.g., 5G, LTE, and WiFi).

2.3.2 Managing Virtualized Wireless Resources

The term wireless resource virtualization refers to the variety of ways that resources are treated within the RAN. The state of the art solutions that we present in this chapter focus on creating a substrate within the BS for monitoring wireless resources of a cellular network delivered to the BS itself. Virtualizing wireless resources in cellular networks fosters several interesting deployment scenarios that are of interest [27, 28]:

- Active RAN sharing: Sharing of RAN resources that enables significant reduction in equipment in low traffic areas and results in at-least 100% increased rollout

speed with a given cost. Further details on RAN sharing can be found in section 2.4.

- MVNO: In the recent past, several MVNOs have emerged as strong players in the cellular market providing enhanced services. Such MVNOs often do not own spectrum and rely on sharing the wireless resources of a MNO in that region. Virtualization encourages the partition of resources in a network infrastructure owned by a third party (i.e., MNO) effectively, thereby encouraging stricter and fine-grained SLAs between MVNOs and MNOs.

- Corporate Bundle Plans: Currently, MNOs offer data plans to enterprises and corporations that allow sharing of bandwidth dynamically across their users. However, no bandwidth guarantees are provided. Virtualization may help realize better guarantees on resource allocation, and hence fosters more sophisticated data plans.

- Controlled evaluation: Virtualization enables MNOs to isolate partial wireless resources to deploy and test novel ideas without affecting the operational networks. Currently, MNOs often use dedicated / small scale deployments to test new ideas.

- Services with Leased Networks (SLNs): With the increased use of wireless and mobile networks for Internet services, we envision application service providers reserving bandwidth with MNOs and paying on the behalf of their users to enhance Quality of Experience (QoE). Virtualization helps in ensuring that such reservations are met.

Another study on managing the virtualized wireless resources has been published in the context of the FLAVIA project [29]. The key concept of the project's work is to expose flexible programmable interfaces enabling service customization and performance optimization through software-based exploitation of low-level operations and control primitives. The separation of the control and data planes in LTE networks is one of the main principles where the project's ideas were based on. In one of the project's dissemination results, the authors propose a generic MAC architecture for both LTE and WiMAX wireless networks [30]. In this paper, both technologies are being analyzed with the goal of finding common functional subsets which can be used as building blocks for a generic and extensible MAC for future mobile cellular networks. To this end, the authors propose a systematic categorization into services, interfaces, functions and primitives as a first step towards achieving generic architecture. In this architecture, it is very interesting to notice, a potential solution of virtualization. The upper part of Layer 2 is considered suitable for hosting the functions that would be responsible for virtualization of the resources delivered to the eNB.

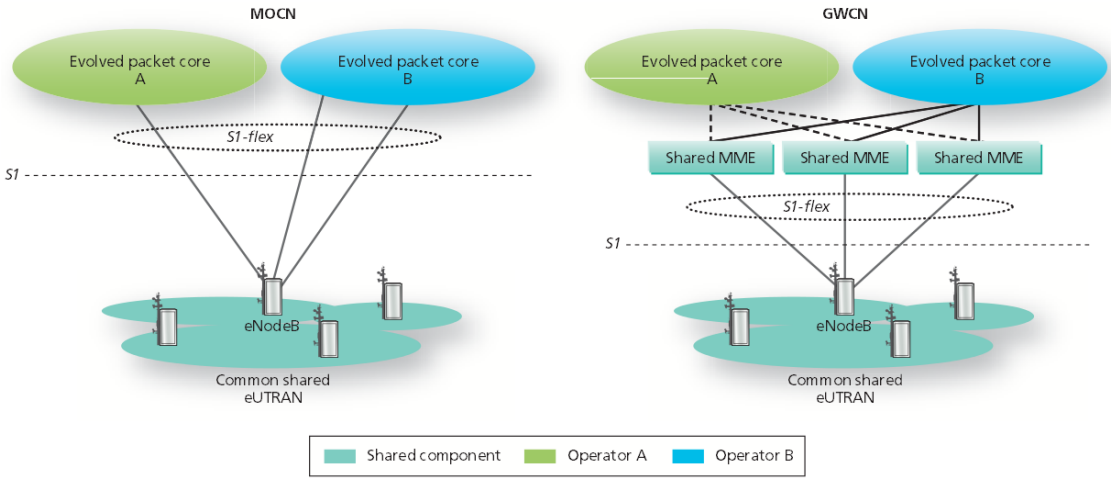


FIGURE 2.8: RAN sharing configurations supported by 3GPP [27, 28].

2.4 Radio Access Network Sharing

2.4.1 RAN Sharing Configurations

Cellular network sharing among operators, is a key building block for virtualizing future mobile carrier networks. 3GPP has recognized the importance of supporting network sharing among operators by defining a set of architectural elements [31] and technical specifications [32]. Indeed, 3GPP has defined and ratified different kinds of architecture with varying degrees of sharing:

- Multi-Operator RAN (MORAN): only equipment is shared;
- Multi-Operator Core Network (MOCN): both spectrum and equipment are shared; and
- GateWay Core Network (GWCN), in which both the RAN and some elements of the CN are shared.

The following two architectural network sharing configurations are of interest in the context of this thesis: GWCN and MOCN. In GWCN configuration, CN operators share control nodes in addition to RAN elements whereas in MOCN, multiple control nodes owned by different operators are connected to a shared RAN. In Fig. 2.8 we present the two general configurations for network sharing as identified by 3GPP [33].

The user behavior in both configurations is the same. No information concerning the configuration is indicated to the UE. If the RAN is shared by multiple operators, the system information broadcasted in each shared cell contains the Public Land Mobile Network (PLMN)-id of each operator (i.e., up to 6 in legacy D-RAN scenarios) and a single

Tracking Area Code (TAC) valid within all the PLMNs sharing the RAN resources [33]. The infrastructure owner provides the underlying physical network whereas by referring to network operator, we denote every operator having its users connected to the RAN (without necessarily owning infrastructure). In both configurations the network sharing agreement between operators is transparent to the end users. Although, operators may share network elements (i.e., RAN / control nodes), radio resource virtualization is required to cover their actual requirements, in isolation per BS. Therefore, in both MOCN and GWCN sharing configurations, virtualization of resources is necessary in order to allow users to have access to the complete set of available resources. Existing network virtualization techniques, can be grouped into solutions for the Evolved Packet Core (EPC) Network and the RAN [34].

2.4.2 Spectrum Sharing

Spectrum sharing is a key technique in RAN virtualization; it can be used at air interface to adapt to traffic load variations of different virtual networks. Many spectrum sharing proposals are designed to adapt the radio interface of the eNB to traffic load variations of distinct virtual networks [27, 35–39]. This is achieved by allowing multiple virtual networks to share the spectrum allocated to a particular physical eNB. A preliminary approach for virtualizing a BS in LTE is described in [35]. A controlling entity called hypervisor was proposed in order to make use of a-priori knowledge (e.g., user channel conditions, operator sharing contracts, traffic load etc.) to schedule the Resource Blocks (RBs) of a BS among different mobile operators. In addition, the authors of [36] evaluate several sharing options, ranging from simple approaches feasible in traditional infrastructure to complex methods requiring a specialized one.

In advancing the basic BS virtualization, works [27, 37] and [38] introduce the concept of Network Virtualization Substrate (NVS) that operates closely to the MAC scheduler. NVS adopts a two-step scheduling process, one managed by the infrastructure provider for controlling the resource allocation towards each virtual instance of an eNB and the second controlled by each virtual instance itself providing scheduling customization within the allocated resources. Additionally, [39] extends NVS solution by investigating the provision of active LTE RAN sharing with Partial Resource Reservation (PRR). In this scheme, each slice is guaranteed a specific minimum share of radio resources to be available to the operator that owns them. The remaining common part is shared among traffic flows belonging to different operators.

The authors in [40] propose a Markovian approach to characterize the resource sharing in multi tenant scenarios with diverse guaranteed bit rate services by considering a slice

aware admission control policy. After describing the Markov model and its implementation and discussing its suitability, the model is applied to study the performance attained in a scenario with two different slices, one for enhanced mobile broadband communications and the other for mission critical services. In general Markovian approaches are well suitable to characterize resource sharing and they have been used in the literature.

2.4.3 Use-cases and Business Requirements

In this section we describe the general on roles in RAN Sharing as defined in [41]. The arrangements for network sharing between the involved entities can vary widely, being influenced by a number of factors including business, technical, network deployment and regulatory conditions. Despite the variety of factors, there is a set of common roles centered on connecting network facilities between the parties involved in a network sharing agreement.

- *Hosting RAN Provider*: It is identified as the owner RAN, which is shared with one or more Participating Operators. The Hosting RAN Provider or in other words the MNO has primary operational access to particular licensed spectrum which is part of the network sharing arrangement. This does not necessarily imply that the MNO owns licensed spectrum but has agreement to operate in that spectrum. Furthermore, the MNO owns a set of RAN nodes in a specific geographic region covered under the network sharing arrangement and provides facilities allowing Participating Operators (or MVNOs) to share the RAN covered under the network sharing arrangement.

- *Participating Operator*: It is identified as the entity that uses shared RAN facilities provided by a Hosting RAN Provider, possibly alongside other Participating Operators (MVNOs). The characteristics of the Participating Operator include the use of a portion of particular shared licensed spectrum to provide communication services under its own control to its own subscribers and the use of a portion of shared RAN in the specific geographic region covered under the network sharing arrangement.

- *Roaming operators*: This category consists of Home Public Land Mobile Network (HPLMN) and Visited Public Land Mobile Network (VPLMN) operators. Roaming agreements between operators, provide similar capability to RAN sharing where a HPLMN subscriber can obtain services while roaming into a VPLMN. This can be viewed as a form of sharing where VPLMN shares the use of its RAN with the HPLMN for each HPLMN subscriber roaming into the VPLMN. The distinction between roaming and RAN sharing is that when roaming, the subscriber

uses the VPLMN when outside of the HPLMN geographic coverage and within the VPLMN geographic coverage. In a RAN sharing arrangement, all the participants provide the same geographic coverage through the Hosting RAN.

- *Operators with multiple roles*: Operators can take on multiple roles at the same time depending on business needs. For the purposes of [41], each specific network set (i.e., spectrum-region-RAN) can be considered independently and combined with other network sets in various combinations. Indicative examples include:
  - An operator has its own spectrum, which does not share and additionally uses the shared RAN in the same region (Participating Operator) provided by Hosting RAN Provider.
  - Two operators set up a joint venture to build and operate a shared network. The two operators are both Participating Operators and the joint venture is a Hosting RAN Provider.
  - Two operators A and B, divide a region covered by a joint spectrum license and each build and operate the RAN in their portion of the region. In the region covered by operator A's RAN, operator A is the Hosting RAN Provider and at the same time Participating operator while operator B is only Participating Operator. In the region covered by operator B's RAN, operators A and B are the Participating Operators and operator B is the Hosting RAN Provider.

According to [41] a Hosting RAN Provider may share RAN resources with Participating Operators in various ways. Therefore, it is assumed that at least a set of radio resources in addition to physical BSs are shared for use by Participating Operators.

2.5 Radio Access Network Slicing

Network slicing in a mobile network is highly related to network sharing, particularly to RAN sharing in the case of mobile networks. It is important to point out that network slicing enables different network architectures for different service needs. In this chapter we focus in particular, on solutions for dynamic resource slicing and the idea of slicing resources between FrontHaul (FH) and BackHaul (BH).

2.5.1 Dynamic Resource Slicing

Dynamic resource slicing is another category of solutions based on the concept of RAN virtualization. The authors of [28] have proposed CellSlice; a dynamic framework to

achieve active RAN sharing by remotely controlling the scheduling decisions, ensuring that each entity receives its share of the wireless resources. This idea does not require the modification of the BS schedulers but it controls the BS scheduling decisions from a remote gateway. Slicing can be done with either a BS-level solution or a gateway-level solution. Compared to BS-level solution, remotely slicing wireless resources makes the proposal easily deployable, enables easier network-wide resource reservations for slices, and guarantees the operation with BS from multiple vendors, some or all of which may not support native virtualization. The work focuses on remote uplink slicing, since wireless resource reservation requests from the clients for uplink transmissions terminate at the BS and they are not visible at the gateways.

With regard to the dynamic resources' slicing, further interesting proposals are presented in [42–44]. In [42–44], the authors present software defined cellular network architectures, allowing remote gateway level controller applications to perform resource slicing without modifying the BSs' MAC schedulers. Such solutions express real-time, fine-grained policies based on subscribers attributes rather than network addresses and locations.

There are different slice resource management models depending on the level of resource isolation, which may handle frequency spectrum as a dedicated medium per slice or shared resource among specific slices [45]. In the dedicated resource model, a RAN slice consists of isolated resources in terms of the control and user plane traffic, MAC scheduler and spectrum. Each slice has access to its own Radio Resource Control, Packet Data Convergence Protocol, Radio Link Control, Medium Access Control (RRC / PDCP / RLC / MAC) instances and a percentage of dedicated Physical Resource Blocks (PRBs) or a subset of channels. Although the dedicated resource model ensures committed resources per slice, i.e., assuring delay and capacity constraints, it reduces resource elasticity and limits the multiplexing gain. Indeed, the dedicated resource model restricts the slice owner to modify the amount of resources (i.e., PRB) committed to a slice during its life-cycle, even if they are not utilized. On the other hand, the shared resource model allows slices to share the control plane, MAC scheduler and spectrum.

In addition network slicing can be associated with service chain embedding problem for diversified 5G requirements, considering the sharing property of VNFs. Leveraging on network function virtualization (NFV), the network operator performs service chain embedding (SCE) to create the logical slices.For instance the authors in [46] present a fine-grained network slicing model for resource and QoS requirements of slices and their traffic flows. They propose an optimization formulation that yields an embedding solution, routing path and resource allocation for each slice.

Another aspect in network slicing solutions lies on how to determine the relationship among traffic demand, amount of resources and end to end delay. In [47] the authors design a two step algorithm that can be used by tenants or service providers to determine the minimal amount of resources allocated to each VNF along a service chain so that the specified end to end delay requirements can be met. They further design a fast search algorithm for tenants to decide whether to adjust the running network slice when traffic demand changes in order to guarantee its QoS. Additionally they propose an auto resizing method for tenants to adjust the slice size in response to traffic change.

2.5.2 5G Slicing and FrontHaul / BackHaul Integration

The emerging 5G networks introduce a heterogeneous FH / BH landscape that consists of various technologies such as optical, millimeter-wave, Ethernet, and IP [26, 48]. Currently, network virtualization in the mobile backhaul relies on dedicated and overlay networks over a shared infrastructure, converging distinct transport network services into a unified infrastructure [49, 50].

The stringent 5G RAN requirements, in terms of device and load density, and high mobility, are expected to shape the transport network layer facilitating enhanced capacity, high availability and an agile control. For the FH / BH this means multi-path connectivity, tighter synchronization, coordination of both radio and transport layers and software defined control. In principle, different BS functional split can offer a particular service performance, requiring a distinct capacity and delay from the transport network layer. An integrated FH / BH architecture such as the one presented in Fig. 2.6, i.e. offering FH / BH services on common links, can assure the desired performance by allowing a different centralization of the control and data planes for each service, while optimizing the network resource efficiency. Network slicing can assure isolation and performance guarantees between the different logical networks that employ a different FH / BH split according to the corresponding BS functional split. Such an integrated FH / BH architecture is based on a unified control plane and on a data plane that relies on network nodes capable of integrating different transport technologies for FH and BH via a common data frame [51].

The authors in [52] design a BH infrastructure virtualization market in which the virtual operator can use the BH nodes of each service provider. They further assume that the renewable energy supplier produces renewable energy and sells it to one of the on grid nodes of each BH. The results show that green BH virtualization can provide significant gains in terms of both network deployment and energy cost saving, making it attractive for the virtual operator to use green virtualized BH links. The results also show that

the proposed decentralized market scheme achieves performance comparable to that of the centralized optimal solution.

The flexible functional split we presented in 2.3.1, can highly impact the performance of network slicing and the optimal split largely depends on the characteristics of the target service. For example, a low latency slice (e.g., ultra reliable low latency - uRLLC service type) may require most RAN functions to run on DU in order to fulfill latency requirements, while in a service slice with high data requirements (e.g., enhanced Mobile BroadBand - eMBB), a higher centralization can enhance the throughput by aggregating RRHs (e.g., enabling CoMP). In the context of network slicing, certain RAN functions can be also shared among different slices as elaborated in [53]. For example, each network slice may have its own instance of RRC (configured and tailored user plane protocol stack), PDCP, and RLC (non-real time functions), while the low RLC (real-time function), MAC scheduling (inter-slice scheduler) and physical layer can be shared. Some network slices may also have their own intra-slice application scheduler or tailor the RLC and PDCP functions to the specific slice type [26]. For example, in a network slice supporting low latency, the header compression may not be used and RLC transparent mode may be configured, while a service requiring high QoS/QoE may activate an acknowledged RLC mode [45].

2.6 Tools for Network Virtualization

2.6.1 Software Defined Networks (SDN)

Software-Defined Networking (SDN) is one of the promising technologies which is expected to solve existing limitations in current networks. SDN provides the required improvements in flexibility, scalability and performance to adapt mobile network to keep up with the expected growth. SDN in mobile cellular networks is directing the current mobile network towards a flow-centric model employing inexpensive hardware and a logically centralized controller. It enables the separation of data forwarding plane from the control plane.

In this paradigm, operators have the flexibility to develop their own networking concepts, optimize the network and address specific needs of the subscribers. The acquisition of virtualization into mobile networks brings economical advantage in two ways. First, SDN requires inexpensive hardware such as commodity servers and switches instead of expensive mobile BH gateway devices. Second, the introduction of SDN to mobile networks allows entering new actors in the mobile network ecosystem such as Independent Software Vendor (ISV), cloud providers, Internet Service Providers (ISP) and

that will change the business model of mobile networks. The provided network services are abstracted from the underlying infrastructure and network behavior is directly programmable [54, 55].

Currently, the most popular specification implementing SDN is OpenFlow. OpenFlow is an open standard that lets network administrators remotely control routing tables [56]. In [42] the authors present a SDN enabled cellular network architecture, called CellSDN, allowing controller applications to express policies based on the attributes of subscribers, rather than network addresses and locations, enables real-time, fine-grained control via a local agent on each switch, and extends switches to support features like deep packet inspection and header compression to meet the needs of cellular data services. It is heavily inspired and follows the high-level vision of OpenRoads (or OpenFlow Wireless [57]) which is a platform for innovation and realistic deployment of services for wireless networks. In fact, [57] is the first SDN wireless network. It is mainly based on WiFi and offers no special support for cellular networks. In contrast, CellSDN [42], addresses specific cellular network requirements such as real-time session management which runs on top of Stream Control Transmission Protocol (SCTP) instead of Transmission Control Protocol (TCP) for paging, UE state tracking, policy enforcement, charging and RRM. Finally in [43] that completes the previous work, the same authors add to the proposed system an entity called CellSDN controller. Its design has as target to separate traffic management from the low-level mechanisms for installing rules and minimizing data-plane state. The traffic management layer determines the service attributes for a UE from the Subscriber Information Base (SIB), and consults the service policy to compute policy paths that traverse the appropriate middle-boxes and optimize traffic management objectives.

2.6.2 Network Functions Virtualization (NFV)

Network Functions Virtualization (NFV) is transforming how network operators architect networks by enabling the consolidation of network services onto industry standard servers. These services can be located in Data Centers, on Network Nodes or at the user premises. NFV involves delivering network functions as software that can run as virtualized instances, being deployed at locations in the network as required, without the need to install equipment for each new service. It is worth pointing out that functional splits as defined in section 2.3.1 constitute a particular definition of NFV at the BS level of cellular networks. NFV is applicable to any network function in both mobile and fixed networks. SDN, forms a concept related to NFV, but they refer to different domains. SDN is focused on the separation of the network control layer from its forwarding layer, while NFV is focused on porting network functions to virtual environments to enable the

migration from proprietary appliance based embodiments to a standard hardware and cloud based infrastructure. Both concepts can be complementary, although they can exist independently. Virtual appliances might be configured using SDN capabilities and they might be connected via overlay network tunnels in clusters based on an application or based on the needs of an organization [58].

After its definition, it is very interesting to see the problem that target to solve. As Service Providers are faced with increased competition from Over-the-Top (OTT) Providers they are seeking new markets to enter. However, as they look to create and launch new services they must grapple with the growing number and complexity of hardware devices in their networks. This creates challenges due to the time it takes to certify equipment and with staffing and training of skilled operators for many devices. It also creates cost pressure with the need for more space and power at a time when these resources are becoming ever more expensive. Making upfront capital outlays for equipment in anticipation of revenue that ramps up over time can stress budgets. As a result Service Providers are looking to change how network services are deployed and some are finding NFV as the answer to their problems.

There is almost no limit to the network functions that can be virtualized. Service Providers are already making use of virtual switching to connect physical ports to virtual ports on virtual servers and using virtual routers and virtualized Internet Protocol Security (IPsec) and Secure Sockets Layer VPN (SSL) gateways to terminate customer traffic cloud data centers. There is a desire to use virtualized network appliances at customer premises as well and functions contained in home or small office routers and set top boxes can be implemented to create virtualized home and small office appliances. These services presently require multiple dedicated hardware appliances on customer premises to deliver services such as fire-walling, web security, IPS/IDS, WAN acceleration and optimization, as well as routing functions. There are many other network services that could be virtualized such as traffic analysis tools and network monitoring tools, as well as application optimization services such as load balancers and application accelerators [59].

Chapter 3

Scalable RAN Virtualization in Multi-Tenant LTE-A Networks

3.1 Introduction

Our first approach toward RAN virtualization is aimed to shed light on the limitations of LTE-A architecture in 3GPP Release 13, as shown in chapter 2.2.1, in a dense multi-tier network deployment. Previous research has motivated us a lot to work on dense cellular deployments since these heterogeneous and densely deployed scenarios are designed to meet the envisaged traffic demands [60].

Considering that operating infrastructure is a significant cost for operators, the densification of access networks and the necessity to reduce the costs will lead to cooperation between them and to the sharing of resources, including infrastructure sharing itself. In this context, the provision of solutions enabling the creation of logically isolated network partitions over shared physical network infrastructure should allow multiple heterogeneous virtual networks to coexist simultaneously and support resource aggregation. This concept defines the principle of network virtualization [61] and explains why Radio Access Network (RAN) virtualization emerges as a key aspect of the future cellular Long Term Evolution-Advanced (LTE-A) networks.

Today's cellular networks have relatively limited support for virtualization. Thus, although Third Generation Partnership Project (3GPP) standardizes necessary functionalities to enable several Core Network (CN) operators to share one RAN [32], neither a detailed implementation of radio resource customization among them nor mechanisms to exploit the network heterogeneity of the dense multi-tier architectures, defined in the

latest release of LTE-A, are provided [62]. Therefore, the particular definition of algorithms implementing RAN virtualization for radio resources in a multi-operator sharing architecture still remains an open issue. In this point we define *RAN virtualization* according to [63], as the way *"in which physical radio resources can be abstracted and sliced into virtual cellular network resources holding certain corresponding functionalities, and shared by multiple parties through isolating each other"*. In turn, *network sharing* is defined as the sharing configuration where *"multiple CN operators have access to a common RAN" [32].*

The main challenges that should be addressed by RAN virtualization in LTE-A are i) the capacity limitation imposed by resource allocation, ii) the complete isolation between multiple coexisting services, and iii) the additional signaling overhead of each proposed solution. These challenges are even more complex to tackle in dense multi-tier scenarios, where Small Cells (SCs) are characterized by reduced coverage areas and therefore make the scenario more prone to geographical traffic non-uniformities [11].

Both dense SC scenarios as well as two tier scenarios where the SCs are overlaid with a macro BS pose different challenges when it comes to RAN virtualization. In particular traffic load and deployment are the foremost aspects to reveal the potential effectiveness of RAN virtualization in such scenarios. Although research solutions proposed so far have been mainly focused on the virtualization of resources in each Base Station (BS)[1], there is still a gap in the literature for solutions that abstract the available resources to deliver them to multiple tenant BSs, considering geographical traffic variations that can occur in dense scenarios.

In the following we present our contribution. Firstly, we introduce our proposal, the Resources nEgotiation for NEtwork Virtualization (RENEV) algorithm, for dynamic virtualization of radio resources spread both in a tier composed only of SCs and a heterogeneous scenario wherein resources are spread in two tiers - low power SCs which are overlaid with the existing macro-only cell. Motivated by the geographical traffic variations, we propose a solution where baseband modules of distributed BSs, interconnected via the logical point-to-point X2 interface, cooperate to reallocate radio resources on a traffic need basis. Our proposal is based on the concept of physical resources transfer, defined as the possibility of reconfiguring the Orthogonal Frequency Division Multiple Access (OFDMA)-based medium access of two BSs, to allow a BS to use a set of subcarriers initially allocated to another BS. Resource customization to various tenants (i.e., in this chapter we regard as tenants the involved BSs), is conducted after appropriate

[1]Throughout the rest of this manuscript, the term BS is used to describe either a macro eNB or a small cell (SC). The exact name of the BS is defined in all the particular cases that require the exact distinction among them.

signaling exchange. RENEV is a virtualization solution that abstracts resources, by customizing them in isolation among different Requesting BSs. Secondly, we identify the basic limitations and signaling overhead caused to the current 3GPP LTE-A architecture. In that sense, RENEV is harmonized and adapted to be compatible with LTE-A multi-operator network sharing configuration. Additionally, an insight on the analysis of the additional signaling overhead is given, since it is a key issue for virtualization, particularly as the network planning becomes denser.

The remainder of this chapter is organized as follows. Section 3.2 introduces the state of the art whereas Section 3.3 states our contribution. Section 3.4 provides an overview of the architectural elements and functions of the scenario and then the proposed algorithm is described. The signaling design considerations associated to each phase of our proposal in current 3GPP architecture are presented in Section 3.5. In Section 3.6 we introduce the analytical framework for network's throughput and in Section 3.7 we calculate the theoretical signaling overhead introduced by RENEV. Both experimental and analytical results are illustrated to show the performance of RENEV in Section 3.8. Finally, conclusions are given in Section 3.9.

3.2 State of the Art

Two possible architectural network sharing configurations have been specified and analyzed in section 2.4: the Gateway CN (GWCN) and the Multi-operator CN (MOCN). In GWCN configuration, CN operators share control nodes in addition to RAN elements whereas in MOCN, multiple control nodes owned by different operators are connected to a shared RAN. Throughout this chapter, the infrastructure owner provides the underlying physical network whereas by referring to network operator, we denote every operator having its users connected to the RAN (without necessarily owning infrastructure). In both configurations the network sharing agreement between operators is transparent to the end users. Although, operators may share network elements (i.e., RAN/control nodes), radio resources virtualization is required to cover their actual requirements, in isolation per BS. Therefore, in both MOCN and GWCN sharing configurations, virtualization of resources is necessary in order to allow operators' users to have access to the complete set of available resources. Existing network virtualization techniques, can be grouped into solutions for the Evolved Packet Core (EPC) Network and the RAN [34]. This chapter is focused on the RAN of an heterogeneous LTE-A deployment, which, in turn, can be divided in *dynamic resources' slicing* and *spectrum sharing*.

With regard to the dynamic resources' slicing, interesting proposals are presented in [28, 42, 43]. CellSlice framework is proposed in [28] to achieve active RAN sharing by

remotely controlling scheduling decisions without modifying BS's schedulers. As for spectrum sharing [27, 35–39], the proposals are designed to adapt the radio interface of the eNB to traffic load variations of distinct virtual networks. This objective is achieved by allowing multiple virtual networks to share the spectrum allocated to a particular physical eNB. A preliminary approach for virtualizing a BS in LTE is described in [35]. Further related works to dynamic resource slicing can be found in section 2.5.1 and to spectrum sharing in section 2.4.2.

Based on the state of the art presented above as well as in section 2.4, virtualization solutions proposed so far have been mainly focused on allocating resources, per operator/service, within a specific BS ([27, 28]). In particular, whereas in some proposals resources are dynamically sliced between services with different QoS characteristics ([35–37]), in other proposals the same resources are virtualized and distributed among different operators with shared access to the same BS ([38, 39]). Such proposals are effective virtualization solutions to address the traffic dynamics in two aspects: service and operator dimensions. In the first case, the variety of services poses challenges to resource allocation, whereas the second dimension is really interesting since the distribution of traffic between different operators is not necessarily uniform. However, none of the aforementioned proposals is able to cope with dynamics in a third aspect of traffic: the geographical dimension.

Heterogeneous networks (HetNets) are characterized by dense deployment of BSs with different transmission power and overlapped coverage areas. In these scenarios, the densification of the network with low-power BSs (i.e., SCs) has clear impacts on the traffic load: i) the distribution of the traffic between BSs is not uniform [11, 64], and ii) the variability of traffic in the short-term, particularly in SCs, is high. As a consequence the overall capacity of the system is usually compromised by spatial non-uniformities. Therefore, even appropriate deployments, which are static in nature, are unable to optimally tackle the spatial variations of the traffic.

3.3 Contribution

RENEV offers a complementary solution to the state of the art and covers gaps found therein by introducing a new dimension in RAN virtualization.

3.3.1 RENEV in Small Cell Deployment

In principle an efficient use of the available radio resources can be achieved if a proper coordination / negotiation of resources is carried out among the BSs. Thus we introduce

RENEV - our proposal for resources negotiation and firstly apply it only among BSs belonging to one tier (i.e., SCs or Home Evolved Universal Terrestrial Radio Access (E-UTRA) NodeBs (HeNBs)). The inclusion of SCs (e.g., HeNBs) along with the irregular traffic distribution poses several challenges in the management of the radio resources. Therefore we propose a solution to tackle the challenges appearing therein.

We consider that physical resource migration among HeNBs is necessary for covering the traffic demands of the existing users. In particular when a HeNB has some spare resources, it is available in order to participate to the resources negotiation process. This process is decentralized since all the existing HeNBs of a topology can participate, as soon as they have spare resources. In such environments our algorithm is responsible for reallocating / transferring radio resources by reconfiguring the OFDMA based radio interface in a decentralized manner. In this manner the baseband part of the BSs is shared and a common Radio Resource Control (RRC) layer for a specific group of BSs is created in a coordinated way. So in the first part of this work, we propose a solution for resource negotiation assigned only in BSs belonging to one tier (i.e., small cells). This is done by the cooperation of Radio Admission Control (RAC) and Radio Bearer Control (RBC) functions. This proposal is based on radio physical resource transfer in isolation and on-demand basis. Furthermore, this approach supports common RRC scheduling between different HeNBs.

RENEV is essentially designed to reconfigure the radio resources of two BSs, independent on the tier wherein they belong, in order to adapt the allocation of resources to the traffic dynamics of an operator. Thus, when there is a tenant BS without enough resources to serve the offered traffic, RENEV should find out if there are unused resources in other neighboring BSs, check if the unused resources could be reallocated, and finally reconfigure the medium access of the two BSs to reallocate them from one to the other (hereinafter also known as transfer of resources). In this scenario, the hierarchical or non-hierarchical operation of the nodes arises as a key aspect. To that end we extend and modify our initial approach in order to solve the problem of how to improve transmission conditions in a two tier heterogeneous (i.e., HetNet) scenario leveraging virtualization principles.

3.3.2 RENEV in HetNet Deployment

Since HetNets pose several challenges we modify and extend our solution for HetNet deployments such that BSs that belong to two tiers (i.e., both macro cell and SCs) are able to reallocate underutilized spectrum to other BSs. Our main contributions can be summarized as follows:

- We introduce RENEV as a solution, that can be employed on former RAN virtualization proposals (e.g., NVS [38] and PRR [39]), in HetNet scenarios composed of two tiers, each one operating on different sets of subcarriers. In these scenarios the geographical traffic non-uniformities render the initial allocation of resources into the BSs insufficient; some BSs are more loaded than others, resulting into areas that require more resources. On the one hand, RENEV is a virtualization solution that customizes resource slices from a BS to another, based on the traffic requirements created by the participating operators. On the other hand, virtualization solutions proposed so far in the literature (e.g., NVS [38] and PRR [39]) only allow resources customization among operators/services within the same physical BS.

- We demonstrate that RENEV could be applied independently on top of existing virtualization solutions (e.g., NVS [38] and PRR [39]), thereby guaranteeing its operation in multi-service multi-operator scenarios [36]. The implementation of RENEV does not impose additional constrains to the virtualization of resources within each tenant BS, proposed by the aforementioned solutions.

- We analytically derive the upper bounds of the throughput with and without RENEV.

- We provide the description and analytical model of the signaling introduced by RENEV, a key point in the dimension of the physical connections that support the logical X2 interface. This analysis arises as a key point in the dimension of the physical connections that support the logical X2 interface.

One of the main principles of network virtualization as a concept, is the division of the control and data planes of a system. RENEV, is based on the fact that the baseband part of the RAN nodes could be shared among different BSs; the target is to concentrate and orchestrate the control plane functionalities to serve a specific group of users. RENEV creates a common control plane among a group of BSs where the available radio resources could be dynamically transferred in the network, according to the users' demands in a holistic way. The control plane of LTE-A in the RAN nodes is concentrated in RRC protocol, which is terminated in the BS on the network side. Its main functionalities are the establishment of the connections with the users, configuration of the radio bearers and their corresponding attributes and control of mobility.

3.4 Resources Negotiation for Network Virtualization (RENEV)

3.4.1 Network Configuration and Assumptions

In this section, we introduce (i) the specific network sharing configuration where the resources virtualization by RENEV is applied and (ii) its architectural elements (i.e, the RAN nodes (BSs), the control nodes and their interconnecting interfaces). In our scenario different CN operators may connect to a shared RAN [32]. We study the GWCN configuration, where different operators may also share the same control node. This sharing configuration, consists of a set of resources belonging to the RAN elements and need to be customized in isolation among users of multiple operators. Regarding the RAN elements (whose resources need to be virtualized), the underlying considered network is a residential region composed of an eNB and a number of open access mode SCs placed throughout its coverage area in clusters, close to each other, in random positions [65].

When we study two tiers we assume that they are initially assigned disjoint frequency bands [66]; however by exploiting the concept of Carrier Aggregation (CA), both tiers can operate on the whole bandwidth [5]. Most RAN nodes maintain standardized connections to each other, for example, BSs are connected to their neighbors using the point-to-point, logical X2 interface to support a direct control and data information exchange. Furthermore, we focus on the downlink, where the RB is the basic time-frequency resources unit. In principle, any RB can be assigned to one or several BSs subject to interference limitations. The eNB is assumed to transmit with a fixed power per RB. The downlink transmitted power per RB is also fixed and equal among the SCs [67].

The BSs are connected to the EPC directly with the Mobility Management Entity (MME) or through an intermediate node, named Home eNB Gateway (HeNB GW) using the S1 interface [68]. These nodes manage BSs to provide a radio network. According to GWCN network sharing configuration [32], these control nodes are shared by different operators as defined by their Service Level Agreement (SLA). Therefore, this sharing configuration may host a scalable number of CN operators owning both CN and RAN nodes.

Based on [62], three ways of interconnection of the tenant BSs arise: (i) a cluster of SCs (i.e., in our test case HeNBs) connected to the same HeNB GW, (ii) a group of eNBs connected to the same MME and (iii) a group of eNBs as well as SCs associated to the same MME. In the first and second case, the HeNB GW and the MME concentrate the

control plane of the SCs and the eNBs respectively. In the last case the MME integrates the control plane of both types of BSs within a certain geographical area. Despite the different cases presented in [62], from a BS's perspective all cases are identical in terms of signaling. This means that the message exchange from the BS-BS communication required by RENEV, is independent from the coverage area and the transmission power of a BS. Therefore, we consider equivalent the cases of message exchange between eNB-SC and SC-SC that is required when executing RENEV. Under these circumstances, in a scenario like this, we further assume that the involved BSs are necessarily deployed over the same geographical area, and therefore connected to the same control node (i.e., MME) which is shared by multiple operators.

3.4.2 Radio Resource Management Functions

The management of spectrum resources allocated to the BSs, relies on their control plane.

The control plane of a BS in LTE-A is logically divided in two entities: baseband and network modules, as defined in the standard in [62]. The former is responsible for bearer setup, to register users from each operator to the network via RRC protocol, whereas the latter connects the BS with the EPC. Radio Resource Management (RRM) is implemented in baseband module of a BS with primary goal to control the use of radio resources in the system, by ensuring QoS requirements of the individual radio bearers and minimization of the overall use of resources.

Focusing on the baseband module, two fundamental functions of the RRM jointly manage the resources of a BS: the Radio Bearer Control (RBC) and Radio Admission Control (RAC) [62]. On the one hand, RBC is responsible for the establishment, maintenance and release of radio bearers. When setting up a radio bearer, RBC considers the overall resource situation and QoS requirements of in-progress sessions [69]. Correspondingly, it is involved in the release of radio resources at session termination. On the other hand, the task of RAC is to admit or reject the establishment requests for new radio bearers. RAC ensures high radio resource utilization by accepting bearer requests from operators as long as radio resources are available. At the same time, it ensures proper QoS for in-progress sessions by rejecting radio bearer requests when they cannot be accommodated [70]. A new bearer will be built only if radio resource in the cell is able to maintain the QoS of the current sessions. It will be released at the end of the communication. Based on the role played by RBC and RAC, any RRM technique aimed to improve the efficiency in the dynamic allocation of the radio resources among BSs must interact with these two functions.

3.4.3 RENEV in Small Cell Deployment

In this section we introduce our initial design of RENEV where resources are managed within one tier (i.e., SCs). The inclusion of SCs (e.g. HeNBs) along with the irregular traffic distribution poses several challenges in the management of the radio resources. In such a context, we study a residential region composed of an eNodeB and a number of HeNBs located close to each other in random positions. When a user is served by a certain HeNB, this latter is called serving HeNB of the user. We consider an open access small cell network and we further assume that the downlink transmitted power is fixed and the same for all the HeNBs.

Physical resource migration among HeNBs, is necessary for covering the traffic demands of the existing users. Since a HeNB has some spare resources, it is available in order to participate to the resources negotiation process. This process is decentralized since all the existing HeNBs of a topology can participate, as soon as they have spare resources. So we propose RENEV for resources negotiation between SCs, by the cooperation of RAC and RBC functions, belonging in RRC of different small cells. This solution is based on radio physical resource transfer in isolation and on-demand basis. Furthermore, our solution supports common RRC scheduling between different HeNBs. RENEV is described by the following steps:

Step 1: If a user can be served by the resources of the serving HeNB then it gets served [41].

Step 2: Otherwise:

- The user enables the RRC connection with the serving HeNB.
- This HeNB finds the nearest [1] and less loaded neighbor HeNB.
- When it finds it, the two RRC functions of the node are enabled; the RAC function is responsible for checking if the node has the available resources and the RBC for establishing the radio bearer.
- A control connection is created between the involved HeNBs via the logical point-to-point X2 interface.
- The serving HeNB leases the demanded resources so the user is getting served. This happens by setting up X2 interfaces and resetting the link resolving security issues for the exchange of HeNB configuration data over the link.

[1] The term "nearest" indicates the neighbor HeNB located geographically closer to the serving HeNB. This fact restricts the effect of the algorithm geographically, in order to avoid instability issues.

The target metric that this algorithm improves is the aggregate system throughput, since the resources negotiation in terms of RBs affect the data rates that are delivered to all terminals in a system.

3.4.4 RENEV in HetNet Deployment

HetNet scenarios pose further challenges in the management of resources. In the scenarios such as the ones described in Section 3.4.1, traffic non-uniformities among BSs make resource allocation a challenging task. A dynamic coordination of radio resources is required to address such kind of variations. This is the objective of RENEV in these environments; customizing resources in terms of RBs, to satisfy new incoming user requests by multiple operators in tenant BSs, while supporting isolation among the reallocated resource slices.

Let us define the number of RBs initially allocated to a particular BS as RB, and the number of RBs required to serve the demand of its associated users as u. By definition, the number of available RBs in this specific BS, denoted as r, can be expressed as $r = RB - u$. As long as $r > 0$, the tenant BS will be able to serve the offered traffic. Conversely, when $r < 0$, the BS will start to degrade users' performance and block UEs' incoming attachment requests.

It is particularly worth noting that in HetNets the significant variability of the traffic among neighboring BSs can lead to the paradox of having some BSs with $r < 0$ and, at the same time, some other BSs with $r \gg 0$. RENEV is defined as the decentralized procedure intended to match the tenant BSs with $r < 0$ and the ones with $r > 0$, and manage the exchange of control messages to reconfigure the allocation of resources among them. For this reason, RENEV is divided into two sequential phases, as shown in Fig. 3.1.

First, the detection phase, where a BS with $r < 0$ seeks among the neighbouring BSs if any of them has $r > 0$. This search is carried out by polling one by one the neighbouring tenant BSs to figure out the amount of available resources. Subsequently, the transfer phase is only executed if the tenant BS with $r < 0$ finds neighboring BSs with $r > 0$. This phase consists in re-configuring the two involved BSs according to the operators' traffic requirements. The details of each phase are stated below, and a proposal of the messages exchanged during the two phases is described in Section 3.5. Before proceeding with the details, we describe the basic nomenclature:

- **Serving BS:** is the node that a User Equipment (UE) is associated to and it is responsible for serving it.

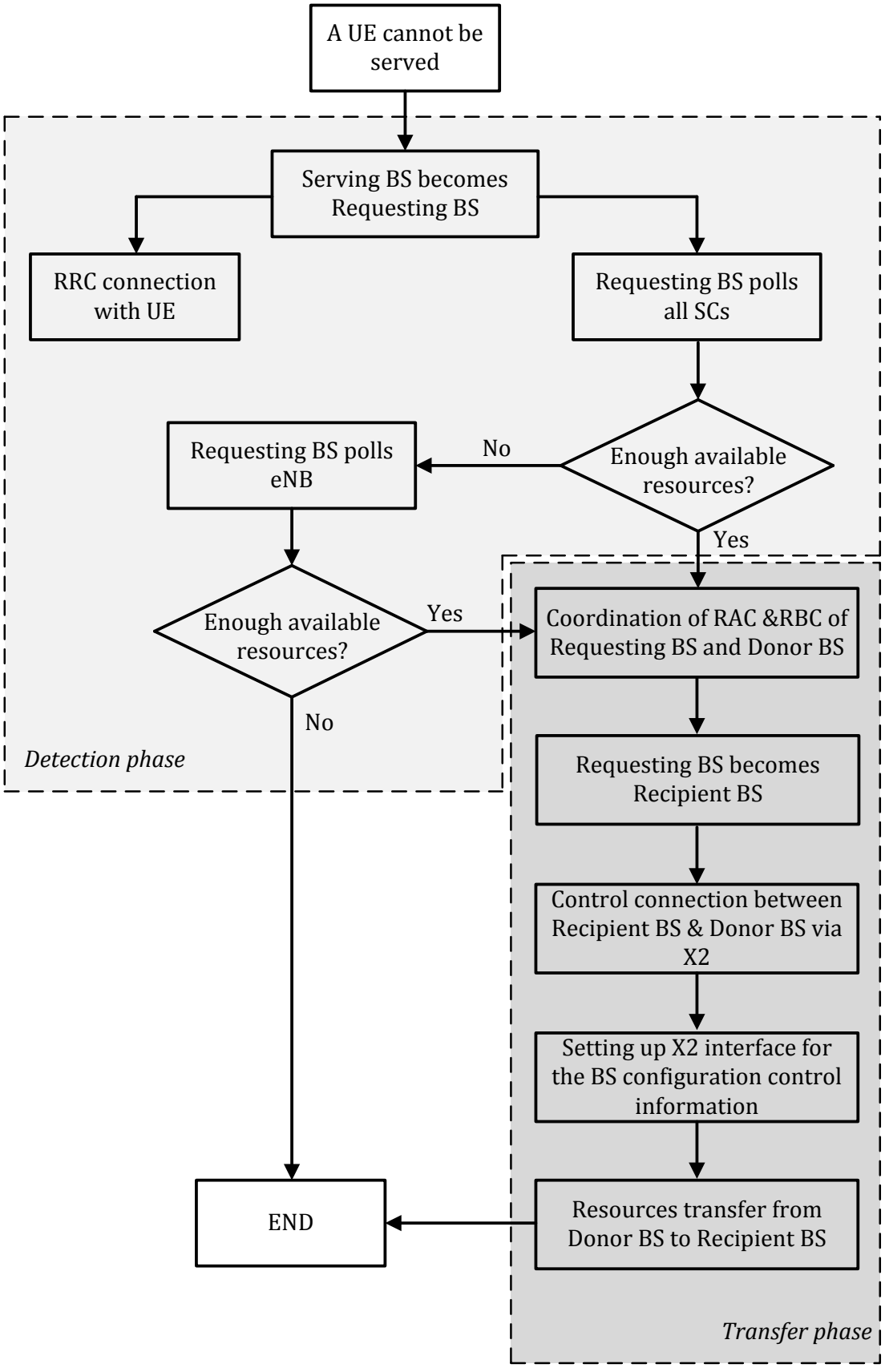


FIGURE 3.1: RENEV for a Heterogeneous Deployment.

- **Requesting BS:** is the node that, after receiving an access request from a UE, determines that the request cannot be accommodated with the available resources. It is precisely at this time, that the node takes the role of Requesting BS and triggers a requesting process among the neighboring BSs to figure out if there are unused resources.
- **Requested BS:** is the node that, after a neighboring Requesting BS triggers a requesting process, receives a request to inform about its unused resources.
- **Donor BS:** is the node that, upon the completion of a requesting process triggered by a Requesting BS, is selected to transfer resources to this Requesting BS.
- **Recipient BS:** is the role taken by a Requesting BS after reconfiguring the radio interface to use the resources transferred from a Donor BS.

Since, in general, spectral efficiency of SCs is higher than spectral efficiency of eNBs, SCs play the role of Requesting BSs. SCs are usually needed for dense deployments in high-traffic environments and therefore, they are more prone to lack resources. This is the main difference in the scale of macrocells and SCs. Thus, if the eNB could play the role of Requesting BS, the RBs transferred from a SC to the eNB could not be reused by any other SC, resulting in a reduction of the capacity. In RENEV, RBs transferred by the eNB can be reused in more than one SC in the SCs tier, given that the involved SCs do not have overlapped coverage areas. When the imbalance between the demanded and the allocated resources comes to an end (i.e., the additional resources transferred by RENEV to a Requesting BS are no longer needed), the resources given by the Donor BSs (resulting from the execution of RENEV) reverts to the initial allocation. As a consequence, the role of Requested BS can be held either by SCs or an eNB. In that sense, SCs can be both Donor and Recipient BSs, whereas eNB is always a Donor BS.

3.4.4.1 Detection phase

If a user from an operator can be served by resources owned by the Serving BS (i.e., $r > 0$), then it is served [70]. Otherwise, the Serving BS, after setting up a RRC connection on the air interface with the user requiring service provision, triggers RENEV by adopting the role of Requesting BS. At this point the detection phase starts (see Fig. 3.1). Next, the Requesting BS scans the local network[2] to find a potential Donor BS by polling BSs around it. The polling procedure undertaken by the Requesting BS may itself be divided into two steps. First, the Requesting BS polls each neighboring

[2]The local network of a BS is defined as the set of BSs deployed in its vicinity. Generally, this local network consists of an eNB and a finite number of SCs under its coverage area.

SC, one by one, to monitor the resources status of the SCs tier. Secondly, if there are not available resources in the SCs tier, the Requesting BS polls the macro eNB. After completing the requesting process, the Donor BS is selected among the set of Requested BSs according to two criteria: load and proximity.

1. **Load:** The Requested BS with more unused resources is selected as the Donor BS. Yet, in order for the Donor BS to be able to accommodate possible further increase of the traffic demand in the short/mid-term future, a Requested BS can only become a Donor BS if the amount of remaining resources after the transfer is above a minimum threshold.

2. **Proximity:** For a set of Requested BSs likely to become the Donor BS, and if more than a single Requested BS has the same amount of unused resources, the Donor BS will be the BS with the minimum distance to the Requesting/Recipient BS. This criterion guarantees that the effect of the algorithm is geographically restricted to limit undesirable instability problems caused by the nature of the wireless medium.

Regarding the implementation details of this phase, when a user is attached to the Requesting BS, the RRC connection establishment is used to make the transition from RRC Idle to RRC Connected mode. This transition is carried out before transferring any application data, or completing any signaling procedures, as shown in Fig. 3.1. RRC establishment procedure is always initiated by the user but it can be triggered by the user or the network [70]. When the Requesting BS scans the network to find a Donor BS, a coordinated control connection of their baseband parts is created via the X2 interface. Every time that a polling procedure between a Requesting BS and a Requested BS is carried out, two messages are exchanged through X2 interface (one from the Requesting BS to the Requested BS, and another one vice versa).

3.4.4.2 Transfer phase

Upon detecting the Donor BS, the transfer of resources from the Donor BS to the Recipient BS takes place via X2 interface. It is worth noting, that the exchange of BS configuration data over the link must be preceded by resetting the link resolving security issues.

In the proposed scheme, RAC and RBC functions, belonging to RRC layer of distinct neighboring BSs, cooperate to provide seamless service to the end users (first action of the transfer phase, dark shaded in Fig. 3.1). We leverage the logical split of a BS

into baseband and network modules and create a common RRC process among the Recipient BS and the Donor BSs. When the Requesting BS finds the Donor BS, RRC functions of the two nodes are enabled; RAC is responsible for checking if the node has available resources and RBC for establishing the radio bearer; it is in that moment that the Requesting BS becomes the Recipient BS. Next the medium access of two involved BSs is reconfigured and spectrum is lent by the Donor BSs through the control communication of the nodes. This process is seamless to end users since RRC connection is maintained with the initial Requesting BS and it is done without the participation of additional BSs or gateways. Finally, the Donor BS leases the demanded resources, which are used by the Recipient BS.

3.4.5 Discussion on RENEV

3.4.5.1 RAN Virtualization Properties

In this subsection, we introduce the key virtualization properties of RENEV, its main differences with conventional joint resource allocation solutions and how it can interact with already proposed RAN virtualization schemes. RENEV provides the virtualization features defined by 3GPP SA1 RSE requirements [38] :

- **Abstraction:** RENEV abstracts the radio resources belonging to a deployment into a pool; these resources are delivered on demand to each Requesting BS according to the operators' needs. In particular, abstraction of resources is accomplished by the communication between Requesting BS and Requested BS (as defined in Section 3.4.4.1). Instead of having the view of the physical radio resources (i.e., RBs) in each BS, RENEV after being triggered creates a set of virtual resources. This set consists of physical resources coming from different Donor BSs and it is accessible by various Requesting BSs according to the existing traffic non-uniformities.

- **Isolation:** RENEV ensures a reserved portion of resources to each Requesting BS that triggers it, to meet the requirements of the operators, in this specific BS. Traffic, mobility and fluctuations in channel conditions of one Requesting BS do not affect the reserved resource allocations of other Requesting BSs. More specifically, isolation is achieved during the Transfer phase (defined in Section 3.4.4.2) where RENEV creates a logical common RRC process among the Recipient BS and the Donor BS. RAC and RBC functions for the Requested BS are enabled, and the lent resources are seamlessly reserved to be used by a particular Recipient BS.

- **Customization:** RENEV offers the flexibility to different operators having access to the sharing configuration, to conquer different part of the shared resources according to the actual requirements. Resource customization is attained during the Detection phase of RENEV (defined in Section 3.4.4.1). When a BS runs out of resources, RENEV is triggered so as resources can be allocated to the Requesting BS that needs them according to the specific traffic load conditions.
- **Resource Utilization:** RENEV guarantees the efficient use of physical radio resources with a rational signaling burden for applying the solution onto the network. The medium access of each pair Requesting - Requested BSs is reconfigured during RENEV. Thus, the spare spectrum is lent by Donor BS through the control communication of the nodes.

3.4.5.2 Differences with Joint Resource Allocation and Generic Resource Sharing

The aforementioned properties distinguish in general virtualization solutions from conventional joint resource allocation ones. As defined in [71], the latter *"apply a joint optimization approach (power control, channel allocation, and user association) for resource allocation in a multi-cell network, which can be invoked at the network planning stage or when the resource status changes"*. Although in such kind of solutions, resources are allocated among cells, the isolation property does not hold. Traffic, mobility and fluctuations in channel conditions of users of one entity affect the resources that would be given to other entities. In RENEV, the customization of resources among tenant Requesting BSs is performed on demand, with the target to serve as many users as possible, belonging to distinct network operators that share the RAN. In our solution, dedicated resources are served and locked per Requesting BSs to be allocated to a user of a certain participating operator. However, a conventional multi-cell joint resource allocation solution, does not isolate any resources for specific operators within the topology. Therefore in RENEV isolated slices of RAN can be assigned to Requesting BSs, to serve the traffic needs of users belonging to distinct operators.

It is important to differentiate RENEV from generic resource sharing approaches. That is because generic resource sharing among multiple operators can be performed with or without virtualization. To highlight the difference, let us consider the Spectrum Sharing (SS) scheme presented in [15]. It represents a traditional resource sharing approach that is done via a "request and release of spectrum" method where the portions that can be allocated are fixed and it is performed into BS level. According to SS, a supply sector belonging to an operator allows access of a portion of its own carrier to a heavily

loaded demand sector (i.e., leased sector) of another collocated operator. Unlike resource sharing via virtualization, this sharing procedure requires spectrum division and reconfiguration - during this process the operators' users are put in a suspended state. While this is a conventional case of resource sharing in a BS, when adding virtualization, the allocation of resources takes place dynamically and the reconfiguration process is not necessary because the supply and the leased sectors share a number of physical RBs. For example in the NVS [38] case, this is because the BS scheduler is modified: this virtualization solution does not require the same operators to be collocated in order to share the resources, neither the suspended state for the users. The trade-off is the added complexity due to the BS MAC scheduler modification. Similarly, RENEV is also performing a virtualization of resources, but in a higher level. Altogether, RENEV achieves resource sharing among operators via virtualization in a process that does not take place in each specific BS but in a set of resources owned by a geographically constrained BS set.

3.4.5.3 Interaction with existing Virtualization Proposals

Also it is worth emphasizing that RENEV operates independently, on top of existing virtualization solutions, such as NVS [38] and PRR [39]. In general, each virtualization mechanism abstracts physical resources to a number of virtual resources, which are then delivered in isolation to different tenants. However, resource virtualization may appear in different levels and distinct solutions can exist that determine how resources are distributed: *within each BS* and *above the BS*.

NVS and PRR are indicative examples for virtualization within each BS. For instance if NVS is implemented, a particular BS will lack resources as soon as the traffic from an operator consumes all resources devoted to it [38]; if PRR is applied, all the traffic load from an operator will be served as long as the shared part of resources belonging to particular BS, is nonempty [39]. Thus, the needs or surpluses of resources within the BS vary, based on the aggregate traffic demand by the operators in a particular BS and how the resources are distributed within it. However looking at the top-down approach, RENEV also virtualizes resources, but from a higher network perspective (i.e., above the BS): instead of performing its tasks per BS level, RENEV abstracts and slices the physical RBs according to the spatial traffic non-uniformities. The delivery of these resources is done on demand according to these requirements as described in section 3.4.4.

All in all, there is no need for explicit communication between RENEV and these solutions. RENEV can first be implemented to virtualize resources among Requesting

BSs of the whole deployment based on the aggregate operators' requirements (due to geographical non-uniformities of their traffic). Then NVS [38] and PRR [39] may be applied in each particular BS, to customize the available resources (made accessible by RENEV when required) among operators.

3.5 Signaling Design Considerations

The additional signaling overhead introduced in the network is a key aspect of the proposed solution, since it could limit its feasibility. This section is intended to analyze in detail the signaling messages exchanged in the network to implement RENEV, as well as its compliance with the current standards and architecture of LTE-A. A short discussion about the time scale of RENEV is also introduced.

Any procedure concerning the accommodation of a new user in a cell, starts with its attachment as explicitly defined in the standard [70]. The attachment of a user to a new cell is characterized by two main processes: firstly, the communication between UE and Serving BS over air (i.e., Uu) interface, and secondly, the communication between Serving BS and the MME to exchange initial UE context setup over S1 interface.

Regarding the message exchange over Uu interface, the user sends the attach request message to the Serving BS, as also defined in the standard [70]. After sending the first message of random access procedure to the network, denoted as RACH preamble, RRC connection is established. The initial UE context setup, consists of an exchange of messages with the purpose of transferring UE context information from the MME to the Serving BS. These messages are exchanged over S1-AP application layer using SCTP. When the appropriate RRC transport container is received by the Serving BS, the establishment of a dedicated SCTP control stream on S1-MME is triggered [72]. The described procedure is nonetheless subject to the availability of resources in the Serving BS. In that sense, RENEV aims to transfer resources from one BS to another to minimize the number of unsuccessful procedures. Therefore, RENEV should be executed after the UE attachment request and before the UE context exchange.

The direct communication between two BSs is conducted via X2, using the X2 Application Protocol (X2-AP) [62]. X2-AP messages are characterized by communication context identifiers and some specific parameters called Information Elements (IEs). These define the source and target BS, as well as characteristics of the transferred message. The messages required to implement RENEV are detailed below.

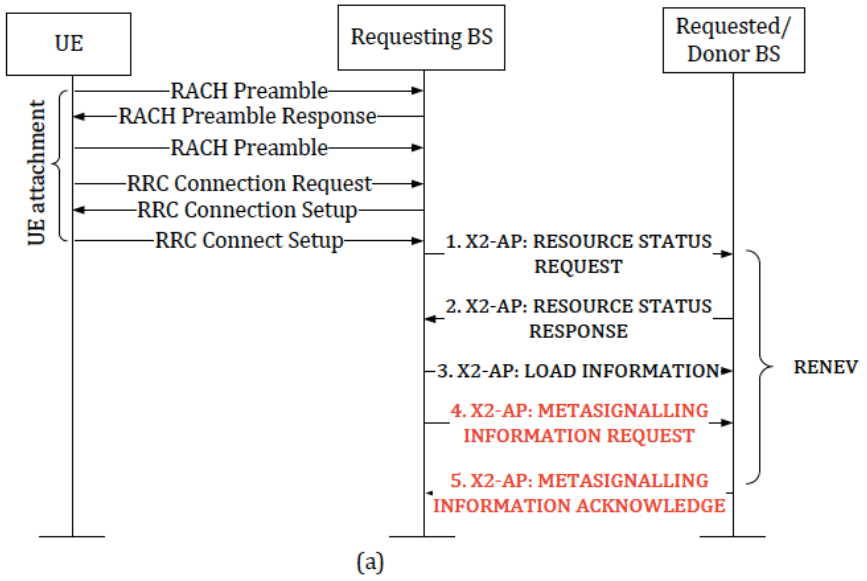


FIGURE 3.2: Call Flow of the messages for UE Attachment and RENEV.

3.5.1 Detection phase signaling

When applying RENEV, the first process to carry out includes the polling procedure to detect spare resources (see Fig. 3.2 and Fig. 3.3, messages 1, 2 and 3). During this operation, the Requesting BS scans the network to find the Donor BS, as shown in Fig. 3.1. For each Requesting BS-Requested BS pair the polling process entails the information exchange about resources and load status [62]. In the standard, the X2-AP defines two Elementary Procedures (EP) for this same purpose, namely the "Resource Status Initiation" and "Load Indication" procedures [62]. The former is defined as a class 1 EP (i.e., it consists of two messages, a request and a response, namely "X2-AP:RESOURCES STATUS REQUEST" and "X2-AP:RESOURCE STATUS RESPONSE" messages), whereas the latter is defined as a class 2 EP (i.e., it consists of a single message, without response, namely "X2-AP:LOAD INFORMATION" message). RENEV makes use of these two EPs, defined by the X2-AP, to implement the detection phase.

As shown in Fig. 3.2 and Fig. 3.3, the Requesting BS sends the standardized "X2-AP:RESOURCE STATUS REQUEST" message to the Requested BS (Fig. 3.2, message 1) asking for the following information (known as IE in the X2-AP nomenclature): the percentage of RBs in use, the load on S1 interface and the hardware load. The Requested BS returns a response and then reports each IE for both uplink and downlink with the standardized "X2-AP:RESOURCE STATUS RESPONSE" message (Fig. 3.2, message 2) [73]. Also, Load Indication procedure is used to transfer interference coordination information between neighboring BSs managing intra-frequency cells. The standardized "X2-AP:LOAD INFORMATION" message (Fig. 3.2, message 3) includes

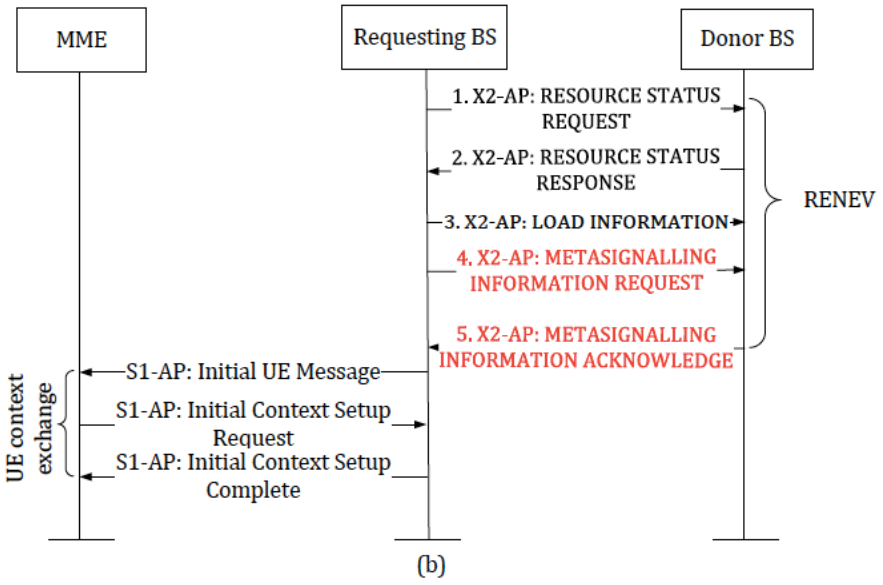


FIGURE 3.3: Call Flow of the messages UE Context Exchange and RENEV.

three IEs for the controlling cell: the transmitted power in every downlink RB, the interference received in every uplink RB, and the list of uplink RBs in which the BS intends to schedule distant mobiles [73]. These control messages are necessary before transferring additional control information for establishing common RRC layer among BSs with RENEV. This procedure is repeated for all neighboring SCs. If none of the requested SCs has enough unused resources, the procedure is repeated with the eNB. Up to this point, all messages used by RENEV in the detection phase are defined in the standard [73].

3.5.2 Transfer phase signaling

We define the transfer of resources as the reconfiguration of a set of unused subcarriers to be vacated by the Donor BS and subsequently used by the Recipient BS. As this procedure is not considered in X2-AP, a new Class 1 EP compatible with the standard should be defined. In this chapter the two proposed messages of the new EP are the messages 4 and 5 (Fig. 3.2). We denote them as "X2-AP:METASIGNALLING INFORMATION REQUEST" and "X2-AP:METASIGNALLING INFORMATION ACKNOWLEDGE", although other possible implementations are not precluded.

Once the Donor BS is selected, the initiating "X2-AP:METASIGNALLING INFORMATION REQUEST" message (message 4 in Fig. 3.2) is transmitted from Requesting BS to the Requested BS to show that resources are required by the former. The message must contain the following IEs: Message Type, Requesting BS X2-AP ID, Requested BS X2-AP ID and the corresponding transparent container. These IEs indicate the number

of necessary RBs to cover the needs of the UE, and the identities of the Requesting and Requested BSs. For its part, the Donor BS returns a response to the Recipient BS via "X2-AP:METASIGNALLING INFORMATION ACKNOWLEDGE" message (see Fig. 3.2, message 5). This message carries all control information needed to execute the actual transfer of resources. The corresponding IEs are the Message Type, Cause, Bearers Admitted List, Bearers Rejected List and the equivalent transparent container. These IEs are necessary to confirm that the requested RBs exist in the Donor BS and that they are available for use by the Requesting BS.

3.5.3 Discussion on the Time Scale of RENEV

One dimension regarding the time scale of RENEV, is related to its duration. The algorithm is triggered every time that a Requesting BS lacks resources. The main RRC functions that have to be triggered per BS, RAC and RBC, are Layer 3 RRM functions. Therefore, the time scale of the algorithm resides on the time scale that RAC and RBC need in order to be activated in each BS. Another dimension of the time scale of RENEV, regards its periodicity of triggering. To begin with, it is expected that too often triggering of RENEV leads to excessive signaling of message exchange. In the second place, parameters such as the number of users or their mobility affect the signaling burden exchanged by RENEV. The ability of exchanging messages over X2 interface resides on the actual implementation of the interface (i.e., over the air wireless, fiber etc.). These are design parameters by the infrastructure owner. To conclude, although frequent triggering of RENEV leads to better adaptation to traffic variations, it also leads to higher message exchange over X2 interface, thereby increasing exponentially the signaling.

3.6 Throughput Analysis

3.6.1 System Model

As described in Section 3.4.1, the scenario consists of a single macro eNB (hereafter denoted as BS_0 and located at the center of the scenario) and a SCs tier, made up of a set of SC clusters, each one consisting of $N \in \mathbb{N}$ SCs (denoted as BS_i, with $1 \leq i \leq N$), randomly distributed on a two-dimensional Euclidean plane $\mathbb{R}^2$. As clusters are not overlapped, there is no loss of generality in assuming one SC cluster within the eNB coverage area, creating a set of N+1 BSs, which is referred to as $B = \{BS_i : 0 \leq i \leq N\}$. We denote as $X \in \mathbb{N}$ the number of the overall users within the deployed scenario. These X users are divided into N+1 traffic layers, each one geographically spanned over the

coverage area of a BS. The coverage area of a BS_i is defined as the region where users are served by this specific BS and all the users are assumed to be connected to the BS from which they receive the best Signal-to-Noise-Ratio (SNR), given that there are available RBs within BS_i. Given the described scenario, if the proportion of users contained within the coverage area of BS_i is denoted as a_i, the number of users within this coverage area may be expressed as $X_i = a_i X$, with $\sum_{i=0}^{N} a_i = 1$. Within each traffic layer, users are distributed uniformly.

3.6.2 General Throughput Formulation

The LTE-A standard defines a discrete set of Modulation and Coding Schemes (MCSs) with the following possible configurations in the downlink for data transmission for both SCs and the eNB: QPSK ($\frac{1}{8}$, $\frac{1}{5}$, $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$), 16-QAM ($\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$) and 64-QAM ($\frac{2}{3}$, $\frac{3}{4}$, $\frac{4}{5}$) [74]. Based on a target bit error rate, the MCS is selected by the BS according to the SNR received by the user. In that sense, given that the transmission rate depends on the applied MCS, the expected transmission rate per RB of a user connected to BS_i is

$$\mathbb{E}[R_i] = \sum_k P(\text{MCS}_i = k) \cdot R_{ik}, \tag{3.1}$$

where $P(\text{MCS}_i = k)$ is the probability of using the k^{th} MCS in BS_i, and R_{ik} is the transmission rate (in bps) achieved within a single RB with the k^{th} MCS. The derivation for $P(\text{MCS}_i = k)$ may be found in Appendix A.1. Note that (3.1) is valid for eNB and SCs. However, due to the overlapping of the coverage areas of the SCs and the eNB, the users located within the coverage area of a SC could be connected to the eNB if the available resources allocated to SCs do not suffice. In other words, a user of the i^{th} traffic layer (with $i \neq 0$) could get connected to BS_0 despite $\text{SNR}_i > \text{SNR}_0$. Hence, if a user within the i^{th} traffic layer (with $i \neq 0$) is served by the eNB, the expected transmission rate per RB is given by

$$\mathbb{E}[R_i^0] = \sum_k P(\text{MCS}_i^0 = k) \cdot R_{ik}, \tag{3.2}$$

where $P(\text{MCS}_i^0 = k)$ is the probability of using the k^{th} MCS in the eNB (i.e., BS_0) with a user in the i^{th} coverage area ($i \neq 0$). Based on this, for a given number of users X_i, there is a group of users associated to BS_i, namely X_i^i, and a group of users associated to BS_0, denoted by X_i^0. Thus, for a given X_i, the expected number of users associated to BS_i is

$$\mathbb{E}[X_i^i] = \min\left(X_i, \frac{RB_i \cdot \mathbb{E}[R_i]}{d}\right), \tag{3.3}$$

where RB_i is the number of RBs allocated to BS_i and d is the specific demand of every single user (in bps). According to (3.3), $\mathbb{E}[X_i^i] = X_i$ if RB_i is enough to serve all the attached users. Otherwise, not all users associated to BS_i will be served. The maximum number of users that can be served is calculated from the expected maximum throughput, defined as the expected throughput per RB (i.e. $E[R_i]$) multiplied by the number of available RBs, RB_i. Thus, the expected maximum number of users is $\frac{RB_i \cdot E[R_i]}{d}$. Finally, by definition, $\mathbb{E}[X_i^0] = X_i - \mathbb{E}[X_i^i]$. According to the definition, the total throughput, expressed as the sum of the throughput of each BS (i.e., $T = \sum_i T_i$), depends on the number of users from every operator connected to each BS, the transmission rate per RB, as well as the amount and the distribution of the available resources. In the following, we assume the use of a first-come first-served policy in each BS. This policy is equivalent to an extreme case of PRR, where 100% of the resources in each BS are shared and delivered on-demand (hereafter denoted as PRR 100%). This assumption (with and without RENEV) results in the upper bound of the aggregate throughput.

3.6.3 Aggregate Throughput with RENEV

Although RENEV negotiates resources in a peer-to-peer fashion among BSs, the procedure can be stochastically modelled as a single pool of resources dynamically allocated to the tenant BSs, when RENEV and PRR 100% are implemented.Let us denote the throughput served by the eNB and generated by the X_0 users, as $T_{R,0}$. This throughput will equal the traffic generated by X_0 users associated to BS_0, subject to the availability of sufficient resources (i.e., RB_0). Thus,

$$T_{R,0} = \min\Big(X_0 \cdot d, \mathbb{E}[R_0] \cdot RB_0\Big). \tag{3.4}$$

Note that, when RENEV is applied, the eNB tends to transfer resources to the SCs, if necessary and feasible, rather than serve users within the coverage area of the SCs. Therefore, $X_i^0 = 0$ for $\forall i \neq 0$. In turn, SCs serve their users with all the resources allocated within the SCs tier, as well as with unused resources in the eNB, RB_0. The application of RENEV may be modeled with two unified pools of resources; one composed of the RBs belonging to the SCs tier (denoted as $RB_T = \sum_{i \neq 0} RB_i$) and one consisting of the RBs from the eNB. Each Requesting BS will receive proportionally to its traffic load, resources from the SCs pool (i.e., $\frac{a_i}{1-a_0} \cdot RB_T$) and the corresponding portion of resources belonging to the eNB pool, denoted as $\mathbb{E}[RB_i^s]$. Therefore, the aggregate throughput generated by the SCs tier, according to the proof provided in

Appendix A.2, can be written as

$$\sum_{i\neq 0} T_{R,i} = \min\left(X\cdot(1-a_0)\cdot d, \sum_{i\neq 0}\mathbb{E}[R_i](\frac{a_i\cdot RB_T}{1-a_0}+\mathbb{E}[RB_i^s])\right). \tag{3.5}$$

Consequently, the expected overall system throughput with RENEV, is given by: $T_R = T_{R,0} + \sum_{i\neq 0} T_{R,i}$.

3.6.4 Aggregate Throughput without RENEV

Alternatively, when RENEV is not applied (still considering a first-come first-served policy per BS, or in other words PRR 100%), there is not any mechanism to reallocate resources, and consequently all BSs can only serve users with their initially allocated RBs. Similarly to (3.4), the throughput of each SC is $T_{NR,i} = \min(X_i\cdot d, \mathbb{E}[R_i]\cdot RB_i),\ \forall i\neq 0$.

As for the eNB throughput, it is divided into two components: the throughput offered by the X_0 users within the coverage area of BS_0 (i.e., $T^0_{NR,0}$); and the traffic offered by users within the coverage area of the SCs that cannot be served by these BSs due to lack of resources (i.e., $T^0_{NR,SCs}$):

$$T^0_{NR,0} = \min\left(X_0\cdot d, \mathbb{E}[R_0]\cdot RB_0\right) \tag{3.6}$$

$$T^0_{NR,SCs} = \min\left(\sum_{i\neq 0}\mathbb{E}[X_i^0]\cdot d, \mathbb{E}[R_i^0]\cdot\left(RB_0 - \frac{T^0_{NR,0}}{\mathbb{E}[R_0]}\right)\right). \tag{3.7}$$

According to (3.7), if the available resources by the eNB (i.e., RB_0) are enough to serve the users in the coverage area of the SCs that cannot be served by them due to lack of RBs (i.e.,$\mathbb{E}[X_i^0]$ with $i\neq 0$), then they are served and their throughput equals $T^0_{NR,SCs} = \sum_{i\neq 0}\mathbb{E}[X_i^0]\cdot d$. Otherwise, not all $\mathbb{E}[X_i^0]$ users are served. The maximum throughput that can be achieved is calculated from the expected maximum throughput per RB (i.e., $\mathbb{E}[R_0^i]$), multiplied with the available remaining RBs in the SC tier. To calculate the latter, we subtract from the total number of RBs belonging to the eNB (RB_0), the ones used to serve the eNB's traffic (i.e., $\frac{T^0_{NR,0}}{\mathbb{E}[R_0]}$). Therefore, the aggregate throughput without RENEV is,

$$T_{NR} = (T^0_{NR,0} + T^0_{NR,SCs}) + \sum_{i\neq 0} T_{NR,i}. \tag{3.8}$$

3.7 Additional Signaling Overhead Analysis

The densification of the network via the deployment of numerous SCs poses challenges in the infrastructure. Specifically, the need for a backhaul to interconnect BSs and forward both data traffic and signaling has emerged as one of the key points that could constrain the feasibility of these scenarios. Focusing on the implementation of RENEV, the whole communication among BSs relies on the existence and capacity of the logical X2 interface (as described in Section 3.5). Although this logical interface is standardized [62], the description of the backhaul physical infrastructure in order to support it, is left open. For such a reason, it is crucial from the infrastructure provider's perspective to assess the additional overhead introduced in the network by RENEV. In the following, we theoretically derive the number of signaling messages exchanged during RENEV operation, as well as the expression for the percentage of successful resources' transfer requests.

Given the system model presented in Section 3.6.1 and the nomenclature used in Section 3.4.4, each BS may be characterized by the number of RBs initially allocated to it as well as the number of used/unused RBs for a particular number of users. Thus, let us define, the number of available resources for a specific BS_i as $r_i = RB_i - u_i$, where RB_i are the RBs initially allocated to BS_i and u_i is the number of RBs required to serve the demand of the users associated to BS_i. The number of required resources, u_i, will be upper and lower bounded as a function of the number of users connected to BS_i, their traffic demand and their received SNR. Therefore, $u_i \in [u_{i,min}, u_{i,max}]$, where $u_{i,min}$ and $u_{i,max}$ are the numbers of RBs being required when all the UEs associated to BS_i use 64QAM$\frac{4}{5}$ (i.e., the maximum throughput per RB) and QPSK$\frac{1}{8}$ (i.e., the minimum throughput per RB) respectively. Based on these definitions, the upper and lower bounds of available resources for the set of BSs of the described system can be defined as $r_{min} = \min_{0 \leq i \leq N}(RB_i - u_{i,max})$ and $r_{max} = \max_{0 \leq i \leq N}(RB_i - u_{i,min})$.

In this context, the system is defined by the set of possible initial states $\mathcal{S} = \{S_1, S_2, \ldots, S_W\}$ and the set of probabilities of occurrence of each state $\pi = \{\pi_1, \pi_2, \ldots, \pi_W\}$, where W stands for the number of possible states. In turn, each state is defined as $S_j = (s_{j,1}, s_{j,2}, \ldots, s_{j,r_{max}-r_{min}+1})$, where $s_{j,k} \in \mathbb{N}_0$ denotes the sum of BSs with a number of available resources equal to $(r_{min} - 1 + k)$ and $S_j \in \mathcal{S}$. As in RENEV the BSs first seek for resources in the SCs tier and subsequently in the eNB, we decouple the analysis into these two steps. Focusing first on the SCs tier (without considering the resources in the eNB), the system may be defined by the set of possible initial states $\mathcal{S}$ and the probability of occurrence π. By definition, $\sum_{k=1}^{r_{max}-r_{min}+1} s_{j,k} = N$. According to the definitions stated above, the number of Requesting BSs in a given state S_j, will be equal

to the number of BSs with negative r_i, also expressed as $n_R(S_j) = \sum_{k=1}^{-r_{min}} s_{jk}$.Therefore, the expected number of Requesting BSs may be written as

$$\mathbb{E}[n_R] = \sum_{j=1}^{W} n_R(S_j) \cdot \pi_j. \tag{3.9}$$

After the operation of RENEV in the SCs tier, the available resources of the Donor BSs will have been transferred to the Requesting BSs to cover their needs. Consequently, the probability of having the system in a particular state S_j after executing RENEV will vary. If we denote by π'_j the probability of being in the state S_j after the RENEV completion in the SCs tier, it holds,

$$\pi'_j = \sum_{n=1}^{W} \pi_n \cdot p_{nj}, \tag{3.10}$$

where p_{nj} is the probability of transiting from state S_n to S_j. Note that not all transitions are feasible since the redistribution of resources among SCs imposes some restrictions. Thus, $p_{nj} \neq 0$ if and only if S_j is contained in the set of feasible future states of S_n, i.e., $S_j \in \mathcal{F}(S_n)$. The detailed definition of $\mathcal{F}(S_n)$, according to the conditions that should hold to satisfy that $S_j \in \mathcal{F}(S_n)$, is introduced in Appendix A.3. Hence, the transition probability, is given by

$$p_{nj} = \begin{cases} 1 & : j = n, \mathcal{F}(S_n) = \emptyset, \\ \frac{1}{|\mathcal{F}(S_n)|} & : j \neq n, S_j \in \mathcal{F}(S_n), \\ 0 & : \text{otherwise}, \end{cases} \tag{3.11}$$

where $|\mathcal{F}(S_n)|$ is the cardinality of the set $\mathcal{F}(S_n)$. Although the SCs tier is the first alternative for RENEV to reallocate the existing resources, not all requests can be covered with the resources of this tier. Thus, and according to (3.9), the expected number of successful requests (i.e., when the needs of the Requesting BSs are covered by the unused resources of the Donor BSs) in the SCs tier may be calculated as

$$\mathbb{E}[n_s] = \sum_{j=1}^{W} n_R(S_j) \cdot [\pi_j - \pi'_j]. \tag{3.12}$$

As RENEV is completed in the SCs tier, all feasible redistribution of resources has been successfully conducted, and the system is found in state $S_j \in \mathcal{S}$, with probabilities π'. However, note that S_j characterizes the scenario without taking into account the resources available in the eNB, i.e., r_0. Therefore, in the second step of the signaling analysis a new set of states, namely $\mathcal{S}''$, must be defined to include r_0. It should be noted that the r_0 resources inserted into the system, may be distributed in different ways. For

instance, if all Requesting BSs are overlapped among them, the new resources will be transferred to the SCs tier only once. Conversely, if not all Requesting BSs overlap with the rest of the Requesting BSs, the r_0 resources will be transferred more than once. Therefore, if we define the number of non-overlapping groups of Requesting BSs as $Q = \{1, 2 \cdots M\}$, where M stands for the number of Requesting BSs (for instance, for S_j we have $M = n_R(S_j)$, the r_0 resources can be transferred to the SCs tier Q times. Thus, for a specific state S_j containing M Requesting BSs, the inclusion of the r_0 resources from the eNB can lead to M possible new states. Specifically, a state S_j results in M new states defined as $S_t'' = (s_{t,1}'', s_{t,2}'', \ldots, s_{t,k}'', \ldots, s_{t,r_{max}-r_{min}+1}'')$, with $s_{t,k}'' = s_{j,k} + Q$ for $k = r_0 - r_{min} + 1$ and $Q = \{1, 2 \cdots M\}$, and $s_{t,k}'' = s_{j,k}$ otherwise. This set of new states is defined for each value of r_0. Therefore, after the inclusion of the resources available in the eNB the system may be described by the set of new possible initial states $\mathcal{S}'' = \{S_1'', S_2'', \ldots, S_L''\}$ and the probability of being initially in these states $\pi'' = \{\pi_1'', \pi_2'', \ldots, \pi_L''\}$, where L stands for the number of possible states. Thus, it holds that

$$\pi_t'' = \pi_j' \cdot P(Q = q|N, M) \cdot P_{eNB}(r_0), \tag{3.13}$$

where $P(Q = q|N, M)$ is the probability of having q non-overlapping groups in a cluster with M Requesting BSs out of N BSs (calculated in Appendix A.4) and $P_{eNB}(r_0)$ is the probability that the eNB has r_0 spare RBs that could be transferred. For a given scenario, the latter is a random variable that depends on the resources allocated to the eNB, the number of users and the traffic demand of each user.

Henceforth, we use the same calculation method that we used for the SCs tier to derive the expected number of successful requests. Firstly, the expected number of Requesting BSs is calculated as in (3.9), using the new probabilities of occurrence π'', denoted as $\mathbb{E}[n_R'] = \sum_{j=1}^{L} n_R(S_j'') \cdot \pi_j''$. After the application of RENEV, the available resources of the eNB will have been transferred to the Requesting BSs. The new transition probabilities from state S_n'' to S_j'' for this phase, according to (3.10), will be equal to $\pi_j''' = \sum_{n=1}^{L} \pi_n'' \cdot p_{nj}'$, where p_{nj}' is calculated with (3.11) and the set of feasible future states $\mathcal{F}(S_n'')$ according to Appendix A.3. Under the conditions stated above, it cannot be assured that all requests can be covered with the resources of the eNB tier. Thus, the expected number of successful requests in the eNB tier may be calculated as $\mathbb{E}[n_s'] = \sum_{j=1}^{L} n_R(S_j'') \cdot [\pi_j'' - \pi_j''']$. Therefore the total expected number of successful requests by both tiers after the completion of RENEV is equal to $\mathbb{E}[n_{s_{total}}] = \mathbb{E}[n_s] + \mathbb{E}[n_s']$, and the probability of successful requests is calculated as $(\frac{\mathbb{E}[n_{s_{total}}]}{\mathbb{E}[n_R]})$. The number of signaling messages exchanged by the BSs depends on the total number of BSs (i.e., $N + 1$), the number of Requesting BSs, and the number of successful requests. In particular, and by observing

Fig. 3.2 and Fig. 3.3, it can be noticed that all Requesting BSs (whose number is in average equal to $\mathbb{E}[n_R]$) exchange 3 messages (messages 1, 2 and 3) with the rest of the $N-1$ SCs. Additionally, the Requesting BSs not being able to obtain resources from the SCs tier (whose number is in average $\mathbb{E}[n'_R]$) exchange the aforementioned three messages with the eNB. Finally, if any of the requests is successful, the Requesting BSs exchange 2 messages (messages 4 and 5 in Fig. 3.2 and Fig. 3.3). Therefore, the expected number of signaling messages exchanged by RENEV may be expressed as: $\mathcal{I} = 3 \cdot (N-1) \cdot \mathbb{E}[n_R] + 3 \cdot \mathbb{E}[n'_R] + 2 \cdot \mathbb{E}[n_{s_{total}}]$.

3.8 Performance Evaluation

3.8.1 RENEV in Small Cell Deployment

3.8.1.1 Simulation Scenario and Parameters

We consider 3GPP HeNB settings for the setup of SC network [62]. The system transmission bandwidth equals 20MHz, corresponding to 100 RBs, and the transmission mode is Single Input Single Output (SISO). We assume that there are 6 SCs not uniformly distributed in an area of 100m $\times$ 100m, belonging to one service provider and one network operator. The propagation path loss in the small cell network is given by the following path loss model:

$$L_{dB} = 37 + 30log_{10}(d) + 18.3f^{(\frac{f+2}{f+1}-0.46)}, \tag{3.14}$$

where d is the distance in meters from the antenna and f is the number of penetrated floors in the propagation path [66, 75, 76]. We assume a dense wireless environment, such as an outdoor urban area, where there are less penetrated walls and floors and therefore we consider $f = 3$ in our work [66]. We assume that the total transmit power including the antenna gain of each HeNB is 32dBm. Shadow fading is modeled as random variable with log-normal distribution of 0 mean and standard deviation 8dB [66]. The received noise power is the one of an Additive White Gaussian Noise (AWGN) channel. In the following, we provide results for a SC network deployment, where the SCs are deployed randomly.

We further consider an open access small cell network and we assume that the downlink transmitted power is fixed and the same for all the HeNBs. In particular such consideration is efficient in 3GPP LTE-A where the same amount of power is transmitted on all RBs and there is no or very limited power control in the downlink [67].

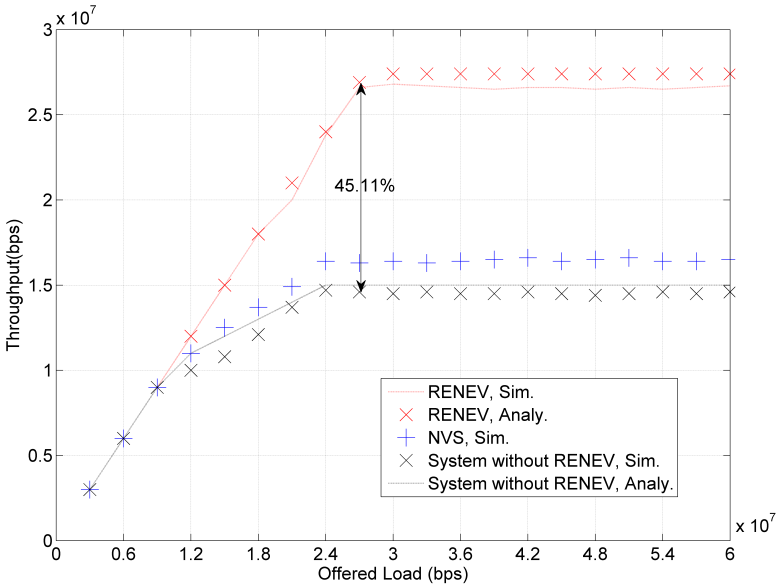


FIGURE 3.4: Aggregate System Throughput for different offered loads in SC tier.

3.8.1.2 Network Performance

In this set of experiments, we compare the system with and without the application of RENEV, to illustrate the benefits gained in terms of network's aggregate throughput. We also compare it with NVS, another framework presented in works [37] and [38] of the state of the art, that opportunistically allocates the unused resources among the existing slices in a BS. We adapt this framework to our scenario creating distinct slices within one tier (i.e., SCs), each one accommodating a certain percentage of the overall RBs that can serve a specific number of users. All users inserted into the topology, download files using File Transfer Protocol (FTP) at an average data rate of 300Kbps in the downlink. Following common practice in commercial cellular networks, FTP requests are always admitted regardless of the system's load conditions.

In Fig. 3.4 the aggregate system throughput is shown with respect to an increasing offered traffic load for the system with and without the application of RENEV as well as for the NVS framework. For low offered load, up to 9Mb/s, the system's behavior is the same; the users' demanded traffic is served in all the cases. However, as the load increases, the system without the application of RENEV is able to serve less traffic load, compared to the system where the algorithm is applied. When saturation is reached (i.e. when the offered load equals 27Mb/s) the achieved throughput raises 45.11%. This could be explained by the fact that in the first case, the available resources are distributed among the group of the participants HeNBs in order to cover the maximum of the users' traffic demand.

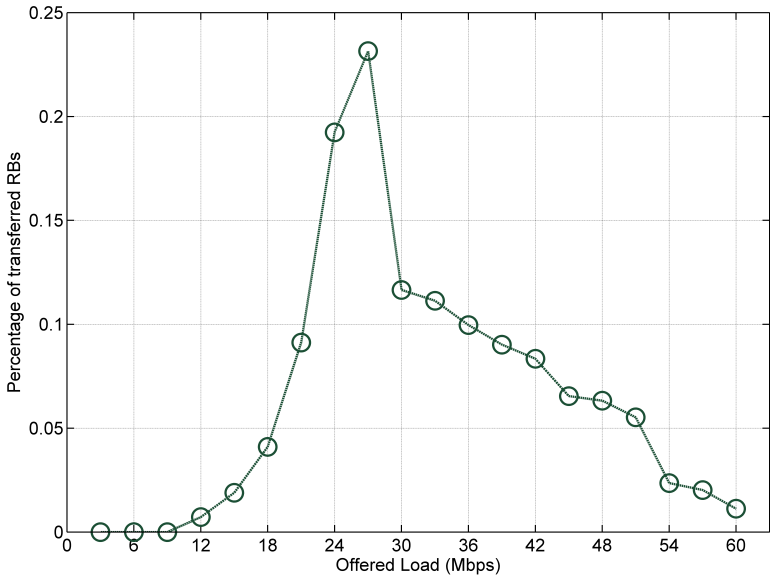


FIGURE 3.5: Percentage of transferred Resource Blocks in SC tier.

In the case where RENEV is not applied, each HeNB controls its own resources and after a while these resources are depleted. System saturation is reached in the case where more load is introduced but the existing RBs are depleted and consequently no more users can be served. NVS reaches higher system throughput than the system without the application of RENEV, due to its capability to allocate the free resources in the slices that contain users that need it. Since here one type of traffic and fixed percentage of resources among the slices are introduced, this solution restricts the number of the resources that are transferred. We could consider that RENEV adds one more dimension to the vision of NVS in order to achieve virtualization in LTE-A environments. NVS is based on the heterogeneity of services and RENEV on the idea of resources transferring in HeNBs that do not own specific percentage of resources to transfer. So, RENEV takes advantage of this capability adding one more degree of flexibility in the resource transferring among the existing flows that share one or more physical BSs.

Figure 3.5 depicts the percentage of the transferred RBs versus the offered load during the application of RENEV. When the demanded traffic reaches 27Mb/s the system requires the highest number of RBs in order to satisfy the existing users; 23.15% of the total RBs belonging to the system are transferred. After this point, although the number of users that require resources is augmented, the number of transferred resources decreases because the system runs out of resources since all of them are already allocated to the existing users. HeNBs are only capable of transferring resources to other HeNBs when they have unused resources. Accordingly, when the offered traffic grows, the possibility of transferring resources to other cells falls. During the application of

RENEV, as the offered load increases the HeNBs request more RBs. However, after a certain point, the successful RB transfer decreases.

RENEV is a decentralized proposal for transferring resources among several HeNBs. This means that when the offered load is augmented, the total number of resources is distributed among the users according to their requirements dynamically as they come from a common pool. This leads to a peer-to-peer common RRC scheduling between the participants HeNBs and also a common control plane for the RAN nodes. With the use of RENEV, all the system's resources are dynamically used according to the users' needs on an isolated and on-demand basis. In this way, the majority of the users is served, as long as spare resources exist. In any other case, the users would receive lower quality of service or they could not get even served at all.

3.8.2 RENEV in HetNet Deployment

3.8.2.1 Simulation Scenario and Parameters

In this environment our simulation consists of an eNB overlaid with a cluster of outdoor HeNBs-LTE femtocells [75] [76], operating on the same carrier [17, 18]. The number of SC clusters per eNB coverage area varies from 1 to optional 4 and the number of SCs per cluster can vary from 1 to 10 depending on the deployment [17]. We choose 6 outdoor HeNBs-LTE femtocells [75] to provide results for the system throughput.

We conduct Monte-Carlo extensive simulations (i.e., a thousand iterations to achieve statistical validity) in a custom made tool implemented in MATLAB®, using random deployments of a SCs cluster placed within eNB coverage area. In each iteration users are distributed independently and non uniformly; i.e., 2/3 of the traffic is dropped within the SC tier [17, 18]. The simulation parameters are listed in Table 3.1; the 3GPP related parameter values are based on [66]. The overall system bandwidth consists of 2 bands of 20 MHz, operating at 2 GHz, each one assigned to each tier using CA. Packet scheduling is proportional fair both at eNB and SCs. We conduct simulations for a full buffer traffic model [17]. In this scenario we assume that users download files using FTP at an average data rate of 300 Kbps.

As discussed in the previous sections, RENEV is a complementary virtualization solution implementable on top of existing solutions. Hence, in the scenario under consideration both NVS [38] and PRR [39] are simulated with and without RENEV. NVS creates distinct slices of spectrum in each particular BS. These slices accommodate equal percentage of the overall RBs, each one residing in a specific traffic flow. PRR framework, guarantees a minimum number of RBs per subframe on average for each traffic flow,

TABLE 3.1: Basic System Parameters used in the HetNet Simulation, RENEV

Parameters	**Settings/Assumptions [66]**
Network layout	Cluster of 6 HeNB LTE Femtocells randomly placed per Macrocell
Inter-site distance/cell radius	Macrocell: 500 m (ISD) Femtocell: 25 m (Cell radius)
Transmit power	Macrocell: 46 dBm, Femtocell: 17 dBm
Bandwidth	20 MHz at 2 GHz for each tier
Path loss	Macrocell: $140.7 + 36.7log10(\text{R[km]})$ Femtocell: $128.1 + 37.6log10(\text{R[km]})$
Shadow fading	Lognormal, $\mu = 0$, std.=8 dB for Macrocell Lognormal, $\mu = 0$, std.=10 dB for Femtocell

which is available when a particular flow wants to use it (i.e., reserved part). The portion of system resources remaining after subtracting the reserved part at each BS, is called shared part and it can be used by any incoming traffic flow. According to [39], an operator requires at least a minimum portion of resources to be reserved for its users within a BS, in order to guarantee QoS for particular traffic slices. In simulations, for users downloading FTP files this percentage is set to 50% [39] corresponding to the scheme named PRR 50%. Although setting a high value for shared part within a BS can lead to more flexible allocation of resources, it comes with the shortcoming of not covering the minimum requirements for QoS imposed by operators. However, we use this maximum degree of flexibility in PRR, having 0% RBs reserved part and 100% shared within each BS (i.e., "PRR 100%"), to calculate the theoretical upper bound of the aggregate throughput.

3.8.2.2 Network Performance

Fig. 3.6 presents the aggregate system throughput (a metric indicated by 3GPP in [18, 66]) with respect to an increasing offered traffic load for NVS as well as PRR 50% and PRR 100% with and without RENEV. As it may be observed, the experimental and theoretical curves for PRR 100% and RENEV+PRR 100% (the upper bound expressions as derived in Section 3.6) match. For offered load equal to 18 Mb/s, the system's behavior is the same for all the depicted schemes; all demanded traffic is served. However, as the load increases all compared schemes are able to serve less users compared to the system where RENEV is applied.

In particular, when saturation is reached due to a lack of resources (i.e., offered load equals 78 Mb/s), the throughput achieved with RENEV + PRR 100% (60.93 Mb/s) represents an increase of 50.68% with respect to PRR 100%. In the first case, the available resources of two tiers are distributed according to traffic demand to cover

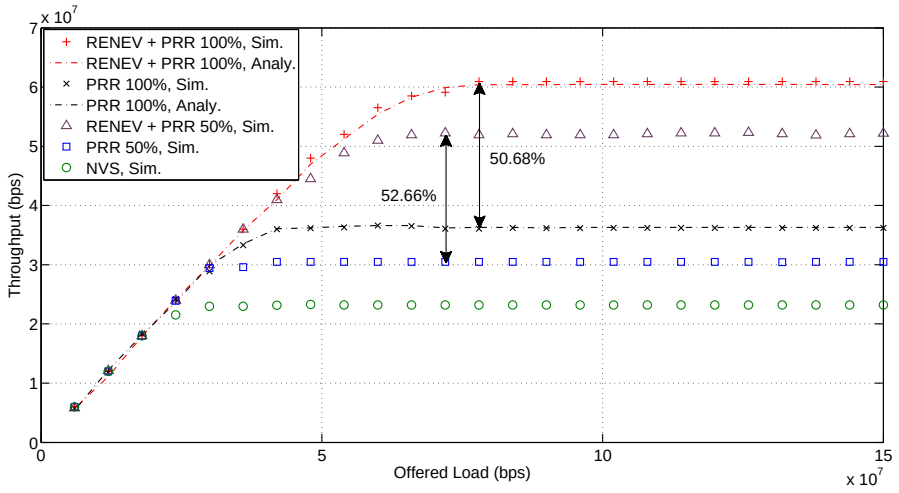


FIGURE 3.6: Aggregate System Throughput for different number of Offered Loads in HetNet.

the maximum number of users' needs; however when RENEV is not applied, each BS manages its own resources which are depleted after a while. At the other extreme, the NVS scheme achieves the poorest performance, since resources from different slices cannot be shared regardless of the traffic demands in each slice. The maximum value in this case is 23.19 Mb/s. As for PRR 50%, with and without RENEV, its performance constitutes an intermediate situation.

Notwithstanding the good results offered by PRR 100% compared to PRR 50% (both of them without the application of RENEV), the authors in [39] expound that a minimum share of the available resources should be reserved for each traffic slice to guarantee minimum QoS requirements. Therefore, PRR 100% is not convenient in terms QoS despite outperforming PRR 50% in terms of aggregate throughput. The same conclusion applies when RENEV is implemented. By inspecting Fig. 3.6, it is particularly worth noting that RENEV + PRR 50% (which does not degrade the QoS requirements of the traffic slices) is able to show higher aggregate throughput than PRR 100%. This behavior is due to the ability of RENEV to compensate not only the traffic spatial non-uniformities but also the QoS loss experienced when sharing the 50% of the resources per BS, instead of the 100% in PRR.

Figures 3.7(a) and 3.7(b) study the percentage of transferred RBs per tier as well as the corresponding served traffic for the case of RENEV + PRR 100% (as depicted in Fig. 3.6). As expected, we observe that the RB transfer first increases, then reaches a specific peak and then decreases for both tiers. The two peaks in Fig. 3.7(a) equal 32.2% of transferred RBs by the SCs tier (achieved for 60 Mb/s) and 32.64% by the eNB (achieved for 78 Mb/s). After these peaks, although the number of users requiring resources is augmented, the transferred resources decrease because both tiers run out of

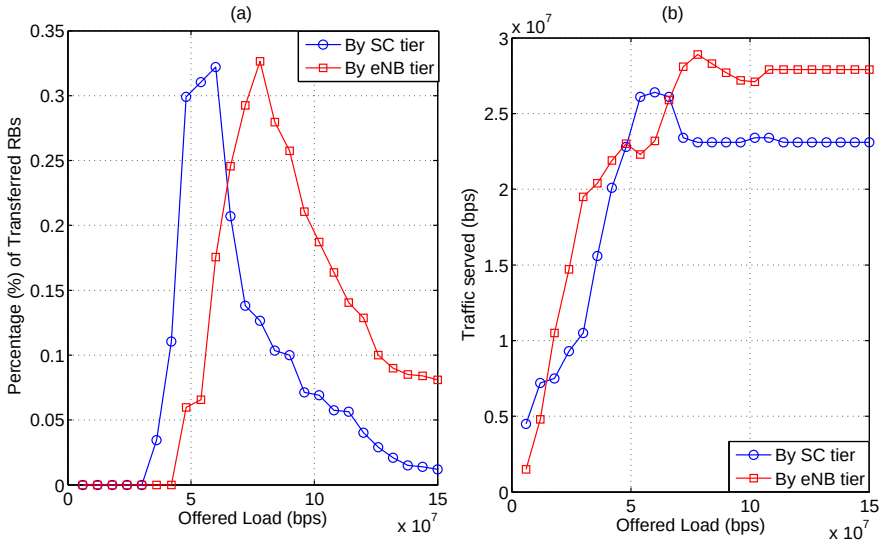


FIGURE 3.7: (a) Percentage of transferred RBs by each tier. (b) Traffic Served by each tier in HetNet.

RBs since all of them are already allocated to the existing users. It is worth noting that the traffic served by each tier (Fig. 3.7(b)) depends on the available number of RBs. In particular, when the percentage of transferred RBs falls, the aggregate throughput in Fig. 3.6 stabilizes since the resources are depleted and the incoming user requests cannot be satisfied.

Finally, when applying RENEV, the resources are provided to the tenant Requesting BSs first by the SCs tier and subsequently, when none of the SCs is able to provide resources, by the eNB, that acts as a donor BS. For this reason, we may observe fluctuation points for the served traffic, among 50 Mb/s and 100 Mb/s in Fig. 3.7(b). In particular, for low traffic load, most transfer of resources is conducted among the SCs. Progressively, as offered traffic increases, it is less probable that SCs provide additional resources. Thus the eNB starts transferring resources to the Requesting BSs. When the maximum load is achieved in the SCs tier (i.e., 60 Mb/s), the probability of finding a Donor BS within this tier falls. On the same time, the eNB (which is still less loaded than the SCs) keeps increasing the percentage of transferred resources till 78 Mb/s. At this point, the eNB is also loaded and the probability of transferring to Requesting BSs decreases. This is translated into the served traffic; the traffic served by the SCs grows thanks to the transfer of resources from the SCs tier and from the eNB. However, when the transfer of resources by the SC tier falls, the increase of eNB transfer of RBs cannot compensate it and the traffic served by the SCs tier decreases. Due to high load in the eNB tier as well, when the transfer of resources decreases, the traffic tends to stabilize to the maximum traffic that can be served by the SCs tier without the transfer of resources. The served traffic by the eNB is also stabilized to the maximum value that can be served by it.

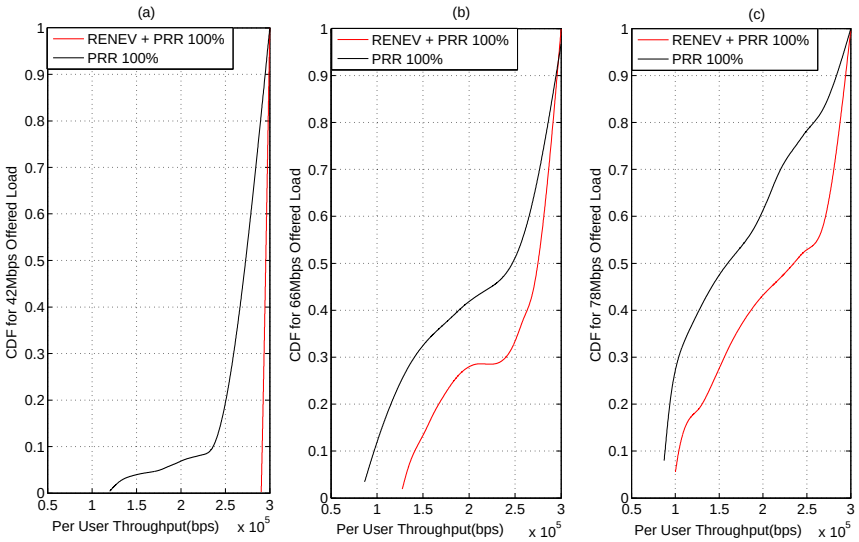
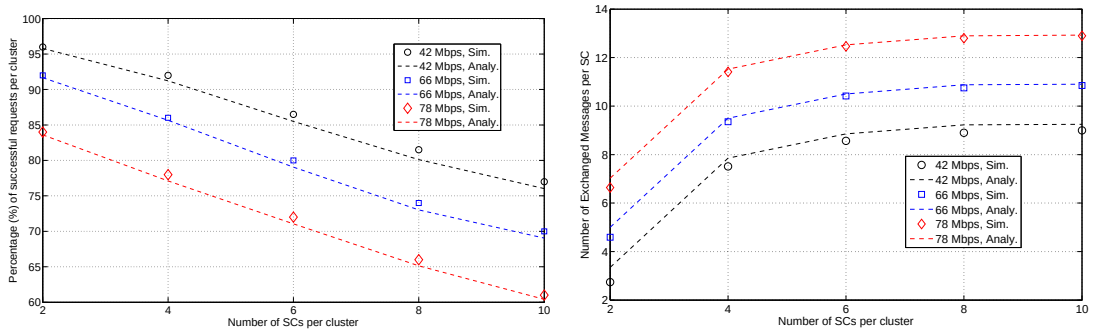


FIGURE 3.8: CDF of user Throughput in HetNet for (a) 42Mb/s, (b) 66Mb/s and (c) 78Mb/s Offered Load.

3.8.2.3 User's Throughput

In Figures 3.8(a), 3.8(b) and 3.8(c) we study the Cumulative Distribution Function (CDF) of user throughput (indicated metric in [18, 66]) for three cases of traffic load: low offered load where the majority of users are served, medium one and the case where the system is saturated; 42 Mb/s, 66 Mb/s and 78 Mb/s correspondingly, as also depicted in Fig. 3.6. In the sequel focus on the scenario with PRR 100% with and without RENEV, since it provides the upper bounds of network's throughput. First, we observe that the gains in throughput acquired in the network side with the application of RENEV, can be translated into merits for the end users. According to Fig. 3.8(a), as the offered load is low, RENEV is able to help the majority of users to achieve the demanded data rate. In particular, the observed slight deviation from 300 Kbps, is due to the fact that some users do not achieve the demanded data rate because of the channel conditions that they experience. However, without applying RENEV the user throughput dispersion is quite high. For instance, 80% of the users achieve throughput values equal or higher to 250 Kbps. The rest 20% of the users achieve values ranging from 120 Kbps to 250 Kbps.

In addition, we observe that higher offered load affects dramatically the user throughput. For example, in Fig. 3.8(b), 72% of the users achieve transmission rate equal or higher than 250 Kbps when RENEV is applied. On the other hand for the same percentage without applying RENEV the lowest user throughput achieved is 130 Kbps. In particular, the transfer of resources defined by RENEV, improves the performance of users with poor links, who are normally located in the cell edge area.



(a) (b)

FIGURE 3.9: (a) Percentage of successful requests for different number of SCs per cluster. (b) Number of exchanged X2 messages per SC in HetNet.

These users are more demanding in terms of required RBs. However, RENEV is able to satisfy such kind of users. For instance, when the system is further loaded (Fig. 3.8(c)) the dispersion among user throughput is quite high, both with and without RENEV. Even in this study case, 50% of the overall users achieve 75% of the demanded transmission rate (with lowest user throughput equal to 102 Kbps). On the contrary, without RENEV, this percentage falls to 52.5% of the demanded data rate.

3.8.2.4 Signaling Overhead

In this set of our experiments, we evaluate the requests and the corresponding messages that are necessary for the transition from a scenario where all resources are initially distributed uniformly among the BSs, to a scenario where the resources are finally distributed according to the existing geographical traffic variations (i.e., upper bound values).

In Fig. 3.9(a) we study the impact of the number of SCs into the percentage of successful requests per cluster, for different traffic offered loads (low, medium, and high as in Fig. 3.8). It is worth noting that in dense scenarios in terms of SCs, the available RBs are quickly depleted, and therefore, the number of successful requests falls. This means that the tenant Requesting BSs cannot attain the demanded resources. For high loaded systems less requests are satisfied since resources are exhausted faster. For example, if a cluster with 6 SCs is considered (scenario analyzed in Fig. 3.6), the percentage of successful requests is 86.5% for 42 Mb/s offered load, 80% for 66 Mb/s and 72% for 78% Mb/s. On the other hand, when 10 SCs are considered within the cluster's surface, this percentage falls to 77%, 70% and 61%, respectively.

Fig. 3.9(b) studies the number of exchanged messages per SC, for the three studied offered loads. In all cases, the experimental results showcase that higher number of

SCs within the cluster, is translated into higher number of exchanged messages over X2 interface. For instance, for a cluster with 6 SCs, we observe in average 8.5 exchanged messages for 42 Mb/s, 10.4 for 66 Mb/s and 12.4 for 78 Mb/s. In particular, as the number of SCs in a cluster increases, the messages among the participant tenant BSs are also increasing even though the rate of increase progressively reduces.

The physical implementation of X2 is still not standardized, so it should be noted that it is the main factor imposing feasibility constraints. In general we note that a particular number of SCs where RENEV can be applied depends on the limits inserted of the actual implementation of X2 and the corresponding capacity reserved for signaling. Fig. 3.9(b) can result quite useful for operators, to calculate the actual signaling for a certain number of SCs per cluster, according to the way they choose to implement X2 (i.e., such as fiber, over-the-air wireless, etc.).

3.9 Conclusion

In this chapter, we proposed RENEV; a scheme that considers the coordination among several BSs (i.e., either belonging to one or multiple tiers) to create an abstraction of systems' radio resources, so that multiple tenants (i.e., BSs) can be served, in a heterogeneous environment. The extensive performance assessment has revealed that gains in system's throughput are translated into gains for the users' throughput as well. With the use of RENEV, system's resources are dynamically distributed according to users' needs on an isolated and on-demand basis. In this way, the majority of the users is served, as long as spare resources exist. Finally, the solution has been evaluated for the signaling overhead that adds into the network for increasing number of SCs per cluster.

Chapter 4

A Capacity Broker Framework for Multi-tenant LTE-A Networks

4.1 Introduction

Mobile communications are entering a new era with the popularity of portable electronic devices, which gave rise to a plethora of new services with ever-increasing resource demands. Lately, Mobile Network Operators' (MNOs) revenues cannot keep pace, considering the cost to operate and upgrade their infrastructure. To date, operational observations show that there are underutilized resources, e.g., 50% of sites carry traffic that yields less than 10% of revenue [77]. Network sharing has been proposed to allocate these underutilized resources among Mobile Virtual Network Operators (MVNOs), providing another revenue source for MNOs. Studies have shown that it can recover up to 20% of operational costs for typical European MNOs and significantly reduce capital expenditures in developing countries (e.g., up to 70% in India) [78].

There are still many challenges to overcome, to achieve a viable network sharing business model appealing to MNOs. First, network sharing should be performed on demand, with resources acquired in the scale of minutes, while allocations are configured via signaling. A centralized resource management entity should facilitate this process. Its role is to assist the MNO owning a shared RAN (i.e., infrastructure provider), to fully exploit the unused capacity. The notion of this entity, referred to as capacity broker, has been introduced in the 3rd Generation Partnership Project (3GPP), from a business perspective [79]. Such a central entity is required to assure synchronization in resource sharing for such short-time scales, while satisfying Service Level Agreements (SLAs). Nevertheless, its integration into the 3GPP management architecture [80] is an open issue. In addition, a key question is how to exploit the functionality of capacity broker

to accomplish an efficient resource allocation, by considering: (i) the global view of network resource utilization, and (ii) the knowledge of the expected traffic volumes, a challenging task due to lack of periodicity in short-term scale. Although many interesting studies on capacity slicing have been carried out, either they study the problem from different layer, or they introduce non-backwards compatible centralized entities with the existing 3GPP architecture.

To that end, our second contribution presented in this chapter concentrates on facilitating resource provisioning between MVNOs, by integrating the capacity broker in the 3GPP network management architecture with a minimum set of enhancements. Furthermore, to fully exploit its range of capabilities, we propose the Multi-tenant Slicing (MuSli) of capacity framework for on-demand resource allocation considering two types of traffic: (i) Guaranteed Quality of Service (QoS) with resources locked for explicit use by a MVNO and (ii) Best-Effort (BE) where resources are pooled and shared by all participants. To accomplish this, we follow a two-step approach: (i) we improve short-term forecasting techniques by extracting traffic variation trends and facilitate the capacity broker with accurate information regarding the expected traffic and (ii) we propose how to slice the available resources into these two types of traffic classes, depending on the forecasting and its respective accuracy.

The remainder of the chapter is structured as follows. The related work is presented in Section 4.2. In Section 4.3 we explain how the capacity broker is integrated in the 3GPP management architecture. Section 4.4 introduces the system model along with the MuSli framework. Section 4.5 analyzes the simulation set-up and the evaluation results. Finally, Section 4.6 concludes the chapter.

4.2 State of the Art

In this section, we provide a brief literature review of the related work. The initial adoption of network sharing in 3GPP, concentrated on passive solutions, wherein MNOs share base station sites, antennas, etc. Active sharing that followed, enabled operators to share network resources for long term periods according to contractual agreements. For active network sharing, 3GPP has specified two architectures in [81]: (i) the Multi-Operator Core Network (MOCN) and (ii) the Gateway Core Network (GWCN). In the former, each operator is sharing eNBs connected to core network elements belonging to each MNO using a separate S1 interface. In the latter, operators share additionally the Mobility Management Entity (MME). Our proposal is compatible with both 3GPP network sharing architectures, while introducing on-demand resource allocation via the means of signaling extensions of 3GPP network sharing management [80].

A preliminary approach for virtualizing an eNB is introduced in [82], by detailing the notion of hypervisor, that performs resource sharing among MNOs considering radio conditions, contracts and traffic load. In advancing the basic eNB virtualization, [38] introduces the Network Virtualization Substrate (NVS) that operates closely to the MAC scheduler. A tailored mixture of reserved and shared resources with respect to NVS component is proposed in [39], in order to flexibly allocate shared resources modifying the MAC scheduler. In this work, we adopt such NVS two-step process, but instead of concentrating on the MAC scheduler for performing resource differentiation, we leverage the capacity broker to provide different resource slices based on the expected traffic volume.

A study adopting the capacity broker paradigm in LTE is detailed in [83], regarding a range of capacity and spectrum sharing options. Unlike such an approach that introduces a new control plane interface to coordinate sharing agreements, our proposal is backwards compatible with the existing 3GPP network management architecture, reusing current interfaces, while introducing a minimum set of enhancements.

The accuracy of short-term load forecasts can significantly affect the capacity broker decisions for resource slicing. A wide range of solutions for short-term load forecasting have been reported in the literature [84], which can be distinguished in two categories. The first one employs characteristics of traffic loads, such as spatial/temporal relevance or self-similarity [85]. The second category employs techniques, such as exponential smoothing to study the intrinsic dimensionality [86], Kalman filtering to capture the evolution of traffic [87] or modern signal processing techniques such as compressive sensing [88]. In this chapter, we investigate which of the above methods fits best the capacity broker paradigm and we provide a set of enhancements, to compensate the lack of periodicity and non-uniformities of a short-term prediction.

4.3 3GPP Network Sharing Management Architecture

The overview of the 3GPP network sharing management architecture [80], in which we integrate the capacity broker and execute MuSli, is depicted in Fig. 4.1. The Master Operator-Network Manager (MO-NM) monitors the shared network via the Master Operator-Shared RAN-Domain Manager (MO-SR-DM) using Type 2 (i.e., Itf-N) interface. In turn, the latter communicates with a set of shared base stations, via Type 1 (i.e., Itf-B) interface. All radio-related functions (i.e., Radio Resource Management, connectivity to core network etc.) take place in the level of the shared base stations. In addition, MO-NM enables the Sharing Operator-Network Manager (SO-NM), to monitor and control the allocated resources to MVNOs via Type 5 interface.

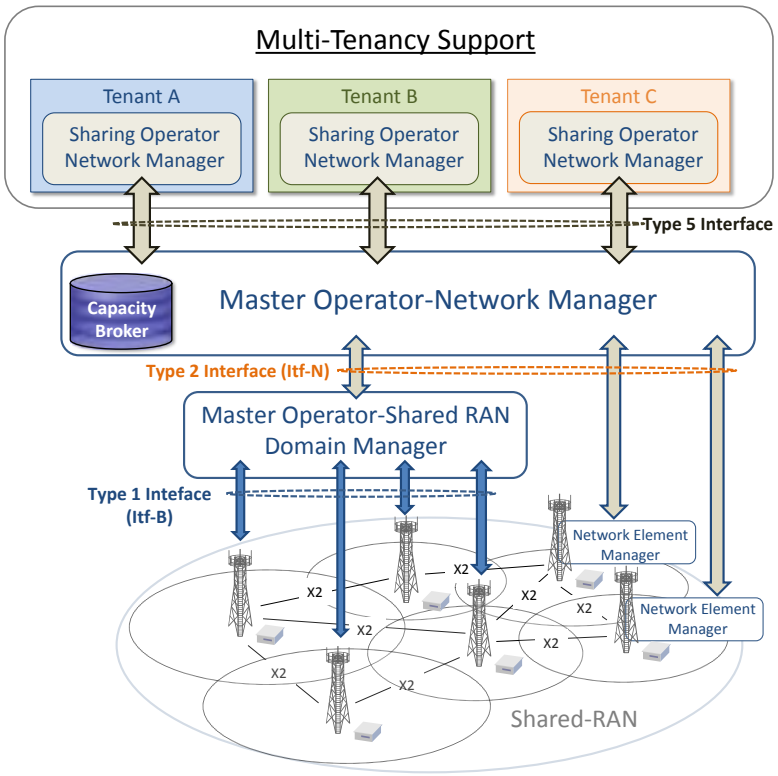


FIGURE 4.1: Capacity Broker in 3GPP Network Sharing Management Architecture.

Given the existing architecture, we propose to place the capacity broker on the MO-NM, to facilitate the allocation of shareable resources, by automatic means and on an on-demand basis, to MVNOs. The capacity broker, by deciding which requests will be accepted, assures synchronization in resource sharing for short-time scales, while satisfying their SLAs. Thus, when co-locating it at MO-NM, it has rapid access to network monitoring information (such as Uplink/Downlink load and performance measurements), as well as to network planning information (i.e., MO-NM has collected this from MO-SR-DM). Then, the MO-NM uses the output of the capacity broker to inform the MO-SR-DM about which specific requests should be accepted and the shared base stations implement their respective radio-related functions. Our proposal requires extensions to Type 1, Type 2 and Type 5 interfaces. Type 1 and Type 2 need to accommodate the tenant identification (i.e., PLMN-id), resource allocation (e.g., Resource Blocks (RBs)), start time and duration of the request. In addition, Type 5, which is typically established upon an agreement, should include the list of MNO's cells involved in the capacity slicing process. All the above interfaces should support resource measurements and performance monitoring per MVNO. To that end, we introduce the PLMN-id within each corresponding packet. For the portion of pooled resources, monitoring information should be shared among all tenants' SO-NM systems.

4.4 Multi-tenant Resource Slicing Framework

This section concentrates on elaborating a resource management framework, called Multi-tenant Slicing (MuSli), to be executed in the capacity broker in coarse time-scales. Its objective is performing resource slicing among incoming requests considering two different traffic classes: guaranteed QoS and BE. The difference between the two aforementioned traffic classes lies in their distinct requirements in terms of radio resources. Thus, whereas guaranteed QoS traffic (usually identified with services such as voice) is characterized by a fixed transmission rate, BE traffic (identified, for instance, with data services) is defined in terms of average demanded data rate as well as more relaxed delay constraints.

In this scenario the management of the shared RAN resources, conducted by the capacity broker, has to deal with two main hurdles: i) the diversity of the traffic requests, and ii) the varying nature of the radio interface. Our methodology consists in using a forecasting procedure to predict the traffic volume in near future for all MVNOs considering the entire deployment and allocating resources with different quality to different traffic classes (e.g., for voice and data).

4.4.1 System Model

Let us define a scenario composed of a set of MVNOs, $\mathcal{V} = \{i : i = 0, \cdots, V\}$ sharing a single RAN. For the sake of simplicity, and without loss of generality, we assume hereafter that MVNO 0 is the owner of the shared RAN. The capacity broker (described in Section 4.3) decides whether to accept or reject the incoming MVNOs' requests. Thus, it manages the shared RAN capacity to serve the capacity requests generated by the MVNOs in $\mathcal{V}$. In this context, the appropriate management of the available capacity is a twofold problem. First, the future capacity usage must be forecasted, and secondly the available expected capacity must be allocated to the set of received requests. According to the described traffic classes, the r^{th} request of the i^{th} MVNO can be defined as $g_{i,r}\{t_{i,r}, T_{i,r}, w_{i,r}\}$ for guaranteed QoS requests or as $b_{i,r}\{t_{i,r}, T_{i,r}, p_{i,r}, \lambda_{i,r}\}$ for BE requests, where $t_{i,r}$ is the request arrival time, $T_{i,r}$ is its duration, $w_{i,r}$ (in bps) is the requested transmission rate in guaranteed QoS traffic, $p_{i,r}$ is the average size of the packets (in bits/packet) and $\lambda_{i,r}$ is the average number of generated packets per second (both for BE traffic). It holds that each MVNO i generates a set of requests $\mathcal{R}_i = \{r : r = 1, \cdots, R\}$. With regard to the shared RAN, we consider a cellular deployment, consisting of a set of sectors $\mathcal{S} = \{s : s = 1, \cdots, S\}$. We denote by $x_{i,s}(t)$, the traffic volume of MVNO i in sector s at time t (expressed in RBs).

Upon the arrival of a request $r \in \mathcal{R}_i$ from MVNO $i \in \mathcal{V}$, the capacity broker must decide if the future availability of resources will suffice to serve the request r based on traffic forecasting. We define the column vector of the previous T_p+1 samples of $x_{i,s}(t)$ as $\mathbf{x^t_{i,s}} = (x_{i,s}(t-T_p), x_{i,s}(t-(T_p+1)), \ldots, x_{i,s}(t))$, where t is expressed in minutes. Likewise, the vector of forecasted traffic volumes for the period $[t+1, t+T_f]$ is defined as $\hat{\mathbf{x}}^{\mathbf{t}}_{\mathbf{i,s}} = (\hat{x}_{i,s}(t+1), \hat{x}_{i,s}(t+2), \ldots, \hat{x}_{i,s}(t+T_f))$. Therefore, the forecasting function, f, can be defined as:

$$\begin{array}{rccc} f & : & \mathbb{R}^{T_p+1} & \longrightarrow & \mathbb{R}^{T_f} \\ & & \mathbf{x^t_{i,s}} & \longrightarrow & \hat{\mathbf{x}}^{\mathbf{t}}_{\mathbf{i,s}} \end{array} \tag{4.1}$$

There is a wide range of forecasting functions that could be used. In Section 4.4.3 we propose some improvements to be applied to the forecasting function, and in Section 4.5.2 results obtained with different forecasting methods are evaluated.

Let us note, that the actual traffic volume can be seen as the forecasted traffic volume plus an error, i.e., $x_{i,s}(t) = \hat{x}_{i,s}(t) + \epsilon_{i,s}(t)$, with $\epsilon_{i,s}(t) \in \mathbb{R}$. Thus, in order to cope with the inaccuracy of the forecasted traffic, we define the Confidence Degree (CD) of the traffic volume of sector s, $\gamma^{\beta}_s(t)$, as the value that will not be exceeded by the actual traffic volume with probability β. Thus, it holds that

$$P[\hat{x}_s(t) + \epsilon_s(t) \leq \gamma^{\beta}_s(t)] = \beta, \tag{4.2}$$

where $\hat{x}_s(t) = \sum_{i \in \mathcal{V}} \hat{x}_{i,s}(t)$ and $\epsilon_s(t) = \sum_{i \in \mathcal{V}} \epsilon_{i,s}(t)$.

4.4.2 MuSli: Algorithm for Multi-tenant Slicing of Capacity

In our proposal, the capacity broker allocates to incoming guaranteed QoS requests, the RBs that are expected to be available based on the forecast traffic volume. Conversely, RBs with higher probability of being used, must be allocated to incoming BE requests. Note that the capacity broker defines the available capacity at time t in sector s and for a given β, as $C^{\beta}_s(t) = C - \gamma^{\beta}_s(t)$, where C is the total capacity of each sector (i.e., both $C^{\beta}_s(t)$ and C expressed as the number of RBs). Due to differences in the requirements of the two traffic classes, MuSli prioritizes guaranteed QoS requests over BE requests.

4.4.2.1 Guaranteed Requests

Let us consider a request $g_{i,r}\{t_{i,r}, T_{i,r}, w_{i,r}\}$ generated by MVNO i to serve a specific user. This user moves around the scenario with a trajectory described by $\mathcal{M}_{i,r} = \{(s_1, \tau_1), \cdots, (s_M, \tau_M)\}$, where the tuple (s_m, τ_m) refers to the m^{th} sector visited by the user ($s_m \in \mathcal{S}$) and the time at which the user enters sector m (i.e., $\tau_m \in [t_{i,r}, t_{i,r} +$

$T_{i,r}$]). For this specific case, the capacity broker should only accept the request if the transmission rate (i.e., $w_{i,r}$ bps), can be guaranteed along $T_{i,r}$. In other words, it would be accepted if

$$\min_{t\in[\tau_m,\tau_{m+1})} \left\{C^{\beta}_{s_m}(t)\right\} \geq \frac{w_{i,r}}{w_{s_m}} \ , \forall(s_m,\tau_m) \in \mathcal{M}_{i,r}, \tag{4.3}$$

where w_{s_m} is the average transmission rate per RB, within sector s_m. Yet, as trajectories are unknown by the capacity broker, the acceptance/rejection decision is performed stochastically. We assume, that at time t_0 a set of new guaranteed traffic requests, namely $\mathcal{G}(t_0)$, reaches the capacity broker. According to the data collected until t_0, the probability that the new traffic will be served by sector s can be calculated as:

$$\alpha_s = \frac{w_s \sum_{i\in\mathcal{V}} ||\mathbf{x^{t_0}_{i,s}}||_1}{\sum_{s'\in\mathcal{S}} w_{s'} \sum_{i\in\mathcal{V}} ||\mathbf{x^{t_0}_{i,s'}}||_1}, \tag{4.4}$$

where $||\cdot||_1$ stands for the 1-norm operand. Initially, the set of accepted requests is empty and denoted by $\mathcal{G}'(t_0) = \emptyset$. Thus, a request $g_{i,r}\{t_0, T_{i,r}, w_{i,r}\} \in \mathcal{G}(t_0)$ is accepted if $F_g(g_{i,r}) \geq 0$ for $\forall t \in [t_0, T_{i,r}]$, where $F_g(g_{i,r})$ yields the available RBs given that $g_{i,r}$ is accepted. Hence, it is expressed as:

$$F_g(g_{i,r}) = \sum_{s\in\mathcal{S}} \alpha_s \left[C^{\beta}_s(t) - \left(\sum_{g_{j,k}\in\mathcal{G}'(t)} \frac{w_{j,k}}{w_s} \right) - \frac{w_{i,r}}{w_s} \right]. \tag{4.5}$$

We calculate (4.5) for all sectors of the deployment (each one weighted by α_s), by subtracting the resources that are needed to serve the already accepted requests and the resources required for the incoming $g_{i,r}$, from the available capacity of sector s in time t. If accepted, $g_{i,r}$ is removed from $\mathcal{G}(t_0)$ and it is included in $\mathcal{G}'(t_0)$. This procedure is repeated for all requests in $\mathcal{G}(t_0)$.

4.4.2.2 Best Effort Requests

BE requests are served after accommodating the guaranteed ones. However, since these requests do not have the strict data rate constraint imposed by the latter, the capacity broker can allocate them resources more flexibly. Let us consider that at time t_0, a set of new BE traffic requests (i.e., $\mathcal{B}(t_0)$), reaches the capacity broker.

For a given request $b_{i,r}\{t_0, T_{i,r}, p_{i,r}, \lambda_{i,r}\} \in \mathcal{B}(t_0)$, the average amount of bits generated along its duration (i.e., $T_{i,r}$), may be expressed as $T_{i,r}p_{i,r}\lambda_{i,r}$ bits. Following the same rationale stated in Section 4.4.2.1, the average number of RBs required to serve this request in sector s, is equal to $\frac{T_{i,r}p_{i,r}\lambda_{i,r}}{w_s T_{sf}}$, where T_{sf} is the sub-frame time of LTE-A (i.e., 0.5 msec). However, the service disruption tolerance of BE traffic allows the capacity

broker to allocate resources more elastically. Therefore, if we define the set of accepted new BE requests at time t_0 as $\mathcal{B}'(t_0)$, which is initially empty (i.e., $\mathcal{B}'(t_0) = \emptyset$), a request $b_{i,r}\{t_0, T_{i,r}, p_{i,r}, \lambda_{i,r}\}$ will only be accepted if $F_b(b_{i,r}) \geq 0$. $F_b(b_{i,r})$ expresses the available RBs given that $b_{i,r}$ is accepted and it is expressed as

$$\begin{aligned} F_b(b_{i,r}) &= \sum_{s\in\mathcal{S}} \alpha_s \left[\int_{t_0}^{t_0+T_{i,r}} \left(C_s^{\beta}(t) - \sum_{g_{j,k}\in\mathcal{G}'(t)} \frac{w_{j,k}}{w_s} \right) dt - \right. \\ &- \left. \left(\sum_{b_{j,k}\in\mathcal{B}'(t)} \frac{\lambda_{j,k} p_{j,k} T_{j,k}}{w_s T_{sf}} \right) - \frac{\lambda_{i,r} p_{i,r} T_{i,r}}{w_s T_{sf}} \right]. \end{aligned} \tag{4.6}$$

We compute (4.6), by subtracting the required resources to serve the already accepted BE requests and the resources to serve $b_{i,r}$, from the available capacity in sector s, along the duration of the request (i.e., $T_{i,r}$). As guaranteed requests precede, the available sector capacity for BE requests is calculated by deducing the resources needed to serve the accepted guaranteed ones. If request $b_{i,r}$ is accepted, then it is removed from $\mathcal{B}(t_0)$ and it is included in $\mathcal{B}'(t_0)$.

4.4.3 Capacity Forecasting

The flexibility of the network sharing management architecture (i.e., detailed in Section 4.3), required to provide short-time scale dynamic provision of resources, poses challenges into traffic forecasting. There are several factors that affect the variation of the traffic along time, such as the mobility of the users, the deployment of the eNBs, etc. In our work, non-uniformities in the prior traffic load are due to gravity points of the mobility model. Given that the time horizon of the forecasting (which is taken into account by the capacity broker to make admission decisions) depends on $T_{i,r}$ of each request, we propose the prior decoupling of the variation trends that exist in $\mathbf{x}^{\mathbf{t}}_{\mathbf{i,s}}$.

In order to conduct the decoupling, the forecasting function, first defined in (4.1), performs the Fast Fourier Transform (FFT) of the traffic vector for each sector, i.e. $\mathbf{X_{i,s}} = \mathcal{F}\{\mathbf{x}^{\mathbf{t}}_{\mathbf{i,s}}\} = \{X_{i,s}(k) : k = 0, \ldots, T_p\}$, where $\mathcal{F}\{\cdot\}$ stands for the FFT transform. After applying the FFT, the capacity broker identifies the set of peaks of $\mathbf{X_{i,s}}$ and then splits it up into a set of components. Hence, for the jth peak of $\mathbf{X_{i,s}}$, located at $k = k_j$, we define $\mathbf{X}^{\mathbf{j}}_{\mathbf{i,s}} = \{X^j_{i,s}(k) : k = 0, \ldots, T_p\}$ where $X^j_{i,s}(k) = \{X_{i,s}(k) \cdot \Lambda_j(k) : k = 0, \ldots, T_p\}$, with $\Lambda_j(k) = 1$ for $k_{j,min} < k < k_{j,max}$ and $\Lambda_j(k) = 0$ otherwise. If a minimum threshold X_{min} is set, the limits $k_{j,min}$ and $k_{j,max}$ are defined as $k_{j,min} = (k_{j-1} + k_j)/2$ and $k_{j,max} = (k_j + k_{j+1})/2$. Finally, the decoupled traffic is generated as $\mathbf{x}^{\mathbf{t,j}}_{\mathbf{i,s}} = \mathcal{F}^{-1}\{\mathbf{X}^{\mathbf{j}}_{\mathbf{i,s}}\}$, where $\mathcal{F}^{-1}\{\cdot\}$ is the Inverse Fast Fourier Transform (IFFT).

The important point to note here, is that each $\mathbf{x}_{\mathbf{i,s}}^{\mathbf{t,j}}$ isolates a component of the traffic variation, and therefore it can be the basis for a more accurate forecasting. Thus, for a given forecasting method $f_{FM} : \mathcal{R}^{T_p+1} \rightarrow \mathcal{R}^{T_f}$, the forecasted vector of sector s assuming that J peaks are identified in $\mathbf{X_{i,s}}$ may be expressed as: $\mathbf{x}_{\mathbf{i,s}}^{\mathbf{t}} = \sum_{j=1}^{J} f_{FM}\left(\mathbf{x}_{\mathbf{i,s}}^{\mathbf{t,j}}\right)$.

In Section 4.5.2, results for different f_{FM} are obtained, i.e., ARIMA, compressive sensing-based method, Kalman Filter and Holt-Winters.

4.4.4 Forecasting Error and Confidence Degree

As stated in (4.2), the forecasting error and the CD are tightly coupled. Specifically, the error $\epsilon_{i,s}(t)$ depends on t, T_p, T_f and f_{FM}. Therefore, in Section 4.5 the error (and consequently the CD, γ_s^{β}) is estimated empirically by applying the following methodology:

- 1000 realizations of $\epsilon_{i,s}(t)$ are collected (i.e., in a deployment with differently loaded cells) for each forecasting method. Next the 1000 sample measurements are used to obtain the empirical density function by employing the Kernel Density Estimation Technique (KDE) [89]. KDE is a non-parametric method, and thus it is not necessary to make assumptions on the $\epsilon_{i,s}(t)$ distribution.
- For computing the CD, a profile of 1000 experimentally estimated capacity values (i.e., $\hat{x}_{i,s}(t)$) is created. This profile is used as an observation. As previously, the KDE is used to obtain the empirical density function.

4.5 Performance Evaluation

4.5.1 Simulation Environment and Parameters

In this chapter our simulation consists of an Urban Micro-cell (UMi) scenario comprising 19 BSs with 3 sector antennas each one (total $S = 57$ sectors), based on the IMT-Advanced evaluation guidelines [90]. Table 4.1 summarizes the detailed system parameters. Users move in the network following the SLAW model, which is a human walk mobility model, considering mobiles moving in confined gravity areas [91]. According to this model, users move among a number of waypoints, which are distributed over the network area according to self similarity rules forming a given number of clusters. Clusters with more waypoints can be seen as hotspots attracting more users.

With regard to the forecasting, we collected the prior data traffic records from 57 sectors with coverage 2000 m^2. Each data record contains: *Time*, *Sector ID* and *RBs*. For our

simulations, we use two traffic models to represent guaranteed QoS and BE traffic, following parameters in [92]. The users generate guaranteed Constant Bit Rate (CBR) VoIP traffic with transmission rate 64 Kb/s, as well as BE traffic FTP requests with file size 0.5 Mbyte every 60 seconds. The inter-arrival rate follows a Poisson distribution.

TABLE 4.1: Basic System Parameters used in the Simulation, MuSli

Parameters	**Settings/Assumptions [90]**
Network layout	19 BSs ($S = 57$ sectors)
Tenants	$V = 2$ (MNO:$i = 0$ and MVNOs:$i = 1, 2$)
Inter-site distance	200 m (ISD)
Bandwidth	20 MHz (100 RBs) 2.5 GHz
Antenna configuration	2 x 2 MIMO
Path loss Model	$36.7log10(\text{d[m]}) + 22.7 + 26log10(\text{fc[GHz]})$
Shadow fading	Lognormal, $\mu = 0$, std.=4 dB

4.5.2 Forecasting Evaluation

For our study, we examine the following short-term capacity forecasting methods: ARIMA [85], compressive sensing-based method [88], Kalman filter [87], and Holt-Winters [86]. To identify the most suitable method for the capacity broker, we generated data that spanned in a two-hour prior time period ($T_p = 120$ minutes) using SLAW mobility model [91] and we obtained a $T_f = 20$ minute forecast. According to SLAW, the generated data capture spatial non-uniformities due to variations in users' trajectories. To compare the performance of the above methods, we consider a set of network instances with different load conditions. We use Root Mean Square Error (RMSE) to measure the forecasting accuracy of the studied methods. RMSE represents the sample standard deviation of the difference between predicted and observed values. The results in Table 4.2 show that the most accurate forecast (in the sense of minimizing RMSE) is the Holt-Winters technique. Applying the decoupling method of Section 4.4.3 (i.e., FFT), outperforms the case of forecasting the prior traffic vector without any decomposition. The highest gain is achieved in methods that leverage the seasonality of the input data (i.e., Holt-Winters and Kalman Filter). Therefore we conclude that Holt-Winters exponential smoothing suits better short-term prediction scenarios.

TABLE 4.2: RMSE of the studied Forecasting Methods

	HW	Kalman	Comp.Bas.Sens.	Arima
Without FFT	4.18	5.25	7.1	9.9
With FFT	2.46	3.97	5.96	7.43

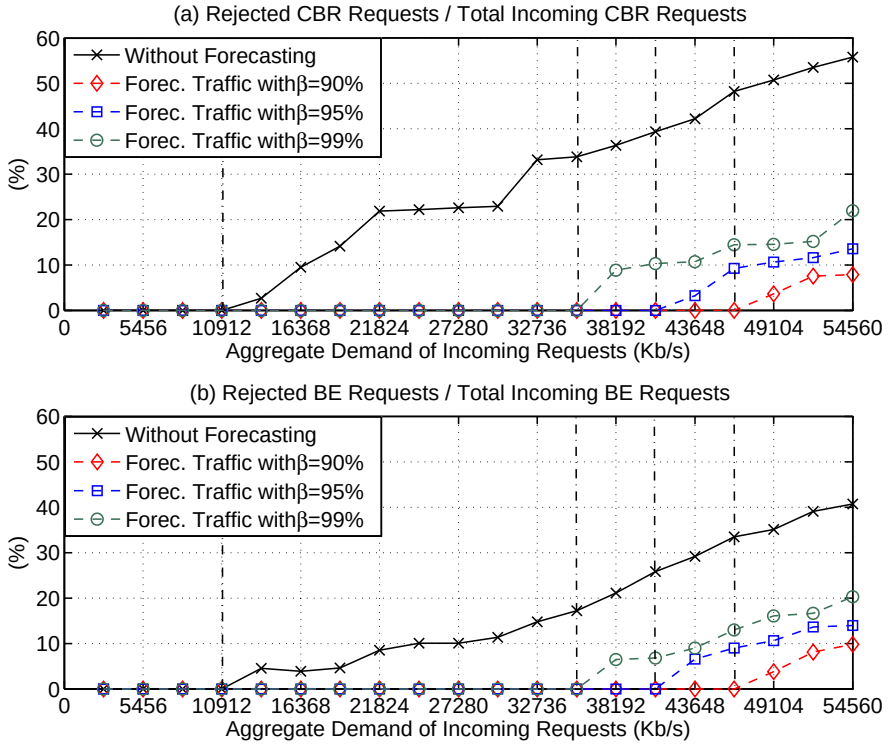


FIGURE 4.2: (a) Rejected Guaranteed Requests and (b) Rejected BE Requests.

4.5.3 MuSli Results

In this section we study the performance of the capacity broker, by executing MuSli for varying forecasting CDs (i.e., where $\beta = \{90\%, 95\%, 99\%\}$). The capacity slicing is applied by considering all cells in the network deployment. In our scenario, MVNOs generate both guaranteed QoS and BE requests, with a traffic mix ratio 20% - 80%. We study different parameters for the time duration of the prediction (i.e., T_f), while augmenting the aggregate demand of incoming requests. At the arrival moment of a request (i.e., t_0), MuSli decides which requests to accept / reject by checking the prediction CD (i.e., β). To evaluate its performance, we compare MuSli with the baseline scenario, where admission for an incoming request is based on resource availability at the arrival moment of a request, t_0. We conducted Monte-Carlo event-based simulations in MATLAB® with 1000 iterations to achieve statistical validity for each forecasting step.

4.5.3.1 Admission of Incoming Requests

We begin the evaluation of MuSli by emphasizing the effect of slicing the overall capacity using various CDs (i.e., β), on the number of accepted / rejected requests. Fig. 4.2

depicts the average percentages of (a) rejected guaranteed QoS (i.e., CBR) and (b) BE (i.e., FTP) requests.

In general, when the capacity broker applies MuSli with different CDs, more requests are accepted compared with the baseline scheme. Even for the case of MuSli with $\beta = 99\%$ for 46376 Kb/s aggregate demand (i.e., the most conservative approach in slicing resources), the capacity broker rejects 10.28% of the incoming guaranteed requests whereas the baseline scenario rejects 39.34%. In particular, we observe that the capacity broker that applies MuSli with high β rejects more requests, since it considers that there is overall less capacity to allocate to them. This is a strict (i.e., conservative) slicing approach that ensures surplus capacity to the operator. The vertical dashed lines denote the limit of offered load that can be accepted without any rejection (i.e., 10912 Kb/s for the baseline scheme, 35464 Kb/s for MuSli with $\beta = 99\%$, 40920 Kb/s for MuSli with $\beta = 95\%$ and 46376 Kb/s for MuSli with $\beta = 90\%$).

In principle, there is a trade-off between service quality assurance and number of served requests. On the safe side, using high β on the predicted traffic, ensures service quality but results into accepting fewer requests, since the capacity broker considers less resources to allocate. Therefore, the capacity broker can tune the CD of the forecasting to treat requests, according to the desired level of certainty in assuring service quality. For this reason, in Fig. 4.2, the capacity broker that applies MuSli with high β rejects more both guaranteed and BE requests compared with MuSli with lower β.

Moreover, when comparing Fig. 4.2(a) and Fig. 4.2(b), BE requests are rejected with lower probability compared to guaranteed ones. This is due to their more relaxed delay constraints in contrast to the VoIP requests that are characterized by stringent requirements.

In general it is considered preferable from the network operator's point of view to sacrifice BE requests due to their elastic demands. It is further noted that BE requests are served after admitting the guaranteed ones. However, since these requests do not have the strict data rate constraint imposed by the latter, the capacity broker can allocate them resources more flexibly. To that end we note that there is a trade-off among service quality and the CD of the forecasting.

4.5.3.2 Resource Block Utilization

In Fig. 4.3, we study the mean percentage of (a) RB utilization and (b) RBs of dropped requests, versus the aggregate demand of incoming requests. In our scenario, a guaranteed request is dropped when it lacks resources at some point along its duration, whereas

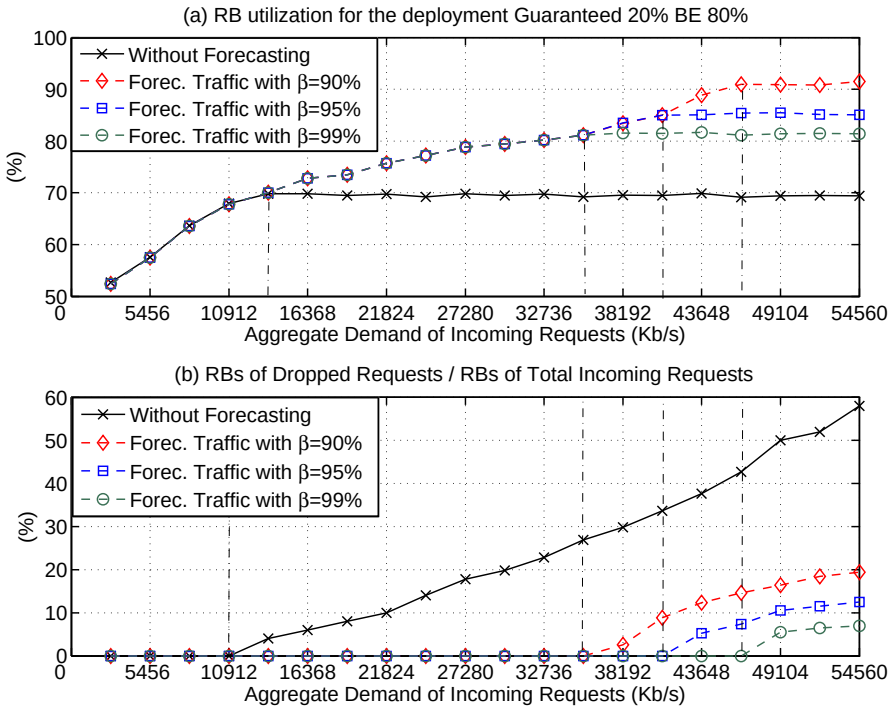


FIGURE 4.3: (a) RB utilization and (b) SLA violation.

a BE request is dropped when its total transmission time is higher than a threshold time [92] (i.e., this parameter is chosen based on the requested service). Given that both these cases result into disregarding the agreed SLA, let us refer to them as SLA violation.

In principle in Fig. 4.3(a), we observe that for low incoming demand (up to 10912 Kb/s), accepting requests based only in current resource knowledge (i.e., baseline approach) results into the same utilization as the one achieved by the capacity broker. As soon as the baseline approach starts rejecting the incoming demand (i.e., starting at 13640 Kb/s as also shown in Fig. 4.2), the RB utilization stabilizes around 69.8%. However, short term traffic prediction can prove to be very useful for higher demands. The capacity broker, by applying MuSli improves the utilization of the network since the accurate prediction enhances the knowledge of the network and therefore resources are used more efficient to serve the incoming requests. All RB utilization curves stabilize at a certain offered load limit, beyond which the capacity broker rejects further incoming requests. This is depicted in Fig. 4.2 as well.

As expected applying MuSli with high β results in restricted utilization compared to MuSli with lower β. This is due to the fact that MuSli with high β considers less overall network resources to allocate to the incoming load. As shown in Fig. 4.2, when using high β more requests are rejected and thus the RB utilization is limited. Furthermore since we are considering the whole deployment, particular overloaded cells (i.e., gravity

points of the mobility model where several users are concentrated) restrict the available resources that the capacity broker can allocate in the slicing process.

Moreover Fig. 4.3(b) illustrates the percentage of RBs of dropped requests due to violation of the SLA. Although MuSli with high β rejects more requests (see Fig. 4.2), it is less likely to have dropped ones (e.g., when the real traffic is higher than the chosen CD, β). For instance, for 43648 Kb/s, an operator can choose Musli with $\beta = 90\%$ to achieve 90% utilization in the cost of having 11% SLA violation. On the contrary, being more conservative and choosing MuSli with $\beta = 99\%$, will result into 81% utilization without any SLA violation. This confirms the trade-off between service quality assurance and number of served requests. Our results show, that for guaranteed services without SLA violation, our proposal yields resource block utilization gain in the range of 15-25% (i.e., while $\beta = 95 - 90\%$ and considering a traffic mix of requests). In this scenario, MuSli achieves higher utilization by accepting first CBR requests according to their starting time and duration and then BE based on the available capacity. As also pointed in section 4.5.3.1 CBR requests are prioritized over BE ones due to their stringent service requirements.

As it can be observed, the network slicing multiplexing gains achieved with MuSli allow for increasing the number of requests can be accepted in the system. In Fig. 4.3(a) requests can be admitted such that the effective capacity of the system and the achievable RB utilization increase accordingly. The cost of this gain is shown in Fig. 4.3(b) where after admitting in the requests, the SLA protection level could be threatened. Each operator should use these figures as guidelines on how they choose to provide the services and the according type of subscriptions that they would like to accommodate.

We remark the following key points: (i) the increasing slope in Fig. 4.3(a) while augmenting the number of requests and (ii) the larger the system capacity, the greater the relative gain. A small number of requests are fully accommodated with or without traffic prediction. As soon as the network becomes congested, i.e., some requests must be rejected, the utilization of our proposal outperforms the baseline scheme without forecasting. On the other hand, when the system capacity is improved (i.e., higher RB utilization), there is more room to accommodate more requests into the system showing better performance. Finally we note that low SLA violation risk levels result in substantive system RB utilization gains.

4.5.4 Impact of Traffic Load Aspects

The capacity broker performs resource monitoring by checking the available resource blocks and MCS of the traffic along with the traffic forecasting methods as described

above. Additionally it supports admission control for different traffic types (e.g., GBR VoIP and BE FTP) and therefore supports multiple classes of SLAs. In our proposal, the capacity broker allocates the RBs that are expected to be available based on the forecast traffic volume to incoming guaranteed QoS requests. On the other hand RBs with higher probability of being used, must be allocated to incoming BE requests.

In this section we would like to further discuss how different dimensions of the traffic load impact on the performance and implementation of the capacity broker.

The first dimension is the traffic type. We observe that the total number of admitted requests increases with the number of BE requests, showing that BE traffic requests are preferred due to the higher flexibility. Additionally low SLA violation risk levels for GBR traffic result in very significant system utilization gains. This aspect has been analyzed in detail in sections 4.5.3.1 and 4.5.3.2.

The second dimension is the geographical traffic dimension. When a high number of users is concentrated in hotspots a high amount of traffic requests are concentrated in particular areas. Therefore the capacity broker receives such an increased amount of distinct traffic requests. This means that slice allocations may overlap and traffic class requirements might not be satisfied incurring in SLA violations. The MNO is responsible to determine the capacity broker to handle requests with a specified SLA based on particular user subscriptions to allocate the desired SLAs. In any case the capacity broker acts as mediator, mapping the SLA requests of multiple tenants with the physical network resources through the interfaces provided by the MO-NM. In such scenario the Type 5 interface as well as the vertical industries / OTT provider APIs should be extended to accommodate on-demand network slice requests with particular SLA and timing requirements. The Itf-N and Itf-B interfaces should also be extended to carry out the configuration of network slices by introducing the necessary signaling considering MVNOs and vertical industries / OTT providers [93].

Finally traffic demand may vary between locations based on the time. This dimension should be taken into account when designing a capacity broker solution. In particular the Itf-N and Itf-B interfaces should also be extended to cope with such scenario. Such interface enhancement and signaling should contain a set of additional information including (i) the amount of resources allocated to a network slice, e.g. physical resources or data rate, (ii) timing, e.g. starting time, duration, or periodicity of a request and time window and (iii) service related information, e.g., mobility (stationary, low, medium, high), data off-loading policies, and service disruption tolerance [93].

4.6 Conclusion

In this chapter, we integrated the capacity broker in the 3GPP management architecture with a minimum set of enhancements. In addition, by leveraging traffic non-uniformities in a shared deployment, we proposed MuSli, a framework to be implemented by the capacity broker in coarse time scales. Along with our proposal, we introduced a decoupling process to extract variation trends in irregular traffic patterns and improve traffic forecasting. MuSli, by deciding how to slice the deployment's capacity among two types of requests (i.e., Guaranteed QoS and BE), improves network's performance by (i) increasing the accepted requests, and (ii) decreasing the underutilized resources. The created slices are self-contained and mutually isolated. Our results can be leveraged by infrastructure owners to flexibly allocate capacity to tenants, considering different types of services and the uncertainty of expected traffic.

Chapter 5

NetSliC: Base Station Agnostic Framework for Network SliCing

5.1 Introduction

5.1.1 Motivation

Fifth Generation (5G) cellular Radio Access Network (RAN) is envisioned to support different categories of services, which require ultra-high reliability and low latency, enhanced mobile broadband or massive connectivity. This type of Quality of Service (QoS) is vastly different from that of legacy mobile broadband applications.

Traditional Distributed RAN (D-RAN) supports the dense deployment of standalone Small Cells (SCs) to increase the area spectral efficiency. However, it has been quickly observed that non-ideal BackHaul (BH) limits the coordination among SCs. Incremental advancements on traditional D-RAN architecture are not able to satisfy these QoS requirements [19].

In order to cope with the increasing capacity demand and new service requirements, the Cloud-RAN (C-RAN) paradigm has been introduced to increase the degree of cooperation between access nodes [22]. In this model RAN functions are split between the Base Band Unit (BBU), hosted in the cloud, and Remote Radio Heads (RRHs) that provide antenna equipment and radio access. However, the centralization of BBU into a common shared BBU pool poses strict capacity requirements to the FrontHaul (FH) (i.e., interface between RRH and BBU). Furthermore the availability of high speed fiber links (i.e., ideal FH), especially in urban deployments, is controversial due to high implementation cost.

As a result of the opposite trends (i.e., initial decentralization and subsequent centralization of functionalities and layers) and the physical hurdles to connect different types of Base Stations, BSs, (i.e., SCs / RRHs) through high capacity and low latency connections, the current deployed RAN is a mixture of BSs, some of them following the D-RAN architecture and some of them the C-RAN architecture.

In addition, Mobile Network Operators (MNOs) that are interested in using 5G to increase capacity of their networks are likely to deploy future 5G New Radio (NR) NodeBs (gNBs) alongside SCs and RRHs. These access nodes follow the Network Function Virtualization (NFV) model. The gNBs are characterized by different functional splits between Central Units (CUs) and Distributed Units (DUs), thus supporting centralization of upper layers of the NR radio protocol stack. Different protocol split options between CU and lower layers of gNBs are possible [7]. The functional split between the CU and lower layers of gNBs depends on the transport network. In a gNB having transport network with high latency (i.e., non-ideal), higher layer splits are applicable whereas for transport network with negligible latency (i.e., almost ideal), lower layer splits are chosen to realize enhanced performance such as centralized scheduling [7, 26].

Unlike previous LTE standard, 5G NR has been standardized as a non-standalone network at a first stage (i.e., included in Rel. 15 and completed in December 2017), and as a standalone network at a subsequent stage (i.e., also included in Rel. 15 and approved in June 2018). In [7] a full set of 5G RAN architecture options are discussed, ranging from the complete standalone 5G gNB directly connected to the Next Generation Core (NGC) to the non-standalone option where the gNB is connected to the Evolved Packet Core (EPC) through an Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Node B (eNB). This range of architecture options means that coexistence and inter-working with previous standards are main design objectives of the 5G RAN to allow a gradual migration from 2G / 3G / 4G to 5G, thereby reducing time to market and minimizing initial MNOs' investment. The deployment of the standalone 5G RAN and NGC able to provide full coverage requires time and a major investment by MNOs. Initial 5G deployments will be multi-standard networks, resulting in a composite of different 3GPP, and eventually non-3GPP, network architectures jointly operated to serve the diverse traffic, thus resulting in intermediate phases where inter-working between 5G and previous standards will be a necessity either at RAN level or at Core Network (CN) level. This progressive migration process is not taken into consideration in the current literature and these intermediate deployments have not been tackled therein.

This composition of BSs with different transport and access capabilities yields complex deployments such as the one depicted in Fig. 5.1. In these deployments distinct types of BSs coexist, i.e., legacy distributed SCs connected with non-ideal BH to the CN,

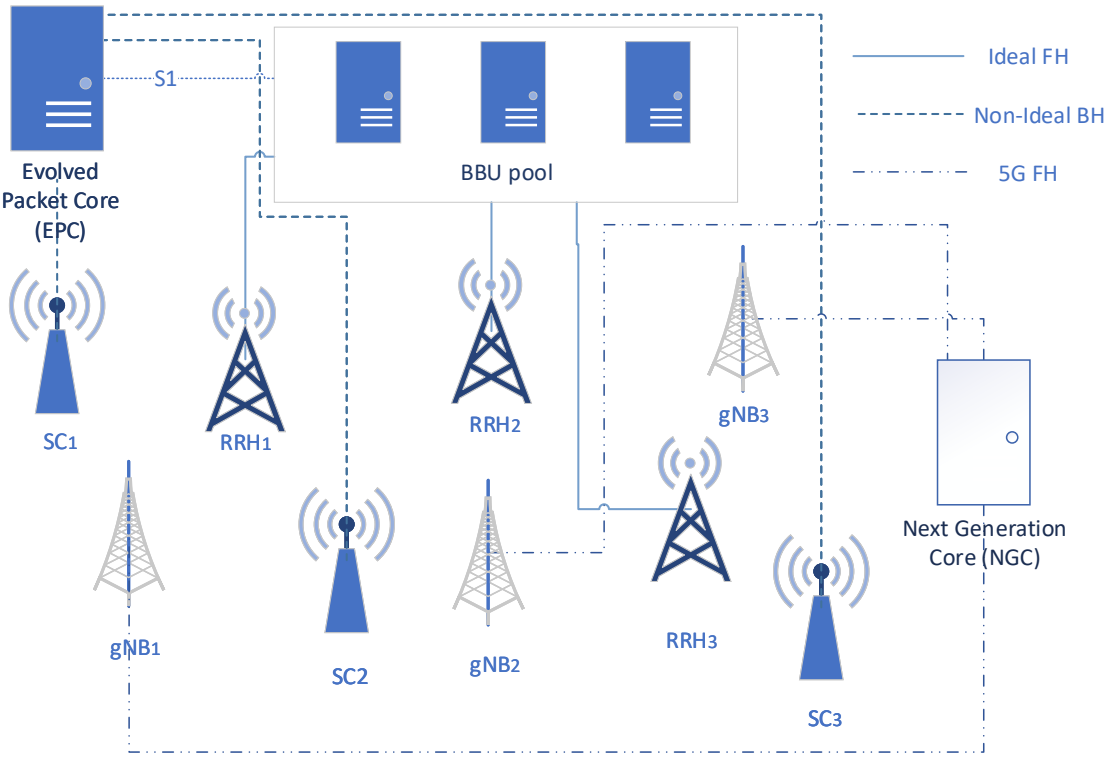


FIGURE 5.1: Scenario comprising SCs with non-ideal BH, RRHs connected to the BBU with ideal FH and gNBs with 5G FH.

RRHs connected with ideal FH to a centralized BBU pool and future gNBs leveraging virtualization with intermediate functional split, connected with 5G FH to the NGC. Several bottlenecks appear therein with regard to the characteristics of the network components. In principle, the air spectrum is scarce. The constrained BH capacity is a limitation, due to the large number of users and the increasing demands for high data rate services. Despite the fact that ideal FH is deployed, it can support a limited number of users due to the constraint in the BBU processing load. Furthermore, the delay of the transport network and the delay imposed by each service type constitute limitations that should be tackled.

Moreover network slicing has been introduced at the BS level, to allow differentiated service treatment depending on distinct service requirements. Leveraging network slicing, MNOs consider users as belonging to different tenant types based on their respective Service Level Agreement (SLA) and subscriptions. To that end, MNOs support different policy enforcement between network slices based on the SLA [7]. A further open issue when realizing network slicing lies on how to choose criteria that fit both the distinct BS requirements and the vastly different SLAs of services. In addition, the flexible functional split impacts on the performance of network slicing and the optimal split depends on the target service.

One of the biggest research challenges is to obtain a mechanism that could perform resource allocation and isolation of wireless slices independently of the BS technology. The air-interface, the spectrum and the capabilities of the transport network are different for each particular BS. To the best of our knowledge, there is not yet a unified approach dealing with any of the above mentioned factors. This dependency on the access and transport capabilities as well as the distinct functional split of each BS will be a problem when slicing a heterogeneous migration network (i.e., consisting of legacy distributed SCs, RRHs connected to a centralized BBU pool and future gNBs with intermediate functional split as shown in Fig. 5.1). Therefore, MNOs must design efficient BS agnostic mechanisms to jointly manage this complexity in a seamless manner.

This calls for a new layer, leveraging virtualization, in a multi-tenant migration scenario not only composed of different traffic types (e.g., users with tight latency requirements and users having high data rate demands) but also BSs adopting network slicing characterized by different access and transport capabilities (i.e., RRHs, SCs and gNBs with various functional splits with ideal and non-ideal transport network). Within this future network deployment (see Fig. 5.1), since large amount of SCs, RRHs and gNBs coexist, the re-selection of access BS becomes a great challenge. Therefore, the creation of network slices shall ensure that the proper amount of resources is reserved per BS based on the tailored SLA of the specific traffic requirements.

5.1.2 Related Work

Although several excellent works have been published on wireless network slicing leveraging virtualization, most of them provide solutions on the architecture and allocation of wireless resources, being less focused on the design of BS agnostic solutions for creating the appropriate network slice for each service type. In addition, few studies propose solutions to be applied in future migration deployments that combine legacy D-RAN with the prominent C-RAN architecture along with scenarios including gNBs with functional splits. Currently many stakeholders put a lot of effort in harmonizing the BH / FH [21], since transport network can rely on different types of technologies (i.e., fiber, millimeter wave and/or microwave). Some studies, such as [94], deal with user association in a multi-tenant scenario, where admission control with respect to network slicing is devised at single cell level in a deployment composed only of SCs. The authors in [95] propose a user association mechanism for multi-tenant RANs that takes into account solely the access capacity in a scenario composed only of SCs. [96] proposes a method for choosing appropriate BSs and physical resources in a C-RAN for users while minimizing the overall energy consumption and reducing the network interference. However, in this scenario differentiated transport capacity and delay capabilities for the involved

BSs are not taken into account. In addition, the authors of [97] propose a dynamic slicing scheme that flexibly schedules radio resources based on the requested SLA, while maximizing user rate and applying fairness criteria. However, the problem of how to perform network slicing in deployments combining legacy with future architectures, where deployed BSs have distinct access and transport network characteristics has not been tackled. Furthermore [98] proposes the joint user association and resource allocation in a multi-cell virtualized wireless scenario. However this work assumes only one type of BS (i.e., legacy SC) and further considers that each user at each transmission instance can connect to no more than one BS. Moreover [99] proposes a dynamic network slicing scheme for multi-tenant C-RAN, which takes into account MNOs' priority, baseband resources, FH and BH capacities, QoS and interference. This scheme leverages advantages of the centralized BBU but it does not consider the coexistence of C-RAN with legacy SCs and future gNBs with functional splits. The authors in [100] introduce a new architecture in compliance with the ETSI-NFV model and the 3GPP specifications to create a fine-grained network slicing solution. They leverage the NFV model (i.e., this layer refers to the NFV infrastructure (NFVI)), which is controlled by the Virtualized Infrastructure Manager (VIM). However this work focuses on a cost aware solution and targets only future deployments without mentioning distinct transport capabilities for the involved access points. Therefore, the current study aims to fill the aforementioned gaps.

5.1.3 Contribution and Structure

In this chapter, taking into account the gaps in the current literature, we present the BS agnostic framework for Network SliCing (NetSliC) to be adopted by the MNOs. Our main contributions are summarized as follows:

- We shed light on how to solve the problem of providing efficient and tailored network slicing for different services in a future migration scenario combining traditional SCs with non-ideal BH capabilities connected to the CN, future RRHs with ideal FH connected to the BBU and gNBs with different functional splits connected to the NGC through 5G FH. The offered services share the same RAN resources while some of them maintain tight latency requirements competing among other service types with different demands.
- NetSliC achieves efficient management of BSs having different access and transport network technologies, by creating customized network slices based on the characteristics of different traffic types in terms of latency and data rates as well as the different BSs capabilities in terms of access and transport network. NetSliC

enhances the management of complex deployments, which will be the result of the migration process toward 5G.

The rest of the chapter is structured as follows. We first introduce our System Model in section 5.2. Building on this model, we formulate the problem to be solved and we propose NetSliC in section 5.3 along with implementation details that make our solution backward compatible with the existing standard. In section 5.4 we conduct a performance evaluation of our framework and the chapter concludes with some final remarks in section 5.5.

5.2 System model

This section expounds the model for each particular component of the scenario under study. In particular, we describe the set of different traffic types, the characteristics of the air interface, the model for the BH and FH connections for each BS, the analysis of the processing load in the BBU pool as well as the definition of functional split for each BS (i.e., SC, RRH and gNB).

5.2.1 Characterization of the traffic and users

The set of traffic services considered in the scenario is denoted by $\mathcal{S}$, and its cardinality by $S = |\mathcal{S}|$. Specifically, in the following each service $s \in \mathcal{S}$ is characterized by the tuple (λ^s, d^s, t^s), where λ^s is the mean packet arrival rate (in packets/sec), d^s is the mean packet length (in bits) and t^s is the allowed delay. Only downlink traffic carried in the Physical Downlink Shared CHannel (PDSCH) is considered.

The set of users generating traffic is defined as $\mathcal{K}$, and the cardinality of the set is expressed as K. As each user generates traffic according to the defined set of services, we further define the set of users with service $s \in \mathcal{S}$ as $\mathcal{K}^s$, where $\cup_{s\in\mathcal{S}}\mathcal{K}^s = \mathcal{K}$ and $\cap_{s\in\mathcal{S}}\mathcal{K}^s = \emptyset$. Therefore, for a user $k \in \mathcal{K}^s$, the mean transmission rate R_k is given by

$$R_k = R^s = \lambda^s \cdot d^s, \ \forall k \in \mathcal{K}^s, \tag{5.1}$$

where R^s is the average transmission rate of the service s. Each user has N_{ue} antennas.

5.2.2 Characterization of the Base Stations

The set of deployed BSs in the scenario is denoted by $\mathcal{B}$, which in turn is divided into RRHs, $\mathcal{B}_C$, distributed SCs, $\mathcal{B}_D$, and gNBs, $\mathcal{B}_G$, with $\mathcal{B} = \mathcal{B}_C \cup \mathcal{B}_D \cup \mathcal{B}_G$. Given the delay

experienced by distributed SCs due to the non-ideal BH connection, in general CoMP techniques for SCs are unfeasible or gains are drastically reduced [101]. Therefore, in the sequel only RRHs connected via ideal FH to the BBU are assumed to implement CoMP techniques. Specifically, the type of CoMP implemented by RRHs is assumed to be Coordinated Scheduling and Coordinated Beamforming (CS / CB) [102, 103]. In CS / CB data to a single user is instantaneously transmitted from one of the RRHs in the CoMP set and scheduling decisions and/or generated beams are coordinated in the BBU to control the created interference. This choice also provides a much reduced load in the BBU since only the scheduling data and no further control information needs to be transferred between the different RRHs that are coordinating with each other. In addition user data does not need to be transmitted from multiple RRHs, and therefore only needs to be directed to one RRH. Both RRHs and distributed SCs use Zero Forcing (ZF) beamforming or Interference Alignment (IA) techniques to cancel intra-cell and inter-cell interference [104–106]. All BSs have N_{bs} antennas.

5.2.3 Downlink Transmission

We define the eigenmodes of the downlink channel as $L_{max} = \min(N_{bs}, N_{ue})$, which means that the MIMO channel between a user $k \in \mathcal{K}$ and a BS $b \in \mathcal{B}$ can be decomposed into L_{max} parallel non-interfering SISO channels (under the assumption of perfect Channel State Information at the Transmitter (CSIT) and at the receiver (CSIR)). The total capacity of the MIMO channel between BS b and user k ($C_{b,k}$) is then the summation of the capacity of each individual parallel SISO channel [107]. Thus, assuming a single Physical Resource Block (PRB), the spectral efficiency is

$$C_{b,k} = \sum_{l=1}^{L} \log_2 \left(1 + \gamma_{b,k}^{l}\right), \tag{5.2}$$

where $\gamma_{b,k}^{l}$ is the SINR received at user k from BS b for the stream / channel l, and the number of streams allocated to each user (L) is smaller than the number of eigenmodes, $L \leq L_{max}$. Assuming downlink zero forcing (ZF) beamforming or alternative Interference alignment (IA) solutions, intra-cell interference is effectively cancelled [104–106]. However, inter-cell interference can only be cancelled if there is cooperation among BSs. Therefore, if we define the set $\mathcal{F}_k$ as the set of BSs that cooperate to give service to user k (through CS / CB), the SINR received by user k from BS b for stream l is expressed as

$$\gamma_{b,k}^{l} = \frac{p_{b,k}^{l} \cdot (g_{b,k}^{l})^2}{\sigma^2 + \sum_{j \notin \mathcal{F}_k} \sum_{v=1}^{L} p_j^v \cdot (g_{j,k}^{v})^2}, \tag{5.3}$$

where $g^l_{b,k}$ is the lth channel gain between BS b and user k, $p^l_{b,k}$ is the transmitted power and σ^2 is the mean noise power. Let us assume that, for a given user k and BS b, the SINR of two streams is approximately the same, i.e., $\gamma^l_{b,k} \approx \gamma^v_{b,k}$. Then, it can be approximated that

$$C_{b,k} \approx L \cdot \log_2 (1 + \gamma_{b,k}), \tag{5.4}$$

where $\gamma_{b,k} = \frac{1}{L} \sum_{l=1}^{L} \gamma^l_{b,k}$. According to [107], this approximation is very accurate. It is known the optimal power allocation is given by the water-filling algorithm [107]:

$$p^l_{b,k} = \left(p_{wf} - \frac{1}{\sigma_i^2} \right)^+, \tag{5.5}$$

where $(x)^+ = \max(x, 0)$, σ_i^2 is the noise and interference power and p_{wf} is the water-filling level, that must hold $\sum_{k \in \mathcal{K}_b} \sum_{l=1}^{L} p^l_{b,k} = P^{max}_{prb}$, and P^{max}_{prb} is the maximum transmission power per PRB. Note that the notation used so far is for a single PRB, but it is applicable to a set of different PRBs. For the sake of simplicity, as done in [103], in the following we assume equal transmitted power per PRB, i.e., $P^{max}_{prb} = \frac{W_{prb}}{W_b} P^{max}_b$, where P^{max}_b stands for the maximum transmitted power per BS, W_b is the BS bandwidth and W_{prb} is the bandwidth of a PRB.

Moreover, in our downlink channel model, we adopt Adaptive Modulation and Coding (AMC) over any radio link. Consequently, the appropriate SINR $\gamma^l_{b,k}$ will eventually define the MCS that will be used over the link. The standard defines a discrete set of MCSs with the following possibilities in the downlink for data transmission: QPSK ($\frac{1}{8}$, $\frac{1}{5}$, $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$), 16-QAM ($\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$) and 64-QAM ($\frac{2}{3}$, $\frac{3}{4}$, $\frac{4}{5}$) and 256-QAM (where gNBs also adopt the same constellation mapping as in LTE SCs) [108]. Based on a target bit error rate, the MCS is selected by the BS according to the SINR $\gamma^l_{b,k}$ received by user k from BS b. In that sense the expected transmission rate per PRB of a user k connected to BS b depends on the applied MCS. The mapping between requested data rate and MCS is executed as indicated in the offline look-up table in [109].

Focusing on a particular user $k \in \mathcal{K}$ cooperatively served by a set of BSs $\mathcal{F}_k$, each cooperating BS serves a fraction of the traffic equal to $\beta_{b,k} \cdot R_k$, with $\sum_{b \in \mathcal{F}_k} \beta_{b,k} = 1$, i.e., each RRH serves a fraction of the traffic $\beta_{b,k} = \frac{1}{|F_k|}$ for a particular user. Therefore, the bandwidth required by BS b to get user $k \in \mathcal{K}^s$ served is given by

$$\begin{aligned} \mu_{b,k} &= \frac{\beta_{b,k} \cdot R_k}{C_{b,k}} \\ &= \frac{\beta_{b,k} \cdot \lambda^s \cdot d^s}{L \cdot \log_2(1 + \gamma_{b,k})}. \end{aligned} \tag{5.6}$$

It is further noted that based on the imposed assumption only RRHs can cooperate.

Therefore, if a user k is served by a distributed SC b, it holds that $\mathcal{F}_k = \{b\}$ and $\beta_{b,k} = 1$. Note that the number of users served simultaneously by a BS over the same spectrum equals $\frac{N_{bs}}{L}$. Based on the used nomenclature, if the set of users completely or partially served by BS b is denoted by $\mathcal{K}_b$ and the subset of users served by BS b and with service s is denoted as $\mathcal{K}_b^s$, then the required bandwidth is

$$\begin{aligned} \mu_b &= \frac{L}{N_{bs}} \sum_{k \in \mathcal{K}_b} \mu_{b,k} \\ &= \frac{L}{N_{bs}} \sum_{s \in \mathcal{S}} \lambda^s \cdot d^s \left(\sum_{k \in \mathcal{K}_b^s} \frac{\beta_{b,k}}{L \cdot \log_2(1+\gamma_{b,k})} \right) \\ &= \frac{1}{N_{bs}} \sum_{s \in \mathcal{S}} \lambda^s d^s K_b^s \theta_b^s, \end{aligned} \tag{5.7}$$

where K_b^s is the cardinal of the set $\mathcal{K}_b^s$ and $\theta_b^s = \frac{1}{K_b^s} \sum_{k \in \mathcal{K}_b^s} \frac{\beta_{b,k}}{\log_2(1+\gamma_{b,k})}$.

5.2.4 Processing load in the BBU pool

The functional split of C-RAN approach consists of the physical separation of BBU and RRHs. With this functional split, C-RAN centralizes most of the cell functions onto a pool of virtual BBUs run in a General-Purpose Processor (GPP) and deploys RRHs connected through high capacity links (i.e., usually though optical fibre connections) [21, 110]. The BBU centralization has multifarious advantages to support advanced interference management techniques, such as enhanced Inter-Cell Interference Coordination (eICIC) and CoMP. However, it also poses challenges in dimensioning FH connections and GPP [111, 112].

The quantification of the load caused by each user in the BBU pool depends on factors such as the bandwidth, the processing platform, the number of allocated PRBs and the Modulation and Coding Scheme (MCS) used by each user. In [113] a general model for the total load incurred by a given user is introduced. Following the rationale stated in [113], we denote the BBU load of a RRH b as η_b. This total load is split up into cell-specific processing load (i.e., η_b^c) and user specific load (i.e., $\eta_{b,k}^u$). The former depends on the bandwidth and the processing platform, whereas the latter mainly depends on the PRBs allocated to the user and the MCS. Based on the dependencies described in [113], the cell-specific processing load is modeled as

$$\eta_b^c \approx \zeta W_b + \phi, \tag{5.8}$$

where ζ is a constant that links the bandwidth allocated to BS b and the resulting computational burden, and ϕ is a constant that depends on the GPP processing architecture.

Similarly, the user specific load can be expressed as

$$\eta^u_{b,k} \approx \omega n_{b,k} W_{prb} \log_2 (1 + \gamma_{b,k}), \tag{5.9}$$

where ω is a constant value, W_{prb} is the PRB bandwidth, $\gamma_{b,k}$ is the SINR received by user k served by BS b, and $n_{b,k}$ is the number of PRBs allocated to user k. Note that $\eta^u_{b,k}$ is proportional to the number of PRBs ($n_{b,k}$) and to the MCS. Although in LTE-A the MCS range is a discrete set, in (5.9) we approximate the MCS by the capacity (i.e., in bit/s) using the Shannon Capacity formula. It is indicated that the higher the SINR, the higher the MCS used and therefore the higher the spectral efficiency. Thus the combination of (5.8) and (5.9) yields

$$\begin{aligned} \eta_b &\approx \eta^c_b + \sum_{k \in \mathcal{K}_b} \eta^u_{b,k} \\ &\approx \zeta W_b + \phi + \omega \Psi_b, \end{aligned} \tag{5.10}$$

where Ψ_b is the total throughput in BS b. Using (5.1),

$$\eta_b = \zeta W_b + \phi + \omega \sum_{s \in \mathcal{S}} \overline{\beta^s_b} K^s_b \lambda^s d^s, \tag{5.11}$$

where $\overline{\beta^s_b}$ is the average $\beta_{b,k}$ ratio $\frac{1}{|\mathcal{F}_k|}$, where $|\mathcal{F}_k|$ denotes the set of RRHs serving a user k according to the definition given in (5.6).

Therefore, the total processing load caused by the set of users connected to BS b, denoted as $\mathcal{K}_b$, is given by

$$\begin{aligned} \eta_b &= \sum_{k \in \mathcal{K}_b} \eta_{b,k} \\ &= \sum_{k \in \mathcal{K}_b} [(\theta + \xi + \omega W_{prb} \log_2(1 + \gamma_{b,k})) n_{b,k} + \phi] \\ &= K_b \phi + (\theta + \xi) n_b + \omega \sum_{k \in \mathcal{K}_b} \beta_{b,k} R_k \\ &= K_b \phi + (\theta + \xi) n_b + \omega \sum_{s \in \mathcal{S}} \lambda^s d^s \sum_{k \in \mathcal{K}^s_b} \beta_{b,k}. \end{aligned} \tag{5.12}$$

5.2.5 FrontHaul characterization

The FH connects the BBU pool and the set of RRHs, and it is usually implemented with high capacity fiber optics due to the stringent capacity requirements. In general, the Common Public Radio Interface (CPRI) is the protocol used to encapsulate baseband signals before the transmission between BBUs and RRHs [114]. The CPRI defines the

so-called Antenna Carrier (AxC) concept as the data required to transmit the In-phase and Quadrature (IQ) data flow (i.e., user plane data) of one antenna for one carrier. Thus, the required transmission rate to transmit an AxC is calculated as [115]

$$R_{AxC} = 2 \cdot N_{IQ} \cdot f_s \cdot r_w \cdot r_c, \tag{5.13}$$

where N_{IQ} is the number of bits used in the quantization process of the In-phase and the Quadrature-phase (i.e., between 8 and 20 bits/sample, though it is usually set to 15 bits/sample), 2 is a multiplication factor to account both In-phase and Quadrature-phase data, f_s is the sampling frequency, r_w is a correction factor due to control data (i.e., a basic frame is composed of 16 words, where only 15 out of them are used for data, i.e., $r_w = 16/15$), and r_c is the line coding factor (line coding with 8B/10B or 64B/66B, i.e., $r_c = 10/8$ or $r_c = 66/64$). The sampling frequency in LTE is determined based on the bandwidth, where $f_s = \{1.92, 3.84, 7.68, 15.36, 23.04, 30.72\}$ MHz for total bandwidth equal to $W_b = \{1.25, 2.5, 5, 10, 15, 20\}$ MHz, respectively [115]. Specifically, the sampling frequency can be expressed as $f_s = \upsilon_s \cdot W_b$, where $\upsilon_s = 1.536$ and W_b is the bandwidth allocated to RRH b. Given that R_{AxC} is defined for a single antenna, the required transmission rate for RRH b with N_{bs} antennas is given by

$$\begin{aligned} R_b^{fh} &= N_{bs} \cdot R_{AxC} \\ &= 2 \cdot N_{bs} \cdot N_{IQ} \cdot \upsilon_s \cdot W_b \cdot r_w \cdot r_c. \end{aligned} \tag{5.14}$$

For instance, a RRH with two antennas, allocated bandwidth of 20 MHz and 15 bits per sample (i.e., $r_w = 16/15$ and $r_c = 10/8$), requires FH with 2.45 Gb/s capacity. This stringent requirement, along with the high deployment cost of fiber optics makes C-RAN architecture unfeasible in some scenarios [111, 116]. The delay introduced by the FH between the BBUs pool and RRH b is henceforth denoted by T_b^{fh}. However, in general this delay can be neglected in high capacity optical connections.

5.2.6 BackHaul characterization

The BH connects the aggregation point (i.e., also known as Point of Presence, PoP) and the distributed standalone SCs [117]. The connection is characterized by a transmission rate, hereafter denoted by R_b^{bh}, a transmission delay T_b^{bh} and a scheduling policy. Since the BH can be implemented with a wide range of technologies, such as wireless or x Digital Subscriber Line (xDSL) [48, 118], none of the technologies is precluded. However, BH capacity is assumed to be constrained.

The BH transports data between the aggregation point and the SCs, but it must also exchange control signals [119]. Let us characterize the signaling between the distributed SC b and the CN by its arrival rate, denoted by λ_b^I, and the average signaling packet size d^I (i.e., the mean size of the signaling packets is assumed to be equal in all distributed SCs). Signaling traversing the BH can be decomposed into X2 U/C-Plane (i.e., communication among BSs, particularly connected to handover procedures or Almost Blank Subframe process), S1 C-plane and Transport protocol overhead [120]. Given the difficulty of modeling the signaling overhead, it is usually expressed as a percentage of the S1 U-plane traffic (i.e., data traffic); therefore, the signaling throughput $\Psi_b^I = \lambda_b^I d^I$ is given by $\Psi_b^I = \alpha_I \Psi_b$, where $\Psi_b = \sum_{s\in\mathcal{S}} K_b^s \lambda^s d^s$[1] is the data throughput of BS b and $\alpha_I < 1$. Thus,

$$\lambda_b^I = \frac{\alpha_I}{d^I}\Psi_b = \frac{\alpha_I}{d^I}\sum_{s\in\mathcal{S}} K_b^s \lambda^s d^s. \tag{5.15}$$

The BH is modeled as a non-preemptive queue system with $S+1$ priority classes (i.e., S services and signaling). In the proposed model, neither the arrival distribution nor the packet size distribution are known *a priori*. However, only loose upper and lower bounds for the performance metrics are known for G/G/1 queueing systems (i.e., general packet arrival and size distributions). Thus, for the sake of simplicity and according to [119], the M/G/1 system is used to analyze the BH (i.e., Poisson distributed arrivals and any packet size distribution). The scheduling policy implemented in the queue prioritizes packets based on their requirements. Hence, signaling packets have the highest priority. As for the services, we assume without loss of generality, that service s has a higher priority than j if $s < j$, with $s, j \in \mathcal{S}$. It can be easily shown that the expected time spent in the queueing system of a packet of service s is given by [121]

$$\tau_b^s = \frac{d^s}{R_b^{bh}} + \frac{\frac{1}{2(R_b^{bh})^2}\sum\limits_{j\in\mathcal{S}\cup\{I\}} K_b^j \lambda_b^j (\mathbb{V}ar(d^j) + (d^j)^2)}{\left(1-\rho_b^I - \sum\limits_{j<s}\rho_b^j\right)\left(1-\rho_b^I - \sum\limits_{j\leq s}\rho_b^j\right)}, \tag{5.16}$$

[1] As it is assumed that distributed SCs do not implement CoMP, for a SC b it holds that $\beta_{b,k} = 1$ $\forall k \in \mathcal{K}_b$. Therefore, $\Psi_b = \sum_{s\in\mathcal{S}} K_b^s \lambda^s d^s$.

where $\rho_b^j = K_b^j \lambda_b^j d^j / R_b^{bh}$, $\rho_b^I = \lambda_b^I d^I / R_b^{bh}$ and $\mathbb{V}ar(d^j)$ is the variance of the packet size of service j. Using (5.15), (5.16) can be rewritten as

$$\tau_b^s = \frac{d^s}{R_b^{bh}} + \frac{\frac{1}{2(R_b^{bh})^2} \sum_{j \in \mathcal{S}} K_b^j \lambda_b^j d^j \varphi_b^j}{\left(1 - \sum_{j \in \mathcal{S}} (\mathbb{1}_{(j<s)} + \alpha_I)\rho_b^j\right)\left(1 - \sum_{j \in \mathcal{S}} (\mathbb{1}_{(j \leq s)} + \alpha_I)\rho_b^j\right)}, \tag{5.17}$$

where $\mathbb{1}_{(j<s)}$ is the indicator function, equal to 1 when the condition $(j < s)$ is true and 0 otherwise, and

$$\varphi_b^j = \left(\frac{\mathbb{V}ar(d^j)}{d^j} + d^j\right) + \alpha_I \left(\frac{\mathbb{V}ar(d^I)}{d^I} + d^I\right). \tag{5.18}$$

As $\rho_b^j = K_b^j \lambda_b^j d^j / R_b^{bh}$, (5.17) can be written as

$$\begin{aligned} \tau_b^s = & \frac{d^s}{R_b^{bh}} + \frac{\frac{1}{2R_b^{bh}} \sum_{j \in \mathcal{S}} \rho_b^j \varphi_b^j}{\left(1 - \sum_{j \in \mathcal{S}} (\mathbb{1}_{(j<s)} + \alpha_I)\rho_b^j\right)\left(1 - \sum_{j \in \mathcal{S}} (\mathbb{1}_{(j \leq s)} + \alpha_I)\rho_b^j\right)} \\ = & \frac{d^s}{R_b^{bh}} + \frac{1}{2R_b^{bh} \vartheta_b^s \left(\vartheta_b^s - \rho_b^s\right)} \sum_{j \in \mathcal{S}} \rho_b^j \varphi_b^j, \end{aligned} \tag{5.19}$$

with $\vartheta_b^s = 1 - \sum_{j \in \mathcal{S}} (\mathbb{1}_{(j<s)} + \alpha_I)\rho_b^j$.

5.2.7 Functional Splits

In our scenario, each BS b has a particular functional split. All the tentative functional splits are discussed in [7] whereas the corresponding values of bandwidth and delay requirements for the respective transport network are defined in [26].

In principle, the two extreme cases of functional splits are the following: in D-RAN standalone SCs all the functions (i.e., PHY, MAC, RLC, PDCP, RRC and S1 transport) are implemented in the DU whereas in C-RAN RRHs the RF functionality is placed in the DU and upper layer functions are in the CU (i.e., Physical (PHY) - Radio Frequency (RF) split). The conditions for the BH that characterizes the SCs are described in section 5.2.6 and the conditions for the FH of the RRHs are described in section 5.2.5.

With regard to the 5G gNBs, a wide granularity of the functional split has been considered [7]. Most of the defined functional splits allow for having Radio Resource Management (RRM) functions like Call Admission Control (CAC) and Load Balancing in the

CU controlling multiple DUs. This permits increased efficiency in inter-cell coordination for RRM functions like the coordination of interference management, load balancing and CAC. However, not all functional splits will be adopted in real deployments comprising densely deployed gNBs, due to the implementation cost of the 5G FH [111, 116].

Therefore, in our study we adopt an intermediate functional split, Radio Link Control (RLC) - Medium Access Control (MAC), which places upper layer functions and RLC in the CU whereas the MAC and PHY layer are in the DU. This split allows resource sharing benefits for both storage and processor utilization and it has a low relative implementation cost [111, 116], thus making it eligible for future deployment. For gNBs in our scenario we choose specific values for T_b^{fh} and R_b^{bh} based on [7, 26].

5.3 NetSliC: Base Station Agnostic Framework for Network SliCing

A BS agnostic framework for network slicing is aimed to build a virtualization layer that abstracts the specificities of each BS (i.e., latency, BH / FH latency and/or bandwidth limitations), thus overcoming the intrinsic complexity of the network. In the context of networks as the one shown in Fig. 5.1, this abstraction process consists in guaranteeing the feasibility of each slice before its configuration, as well as its reconfiguration when required while enhancing the network spectrum utilization. To that end in this section we formulate the problem and define the conditions that should be fulfilled to create network slices that guarantee the distinct service demands. These conditions are the constraints that should apply to each BS irrespective of their transport and access requirements. In these complex deployments several bottlenecks appear with regard to the characteristics of the network components:

- Limited wireless access capacity (i.e., spectrum availability).
- BH limited capacity and latency in D-RAN architectures.
- Limited processing capacity in the BBU pool of C-RAN architectures.

Our proposed framework takes into consideration all these requirements, which are explained in detail in the following. In addition we describe the heuristic algorithm to address the formulated problem, NetSliC, as well as further implementation details that render it backward compatible with the existing standard.

5.3.1 Problem Formulation

In order to retrieve an efficient slice configuration that satisfies the conditions while maximizing the total spectrum utilization, we formulate the following optimization problem:

$$\max_{Q_{b,k}} \quad \sum_{b\in B, k\in K} \mu_{b,k} Q_{b,k} \tag{5.20}$$

$$\begin{aligned}
\text{subject to} \quad & Q_{b,k} \in [0,1], \forall b \in \mathcal{B}, \forall k \in \mathcal{K} && \text{(5.21a)}\\
& W_b \geq \mu_b, \forall b \in \mathcal{B} && \text{(5.21b)}\\
& \eta \leq \eta^{th}_{max}, \forall b \in \mathcal{B}_{\mathcal{C}} && \text{(5.21c)}\\
& \tau^s_b \leq t^s, \forall b \in \mathcal{B}. && \text{(5.21d)}
\end{aligned}$$

where $Q_{b,k}$ in (5.21a) is set to one if user k is admitted in BS b and zero otherwise. Constraint (5.21b) reassures that the requirements in the access network are fulfilled as further detailed in 5.3.2.1. Constraint (5.21c) ensures that a user will be associated to a particular BS while respecting the processing capability of the BBU as explained in 5.3.2.2. Finally constraint (5.21d) reassures that the service delay by each user is respected as further detailed in 5.3.2.3.

The problem of (5.20) is an NP-hard, non-linear, integer problem [122]. The solution of problem in (5.20) typically requires searching big search trees of possible configurations of BS types and user associations ($Q_{b,k}$) that result in different number of associated users per BS and therefore different spectrum usage values. In order to address the problem, while taking into consideration the facts that (i) certain BSs cooperate to serve users and (ii) the exhaustion of resources in a BS, in a BH connection or in the BBU pool should lead to re-association of particular users to create efficient slices in terms of spectrum usage, we propose a greedy heuristic scheme, for BS agnostic Network SliCing, NetSliC. In our proposal, NetSliC, as soon as we create an initial slice configuration[2], by occupying the required resources in each BS such that the first condition is satisfied for each service, we move on to fulfilling the next one, thus updating the previous slice configuration. To that end, a solution is never unique. We aim at finding an efficient slice configuration (i.e., in terms of spectrum usage) that satisfies each condition and further we relax this slice configuration in order to satisfy any following one.

[2]The term "slice configuration" in [7] refers to resource allocation / assignment into isolated slices per BS to achieve differentiated handling of traffic for services with different SLA.

5.3.2 Conditions

5.3.2.1 Condition 1: Bandwidth in the air interface

Initially, we fulfill the requirements in the access network (i.e., air interface, Uu) for each BS $b \in \mathcal{B}$ by reserving an initial set of radio resources, such that

$$W_b \geq \mu_b = \frac{1}{N_{bs}} \sum_{s \in \mathcal{S}} \lambda^s d^s K_b^s \theta_b^s, \tag{5.22}$$

where Wb is the total bandwidth available for BS b and μ_b is the bandwidth required by the same BS b (see expression (5.7)). The candidate serving BSs for a particular UE in our scenario is either a SC, a RRH or a gNB (i.e., $\mathcal{F}_k = \{b\}$, $\beta_{b,k} = 1$) or a set of RRHs performing CoMP (i.e., $\mathcal{F}_k = \{b_1, b_2, \cdots b_n\}$).

5.3.2.2 Condition 2: Processing Load in the BBU

In the following, we aim at accommodating the processing load in the BBU. This load is relevant only for the users that are served by the RRHs. Therefore, the total processing load caused in the BBU, by all the connected UEs to the RRHs in a deployment, is given by

$$\eta = \sum_{b \in \mathcal{B}_C} \eta_b, \tag{5.23}$$

where η_b is the computational burden caused by BS b as in (5.12). It is noted that both radio and computing resources across all RRHs are centrally processed by the same BBU pool. A user can be accommodated in a RRH if the following holds

$$\eta \leq \eta_{max}^{th}, \tag{5.24}$$

wherein η_{max}^{th} is the threshold set by the MNO based on the BBU capabilities.

5.3.2.3 Condition 3: Delay

Finally, we check whether the service delay requirement is satisfied for each BS. The FH of the RRHs is constrained by its maximum allowable rate R_b^{fh} and the maximum allowable delay T_b^{fh}. The BH respectively is constrained by R_b^{bh} and T_b^{bh}. It should be noted that the FH for the RRHs is characterized by ideal conditions (i.e., $T_b^{fh} \approx 0$ms), whereas the BH for the SCs is non-ideal, i.e., $T_b^{bh} \geq 0$ ms. The 5G FH for the gNBs is also characterized by non-ideal conditions, such that $T_b^{fh} \geq 0$ ms, as defined in section 5.2.7.

The expected time spent in the queueing system of a packet of service s is denoted by τ_b^s (5.19). Each service has a predefined maximum allowable delay for a packet of service s, i.e., t^s. For the RRHs it holds that $\tau_b^{s,fh} = \tau_b^s + \tau_{sch}$, where τ_{sch} is the expected delay related to scheduling. Therefore it holds

$$\tau_b^s \leq t^s. \tag{5.25}$$

as defined in (5.21d). Let us also define the time to timeout for each service $s \in \mathcal{S}$ in BS $b \in \mathcal{B}$ as $\delta_b^s = t^s - \tau_b^s$. We define a certain allowable threshold for this metric, which depends on the traffic type, i.e., $\delta_b^{s,th}$. Hence, the third condition that should hold is also expressed as

$$\delta_b^s \geq \delta_b^{s,th}. \tag{5.26}$$

A user $k \in \mathcal{K}^s$ cannot be served in cell $b \in \mathcal{B}$ if $\delta_b^{s,th}$ is violated.

5.3.3 Algorithm description

Based on the previously defined constraints / conditions we now propose our solution: a greedy heuristic algorithm for BS agnostic Network SliCing, NetSliC. Algorithm 1 is the pseudocode summarizing the creation of network slices in terms of finding an efficient slice configuration per BS that fulfills the distinct service requirements and on the same time increases the spectrum usage. This process is the outcome of iteratively satisfying the previously defined conditions.

Let us describe how NetSliC operates at a high level. Initially, each user presents their particular service requirements, i.e., (λ^s, d^s, t^s) to the network. First we create a candidate BS list for each user $k \in \mathcal{K}$ by sorting the candidate serving BSs for each user according to the achieved SINR $\gamma_{b,k}^l$ (i.e., see Algorithm 1, step 2). Thus, we create a list for each user with the optimal serving BSs in terms of SINR.

The initial slice configuration is done using the legacy SINR-based Slicing (SINR-S); the resources of each BS are statically divided among the distinct service types and the users are associated with the BS with the highest SINR (i.e., see Algorithm 1, steps 3 - 4).

Then we allocate the required radio resources in each BS such that condition 1, i.e., (5.22), is satisfied. In order to achieve this purpose we offload the users with maximum throughput consumption to the next candidate BS until (5.22) holds (i.e., see Algorithm 1, steps 5 - 9).

Algorithm 1 NetSliC

1: Input: (λ^s, d^s, t^s)
2: Initialize: $\forall k \in \mathcal{K}$ sort $b \in \mathcal{B}$ with $\max(\gamma^l_{b,k})$ ▷ Create candidate BS list based on SINR for each UE
3: Step 1: SINR-based slice configuration
4: Associate $k \in \mathcal{K}$ with $b \in \mathcal{B}$ wherein $\max(\gamma^l_{b,k})$
5: Step 2: Air Capacity based slice configuration
6: $\forall b \in \mathcal{B}$ Check condition 1 ▷ (5.22)
7: **repeat** $\forall b \in \mathcal{B}$
8: **if** (5.22) false **then** Associate k with $\max(\mu_{b,k})$ with the next candidate BS from the candidate list
9: **until** (5.22) holds
10: Step 3: Processing Load based slice configuration
11: $\forall b \in \mathcal{B}_C$ Check condition 2 ▷ (5.24)
12: **repeat** $\forall b \in \mathcal{B}_C$
13: **if** (5.24) false **then** Associate k with $\max(\eta^u_{b,k})$ with the next candidate BS from the candidate list
14: **until** (5.24) holds
15: Step 4: Delay based slice configuration
16: $\forall b \in \mathcal{B}$ Check condition 3 ▷ (5.26)
17: **repeat** $\forall b \in \mathcal{B}$
18: **if** (5.26) false **then** Associate k with $\max(\delta^s_b)$ with next candidate BS from the candidate list
19: **until** (5.26) holds

In the following, in order to fulfill the processing load requirements in the BBU pool, i.e., (5.24), we update the previous slice configuration. In steps 10 - 14 we associate the users with the highest processing load consumption in the BBU to the next candidate BS (i.e., SC or gNB) from the initial candidate list. In this phase we check only the RRHs since only these nodes result to BBU processing load. It should be pointed out that threshold η^{th}_{max} (i.e., condition 2) is set by the MNO and the capabilities of the BBU.

Finally, we check all the BSs to be sure that (5.26) is satisfied (i.e., steps 15 - 19). According to our third condition, a user is dropped if $\delta^{s,th}_b$ is violated (i.e., if δ^s_b expires). In general, in our system users are dropped as soon as the queue starts to grow, and therefore τ^s_b increases. For the BSs that (5.26) does not hold, we update the previous slice configuration iteratively, by offloading the users with maximum delay to their next candidate BS, such that (5.26) holds. The threshold $\delta^{s,th}_b$ (i.e., condition 3) is set by the MNO based on their traffic policies and the subscriptions / SLAs for each service type.

All in all if the described conditions are satisfied a user k is served, otherwise the user is dropped. It is further noted that the exhaustion of resources in a BS, in a BH connection or in the BBU pool leads to a re-association / offloading of specific users, thus balancing

the load. To that end, we arrive at the final slice configuration that fulfills all the constraints.

NetSliC is defined as a network slicing solution aimed to face the heterogeneity of the network, including the heterogeneity of the traffic requirements and the diversity of BSs types; therefore, not only throughput and processing load are considered in the algorithm but also delay and technology limitations. As it can be observed in Algorithm 1, also the delay of each node and the ability of a specific technology to support a traffic type is considered.

When NetSliC is run, those users that can not meet the required SLA are dropped. This is the reason why the number of dropped users is able to capture how often users are unable to meet the SLA requirements or, in other words, to which extent the network creates customized slices able to serve the traffic. NetSliC is then able to balance and steer the traffic based on the imposed requirements and the network and nodes' restrictions / capacity.

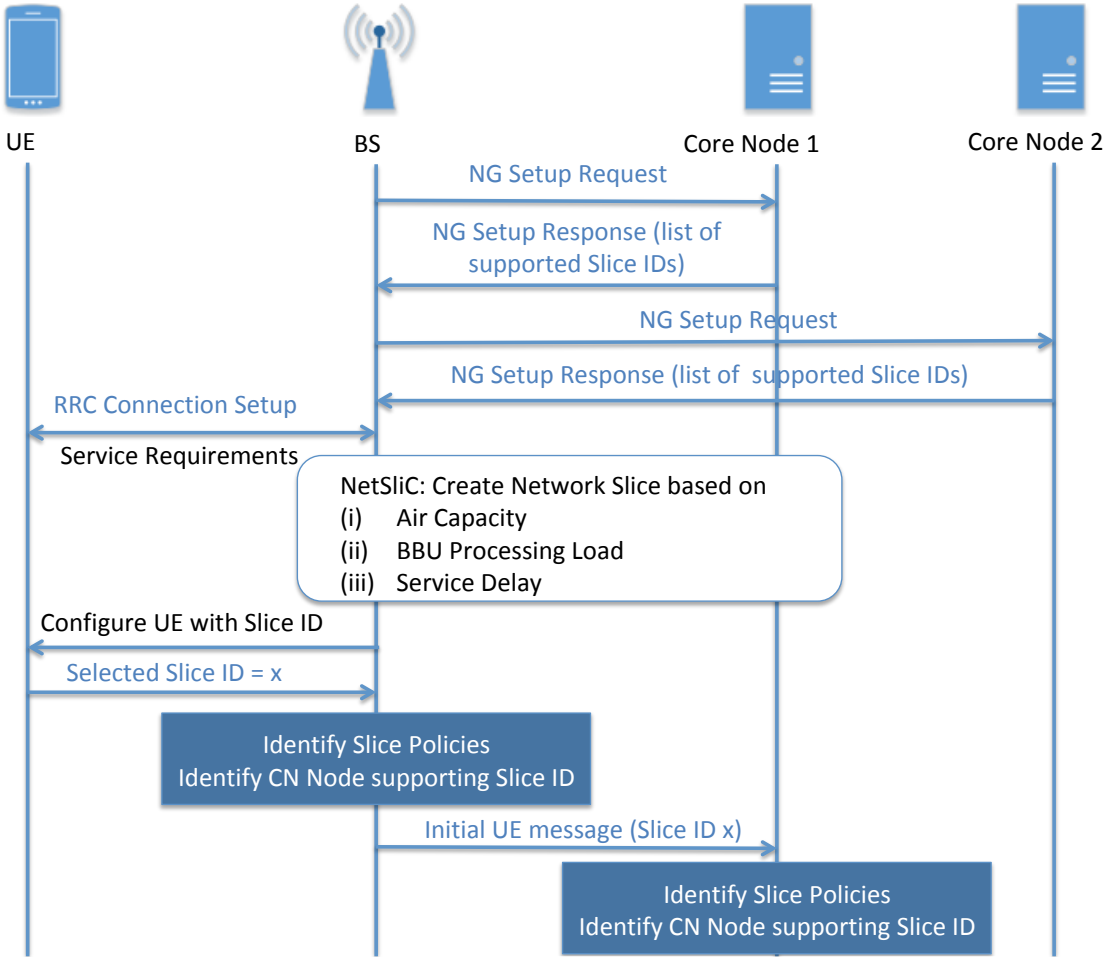


FIGURE 5.2: Message flow for implementing NetSliC in the standard [6].

5.3.4 Complexity and Convergence

The asymptotic computational complexity of Algorithm 1 is $\mathcal{O}(B \log B)$. In the initialization phase of NetSliC sorting all BSs of the set $\mathcal{B}$ is of complexity $\mathcal{O}(B \log B)$ where $B = |\mathcal{B}|$. Then in each one of the four steps of NetSliC the maximum amount of users completely or partially associated with BS b is $\max(K_b)$. The sequential repeat loops require $\mathcal{O}(\max(K_b)B) = \mathcal{O}(B)$ each in the worst case. Since the steps take place sequentially the final complexity is calculated as follows: $\mathcal{O}(B \log B) + \mathcal{O}(B) + \mathcal{O}(B) + \mathcal{O}(B) + \mathcal{O}(B) = \mathcal{O}(B \log B)$.

It is noted that for realistic scenarios, NetSliC provides results in acceptable executable time. We would like to further comment that the the runtime overhead depends on factors such as the computational platform on which the algorithm is run or the number of BSs. In a nutshell, it depends on the scenario and on the computational capacity of the Management and Orchestration (MANO) entity, which is responsible for mapping the service requirements established in a Service Level Agreement (SLA) into the elements of the three architecture layers to create the network slices. In that sense, and given the importance of the matter, the computational complexity as presented above is a common approach to analyze the runtime overhead.

Additionally we would like to discuss further about the time scale of certain functionalities related to our proposal. In fact, admission control and user association process in LTE / LTE-A systems normally need tens of seconds to complete a user handover. To that end NetSliC is compatible with scenarios that will consist of legacy LTE-A SCs along with future gNBs. NetSliC can be executed periodically, or triggered when certain conditions are met (e.g., when new or handover users are requesting admission, or the channel has varied significantly). The events in a real system where mobility is considered have very different time granularities. For instance, scheduling of radio resources is conducted in intervals of milliseconds (e.g., 1 ms in LTE-A), link adaptation to radio conditions is done in intervals of hundreds of milliseconds to seconds (e.g., milliseconds in LTE), new data is generated at intervals of seconds and users move from or into a cell at tens of seconds or minutes.

With regard to convergence, it is pointed out that NetSliC always finds a solution, since its execution is terminated when all conditions have been checked. A step σ represents the index of necessary iterated applications of NetSliC so that a solution gets extracted. Thus the solution that NetSliC extracts at a step σ, is a slice configuration for each BS with particular user associations that result into the same (i.e., or improved, in comparison to step $\sigma - 1$), spectrum usage (b/s) respecting the constraints in the air capacity, the BBU pool and the delay imposed by each service type. A solution is

always found since NetSliC terminates its execution if a substantially identical value of throughput (i.e., while respecting the constraints in the air capacity, the BBU processing load and the service delay) is found repeatedly. Results on the convergence of our solution are illustrated in section 5.4.4.

5.3.5 Implementation Details

Fig. 5.2 depicts how we integrate our solution in the standard [7], by providing the proper service slice to each UE independently of the type of BS (i.e., SC, RRH or gNB). It is pointed out that NetSliC is implemented at the level of the BS, as shown in Fig. 5.2. When NetSliC is applied, network slices are created according to the processing load, traffic or delay within the list of candidate BSs. In particular, with regard to the SCs and gNBs our solution is applied at the BS level. Regarding the RRHs connected to the BBU, NetSliC is applied directly at the BBU since all the upper layer functions are placed therein.

After creating the slices with NetSliC, network slice selection is realized by configuring the UEs with a list of slice identifiers (IDs) to which they are allowed to access. Therefore, each UE of a distinct service type has access only to a subset of resources within each BS according to the defined slice. When the UE requests to access the BS it presents a slice ID. By receiving the slice ID, the BS is able to identify the policies that apply to the selected slice and assign radio resources accordingly. In the CN the BS identifies the CN node that supports the slice ID presented by the UE. To select the appropriate configuration for the traffic for each network slice, RAN receives a slice ID indicating which of the configurations applies for this specific network slice [7].

5.3.6 Discussion on Optimality Degree

We would like to further analyze optimality issues with regard to our proposal. First of all NetSliC provide an efficient slice configuration that satisfies the requirements of a particular set of users, as we also noted in chapter 5.3.1. We use the term efficient slice configuration to describe a final configuration that increases the number of admitted users and therefore enhances the spectrum usage. In its user association part, NetSliC takes into consideration the three defined conditions and as soon as they are satisfied a final slice configuration is determined. The final slice is also determined by the re-association / offloading of certain users (e.g., users that have high data rate demands, high processing load contribution in the BBU and not satisfied service delay requirements). This solution (i.e., slice configuration) is not unique and it varies based on the introduced traffic load, amount and involved BS category (i.e., SCs, RRHs or gNBs).

It should be pointed out that optimality constitutes a general issue for heuristic algorithms, thus for NetSliC as well. This is because, even though heuristic schemes can accelerate the searching process of a solution, optimality is not guaranteed [123]. A greedy algorithm, as the name suggests, always makes the choice that seems to be the best at that moment. This means that it makes a locally-optimal choice in the hope that this choice will lead to a globally-optimal solution. The objective of a heuristic is to produce a solution in a reasonable time frame that is good enough for solving a problem at hand. This solution may not be the best of all the solutions to this problem, or it may simply approximate the exact solution. In this context NetSliC always finds a solution, since its execution is terminated when no further user associations take place. This yields the final spectrum usage and BBU processing load values. The characteristics of the involved BSs as well as the UE demands determine the final slice configuration which satisfies the input set of users and their particular requirements. More specifically, NetSliC falls into the category of greedy heuristic algorithms, since its approach is to follow a certain sequence of steps and to make a choice of decisions among a class of possible ones (i.e., decisions) at each stage.

5.4 Performance Evaluation

In this section, we demonstrate the effectiveness of NetSliC from the overall system and slice viewpoint by simulation. Within the simulation, we compare the performance of our scheme with other benchmark schemes for network slicing from several aspects. In addition we prove the convergence of NetSliC both for a static and a mobile scenario.

5.4.1 Simulation Scenario and Parameters

In the following we present in detail the simulation environment of this chapter. First let us define the service requirements of each user in terms of (λ^s, d^s, t^s). We study a scenario with two types of services: Guaranteed Bit Rate (GBR) traffic VoIP (64 Kb/s) with $d^s = 1280$ bits every 20 ms ($\lambda^s = 50$ packets/sec) with acceptable SLA $t^s = 45$ ms [124] and Best Effort (BE) FTP (300 Kb/s) with $d^s = 300000$ bits generated every 100 ms ($\lambda^s = 100$ packets/sec), $t^s = 1000$ ms allowable delay. We consider that the former has low latency requirements and the latter high data rate. The rate of VoIP transmission is relatively low, and the main requirement is ensuring a high reliability level, with a Probability Error Rate (PER) typically lower than 10^{-5} [125]. The aim of FTP service is to maximize the data rate, while guaranteeing a moderate reliability, with PER on the order of 10^{-3} [126].

We conduct Monte-Carlo extensive simulations (with a thousand iterations to achieve statistical validity) in a custom made simulation tool implemented in MATLAB®. The deployment under study is Urban Micro (UMi), based on IMT Advanced evaluation guidelines for IMT-2020 [127], consisting of scalable number of SCs with transmission power 35 dBm, RRHs connected to BBU and medium range gNBs with intermediate RLC-MAC functional split with transmission power 46 dBm, in space area of 2000 m × 2000 m.

We consider distant dependent path-loss and shadow fading in our channel model. Thus, due to different considered SCs, RRHs, gNBs and users locations, path-loss varies depending on the distance between a user and a BS. Fading losses that we consider are extracted randomly, derived by a log-normal function with a standard deviation of 8 dB (as in [127]). Taking into consideration all the different iterations, we provide average results (i.e., mean values).

We use the following path-loss model $PL_{UMi-LOS} = 32.4+21log_{10}(d[m])+20log_{10}(f_c[GHz])$ where $f_c = 2.5$ GHz for SCs and RRHs and $f_c = 6$ GHz for gNBs [127]. This model holds for 0.5 GHz $\leq f_c \leq$ 100 GHz. This renders it suitable for higher frequencies that will be key characteristic for future 5G deployments. The noise power spectral density and noise figure are respectively set to -174 dBm/Hz and 9 dB [127].

In this study we adopt the values of [26, 116] to define the transport network bandwidth and delay for each BS (i.e., SC, RRH and gNB). The functional split where the BBU is totally centralized in the cloud corresponds to C-RAN (i.e., C-RAN RRH with $T_b^{fh} = 250\mu s \approx 0$ ms and $R_b^{fh} = 2457.6$ Mb/s) whereas the split where all the functionalities are in the DU corresponds to the traditional D-RAN (i.e., LTE-A SC with $T_b^{bh} = 30$ ms and $R_b^{bh} = 151$ Mb/s) [26]. With regard to the gNBs with RLC-MAC split, it holds that $T_b^{fh} = 6$ ms and $R_b^{bh} = 151$ Mb/s [26].

To that end, NetSliC has an overview of the physical nodes and their transport and access links and translates these into slices according to the offered services. Thus, in a scenario composed of three different types of access nodes, i.e., SCs, RRHs and gNBs, NetSliC manages three sets of resources: (i) a resource pool for RRHs, instantly updated (ii) a resource pool for gNBs, which is updated every $T_b^{fh} = 6$ ms and (iii) each standalone SC with constrained resources and non-ideal transport network with delay $T_b^{bh} = 30$ ms.

5.4.2 Overall Network Throughput

5.4.2.1 Average Values

We present mean values of the numerical simulation results illustrating the tradeoffs between the gains in terms of the achieved overall network throughput (a metric indicated in [26]) and the processing load occurred in the BBU pool. We consider a scenario where we increase the offered load (i.e., 50% VoIP and 50% FTP traffic) in a deployment consisting of 10 SCs and 10 RRHs. In this setting, we compare the performance of NetSliC with the following two baseline schemes for network slicing:

- SINR Slicing (SINR-S): the resources of each BS are statically divided among the distinct service types wherein the users are associated with the BS with the highest SINR [128].
- Minimum Rate Slicing (MIN-RATE-S): users of each service type are associated with the BS that guarantees the minimum transmission rate requirements [126].

We evaluate our proposal with regard to the following parameters:

- Average throughput that is defined as the average amount of correctly transmitted data over a time period.
- Baseband Unit pool processing load, which is defined as the number of operations per time period required to serve the traffic in a scenario with a C-RAN architecture (i.e., RRHs connected to a centralized Baseband Unit pool through a FH connection).

With these two metrics, it is possible to understand how much traffic the network is able to serve and, simultaneously, how high the centralized processing load is. It is worth noting that both metrics are relevant in the sense that the former defines the capacity of the network, whereas the latter determines the dimensioning of the central processing unit.

The results are presented in Fig. 5.3. In principle we observe a significant throughput gain (i.e., increase in spectrum usage), e.g., for 25.48 Mb/s medium offered load in Fig. 5.3(a), NetSliC with $\delta_b^s = 2$ ms for VoIP outperforms the baseline SINR-S with 50.15% gain and the MIN-RATE-S with 26.91% gain. The reason is that none of baseline algorithms (SINR-S and MIN-RATE-S) has the ability to flexibly update the slice creation based on the particularities of traffic while abstracting the specificities of each BS. In

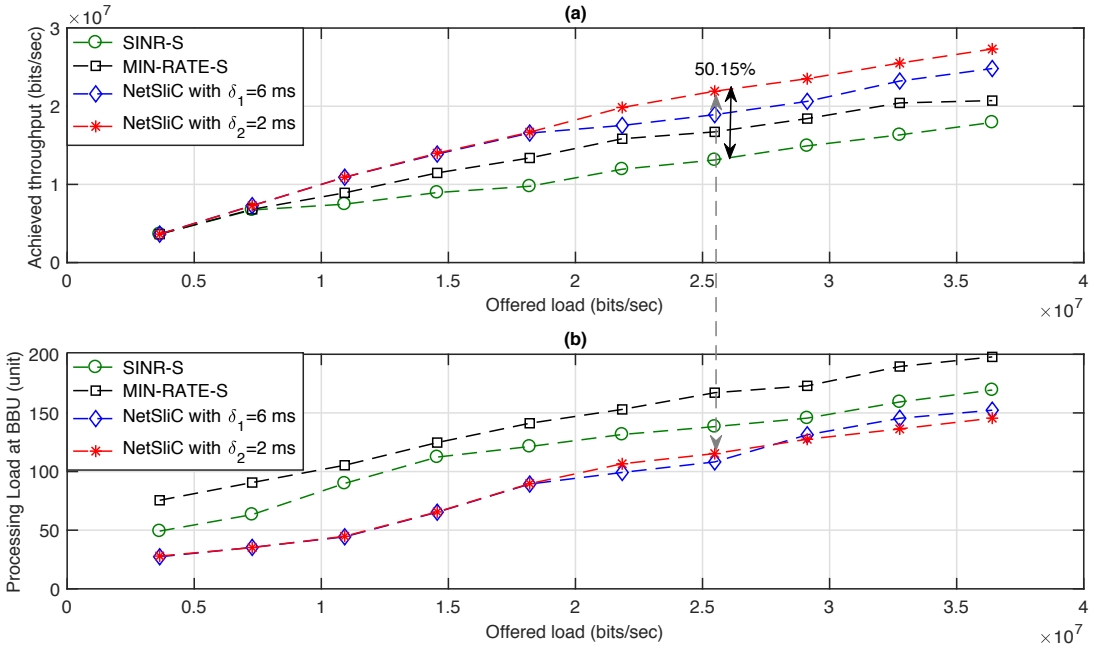


FIGURE 5.3: (a) Overall Network Throughput vs. (b) Processing load in the BBU while serving VoIP and FTP (50% - 50%) traffic with different δ in a deployment with 10 SCs and 10 RRHs and comparison with baseline schemes.

particular in SINR-S the capacity is reserved based on the BS and user position. To that end, there is no consistent approach of how resources are occupied per BS and therefore spectrum usage efficiency is sacrificed. In MIN-RATE-S users have the tendency to connect to RRHs, which in general are able to satisfy the minimum rate requirements. This may result to higher throughput for certain offered loads but with the tradeoff of having heavier burden in the BBU pool. In particular in Fig. 5.3 we confirm that although MIN-RATE-S achieves higher throughput than SINR-S, this has an impact on the BBU pool. However in NetSliC, the capacity of the SCs is firstly reserved for VoIP (i.e., service with low latency requirements) due to the application of the third condition that fulfills the VoIP latency demands. Then the RRHs cooperate performing CoMP to accommodate as many FTP users as possible. Thus NetSliC performs an adaptation to the slice configuration based on the transport and access characteristics of each BS as well as the service requirements. In section 5.4.3 we perform a thorough performance evaluation on the slice behavior. We further point out that in Fig.5.3 we denote the parameter δ only for the VoIP traffic, since it has stricter latency requirement in comparison with BE FTP.

Besides the overall network gain, it is also interesting to look at the processing load occurred in the BBU in Fig. 5.3(b). NetSliC saves up to 30.57% of BBU resources, for medium offered load of 25.48 Mb/s, in comparison to the baseline schemes because of

serving the VoIP traffic by using the capacity of the SCs. We note that MIN-RATE-S, which creates slices that guarantee the minimum transmission rate, has the heaviest impact on the BBU pool due to the fact that users are mostly associated with the RRHs. We also observe in Fig. 5.3(b) that after 29.12 Mb/s offered load the increase in network throughput has a substantial burden in the processing load. Therefore, this trade-off shall be taken into account by the MNO. The values in Fig. 5.3(b) can be used by the MNOs to decide about the required installed BBU pool computational capacity (and the consequent threshold η_{max}^{th} setup) based on the traffic that they wish to serve and the cost of the RRH deployment.

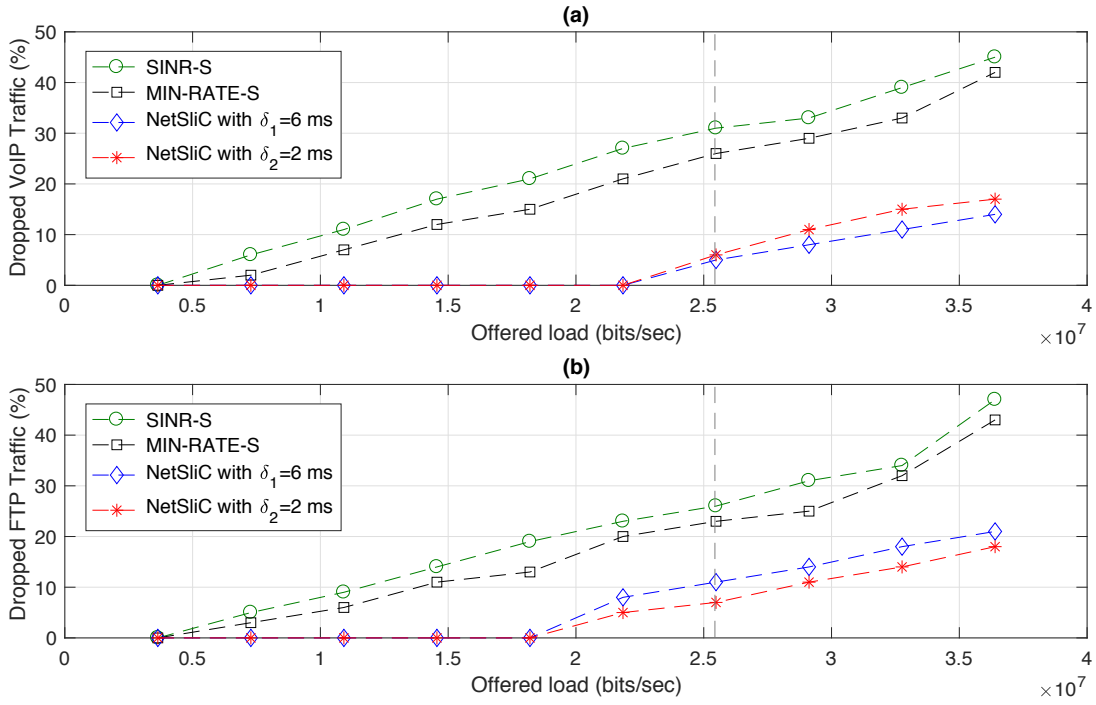


FIGURE 5.4: (a) Dropped VoIP and (b) FTP traffic while serving VoIP and FTP (50% - 50%) traffic with different δ in a deployment with 10 SCs and 10 RRHs and comparison with baseline schemes.

In Fig. 5.4(a) we observe that NetSliC with higher δ_b^s protects VoIP traffic (i.e., less dropped VoIP users) by performing a more conservative approach in network slicing. On the other hand when NetSliC uses lower δ_b^s, network resources are used more efficiently but at the same time when the network is loaded it is more prone to drop users (i.e., precarious approach to network slicing). In Fig. 5.4(a), we observe that when we run NetSliC with lower $\delta_b^s = 2$ ms, VoIP dropping increases (up to 6% for medium offered load 25.48 Mb/s) whereas in Fig. 5.4(b) we note that FTP dropping decreases (resulting in 7% dropped FTP for medium offered load). This is because NetSlic with higher $\delta_b^s = 6$ ms assures the QoS for VoIP traffic by preferring to drop BE FTP traffic instead. Although NetSliC with lower $\delta_b^s = 2$ ms achieves higher throughput as shown in Fig. 5.3(a), when the system is loaded more VoIP users are dropped.

Despite the fact that as the input traffic increases the network drops more users, NetSliC still outperforms the baseline schemes. The value δ_b^s in (5.26) is set by the MNO based on their traffic policies and the subscriptions / SLAs for each service type. Therefore, the MNO can set δ_b^s, to differentiate the treatment of distinct service types, according to the desired level of certainty in assuring QoS.

5.4.2.2 95% Confidence Interval

We also run the same scenario by including the 95% confidence interval of the simulations. As it can be observed in Fig. 5.5, the deviation of the simulation results with regard to the mean value is very small. As an example, we include Fig. 5.5 in the following:

5.4.3 Slice Performance

5.4.3.1 SCs and RRHs

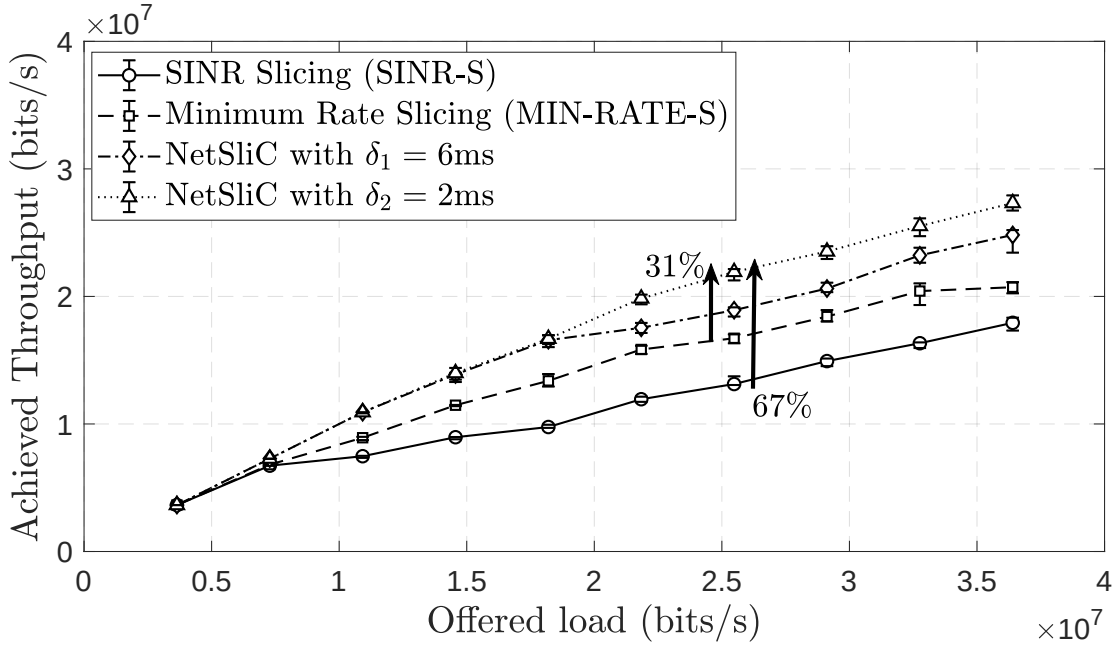
We now elaborate on slice performance by studying the average throughput per slice when we run NetSliC with $\delta = 2$ ms in a deployment consisting of 10 SCs and 10 RRHS. Fig. 5.6 presents the average traffic served by each slice in order to show the type of BS that NetSliC chooses to serve the different types of users.

Fig. 5.6(a) shows that the BH of the SCs is reserved for the VoIP (i.e., low latency requirements), whereas the ideal FH of the RRHs is used to allow CoMP, to boost throughput serving the BE FTP service (i.e., higher data rate requirements), as shown in Fig. 5.6(b). The reason lays in the third condition, (5.26), applied by NetSliC; our solution serves VoIP users in standalone SCs to satisfy the low latency requirements.

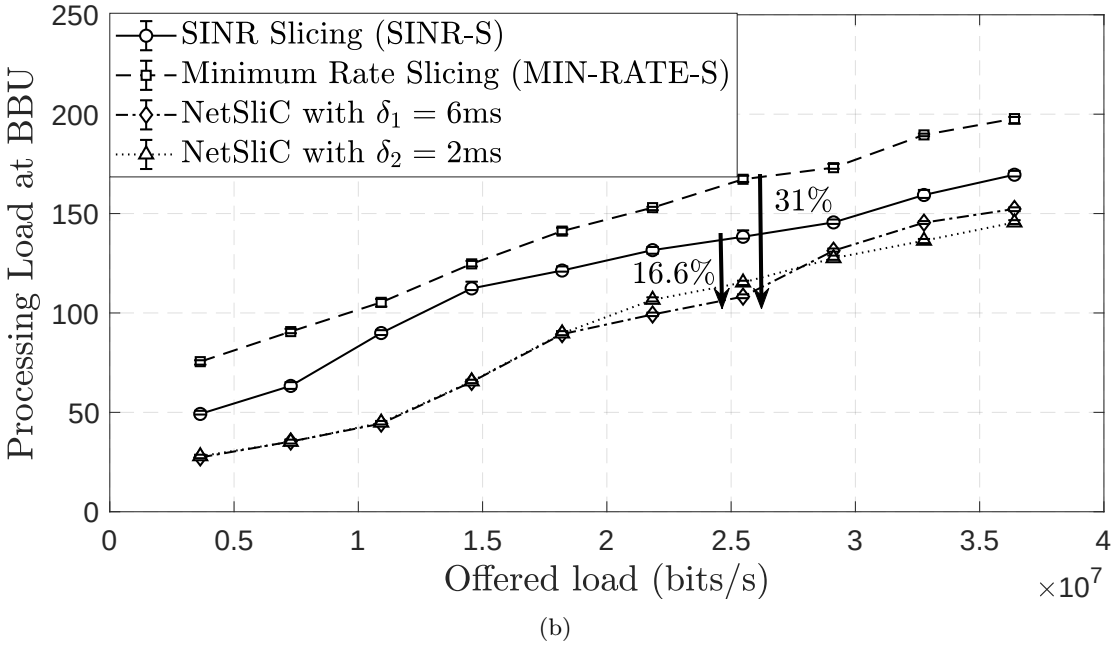
For instance, in the scenario with 19.8 Mb/s achieved throughput in Fig. 5.3 (i.e., medium offered load 21.84 Mb/s) 93% of the VoIP traffic is served by the SCs (i.e., 3.57 Mb/s) whereas the RRHs perform CoMP and serve 87% of FTP traffic (i.e., 13.88 Mb/s) that has higher data requirements.

5.4.3.2 SCs, RRHs and gNBs

We study the performance of NetSliC with $\delta = 2$ ms in a deployment composed of 10 SCs, 5 RRHs and 5 gNBs. Fig. 5.7 summarizes the results with regard to the average throughput per slice, to show how the introduction of gNBs in this dense deployment affects the served traffic.



(a)



(b)

FIGURE 5.5: (a) Overall Network Throughput vs. (b) Processing load in the BBU while serving VoIP and FTP (50% - 50%) traffic with different δ in a deployment with 10 SCs and 10 RRHs and comparison with baseline schemes. Mean values and 95% confidence interval.

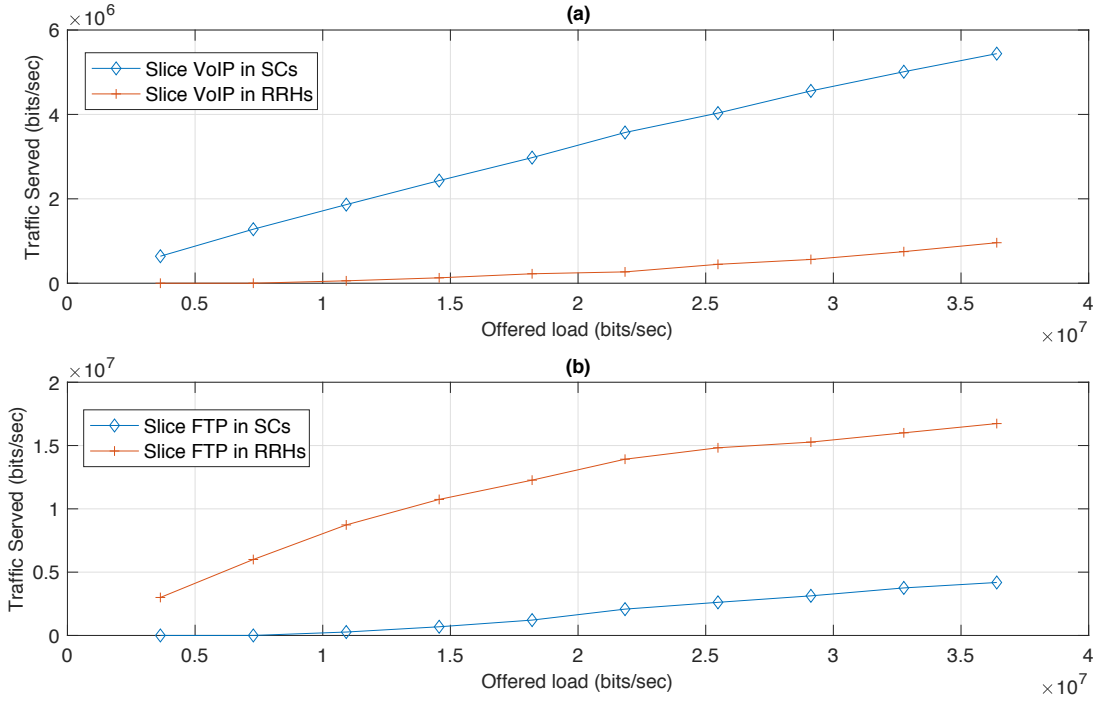


FIGURE 5.6: Average traffic throughput per slice in a deployment with 10 SCs and 10 RRHs for VoIP (a) and FTP (b).

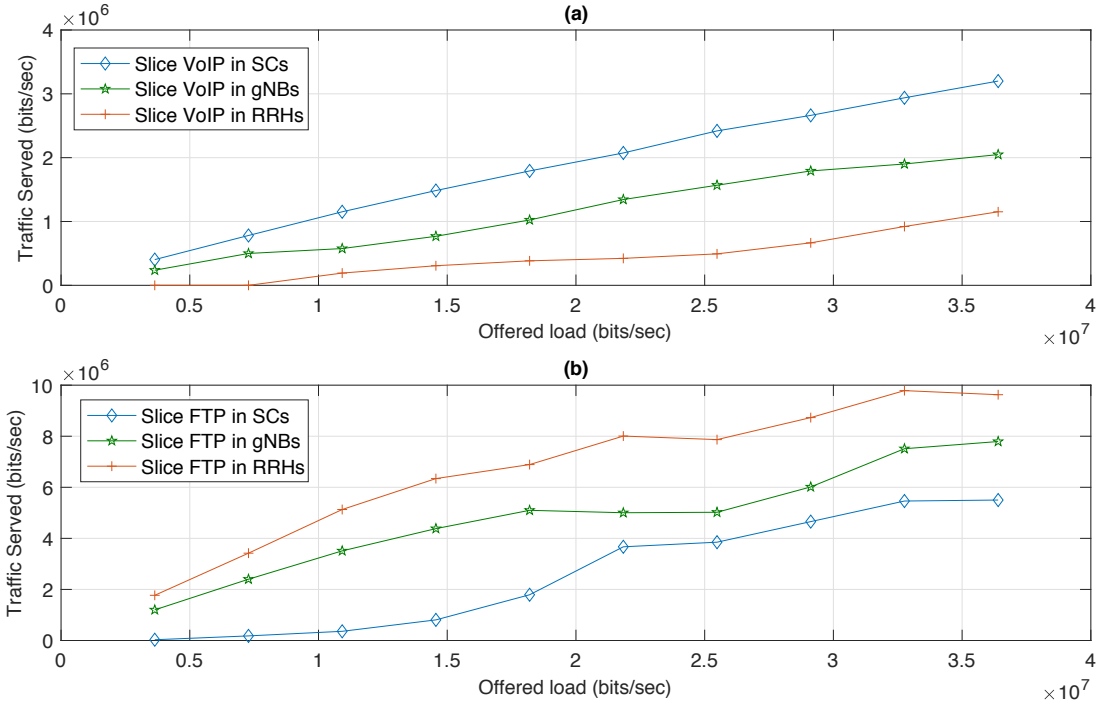


FIGURE 5.7: Average traffic throughput per slice and type of BS in a deployment with 10 SCs, 5 RRHs and 5 gNBs for VoIP (a) and FTP (b).

By observing figures 5.6 and 5.7 we draw the conclusion that certain services such as the VoIP, which has low latency requirements, require a certain split to support them. We observe that NetSliC decides that VoIP service requires most RAN functions to run on the DU to fulfill the latency requirements (i.e., SC and gNB with intermediate functional split). Then a decision that enhances throughput by aggregating RRHs and performing CoMP is taken for the FTP slice (i.e., it requires a higher centralization). This is because NetSliC satisfies the delay requirements of each service type, thus prioritizing VoIP users in this particular scenario.

For example, in Fig. 5.7(a) we observe that for medium offered load, (i.e., 21.84 Mb/s), the main burden of creating network slices that satisfy VoIP low latency requirements is shared between the standalone SCs and the gNBs, with 54% (i.e., 2.07 Mb/s) and 35% (i.e., 1.34 Mb/s) of the VoIP throughput respectively (i.e., total 3.84 Mb/s). In addition, we inspect that NetSliC accommodates VoIP service in gNBs to improve the BH delay for the SCs since the load is low. Fig. 5.7(b) shows that a smaller number of RRHs results into less cooperation and therefore less BE FTP traffic is served by the RRHs, i.e., 48% (i.e., 8 Mb/s).

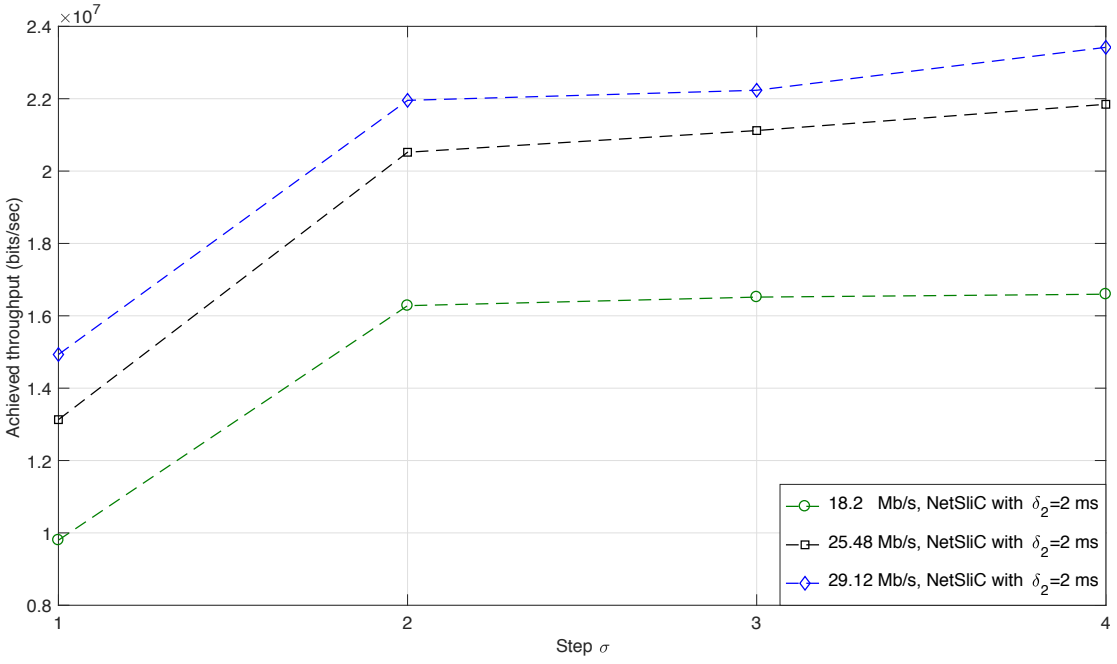


FIGURE 5.8: Convergence for a static scenario with different offered loads in a deployment with 10 SCs and 10 RRHs.

5.4.4 Convergence Study

5.4.4.1 Static Scenario

In principle, we study the convergence of NetSliC for a static scenario. We run a simulation on a deployment composed of 10 SCs and 10 RRHs for different traffic loads (i.e., 18.2 Mb/s, 25.48 Mb/s, 29.12 Mb/s). In Fig. 5.8 we present how many iterations / steps σ are needed to extract the slice configuration that yields the final spectrum usage / achieved throughput value (i.e., solution). On average we note that $\sigma = 4$ required steps for convergence for the scenario with static users, as explained in detail in Section 5.3.4.

5.4.4.2 Mobile Scenario

We further study a mobile scenario, which captures small-scale variations, where the users are moving at a reasonable speed but for small amounts of time (i.e., in seconds). In particular we provide results for moving pedestrian users with fixed and identical speed $v = 3$ km/h in randomly and uniformly distributed directions based on the mobility model for UMi, described in [127]. In our simulation, we introduce 25.48 Mb/s medium offered load (i.e., 50% VoIP and 50% FTP traffic). Each user chooses a random destination within the scenario, which consists of 10 SCs and 10 RRHs. The simulation duration is 1000 sec and we collect statistics every 50 seconds. The shadowing factor is given by a log-normal function with standard deviation of 8 dB (as in [127]) updated every second, and fast fading follows a Rayleigh distribution (i.e., dependent on user speed and angle of incidence [127]). It is noted that when mobility is considered, the slice configuration varies along time. While users are moving within the scenario, NetSliC readapts the slice configuration by allocating resources to the corresponding BS such that the maximum amount of users is satisfied. We further define the periodicity of triggering NetSliC as $\Delta(t)$ (i.e., how often we run NetSliC).

In Fig. 5.9 we present results with regard to the periodicity that we run NetSliC, $\Delta(t)$, and its effect on the percentage of dropped traffic. In Fig. 5.9 we observe that the choice of $\Delta(t)$ during mobility, affects the number of dropped VoIP users. Our first observation is that the lower the value of $\Delta(t)$ (i.e., the more often NetSliC is run), the higher the achieved VoIP throughput (i.e., the lower the number of dropped VoIP users). Quite similar values of dropped VoIP traffic are observed for $\Delta(t) = 10$ sec till $\Delta(t) = 50$ sec. On the contrary when NetSliC is triggered every $\Delta(t) = 100$ sec, we observe high percentage of dropped VoIP users. This is due to the fact that the more often we run NetSliC, the better tracking of the variations due to mobility and/or channel changes

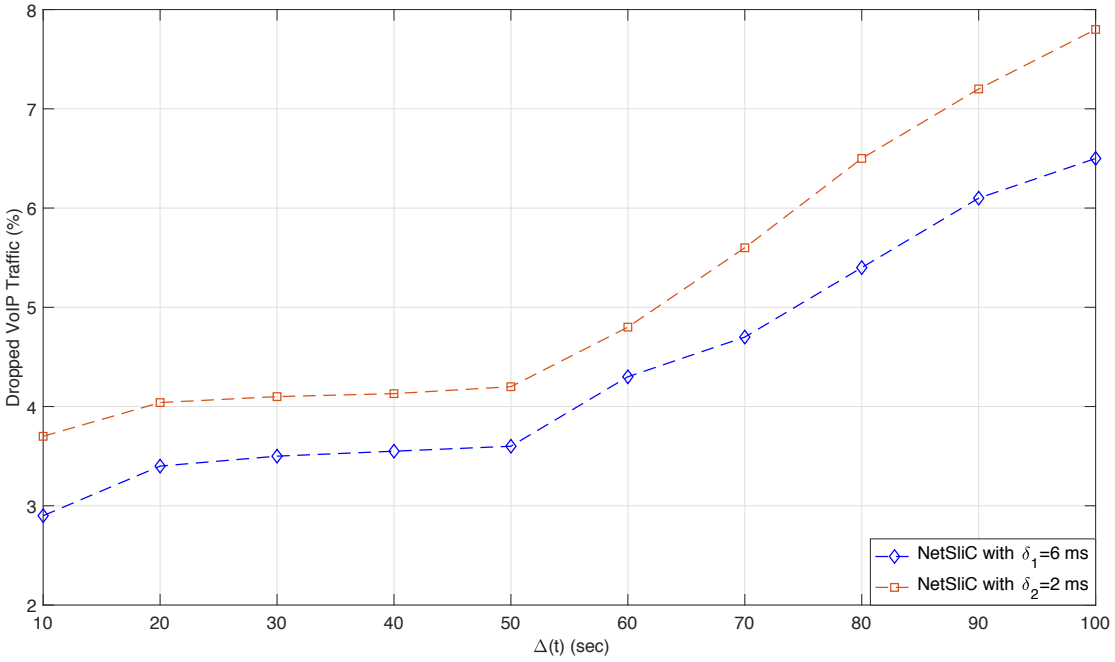


FIGURE 5.9: Choice of periodicity of triggering, $\Delta(t)$, in a scenario with mobility (25.48 Mb/s offered load) in a deployment with 10 SCs and 10 RRHs.

can be done. In this experiment we focus only in VoIP traffic since it presents low (i.e., tight) latency requirements. It is interesting to notice that the cooperating RRHs help NetSliC being triggered less often with satisfactory percentage of VoIP dropped users (e.g., $\Delta(t) = 50$ ms) despite the changes that take place due to user mobility. NetSliC prefers to keep cooperating RRHs serving a particular moving FTP user whereas VoIP traffic is connected to the standalone SCs to satisfy the stringent delay requirements. To that end, it is left open to the MNO to choose the periodicity of triggering $\Delta(t)$ for NetSliC in case of user mobility; Fig. 5.9 can be used as a guideline this purpose.

Furthermore we would like to further elaborate on the impact of mobility patterns on the dropped traffic when using our solution. NetSliC is designed to deal with mobility and intrinsic mobility. It is noted that when mobility is taken into account, the slice configuration varies along time. Variation in mobility parameters such as user speed and direction determine how often the slice configuration changes but do not affect the operation of our solution. Therefore different mobility patterns will result to slice configurations that alter in different pace. Additionally it is worth commenting on the existence of hotspots when mobility is considered. When a lot of user traffic is concentrated in hotspots NetSliC copes with the service dynamics leveraging the characteristics of the diverse BSs. Therefore traffic is accommodated by creating the most efficient in terms of spectrum usage and BBU processing load slice even when traffic is concentrated in particular hotspots.

The creation of network slices ensures that the proper amount of resources is reserved per BS based on the SLA of the specific traffic, including transport network capacity and processing load in the BBU pool when needed. As it can be observed, the inherent flexibility of NetSliC and the ability to adapt to different traffic profiles and BS capabilities leads to higher achieved throughput (i.e., and lower BBU processing load) even when mobility is considered.

Fig. 5.10 presents the dropped VoIP and FTP traffic for the described mobile scenario under study along time (sec). In particular in Fig. 5.10 we observe that while users are moving within the scenario (i.e., total 1000 sec of movement are observed), NetSliC readapts the slice configuration by allocating resources to the corresponding BS such that the maximum amount of users is satisfied. We further notice that NetSliC still protects VoIP traffic that has low latency requirements. Although in Fig. 5.10(a) the differences between NetSliC with $\delta_1 = 6$ ms and NetSliC with $\delta_2 = 2$ ms are small, we observe that when δ_b^s is higher (i.e., $\delta_1 = 6$ ms), less VoIP users are dropped. In particular and as confirmed by Fig. 5.10(b), NetSliC with $\delta_1 = 6$ ms prefers to drop FTP users instead.

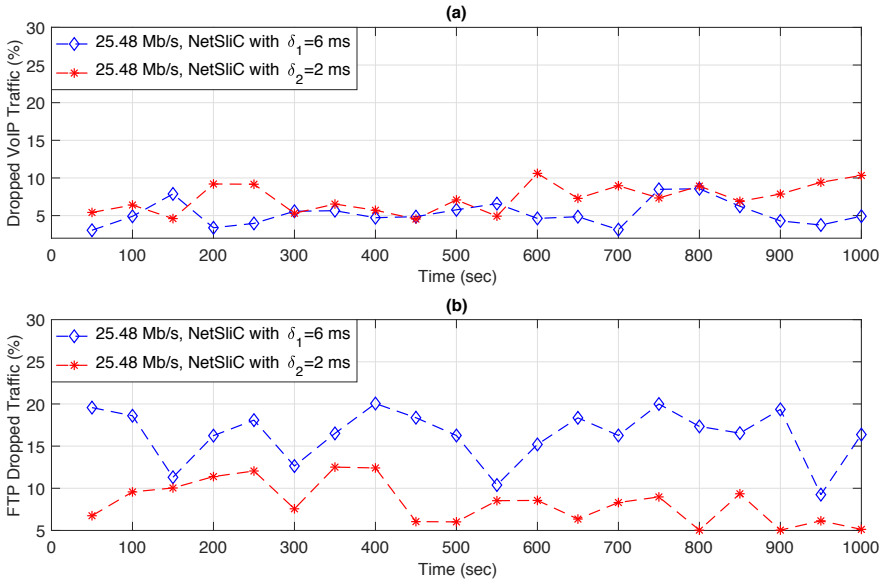


FIGURE 5.10: (a) Dropped VoIP and (b) FTP traffic while serving mobile VoIP and FPT users (50% - 50%) with different δ in a deployment with 10 SCs and 10 RRHs.

It is further noted that different network slices have different characteristics and requirements in terms of mobility, latency and reliability. For instance, in railway communications, many handovers could be triggered by a high-speed train during a short time; while in IoT applications, reliable and/or low-latency communications should be guaranteed for many devices with low or no mobility.

5.5 Conclusion

In this chapter, we have proposed NetSliC, a BS agnostic scheme that creates network slices taking into account the distinct service requirements. The proposed framework considers the bottlenecks in the air capacity, BH and FH transport network capacity and delay as well as the delay requirements imposed by the different service types. The extensive performance assessment has revealed interesting tradeoffs between throughput and processing load in the BBU pool. In the simulated scenario, composed of heterogeneous nodes, it has been shown that the average throughput gain can reach around 67%, while the BBU pool processing load drops around 16.6% compared with the baseline SINR-based proposal. Although the average throughput gain with respect to MIN-RATE-S is lower (around 30%), this is translated into a larger gain in terms of BBU pool processing load (around 30%).

Furthermore, we underline the fact that the standalone SCs, having functions in the DU, are used for the low latency service network slices whereas the ideal RRHs are used to perform CoMP and enhance throughput of the data rate demanding services. In addition, NetSliC in a deployment with gNBs assists the standalone SCs by creating network slices that satisfy low latency service requirements (i.e., VoIP) and therefore sharing the burden of serving this traffic type.

Chapter 6

Conclusions and Future Challenges

This chapter summarizes the main contributions of this dissertation, while it also provides some potential research lines for future investigation. Section 6.1 contains the most significant remarks from each chapter, while section 6.2 reveals some open issues in relation to the contributions of the present thesis.

6.1 Conclusions

One of the greatest challenges for cellular networks in the near future is their scalability and sustainability. In this thesis, we have made a first attempt to answer the question of what constitutes virtualization and network slicing in the RAN.

With regard to the first research direction, presented in Chapter 3, we proposed RENEV, a complementary solution to the state of the art, which covers gaps found therein by introducing a new dimension in RAN virtualization at the BS level. In our work we allow BSs that belong to two tiers (i.e., macro cell and small cells) to reallocate underutilized spectrum to other BSs. Our scheme considers the coordination among several BSs to create an abstraction of system radio resources, so that multiple BSs with load variations can be served, in a heterogeneous environment. The extensive performance assessment has revealed that gains in system throughput are translated into gains for the user throughput as well. With the use of RENEV, system resources are dynamically distributed according to users needs on an isolated and on demand basis. We have also evaluated the solution for the signaling overhead that adds into the network for increasing number of SCs per cluster.

For the support of our second research direction, presented in Chapter 4, we focus on introducing a capacity broker in 3GPP network sharing architecture, while introducing on-demand resource allocation via the means of signaling extensions of 3GPP network sharing management. This solutions solves the problem of how to achieve balanced sharing of resources in an architecture shared by several MVNOs. In addition, by leveraging traffic non-uniformities in a shared deployment, we proposed MuSli, a framework to be implemented by the capacity broker in coarse time scales. Along with our proposal, we introduced a decoupling process to extract variation trends in irregular traffic patterns and improve traffic forecasting. MuSli, by deciding how to slice the deployment capacity among two types of requests (i.e., Guaranteed QoS and BE), improves network performance by (i) increasing the accepted requests, and (ii) decreasing the underutilized resources. Our results can be leveraged by infrastructure owners, to flexibly allocate capacity to tenants, considering different types of services and the uncertainty of expected traffic.

Finally in the last research direction of this thesis, presented in Chapter 5, we focus on how to create a virtualization layer for managing resources between BSs with different characteristics in an agnostic manner. The idea behind our proposal is that emerging and future architectures eventually become more complex; MNOs decide whether to deploy new access nodes (i.e., RRHs or gNBs) or use the legacy ones (i.e., SCs) based on the profit and the scalability of the network. To that end we propose NetSliC that considers the bottlenecks in the air capacity, BH and FH transport network capacity and delay for each BS as well as the delay requirements imposed by the different service types. The extensive performance assessment has revealed interesting tradeoffs between throughput and processing load in the BBU pool. Furthermore, we underline the fact that the standalone SCs, having functions in the DU, are used for the low latency service network slices whereas the ideal RRHs are used to perform CoMP and enhance throughput of the data rate demanding services. In addition, NetSliC in a deployment with gNBs assists the standalone SCs by creating network slices that satisfy low latency service requirements (i.e., VoIP) and therefore sharing the burden of serving this traffic type.

6.2 Future Challenges

The research contributions presented in this work can be the starting point of new research lines for investigation. Similar to other emerging technologies, there is no doubt that RAN virtualization and slicing brings forward a significant potential, but also introduces several technical and business challenges. The future cellular networks

will focus on the different business applications and user experience other than just the pursuit of the greater bandwidth and volume. This will raise the requirement to build service oriented networks to quickly and efficiently respond to user needs, as well as to offer consistent and high quality services for different use cases. There are still numerous open research problems and implementation challenges to be addressed. In the following list, we summarize the main ideas that we have identified for future work:

- **Slice Security**: When it comes to slice security one size fits all model is not applicable. Although a fundamental premise of network slicing is that the network is carved into discrete, self-contained slices, in many cases each slice must still leverage network-wide resources. As such, while unique security parameters can be defined for network slices individually, there are security parameters that must be applied to shared network resources. As such, there is an open research direction to bridge the gap between a network-wide security policy and a security policy that must be applied to an individual slice. In particular, there is a high interest on how to enhance security and privacy protection for IoT services powered by 5G.

- **Network Slicing Techno-economics Aspects**: Network slicing offers MNOs a way to provide premium services to multiple customers from fields as diverse as public safety, industrial automation and healthcare. An early lesson from the modeling work is how critical it is for the operator's sales teams to understand deeply the value of 5G network slicing. Sales must be dedicated to finding customers with both the specialized requirements and an appreciation of the benefits of a service based on network slicing, indicating they would be willing to pay a premium price. Except from that there is an open research direction on what are the gains of network slicing and which are the advantages of functional splits for serving these premium users. Studies on the techno-economics aspects of these technologies will be very helpful to the MNOs when deciding whether to deploy them or not.

- **Further enhancements of the capacity broker**: Open issues lie in the field of the capacity broker. In particular further study on the degree of certainty in resource provisioning can be done, based on the density of the deployment and the variation of mobility.

- **Network Automation and integration of Artificial Intelligence**: Artificial Intelligence (AI) is the latest piece of the puzzle that telecom operators must put together as they evolve their networks from physical to cloud-based, virtualized infrastructures. The area where AI can make the biggest impact is operational automation: letting networks run themselves. Instead of spending substantial

amounts on managing, maintaining and fixing them, networks could become 'Zero touch'. 'Zero-touch deployment' and 'zero-touch provisioning' in the telecom industry extends this concept of automation beyond the initial installation phase to cover the entire lifecycle of network operations including planning, delivering, monitoring, updating and, ultimately, decommissioning of services. This will move telecom networks from today's automatic functions to fully-autonomous operations that bring significant top-line revenue improvements as well as sustainable reductions in operational costs.

- **End-to-end slice creation and management**: The creation of slices comprising both RAN and CN has not been successfully solved and there is plenty of room for research in this area.

- **Dynamic functional split**: The dynamic allocation of Network Functions (NF) is a current research topic. The joint optimal NF allocation and slice creation must be studied to fully exploit the potential of SDN driven cellular networks.

Concluding, this thesis has advanced the state of the art first by investigating the notion of RAN virtualization at the BS level, second by introducing the capacity broker context to create slices that accommodate the demands of several MVNOs and finally by proposing a BS agnostic virtualization layer that creates network slices for distinct traffic types in a scenario with various types of BSs with different access and transport characteristics. The three parts of the thesis were treated separately but a network were all parts can be combined is not precluded. The road ahead lies open for further study, following the new research lines that have been analyzed.

Appendix A

Appendix A: Calculations for Chapter 3

A.1 MCS Selection Probability

Let us denote by $x_i \in \mathbb{R}^2$ the location of BS_i and $y \in \mathbb{R}^2$ a random location in the scenario. The signal strength received from BS_i at location y, expressed in dB, may be written as $p_i(y) = P_{T_i} - L_i(y) - S_i(y)$, where P_{T_i} is the constant that includes antenna gains and transmitted power of BS_i, $L_i(y)$ is the path loss from x_i to y, and $S_i(y)$ is the slow fading. The SNR received at y from BS_i, when no interference is received, is given by $\mathrm{SNR}_i(y)_{dB} = p_i(y) - N_0$, where N_0 represents the noise average power. Throughout the rest of the analysis, taking into account the transmission power and coverage area of each BS as well as that subcarriers are not utilized by neighboring cells, we assume that interference is imperceptible among them [18]. Without loss of generality, the dependency of the several variables on the location y will be omitted in the sequel. Yet, all expressions are still derived for a random location y. Therefore, let SNR_{max} be the highest SNR received from a BS in B at a random location y, where $\mathrm{SNR}_{max} = \max_{BS_i \in B} \mathrm{SNR}_i$.

Focusing on the adaptive MCS mechanism, the k^{th} MCS is selected by BS_i if and only if $\mathrm{SNR}_k^{min} \leq \mathrm{SNR}_i < \mathrm{SNR}_k^{max}$, where SNR_k^{min} and SNR_k^{max} stand for the minimum and maximum thresholds of MCS k, respectively. Therefore, the probability of using a certain MCS could be expressed as:

$$P(\mathrm{MCS}_i = k) = \frac{P(\mathrm{SNR}_k^{min} \leqslant \mathrm{SNR}_i < \mathrm{SNR}_k^{max} \cap \mathrm{SNR}_i = \mathrm{SNR}_{max})}{P(\mathrm{SNR}_i = \mathrm{SNR}_{max})}. \tag{A.1}$$

Since the SNR of a particular BS_i is considered independent from the SNR of the rest BSs,

$$P(\mathrm{SNR}_i = \prod_{j \neq i} P(\mathrm{SNR}_i > \mathrm{SNR}_j) = \prod_{j \neq i} P(S_j > S_i + \mu_{ij}), \tag{A.2}$$

where $\mu_{ij} = P_{T_j} - P_{T_i} + L_i - L_j$. After a convenient change of variables, (A.2) is equal to $F_{S_i}(\frac{\sigma_i \cdot \mu_{ij}}{\sigma_j \cdot \sqrt{2}})$, where F_{S_i} denotes the Cumulative Distribution Function (CDF) of the random variable S_i expressing the shadowing, whereas σ_i and σ_j denote the standard deviations of the shadowing of BS_i and BS_j.

Correspondingly, the numerator of (A.1) is derived as $P(\mathrm{SNR}_k^{min} \leqslant \mathrm{SNR}_i < \mathrm{SNR}_k^{max} \cap \mathrm{SNR}_i = \mathrm{SNR}_{max}) = \prod_{j \neq i} P(\mathrm{SNR}_k^{min} \leqslant \mathrm{SNR}_i < \mathrm{SNR}_k^{max} \cap \mathrm{SNR}_i > \mathrm{SNR}_j)$. By substituting the values $S^0 = P_{T_i} - SNR_k^{max} - L_i$ and $S^1 = P_{T_i} - SNR_n^{min} - L_i$, the previous equation is expressed as follows:

$$P(S^0 \leqslant S_i < S^1 \cap S_j > S_i + \mu_{ij}) = (F_{s_i}(S^1) - F_{s_i}(S^0)) - \int_{S^0}^{S^1} F_{s_i}(s_i + \mu_{ij}) f_{s_i}(s_i) ds_i. \tag{A.3}$$

where $f_{s_i}(s_i)$ is the Probability Distribution Function (PDF) of S_i.

A.2 Derivation of throughput by users in SCs tier

For deriving the throughput achieved by the users located within the SCs tier with RENEV, let us divide the process according to the source that provides RBs to the Requesting BSs. First resources are redistributed within the SCs tier to serve the demanded traffic. In the case that these are not enough, resources are granted from the eNB. To begin with, SCs tier redistributes its RBs to accommodate the demanded traffic. If the overall traffic is less or equal to the SCs capacity, all users can be served. The overall resources within this tier, are equal to $RB_T = \sum_{i \neq 0} RB_i$. What is more, the average transmission rate for this case, equals $\mathbb{E}[R_{TOT}] = \frac{1}{1-a_0} \cdot \sum_{i \neq 0} a_i \cdot \mathbb{E}[R_i]$, where a_i denotes the percentage of users located within the coverage area of BS_i and $\mathbb{E}[R_i]$ the expected transmission rate in BS_i. Thus, if $\sum_{i \neq 0} X_i \cdot d \leq RB_T \cdot \mathbb{E}[R_{TOT}]$, all users located in the SC tier (i.e., $\sum_{i \neq 0} X_i = X \cdot (1 - a_0)$) will be served by the SCs' resources. It follows that $\sum_{i \neq 0} T_{R_i} = d \cdot \sum_{i \neq 0} X_i$.

Once SCs' resources (i.e., RB_T) are depleted, $X \cdot (1 - a_0)$ users within the Requesting BSs, will require further resources from the eNB tier. Therefore, the expected number of users to be served with resources from the eNB is $\mathcal{E} = X \cdot (1 - a_0) - \frac{RB_T}{d} \cdot \mathbb{E}[R_{TOT}]$. Thus, each Requesting BS_i will have to serve $\mathcal{E}_i = \frac{a_i}{1-a_0} \cdot \mathcal{E}$ users. Let us denote as

$\mathbb{E}[RB_i^s]$, the amount of resources from the eNB that can be given to each Requesting BS_i.

Based on this, a particular Requesting BS_i, will serve all this traffic (i.e., $\mathcal{E}_i \cdot d$) in the case where $\mathcal{E}_i \cdot d \leq \mathbb{E}[RB_i^s] \cdot \mathbb{E}[R_i]$ holds. In contrast, the traffic served by Requesting BS_i with RBs from the eNB tier will be equal to $\mathbb{E}[RB_i^s] \cdot \mathbb{E}[R_i]$. If $\mathcal{E}_i \cdot d$ is served, in total the throughput achieved within this tier, will equal $\sum_{i \neq 0} \mathcal{E}_i \cdot d = d \cdot \mathcal{E}$. Otherwise, it will be yielded by the summation of the traffic served in each Requesting BS with resources from the eNB (i.e., $\sum_{i \neq 0} \mathbb{E}[RB_i^s] \cdot \mathbb{E}[R_i]$). Consequently the throughput generated by the users in the SCs tier, served both with resources redistributed within the SCs tier and resources transferred from the eNB, will be equal to

$$\sum_{i \neq 0} T_{R_i} = \min \left(X \cdot (1 - a_0) \cdot d, RB_T \cdot \mathbb{E}[R_{TOT}] + \sum_{i \neq 0} \mathbb{E}[RB_i^s] \cdot \mathbb{E}[R_i] \right). \tag{A.4}$$

It remains to show how we calculate the number of eNB resources, that can be given to each Requesting BSs (i.e., $\mathbb{E}[RB_i^s]$), included in (A.4). For the sake of simplicity and without loss of generality, we can assume circular cluster's surface containing N circular shaped SCs. Therefore, the area of the cluster is $A = \pi \cdot R_c^2$, where R_c is the cluster radius, and $A_i = \pi \cdot R_{SC_i}^2$ holds for a circular shaped coverage area with Radius R_{SC_i}. For a particular Requesting BS_i, located randomly within the cluster, there will be an overlap if the distance between BS_i and another BS_j is less than $R_{SC_i} + R_{SC_j}$. Thus the probability of overlap among two Requesting BSs is derived as

$$P_o = \frac{\pi \cdot (R_{SC_i} + R_{SC_j})^2}{\pi \cdot R_c^2} = \left(\frac{R_{SC_i} + R_{SC_j}}{R_c} \right)^2. \tag{A.5}$$

Then, the probability for a Requesting BS_i of having n_i overlaps is described by a binomial random variable as follows:

$$P(n_i = n) = \binom{N-1}{n} \cdot P_o^n \cdot (1 - P_o)^{N-1-n}. \tag{A.6}$$

Although log-normal shadowing is considered, our assumption of circular coverage SCs has been validated by simulations (for 17 dBm SC transmission power, $\mu = 0$ dB and $\sigma_S = 10$ dB as indicated in Table 3.1). We assume that a Requesting BS_i with n_i overlapping BSs, receives$(RB_s \cdot \frac{1}{n_i+1})$ RBs. The expected value of this term is equal to

$$\mathbb{E}[\frac{RB_s}{n_i + 1}] = \sum_{n_i=0}^{N-1} \frac{RB_s}{n_i + 1} \cdot \binom{N-1}{n_i} \cdot P_o^{n_i} \cdot (1 - P_o)^{N-1-n_i}, \tag{A.7}$$

where a convenient change of variables can be applied, $m = n_i + 1$, so as (A.7) equals

$$\mathbb{E}[\frac{RB_s}{n_i+1}] = \frac{RB_s}{P_o} \cdot \sum_{m=1}^{N} \frac{(N-1)!}{m!(N-m)!} \cdot P_o^m \cdot (1-P_o)^{N-m} = \frac{RB_s}{NP_o} \cdot [1-(1-P_o)^N], \tag{A.8}$$

which is valid for all Requesting BSs since each one is assumed to receive $\frac{RB_s}{n_i+1}$. However this is not true in the case that each Requesting BS accommodates different portion of users (i.e., a_i). This difference in the Requesting BS load, implies different traffic demands and hence unequal percentage of resources to be allocated. Let us assume, as previously, that Requesting BS_i overlaps with n_i BSs. The number of the possible ways of overlapping equals $\binom{N-1}{n_i}$. Each of these ways can occur with probability $(P_o^{n_i} \cdot (1-P_o)^{N-1-n_i})$. Let us define the set $\mathcal{O}_{ic}^{n_i}$ as a particular set of n_i overlapping Requesting BSs with Requesting BS_i. All Requesting BSs in $\mathcal{O}_{ic}^{n_i}$, as well as Requesting BS_i, will share RBs according to the proportion of users that each one accommodates. Thus, the percentage of resources achieved per Requesting BS_i is equal to $\frac{a_i}{a_i + \sum_{BS_k \in O_{ic}^{n_i}} a_k}$. Therefore

$$\mathbb{E}[RB_i^s] = RB_s \cdot \sum_{n_i=0}^{N-1} P_o^{n_i} \cdot (1-P_o)^{N-1-n_i} \cdot \sum_{c=1}^{\binom{N-1}{n_i}} \frac{a_i}{a_i + \sum_{BS_k \in \mathcal{O}_{ic}^{n_i}}} \tag{A.9}$$

It should be noticed that for equal percentage of users in each Requesting BS (i.e., when $a_i = a_k$, $\forall i \neq k$), (A.9) is equal to (A.8).

A.3 Set of Feasible Future States

RENEV is intended to redistribute the unused resources of the possible Donor BSs among the Requesting BSs. Therefore, not all transitions from state S_n to state S_j are feasible. The set of feasible future states for a given state S_n, $\mathcal{F}(S_n)$, is defined as the set of states to which S_n could transit after performing RENEV. Based on the definition of states S_n, S_j and RENEV algorithm, the following conditions must be accomplished to assure that $S_j \in \mathcal{F}(S_n)$:

- The amount of resources is constant in the initial and the final states: $\sum_{k=1}^{r_{max}-r_{min}+1} s_{j,k} \cdot (r_{min}-1+k) = \sum_{k=1}^{r_{max}-r_{min}+1} s_{n,k} \cdot (r_{min}-1+k)$.
- After performing RENEV, the number of Requesting BSs should be smaller. Therefore, $n_R(S_j) < n_R(S_n)$.

- The number of requested RBs in the final state should be less than the corresponding number in the initial state: $\sum_{k=1}^{-r_{min}} s_{jk} \cdot (r_{min} - 1 + k) < \sum_{k=1}^{-r_{min}} s_{nk} \cdot (r_{min} - 1 + k)$.

- In state S_j (i.e., final state) there are not new Requesting BSs. Therefore, $\forall s_{n,k} = 0$ and $k \leq -r_{min}$, then $s_{j,k} = 0$. Likewise, $\forall s_{n,k} \neq 0$ and $k \leq -r_{min}$, it holds that $s_{j,k} \leq s_{n,k}$.

- The number of RBs transferred by the Donor BSs is equal to the number of RBs received by the Requesting BSs: $\sum_{k=2-r_{min}}^{r_{max}-r_{min}+1}(s_{n,k} - s_{j,k}) \cdot (r_{min} - 1 + k) = \sum_{k=1}^{-r_{min}}(s_{n,k} - s_{j,k}) \cdot (-r_{min} + 1 - k)$.

- The absolute value of the highest amount of requested RBs in the initial state (i.e., negative value) should be lower or equal than the minimum amount of available RBs such that if $k' = r_{max} - r_{min} + 1$, $\forall k \leq -r_{min}$, $\forall s_{n,k} \neq 0$ then $s_{jk} \neq 0$, $\forall k \geq |r_{min} - 1 + k'|$.

- As RENEV is completed, all possible redistribution of resources has been done. Therefore, there is not any possible Donor BS that could cover the needs of a Requesting BS. Thus, $\forall s_{j,k} \neq 0$ and $k \leq -r_{min}$, it is true that

$$\forall s_{j,k} \neq 0, k \leq -r_{min} \Rightarrow \sum_{m=-2r_{min}+2-k}^{r_{max}-r_{min}+1} s_{j,m} = 0. \tag{A.10}$$

A.4 Derivation of $P(Q = q|N, M)$

The probability that two BSs within the cluster are overlapping is derived in (A.5), denoted as P_o, and the probability that a specific BS_i in the cluster is overlapped with n BSs (no overlapping among different clusters is assumed), denoted as $P(n_i = n)$, is derived in (A.6). Note that, for a given state S_j, if BS_i is assumed to be a Requesting BS, the probability that a BS different from BS_i is a Requesting BS equals $P_{N,M} = \frac{M-1}{N-1}$, where $M = n_R(S_j)$. Let us denote with m_i, the number of Requesting BSs overlapping BS_i. Henceforth, $P_{RB}(m_i = m|N, M)$ denotes the probability that m Requesting BSs overlap BS_i, given that M out of N BSs are Requesting BSs it can be expressed as

$$P_{RB}(m_i = m|N, M) = \sum_{k=m}^{N-1} \left(P(n_i = k) \cdot \binom{k}{m} \cdot P_{N,M}^{m} \cdot (1 - P_{N,M})^{k-m} \right). \tag{A.11}$$

In RENEV, the eNB will only transfer the same resources to two different Requesting BSs if they do not overlap. Approximately, we could claim that the available resources of the eNB can be transferred to a specific SCs cluster, as many times as the number of non-overlapping groups of Requesting BSs. Therefore, we are interested in figuring out

the number of non-overlapping groups of Requesting BSs within the cluster, denoted as $Q = \{1, 2, \cdots M\}$. For instance, when all Requesting BSs overlap altogether, $Q = 1$; when there are two non-overlapping groups of BSs, $Q = 2$ (i.e., BSs are overlapped within each group but non-overlapped with the BSs of the other group); finally, when all Requesting BSs are not overlapped, $Q = M$. If we assume that all BSs within each group overlap with each other, the probability of having Q non-overlapping groups of BSs can be approximated by

$$P(Q = q|N, M) \simeq \begin{cases} P_{RB}(m_i = N - 1|N, M) & \text{if } q = 1, \\ \sum_{k=0}^{M-1} P_{RB}(m_i = k|N, M) \cdot P_{RB}(m_i = M - q - k|N, M)) & \text{if } q = 2, \\ \sum_{k=0}^{M-Q} P_{RB}(m = k|N, M) \cdot P(Q = q - 1|N - 1 - k, M - 1 - k) & \text{if } q > 2. \end{cases} \tag{A.12}$$

Bibliography

[1] Sina Khatibi, Luísa Caeiro, Lúcio S. Ferreira, Luis M. Correia, and Navid Nikaein. Modelling and implementation of virtual radio resources management for 5G Cloud RAN. *EURASIP Journal on Wireless Communications and Networking, 2017:128*, 07 2017. URL http://www.eurecom.fr/publication/5270.

[2] NGMN Alliance. 5G White Paper, Next Generation Mobile Networks. Technical report, NGMN, 2015.

[3] 3GPP. 3GPP TR 22.891 Draft, Feasibility Study on New Services and Markets Technology Enablers, Nov. 2015.

[4] 5GPPP. 5G Vision: The 5G Infrastructure Public Private Partnership: The next generation of communication networks and services, Feb. 2015.

[5] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder. 5g: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications*, 35(6): 1201–1221, June 2017. ISSN 0733-8716. doi: 10.1109/JSAC.2017.2692307.

[6] H. Li, K. Ota, and M. Dong. Eccn: Orchestration of edge-centric computing and content-centric networking in the 5g radio access network. *IEEE Wireless Communications*, 25(3): 88–93, JUNE 2018. ISSN 1536-1284. doi: 10.1109/MWC.2018.1700315.

[7] 3GPP. Tr 38.801 study on new radio access technology: Radio access architecture and interfaces. Technical Report 14.0.0, 3GPP, March 2017.

[8] M. Ali, S. Qaisar, M. Naeem, and S. Mumtaz. Joint user association and power allocation for licensed and unlicensed spectrum in 5g networks. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–6, Dec 2017.

[9] Haris Pervaiz, Muhammad Ali Imran, Shahid Mumtaz, Anwer-al Dulaimi, and Nikolaos Thomos. Editorial: Spectrum extensions for 5g and beyond 5g networks. *Transactions on Emerging Telecommunications Technologies*, 29(10):e3519, 2018. doi: 10.1002/ett.3519. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.3519.

[10] X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain. Network slicing for 5g: Challenges and opportunities. *IEEE Internet Computing*, pages 1–1, 2018. ISSN 1089-7801.

[11] K. Ait Ali, O. Baala, and A. Caminada. On the spatio-temporal traffic variation in vehicles mobility modeling. *Vehicular Technology, IEEE Transactions on*, PP(99):1–1, 2014. ISSN 0018-9545. doi: 10.1109/TVT.2014.2323182.

[12] C. Liang and F. R. Yu. Wireless virtualization for next generation mobile cellular networks. *IEEE Wireless Communications*, 22(1):61–69, February 2015. ISSN 1536-1284.

[13] 3GPP. Overview of 3gpp release 10 v0.1.8, March 2013. URL http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/.

[14] 3GPP. Overview of 3gpp release 11 v0.1.4, March 2013. URL http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/.

[15] 3GPP. Overview of 3gpp release 12 v0.0.8, March 2013. URL http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/.

[16] 3GPP. Ts 36.300 v11.4.0 release 11. Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 11), March 2013.

[17] 3gpp tr 36.872. Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects, December 2013.

[18] 3gpp tr 36.842. Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects, January 2014.

[19] X. Wang et al. Virtualized cloud radio access network for 5g transport. *IEEE Communications Magazine*, 55(9):202–209, 2017.

[20] NGMN. Suggestions on potential solutions to c-ran by ngmn alliance. Public Publication Version 4.0, NGMN Board, Januart 3rd 2013.

[21] J. Huang and Q. Wang. Further study on critical "c-ran" technologies. Technical Report v1.0, NGMN Alliance, UK, March 2015.

[22] J. Wu et al. Cloud radio access network (c-ran): a primer. *IEEE Network*, 29(1):35–41, Jan 2015. ISSN 0890-8044.

[23] S. Hamalainen, H. Sanneck, and C. Sartori. *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*. Wiley, 2011. ISBN 9781119963028. URL http://books.google.es/books?id=vM2gUROYpNIC.

[24] Huawei. Cloud ran introduction. Bundang, Korea, September 6th 2011.

[25] A. Garcia-Saavedra et al. Wizhaul: On the centralization degree of cloud ran next generation fronthaul. *IEEE Transactions on Mobile Computing*, pages 1–1, 2018. ISSN 1536-1233.

[26] Small cell virtualization functional splits and use cases. Technical Report Release 7.0, Document: 159.07.02, Small Cell Forum, 2016.

[27] Ravi Kokku, Rajesh Mahindra, Honghai Zhang, and Sampath Rangarajan. Nvs: A substrate for virtualizing wireless resources in cellular networks. *IEEE/ACM Trans. Netw.*, 20(5):1333–1346, 2012.

[28] Ravi Kokku, Rajesh Mahindra, Honghai Zhang, and Sampath Rangarajan. Cellslice: Cellular wireless resource slicing for active ran sharing. In *COMSNETS*, pages 1–10, 2013.

[29] Flavia fp7 project: Flexible architecture for virtualizable future wireless internet access. http://www.ict-flavia.eu.

[30] A. Maeder, V. Mancuso, Y. Weizman, E. Biton, P. Rost, X. Perez-Costa, and O. Gurewitz. Flavia: Towards a generic mac for 4g mobile cellular networks. In *Future Network Mobile Summit (FutureNetw), 2011*, pages 1–9, 2011.

[31] 3gpp tr 22.951 Service aspects and requirements for network sharing, October 2014.

[32] 3gpp ts 23.251. Network Sharing; Architecture and functional description (Release 11), 2013.

[33] 3GPP. Ts 23.251 version 11.4.0 release 11, January 2013. URL http://www.etsi.org/deliver/etsi_ts/123200_123299/123251/11.04.00_60/ts_123251v110400p.pdf.

[34] C. Liang and F.R. Yu. Wireless network virtualization: A survey, some research issues and challenges. *Communications Surveys Tutorials, IEEE*, PP(99):1–1, 2014. ISSN 1553-877X. doi: 10.1109/COMST.2014.2352118.

[35] 4WARD. The FP7 4WARD Project. URL http://www.4ward-project.eu/.

[36] J.S. Panchal, R.D. Yates, and M.M. Buddhikot. Mobile Network Resource Sharing Options: Performance Comparisons. *Wireless Communications, IEEE Transactions on*, 12(9):4470–4482, September 2013. ISSN 1536-1276. doi: 10.1109/TWC.2013.071913.121597.

[37] NEC. RAN sharing, NEC's approach towards Active Radio Access Network Sharing. White Paper, 2013.

[38] X. Costa-Perez, J. Swetina, Tao Guo, R. Mahindra, and S. Rangarajan. Radio Access Network Virtualization for Future Mobile Carrier Networks. *Communications Magazine, IEEE*, 51(7):–, 2013.

[39] Tao Guo and Rob Arnott. Active LTE RAN Sharing with Partial Resource Reservation. In *Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th*, pages 1–5, Sept 2013.

[40] I. Vila, O. Sallent, A. Umbert, and J. Perez-Romero. An analytical model for multi-tenant radio access networks supporting guaranteed bit rate services. *IEEE Access*, 7: 57651–57662, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2913323.

[41] 3GPP TR 22.852. Study on RAN Sharing Enhancements, 2013.

[42] L.E. Li, Z.M. Mao, and J. Rexford. Towards Software-Defined Cellular Networks. In *European Workshop on Software Defined Networking (EWSDN)*, pages 7–12, 2012.

[43] Xin Jin et al. CellSDN: Software-Defined Cellular Core Networks. April 15th - 17th 2013.

[44] Dmitry Drutskoy, Eric Keller, and Jennifer Rexford. Scalable network virtualization in software-defined networks. *Internet Computing, IEEE*, 17(2):20–27, March 2013. ISSN 1089-7801. doi: 10.1109/MIC.2012.144.

[45] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck. Network slicing and softwarization: A survey on principles, enabling technologies and solutions. *IEEE Communications Surveys Tutorials*, pages 1–1, 2018.

[46] T. Truong-Huu, P. Murali Mohan, and M. Gurusamy. Service chain embedding for diversified 5g slices with virtual network function sharing. *IEEE Communications Letters*, 23 (5):826–829, May 2019. ISSN 1089-7798. doi: 10.1109/LCOMM.2019.2900888.

[47] Q. Xu, J. Wang, and K. Wu. Learning-based dynamic resource provisioning for network slicing with ensured end-to-end performance bound. *IEEE Transactions on Network Science and Engineering*, pages 1–1, 2018. ISSN 2327-4697.

[48] Backhaul technologies for small cells: Use cases, requirements and solutions. Technical Report Release 7.0, Document: 049.07.02, Small Cell Forum, 2014.

[49] N.M.M.K. Chowdhury and R. Boutaba. Network virtualization: state of the art and research challenges. *Communications Magazine, IEEE*, 47(7):20–26, 2009.

[50] N. M. Mosharaf Kabir Chowdhury and Raouf Boutaba. A survey of network virtualization. *Computer Networks*, 54(5):862–876, 2010.

[51] X. Li, R. Casellas, G. Landi, A. de la Oliva, X. Costa-Perez, A. Garcia-Saavedra, T. Deiss, L. Cominardi, and R. Vilalta. 5g-crosshaul network slicing: Enabling multi-tenancy in mobile transport networks. *IEEE Communications Magazine*, 55(8):128–137, AUGUST 2017. ISSN 0163-6804. doi: 10.1109/MCOM.2017.1600921.

[52] D. Li, L. Gao, X. Sun, F. Hou, and S. Gong. A cellular backhaul virtualization market design for green small-cell networks. *IEEE Transactions on Green Communications and Networking*, 3(2):468–482, June 2019. ISSN 2473-2400. doi: 10.1109/TGCN.2019.2904975.

[53] 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. ISSN 0163-6804. doi: 10.1109/MCOM.2017.1600920.

[54] Chung-Sheng Li and Wanjiun Liao. Software defined networks [guest editorial]. *IEEE Communications Magazine*, 51(2):113, 2013.

[55] ONF. Open networking foundation. Home Page https://www.opennetworking.org/, 2013.

[56] Masum Z. Hasan. Tutorial on programmable cloud computing and networking. Ieee globecom, Anaheim, California, December 2012 2012.

[57] Kok-Kiong Yap, Masayoshi Kobayashi, Rob Sherwood, Te-Yuan Huang, Michael Chan, Nikhil Handigol, and Nick McKeown. Openroads: empowering research in mobile networks. *SIGCOMM Comput. Commun. Rev.*, 40(1):125–126, 2010.

[58] NFV. Network Functions Virtualization: An Introduction, Benefits, Enablers, Challenges and Call for Action. Network Functions Virtualization - Introductory White Paper, October 22-24 2012.

[59] MICHAEL LEONARD. Network Functions Virtualization is Changing How Services are Delivered, December 12th 2012. URL http://forums.juniper.net/t5/Data-Center-Directions/Network-Functions-Virtualization-is-Changing-How-Services-are/ba-p/171784.

[60] 5G Radio Access: Research and Vision. White paper, Ericsson AB, 2013.

[61] Jorge Carapinha and Javier Jiménez. Network virtualization a view from the bottom. In *Proceedings of the 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures*, NY, USA, 2009. ISBN 978-1-60558-595-6.

[62] 3gpp ts 36.300. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) (Release 12), September 2014.

[63] Chengchao Liang and F.R. Yu. Wireless virtualization for next generation mobile cellular networks. *Wireless Communications, IEEE*, 22(1):61–69, Feb. 2015. ISSN 1536-1284.

[64] B. Soret and K.I. Pedersen. Macro cell muting coordination for non-uniform topologies in lte-a hetnets. In *Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th*, pages 1–5, Sept 2013. doi: 10.1109/VTCFall.2013.6692277.

[65] B.A. Bjerke. Lte-advanced and the evolution of lte deployments. *Wireless Communications, IEEE*, 18(5):4–5, October 2011. ISSN 1536-1284. doi: 10.1109/MWC.2011.6056684.

[66] 3gpp tr 36.913. Requirements for further advancements for E-UTRA, September 2012.

[67] N.A. Ali, A.E.M. Taha, and H.S. Hassanein. *LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks*. ITPro collection. Wiley, 2011. ISBN 9781119970453. URL http://books.google.es/books?id=WN-h7OskezIC.

[68] 3gpp ts 36.412. E-UTRAN S1 signalling transport, September 2014.

[69] 3gpp tr 25.813. E-UTRA and E-UTRAN Radio interface protocol aspects, October 2006.

[70] 3gpp ts 36.331, Radio Resource Control (RRC); Protocol specification, September 2014.

[71] Zhuo Li, Song Guo, Deze Zeng, A. Barnawi, and I. Stojmenovic. Joint resource allocation for max-min throughput in multicell networks. *Vehicular Technology, IEEE Transactions on*, 63(9):4546–4559, Nov 2014. ISSN 0018-9545.

[72] 3gpp ts 36.413. E-UTRAN S1 Application Protocol (S1AP), Semtember 2013.

[73] 3gpp ts 36.423. E-UTRAN X2 Application Protocol (X2AP), March 2014.

[74] 3gpp ts 36.213. E-UTRA Physical layer procedures, March 2014.

[75] Yi Zhong and Wenyi Zhang. Multi-channel hybrid access femtocells: A stochastic geometric analysis. *Communications, IEEE Transactions on*, 61(7):3016–3026, July 2013. ISSN 0090-6778.

[76] URL http://www.smallcellforum.org/about/about-small-cells/small-cell-definition/.

[77] K.K Larsen. Network Sharing Fundamentals. Jul. 2012.

[78] GSMA. Network Infrastracture Sharing. Sep. 2012.

[79] 3GPP TR 22.852, Study on RAN Sharing enhancements, rel.12, Sept. 2014.

[80] 3GPP TS 32.130 Telecommunication management; Network Sharing; Concepts and requirements, rel.12, Dec. 2014.

[81] 3GPP TS 23.251, Network Sharing; Architecture and Functional Description, Mar. 2015.

[82] Y. Zaki et al. LTE Wireless Virtualization and Spectrum Management. In *IFIP WMNC, Budapest*, Oct. 2010.

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

[84] Wang Jiewu et al. User traffic collection and prediction in cellular networks: Architecture, platform and case study. In *IEEE IC-NIDC, Beijing*, Sep. 2014.

[85] Yantai Shu et al. Wireless traffic modeling and prediction using seasonal arima models. In *IEEE ICC, Anchorage*, volume 3, May 2003.

[86] D. Tikunov and T. Nishimura. Traffic prediction for mobile network using Holt-Winter's exponential smoothing. In *15th SoftCOM, Dubrovnik*, Sep. 2007.

[87] A. Yadav et al. A constant gain kalman filter approach to target tracking in wireless sensor networks. In *IEEE ICIIS, Chennai*, Aug. 2012.

[88] Rongpeng Li et al. Energy savings scheme in radio access networks via compressive sensing-based traffic load prediction. *Transactions on Emerging Telecommunications Technologies*, 25(4), Apr. 2014.

[89] J. Hwang et al. Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing,*, 42(10):2795–2810, Oct 1994.

[90] ITU-R. Guidelines for evaluation of radio interface technologies for IMT-Advanced. Report itu-r m.2135-1, Dec. 2009.

[91] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. Slaw: A new mobility model for human walks. In *IEEE INFOCOM 2009*, pages 855–863, April 2009. doi: 10.1109/INFCOM.2009.5061995.

[92] 3GPP TR 36.814 Further advancements for E-UTRA physical layer aspects, Mar. 2010.

[93] K. Samdanis, X. Costa-Perez, and V. Sciancalepore. From network sharing to multi-tenancy: The 5g network slice broker. *IEEE Communications Magazine*, 54(7):32–39, July 2016. ISSN 0163-6804. doi: 10.1109/MCOM.2016.7514161.

[94] T. Guo and R. Arnott. Active lte ran sharing with partial resource reservation. In *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, pages 1–5, Sept 2013.

[95] J. Pérez-Romero et al. Admission control for multi-tenant radio access networks. In *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1073–1078, May 2017.

[96] H. Zhang et al. User association scheme in cloud-ran based small cell network with wireless virtualization. In *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 384–389, April 2015.

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

[98] S. Parsaeefard, R. Dawadi, M. Derakhshani, and T. Le-Ngoc. Joint user-association and resource-allocation in virtualized wireless networks. *IEEE Access*, 4:2738–2750, 2016. ISSN 2169-3536.

[99] Y. L. Lee, J. Loo, T. C. Chuah, and L. Wang. Dynamic network slicing for multitenant heterogeneous cloud radio access networks. *IEEE Transactions on Wireless Communications*, 17(4):2146–2161, April 2018. ISSN 1536-1276.

[100] R. Addad, M. Bagaa, T. Taleb, D. L. Cadette Dutra, and H. Flinck. Optimization model for cross-domain network slices in 5g networks. *IEEE Transactions on Mobile Computing*, pages 1–1, 2019. ISSN 1536-1233.

[101] T. Biermann et al. How backhaul networks influence the feasibility of coordinated multipoint in cellular networks [accepted from open call]. *IEEE Communications Magazine*, 51 (8):168–176, August 2013. ISSN 0163-6804. doi: 10.1109/MCOM.2013.6576356.

[102] S. Bassoy et al. Coordinated multi-point clustering schemes: A survey. *IEEE Communications Surveys Tutorials*, 19(2):743–764, Secondquarter 2017. ISSN 1553-877X. doi: 10.1109/COMST.2017.2662212.

[103] P. Marsch and G. P. Fettweis. *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge University Press, UK, 2011.

[104] M. Li et al. Multicell coordinated scheduling with multiuser zero-forcing beamforming. *IEEE Transactions on Wireless Communications*, 15(2):827–842, Feb 2016. ISSN 1536-1276. doi: 10.1109/TWC.2015.2479226.

[105] J. Shin and J. Moon. Regularized zero-forcing interference alignment for the two-cell mimo interfering broadcast channel. *IEEE Communications Letters*, 17(7):1336–1339, July 2013. ISSN 1089-7798. doi: 10.1109/LCOMM.2013.060513.122713.

[106] W. Shin et al. Coordinated beamforming for multi-cell mimo-noma. *IEEE Communications Letters*, 21(1):84–87, Jan 2017. ISSN 1089-7798. doi: 10.1109/LCOMM.2016.2615097.

[107] A. Goldsmith et al. Capacity limits of mimo channels. *IEEE Journal on Selected Areas in Communications*, 21(5):684–702, June 2003. ISSN 0733-8716. doi: 10.1109/JSAC.2003.810294.

[108] 3GPP. Tr 38.804, study on new radio access technology radio interface protocol aspects, 3gpp, tech. report. Technical Report 1.0.0, 3GPP, 2017.

[109] 3GPP. Tr 36.942, evolved universal terrestrial radio access (e-utra); radio frequency (rf) system scenarios. Technical Report 13.0.0, 3GPP, 2016.

[110] J. Wu et al. Cloud radio access network (c-ran): a primer. *IEEE Network*, 29(1):35–41, Jan 2015. ISSN 0890-8044. doi: 10.1109/MNET.2015.7018201.

[111] A. Checko et al. Cloud ran for mobile networks - a technology overview. *IEEE Communications Surveys Tutorials*, 17(1):405–426, Firstquarter 2015. ISSN 1553-877X. doi: 10.1109/COMST.2014.2355255.

[112] C. Fan et al. Advances and challenges toward a scalable cloud radio access network. *IEEE Communications Magazine*, 54(6):29–35, June 2016. ISSN 0163-6804. doi: 10.1109/MCOM.2016.7497763.

[113] N. Nikaein. Processing radio access network functions in the cloud: Critical issues and modeling. In *Proceedings of the 6th International Workshop on Mobile Cloud Computing and Services*, pages 36–43, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3545-4.

[114] Common public radio interface (cpri); interface specification, 2015.

[115] A. de la Oliva et al. An overview of the cpri specification and its application to c-ran-based lte scenarios. *IEEE Communications Magazine*, 54(2):152–159, February 2016. ISSN 0163-6804. doi: 10.1109/MCOM.2016.7402275.

[116] C. Y. Yeoh et al. Performance study of lte experimental testbed using openairinterface. In *2016 18th International Conference on Advanced Communication Technology (ICACT)*, pages 617–622, Jan 2016.

[117] J. Robson. Small cell backhaul requirements. Technical Report v1.0, NGMN Alliance, UK, June 2012.

[118] E. Metsala and J. Salmelin. *Mobile Backhaul*. John Wiley & Sons, UK, 2012.

[119] E.Metsala and J.Salmelin. *LTE Backhaul; Planning and Optimizing Mobile Backhaul for LTE*. John Wiley & Sons, UK, 2016.

[120] J. Robson. Backhaul provisioning for lte-advanced & small cells. Technical Report v0.0.14, NGMN Alliance, UK, 2014.

[121] J. L. Jain et al. *A Course on Queueing Models*. Chapman & Hall/CRC, USA, 2007.

[122] Raymond Hemmecke, Matthias Koeppe, Jon Lee, and Robert Weismantel. 50 years of integer programming 1958-2008. *50 Years of Integer Programming 1958-2008*, 06 2009. doi: 10.1007/978-3-540-68279-0_15.

[123] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. ISBN 0262033844, 9780262033848.

[124] Salma Rattal and Mohammed Moughit. Performance analysis of hybrid codecs g . 711 and g . 729 over signaling protocols h . 323 and sip. 2013.

[125] O. Awoniyi and F. A. Tobagi. Packet error rate in ofdm-based wireless lans operating in frequency selective channels. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, pages 1–13, April 2006.

[126] Petar Popovski et al. 5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view. *CoRR*, abs/1804.05057, 2018.

[127] Report itu-r m.2412-0, guidelines for evaluation of radio interface technologies for imt-2020. Technical Report M.2412-0, ITU-R, 2017.

[128] P. Caballero et al. Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads. *IEEE/ACM Transactions on Networking*, 25(5):3044–3058, Oct 2017. ISSN 1063-6692.