UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

On-demand RAN Slicing Techniques for SLA Assurance in Virtual Wireless Networks

Behnam Rouzbehani

Supervisor: Doctor Luís Manuel de Jesus Sousa Correia

Co-Supervisor: Doctor Maria Luísa Pedro Brito da Torre Caeiro

Thesis approved in public session to obtain the PhD Degree in

Electrical and Computer Engineering

Jury final classification: **Pass with Distinction**

2019

TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

On-demand RAN Slicing Techniques for SLA Assurance in Virtual Wireless Networks

Behnam Rouzbehani

Supervisor: Doctor Luís Manuel de Jesus Sousa Correia

Co-Supervisor: Doctor Maria Luísa Pedro Brito da Torre Caeiro

Thesis approved in public session to obtain the PhD Degree in

Electrical and Computer Engineering

Jury final classification: **Pass with Distinction**

Jury

Chairperson: Doctor Isabel Maria Martins Trancoso, Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee:

Doctor Paulo da Costa Luís da Fonseca Pinto, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Doctor Luís Manuel de Jesus Sousa Correia, Instituto Superior Técnico, Universidade de Lisboa

Doctor Rui Manuel Rodrigues Rocha, Instituto Superior Técnico, Universidade de Lisboa

Doctor Hamed Ahmadi, School of Computer Science and Electronic Engineering, University of Essex, UK

Doctor António Manuel Raminhos Cordeiro Grilo, Instituto Superior Técnico, Universidade de Lisboa

2019

To my beloved family for their unfailing faith and support.

Acknowledgements

First and above all, I would like to express my sincere gratitude towards my supervisor Prof. Luis M. Correia for suggesting the problem, his valuable guidance and constant encouragement throughout the development of my work, which were beyond the call of duty.

My highest gratitude also goes to my co-supervisor Prof. Luísa Caeiro for her continuous support, availability and constructive suggestions, which were determinant for the accomplishment of this work.

Several people have contributed to this work by inspiring discussions and fruitful collaborations. I want to mention namely our Group for Research on Wireless (GROW), in particular Kenan Turbic, Mojgan Barahman, Ema Catarré and Vera de Almeida, as well as my former colleague Dr. Sina Khatibi.

Finally, special thanks to my family and friends, especially Shaghayegh Monfared, who has spared my company on many occasions when I was committed to this work. Above all, my heartfelt thanks are to my parents and my brother for their love, encouragement, moral support, patience and their faith in me.

Abstract

This thesis addresses a service-based approach of Radio Access Network (RAN) slicing for on-demand Radio Resource Management (RRM) in virtual wireless networks, from a high-level perspective. In this regard, the functionalities related to service orchestration, such as satisfying contracted Service Level Agreements (SLAs) and customising radio bearers, are managed by a centralised virtualisation platform, called Virtual-RRM (VRRM), inside the Service Data Adaptation Protocol sublayer, which has been recently introduced in 5G New Radio, while the mapping of the service demands to be addressed by the underlying physical Radio Access Technologies (RATs) is performed in another individual management entity, called Common-RRM (CRRM), which is controlled by Infrastructure Providers (InPs). The proposed model realises a separation in the role of Virtual Network Operators (VNOs) and InPs, since VNOs do not own or control any physical infrastructure, and InPs are not aware of VNOs' individual policies for service management. The efficient interaction between VRRM and CRRM not only reduces the complexity of resource management, but it also maximises the utilisation of the aggregated capacity, obtained from Radio Resource Units, up to 100% in case of existing demands. Results for 3 types of SLAs and the 4 types of service classes show a level of performance isolation among VNOs according to their customised service requirements, and that there is capacity share proportional to the serving weights in order to satisfy the proportional fairness definition. Furthermore, the algorithms of RAT selection and load balancing are capable of efficiently distribute demand among the available RATs, in order to satisfy InPs' policies separately.

Keywords – Radio Access Network slicing, Radio Resource Management, Service Level Agreement, Virtualisation, Performance isolation, Proportional fairness.

Resumo

Esta tese aborda uma perspetiva de alto nível orientada ao serviço das Redes de Acesso Rádio (RAN), para a segmentação a pedido da Gestão de Recursos Rádio (RRM) em redes sem fios virtuais. As funcionalidades relacionadas com a orquestração de serviços, como a satisfação dos Acordos de Níveis de Serviço (SLAs) e a customização de portadoras rádio, são geridas centralmente por uma plataforma de virtualização, designada por Gestão de Recursos Rádio Virtual (VRRM), dentro da camada de Protocolos de Adaptação do Serviço de Dados, que foi recentemente introduzida na Novo Rádio 5G, enquanto o mapeamento dos pedidos de serviços para serem processados pela subjacente Tecnologias de Acesso Rádio (RAT) físicas por outra entidade de gestão virtual, a Gestão de Recurso Rádio Comum (CRRM), que é controlada pelos Fornecedores de Infraestrutura (InPs). O modelo proposto efetua a separação dos Operadores de Rede Virtuais (VNOs) e os Fornecedores de Infraestrutura, uma vez que os primeiros não possuem nem controlam infraestrutura física, e os segundos não conhecem as políticas individuais de gestão de serviço dos primeiros. A Integração eficiente entre VRRM e CRRM não só reduz a complexidade da gestão de recursos, mas também maximiza o uso da capacidade agregada, obtida das Unidades de Recursos Rádio, até 100% em caso de pedidos existentes. Os resultados para 3 tipos de SLAs e 4 tipos de classes de serviço mostram um nível de desempenho no isolamento entre VNOs de acordo com os requisitos de serviço customizados, e que existe uma capacidade repartida proporcional aos pesos dos serviços, de modo a satisfazer a definição de justiça proporcional. Além disso, os algoritmos para seleção de RAT e equilíbrio de carga são capazes de distribuir eficientemente os pedidos entre as RATs disponíveis, satisfazendo separadamente as políticas dos InPS.

Palavras-chave: Segmentação de Redes de Acesso Rádio, Gestão de Recursos Rádio, Acordo de Nível de Serviço, Virtualização, Independência de Desempenho, Justiça Proporcional.

چکیده

هدف این پایان نامه ارائه و بررسی یک مدل مبتنی بر سرویس، برای تقسیم و مدیریت منابع شبکه دسترسی رادیوئی در شبکه های بی سیم مجازی از منظر لایه های بالایی است. در این راستا، ویژگی های مربوط به تنظیم و مدیریت سرویس، مانند سطح رضایت قراردادها، سطح خدمات و سفارشی سازی ارائه دهنده های رادیویی توسط یک پلتفرم مجازی سازی متمرکز که درون زیر بستر پروتکل سرویس داده قرار دارد و اخیرا در نسل پنجم شبکه های سلولار تعریف شده، مدیریت و کنترل می شود. این در حالی است که مدیریت زیرساخت فیزیکی شبکه برای تطبیق و ارسال ترافیک درخواست سرویس در لایه های بالایی به عهده یک کنترلر مرکزی دیگریست که مستقلا توسط صاحبان و ارائه دهندگان زیرساخت فیزیکی مدیریت می شود. مدل ارائه شده، ایده جداسازی اپراتورهای مجازی و ارائه دهنده های زیرساخت های فیزیکی را محقق می کند زیرا هیچ یک از این دو از سیاست های دیگری در زمینه کنترل و مدیریت سرویس و منابع اطلاع ندارد. تعامل کارآمد بین این دو کنترلر مرکزی، نه تنها پیچیدگی رویکرد تخصیص منابع رادیوئی را کاهش می دهد، بلکه همچنین استفاده از ظرفیت جمع آوری شده از واحدهای رادیو را در صورت تقاضای موجود به حداکثر می رساند. نتایج بررسی برروی 3 نوع قرارداد سرویس و 4 نوع کلاس سرویس نشان از سطح عملکرد ایزوله اپراتورها با توجه به الزامات خدمات سفارشی خود داشته و زمانی که ممکن باشد، به منظور محقق شدن چارچوب اعتدال تناسبی، ظرفیت ها با وزن های ارائه شده سرویس ها متناسب است. علاوه بر این، الگوریتم های انتخاب شبکه دستری رادیوئی و متعادل سازی بار می تواند به طور موثر توزیع خواسته ها در میان شبکه های در دسترس، به منظور تضمین سیاست های ارائه دهندگان زیرساخت فیزیکی، به طور جداگانه پیاده سازی کند.

کلمات کلیدی – تقسیم شبکه رادیو دسترسی، مدیریت منابع رادیویی، توافقنامه سطح خدمات، مجازی سازی، استقلال عملکرد، اعتدال تناسبی.

Table of Contents

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
AIV	Air Interface Variants
AP	Access Point
BBU	Base Band Unit
BG	Best Effort with Minimum Guaranteed
BE	Best Effort
BS	Base Station
BSC	Base Station Controller
BTS	Base Transceiver Station
CA	Carrier Aggregation (CA)
CDMA	Code Division Multiple Access
CN	Core Network
CoMP	Coordinated Multi-Point
CNF	Core Network Functions
C-RAN	Cloud-Radio Access Network
CRRM	Common-RRM
CS	Circuit Switch
DL	Downlink
EDGE	Enhanced Data rates for GSM Evolution
EGPRS	Enhanced General Packet Radio Service
eMMB	Enhanced Mobile Broadband
eNodeB	Evolved NodeB
FDD	Frequency Division Duplexing
FDMA	Frequency Division Multiple Access
GBR	Guaranteed Bit Rate
GGSN	Gateway GPRS Support Node
GMSK	Gaussian Minimum Shift Keying
GPRS	General Packet Radio Service
GSM	Global System for Mobile
Het-Net	Heterogeneous Networks
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSUPA	High Speed Uplink Packet Access
IaaS	Infrastructure as a Service
InP	Infrastructure Provider

IP	Internet Protocol
IRACON	Inclusive Radio Communication Networks for 5G and beyond
ISP	Internet Service Provider
KPI	Key Performance Indicator
LAN	Local Area Network
LRRM	Local RRM
LTE	Long Term Evolution
LTE-A	LTE-Advanced
MAC	Medium Access Control
MCS	Modulation and Coding Scheme
MME	Mobility Management Entity
mMTC	Massive Machine-Type Communications
MNO	Mobile Network Operators
MIMO	Multiple-Input Multiple-Output
MSC	Mobile Switching Centre
MT	Mobile Terminal
MTC	Machine Type Communication
MU-MIMO	Multi-User MIMO
MS	Mobile Station
MVNE	Mobile Virtual Network Enabler
MVNO	Mobile Virtual Network Operator
MVNP	Mobile Virtual Network Provider
NaaS	Network as a Service
NB-IoT	Narrow-Band Internet of Things
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PDN-GW	Packet Data Network Gateway
PHY	Physical
PRACH	Physical Random Access Channel
PRB	Physical Resource Block
PS	Packet Switch
QoE	Quality of Experience
QoS	Quality of Service
QCI	QoS Class Indicator
QPSK	Quadrature Phase Shift Keying
RAN	Radio Access Network
RAT	Radio Access Technology
RRU	Radio Resource Units
RRM	Radio Resource Management

RN	Relay Node
RNC	Radio Network Controller
RNF	Radio Network Function
RRH	Remote Radio Head
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDAP	Service Data Adaptation Protocol
SDN	Software Defined Networking
SDWN	Software Defined Wireless Networking
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SLA	Service Level Agreement
SP	Service Provider
TD	Technical Document
TDD	Time Division Duplexing
TDMA	Time Division Multiple Access
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
URLLC	Ultra-Reliable and Low Latency Communications
UTRAN	UMTS Terrestrial Radio Access Network
VM	Virtual Machine
VNO	Virtual Network Operator
VNP	Virtual Network Provider
VRRM	Virtual-RRM
WCDMA	Wideband Code Division Multiple Access
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Networks

List of Symbols

a_m	Coefficients of the polynomial fit
α_p	the path loss exponent
γ	SINR value
γ_v	Weight of VNO v assigned by InP
γ_{j,v_s}^{RAT}	Assigned weight to RATj, for performing service s, provided by VNO v
δ_j^{RAT}	Load balancing factor of RAT j
δ_s	Serving weight, assigned to service s
λ_{v_s}	Tuning weight associated with service s, provided by VNO v
μ	Lagrange multiplier
$\mu_{v_s}^{RAT}$	Assigned weight to service v_s, to define service-to-RAT allocation priorities
f_{CRRM}	The objective function of CRRM
f_{VRRM}	The objective function of VRRM
$I_{pf_{v_s}}$	Proportional fairness index of service s, from VNO v
N^{RAT}	Number of RATs
N^{srv}	Number of services
N_v^{srv}	Number of services provided by VNO v
$N_{v_s}^{usr}$	Number of users performing service s, from VNO v
$N_{v_s}^{usr_{cell}}$	Number of users with access only to cellular RATs, performing service v_s
$N^{usr_{net}}$	Total number of offered users in the network
$N_{v_s}^{usr_{tot}}$	The total number of offered users in service s of VNO v
p	PDF function
p_{VRRM}^{tot}	Percentage of total assigned data rate
$p_{v_s}^{usr_{net}}$	Percentage of served users from service s of VNO v
$p_{v_s}^{usr_{srv}}$	Percentage of served users in each service s of VNO v
P	CDF function
$\boldsymbol{P}_{v_s}^{usr_{net}}$	Vector of cumulative percentage of served users from service s of VNO v
R_{j,v_s}^{RAT}	Allocated data rate from RATj to service s, from VNO v
$R_j^{RAT_{tot}}$	Total available capacity of RAT j
R_j^{RRU}	Approximated data rate of a single RRU
$R_j^{RRU_H}$	The higher bound for the RRU's data rate
$R_j^{RRU_L}$	The lower bound for the RRU's data rate
$R_j^{RRU_{max}}$	The maximum achievable data rate for a single RRU from each RAT

$R_{v_s}^{srv}$	Total served data rate of service v_s
$R_{v_s}^{srv_{max}}$	Maximum assignable data rate to the user of service s, from VNO v
$R_{v_s}^{srv_{min}}$	Minimum assignable data rate to the user of service s, from VNO v
$R_{v_s,i}^{usr}$	Data rate of user i from service s, from VNO v
$R_{v}^{vno_{max}}$	Maximum assignable data rate to the user of service s, from VNO v
$R_{v}^{vno_{min}}$	Minimum contracted data rates of VNO v
$R_{v_s}^{Srv_{pf}}$	Proportional fair value of the data rate for service s, from VNO v
$R_{v_s}^{srv_{tot}}$	Total data rate of service s, from VNO v
$R_{VRRM}^{VNO_v}$	VRRM capacity share
$\mathbf{R^{RAT}}$	Vector of assigned data rates from different RATs
R_j^{RAT}	Load share of RAT j
$\mathbf{R^{srv}}$	Vector of serving data rates
$\mathbf{R}_{\boldsymbol{v_s}}^{\mathbf{srv_{RAT}}}$	Vector of Assigned data rate to service s of VNO v, from different RATs
R^{CRRM}	Total capacity obtained by aggregation of RRUs from different RATs
R^{VRRM}	Total offered capacity from CRRM to VRRM
$R^{VRRM_{cell}}$	Available capacity of VRRM, provided by cellular RATs
$S_{v_s,i}^{usr}$	Users' satisfaction ratio of user i, performing service s, from VNO v
$S_{v_s,i}^{usr_{Int}}$	Interactive users' satisfaction ratio of user i, performing service s, from VNO v
$w_{v_s,i}^{usr}$	Assigned weight to user i, performing service s, from VNO v
$\mathbf{w^{usr}}$	Vector of users' weights

List of Software

CVX	CVX turns MATLAB into a modelling language and used for solving convex optimisation problems.
MATLAB	MATLAB is used for optimisation and plotting figures.
MS Visio 2013	Visio is used to edit several figures presented in this thesis.
MS Word 2013	Word is used to edit this thesis and all associated document such as publications.

Chapter 1

Introduction

This chapter provides the thesis overview. Section 1.1 presents a brief history of the mobile cellular networks evolution in the last two decades. Section 1.2 clarifies the key aspects of the thesis motivation and objectives. Section 1.3 highlights the novelty of the work and the concepts explored in the thesis. Section 1.4 provides an overview of the pursuit research strategy, where projects contributions and published work are highlighted. Finally, Section 1.5 defines the dissertation contents.

1.1 Brief History

Over the past few decades, wireless and mobile communication technologies have experienced enormous developments to accommodate the ever-increasing demands of mobile users for service connectivity. Figure 1.1 presents a short chronological history of cellular radio systems from their first generation, 1G, in the 1980s, until the 5th generation, 5G in 2020s. The major steps are shown in the figure and are explained thereafter.

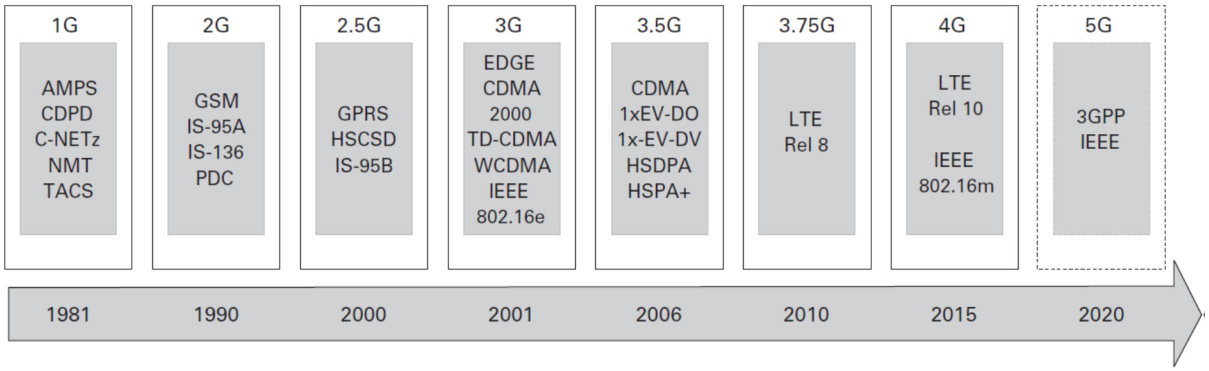


Figure 1.1 – Evolution of cellular standards (extracted from [OMM16]).

The first commercial implementation of a 1G system took place in 1981. This generation is also known as *analogue systems*, since it was using analogue technology, typically frequency modulated radio signals with a digital signalling channel [OMM16].

The introduction of 2G was characterised by the implementation of digital transmission and switching technologies. Digital communications made considerable improvements in Voice quality and network capacity, and presented growth in the form of supplementary services and advanced applications, such as SMS for storage and forwarding of written information. The primary purpose of 2G's Global System for Mobile communications (GSM), was to create a digital Voice telephony network that allowed international roaming across Europe. GSM is based on a hybrid Time Division Multiple Access (TDMA)/Frequency Division Multiple Access (FDMA) method [Hill02]. The evolution of GSM introduced packet-switched data services in addition to Voice and circuit-switched data, designated as General Packet Radio Service (GPRS), and later it evolved further to Enhanced Data Rates for Global Evolution (EDGE) and its associated component Enhanced General Packet Radio Service (EGPRS), mainly by addition of higher order modulation and coding schemes.

Universal Mobile Telecommunication System (UMTS) was the major 3G mobile communications standard. New specifications were developed within the framework of the 3rd Generation Partnership Project (3GPP), which is known as 3.5G, Figure 1.1. For this evolution, two Radio Access Network (RAN) approaches and an evolution of the core network were taken. One of the RAN approaches was High Speed Packet Access (HSPA) to support enhanced packet data rate in both up- and downlinks (UL and DL), based on Wideband Code Division Multiple Access (WCDMA), and quickly evolved to handle higher data rates with the introduction of Multiple-Input Multiple-Output (MIMO) technology. 3GPP standards followed the philosophy of adding new features, while still maintaining backward

compatibility. This has been further applied in the evolution of HSPA known as HSPA+, which supports carrier aggregation for higher peak data rates without affecting existing terminals in the market [HoTo07].

The next evolution, commercially accepted as 4G, is called Long Term Evolution (LTE) [SBT11], and is composed of a new air interface based on Orthogonal Frequency Division Multiple Access (OFDMA) and a new architecture and core network called the Evolved Packet Core (EPC). LTE offers significant improvements in capacity, and was designed to transition cellular networks away from circuit-switched technology, which provided a major cost reduction from previous generations. The first LTE specifications were approved in 3GPP as LTE Release 8, and then several technical features were added, such as higher order MIMO and carrier aggregation, to improve capacity and throughput. The standardisation of LTE continued to Release 11 to Release 13, and is expected to proceed beyond. Release 11 refined some of Release 10 capabilities, by enhancing carrier aggregation, relaying and interference cancellation; new frequency bands were added, and the use of Coordinated MultiPoint (CoMP) transmission and reception was defined. LTE Release 12, which was concluded in March 2015, added several features to improve the support of heterogeneous networks, even higher order MIMO, and aggregation between Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD) carriers. Several features for the offloading of backhaul and core networks were also defined. Furthermore, in Releases 12 and 13, new solutions (like Narrow-Band Internet of Things (NB-IoT)) were introduced, in order to support massive Machine Type Communication (MTC) devices, such as sensors and actuators [OMM16].

The introduction of 5G will allow cellular and wireless networks to match data rates and use cases that are currently handled by fibre access. 5G planning, as shown in Figure 1.2, aims at a higher capacity than 4G, allowing a higher density of mobile broadband users and supporting device-to-device, ultra-reliable, and massive machine communications [NGMN15]. It also aims at lower latency than 4G equipment and lower battery consumption, for better implementation of IoT.

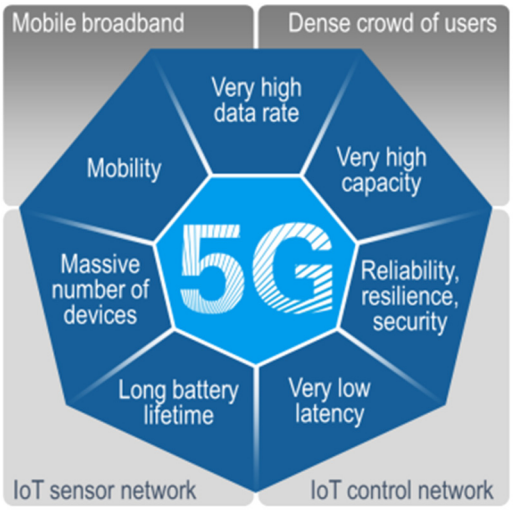


Figure 1.2 – 5G planning goals (extracted from [Rohd18]).

One major enabler technology, which is envisioned for 5G, is network virtualisation and software-based design, as a transformation that is not expected to be completed in the 5G timeframe. In this regard, two

fundamental projects, 4WARD [4WAR10] and SAIL [SAIL13] in Europe had major contributions in the adoption of network virtualisation and cloud as core strategies for future implementations of service-based architectures, which includes 5G. 4WARD focused on a generalised approach to allow virtualisation of different resources that form a unified framework, and supports both wireline and wireless resources, while a key objective of SAIL was to expand the concept of network virtualisation to accommodate cloud providers as Virtual Network Operators (VNOs). Beyond 5G, the biggest opportunity and challenge will be to finish an overall industry transformation to a software-centric vision, in which commercial network equipment is flexible and easily designed, implemented, deployed, upgraded, managed, maintained, and programmed using optimisation techniques, machine learning and artificial intelligence. These are very comprehensive and difficult tasks that will require another decade or so to be completed [OMM16].

In addition to cellular networks, the current wireless technologies include Wireless Local Area Networks (WLANs) and Wireless Metropolitan Area Networks (WMANs). These systems, operating under IEEE802.11 and IEEE802.16, are considered to be part of the wireless and mobile systems evolution, providing extra capacity and contributing to the optimisation of the diverse wireless resources utilisation.

1.2 Thesis Motivation and Objectives

The traditional mobile communication networks employ one-size-fits-all methods to provide services to mobile users, regardless of the communication requirements of vertical services. This design attitude cannot offer differentiated services, hence, it is necessary for the research community to explore new techniques to address the challenges associated with supporting vertical industries in 5G and beyond networks. In addition, network parameters, such as users' traffic load or channel condition changes in time and place and the growth of service demand over mobile networks, lead to traffic imbalance in different areas [NaWK12]. From the operators' viewpoint, the establishment of the current RAN, which is done based on the peak hour traffic, enforces additional resources and cost. Operators have not witnessed a comparable increase in their revenue, since they have to upgrade their infrastructure equipment to provide the explosive data rate demands, which will considerably increase their total cost of ownership, as well as complicate the maintenance of different co-existing mobile network generations.

The concept of RAN slicing, which is a fundamental feature of 5G, is foreseen as a promising approach to address these challenges for the future implementation of mobile communications, enabling an increased statistical multiplexing gain by a more efficient management of resources, together with a better performance isolation of network entities [NGMN18]. Slices are self-contained logical instances of the network, sharing the same physical infrastructure in an isolated manner, according to the concept of end-to-end network virtualisation [RMM17]. RAN slicing allows different network operators to define their own protocols and regulations over the common set of Radio Resource Units (RRUs), which accordingly promotes the notion of multi-tenancy and increases the possibility of providing a wide range of customised services.

Radio Resource Management (RRM) as one of the key functionalities of cellular networks, has a direct influence on the Quality of Service (QoS) of users, as well as on the performance of higher layers [XZS17]. With the introduction of new applications, which has led to the increased number of connected users, the problem of RRM has become particularly challenging, since the various applications have different and often conflicting needs [LiYu15]. Therefore, each service has to be managed independently, and the available radio resources must be allocated on a priority basis. This is the point when the context of resource slicing becomes interesting. Each resource slice is defined to address the specific requirements of a service with a certain degree of performance isolation. A network slice can scale up or down as service requirements and the number of users change to ensure that, regardless of the variation of network status, the desired performance level of independent slices is always met and the configurations among various slices do not affect each other [LAL17]; therefore, the reliability and security of each slice can be enhanced. Such a flexible mechanism of radio resource slicing and management can be achieved through the virtualisation of radio resources. 5G RAN slicing can be implemented through the logical abstraction of physical radio resources, such as spectrum, and physical hardware, such as a base station. This way, an RRM algorithm should not only maximise the performance of different slices, but also the usage of the overall pool of shared resources provided by different Radio Access Technologies (RATs) [Aija16].

Regarding the economic perspective, network slicing has the potential to increase the expected revenues of Mobile Network Operators (MNOs). In traditional network planning, radio resources are underutilised, the main reason being that according to given Key Performance Indicators (KPIs), large portions of the available resources are typically reserved for some use cases with a small demand, since MNOs are only able to provide undefined resource bundles for the general utilisation of all applications [LAL17]. However, on the other hand, by deploying virtualisation techniques for RAN slicing, VNOs as network tenants can specify different bundles for several applications based on their own policies over the same set of radio resources. This way, RRM will be more flexible and efficient, and with the same network infrastructure, the revenue will increase. In this respect, the main difference between the services that can be provided by future mobile implementations, compared to the majority of current technologies, such as LTE, lies in the *granularity* of the level of service customisation [MBQB18]. While in LTE all packets in a bearer are treated the same way, service flows in Virtual RAN can be flexibly customised in a more granular way to sufficiently differentiate in between their requirements.

Although there are quite extensive studies to address different challenges in traditional wireless networks, when it comes to the evolutionary technologies for future wireless communications, the existing techniques have to be modified in order to accommodate the specifications of new services and architectures. However, speaking about service-oriented approaches of RAN slicing and resource allocation, to the best of the author’s knowledge, there is no significant effort to consider all the key parameters in a single model, such as differentiated service requirements, which needs to be defined in Service Level Agreement (SLAs) and VNOs policies of service management, performance isolation in a multi-tenant network, and provision of fairness metrics. Most of the public available works just focus on addressing a particular aspect, while neglecting the rest of complementary assumptions.

Consequently, according to the aforementioned potentials, the main objective of this work in general, was to develop a RAN slicing technique for RRM in virtual mobile networks, to address customised requirements of different services in a multi-tenant scenario, by sharing a common pool of aggregated radio resources provided by a physical infrastructure. More specifically, regarding the proposed objectives of Virtual RRM (VRRM), this entity has to centrally process all service demands, which are delivered from different VNOs, and then calculate the optimal amount of capacity to be shared among various service slices in order to fulfil some important parameters, including the internal policy of the associated VNO in terms of customised range of data rate, service priority, the contracted SLA between VNOs and Infrastructure Provider (InP), satisfying the framework of proportional fairness in between all the existing service demands when instantiating a service slice, as well as maximising the use of available aggregated capacity.

Additionally, a level of performance isolation among VNOs has to be established, to make sure that variation in different network parameters, such as traffic load, channel conditions and underlying physical infrastructure, will not have a considerable effect on satisfying the guaranteed SLA types. Concerning the key objectives of Common-RRM (CRRM), this entity is in charge of mapping the resources and capacity onto the underlying physical RATs to satisfy VRRM's demands. The mechanism of service-to-RAT association takes the RATs' load, their suitability for the corresponding services, as well as energy efficiency, into consideration for decision making.

1.3 Novelty

Since the concept of RAN slicing is rather new and still in the development phase, most of the available studies in this area focus on improving one or a few design parameters in the suggested models, while neglecting the assumptions correlated to the key fundamental aspects that can be potentially enabled through RAN virtualisation and slicing. Accordingly, considering the lacks in literature, the novelty and contribution of this work have two folds. On the one hand, the aim of designing the proposed RAN slicing technique is to include more specific parameters and assumptions as a single model, and on another hand, it is to extend and improve some of the existing models with research directions similar to this thesis work. Therefore, the main contribution can be seen as listed in what follows.

The model is capable of realising the idea of *separation* between the roles of InPs and VNOs in a service-oriented manner, which deals with the high-level management of Virtual RAN. The functionalities related to service management, such as *SLA enforcement*, are *decoupled* from the underlying RATs to be managed independently inside the Service Data Adaptation Protocol (SDAP) sublayer, according to the recent 3GPP standard for 5G New Radio [3GPP18]. The conceptual framework is in agreement with the recommended 3GPP model for *wholesale-only* network sharing [CSGM13], in which InPs do not provide service to end-users (in contrast with existing Heterogeneous Networks (Het-Nets)), rather selling capacity to tenants, i.e., VNOs that do not own the infrastructure.

The design of the VRRM model takes various key parameters of RAN slicing into consideration. It maintains a level of performance *isolation* among VNOs in a *multi-tenant* scenario with a range of differentiated service requirements, while satisfying all *SLAs* and internal policies of each VNO in service orchestration *on-demand*, and regardless of the variation of different network parameters, such as traffic load and channel conditions. This entity is also capable of *maximising* the utilisation of available capacity, which is obtained by aggregation of the physical RRUs from different RATs. Furthermore, the network-wide virtualisation platform of VRRM not only provides dynamic resource sharing across service slices, but also imposes minimum changes to traditional Base Stations (BSs), while expressing a higher statistical *multiplexing gain* compared to traditional RRM schemes in Het-Nets, with the same amount of radio resources.

The proposed models of VRRM in previous work [Khat16] is extended and modified in order to include the management of infrastructure as well. This way, the task of mapping the demanded capacity from different virtual slices onto the underlying physical RATs is defined in CRRM, in order to promote the notion of *end-to-end slicing*, since it was not covered previously. To accommodate this function, a new mechanism of cooperation between the two entities, VRRM and CRRM, is also proposed to define which information has to be exchanged in order to achieve an efficient interaction, while keeping a desired level of isolation.

Regarding the fairness problem, while in [Khat16], the definition of fairness is to minimise the deviation of services data rates from a nominal fixed value, the framework has been changed to a more flexible and accurate one, to cope with the concept of *proportional fairness*. This change comes with the price of a more complex objective function compared to the previous linear one. However, first of all, since the VRRM functionalities, such as the ones related to service orchestration for different VNOs, are *decoupled* from the underlying CRRM tasks associated with physical resource management among different RATs, the *complexity* of resource management approach is reduced, and second, as VRRM deals with the high-level network management including VNOs' policies, which does not require to be revised so often, this level of complexity is tolerable.

1.4 Research Strategy and Impact

The work developed in this thesis has done within the scope of a European project, COST Action CA15104, Inclusive Radio Communication Networks for 5G and Beyond (IRACON). Although contributing to this project has involved some additional effort beyond the main work of this thesis, it enabled the sharing of knowledge, visions and experience with multiple researchers from international research centres and universities.

In the development of this thesis, IRACON naturally had a significant influence over many decisions taken. Reciprocally, the research activity carried out in the thesis had also impressions on this project, which led to publishing and presenting several scientific papers in international conferences and

journals, as well as Technical Documents (TDs) within the framework of the project. Furthermore, sharing and discussing the obtained results with other researchers in project meetings opened potential directions for future collaborations based on the ideas of this thesis.

The work presented in the thesis was disseminated in several papers that have been published or submitted to several conferences and journals:

- International Journals:
    - B. Rouzbehani, L.M. Correia and L. Caeiro, "On the Benefits of RAN Slicing for RRM in Virtual Wireless Networks", (under 2nd round of review) *IEEE Access,* Oct. 2018.
    - B. Rouzbehani, L.M. Correia and L. Caeiro, "A Service-Oriented Approach for Radio Resource Management in Virtual RANs", *Journal of Wireless Communications and Mobile Computing,* Vol. 2018, Jul. 2018, pp. 1-13.
- International Conferences:
    - B. Rouzbehani, L.M. Correia and L. Caeiro, "An Optimised RRM Approach with Multi-Tenant Performance Isolation in Virtual RANs", in *Proc. of PIMRC'18 – 29th IEEE Symposium on Personal, Indoor and Mobile Radio Communications,* Bologna, Italy, Sep. 2018.
    - B. Rouzbehani, L.M. Correia and L. Caeiro, "Radio Resource and Service Orchestration for Virtualised Multi-tenant Mobile Het-Nets", in *Proc. of WCNC'18 – 19th IEEE Wireless Communications and Networking Conference,* Barcelona, Spain, Apr. 2018.
    - B. Rouzbehani, L.M. Correia and L. Caeiro, "A Fair Mechanism of Virtual Radio Resource Management in Multi-RAT Wireless Het-Nets", in *Proc. of PIMRC'17 – 28th IEEE Symposium on Personal, Indoor and Mobile Radio Communications,* Montreal, QC, Canada, Oct. 2017.
    - B. Rouzbehani, L.M. Correia and L. Caeiro, "A modified proportional fair radio resource management scheme in virtual RANs", in *Proc. of EuCNC'17 – IEEE 26th European Conference on Networks and Communications,* Oulu, Finland, Jun. 2017.

The main contributions made within IRACON were as follows:

- B. Rouzbehani, L.M. Correia and L. Caeiro, "An Energy Efficient Service-based RAN Slicing Technique in Virtual Wireless Networks", CA15104 TD (18) 08019 at IRACON 8th MC and Technical Meeting, Podgorica, Montenegro, Oct. 2018.
- B. Rouzbehani, L.M. Correia and L. Caeiro, "An SLA-Based Method for Radio Resource Slicing and Allocation in Virtual RANs", CA15104 TD (18) 07034 at IRACON 7th MC and Technical Meeting, Cartagena, Spain, Jun. 2018.
- B. Rouzbehani, L.M. Correia and L. Caeiro, "An Optimised RRM Approach with Multi-tenant Performance Isolation in Virtual RANs", CA15104 TD (18) 06013 at IRACON 6th MC and Technical Meeting, Nicosia, Cyprus, Jan. 2018.
- B. Rouzbehani, L.M. Correia and L. Caeiro, "Radio Resource and Service Orchestration for Virtualised Multi-tenant Mobile Het-Nets", CA15104 TD (17) 05011 at IRACON 5th MC and Technical Meeting, Graz, Austria, Sep. 2017.

- B. Rouzbehani, L.M. Correia and L. Caeiro, “A Fair Mechanism of Virtual Radio Resource Management in Multi-RAT Wireless Het-Nets”, CA15104 TD (17) 04001 at IRACON 4th MC and Technical Meeting, Lund, Sweden, May 2017.
- B. Rouzbehani, L.M. Correia and L. Caeiro, “A Modified Proportional Fair Radio Resource Management Scheme in Virtual RAN”, CA15104 TD (17) 03037 at IRACON 3th MC and Technical Meeting, Lisbon, Portugal, Feb. 2017.
- B. Rouzbehani, L.M. Correia and L. Caeiro, “A Model for Virtual Radio Resource Management in C-RAN”, CA15104 TD (16) 02012 at IRACON 2nd Technical Meeting, Durham, United Kingdom, Oct. 2016.

1.5 Content

This thesis is structured into seven chapters and one annex. The current chapter gives an introduction to the thesis by presenting a brief history of mobile and wireless networks in Section 1.1. Section 1.2 addresses the motivation and objectives of the thesis, followed by the highlighted novel aspects and concepts explored in Section 1.3. Section 1.4 provides an overview on the research strategy and the European project contributions and published work. Finally, this Section 1.5 covers the dissertation contents in detail.

Chapter 2 gives an overview of the fundamental concepts and definitions used to develop the thesis. Section 2.1 provides a brief introduction to the key parameters of network design. The framework of wireless network virtualisation is explained in Section 2.2, followed by the concept of network slicing in Section 2.3. Then the strategies of RRM among RAN slices are presented in Section 2.4. Finally, a state of the art of wireless network virtualisation and RAN slicing techniques for RRM is given in Section 2.5.

The details of the proposed analytical model for RAN slicing and service management are presented in Chapter 3. Section 3.1 defines the network architecture for the model, together with the functionalities of each entity. The main assumptions and inputs are provided in Section 3.2. Analytical models of VRRM, CRRM, mechanisms of interaction between the two, as well as the proposed approach of capacity aggregation from the underlying RATs, are presented in Section 3.3. In order to evaluate the model, a practical scenario is presented in Section 3.4, followed by the assessment of the model in terms of some key evaluation metrics, provided in Section 3.5, which concludes this chapter.

The focus of Chapter 4 is on explaining the details of model implementation in the simulator. In this regard, Section 4.1 gives an overview of simulator implementation, followed by a brief introduction to CVX solver in Section 4.2, as the main program to solve the convex optimisation problems. Different processes and key function blocks are explained in detail in Section 4.3, followed by the evaluation of the developed simulator in terms of accuracy and convergence, which is provided in Section 4.4.

Chapter 5 considers different variations in the input parameters of the reference scenario and evaluates the obtained results. In this respect, the reference scenario with some assumptions are defined in

Section 5.1, and then the effect of some key parameters, such as traffic load, SLA and number of active RATs on the model's performance is studied. Furthermore, by defining two different channel conditions, the effect of channel quality on the performance of the model is provided in Section 5.2. Finally, by choosing the same set of services for three different types of VNOs, the performance of the model in terms of guaranteeing a level of isolation among VNOs' policies is assessed in Section 5.3.

The main idea of Chapter 6 is to make a comparison between the efficiency of the VRRM model in the proposed thesis work, and the typical RRM deployed by Het-Net MNOs. Section 6.1 provides the assumptions for a scenario with a realistic traffic pattern in a 24-hour time span. The achieved statistical multiplexing gain is then compared between the VNOs and MNOs in residential and business areas. CRRM efficiency in terms of mapping service demands onto the underlying RATs is then analysed in Section 6.3 for this period of time. Section 6.4 and Section 6.5 are intended to make a comparison between the users' throughput and satisfaction between the VNOs and MNOs. Finally, the efficiency of VRRM model in terms of satisfying proportional fairness is addressed in Section 6.6.

The main conclusions of the thesis are presented in Chapter 7. Section 7.1 presents a summary of the thesis. Section 7.2 gives the main results obtained from the proposed model. The novelty and the key contributions are presented in Section 7.3, followed by Section 7.4 which aims at pointing out the main aspects that can be considered to be addressed in future work.

At last, there is one annex, Annex A, which presents the system Signal to Interference plus Noise Ratio (SINR) and data rate model for different mobile networks.

Chapter 2

Systems Overview

This chapter begins with an overview of mobile and wireless cellular networks and WLAN systems, and presents their main specifications in terms of network architecture, radio interface, coverage and capacity. Then, it continues by providing models, requirements and framework of network virtualisation. Key principles of network slicing, including RAN slicing techniques, are given in the following section. Finally, this chapter is concluded with the state of the art of virtualisation and RAN slicing approaches for RRM in wireless networks.

2.1 Cellular and WLAN Networks

2.1.1 Network Architecture

The general network architecture of key cellular technologies, comprising GSM, UMTS and LTE, along with connections, is presented in Figure 2.1. The functionalities of each node are explained in what follows [BAEK13].

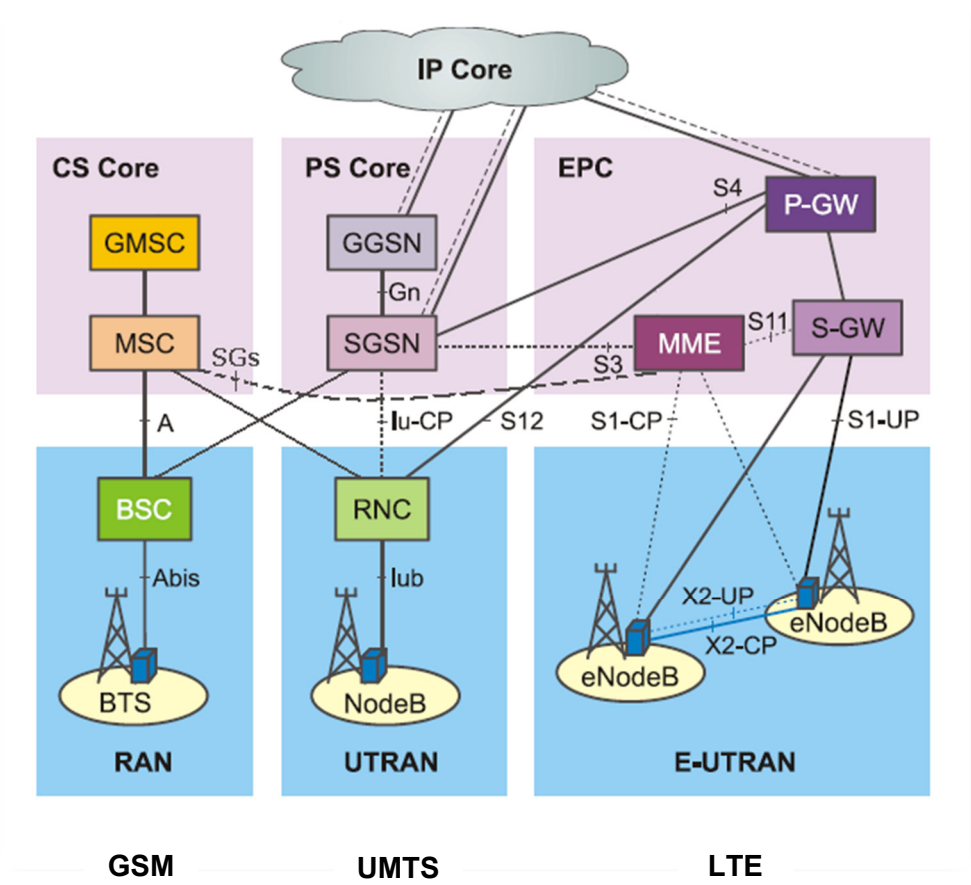


Figure 2.1 – The general architecture of a GSM/UMTS/LTE network (adapted *from* [Jacn14]).

The Mobile Switching Centre (MSC), which is connected to all main databases, is responsible for the key functions of Circuit Switch (CS) based services, being called the Gateway MSC (GMSC) when connected to other networks, either fixed or mobile. The Packet Switch (PS) elements in the 2G/3G core networks are mainly the Serving and the Gateway GPRS Support Nodes (SGSN and GGSN): the former is responsible for the delivery of data packets from and to Mobile Terminals (MTs), its tasks also including packet routing, mobility management (attach/detach and location management), logical link management, authentication and charging functions; the latter is responsible for interworking between the General Packet Radio Service (GPRS) network and external PS ones, e.g., Internet [BAEK13]. In LTE, a simplified flat packet-oriented network architecture was adopted: unlike GSM and UMTS, the Evolved NodeB (eNodeB), an evolution of UMTS' NodeB, is directly connected to the Serving Gateway (S-GW) with no separate control element. The S-GW is responsible for routing packets and acting as

mobility anchor during inter-eNodeB handovers, and also between LTE and other 3GPP technologies. The Mobility Management Entity (MME) executes a wide set of functions, such as the ones related to users' mobility, roaming, security and tracking [3GPP10]. The Packet Data Network Gateway (P-GW) is responsible for assigning Internet Protocol (IP) addresses to users, as well as classifying traffic into different QoS classes, and also acting as the mobility anchor point for inter-working with non-3GPP technologies, e.g., WLAN and Worldwide Interoperability for Microwave Access (WiMAX) [3GPP13b].

S1 is the interface among eNodeB, MME and S-GW, which has been separated into an S1-CP (control) and S1-UP (user plane) part [BAEK13], and X2 is the interface among eNodeBs. The X2-CP consists of a signalling protocol and the X2-UP is used to support loss-less mobility (packet forwarding). S11 is the control plane interface between the MME and S-GW. S3 interface is used to manage mobility between LTE and UMTS. S4 is the control and user plane interface between P-GW and SGSN. S12 interface is used between S-GW and UTRAN for direct user plane tunnelling during E-UTRAN and UTRAN handovers. The Gn interface is used for transferring packets between SGSN and GGSN in the transmission plane. Iu-CP is the signalling interface between RNC and SGSN. The Iub interface allows the RNC and the NodeB to negotiate about radio resources, for example, to add and delete cells controlled by the NodeB. In order to make or receive calls, the User Equipment (UE) may change its radio access technology from LTE to a 2G/3G technology, which supports CS-based services. This is accomplished by the SGs interface between the MME and the MSC. The GSM A interface provides two distinct types of information, signalling and traffic, between the MSC and the BSC, while the A-bis interface is responsible for transmitting traffic and signalling information between the BSC and BTS, and is the first actual physical connection for establishing a call.

WiFi WLAN's architecture has two basic operation modes, infrastructure and ad hoc [Sing10]:

- The infrastructure mode supports a wireless physical layer and a Medium Access Control (MAC) one, which are transparent to the wired Local Area Network (LAN) upper layers. In this mode, an Access Point (AP) covers a particular area, called the basic service area, and all communications either with Mobile Stations (MSs) or with the Internet, are made through the AP, which is connected to a LAN segment and enables access to the wired network.
- In the ad-hoc topology, MSs, which are in the range of each other, can communicate directly without a wired infrastructure. This is useful for data transfer among closely located users. One important feature in this scenario is the possibility of coverage extension. This can be accomplished by having an MS connected to the outside world via the infrastructure mode and at the same time to the other MSs in the ad-hoc one.

2.1.2 Radio Interface

In cellular and WLAN networks, the radio interface has been significantly improved throughout the years, leading to technological enhancements when established developments are publicised.

GSM uses both FDMA and TDMA [AgZe15], each frequency band being divided into 200 kHz carriers, allowing 8 users to share it. GPRS supports packet data transmission by employing the same modulation technique as GSM, Gaussian Minimum Shift Keying (GMSK), achieving a theoretical

maximum data rate of 171.2 kbps, using the aggregation of all 8 timeslots. EDGE enables to reach a higher data rate, of 384 kbps by introducing 8-Phase Shift Keying (PSK) modulation to coexist with GMSK.

UMTS is based on WCDMA, initially employing the Quadrature Phase Shift Keying (QPSK) modulation, to achieve a maximum theoretical throughput of 2 Mbps [Cox12]. Applying a chip rate of 3.84 Mcps leads to a carrier bandwidth of approximately 5 MHz. The enhanced performance in terms of data rate, system's capacity and latency introduced by High Speed Downlink Packet Access (HSDPA) and High Speed Uplink Packet Access (HSUPA), i.e., HSPA, improved user's UL and DL data rates up to 5.76 Mbps and 14 Mbps respectively. Further data rate increments became possible by releasing HSPA+, providing speeds of up to 22 Mbps in UL and 42 Mbps in DL. Technically, these capabilities were mainly achieved through employing higher order modulation 64QAM (DL) and 16QAM (UL), as well as 2×2 MIMO, used only in the DL [3GPP14].

LTE provides full IP-based functionalities through an efficient packet-based radio access, enabling low latency and low-cost network [Cox12]. Two multiple access techniques are applied: OFDMA for DL, in order to achieve good performance in frequency selective channels, and Single Carrier Frequency Division Multiple Access (SC-FDMA) for UL. The throughput depends on channel bandwidth, which is variable from 1.4 MHz to 20 MHz depending on the available spectrum. QPSK, 16-QAM and 64-QAM can be used to modulate data streams for both DL and UL, therefore, higher data rates up to 100 Mbps in DL and 50 Mbps in UL (even higher with MIMO) are achievable.

In LTE-Advanced (LTE-A), the main focus is on providing more capacity in a cost-efficient way. An increased number of active users with DL and UL peak data rates of 1 Gbps and 0.5 Gbps respectively is accessible through new functionalities, mainly comprising of Carrier Aggregation (CA), higher order MIMO configurations, i.e., 8×8 schemes for DL and 4×4 schemes for UL, Multi-User MIMO (MU-MIMO), and Het-Net deployments. LTE-A supports the aggregation of up to 5 component carriers, each having a bandwidth from 1.4 MHz to 20 MHz, thus, achieving a maximum aggregated bandwidth of 100 MHz (5 × 20 MHz) and is backward compatible with the legacy LTE, meaning that a given LTE user is able to communicate by using a bandwidth not greater than 20 MHz [AgZe15].

5G is the next major generation of mobile telecommunications standards. The initial specifications of 5G are defined by the ITU IMT-2020 standard [ITUR17]. A theoretical peak DL capacity of 20 Gbps is achievable with millimetre waves of 15 GHz and higher frequency bands, while 5G New Radio can include lower frequencies, from 0.6 GHz to 6 GHz, by supporting a maximum modulation order of 256-QAM. Speed in the lower frequencies is higher than the new 4G systems, being 15% to 50% faster, which makes 5G an evolutionary technology. 5G is not just about higher throughputs, but also supporting low-latency real-time applications, providing higher reliability and energy efficiency.

IEEE 802.11ac, as the newest amendment of WLAN, is intended to enhance the user experience by providing data rates up to 7 Gbps in the 5 GHz band, more than 10 times the speed that was standardised previously [IEEE13]. It adds channel bandwidths of 80 MHz and 160 MHz, with both contiguous and non-contiguous 160 MHz channels for flexible channel assignment, as well as higher

order 256-QAM modulation, and introduces MU-MIMO to support multiple concurrent DL transmissions. By employing smart antennas, MU-MIMO enables more efficient use of spectrum and higher capacity, supporting up to 4 simultaneous user transmissions, which is particularly useful for devices with a limited number of antennas, such as smartphones. Table 2.1 presents a performance comparison among the latest technologies of cellular and WLAN systems.

Table 2.1 – Basic comparison between the latest standards of cellular and WLAN systems (adapted from [IEEE13], [AgZe15] and [ITUR17]).

Standard	**GSM**	**UMTS**	**LTE-A**	**5G**	**WLAN**
Access Technique	FDMA /TDMA	CDMA	Layered OFDMA	Cyclic Prefix OFDM	CSMA/CA
Coding and Modulation Schemes	GMSK, 8PSK	Up to 64-QAM, QPSK	Up to 64-QAM	Up to 256-QAM	BPSK, QPSK, [16, 64, 256]-QAM MU-MIMO
Maximum Throughput [Mbps]	0.384	42	1 000	20 000	7 000
Frequency Bands [GHz]	0.85, 0.9, 1.8, 1.9	2.1	0.6 to 6.	0.6 to 6., and 24. to 80.	2.4, 5.4
Channel Bandwidths [MHz]	0.2	5	10, 20, 40, 80, 100	100, 400	20, 40, 80, 160
Latency Time [ms]	300	30	< 5	<1	< 5

2.1.3 Coverage and Capacity

Coverage refers to the geographical area wherein services are offered, while capacity refers to the highest aggregated peak rate (maximum theoretical throughput) for the area served by a BS or AP, throughput usually designating the data rate delivered to end users in a certain time duration and area. Coverage and capacity play fundamental roles in network design and optimisation. For cellular and WLAN networks, coverage and throughput are mainly related to bandwidth, transmit power, and network planning [LiYu14], major differences in coverage and capacity planning existing among technologies.

In GSM, coverage planning is made after the dimensioning phase and the only limitation in coverage is caused by path loss; BS sites can be planned fairly well without knowing the capacity requirement [LeMa04]. Each user has its own time slot, meaning that capacity depends on the number of RF transceivers installed in the BS, therefore, with the increase or decrease in the number of transceivers within a BS, the capacity of a cell can be changed to a higher or smaller value.

In UMTS, users share the same frequency band and are always interfering with each other, which means that capacity and interference are user dependent. Capacity depends more on the load in DL than in UL, since in the latter each user has its own power amplifier to transmit, while in the former the limited transmitted power of the BS is shared among all users [AgZe15]. On the other hand, coverage is UL limited, since the MT's power level is limited. As a result, the coverage and capacity planning have to be made together, because of capacity requirements and traffic distribution influence on coverage.

In LTE, cell capacity can be defined in terms of the maximum aggregated data rate, which can be served by a cell at a given time, depending on the number of Physical Resource Blocks (PRBs) allocated to all active users within that cell [Cox12]. Capacity should be evaluated according to the different service needs, in order to determine how many users can be concurrently served under specific QoS requirements, capacity is limited not only by bandwidth but also by each service requirement, which may include a variety of performance parameters, such as throughput, latency and SINR.

LTE-A introduces completely new technologies like Relay Nodes (RNs), CA and Het-Net, which significantly affect the current homogenous network planning and optimisation:

- RN is a new type of node entity in the RAN architecture, defined in Release 10, being backwards compatible with Release 8, targeted to improve data rates for cell edge users and other users with poor radio coverage, such as indoor ones [BBRR10]. RNs are wirelessly connected to a donor eNodeB and support full eNodeB functionalities, including encoding, decoding and packet scheduling, being clearly easier to deploy than traditional RF repeaters from the interference viewpoint. Currently, their deployment is limited to stationary and single hop relays; since RNs extend the deployment's coverage area, they should be considered in the network planning process for coverage and capacity analysis.
- In order to enhance network capacity, LTE-A defines Het-Net deployments, which consists of a normal macro-cell layout with some low power nodes placed throughout it [Cox12]. Accordingly, a Het-Net has completely different interference characteristics for a homogeneous deployment, and attention must be paid to these differences when planning the network. LTE-A also supports cell range expansion, where a cell selection bias is allocated to a pico-cell, which can be done for capacity planning, as it can improve throughput performance through an offloading effect towards the pico-cell; however, range expansion may lead to strict interference conditions for UEs in the extended area, which needs to be evaluated in the planning process.
- CA allows combining lower and higher bands, leveraging better coverage of the former with higher availability of the latter, allowing Het-Nets to make the best use of spectrum [Cox12]. In Release 10, 5 possible CA deployment scenarios along with their coverage patterns are defined.

In cellular networks, the cell size is related to the radio frequency range, therefore, putting some requirements on the radio interface design. Cell layers can be classified into four categories according to their coverage radius: macro-, micro-, pico- and femto-cells [Koro11]. Macro-cells are outdoor ones located in rural or suburban environments, with a large radius up to 35 km to support moderate MT speeds and narrowband services. Micro-cell sites are mainly situated in urban areas with a typical radius of up to 1 km to support low MT speeds and narrowband services. Pico-cells are small ones with a

radius of less than 100 m, suited for indoors, providing high traffic capacity for low MT speeds and wideband services [ITUR94]. Femto-cells are low power indoor BSs with a radius of less than 50 m, to provide Voice and broadband services to mobile users, and to improve both capacity and coverage inside buildings. Table 2.2 gives a general comparison among cell categories according to the cell radius, *R*, relative BS antenna position to neighbouring buildings, Δh, and transmitted power, P_t.

Table 2.2 – A general comparison among different cell categories (adapted from [Corr07]).

Cell	*R* [km]	Δh	P_t [dBm]
Macro	> 3	> 0	[50, 60]
Micro	0.1 – 1	≤ 0	[30, 47]
Pico	< 0.1	<< 0	[22, 33]
Femto	< 0.05	<< 0	[10, 25]

WLANs have a very large capacity compared to cellular networks; however, this characteristic is accomplished by sacrificing other aspects, such as coverage [WHWX05]. The deployment of WLAN networks is made mostly in indoor areas, which naturally decreases network coverage. Another limiting factor is the transmitted power level, which is relatively low, typically set to be less than 100 mW. In order to guarantee users mobility, the network coverage must be homogenous, therefore, in large networks; frequency reuse factor techniques can be applied.

WLANs can be considered as the best candidate RAT, supporting all types of applications using IP in a low mobility environment [AgZe15]. Currently, other types of networks, including cellular ones, are taking advantage of this high performance. When a user (of a given operator) detects the existence of a WLAN (supported by the same operator), it can handover to the WLAN, expecting a higher level of QoS. However, this technique is not applicable to all situations, specifically for intermediate or high user mobility scenarios, but only for the cases that MTs operate in a quasi-stationary manner.

Another typical feature of WLANs is that user throughput decreases when the distance between a user and an AP increases, leading to a non-uniform throughput coverage. Therefore, WLANs deployment should be carefully designed, in order to maintain user satisfaction in all areas to be covered.

2.1.4 Services and Applications

A service can be defined as a set of capabilities, which allows end users to establish applications, while an application is characterised by parameters associated with that service, communication links and traffic [Corr13]. Different services according to their specific QoS needs can be grouped into four major classes as defined in UMTS: Conversational, Streaming, Interactive, and Background. Table 2.3 summarises the key characteristic of each service.

Conversational has the most stringent QoS requirements among all, having low delay tolerance, and because of the real-time Conversational pattern, the time relation (variation) in between information

entities of the stream must be preserved, consequently, buffering is not an option to improve the offered QoS in this class of service. User's satisfaction can be achieved as long as sufficient radio resources are allocated to guarantee the minimum required data rate, and a negligible delay is maintained to avoid discontinuities in the Interactive Voice applications.

Similar to Conversational, Streaming is also characterised by the fact that the time relations (variation) between information entities (i.e., samples, packets) within a flow have to be preserved, although it does not have any requirements on low transfer delay, and the acceptable delay variation is much higher than the Conversational one. User's satisfaction can be achieved as long as the offered data rate is greater than or equal to a fixed nominal one.

Table 2.3 – Service classes and QoS requirements (extracted from [Corr13]).

Service Class	Conversational	Streaming	Interactive	Background
Real-time	Yes	Yes	No	No
Symmetric	Yes	No	No	No
Guaranteed Bit Rate	Yes	Yes	No	No
Affordable Delay	Minimum (Fixed)	Minimum (Variable)	Moderate (Variable)	High (Variable)
Buffer	No	Yes	Yes	Yes
Bursty	No	No	Yes	Yes
Example	VoIP	Video	Web	E-mail

Interactive services are more delay tolerant compared to the previous ones, which allows a reduction in the data rate assignment or even temporarily pausing traffic flows. It can be claimed that the data rate offered to each session of Interactive service has to be higher than or equal to a nominal one; furthermore, the serving time of each session has to be less than a maximum acceptable one.

Background services are characterised by the fact that destination is not expecting the data within a certain time, and therefore they are almost insensitive to delay and serving time, which puts them with the lowest priority among all others, traffic being usually handled during low network load periods.

LTE defines other types of QoS classes, each one being characterised by a QoS Class Indicator (QCI) and serving priority [BaLu11]. Table 2.4 presents the key parameters of LTE service classes, the different services being classified into two main categories according to their resource types:

- Minimum Guaranteed Bit Rate (GBR): these services have an associated GBR value, for which dedicated radio resources are always allocated. Bit rates higher than the GBR may be offered in case the resources are available at the time. If this is the case, a maximum bit rate value sets an upper bound to limit the bit rate, which can be assigned to that service.
- Non-GBR: for these services, no particular bit rate is guaranteed, and accordingly no bandwidth resources are allocated permanently to them.

Table 2.4 – LTE traffic classes (extracted from [3GPP16]).

QCI	Resources Type	Priority	Delay Budget [ms]	Error Loss Ratio	Example Services
1	GBR	2	100	10^{-2}	Conversational Voice
2		4	150	10^{-3}	Conversational Video (Live Streaming)
3		3	50	10^{-3}	Real-time Gaming
4		5	300	10^{-6}	Non-Conversational Video
65		0.7	75	10^{-2}	Mission Critical user plane push-to-talk Voice (e.g., MCPTT)
66		2	100	10^{-2}	Non-Mission-Critical user plane push-to-talk Voice
75		2.5	50	10^{-2}	V2X messages
5	Non-GBR	1	100	10^{-6}	IMS Signalling
6		6	300	10^{-6}	Video (Buffered Streaming)
7		7	100	10^{-3}	Voice, Video (live Streaming)
8		8	300	10^{-6}	TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive Video, etc.)
9		9			
69		0.5	60	10^{-6}	Mission Critical delay sensitive signalling (e.g., MC-PTT signalling)
70		5.5	200	10^{-6}	Mission Critical Data (e.g. example services are the same as QCI 6/8/9)
79		6.5	50	10^{-2}	V2X messages

This standardisation ensures that an LTE operator can expect identical traffic handling behaviour all over the network, regardless of the manufacturers of the eNodeB equipment.

Different services have different QoS priorities according to their requirements, therefore, they should be treated differently when there is not enough available network capacity to meet all users' needs. This means that, while deploying a new service in the network, the algorithm of resource management will select a decision corresponding to the specific QoS requirements of that service and the offered traffic carried throughout the network. If data rates reduction or temporary delaying strategies have to be applied in the network, the first services to be suffered are the ones with the lower QoS priority.

Concerning these QoS classes, different services can be considered. Each service is characterised by its maximum, average and minimum data rates, and its size or duration. These values for the listed services are presented in Table 2.5.

Table 2.5 – Service characteristics (modified from [Khat16]).

Service Class	Service	Data Rate [kbps]			Duration [s]	Size [MB]
		Min.	Average	Max.		
Conversational	Voice	12.2	32.0	64.0	90.0	–
	Video Calling	300.0	1000.0	5000.0	90.0	–
	M2M ITS	500.0	1000.0	–	–	2.0
Streaming	Video Streaming	1000.0	5120.0	13000.0	7200.0	–
	Music	64.0	128.0	320.0	120.0	–
	M2M Surveillance	64.0	200.0	384.0	–	5.50
Interactive	File Sharing	300.0	1024.0	–	–	5.0
	Web Browsing	100.0	500.0	–	–	3.0
	Social Networking (Facebook)	200.0	1000.0	–	–	6.0
	M2M e-Health	–	200.0	–	–	5.6
Background	Email	100.0	512.0	–	–	0.3
	File Transfer Protocol	200.0	1024.0	–	–	5.0
	M2M Smart Meters	100.0	200.0	–	–	0.01

Furthermore, regarding 5G service requirements, diverse use cases and applications are considered, covering not only the traditional services for mobile subscribers but also applications for a number of vertical industries, like the automotive, energy, eHealth or manufacturing sectors. All 5G use cases and applications can be assigned to one or more of the following three main usage scenarios, which are also shown in Figure 2.2 [MBQB18]:

- **Enhanced Mobile Broadband (eMBB)**, addressing human-centric use cases for access to multimedia content and data services. This usage scenario embraces a number of use cases and deployment scenarios with quite diverging requirements, e.g., in a hotspot scenario, extreme high throughputs and low-latency communications are in the foreground, while for wide area coverage the customer Quality of Experience (QoE) with reliable and moderate data rates over the coverage area is in focus.
- **Massive Machine-Type Communications (mMTC)**, characterised by the wireless connectivity of billions of network-enabled devices with prioritisation on wide area coverage and deep indoor penetration, typically transmitting non-delay-sensitive data at low rates. Usage scenarios include smart cities, smart buildings, or sensor networks for farming and agriculture.
- **Ultra-Reliable and Low Latency Communications (URLLC)**, having stringent requirements on latency and availability. Examples are the wireless automation of production facilities, monitoring of critical infrastructures in a smart grid, remote medical surgery, remote robotics, the tactile Internet, or vehicular traffic efficiency and safety.

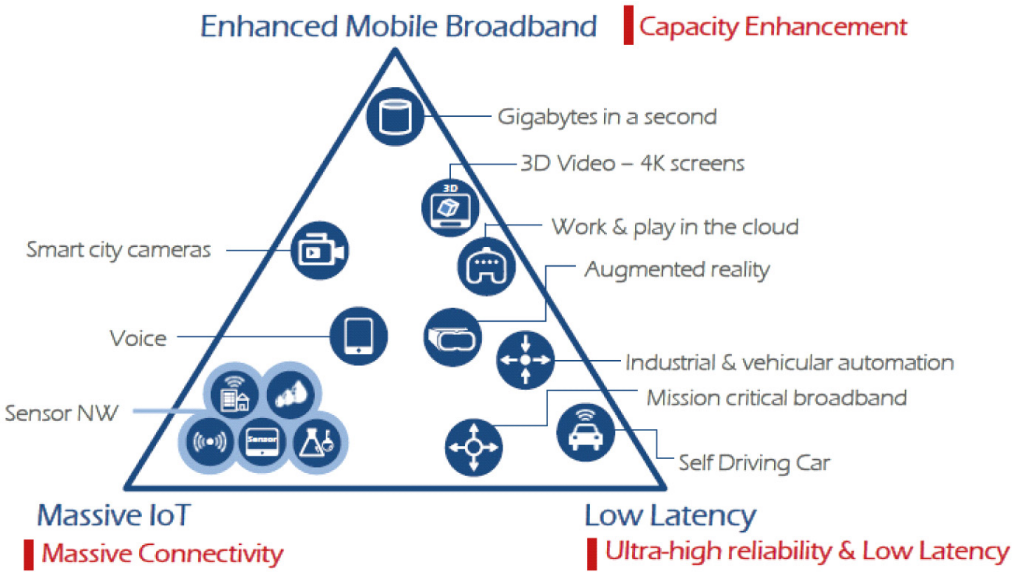


Figure 2.2 – Three main 5G use cases (extracted from [MBQB18]).

eMBB applications require a mixture of frequency bands, including lower bands for coverage purposes as well as for low to medium data traffic, and higher ones with large contiguous bandwidths to deal with the expected extremely high traffic demand. Licensed spectrum is essential to guarantee coverage obligations and a minimum QoS for customers. mMTC applications mainly demand a frequency spectrum range below 6 GHz, and spectrum below 1 GHz is needed in particular for wide area coverage and reliable outdoor to indoor penetration, therefore, the exclusively licensed spectrum is the preferred option. In addition, higher frequency bands and other licensing regimes may be considered, subject to the specific mMTC requirements. URLLC applications require high and reliable spectrum availability, thus, licensed spectrum is considered most appropriate for these services. Particularly for high-speed vehicles and in rural environments, spectrum below 1 GHz is well suited [MBQB18].

2.2 Virtualisation

This section starts with an overview of the concept of network virtualisation and the main projects done in this area. Then, two business models of network virtualisation are introduced and the role of each part in the models is explained. A general framework of network virtualisation based on the architectures of existing studies is given afterwards. Finally, Software-Defined Networking (SDN) is introduced in the last sub-section, as a complementary technology to network virtualisation, along with a possible architecture of the joint design, leveraging both techniques.

2.2.1 Network Virtualisation and Related Projects

Recently, virtualisation has moved from server virtualisation to network one [WTL13]. Server virtualisation, also known as host one, enables multiple users to share the same server through Virtual

Machines (VMs), by means of abstracting and decoupling the computing functionalities from the underlying hardware. With server virtualisation, on-demand and flexible management of computing resources are made possible; however, it does not involve any virtualisation of the network fabric, such as switches and routers. On the other hand, network virtualisation allows multiple isolated virtual networks to share the same physical network infrastructure, as shown in Figure 2.3, decoupling the network infrastructure from the services it provides, which allows virtual networks with truly differentiated services to coexist on the same infrastructure, maximising its reusability.

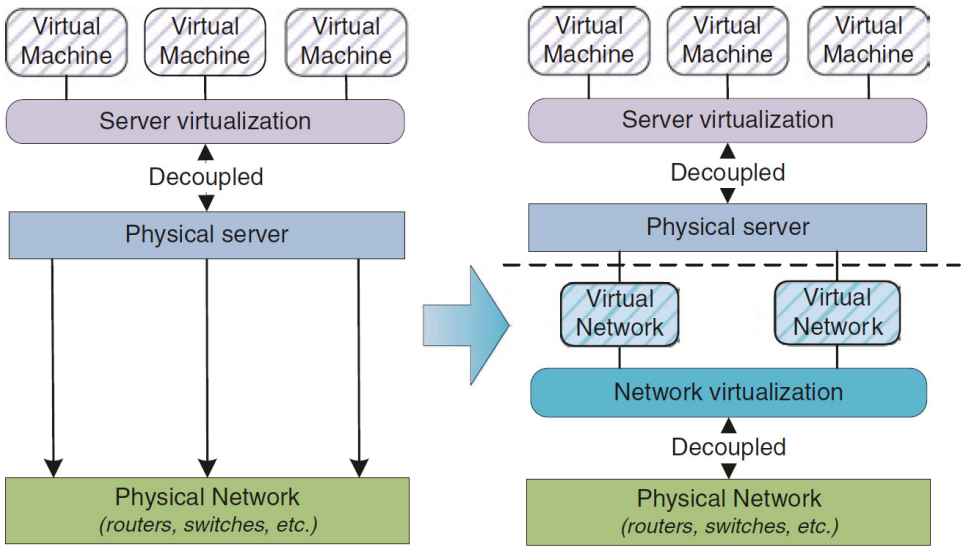


Figure 2.3 – Server versus network virtualisations (extracted from [WTL13]).

Network virtualisation, along with other recent development in SDN technologies, has become involved in cloud computing applications. As an instance, network virtualisation enables more flexible management of the network interconnection between physical servers in a large data centre. A new application of network virtualisation can ultimately allow cloud services to extend beyond the data centre, into the network infrastructure.

In recent years, several projects have been conducted around the world in the area of network virtualisation. A brief introduction to the most important ones, including CABO, 4WARD, SAIL, PlanetLab and GENI, is given as follows:

- CABO is the first project enabling full virtualisation, and the first one in which virtual routers can be mapped from one physical node onto another [FGR07]. The concept of separation between InPs and Service Providers (SPs) is promoted and improved by an integrated project. It also provides guarantees and customisation to SPs, in order to support end-to-end services.
- 4WARD introduced more detailed business models in addition to InPs and SPs, including Virtual Network Providers (VNPs) and VNOs, which provides more opportunities to the market [4WAR10], including significant work on resource allocation, and supporting the virtualisation of heterogeneous networking technologies. Another contribution was to implement network virtualisation not only in experimental networks and testbeds but also in realistic networks.
- SAIL was an integration and further enhancement to 4WARD [SAIL13]. The main objective was to explore the concepts of future internet, and to find solutions to implement it in the current

networks, in order to speed up their adoption and evolution towards the networks of the future. SAIL aimed at developing more responsive and efficient solutions to network related virtualisation scenarios, distributing the virtualisation concept all over the network. To this end, it enabled the delivery of cloud networking services by means of unified virtualisation techniques at network and resource levels, built on resource mobility concepts provided by 4WARD.

- PlanetLab was a global overlay network to develop and access broad coverage network services [SPBP06]. It proposed a concept of slice-ability, to allow multiple services to run concurrently and continuously, each in its own slice of PlanetLab. This was a key ability and design principle for the realisation of both wired and wireless networks virtualisation.
- GENI, which was one of the largest and most complex virtualisation-enabled testbeds in development, introduced virtualisation to wireless networks [PABC06]. The main goal of this project was to achieve a large-scale deployment of expanded testbeds to support Het-Nets, while aimed at exploring new design architectures for the future Internet. In GENI, virtualisation and slicing techniques were proposed by utilising TDMA and FDMA. Moreover, GENI gave researchers the opportunity to create customised virtual networks.

2.2.2 Models and Requirements of Wireless Network Virtualisation

With the tremendous growth in wireless traffic, it is natural to extend the concept of virtualisation to wireless networks. Wireless network virtualisation can be considered as a new technology, in which the physical wireless network infrastructure and physical radio resources can be abstracted and sliced into virtual wireless network resources to achieve certain corresponding functionalities, and to be shared by multiple parties, while their own functions and communication protocols are isolated from each other [LiYu14]. In other words, virtualising wireless network is to realise the process of abstracting, slicing, isolating and sharing the wireless networks.

In network virtualisation, some parties own physical resources, while others use virtual resources. The role of each party can be described as a business model in the network market. Generally, there are two logical rules after network virtualisation, MNO and SP. MNOs own all the infrastructure and radio resources of the physical substrate in networks, including Core Networks (CNs), Transmission Networks (TNs), RAN and the licensed spectrum. MNOs perform the virtualisation of the physical substrate into some virtual network resources, then SPs lease, operate and program the virtualised resources, in order to allocate them to users and satisfy the requirements of end-to-end services. The roles in this model can be further decoupled into a more particular role, comprising InP, Mobile Virtual Network Provider (MVNP), Mobile Virtual Network Operator (MVNO) and SP as shown in Figure 2.4. The functionalities of each role are explained in what follows:

- InP: it owns the physical infrastructure and network resources, but spectrum resources in some cases may or may not be owned by InP [LiYu14].
- MVNP: it is also named as Mobile Virtual Network Enabler (MVNE) in some references (e.g., [FoMD11]), lease the physical resources from InP, to create virtual resources. Some MVNPs may have their own licensed spectrum, and therefore, they do not need to request it from InP.

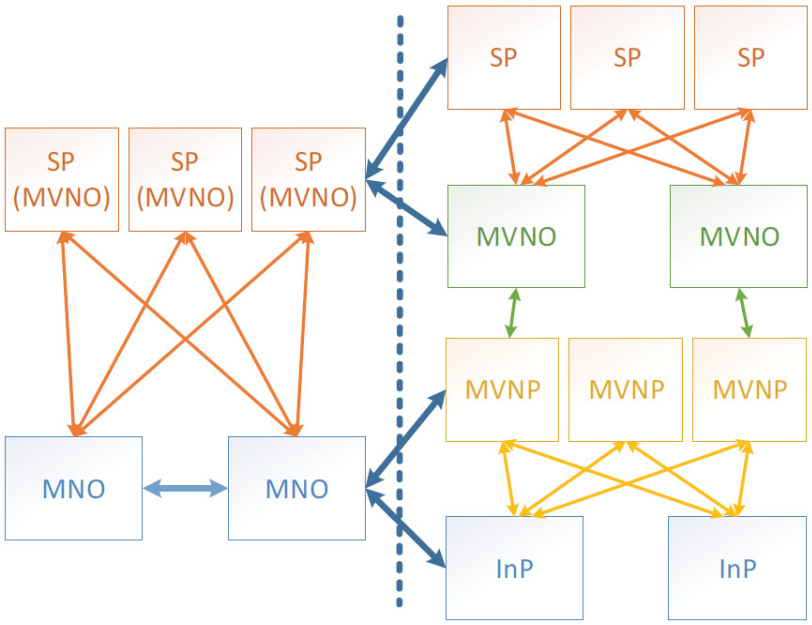


Figure 2.4 – Two and four level business models of network virtualisation (extracted from [LiYu14]).

- MVNO: it operates and allocates the virtual resources to SPs. However, in some approaches, the roles of MVNO also include the roles of MVNP. According to the evolving concept of the so-called XaaS in cloud computing [Xu12], the services provided in InP and MVNO can be fitted to this model as Infrastructure as a Service (IaaS) and Network as a Service (NaaS) respectively.
- SP: it offers services to its subscribers based on the virtual resources provided by MVNOs. In other words, virtual resources that are requested by SPs, managed by MVNOs and created by MVNPs, physically run by InPs in the infrastructure. Considering the XaaS concept, IaaS can be provided to MVNOs, and they can provide NaaS to SPs.

These two business models can be applied to both cellular and WLAN networks. For commercial cellular networks, in which WLAN access technology is also integrated as a supplement to cover the traffic in hotspot areas, WLAN networks can follow the discussed models. There is also another non-commercial model for WLAN, called testbed-as-a-service, in which MNOs and SPs are separated, [LPPA11].

2.2.3 Framework for Wireless Network Virtualisation

Generally, the framework of wireless network virtualisation (based on the architectures of existing studies), is composed of four main components: radio spectrum resource, network infrastructure, wireless virtual resource, and virtualisation controller, as shown in Figure 2.5 [LiYu14]. This framework defines the components, basic ideas and relationship in network virtualisation. The role of each component can be explained as follows:

- Radio spectrum resource, as one of the most important resources in wireless communications, is usually referred to as either licenced spectrum or some dedicated free one. With the evolving concept of cognitive radio, the range of radio spectrum is extended from a dedicated spectrum to white spectrum (the idle spectrum unused by the owner, which can be used by the users).

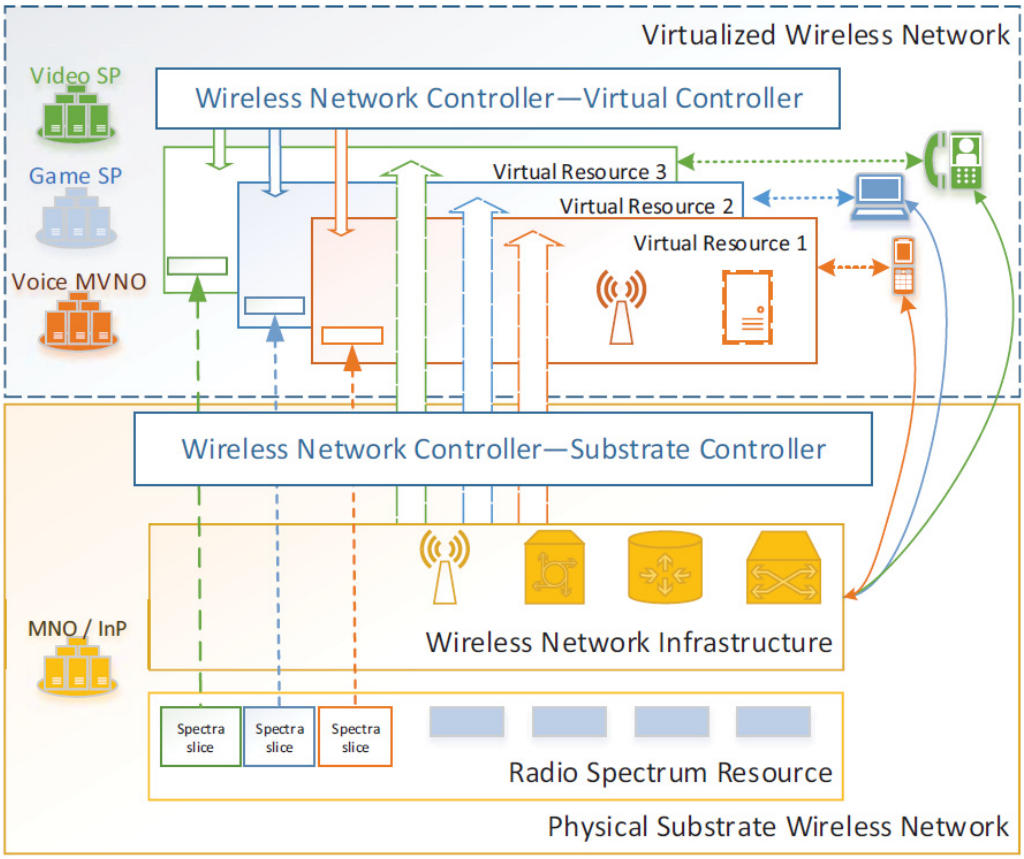


Figure 2.5 – A general framework of wireless network virtualisation (extracted from [LiYu14]).

- Network infrastructure is referred to as whole established physical components in substrate networks, comprising sites (towers and antennas), BSs (macro- to small-cells, relay, RF, baseband processors, radio resource controllers, etc.) in cellular networks, and APs in WLANs, core elements (gateway, switchers, routers, etc.), transmission networks (backhaul and links between RANs and CN).
- The virtualisation of wireless resources is achieved through slicing network infrastructure and spectrum into virtual slices, in a way that, a full or ideal virtual slice represents a universal virtual network, including all the virtual entities sliced by each element in the network infrastructure. However, in reality, a complete virtual slice may not be always necessary (e.g., some MVNOs may have their own CN and only need RAN slices [3GPP13]), therefore, based on different requirements, various levels of slicing and virtualisation, mainly comprising spectrum-level, infrastructure-level and full network-level can be customised.
- The wireless network controller consists of two parts: a substrate and a virtual controllers. The substrate is used by MNOs and InPs to virtualise and manage the substrate physical network, while the virtual one is available to MVNOs and SPs to manage the virtual slices and networks [LiYu14]. Specifically, MNOs use the wireless virtual controller to create and embed the virtual slices onto the physical substrate, while SPs use it to manage and program virtual slices, in order to customise their own end-to-end services.

2.2.4 Interworking of SDN and Wireless Network Virtualisation

SDN is an innovative, emerging approach, allowing network administrators to manage network services through the abstraction of lower-level functionality [YLJZ14], by decoupling the control plane from the data one. SDN enables the abstraction of network control functions into a logically centralised control plane. Consequently, network devices turn into simple packet forwarding and processing ones (the data plane) that can be programmed via an open interface, such as OpenFlow [MABP08].

Although SDN and network virtualisation are considered as different technologies in technical details and approach implementation, they greatly complement each other and achieve a promising direction for the study on future mobile and wireless networks [YLJZ14]. Decoupling control from data planes in SDN, while centralising the control one, can considerably improve programmability and customisation of network virtualisation. Meanwhile, virtualisation enhances scalability and flexibility, leading to the improvement of resource utilisation in SDN. These two approaches are mutually beneficial but independent, i.e., a network can be virtualised without a software defined approach and vice-versa.

A possible architecture of the joint design, leveraging both Software Defined Wireless Networking (SDWN) and network virtualisation techniques, is shown in Figure 2.6. This architecture introduces a logically centralised control plane according to SDWN, while network virtualisation is also supported to enable several concurrent virtual networks. The central control plane integrates network slicing by providing multiple SDNW controllers, each of which corresponds to one virtual network and schedules rules of its virtual network devices.

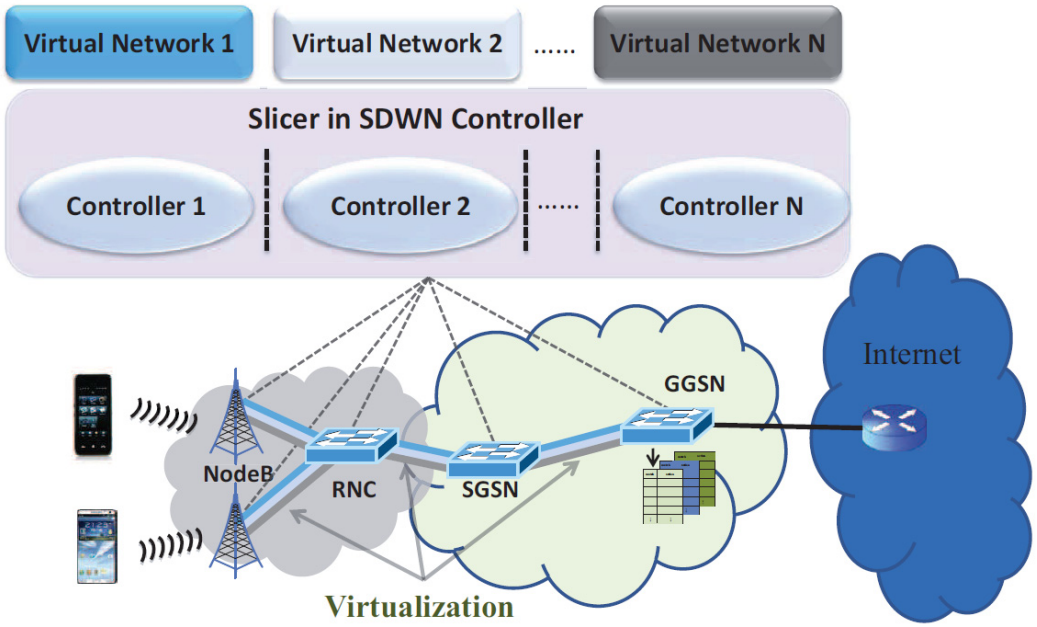


Figure 2.6 – Joint design for software defined wireless networking and wireless network virtualisation (extracted from [YLJZ14]).

2.3 Networks Slicing

2.3.1 Key Principles

The 5G network is promising to upgrade not only mobile broadband services, but also enable the support of services for the so-called “vertical industries” (e.g., health, transportation, factories, and energy) with their own requirements and needs, which can be highly opposing [MBQB18]. Their operational requirements are translated into different KPIs, such as user experienced data rate, reliability, communication efficiency, availability, and energy consumption, which have to be satisfied in specific environments characterised by diverse parameters, such as expected data traffic, density and types of available RATs.

Network slicing is introduced as one of the key enablers to support the required level of flexibility in 5G networks. In the latest specifications, 3GPP defines a network slice to be “a logical network that provides specific network capabilities and network characteristics” [3GPP17]. In [3GPP17b], it is specified that a slice is a network that is created by an operator and customised to provide an optimised solution for a specific market scenario that demands particular requirements with an end to end scope. These definitions suggest the ability to deploy multiple logical networks, possibly over the same physical infrastructure. This level of flexibility is needed to support the diverse requirements and KPIs of 5G use cases, as well as to reduce the cost for network deployment and operation. Therefore, network slicing is intended to be one of the main features of 5G networks, realised by introducing solutions based on virtualisation and functional modularisation.

In this regard, a comparison between 5G and legacy cellular networks is provided in Figure 2.7.

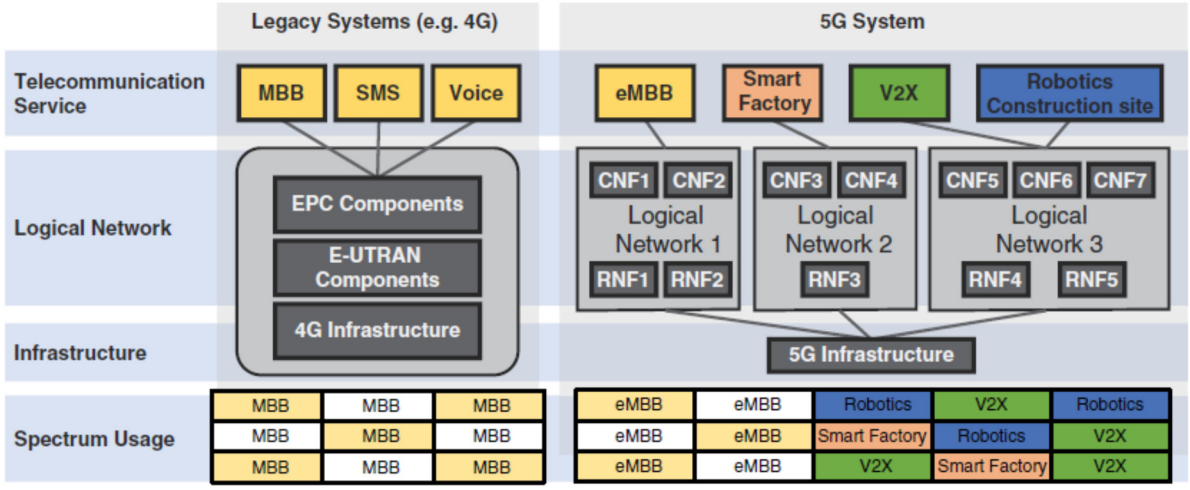


Figure 2.7 – A comparison between legacy cellular networks and slice enabled 5G (extracted from [MBQB18]).

On the left side, where legacy systems are depicted, one can see that the same network functions over monolithic network elements are used to support all telecommunication services, such as Voice or MMB. Such homogeneous management of all services, even with the definition of different QoS classes, is not

an acceptable compromise for 5G services, which is due to the fact that different KPIs like ultra-high reliability and ultra-low latency cannot be deployed for all vertical services in a feasible technical and economical way [MBQB18]. In 5G networks, multiple verticals as logical networks will be running on top of the infrastructure, being composed of Core Network Functions (CNF) and Radio Network Functions (RNF), and will run over the same underlying physical components.

It is notable that dedicated spectrum may be allocated to each slice, or several slices may share the same spectrum, while being managed to meet the SLAs according to the customised requirements of each vertical. This implies that 5G has to be flexible and highly *granular* to handle and prioritise different traffic types, or even different packets belonging to the same traffic type. In this respect, a key limitation in the LTE QoS architecture [LiYu14] is that the finest granularity of differentiating mobile data is on the level of *radio bearers*. Within one bearer, which can be established to reflect GB or non-GB traffics, and characterised by a QoS class identifier (reflecting priority, acceptable delay, and packet loss rate), all packets are treated in the same way.

The key enablers for the dynamic deployment of slices are considered to be function virtualisation, allowing the virtualisation of sets of network functions and organising them into virtual blocks that could be connected together to create communication services, and SDN, which is used for separating control plane and user plane, allowing full programmability of the network [ETSI18].

Regarding control and service orchestration, the concept of *network-wide orchestration* is applied to 5G, which is correlated to the replacement of the individual device configuration, by a more powerful service management mechanism, which is able to provide network-wide services definition, configuration, deployment and monitoring [MBQB18]. By using this concept, services are not configured and managed in a node-by-node fashion, as it is originally done in legacy systems such as 3G or 4G. The increased complexity that 5G will impose on the delivery and maintenance of services, supported by the slicing framework, forces to manage the network in an integrated and coordinated approach, and not as a collection of individual boxes and layers.

Therefore, *higher-level abstractions* and automated procedures are needed to manage and configure every single component of the whole network service at once. The main advantage of the network-wide orchestration is that it gives a single point of integration, providing a *centralised* representation of the distributed network no matter the number of resources involved or where they could be physically placed. This provides great opportunities for accessing KPI measurements and automation, as well as possibilities for the deployment of advanced services across all network domains. This way, the associated complexity of deploying more granular customised services can be simplified by decoupling the underlying automation and monitoring tasks, from the orchestration module. According to the slicing concept, 5G has to consider two levels of orchestration among the slices: *intra-operator slice orchestration*, which deals with the management of resources to address different requirements of network slices belong to one operator, and *inter-operator slice orchestration*, referring to the orchestration of resources among different operators' slices [5GPP16].

2.3.2 Impact of Slicing on the RAN

Slicing support for the RAN has been extensively researched during the past few years. In this section, the latest status of 3GPP standardisation activities as well as elaborate on the key principles are discussed. As currently stated by 3GPP [3GPP18], slicing in the RAN can be realised by MAC scheduling and by providing different configurations for the network functions. The different treatment among slices can be achieved by using specific identifiers in signalling messages to indicate specific slices. The system supports policy enforcement among slices as per SLAs, and is able to apply the best RRM approach to support the slice-specific SLA. Since the RAN can support multiple slices, resource isolation mechanisms have to be deployed accordingly. The MAC-layer scheduling requires to be aware of slice definition and user membership in order to apply the RRM policies.

One aspect about network slicing that is especially relevant in RAN is the notion of *sharing the same radio resources and physical infrastructure* among multiple slices, ideally to the largest achievable level, The majority of proposed 5G use cases are expected to be only economically feasible if they can exploit significant interactions with other use cases that do not require dedicated infrastructure. However, there may be cases where some physical separation of slices is demanded by involved parties or even put by some other administrative domains according to specific laws.

Figure 2.8 explains some illustrative RAN slicing scenarios with different levels of resource sharing and infrastructure reuse among slices, from a full slice separation (left) to maximum multi-slice RAN integration (right) [MBQB18].

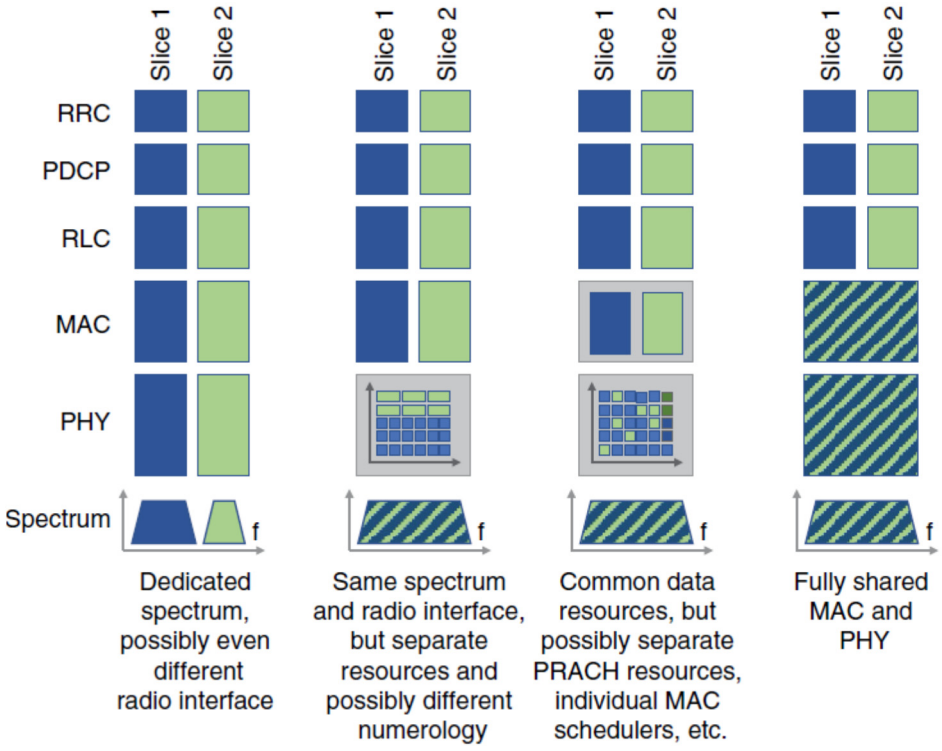


Figure 2.8 – RAN slicing scenarios with different levels of resource sharing and isolation (extracted from [MBQB18]).

On the very left, one can see the case where two slices use dedicated spectrum and possibly different radio interface specifications as well, which can be considered as the legacy case. The second one from left, represents the case when slices share the same spectrum, but still use strictly separated physical resources within each slice; in this case, each one would likely also have dedicated MAC instances and schedulers for different slices. As a further extension of sharing concept, multiple slices can share the same spectrum, and most of the resources, but still have some dedicated resources, such as dedicated Physical Random Access Channels (PRACH), in order to guarantee some strict slice-specific QoS service requirements. In the last case on the right, it is possible to have a fully shared and integrated MAC and physical layer (PHY) for both slices. Note that if the PHY is largely shared, one could further consider having a common MAC instance across multiple slices, for instance enabling multi-slice MAC scheduling. Many different customisations are possible in this respect, i.e., one may, for instance, propose that the two slices use a same high-level MAC scheduler that allows for some flexible resource split among slices, while the slices involve some finer-granular dedicated schedulers per slice.

2.3.3 RRM among Slices

Pooling and sharing the scarce radio resources among network slices in an efficient way is an important target. The basis for allocating resources in a slice-aware approach is to monitor the status of the network slices with respect to their SLAs [LAL17]. This could take place in a new entity of the RAN, such as an access controller, which has to be aware of the existing network slices and their SLAs, as well as which data stream belongs to which network slice. Corresponding information can be obtained via signalling from the CN. The enforcement of the network slice specific requirements can be realised with different levels of complexity. Figure 2.9 shows an example of RAN slicing scenario with fully shared PHY and MAC, assuming that two different Air Interface Variants (AIVs) are integrated into one 5G interface. Based on the outcome of the SLA monitoring, the QCIs of the individual data streams are adjusted and traffic steering is executed. Enforcement of SLAs happens by adapting the QoS classes of individual data streams. For instance, if the SLA of a network slice guaranties a certain latency, any data stream of this slice could be mapped onto a corresponding QoS class [MBQB18].

A sublayer above the PDCP level is introduced by 3GPP [3GPP18], which handles the mapping of QoS flows onto certain data radio bearers. The principle of this newly introduced SLA control loop is shown in Figure 2.10. The SLA controller would be a functionality of the newly introduced sublayer in the 5G architecture, while the SLA monitoring functionality could be handled in a central unit. Slice-specific multiple thresholds are defined to protect the SLA of each slice, together with additional weights to prioritise the service slices. These thresholds are related to relevant slice-specific KPIs, which have to be satisfied, such as guaranteed throughput or latency, but also business-driven additional agreements have to be protected. Realising specific data flow to meet the QoS requirements is a task of the MAC scheduler. With this approach, slice-specific QCI adaptation is *transparent* to the MAC scheduler. Therefore, no further adaptation is needed to the scheduler, and a common MAC scheduler could be used for multiple slices [MBQB18].

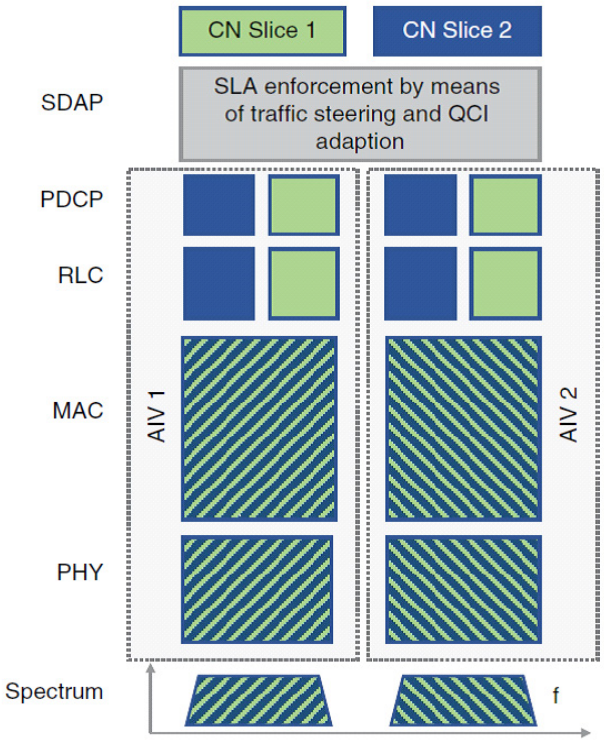


Figure 2.9 – An example of implementation for multi-slice RRM (extracted from [MBQB18]).

More specifically, the SLA controller adapts the data flows with the derived QCI classes and forwards it to the destination BSs. The adapted data packets will then be received and arranged into the new corresponding QCI-related buffers, which are considered as an input for the MAC scheduler. The scheduling decisions are taken based on these new buffer statuses to assign a certain amount of bandwidth from the underlying RATs. RRM functionality needs to monitor the statistics of the flows, based on QoS KPIs and send it back to the SLA monitor.

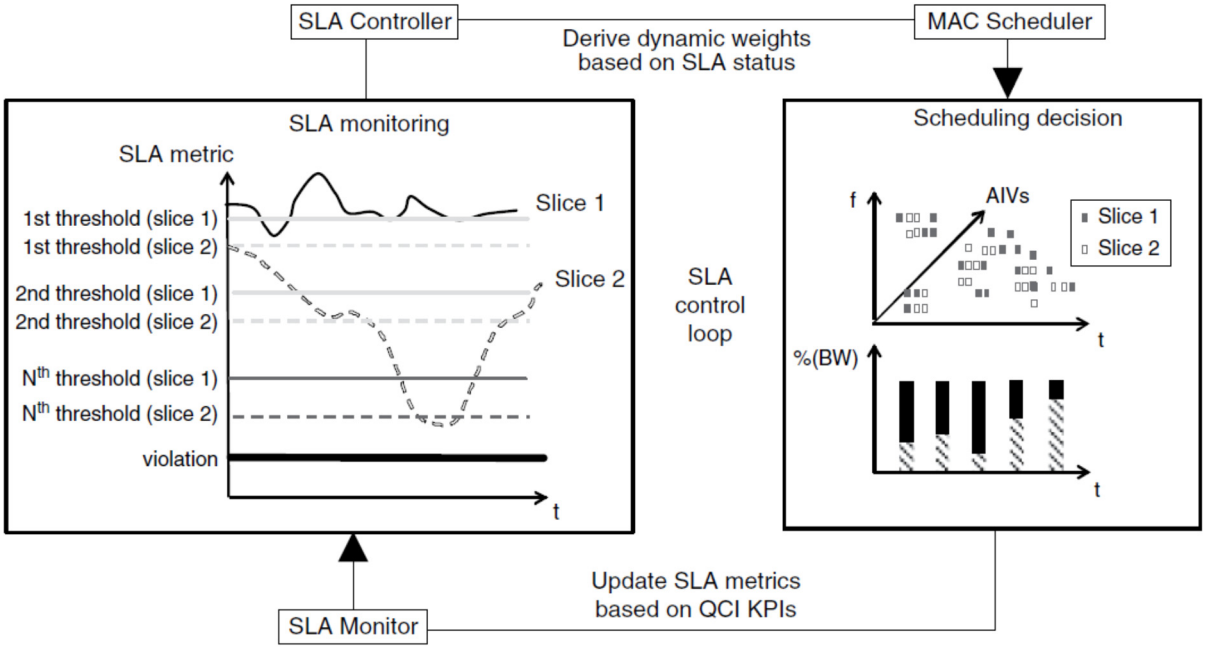


Figure 2.10 – SLA control loop (extracted from [MBQB18]).

2.4 State of the Art of Virtualisation and RAN Slicing for RRM

Although there are quite extensive studies to address different challenges in traditional wireless networks, when it comes to the evolutionary technologies for future wireless communications, the existing techniques have to be modified in order to accommodate the specifications of new services and architectures. Specifically concerning the problem of RAN slicing and RRM in virtual wireless networks, which is a rather new topic, there is a lack of effort to thoroughly cover the key parameters, such as customised service provisions, isolation between virtual slices, fairness and the mechanisms of interaction between different entities, however the research in this area is growing rapidly.

Current works propose different models of the allocation problem, from *low-level* and highly technology-dependent ones (i.e., resource-based models) to more general *high-level* models (such as throughput-based models) [KMZR12]. As an example for the LTE case, a number of works propose resource requests comprising of PRBs to be assigned to each slice [ZZGT10]. This category has the advantage that the requests are given in allocation units, simplifying the implementation of the allocation procedure. However, such low-level requests are difficult to be controlled by high-level management entities such as operators or service providers. A more high-level model considers a fraction of resources per request. In this case, there is not an exact demand of a number of PRBs but a relative percentage of resources. In the case of LTE, this would translate easily to PRBs, but this model could also be used for other technologies where resources are quantified differently [RBS16].

In [ZZGT10] and [ZZGT11], a framework for LTE virtualisation is suggested. The authors propose an architecture for virtualising the LTE BS (eNodeBs in LTE architecture) in a multi-tenant scenario with different operators sharing the same physical resources. The model is based on proposing a hypervisor, which hosts different virtual nodes, allocates the resources and is responsible for spectrum sharing and data multiplexing. The hypervisor uses the PRB as the minimum resource granularity which can be allocated and assigns them among different virtual nodes, instead of distributing them among the users (as performed by traditional schedulers). PRBs are scheduled to different virtual eNodeBs based on predefined contracts, which specify different guarantees for the operator owning a virtual eNodeB. After the hypervisor allocates PRBs to the virtual eNodeBs, each of them allocates the PRBs to its connected users. In order to show the benefits of the proposed approach, it is assumed that multiple operators have their traffic peaks at different moments of time. Although the scheduler handles the coexistence of different slices over a shared physical eNodeB, the proposed model does not provide any explicit mechanism for isolation between the operators.

In [LZLZ12], the framework from [ZZGT11] is taken and extended through a more detailed algorithm for scheduling PRBs in the virtual nodes. The objective of the model is to flexibly allocate PRBs based on the estimated demand of the slices. The demand is estimated separately for GBR traffic and non-GBR traffic. The allocation goal is to satisfy the GBR demands and then to allocate PRBs to non-GBR traffic proportionally. If the total demand overloads the number of available resources, the assignment is done proportionally to each slice demand. One interesting aspect of this solution is introducing a load balancing mechanism to distribute the load of a slice among different eNodeBs. The main idea is that if

an eNodeB is overloaded and a neighbour eNodeB has available resources, a user is selected to be migrated to the unloaded eNodeB.

In all the previously mentioned works, slicing techniques are based on the number of PRBs that are guaranteed and provided to operators, which seems to be a too low-level approach. This would raise problems for a tenant slice, which is not supposed to have enough knowledge or information from the underlying physical RATs to define the optimum number of PRBs for service demands. Moreover, this does not assure and maintain a fixed performance when considering variable rates and variable channel conditions [RBS16]. Agreements related to more high-level variables such as a percentage of the total resources would be more appropriate in this regard. In addition, slicing based on PRB scheduling will require a major modification in the scheduler, which is not a trivial task, because of the complexity of the scheduling algorithms. On the other hand, slice scheduling and traffic shaping techniques do not need to modify the PRB scheduler and are generally easier to deploy [RBS16]. Also, as these are higher level mechanisms, the allocations are made on fractions of the total resources or on guaranteed bandwidths.

A high-level RAN slicing approach is proposed in SoftRAN [GPLK13], introducing a big virtual BS that logically groups geographically close physical BSs. The idea is that these physical BSs can be centrally managed to facilitate radio resource allocation and interference mitigation. To this end, the authors propose an abstraction view of the radio resources through the virtualisation of the physical ones. In this work, the control plane of the wireless devices is decoupled from the hardware. However, service customisation and fairness issues are not taken into consideration.

An example scenario to apply slicing based on QoS requirements in future 5G networks is given in [ErDa15]. This scenario consists of a future network operator, offering differentiated types of services depending on the specific use case. For example, a high-throughput service for smartphones, a low-rate non-critical service for IoT or machine to machine communications and a low-latency service for critical real-time communications. Therefore, the scenario is a combination of these use cases, each one with its specific requirements, and the operator has to provide service and management for all of them jointly. However, there is a lack of effort to define an analytical model or implementation approaches in the real world.

In [TAV16], a model for RAN virtualisation is proposed, which provides a solution for the mapping of virtual network elements onto radio resources, as well as providing an algorithm of radio resource negotiation and allocation in a general term, which does not include specifications of different services into account. The problem of slice scheduling and performance isolation in Virtual RAN is addressed in [RBHR16], but, like the previous work, different needs of applications in network slicing, as well as fairness, are neglected. Although a fairness concept based on roaming price is defined in [FGYK17], and some QoS parameters are addressed in the proposed generalised virtualisation framework, the matter of isolation between services and operators is not covered.

Economic models for pricing based on optimising an objective function have been proposed with the purpose of balancing network throughput and users' fairness as two competing interests, [SAR10]. Among the available mechanisms, proportional fairness has proved to be an effective approach when the objective is to maximise the average long-term users' data rate [YuZh14], [WCLM99]. In this regard,

it is suggested that by employing a logarithmic utility function in the slice scheduler of Virtual RAN, an effective mechanism of fairness based on the concept of proportional fairness can be achieved among virtual slices [CSGM13]. A model of RRM is proposed in [SYHT16] for 5G wireless infrastructures, which aggregates users' traffic across a Wi-Fi/LTE network, by considering an α-fair mechanism in the objective function, including proportional fairness as its special case; however, there is no effort to address the specifications of different services. Another approach for cooperative RRM in 5G heterogeneous cloud RANs is proposed in [GMF15], which considers a modified max-min technique as an alternative to proportional fairness, although still the evaluation of the effect of different service parameters on model performance is neglected.

Chapter 3

Models and Algorithm Development

The main focus of this chapter is on presenting the analytical models for radio resource management, RAT selection and load balancing in virtual RANs, as well as the level of interaction between different components of the network and their functionalities. In order to evaluate the models, a reference scenario is defined and the performance of each component of the model is assessed by some specific metrics. Primary results obtained from the assessments confirm that the model is capable of satisfying the proposed set of goals.

3.1 Network Architecture

Network architecture is an important step towards understanding system functionalities. The proposed architecture shown in Figure 3.1 refers to the virtualisation of a heterogeneous wireless network, which comprises of two main components: managing data rate allocation among the different service classes of the VNOs in the virtual one, as well as radio resource scheduling and load balancing among different RATs in the physical one.

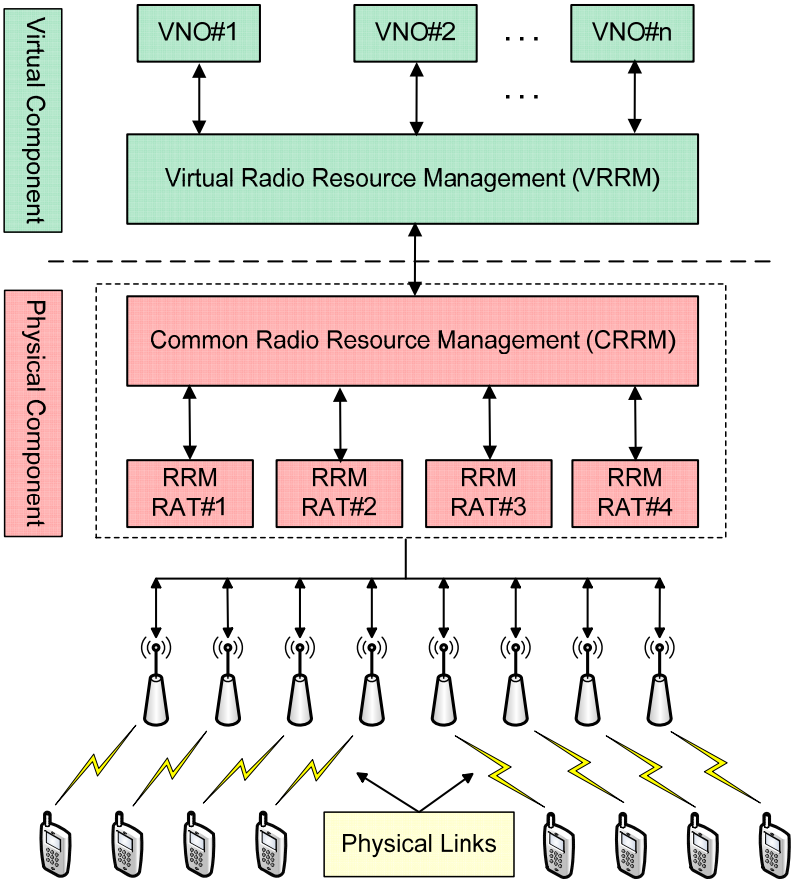


Figure 3.1 – Architecture of the proposed model for resource management in virtual RAN.

A description of the different network components is provided in what follows:

- VNOs are network operators that do not own the physical infrastructure and need wireless connectivity to serve their subscribers. It is assumed that each VNO can support users from different classes of service. VNOs ask for the total required capacity (data rate) of each service from VRRM according to the priority of each service and contracted SLAs. In this way, they do not have to deal with the physical resources to serve their users.
- Each VNO is created inside a VM. The data communication between VMs and the underlying hardware is carried out through virtual channels, which are logical isolated paths to share the capacity of the network among VNOs. The isolation among virtual paths ensures that traffic, mobility or fluctuations in physical channel conditions of a VNO's subscribers do not affect the service level of the other ones.

- VRRM is in charge of translating VNO's requirements to lower level physical parts of the network, as well as assigning the demanded data rate to VNOs. The goal in data rate allocation procedure of VRRM is to maximise the utilisation of the aggregated virtualised capacity, fulfilling the SLAs while considering the priority of various services from different VNOs, as well as implementing an efficient and flexible fairness algorithm to satisfy all users. Accordingly, VRRM as a centralised entity is supposed to be capable of sharing the aggregated capacity provided by the shared physical infrastructure among different VNOs with specific SLAs.
- CRRM in the physical part of the network provides the required capacity for VRRM by demanding a portion of physical resources from different RATs. The main goals of CRRM include RAT selection and load balancing. The CRRM algorithm is in charge of maximising the global utilisation of RRUs, while associating the most suitable RATs among the available ones, to different services considering their priority, which also helps to keep signalling overheads as low as possible. Furthermore, this algorithm should be capable of avoiding the unnecessary congestions of BSs resulting from high load scenarios.
- Local RRMs, in the lower level, are responsible for optimising radio resource usage in a single RAT, being in charge of allocating physical radio resource parameters, such as power, frequency bandwidth, and time-slots, to end-users upon receiving requests. The set of policies delivered by CRRM to each local RRM are used as decision guidelines for them; in addition to CRRM's policies, resource allocation in each local RRM has to meet the QoS requirements of each service.
- Physical channels are a group of RRUs allocated to end-users from physical BSs (the RRU is the minimum radio unit that can be allocated to an end-user in a BS, depending on the RAT, e.g., a time-slot in TDMA or a code in CDMA). The mapping between physical channels onto virtual ones is necessary to compute the VNO's required aggregated capacity and can be considered as a metric to monitor the level of satisfaction between VNOs and InPs.

In order to point out the main differences between traditional Het-Nets and the proposed model for Virtual-RANs, as well as to explain the functionalities of the involved parties in the RRM model, the hierarchical network architecture is shown in Figure 3.2, which is consistent with the suggested 3GPP business model for *wholesale-only* network sharing, in which InPs do not offer service to end-users, rather selling capacity to businesses that do not own the infrastructure [CSGM13].

The service connectivity request in typical Het-Nets is directed to CRRM, as the usual entity for network management, and processed centrally to be assigned to a suitable RAT according to a decision criterion. On the other hand, in the proposed architecture for Virtual-RAN, demand for a specific service and capacity directly goes to the linked VNO as the service provider. This capacity needs to be delivered respecting the SLA agreements between user and VNO.

In contrast to the existing Het-Nets, the role of network operators is separated from InPs, accordingly, VNOs on top of the hierarchy, do not own the infrastructure, rather sharing the radio resources from different RATs owned by InPs. As a result, from their perspective, it is not important by which technology they are being served, as long as the SLAs are satisfied. VNOs ask for Capacity-as-a-Service from a

centralised virtualisation platform called VRRM [KhCo17], which does not exist in the current architecture of Het-Nets, and is placed on the top of CRRM. VRRM is in charge of managing the total available capacity provided by CRRM, through aggregating all the RRUs from different RATs, which are OFDM (based on Wi-Fi), OFDMA (based on LTE), CDMA (based on UMTS) and TDMA (based on GSM), and sharing the capacity into separated slices associated with different services of the VNOs.

By providing isolation and element abstraction, VRRM enables each VNO to deploy its own protocol stack over the same set of RRUs per RAT (e.g., resource-blocks in LTE, codes in UMTS, time-slots in GSM, and carriers in Wi-Fi) [LiYu14]. Low-level physical RRUs are aggregated, abstracted and virtualised into isolate slices, in order to provide high-level resources so that a BS or AP can be shared among various VNOs, therefore, promoting the notion of multi-tenancy in a heterogeneous virtualised environment with existing several access techniques. Accordingly, the total available radio spectrum is considered as a whole resource to be split into virtual *resource slices*, each slice being bandwidth-based and associated with one individual service of a VNO, based on the certain demanded data rate of the provided applications.

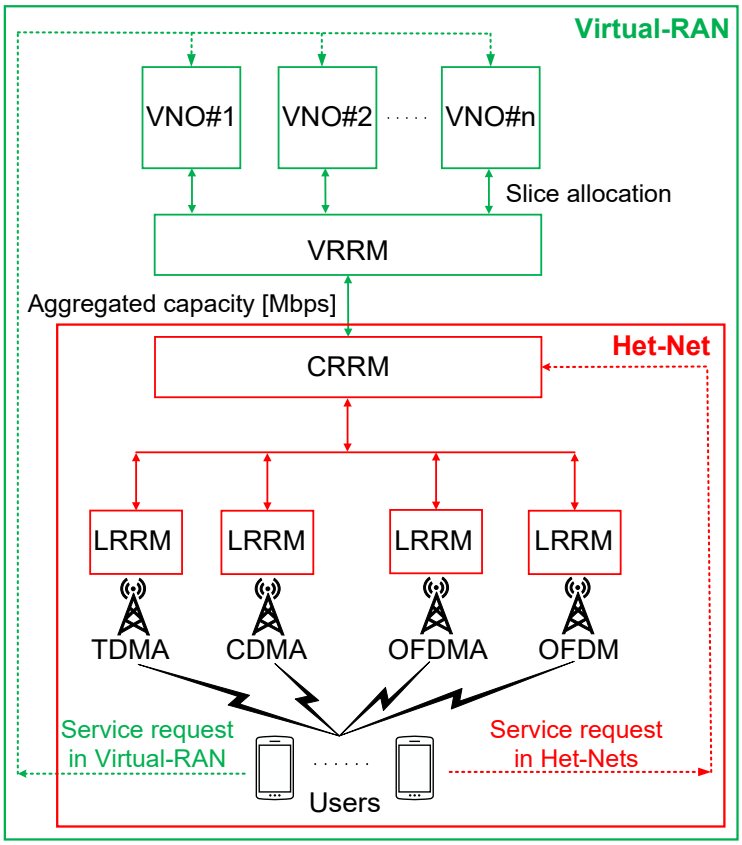


Figure 3.2 – Differences between the traditional heterogeneous network and virtual RAN.

VRRM has to closely interact with CRRM, by translating VNOs' requirements and different SLAs into a set of management policies for the lower levels. These policies contain information about the demanded capacity of each service, as well as their priorities. In return, CRRM provides VRRM with the monitoring reports and information (e.g., the available aggregated capacity) to enhance its performance. CRRM is in charge of managing the resources of different RATs to satisfy the requests of VRRM, by demanding each RAT to provide a portion of available capacity to be assigned to end users. The service-to-RAT

association mechanism takes the load and suitability of each RAT for performing specific services into account. The performance of CRRM is being optimised, based on the information coming from Local-RRMs (LRRMs), which are in charge of managing the allocation of local RRUs from each RAT to the connected end users.

3.2 Assumptions and Inputs

3.2.1 Virtual Network Operators SLAs

The VRRM model has to consider both customised service requirements, such as priority, class of service and acceptable data rates, as well as different contracted SLAs between the InP and various VNOs. It is assumed that all VNOs can provide the four service classes, namely, *Conversational*, *Streaming*, *Interactive* and *Background*.

Conversational (e.g., VoIP) and Streaming (e.g., Video) service classes are delay sensitive, however, they have almost constant data rates, therefore, assigning data rates higher than the contracted ones does not improve users' QoE. In contrast, for Interactive (e.g., file sharing) and Background (e.g., Email) service classes, the increase in data rate assignments can actually enhance users' QoE. Based on the service set and requirements, VNOs may have different SLAs, which can generally be categorised into three types of contract [KhCo14]:

- Guaranteed Bitrate (GB), the highest priority category, for which minimum and maximum thresholds for data rate assignment have to be always guaranteed, regardless of the variation of traffic load and network status; therefore, users are always expecting a good QoE in return of a relatively higher service price.
- Best effort with minimum Guaranteed (BG), the second highest priority, for which just a minimum data rate is guaranteed, and higher data rates are served in a best effort manner; from the users' viewpoint, a service with acceptable QoS and affordable price is expected.
- Best Effort (BE), the least prior type, for which there is no level of service guarantees and users are served in a pure best effort manner; consequently, in extreme case of high traffic loads, BE users are the first ones to suffer.

Figure 3.3 represents the range of data rates to be assigned to the three aforementioned categories of SLAs, as well as the priority of each. The lower bounds for GB and BG SLAs are the minimum guaranteed data rates. Similarly, the upper bound for BG and BE is the aggregated capacity provided by CRRM to VRRM, while the maximum guaranteed data rate limits the data rate assigned to GB services.

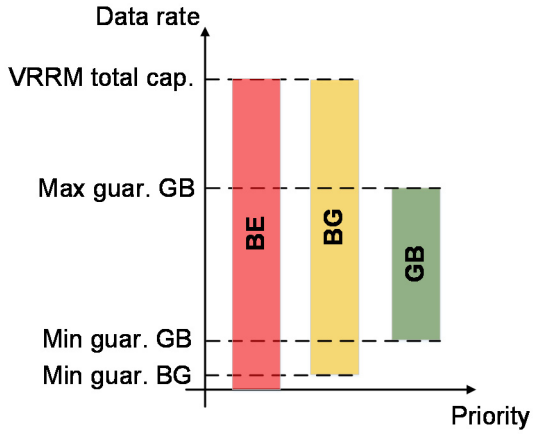


Figure 3.3 – Comparison between different SLAs.

3.2.2 Discussion about the Granularity and Time Scale

Network traffic can be viewed from three different levels as shown in Figure 3.4. According to IETF recommendations for traffic management, radio resource management can be performed at flow level, with the time scale of seconds [Rais04], however, the default time scale value used by the most of commercial BS schedulers at flow level is set to one second [CMKR13]. Regarding the VRRM time scale, in the framework of the CROWD project for resource management of SDN-based 5G RAN, it is recommended to perform it at the time scale of IP flows [CROWD17], accordingly in this work, one second is considered for the time interval of VRRM's algorithm operation, as each virtual slice is comprised of sets of traffic flows to be scheduled.

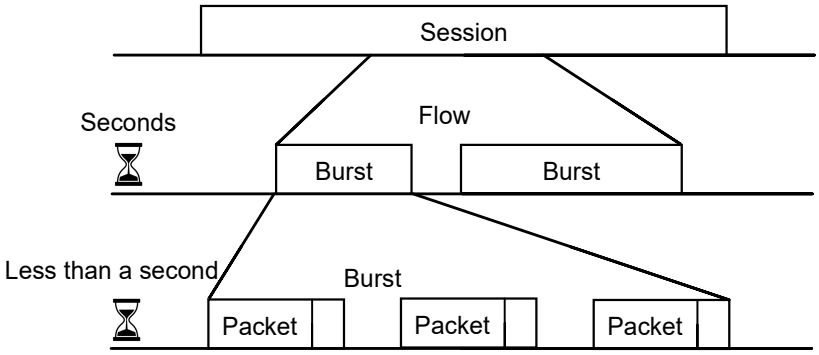


Figure 3.4 – Traffic levels in IP networks (updated from [Bidg12]).

CRRM's operation timescale is highly dependent on the level of its interaction with LRRMs. Given that CRRM has the perspective of all RATs, it can be proposed that the interaction and operation for some higher level objectives, such as RAT selection and load balancing, occur in the order of seconds to minutes [PSAD05], as shown in Figure 3.5.

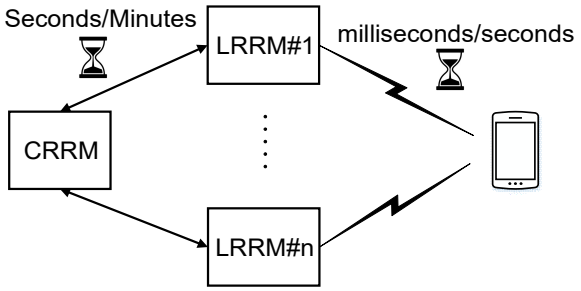


Figure 3.5 – Interaction between CRRM and LRRMs (extracted from [PSAD05]).

3.3 Analytical Models and Design Methods

3.3.1 VRRM Model

The primary goal of VRRM is to maximise the usage of the aggregated capacity, which is calculated and provided by CRRM, in order to satisfy the contracted SLAs to the highest possible level, considering services' priority, while distributing capacity in a fair manner, subject to some constraints, including maximum achievable capacity, predefined SLA thresholds, and users access to different types of RATs.

To realise these goals, the VRRM's analytical model is formulated as a constrained concave optimisation problem [BoVa09]. The objective function, f_{VRRM}, is defined with the aim of balancing between the efficiency and fairness when allocating the resources in a network with heterogeneous services, being a measure of efficiency, mapping a portion of network bandwidth that users use onto a real number, and quantifying the expected users' satisfaction, given the allocated resources. This way, VRRM builds a bridge between the functionalities of MAC and higher layers, by optimising the allocation of radio resources for different applications [MiSo15].

Among all the options for a utility function, the most important and widely used family of utility functions is the generalised *power family*, which includes the **linear**, ***n*th-root**, and **logarithmic** ones, therefore these three types were selected as study use cases, in order to formulate the problem in the context of convex optimisation.

- **Linear**, as a choice of utility, defines a direct relation between the portion of allocated capacity to a user and the corresponding satisfaction of that user from the provided service, meaning that the level of user's satisfaction linearly increases or decreases with an increase or decrease in the allocated data rates, respectively.
- ***n*th-root**, as a strictly concave choice of utility, shows the property of diminishing returns [BoVa09], meaning that users' satisfaction increases with the decreasing rate, as the amount of allocated bandwidth increases, hence, it can better address human nature in decision making.
- **Logarithmic**, considered as a common choice of a concave utility function, which also copes with the criterion of proportional fairness, has proved to be an effective approach when the

objective is to maximise the average long-term users' data rate, while sharing the capacity among users in accordance with predefined weights [Kell08].

The linear function is chosen as a special case, since it is both convex and concave. The resulting optimisation task is classified as a linear programming problem, therefore, one can formulate the problem of the objective function of VRRM, $f_{VRRM}(\mathbf{R^{srv}})$, as the normalised weighted sum of the total data rate for different services as follows:

$$f_{VRRM}(\mathbf{R^{srv}}) = \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \frac{R^{srv}_{v_s[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}} \tag{3.1}$$

where:

- N^{srv}: number of services,
- λ_{v_s}: tuning weight associated with service s, provided by VNO v, to prioritise data rate assignment,
- R^{VRRM}: total offered capacity from CRRM to VRRM.
- $R^{srv}_{v_s}$: total served data rate of service v_s,
- $\mathbf{R^{srv}}$: vector of serving data rates, which can be written as:

$$\mathbf{R^{srv}} = [R^{srv}_1, R^{srv}_2, \dots, R^{srv}_{N_{srv}}]^T \tag{3.2}$$

By putting an upper bound to R^{VRRM}, as a constraint to limit the total assigned data rate, the optimisation problem can be written as (3.3), the left side of the constraint representing the total data rate allocated to all existing services, being apparent that this value cannot surpass the aggregated virtualised capacity, which is provided by CRRM,

$$\begin{aligned} &\underset{\mathbf{R^{srv}}}{Max}\ f_{VRRM}(\mathbf{R^{srv}}) \\ &s.t: \sum_{v_s=1}^{N^{srv}} R^{srv}_{v_s[\text{Mbps}]} \le R^{VRRM}_{[\text{Mbps}]} \end{aligned} \tag{3.3}$$

Tuning weights λ_{v_s} are supposed to define service priorities as well as tuning the portion of the data rate to be allocated to each service, however, a linear utility function is not suited to reach this goal, as it blindly trades the fairness for overall system throughput, thus being hard to manage. In [Khat16], a model for VRRM based on solving a linear objective function is considered; the author puts effort to overcome this problem by defining an extra constraint as a fairness mechanism to the original formulation, in order to somehow control the resource sharing among the different VNOs. Nevertheless, there are two major concerns associated with this modified function: first of all, it is still not properly tuneable with the pre-defined serving weights, and second, introducing an unnecessary constraint to the problem increases its complexity. To show that the linear objective fails in properly distributing the available capacity among different services, assume that λ_{v_s} has the highest value among all serving weights, in which case, it is obvious that the solution vector is

$$\mathbf{R^{srv}} = [0,0, \dots, R^{VRRM}_{[\text{Mbps}]}, \dots, 0]^T \tag{3.4}$$

It turns out that the model with linear objective assigns all the available capacity to the service with the highest priority and serving weight, regardless of the variation of the other serving weights, which is not

desired. To overcome this problem, instead of adding more constraints to the original model, the choice of the objective function can be changed to a strictly concave utility. The n^{th}-root function (such as square or cube root), as already discussed, can be considered as a common choice for the strictly concave utility, in which case, (3.1) can be modified to

$$\underset{\mathbf{R^{srv}}}{Max} \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \sqrt[n]{\frac{R^{srv}_{v_s[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}}} \tag{3.5}$$

To solve this optimisation problem, one may use the standard technique based on Lagrange multipliers [BoVa09], as a strategy for finding the maxima of a function subject to inequality constraints. Since both objective and constraint have continuous first partial derivatives, a new variable, the Lagrange multiplier, is introduced to form the $L(\mathbf{R^{srv}}, \mu)$ Lagrangian as follows:

$$L(\mathbf{R^{srv}}, \mu) = \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \sqrt[n]{\frac{R^{srv}_{v_s[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}}} + \mu \left(1 - \sum_{k=1}^{N^{srv}} \frac{R^{srv}_{v_s[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}} \right) \tag{3.6}$$

where:

- μ: Lagrange multiplier corresponding to the inequality constraint.

by taking derivatives with respect to the variables, one obtains:

$$\begin{cases} \dfrac{\partial L(\mathbf{R^{srv}}, \mu)}{\partial R^{srv}_{v_s}} = 0 \\ \dfrac{\partial L(\mathbf{R^{srv}}, \mu)}{\partial \mu} = 0 \end{cases} \longrightarrow \begin{cases} \dfrac{\lambda_k}{\sqrt[n]{\left(\dfrac{R^{srv}_{v_s[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}}\right)^{n-1}}} = \mu, \qquad v_s \in \{1, 2, \dots, N_{srv}\} \\ R^{VRRM}_{[\text{Mbps}]} - \sum_{v_s=1}^{N^{srv}} R^{srv}_{v_s[\text{Mbps}]} = 0 \end{cases} \tag{3.7}$$

by jointly solving the two equalities in (3.7), the final solution is derived as follows:

$$R^{srv}_{v_s[\text{Mbps}]} = \frac{\sqrt[n]{\lambda_{v_s}^{n-1}}}{\sum_{v_s=1}^{N^{srv}} \sqrt[n]{\lambda_{v_s}^{n-1}}} R^{VRRM}_{[\text{Mbps}]} \tag{3.8}$$

The Logarithmic utility function captures resource allocation according to the criterion of proportional fairness [Kell08], which is a compromise-based scheduling algorithm, based on maintaining a balance between two competing interests: trying to maximise the total wireless network throughput, while at the same time providing all users with at least a minimal level of service. The proposed utility in (3.1) can be modified to the logarithm objective function as follows:

$$\underset{\mathbf{R^{srv}}}{Max} \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \log \frac{R^{srv}_{v_s[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}} \tag{3.9}$$

Once more, by using the Lagrange multiplier technique to solve the problem and the Lagrangian function can be expressed as [BoVa09]:

$$L(\mathbf{R^{srv}}, \mu) = \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \log \frac{R^{srv}_{v_s[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}} + \mu \left(1 - \sum_{v_s=1}^{N^{srv}} \frac{R^{srv}_{v_s[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}} \right) \tag{3.10}$$

by taking derivatives with respect to the variables, two equalities are obtained as:

$$\begin{cases} \dfrac{\partial \mathrm{L}(\mathbf{R^{srv}}, \mu)}{\partial R_{v_s}^{srv}} = 0 \\ \dfrac{\partial \mathrm{L}(\mathbf{R^{srv}}, \mu)}{\partial \mu} = 0 \end{cases} \longrightarrow \begin{cases} \dfrac{\lambda_{v_s}}{R_{v_s[\mathrm{Mbps}]}^{srv}} = \mu, \qquad v_s \in \{1, 2, \dots, N_{srv}\} \\ R_{[\mathrm{Mbps}]}^{VRRM} - \sum_{v_s=1}^{N^{srv}} R_{v_s[\mathrm{Mbps}]}^{srv} = 0 \end{cases} \tag{3.11}$$

By solving the two equalities in (3.11), one gets the final solution, the allocated data rate of each service being proportional to its serving weight,

$$R_{v_s[\mathrm{Mbps}]}^{srv} = \frac{\lambda_{v_s}}{\sum_{v_s=1}^{N^{srv}} \lambda_{v_s}} R_{[\mathrm{Mbps}]}^{VRRM} \tag{3.12}$$

By comparing the results in (3.12) and (3.8), it is apparent that there is a relation between the service weights of the logarithmic utility function and the n^{th}-root one. Considering the vector of serving weights for the n^{th}-root utility function, it is possible to have exactly the same set of solution of the n^{th}-root function, by using the logarithmic one, if the service weights of the latter are adjusted as follows,

$$\lambda_{v_{s_{log}}} = \left(\lambda_{v_{s_{nth}}}\right)^{\frac{n}{n-1}}, \qquad \forall\, k \in \{1, 2, \dots, N_{srv}\} \tag{3.13}$$

where:

- $\lambda_{k_{log}}$: equivalent service weights of the logarithmic utility function,
- $\lambda_{k_{nth}}$: given service weights of the n^{th}-root utility function.

Considering the main objective of VRRM in this work, which is to maximise the usage of the aggregated virtualised capacity provided by CRRM, while maintaining a level of fairness in the process of resource allocation among users according to their priority of requested services, the goal can be formulated as a maximisation concave utility [Song99], in the form of a weighted logarithmic function to cope with the criterion of proportional fairness. Furthermore, as shown in (3.12), the serving weights of the logarithmic objective can be flexibly tuned to include the same behaviour of the n^{th}-root function.

$$\underset{\mathbf{w^{usr}}}{\mathrm{Max}}\, f_{VRRM}(\mathbf{w^{usr}}) = \underset{\mathbf{w^{usr}}}{\mathrm{Max}} \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \log\left(\sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} \frac{R_{v_s\,[\mathrm{Mbps}]}^{srv_{max}}}{R_{[\mathrm{Mbps}]}^{VRRM}}\right) \tag{3.14}$$

where:

- $N_{v_s}^{usr}$: number of users performing service s, from VNO v,
- $R_{v_s}^{srv_{max}}$: maximum assignable data rate to the user of service s, from VNO v,
- $w_{v_s,i}^{usr}$: assigned weight to user i, performing service s, from VNO v, ranging in [0, 1],
- $\mathbf{w^{usr}}$: vector of users' weights, to obtain the long-term average data rate of users,

$$\mathbf{w^{usr}} = [w_{1,1}^{usr}, w_{1,2}^{usr}, \dots, w_{1,N_1^{usr}}^{usr}, \dots, w_{N^{srv},1}^{usr}, w_{N^{srv},2}^{usr}, \dots, w_{N^{srv},N_{N^{srv}}^{usr}}^{usr}] \tag{3.15}$$

The tuning weight is one of the parameters to realise the isolation between InP policies and VNOs internal decision for capacity sharing among their own services, which is composed of a multiplication of two independent positive integer numbers. Given (3.16), γ_v is defined by InP and assigned to VNO v according to the type of its SLA agreements to VNOs' priorities in capacity sharing, while δ_s is a serving weight, assigned to service s, performed by VNO v, to project the internal policy of each VNO in

distributing capacity among the services provided by that VNO. The general framework can be defined as follows: a service with a higher priority receives a higher serving weight, and the lowest value among all the services that each VNO provides is always assumed to be 1.

$$\lambda_{v_s} = \gamma_v \ \delta_s \tag{3.16}$$

There are some constraints associated with the problem, and the objective function has to be solved respecting these constraints. Considering that each VNO has a specific policy to define a *customised* range of data rate variation for each service, the average long-term data rate assigned to each user has to fall within this acceptable data rate interval:

$$R_{v_s\,[\text{Mbps}]}^{srv_{min}} \leq w_{v_s,i}^{usr}\ R_{v_s\,[\text{Mbps}]}^{srv_{max}} \leq R_{v_s\,[\text{Mbps}]}^{srv_{max}} \tag{3.17}$$

where:

- $R_{v_s}^{srv_{min}}$: minimum assignable data rate to the user of service s, from VNO v.

Another constraint is related to the SLA types of each VNO, which can be expressed as a range of contracted capacity between VNOs and InPs. This way, the aggregated capacity that has to be provided to all the users of a specific VNO needs to be varied within the acceptable range of SLA thresholds:

$$R_{v\,[\text{Mbps}]}^{vno_{min}} \leq \sum_{v_s=1}^{N_v^{srv}} \sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr}\ R_{v_s\,[\text{Mbps}]}^{srv_{max}} \leq R_{v\,[\text{Mbps}]}^{vno_{max}} \tag{3.18}$$

where:

- $R_v^{vno_{min}}$, $R_v^{vno_{max}}$: minimum, maximum contracted data rates of VNO v (agreement between InP and VNO),
- N_v^{srv}: number of services provided by VNO v.

In addition, there is a logical constraint similar to the one discussed in (3.3), which indicates that the whole bandwidth allocated to all users cannot exceed the total aggregated capacity provided by CRRM. Therefore, the entire VRRM bandwidth assigned to all users is actually subject to an upper bound defined by the InP,

$$\sum_{v_s=1}^{N^{srv}} \sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr}\ R_{v_s\,[\text{Mbps}]}^{srv_{max}} \leq R_{[\text{Mbps}]}^{VRRM} \tag{3.19}$$

The last constraint is due to the fact that the access of users to different RATs depends on their location, therefore, assuming that a certain number of users is enjoying full coverage of all available RATs, while the rest of them have just access to the cellular ones and not Wi-Fi. Accordingly, the limit to the access of the latter group to cellular RATs is defined as

$$\sum_{v_s=1}^{N^{srv}} \sum_{i=1}^{N_{v_s}^{usr_{cell}}} w_{v_s,i}^{usr}\ R_{v_s\,[\text{Mbps}]}^{srv_{max}} \leq R_{[\text{Mbps}]}^{VRRM_{cell}} \tag{3.20}$$

where:

- $N_{v_s}^{usr_{cell}}$: number of users with access only to cellular RATs, performing service v_s,

- $R^{VRRM_{cell}}$: available capacity of VRRM, provided by cellular RATs.

The proposed problem is solved by CVX [CVX17], which is a modelling system based on MATLAB, developed by Stanford University for disciplined convex programming. The method used to solve the problem is primal-dual interior point.

For a better understanding of the VRRM algorithm, the geometrical vision of the objective function and constraints are shown in Figure 3.6, under simplified assumptions of the canonical scenario. All services are categorised by their service classes as either GB or BG, and users have access to all existing RATs. The objective would be to find the corresponding data rates of R_{BG} and R_{BG} that maximise the utilisation of total aggregated capacity. All strait lines represent the linear constraints of the problem, creating the bounded feasible region. Each point in this region satisfies all the constraints. The iso-objective curves correspond to the different values assigned to the logarithmic objective function when the ratio of tuning weights for GB and BG services is 5/2. By moving from the left to right sides of the figure these values increase. As the objective is to find the maximum values of R_{BG} and R_{GB}, the solution is achieved when the objective curve is tangent to the maximum capacity constraint, R^{CRRM} (assuming that 100% of CRRM capacity is provided to VRRM), this point being unique and located on the border of the feasible region. One can see that the capacity is allocated to each service class in accordance with their serving values.

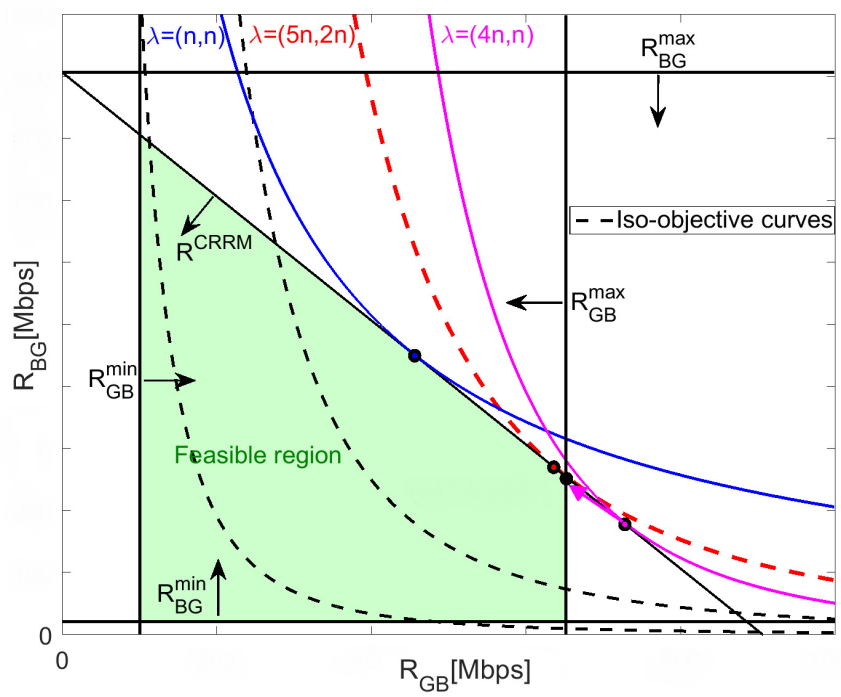


Figure 3.6 – Geometric expression of VRRM algorithm.

As another example, when the serving weights for both categories are equal, the unique solution that is marked on the figure, equally divides the capacity between the two. However, this is not the case when the ratio of serving weights is comparatively higher. If the maximiser is located outside of the feasible region, the share of data rate for GB category will be proportional to its serving weight, but higher than the contracted maximum threshold. Accordingly, the solution, in this case, will be shifted to the closest

vertex to this point, which is the one that satisfies the highest level of contracted GB services and allocates the rest of capacity to the remaining BG ones.

A general trend of data rate assignment by considering an extreme situation is shown in Figure 3.7. The first part of the figure, before Th_C, represents the case when there is still enough capacity to serve all users. At the start, GB users are well served with the "constant maximum data rate" defined in the SLAs and the rest of network capacity is shared among BG ones. Consequently, the general expectation is to observe for BG users a data rate higher than for GB ones when the number of users is relatively low. As the traffic load increases up to Th_C, all the serving rates of different services reach their minimum thresholds. Meanwhile, Conversational can potentially be served with the highest data rate mostly, owing to the highest priority and comparatively low data rate requirements.

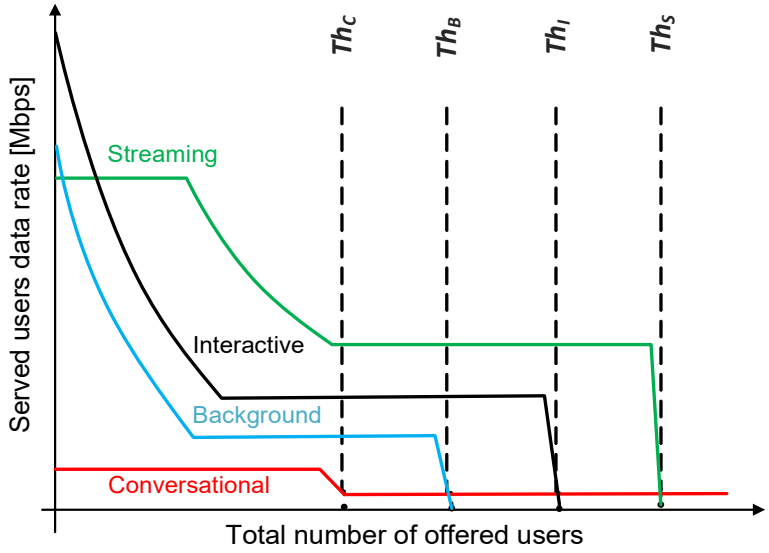


Figure 3.7– General policy of data rate allocation to served users.

As an extension for the proposed VRRM model to handle the extreme cases, when the aggregated capacity is not enough to serve all users, a logical decision would be to start delaying just enough number of users with the lowest priorities, in order to release capacity to serve the rest of the users with the minimum acceptable data rate. Therefore, by increasing the total number of users after Th_C, the algorithm starts delaying Background users, as they have the least priority among all; this process continues until there are no Background users remaining in the system to be delayed at Th_B. The same procedure applies to the Interactive service between Th_B and Th_I, and the Streaming one between Th_I and Th_S, respectively. After Th_S, there are just Conversational users left to be served with the minimum acceptable data rate.

3.3.2 Interaction between CRRM and VRRM

As explained in Section 3.1, VRRM is a centralised entity, responsible for allocating the required capacity of each service defined by the associated VNO, based on the specific policies of that VNO regarding the contracted SLAs and serving priorities. In order to do that, first CRRM should provide VRRM with the information about the maximum available (aggregated) capacity, which is obtainable from all the

available RATs. Then, based on the available capacity, VRRM finds the optimum way of distributing it among different VNOs, to satisfy their requirements.

It is notable that CRRM does not pass the information about the available capacity of individual RATs to VRRM, since it is proposed that VNOs do not own the physical infrastructure, and from their perspective, it is not important which technology provides the contracted data rate. The effect of provided capacity from CRRM to VRRM is expressed in (3.20), the left side of the inequality representing the total demanded capacity of VRRM, which is upper-bounded by the aggregated capacity of CRRM. Therefore, (3.21) can be written to emphasise that the available capacity of VRRM is always a portion of the total network capacity that is provided by CRRM, according to the InPs' policies for RAT selection and load balancing. It is also clear that when users experience a good channel condition, the total aggregated capacity obtained from RRUs is higher, accordingly, on average, the provided data rate (assigned by VRRM) to each service slice in this case will be higher than the one when the channel condition is relatively worse.

$$R^{VRRM}_{[\text{Mbps}]} \leq R^{CRRM}_{[\text{Mbps}]} \tag{3.21}$$

where:

- R^{CRRM}: total capacity obtained by aggregation of RRUs from different RATs.

On the other hand, the total demanded data rate of each service slice, which is obtained by VRRM, should then be provided by suitable RATs in the physical part, according to the specification of each technology and requirements of different services. This way, VRRM also affects the decision of CRRM regarding the RAT selection and load balancing as shown in Figure 3.8.

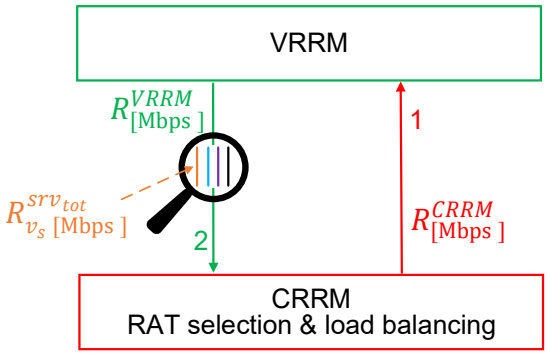


Figure 3.8 – Interaction between CRRM and VRRM.

It is also notable that VRRM does not pass the information about the policies of individual VNOs, such as their serving weights to CRRM. Information about the available aggregated capacity, R^{CRRM}, is exchanged in the first step, provided by CRRM as a decision-making parameter of VRRM, then in the next step, VRRM calculates the total demanded capacity of each service, $R^{srv_{tot}}_{v_s}$, to be sent back to CRRM for the RAT selection process. R^{VRRM}, represents the aggregation of all the calculated demands performed by VRRM.

3.3.3 RAT Selection and Load Balancing Mechanism of CRRM

In order to take advantage of the specifications of different technologies in a wireless Het-Net and achieve multiplexing gain, it is necessary to define a precise cooperation mechanism among RATs, which includes the implementation of efficient CRRM strategies. Among all the established functionalities for CRRM, RAT selection is known as the most essential one [KGJ13]. Different techniques are generally categorised under user- and network-centric approaches: the objectives of the former take user's preference parameters into account, such as lower service prices and higher data rates, whereas the latter considers the desired parameters from the operator's perspective, such as load of each RAT and efficient resource utilisation.

The proposed technique used for RAT selection and load balancing in this work is fundamentally a network-centric approach, since the main objective is to maximise the global utilisation of radio resources. However, as the contracted SLAs are a parameter in the calculation of the demand provided by VRRM, a pre-established user's satisfaction level is also projected as a constraint in the model of CRRM, which is an important parameter from the user's viewpoint.

Furthermore, a service-based policy is also implemented in the RAT selection process, to ensure that the mapping of the various services onto the different access technologies is based on the feasibility and suitability of both characteristics [Hoss09]. As an example, it is not feasible to serve a Video user with GSM, or it is preferable to use Wi-Fi for file sharing as it provides a relatively higher bandwidth compared to the other RATs. Therefore, for a specific service type, a prioritised list of suitable RATs is defined in the model to arrange the available common radio resources [Piao07], which is taken for all services and saved in the rule database, as shown in Figure 3.9.

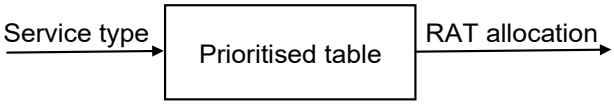


Figure 3.9 – Service-based RAT selection.

To address the assumptions regarding the maximisation of the total utilisation of radio resources, as well as the distribution of the traffic load among existing RATs by considering a service-based policy, the objective of CRRM can be expressed as a weighted concave function, f_{CRRM}, which in this case is chosen to be either a *linear* or a *logarithmic* function, mapping a portion of the available capacity from each RAT onto the most suitable service. While a linear function purely reacts on matching the most suitable RAT to a specific service demand, without considering the distribution of traffic loads of different RATs, a logarithmic utility function has a natural load balancing property to consider both suitability and loading situation with the implementation price of a more complex algorithm,

$$\underset{\mathbf{R^{RAT}}}{Max}\, f_{CRRM}(\mathbf{R^{RAT}}) = \underset{\mathbf{R^{RAT}}}{Max} \sum_{j=1}^{N^{RAT}} \sum_{v_s=1}^{N^{srv}} \gamma_{j,v_s}^{RAT}\, U\left(\frac{R_{j,v_s\,[\mathrm{Mbps}]}^{RAT}}{R_{[\mathrm{Mbps}]}^{VRRM}}\right) \tag{3.22}$$

where:

- f_{CRRM} : the objective function of CRRM for RAT selection and load balancing,
- $U(.)$: choice of utility function, which can be linear or logarithmic,

- N^{RAT}: number of RATs,
- γ_{j,v_s}^{RAT}: assigned weight to RATj, for performing service v_s, to define the priorities and to balance the load among different services and RATs,
- R_{j,v_s}^{RAT}: allocated data rate from RATj to service v_s,
- $\mathbf{R^{RAT}}$: vector of assigned data rates from different RATs,

$$\mathbf{R^{RAT}} = [R_{1,1}^{RAT}, R_{1,2}^{RAT}, \dots, R_{1,N^{srv}}^{RAT}, \dots, R_{N^{RAT},1}^{RAT}, R_{N^{RAT},2}^{RAT}, \dots, R_{N^{RAT},N^{srv}}^{RAT}]^T \quad (3.23)$$

γ_{j,v_s}^{RAT} is composed of two different factors: one to control the load of RAT j (in the case of a logarithmic objective), and another to define the suitability and priority of that RAT to be assigned to service v_s (service s, from VNO v), both parameters being positive integer numbers,

$$\gamma_{j,v_s}^{RAT} = \delta_j^{RAT}\, \mu_{v_s}^{RAT} \quad (3.24)$$

where:

- δ_j^{RAT}: load balancing factor of RAT j,
- $\mu_{v_s}^{RAT}$: assigned weight to service v_s, to define service-to-RAT allocation priorities.

Figure 3.10 shows the mechanism of distributing the traffic load associated with different services, among the existing RATs. When a new request for service is made to the network, if the load of the most preferred cell is less than the threshold, it will be assigned to the user; otherwise, the user will be directed to the next highest priority RAT, which has also enough capacity to serve the user's demand.

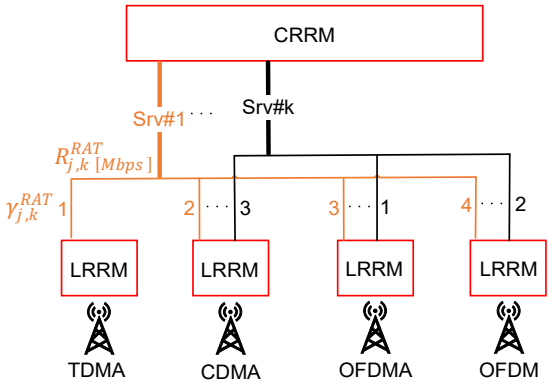


Figure 3.10 – CRRM mechanism of RAT selection and load balancing.

Furthermore, there are some constraints that should be taken into consideration while optimising the objective function. An important constraint is the one that projects the demanded data rate of each service from VRRM and plays an important role in the interaction between CRRM and VRRM. From Figure 3.10, it is noticeable that the distribution of traffic load for each service among the different RATs should sum to the demanded data rate of that specific traffic slice, provided by VRRM. It should not be higher than that, since the extra capacity will be wasted, nor lower, because contracted SLAs cannot be addressed, is expressed as follows:

$$\sum_{j=1}^{N^{RAT}} R_{j,v_s\,[\mathrm{Mbps}]}^{RAT} = R_{v_s\,[\mathrm{Mbps}]}^{srv_{tot}} \quad (3.25)$$

The next constraint is related to the capacity of each RAT: logically, the total demand that is directed to each RAT cannot exceed the total available capacity of that RAT, the constraint being:

$$\sum_{v_s=1}^{N^{srv}} R_{j,v_s\,[\mathrm{Mbps}]}^{RAT} \leq R_{j\,[\mathrm{Mbps}]}^{RAT_{tot}} \tag{3.26}$$

where:

- $R_j^{RAT_{tot}}$: total available capacity of RAT j (a portion of the whole RAT's bandwidth).

The last constraint is just defined for the sake of simulation, to ensure that all services loads requested to RATs are positive values:

$$R_{j,v_s\,[\mathrm{Mbps}]}^{RAT} \geq 0 \tag{3.27}$$

The graphical solution for the linear optimisation problem of (3.22) is shown in Figure 3.11. In order to be able to picture it in a two-dimensional graph, all the resources from the different RATs are categorised as cellular and Wi-Fi, with corresponding capacities of R_C^{RAT} and R_W^{RAT}, respectively. The feasible region, which is filled in with green is the solution space of the problem, including all possible points that satisfy the constraints. The boundaries of feasible region come from the maximum capacity of RATs (horizontal and vertical solid lines in black), as well as the demanded capacity of VRRM (solid line in blue).

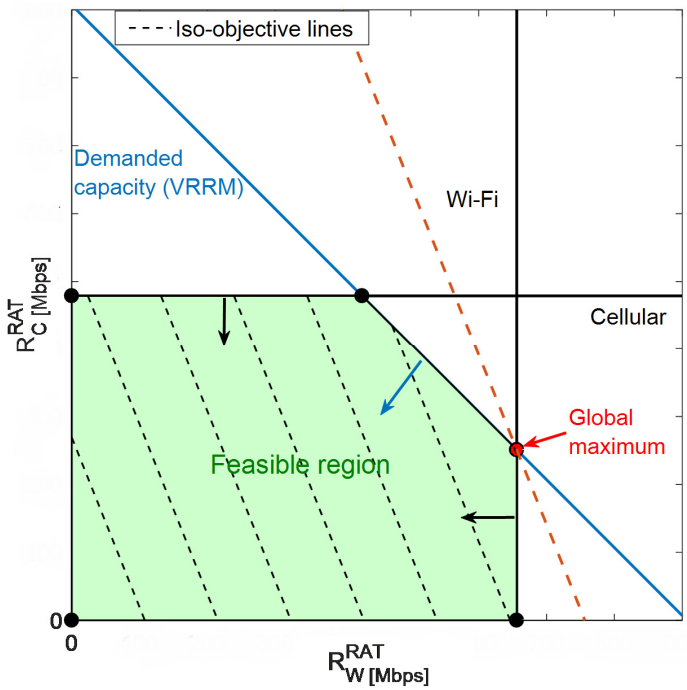


Figure 3.11 – Illustration of the general behaviour of CRRM.

Each of the feasible convex region vertices can be a candidate solution to globally maximise the objective function. Iso-objective lines of the utility function are also represented by dashed lines for the case where the serving weight of Wi-Fi is higher than cellular ones. By moving to the right side of the feasible region, one can see that the values of the objective increases, until it reaches the vertex that maximises the objective, marked as red. The solution is logical and predictable, as it uses the maximum

available capacity of Wi-Fi (preferred RAT according to the serving weight) and the remaining demanded bandwidth is provided by cellular RATs.

3.3.4 Calculating the Aggregated Capacity from Physical RRUs

The achievable data rate of users is highly dependent on SINR, since it has a direct impact on spectral efficiency. Consequently, users with better radio links can be served with higher data rates. In order to reach the optimum performance, efficient data rate adaption techniques are required to change the Modulation and Coding Schemes (MCSs), based on the variation of SINR values. The general concept is to use higher modulation and code rates when the channel condition is good, and alternatively lower ones when the channel state is relatively worse.

It is possible to obtain the data rate that each physical link can carry, as a function of SINR. In [Caei14], there is an approximation for several access technologies to find the data rate from the throughput-versus-SINR curve of each RAT. To perform that, data rate is considered constant in SINR intervals, and each SINR threshold corresponds to the switch points, which are often hard-coded in the transmitter. Inversely, the SINR as a function of data rate can also be presented by a piecewise step function. Using the threshold values provided by [Caei14], the general model to approximate the data rate of a single RRU as a function of SINR can be expressed as follows:

$$R^{RRU}_{j[\mathrm{Mbps}]}(\gamma) = \begin{cases} R^{RRU}_{1[\mathrm{Mbps}]} & \text{if } \gamma_{0\,[\mathrm{dB}]} \le \gamma_{j\,[\mathrm{dB}]} \le \gamma_{1\,[\mathrm{dB}]} \\ R^{RRU}_{2[\mathrm{Mbps}]} & \text{if } \gamma_{1\,[\mathrm{dB}]} \le \gamma_{j\,[\mathrm{dB}]} \le \gamma_{2\,[\mathrm{dB}]} \\ R^{RRU}_{3[\mathrm{Mbps}]} & \text{if } \gamma_{2\,[\mathrm{dB}]} \le \gamma_{j\,[\mathrm{dB}]} \le \gamma_{3\,[\mathrm{dB}]} \\ \dots & \\ R^{RRU}_{n[\mathrm{Mbps}]} & \text{if } \gamma_{n-1\,[\mathrm{dB}]} \le \gamma_{j\,[\mathrm{dB}]} \le \gamma_{n\,[\mathrm{dB}]} \end{cases} \tag{3.28}$$

where:

- γ: SINR value,
- $R_j^{RRU}(\gamma)$: approximated data rate of a single RRU, as a function of SINR, for various intervals.

However, for the sake of simplicity, the discrete function of (3.28) can be estimated by an equivalent continuous fifth-degree polynomial, the fitting being performed by a least-square technique, using MATLAB [WoPh12]. In this work, one needs to express SINR as a function of data rate, therefore, using the inverse function of (3.28) one gets:

$$\gamma^{5th}_{[\mathrm{dB}]}\left(R_j^{RRU}\right) = \sum_{m=0}^{5} a_{m[\frac{\mathrm{dB}}{\mathrm{Mbps}^m}]} \times \left(R^{RRU}_{j[\mathrm{Mbps}]}\right)^m \tag{3.29}$$

where:

- a_m: coefficients of the polynomial fit, obtained from Annex A tables of [Khat16].

The inverse of the piecewise function in (3.28) and the fifth-degree polynomial fit of (3.29) are represented in Figure 3.12 for a single LTE RRU, when SINR is larger than 1 dB for a channel.

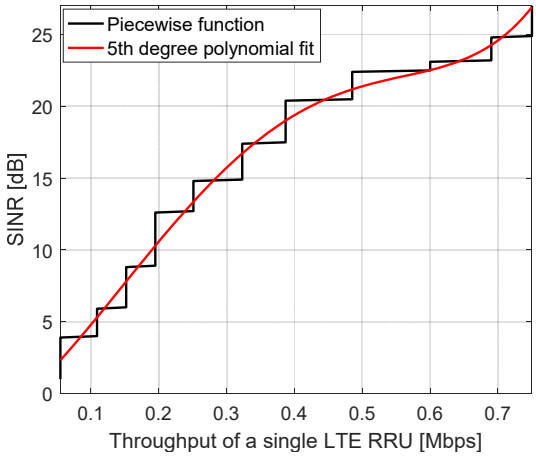


Figure 3.12 – 5th-degree polynomial fit on the piecewise function of SINR.

The PDF of a single RRU's data rate, assuming that this data rate is bounded between lower and higher values, can be expressed as follows (the analytical details are given in Annex A):

$$p_{\mathrm{R[Mbps]}}\left(R_j^{RRU}\middle|R_j^{RRU_L}\le R_j^{RRU}\le R_j^{RRU_H}\right) = \frac{\frac{0.2}{\alpha_p}\left(\sum_{m=1}^{5} m a_m \left(R_{j[\mathrm{Mbps}]}^{RRU}\right)^{m-1}\right) e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU}\right)^m}}{e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU_L}\right)^m} - e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU_H}\right)^m}} \tag{3.30}$$

The goal of deriving the PDF of data rate for each RRU is to find the total capacity of each RAT, as well as the aggregated network capacity, being provided from CRRM to VRRM, therefore, the next step is to obtain the distribution functions associated with the total capacity of each RAT. Assuming that RAT j can assign a N_j^{RRU} number of RRUs to connected users and that all channels are independent, the data rates of all RRUs, which are random variables, are independent as well, therefore, the PDF of the accumulated data rate of each RAT can be expressed as the convolution of all its RRU's PDFs [Khat16]:

$$p_{R[Mbps]}\left(R_j^{RAT_{tot}}\right) = p_{R[Mbps]}\left(R_j^{RRU_1}\right) * p_{R[Mbps]}\left(R_j^{RRU_2}\right) * \cdots * p_{R[Mbps]}\left(R_j^{RRU_{N_j^{RRU}}}\right) \tag{3.31}$$

According to the central limit theorem [PaPi02], the convolution PDF of statistically independent random variables tends to a Gaussian Distribution, as the number of random variables tends to infinity, regardless of the initial PDFs. Since the random variables in this work are RRUs, which are proposed to vary between a minimum and a maximum value for all RATs, (3.31) is approximated by fitting a Truncated Gaussian distribution for each RAT. Suppose that $R_j^{RAT_{tot}}$ has a Gaussian Distribution, $R_j^{RAT_{tot}} \sim N\left(\overline{R_{b_j}}, \sigma_j\right)$, and that R_j^{RRU} lies in the interval $R_j^{RRU} \epsilon \left[R_j^{RRU_H}/2,\ R_j^{RRU_H}\right]$, then, the distribution parameters associated with each RAT within 95% confidence bounds are given in Table 3.1, where $R_j^{RAT_{tot}} \epsilon \left[R_{b_j}^{min}, R_{b_j}^{max}\right]$.

As an example, Figure 3.13 represents the PDF of the aggregated capacity of an LTE BS along with the approximated Truncated Gaussian fit; the fittings are well matched with the convolution PDFs. By sampling from Figure 3.13 over time, Figure 3.14 is generated; the mean values of capacity for different path loss exponents are almost the same as the medians in Figure 3.13.

Table 3.1 – Parameters of the Truncated Gaussian Distributions.

RATs	$R_{b_j[Mbps]}^{min}$	$R_{b_j[Mbps]}^{max}$	$\overline{R_{b_j[Mbps]}}$	$\sigma_{j[Mbps]}$
OFDM (Wi-Fi)	3352	6704	5116	67.1
OFDMA (LTE)	2400	4800	3655	61.2
CDMA (UMTS)	44.1	88.2	66.2	1.63
TDMA (GSM)	0.62	1.24	0.94	0.053

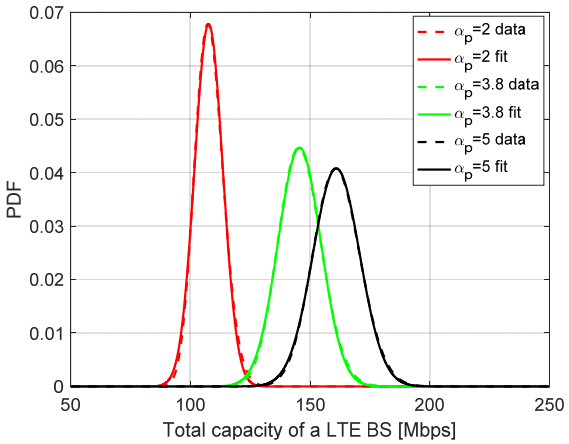


Figure 3.13 – PDF of the total capacity that LTE BS provides.

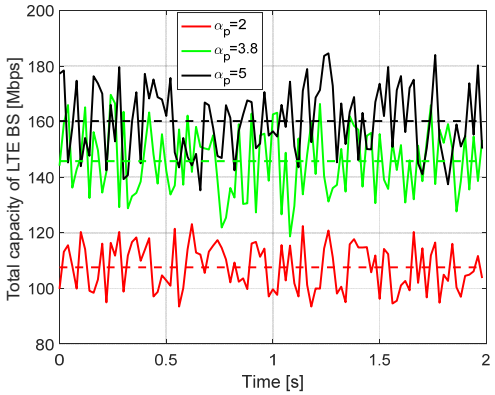


Figure 3.14 – Sampled capacity from the PDF of LTE BS.

As already discussed, the serving data rate of each RRU is in direct relation with SINR. More RRUs are needed to be assigned to the users with low SINR to provide the same data rate as the ones with higher SINR, therefore, the system throughput decreases if users suffer from a bad channel condition. Furthermore, the provided capacity has a direct impact on the performance of VRRM in terms of satisfying SLAs: if SINR is low, the capacity available to VRRM will also be low and consequently the data rate provided to VNOs will be closer to the lowest contracted level, or even in an extreme case it cannot be satisfied.

Figure 3.15 summarises the steps taken in this section to calculate the aggregated capacity of each RAT, required for RAT selection and load balancing model of CRRM. The physical connection of each user is basically comprised of four blocks. In the first step, the relation between the serving data rate of each RRU from different RATs and the SINR of users are approximated by a polynomial fit, based on the different MCS levels of each technology. Then, the PDF of the serving rate for a single RRU is calculated and the total aggregated PDF for each RAT is approximated by fitting a Truncated Gaussian distribution. Sampling from this distribution over time provides the available capacity of each RAT, as well as the total system capacity.

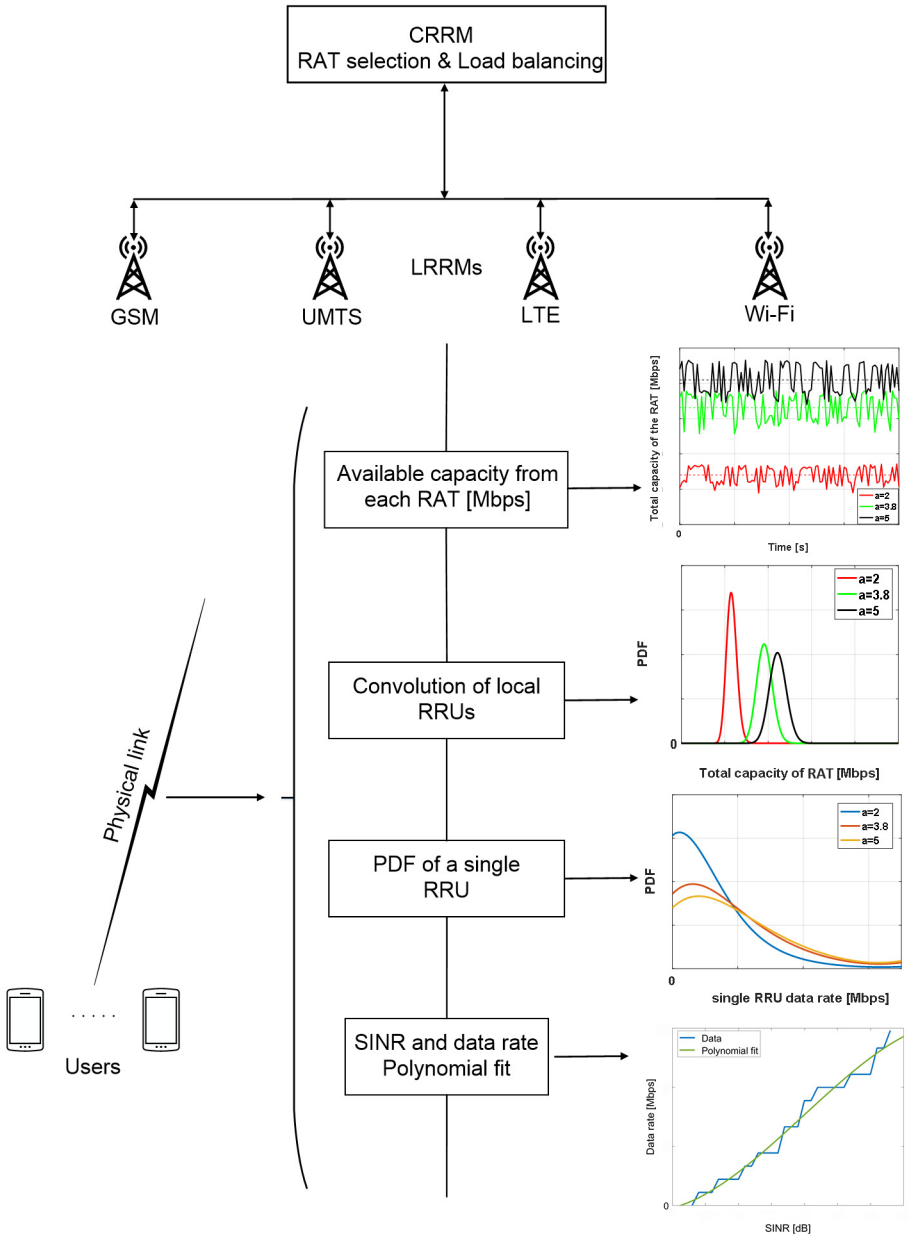
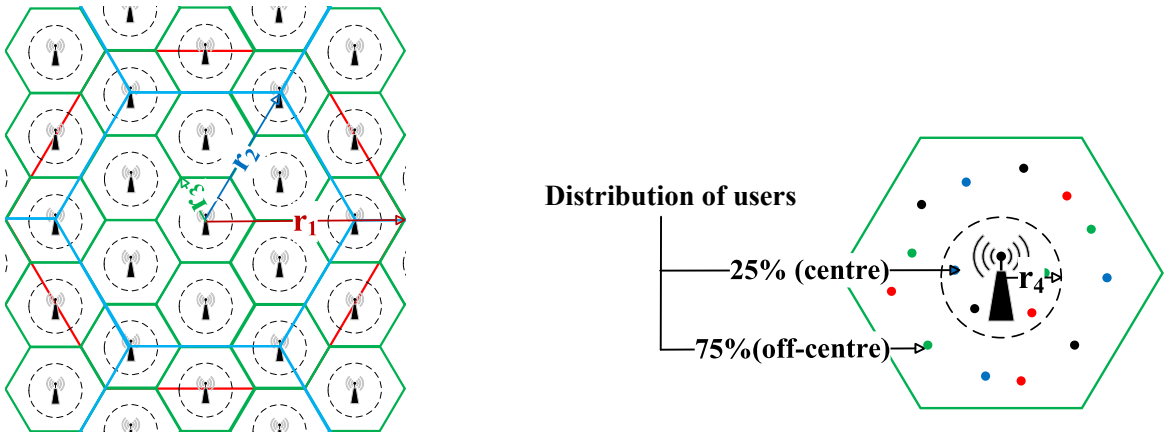


Figure 3.15 – Taken steps to define the available capacity of the system.

3.4 Canonical Scenario

In order to evaluate the model, a canonical scenario with some assumptions was considered as follows:

- The area under analysis is uniformly covered by all existing cellular RATs, namely GSM, UMTS and LTE; furthermore, a Wi-Fi AP is placed at the centre of each LTE cell site to boost capacity. An area of 1 km^2 is used for simulation purposes.
- All RRUs from the available RATs can be aggregated and virtualised into isolated slices.
- Regarding the users distribution, it is assumed that 25% of those inside each LTE cell have access to Wi-Fi, while the remaining 75% are located within the LTE off-centre areas of Figure 3.16, the reason being the fact that users tend to gather around the areas where there is Wi-Fi coverage to enjoy a higher data rate, as well as a cheaper price.
- Users' terminals can support all available RATs, and each user performs one specific service during simulation.
- The aggregated data rate available from CRRM to VRRM is assumed to be 80% of the maximum available capacity, which can be provided by all RATs.



(a) Coverage of existing RATs. (b) Distribution of users.

Figure 3.16 – Network layout for the canonical scenario (r_1=1.6 km, r_2=1.2 km, r_3=0.4 km, r_4=0.05 km).

Concerning the RATs for the reference scenario, it is assumed that the deployed Remote Radio Heads (RRHs) are offering full coverage for the cellular networks, supporting OFDMA (based on LTE), CDMA (based on UMTS), and FDMA/TDMA (based on GSM). The proposed cellular layout for the reference scenario is shown in Figure 3.16, with the following characteristics:

- OFDMA is the smallest cell, working at 2.6 GHz, with a deployed 4×4 MIMO antenna and a coverage radius of 0.4 km, being based on a 20 MHz LTE bandwidth, with 100 RRUs, which can be assigned to traffic bearers [DaPS11].
- UMTS (HSPA+) works at 2.1 GHz, the cell radius being 1.2 km, with 3 carriers, each one with 16 codes; only 45 out of all 48 codes are available for assignment to users [SBT11].
- The TDMA cell, based on GSM EDGE is the biggest one, with a radius of 1.6 km, and 3 carriers, each one with 8 time-slots, [Saut10]; 21 out of 24 time-slots are available to be assigned to users.

Along with cellular networks, Wi-Fi (OFDM) coverage is provided by means of the IEEE802.11ac standard APs, configured with a 80 MHz channel bandwidth. It is assumed that the AP covers a cell with a radius of 50 m, and it is able to support up to 4 spatial Streaming channels, at 40 MHz [BeKM13], [Cisc16]. A summary of the characteristics of the different RATs is presented in Table 3.2.

Table 3.2 – Summary of BS characteristics from different RATs (extracted from [Cisc16]).

RAT	Number of BSs/APs	Cell Radius [km]	RRU	Number of RRUs/BS	Data Rate/RRU [Mbps]	Data Rate/BS [Mbps]	Total Data Rate [Mbps]
OFDM (Wi-Fi)	16	0.05	Carrier	432 (108 sub-carriers × 4 SS)	0.97	336 (420×0.8)	5 376
OFDMA (LTE)	16	0.4	Resource Block	400 (100 Resource blocks × 4 SS)	0.75	240 (300×0.8)	3 840
CDMA (UMTS)	~ 1.7	1.2	Code	45 (3 carriers × 15 codes)	1.4	50.4 (63×0.8)	85.7
TDMA (GSM)	1	1.6	Time-slot	21 (3 carriers × 7 time-slots)	0.059	1 (1.24×0.8)	1

Regarding data rate allocation, the following assumptions are made:

- A single VNO with the capability of serving 4 different services, each one from a different class of service is considered: Voice (Conversational), Video (Streaming), Web (Interactive) and Email (Background).
- Minimum and maximum thresholds for assigning the demanded data rates to users are defined in the SLAs. Voice and Video, with higher priorities, are supposed to be of the GB type, while the other two are categorised under the BG one; for Web and Email, the data rate upper bound is different in the centre and off-centre of LTE cells. Since Wi-Fi is just accessible in the cell centre, the total available data rate in that part is higher compared to the off-centre one. The characteristics of service parameters are summarised in Table 3.3.

Table 3.3 – Assumptions for service parameters.

Class	Service	Range of data rate variation [Mbps]		User mix [%]	SLA
		Centre	Off-Centre		
Conversational	Voice	[0.032, 0.064]		20	GB
Streaming	Video	[2, 13]		50	GB
Interactive	Web	[1, 907]	[1, 383]	20	BG
Background	Email	[0.1, 907]	[0.1, 383]	10	BG

Also concerning the assumptions for allocating the most suitable RATs to different services, the prioritised table of RAT selection for some of the most common services is presented in Table 3.4, while the objective function is supposed to be linear. For each service, the RAT that is numbered as 1 is the

most preferred for that service, NA representing the case where it is not feasible to associate a particular service to a specific RAT.

Table 3.4 – Prioritised table of RAT selection.

Services	Priority of RATs			
	TDMA (GSM)	CDMA (UMTS)	OFDMA (LTE)	OFDM (Wi-Fi)
Voice (Conversational)	1	2	3	4
Video (Streaming)	NA	3	1	2
Web (Interactive)	4	3	2	1
Email (Background)	4	3	2	1

3.5 Assessment of the Model Implementation

3.5.1 Evaluation Metrics for VRRM and CRRM

In order to evaluate the performance of the proposed VRRM model, several evaluation metrics can be used, since the definition of a good performance from network and users' viewpoints can be different. For instance, VRRM takes care of the global network performance, therefore, maximising the overall resource usage is of great importance for this purpose, while, on the other hand, VNO's subscribers just put value on the service level that is provided to them, therefore, the evaluation metrics for network and users are separated.

The metrics used for network performance evaluation are as follows:

- **Percentage of total assigned data rate –** one of the most important metrics, showing the total network throughput in terms of data rate allocation, the values closer to 100% obviously leading to a better VRRM performance:

$$p^{tot}_{VRRM[\%]} = 100\ \frac{\sum_{v_s=1}^{N^{srv}} \sum_{i=1}^{N^{usr}_{v_s}} w^{usr}_{v_s,i} R^{srv_{max}}_{v_s\,[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}} \tag{3.32}$$

- **VRRM capacity share –** the percentage of capacity allocated to each VNO, out of the total available VRRM one, being a key performance metric from VNOs and VRRM perspective:

$$R^{VNO_v}_{VRRM[\%]} = 100\ \frac{\sum_{v_s=1}^{N^{srv}_{VNO_v}} \sum_{i=1}^{N^{usr}_{v_s}} w^{usr}_{v_s,i} R^{srv_{max}}_{v_s\,[\text{Mbps}]}}{R^{VRRM}_{[\text{Mbps}]}}, \qquad \forall\, v \in \{1, 2, \dots, N^{VNO}\} \tag{3.33}$$

- **Total data rate of each service –** it shows the total data rate assigned to each service slice of a VNO, being important from both VRRM and VNOs' viewpoints:

$$R_{v_s\,[\mathrm{Mbps}]}^{srv_{tot}} = \sum_{i=1}^{N_k^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\mathrm{Mbps}]}^{srv_{max}}, \quad \forall\, v_s \in \{1, 2, \dots, N^{srv}\} \tag{3.34}$$

- **Percentage of served users –** the percentage of served users performing a specific service to the total number of offered ones, which is an essential metric for VNOs:

$$p_{v_s[\%]}^{usr_{net}} = 100\, \frac{N_{v_s}^{usr}}{N^{usr_{net}}}, \quad \forall\, v_s \in \{1, 2, \dots, N^{srv}\} \tag{3.35}$$

where:

- $N^{usr_{net}}$: total number of offered users in the network.

- **Cumulative percentage of served users –** it shows the cumulative ratio (in percentage) of served users over the first v_s services, over the total number of offered ones, which is another important metric from VNOs' viewpoint:

$$P_{v_s[\%]}^{usr_{net}} = \begin{cases} p_{v_s[\%]}^{usr_{net}} & if \quad v_s = 1 \\ \sum_{i=1}^{v_s-1} p_{i[\%]}^{usr_{net}} + p_{v_s[\%]}^{usr_{net}} & if \;\; v_s \in \{2, \dots, N^{srv}\} \end{cases} \tag{3.36}$$

- **Percentage of served users in each service –** it is defined as the ratio of served users in each service to the total number of offered ones in that service:

$$p_{v_s[\%]}^{usr_{srv}} = 100\, \frac{N_{v_s}^{usr}}{N_{v_s}^{usr_{tot}}}, \quad \forall\, v_s \in \{1, 2, \dots, N^{srv}\} \tag{3.37}$$

where:

- $N_{v_s}^{usr_{tot}}$: the total number of offered users in service v_s.

- **Proportional fairness index –** it is defined as the normalised deviation of allocated data rate to service v_s, from its proportional fair value according to the ratio of service weights

$$I_{pf_{v_s}} = \frac{\left| R_{v_s}^{srv} - R_{v_s}^{Srv_{pf}} \right|}{R^{VRRM}} \tag{3.38}$$

where:

- $R_{v_s}^{srv}$: allocated data rate to service v_s,
- $R_{v_s}^{Srv_{pf}}$: proportional fair value of the data rate for service v_s.

The metrics that are important from a user's viewpoint are as follows:

- **Data rate of each user –** the data rate allocated to a user is an important QoS metric from both users' and VNOs' viewpoints, having a direct impact on the satisfaction level of the served users:

$$R_{v_s,i\,[\mathrm{Mbps}]}^{usr} = w_{v_s,i}^{usr} R_{v_s\,[\mathrm{Mbps}]}^{srv_{max}}, \quad \forall\, v_s \in \{1, 2, \dots, N^{srv}\}, \quad \forall i \in \{1, 2, \dots, N_{v_s}^{usr}\} \tag{3.39}$$

- **Users' satisfaction ratio –** it is an important metric from VNOs' perspective, the definition depending on the different service classes. For Conversational and Streaming, it is expressed in (3.40), i.e., the ratio between user's average long-term data rate and maximum achievable

data rate of that specific service. For Interactive, this parameter is given in (3.41) together with an illustration in Figure 3.17: if the assigned data rate is higher than the minimum required one, R_{min}^{usr}, the satisfaction is linearly increased up to a certain threshold, called Th_{Int}^{Sat}, and from this point onward, the user will be fully satisfied. The satisfaction ratio is not defined for Background, given its characteristics.

$$S_{v_s,i}^{usr} = \frac{R_{v_s,i\,[\text{Mbps}]}^{usr}}{R_{v_s\,[\text{Mbps}]}^{srv_{max}}}, \quad \forall\, v_s \in \{1,2,\dots,N^{srv}\}, \quad \forall i \in \{1,2,\dots,N_{v_s}^{usr}\} \tag{3.40}$$

$$S_{v_s,i}^{usr_{Int}} = \begin{cases} 0 & if\ \ R_{v_s,i\,[\text{Mbps}]}^{usr} \le R_{min\,[\text{Mbps}]}^{usr} \\ \dfrac{R_{v_s,i\,[\text{Mbps}]}^{usr}}{Th_{Int\,[\text{Mbps}]}^{Sat} - R_{\text{min}[\text{Mbps}]}^{usr}} & if\ R_{min\,[\text{Mbps}]}^{usr} < R_{v_s,i\,[\text{Mbps}]}^{usr} < Th_{Int[\text{Mbps}]}^{Sat} \\ 1 & if\ \ R_{v_s,i\,[\text{Mbps}]}^{usr} \ge Th_{Int[\text{Mbps}]}^{Sat} \end{cases}, \tag{3.41}$$

$$\forall\, v_s \in \{1,2,\dots,N_{Int}^{srv}\}, \quad \forall i \in \{1,2,\dots,N_{v_s}^{usr}\}$$

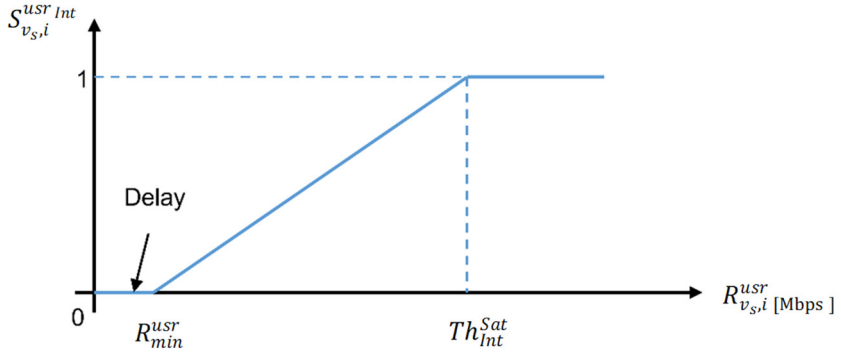


Figure 3.17 – Interactive user's satisfaction ratio.

To evaluate the performance of the CRRM algorithm, some evaluation metrics have been defined, which are important from InPs' viewpoint, as VNOs are not dealing with physical infrastructures:

- **Assigned data rate to each service from different RATs –** an important metric to assess the performance of CRRM in the contribution of each RAT in data rate allocation to a specific service, represented in the form of a set, being clear that the summation of all the values in the set should be exactly equal to the demanded data rate of that service requested by VRRM:

$$\mathbf{R}_{v_s}^{\mathbf{srv_{RAT}}} = [R_{1,v_s[\text{Mbps}]}^{RAT}, R_{2,v_s[\text{Mbps}]}^{RAT}, \dots, R_{N^{RAT},v_s[\text{Mbps}]}^{RAT}], \quad \forall\, v_s \in \{1,2,\dots,N^{srv}\} \tag{3.42}$$

- **RATs' load share –** another important parameter being the percentage of assigned capacity from each RAT out of the total available capacity of that RAT:

$$R_{j[\text{Mbps}]}^{RAT} = 100\ \frac{\sum_{v_s=1}^{N^{srv}} R_{j,v_s\,[\text{Mbps}]}^{RAT}}{R_{j\,[\text{Mbps}]}^{RAT_{tot}}}, \qquad \forall\, j \in \{1,2,\dots,N^{RAT}\} \tag{3.43}$$

3.5.2 Assessment of the VRRM Model

In order to evaluate the performance of the proposed VRRM model, the sensitivity of the algorithm to the variation of the tuning weights was studied in the first step. As a numerical example to assess the

effect of service weights, one took the variation of the serving weight of Conversational services from the lowest possible value according to the proposed framework for service weight assignment (i.e., 4), to a maximum value of 100, while keeping the same ratio of the weights among the other services, thus, the vectors of tuning weights used for testing the VRRM performance can be expressed as follows:

$$\boldsymbol{\lambda} = [\lambda_{con}, \lambda_{str}, \lambda_{int}, \lambda_{bkg}] = [4n, 3n, 2n, n], \qquad n \in \{1, 5, 10, 15, 20, 25\} \tag{3.44}$$

where:

- λ_{srv}, $srv \in \{con, str, int, bkg\}$: tuning weights of Conversational, Streaming, Interactive and Background services, respectively.

The initial results for test scenarios with a single user from each service are presented in Figure 3.18. Solid lines represent the case when the users are gathered around the AP in the centre of the LTE cell, while the dashed lines are associated with the scenario when the users are located in the off-centre areas. As the objective is to maximise the resource utilisation, and the available VRRM capacity in both cases is enough to serve the VoIP and Video users with the maximum achievable data rate defined in the SLAs, these two users are served with the highest data rate, i.e., 64 kbps and 13 Mbps respectively, independent from the variation of the serving weights.

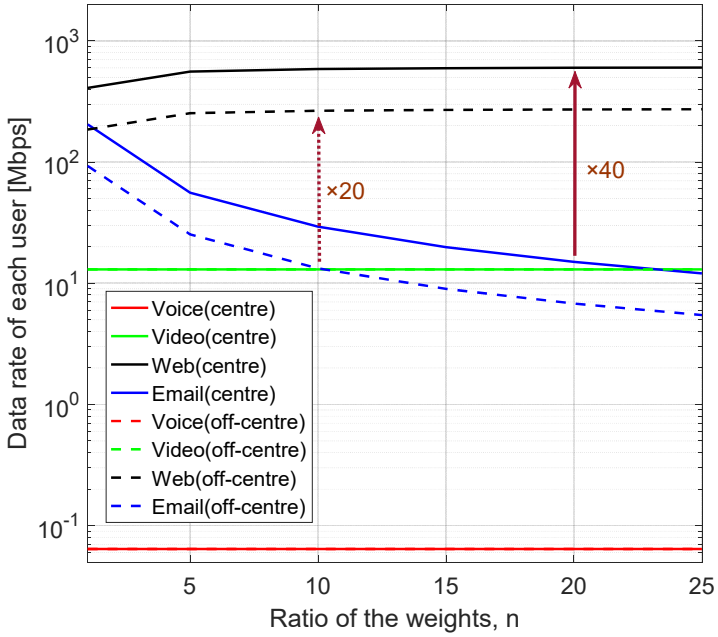


Figure 3.18 – Effect of changing serving weight on the data rate of single users.

For the Web and Email users the situation is different. As these services are categorised under BG SLAs, the rest of the available VRRM capacity is shared between the two in both centre and off-centre areas. By increasing the ratio of the service weights, it is apparent that the data rate of Email drops, but the decreased amount of data rate is added to the Web user, in order to exploit the whole available capacity. The ratio of the assigned data rates to Web and Email users matches exactly the ratio of service weights between these two services. As an example, one can see that when *n*=10, the ratio of the assigned data rate for Web is 20 times higher than for Email in both scenarios. One can also notice that for each set of weights, the sum of data rates is equivalent to the total available VRRM capacity,

showing that the aggregated capacity provided by CRRM is used independently of the variation of service weights and configurations for different scenarios.

As a numerical example, suppose that the vector of service weights is $\boldsymbol{\lambda} = [50, 30, 6, 1]$, Figure 3.19 shows the numerical trend of VRRM data rate assignment, for both centre and off-centre users, which is plotted in Figure 3.7. Data rate values for all the users located in the centre are represented with solid lines and for off-centre ones with dashed lines, being clear that for the same service, the data rate allocated to the users positioned in the centre is always higher than or equal to those located off-centre, since Wi-Fi coverage is only available at the centre.

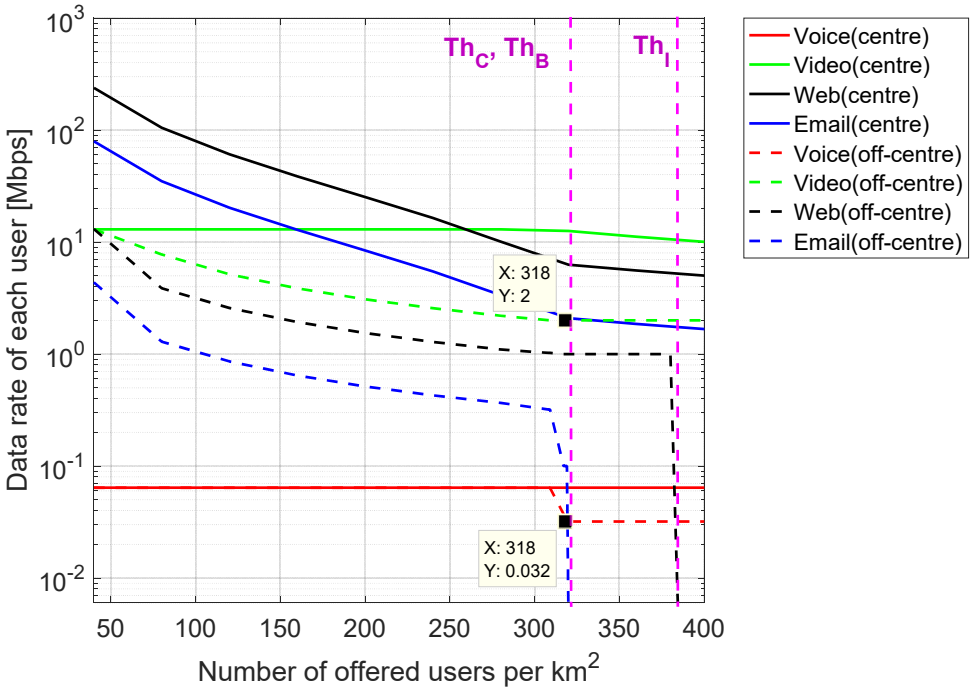


Figure 3.19 – Data rate of each user vs. the number of offered users.

When the number of users is comparatively low, one can see that all GB ones are well served. As traffic load increases, the data rates of off-centre users drop to the lowest contracted level, which were defined in Table 3.3, while all offered users are still served with the minimum data rate at Th_C. It is also noticeable that Voice is the last service to drop from 64 kbps to 32 kbps just before Th_C, since it has the highest priority and the lowest demanded data rate compared to the other services.

The area between Th_C and Th_B is very narrow and the two lines are almost superimposed, because many Background users with the data rate as low as 0.1 Mbps should be delayed to keep the data rate of other services, specially Streaming users, at the lowest possible level, and the same applies to the area between Th_B and Th_I.

Figure 3.20 shows the total data rate assigned to each service slice, as another evaluation metric, being apparent that the different thresholds are compatible with the ones from Figure 3.19. One can notice that when there is no limitation in data rate assignment, in both centre and off-centre areas, the ratio of total allocated data rate to Video, Web and Email is proportional to their service weights, 30, 6 and 1, respectively. To evaluate the total used network capacity, as one of the main parameters, the summation

of values for data rates at each point for both centre and off-centre users is almost equal to the total available capacity of VRRM for 1 km^2. Consequently, independent of the variation of traffic load and of the configuration of several parameters in the proposed scenario, the algorithm of VRRM is capable of distributing almost all the available capacity to address the demanded data rates based on the SLAs.

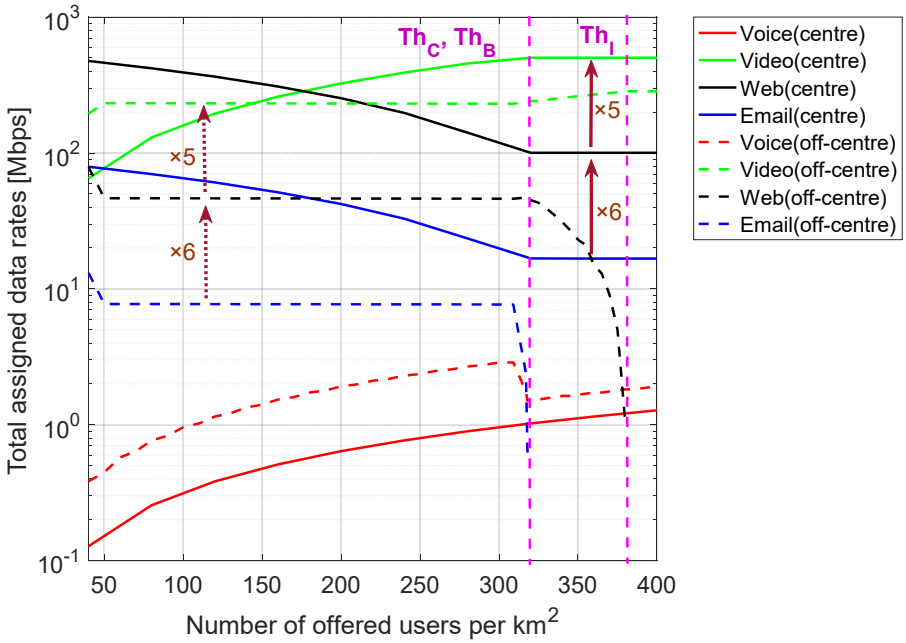


Figure 3.20 – Total data rate of each service vs. the number of offered users.

As another evaluation metric, the percentage of served users is shown in Figure 3.21. Since all users are served before *Th*$_C$, the values for both centre and off-centre ones in this part remains constant and each value is associated with the one defined for the user mix in Table 3.3. The percentage of served users in each service, for the ones located in the centre, does not change by increasing the number of offered users up to 400, as capacity is enough to serve them all.

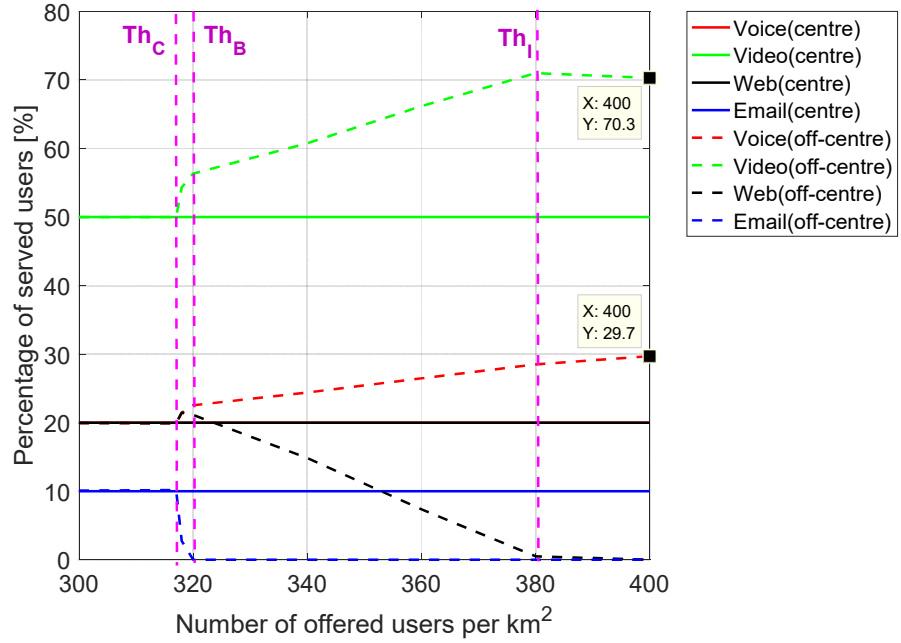


Figure 3.21 – Percentage of served users vs. number of offered users.

However, between Th_C and Th_B, the values for off-centre Email goes to zero, meaning that all the corresponding users are delayed, to release just enough capacity to serve the rest of off-centre users, leading to a slight increase in the values of other services. The same process happens for off-centre Web between Th_B and Th_I. At Th_I, all the available capacity for off-centre is divided between Video and Voice users. At this point, the number of Video users is 2.5 times higher than Voice ones, which is proportional to their ratio of traffic share, 50% and 20% for Video and Voice, respectively.

The cumulative percentage of off-centre served users is shown in Figure 3.22. One can notice that the width of the coloured areas is exactly compatible with the values from Figure 3.21. As the number of offered users increases after Th_C, the areas related to BG services shrink and disappear at Th_I, and GB ones expand. After Th_I, as the algorithm starts to delay Video users, the associated green area starts to shrink as well, while the red one related to Voice increases continuously.

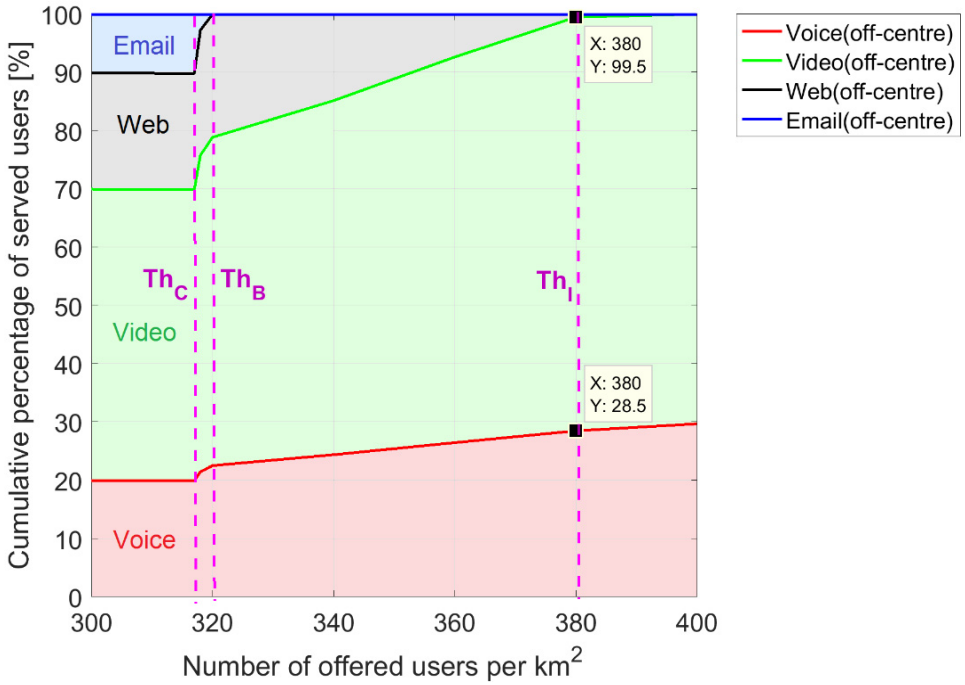


Figure 3.22 – Cumulative percentage of off-centre served users vs. number of offered users

The percentage of served users in each service, for both centre and off-centre areas, is presented in Figure 3.23. Centre users are 100% served, which is also the case for off-centre ones before Th_C. However, by increasing the traffic load after this point, the number of off-centre served users starts to decrease and reaches zero for BG services respectively, according to their priority.

Figure 3.24 presents the users' satisfaction ratio of GB services. Centre users are much more satisfied than off-centre ones, since Voice users in the centre are always served with the maximum achievable data rate, i.e., 64 kbps, and Video users are almost always served with more than 80% of the maximum date rate of 13 Mbps, while the experienced data rate of off-centred Video users is constantly decreasing and reaches to minimum acceptable level after Th_C, which is also the case for off-centre Voice, since they are served with the minimum acceptable satisfaction ratio of 0.5 after this point.

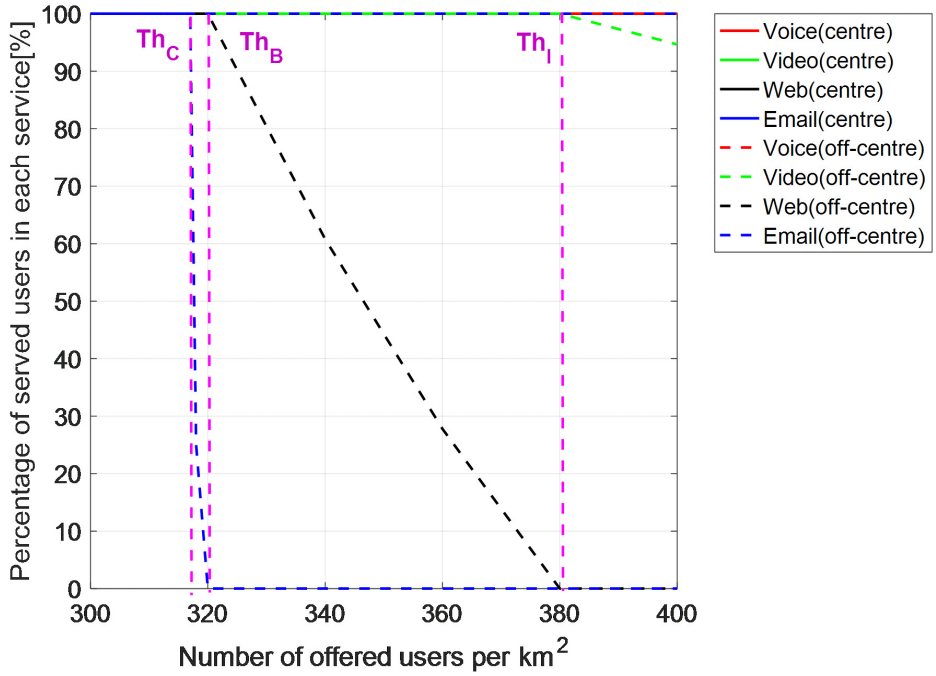


Figure 3.23 – Percentage of served users in each service vs. number of offered users.

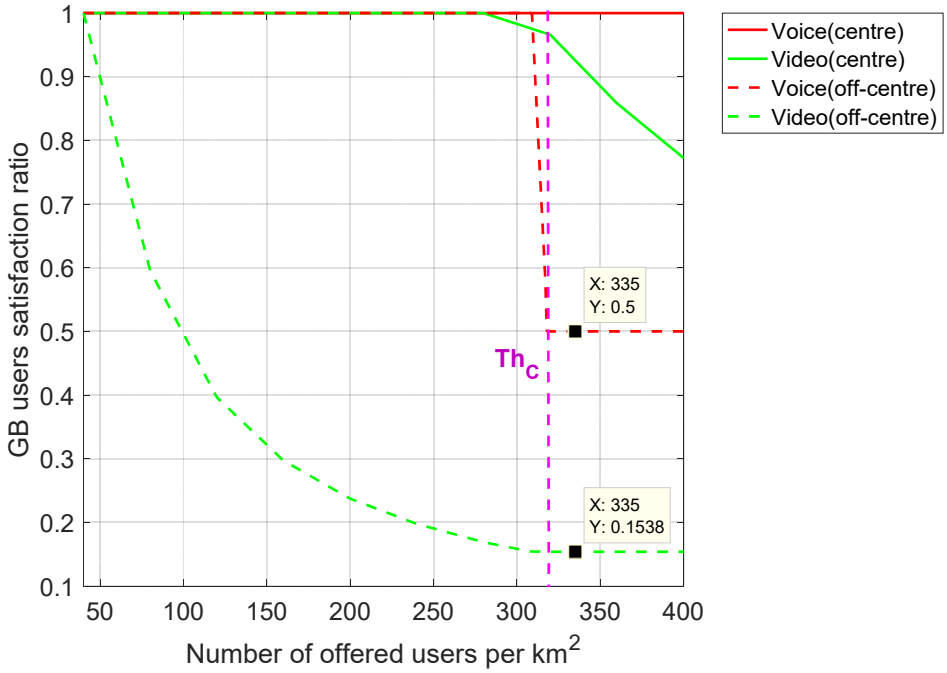


Figure 3.24 – Users' satisfaction ratio for GB services vs. number of offered users.

3.5.3 Assessment of the CRRM Model

In this section, the performance of the CRRM model, as well as the interaction between CRRM and VRRM, are evaluated under the assumptions of the canonical scenario. According to the first evaluation metric, the allocated data rate to Voice traffic by the associated suitable RATs, is shown in Figure 3.25. One can see that the demanded data rate of Voice is provided by GSM and UMTS for both centre and off-centre users; as GSM is more suitable to serve Voice, the CRRM algorithm gives the higher priority

to this RAT, however, the demanded data rate for Voice is higher than the total available GSM capacity, therefore, the rest of the Voice traffic is served by UMTS, which is the second highest priority.

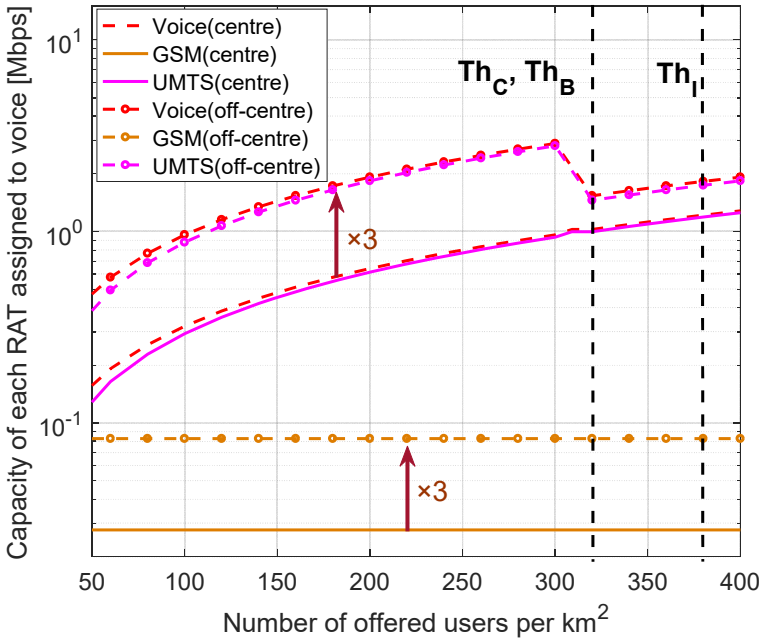


Figure 3.25 – Assigned data rate to Voice from different RATs vs. number of offered users.

One can also notice that before Th_c, when there is no limitation in data rate assignment, the capacity share of off-centred users for both RATs is three times higher than centred ones, which is in accordance with the ratio of traffic (i.e., 75% and 25% accordingly) in these two areas.

The assigned data rate to Video from different RATs is presented in Figure 3.26. Although the number of users in off-centre areas is three times higher than centre ones, the demanded data rate in the centre is mostly higher, except for lower traffic rates, as this group of users has access to Wi-Fi.

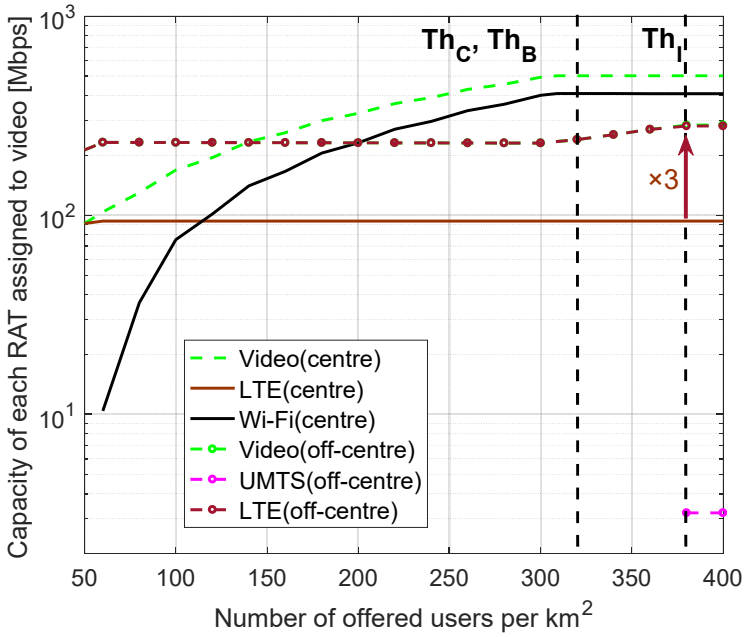


Figure 3.26 – Assigned data rate to Video from different RATs vs. number of offered users.

Therefore, for off-centre users, the available LTE capacity as the most preferred RAT is enough to serve the demanded data rate until Th_I; at this point, all the capacity of LTE is shared between the two groups of Video users, proportional to their traffic mix. After Th_I, off-centre users have to start using capacity from the second available preferred RAT, i.e. UMTS, while at each point for centre users the capacity of LTE and Wi-Fi is sufficient to address the total demand.

For Web users, the allocated capacity from different RATs is presented in Figure 3.27. The whole traffic of centre users can be served just by Wi-Fi as the most preferred RAT. Regarding off-centre users, before Th_B, LTE as the most preferred RAT can cover the demand; but, by increasing the required data rate of off-centre Video users between Th_B and Th_I, there is not enough capacity from LTE to serve the rest of the off-centre traffic. Therefore, the CRRM algorithm has to choose the second suitable RAT, UMTS, besides LTE to address all the demanded capacity of this group of Web users.

Regarding capacity allocation for Email users, the trend is shown in Figure 3.28. The demand of centre users is mostly provided by Wi-Fi, except for a tiny portion that is covered by UMTS. Furthermore, for off-centre users, the whole requested data rate is allocated from LTE and UMTS, which are the first and second preferred RATs, respectively. It is also noticeable that the shares of allocated capacity from UMTS to off-centre and centre users are proportional to their traffic mix.

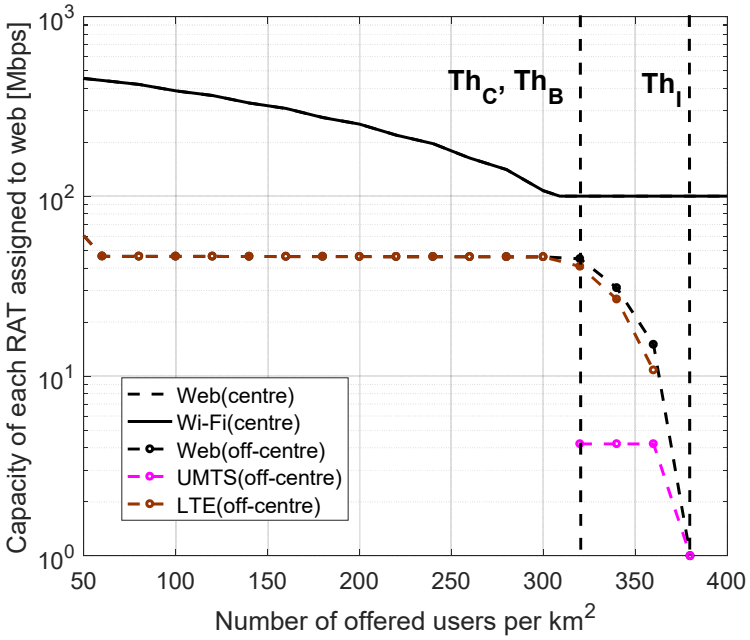


Figure 3.27 – Assigned data rate to Web from different RATs vs. number of offered users.

From another viewpoint, in order to determine which RAT is loaded by which services, a second evaluation metric is used for all available RATs. Starting from GSM, the whole capacity is just dedicated to Voice, as this RAT provides significantly lower data rate compared to the other RATs, and the total demanded Voice data rate as the highest priority service is higher than the total capacity of GSM. Furthermore, the delivered capacity to off-centre users is three times higher than the centre ones, which confirms the results obtained in Figure 3.25.

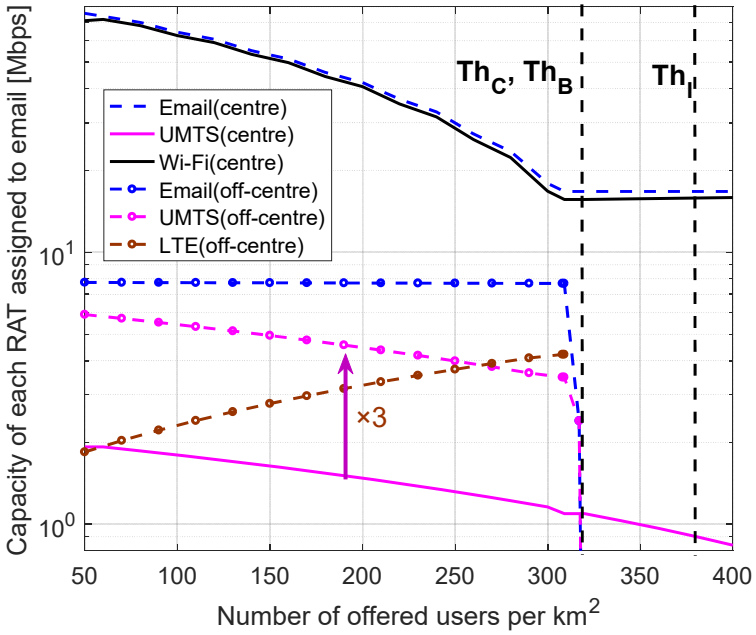


Figure 3.28 – Assigned data rate to Email from different RATs vs. number of offered users.

The percentage of load sharing for UMTS is presented in Figure 3.29. In centre areas at all points just Voice and Email users are being served, which is compatible with the results from Figure 3.25 and Figure 3.28, and this is also the case for off-centre users before Th_B; however, after this point as off-centre Email users are delayed the situation is different. The released capacity of UMTS is allocated to off-centre Web between Th_B and Th_I, which is compatible with the results of Figure 3.27. After Th_I when all off-centre Web users are delayed, the capacity of UMTS has to be shared between the two remaining off-centre services, i.e., Voice and Video to maximise the usage of the available UMTS capacity.

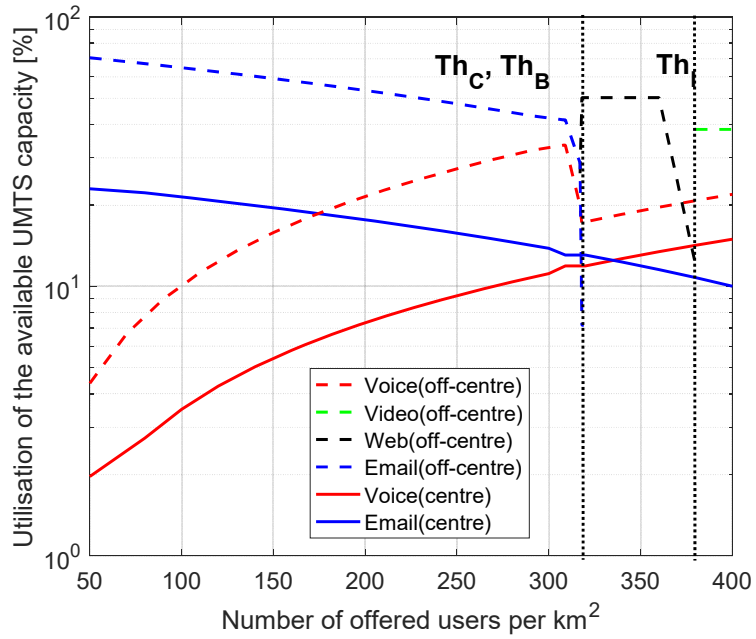


Figure 3.29 – Share of UMTS capacity among different services vs. number of offered users.

Figure 3.30 represents the distribution of LTE capacity among the available services. A significant portion of the available capacity is dedicated to Video, as it is the most suitable RAT for this high priority

service, with the off-centre ones taking more advantage from LTE. This is exactly compatible with the results from Figure 3.26, and the reason is they have not access to Wi-Fi to cover some part of the demand. However, as the LTE capacity for off-centre Video users is more than enough to address the whole demand, Web and Email users in off-centre areas can also benefit from the remaining capacity.

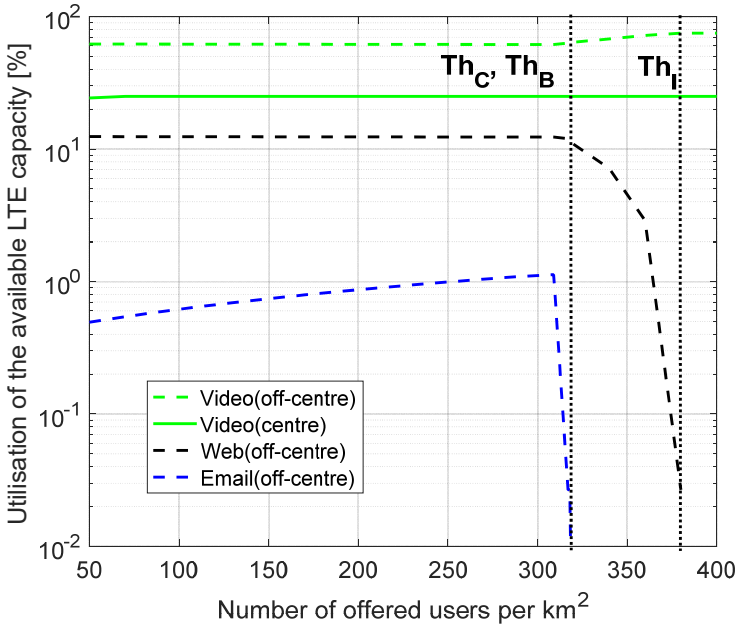


Figure 3.30 – Share of LTE capacity among different services vs. number of offered users.

Regarding the usage of Wi-Fi capacity, the results are shown in Figure 3.31. Just centre users are connected to this RAT and Voice users are excluded, as Wi-Fi is defined to be the least suitable RAT for serving Voice among the other services. It is also noticeable that under low traffic loads, Wi-Fi capacity is mostly shared among BG services, as most of the Video demand has already been covered by LTE. However, this trend is reversed by increasing the traffic load, because the demanded data rate will drop for BG services and increase for Video users.

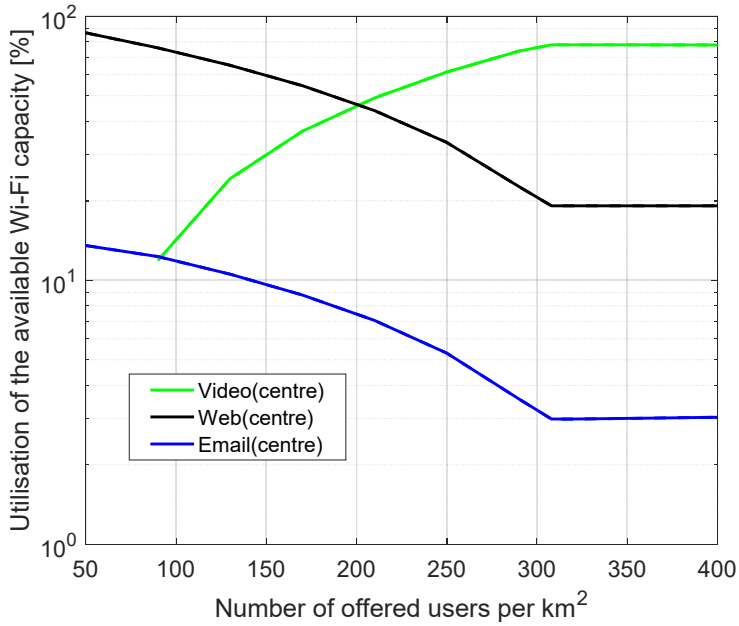


Figure 3.31 – Share of Wi-Fi capacity among different services vs. number of offered users.

According to the proposed assumptions for matching between RATs and services defined in Table 3.4, it is necessary to show if the results are compatible with the assumptions, or not, therefore, a matching table is presented in Table 3.5. Each element of the table can take three different entries to identify if a service is associated with a specific RAT or not, in case the RAT is marked as suitable for the service, or if this is not the case. At most two RATs (from the four available ones) are assigned to a specific service, for both centre and off-centre areas, which are exactly the highest priority ones. It is just notable that for off-centre Web and Email users, as the first priority RAT is Wi-Fi, which is not accessible for off-centre users, they are alternatively connected to LTE and UMTS, as the second and third most suitable RATs respectively.

Table 3.5 – RATs and services matching, before Th_I.

RAT	Voice		Video		Web		Email	
	Centre	Off-centre	Centre	Off-centre	Centre	Off-centre	Centre	Off-centre
GSM	✓	✓	NA	NA	×	×	×	×
UMTS	✓	✓	×	×	×	✓	✓	✓
LTE	×	×	✓	✓	×	✓	×	✓
Wi-Fi	×	NA	✓	NA	✓	NA	✓	NA

Chapter 4

Models Implementation in Simulator

This chapter aims to present the most relevant functional blocks proposed and implemented into the VRRM and CRRM simulators, together with the main assumptions taken and the implementation details. Section 4.1 gives an overview of the simulator in terms of design parameters, main function blocks and output parameters. Section 4.2 briefly introduces the CVX solver as the core for solving the optimisation problems. Then, Section 4.3 explains the algorithms and functionalities of the main process blocks, which were developed in MATLAB and CVX. Finally, the assessment of the simulator in terms of validity and accuracy of results is provided in Section 4.3.

4.1 Simulator Overview

The simulator aims to evaluate different parts of the RAN slicing model, as well as performing the calculations that are proposed in Chapter 3. In general, the structure of the simulator is comprised of three main components: calculation of the aggregated capacity from physical RRUs, service management approach for VRRM, and RAT mechanisms for selection and load balancing of CRRM. To make the calculations, a scenario configuration had to be done, taking the parameters to be varied into account. Therefore, before running the simulator, the input parameters were configured, regarding the scenario under analysis. The configuration of the parameters was done in a MATLAB script, which initialises the variables and runs the simulator. The following parameters can be configured in the simulator as input:

- number and type of available RATs (GSM, UMTS, LTE and Wi-Fi) and number of BSs or APs from each one;
- number of available RRUs in each RAT;
- SINR versus throughput tables related to RRUs of available RATs for different modulation and coding schemes;
- percentage of available capacity from different RATs, according to InPs' policies (70% to 80%);
- minimum and maximum thresholds of data rate associated with RRU from different RATs;
- type of channel conditions based on the scenario (good, bad or undefined);
- access of users to different RATs (cellular, Wi-Fi or both);
- table of suitability weights associated with service-to-RAT assignment, for different services;
- load balancing factor of different RATs to tune the load of system according to InPs' policy;
- number of offered users, which can also be defined as a function of the time of the day (during 24 hours) and the physical location of users (residential or office areas);
- service types (such as Voice, Video and IoT);
- service weights to define the priority of services according to VNOs' strategies;
- service classes (Conversational, Streaming, Interactive and Background);
- users mix of different services;
- the relation between users' satisfaction and assigned data rate as an input function;
- definition of proportional fairness index as an input function;
- customised range of services data rates according to VNOs' policies;
- types of SLA agreements between the VNOs and InPs (GB, BG and BE);
- contracted capacity range of VNO SLAs based on the types of contracted SLAs;
- VNOs' weight to define the priority of VNOs based on their SLA types and InPs' policies;
- percentage of total available capacity provided from CRRM to VRRM.

Regarding the output of the simulator, it comprises the useful parameters for the analysis to evaluate different properties of the model. The parameters obtained in the output are as follows:

- PDF of each RRU from different RATs;
- convolution PDF of all the RRUs from different RATs as a Truncated Normal Distribution;

- total allocated capacity to different service slices (on-demand utilisation of VRRM capacity);
- the capacity share of each VNO, defined by VRRM at a specific network state;
- the data rate assigned to each service slice based on VRRM slice scheduling mechanisms;
- the average long-term data rate of users performing different services;
- number (or percentage) of served (or delayed) users from each service and VNO, calculated by admission control and delay process algorithms;
- users' weights that are performing different services;
- users' satisfaction ratio;
- proportional fairness index of different service slices;
- number and type of active RATs based on the InPs' policies;
- load of each RAT and of the whole network, obtained by CRRM;
- share of capacity from a single RAT to serve various service slices, obtained by CRRM.

Figure 4.1 illustrates the overview of the simulator, in terms of the main algorithms involved, as well as its main functionalities. After the input parameters being configured, they are loaded into the simulator, through the initialisation script. The simulator comprises of three main components: the first one is capacity aggregation unit, which fits a 5-degree continuous polynomial function on the SINR to data rate data of each RAT, based on the least square technique, and then calculates the PDFs of each single RRUs, as well as the resulting convolution PDF of all RRUs' PDFs. The second one is the VRRM module, which is defined inside the CVX solver. In order to ensure that the VRRM optimisation problem is feasible, a mechanism of admission control and delay process is introduced before running the optimisation. The last module is CRRM, which is in charge of distributing the traffic load of demanded services among the available RATs. The CRRM optimisation (solved in CVX) provides the best solution of service-to-RAT mapping, based on InPs policies, such as RAT suitability and energy efficiency.

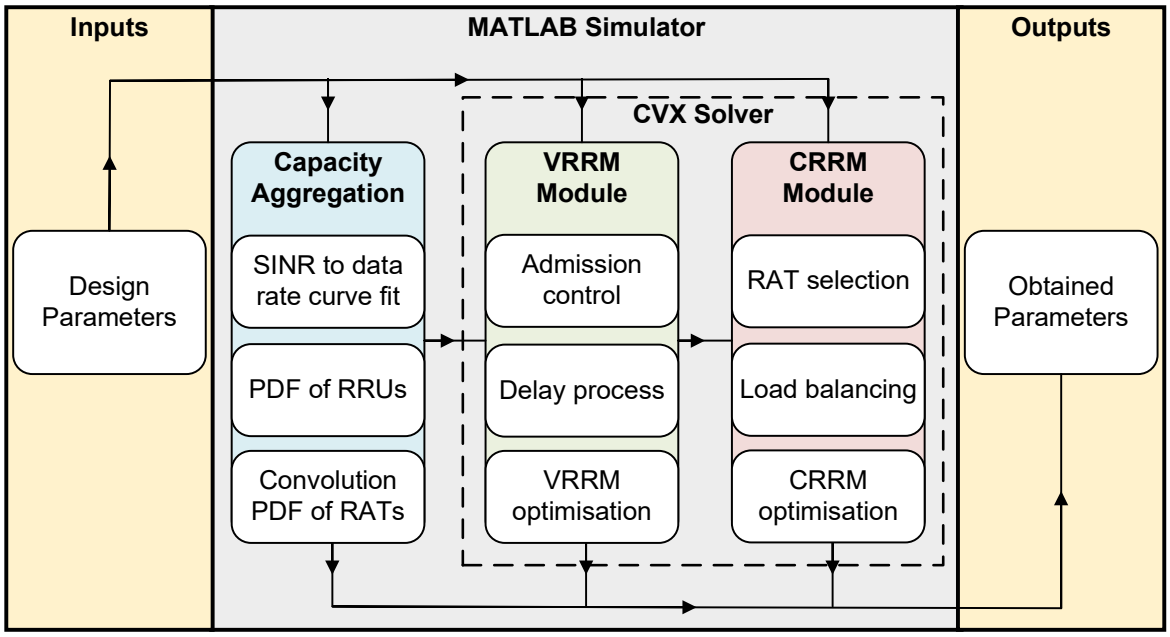


Figure 4.1 – Simulator overview.

4.2 Overview of CVX Solver

The key parts of modelling in this work, including VRRM and CRRM optimisation functions, are defined and solved in CVX, therefore in this section, a brief overview on the performance of this simulator/solver is provided. CVX is a modelling system for constructing and solving disciplined convex programs. CVX supports a number of standard problem types, including linear, quadratic, second-order cone and semidefinite programs. This solver is implemented in MATLAB, effectively turning MATLAB into an optimisation modelling language. Model specifications are constructed using common MATLAB operations and functions, and standard MATLAB code can be freely mixed with these specifications. This combination makes it simple to perform the calculations needed to form optimisation problems, or to process the results obtained from their solution. For example, it is easy to compute optimal trade-off curves by forming and solving a family of optimisation problems in terms of varying the constraints.

Within a CVX specification, optimisation variables have no numerical value; instead, they are special MATLAB objects. This enables MATLAB to distinguish between ordinary commands and CVX objective functions and constraints. As CVX reads a problem specification, it builds an internal representation of the optimisation problem. If it encounters a violation of the rules of disciplined convex programming (such as an invalid use of a composition rule or an invalid constraint), an error message is generated. When MATLAB reaches the *cvx_end* command, it completes the conversion of the CVX specification to a canonical form and calls the underlying core solver to solve it.

If the optimisation is successful, the optimisation variables declared in the CVX specification are converted from objects to ordinary MATLAB numerical values that can be used in any further MATLAB calculations. In addition, CVX also assigns a few other related MATLAB variables. One, for example, gives the status of the problem (i.e., whether an optimal solution was found, or the problem was determined to be infeasible or unbounded). Another gives the optimal value of the problem. Dual variables can also be assigned.

Numerical results of CVX are computed within a predefined precision or tolerance. CVX actually considers *three* different tolerance levels $\epsilon_{\text{solver}} \leq \epsilon_{\text{standard}} \leq \epsilon_{\text{reduced}}$ when solving a model:

- The *solver tolerance* ϵ_{solver} is the level requested of the solver, and the solver will stop as soon as it achieves this level, or until no further progress is possible.
- The *standard tolerance* $\epsilon_{\text{standard}}$ is the level at which CVX considers the model solved to full precision.
- The reduced tolerance $\epsilon_{\text{reduced}}$ is the level at which CVX considers the model "inaccurately" solved, returning a status with the *Inaccurate* prefix; if this tolerance cannot be achieved, CVX returns a status of *Failed*, and the values of the variables should not be considered reliable.

The CVX default values of $[\epsilon_{\text{solver}},\ \epsilon_{\text{standard}},\ \epsilon_{\text{reduced}}]$ are set to $[\epsilon^{1/2},\ \epsilon^{1/2},\ \epsilon^{1/4}]$, where $\epsilon = 2.22 \times 10^{-6}$ is the machine precision. These tolerance levels are chosen in this work, since they are quite sufficient for most of the applications, including the ones proposed in VRRM and CRRM models.

4.3 Simulator Development Procedure

The details of model implementation in MATLAB and CVX are presented in this section. As mentioned before, the two core parts of this thesis related to VRRM and CRRM optimisation were designed, implemented and solved in CVX solver, while the rest of the codes were developed in MATLAB (outside the CVX module). The main parts of the model, containing 3 groups of function blocks, with separated inputs and outputs, categorising the processes associated with *capacity aggregation*, *VRRM* and *CRRM*. The details of each main part, together with the sub-function blocks are presented in Figure 4.2 and explained as follows. The core optimisation problems of VRRM and CRRM are shown in dashed boxes to specify the functionalities that are processed in CVX.

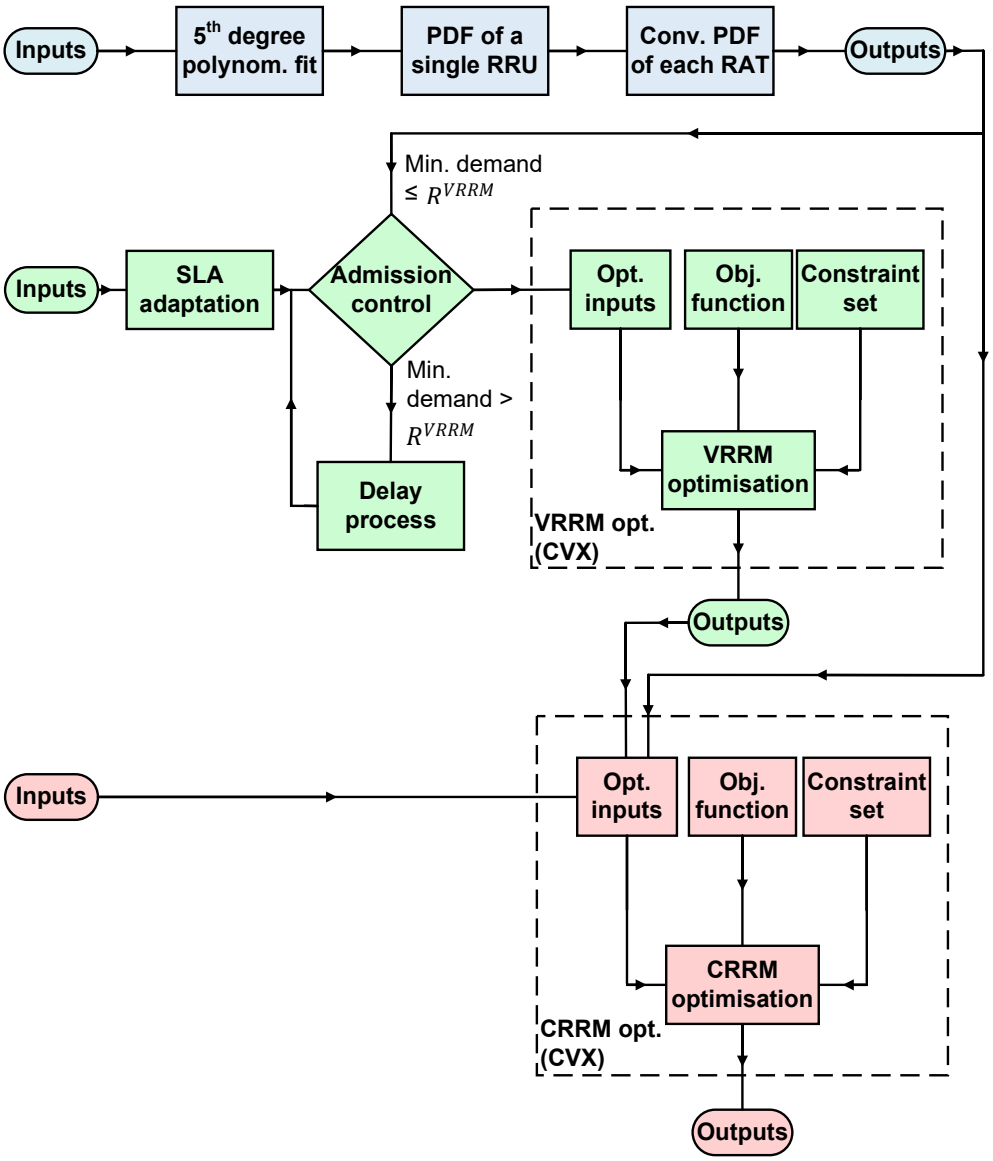


Figure 4.2 – Detailed flowchart of the whole model implementation.

One of the main inputs of VRRM and CRRM is the information about the aggregated capacity that can be obtained from the RRUs of different underlying RATs. This is the first step of the simulator development process that is shown in the top part of Figure 4.2. The inputs of this part are throughput-

versus-SINR tables of single RRUs from all available RATs, which are introduced as piecewise step functions. In the next step, for each RAT, a continuous 5-degree polynomial function is fitted to the resulting step function using the least-square technique in MATLAB. The next process block is related to defining the PDF function of single RRUs from different RATs, which are analytically developed in Section 3.3.4. The following step is to obtain the aggregated capacity of each RAT by convolution of all the RRUs' PDFs of that specific RAT. In the end, by randomly selecting from the convolution PDF of each RAT, the total aggregated capacity of that RAT can be obtained. The information about the individual aggregated capacity of all the RATs is provided as an input to CRRM, since this entity has to control the load of each RAT. Then by summing over all the capacities of different RATs, the total aggregated capacity of the network can also be obtained as another output. Usually, a portion of total network capacity (between 70% and 80%) is provided to VRRM, since this entity is not supposed to deal with the management of the underlying infrastructure.

Regarding the second module, VRRM, the inputs are related to VNOs specifications and service management policies, such as the number of services, serving weights, range of acceptable data rate for each service, priorities and the contracted SLAs, traffic mix and number of offered users. Before solving the core optimisation problem of VRRM, in order to avoid over provisioning in the capacity allocation of different VNOs under low traffic loads, a mechanism called *SLA adaption* is proposed in the first place. The implementation process of this approach is shown in Figure 4.3 as a flowchart.

In the beginning, the total minimum demands of all VNOs, as well as the maximum demand of VNO GB is calculated, based on the number of users and the customised range of services data rates (the results of this calculation are also used in the admission control and delay process). In the next step, VNOs are checked one by one, to find the potential special case that can happen in VNO GB, when the total maximum demand of all users in VNO GB is less than the minimum contracted capacity of this VNO. In case of spotting this situation, the minimum SLA threshold for VNO GB is updated to its maximum demand, in order to avoid wasting resources; this way the algorithm assigns enough capacity to satisfy all GB users to the highest achievable level, while sharing the rest of the capacity among BG and BE VNOs. It is also notable that this situation cannot happen for VNOs BG and BE, since there is no upper limit threshold defined for capacity allocation in these VNOs.

The second unit represents the process of admission control and feasibility check, which is associated with a priority-based mechanism of delay. This is a necessary step before solving the main VRRM optimisation problem, in order to guarantee the feasibility of the constraints. The whole process is shown in Figure 4.4. The problem becomes unfeasible when the total minimum thresholds of guaranteed demands get higher than the available VRRM capacity. When this happens under high traffic loads, the network will be congested, therefore a logical decision in the first step is to delay all BE users, since there is no guarantee for these ones. The next step is to figure out, besides VNO BE, which services are going to suffer, therefore, by comparing the minimum SLA level and the minimum demand for the remaining active VNOs, if the demand surpasses the contracted SLA it specifies that the VNO's capacity share is not enough to serve all its users with at least the minimum level of the proposed data rate, and

since all BE users are already delayed, VRRM has to start delaying the users of that VNO one by one, from the lowest priority services, until the VNO's capacity becomes sufficient to address all minimum users' data rates of that VNO.

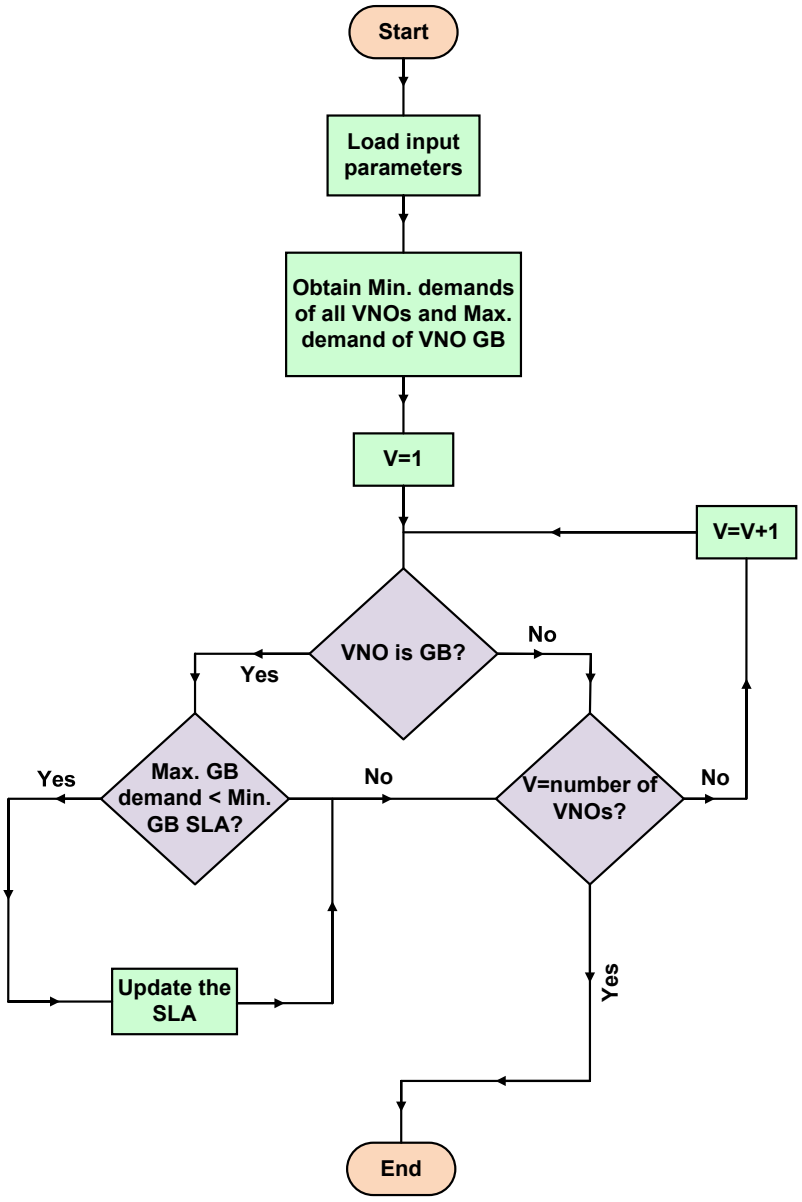


Figure 4.3 – Flowchart of SLA adaptation under low traffic loads.

Finally, in order to solve the main VRRM optimisation problem, three types of information are needed: pre-processed inputs, which are the main inputs after being checked (and potentially modified) by SLA adaptation and admission control blocks; objective function to define the relation between inputs and outputs; and the set of constraints to limit and shape the output results according to the specific policies of each VNO, such as the range of acceptable data rates or types of SLAs. The output of VRRM optimisation block is the demanded capacity that is needed to address the bandwidth requirements of each service slice, to the highest achievable level. These demands are then forwarded as a set of inputs to be provided by CRRM.

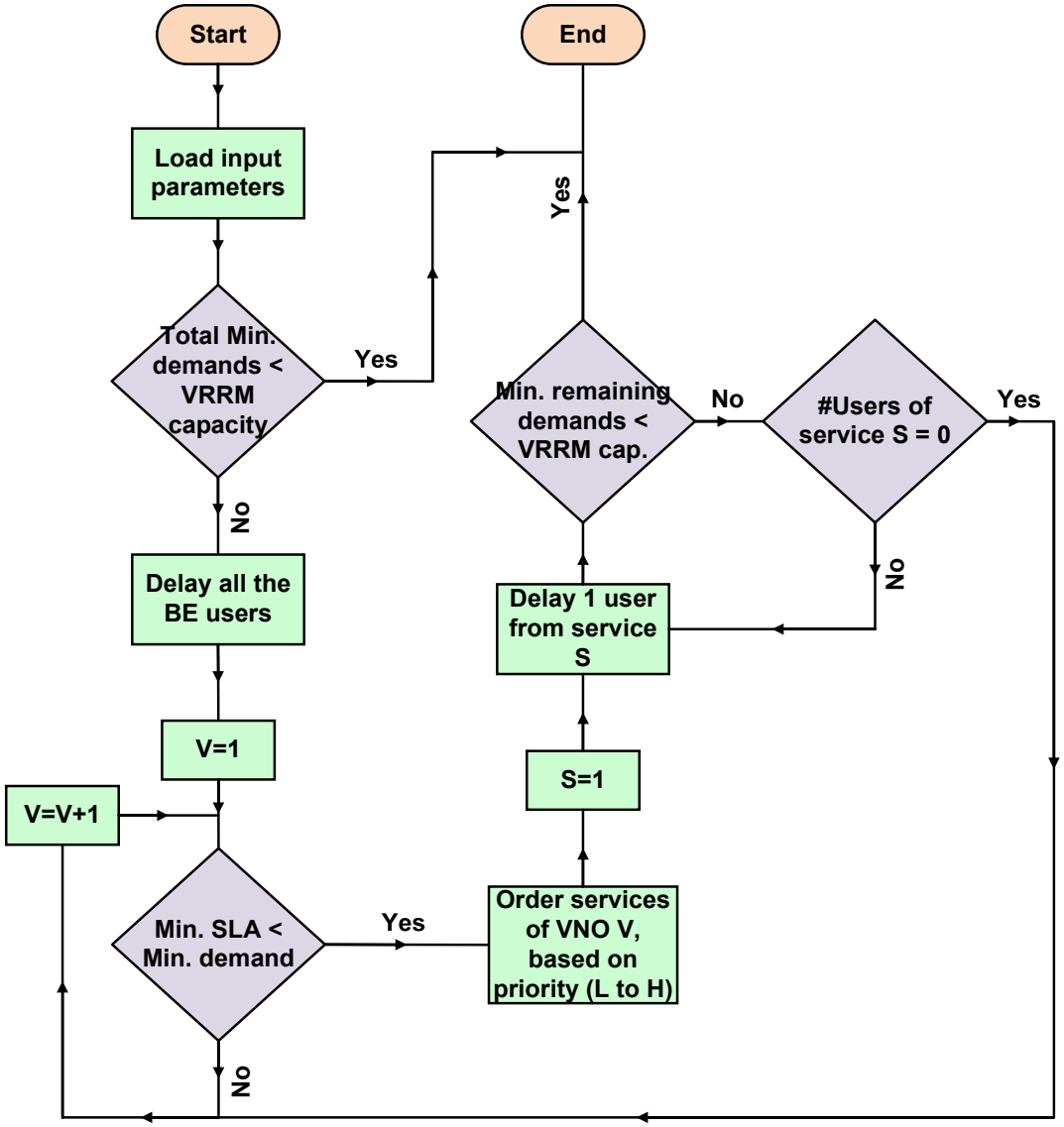


Figure 4.4 – Flowchart of admission control and delay process.

Regarding the third module, CRRM, three types of input parameters are needed to solve the optimisation problem: first, the *InPs' policies* in terms of loading factor, suitability and availability of a RAT; second, the information about the *aggregated capacity* of each RAT, which can be obtained from the specific number of RRUs; and third, the *demanded data rates* of all services from different VNOs, calculated by VRRM to be mapped and provided by the underlying RATs. Here, there is no need to check the feasibility of conditions, since service demands (thanks to the control mechanisms of VRRM) are always less than the total aggregated network capacity. Besides the input parameters of CRRM, the choice of the objective function (linear or logarithmic), as well as the set of constraints, i.e., the relation between the service demands and the maximum capacity of the RATs has to be defined in CVX to run and solve the CRRM optimisation problem. The output solution of CRRM is presented in Figure 4.5. The underlying RATs are virtually sliced into isolated instances to address the most suitable service demands of different VNOs; this way, VRRM and CRRM modules are interacting with each other.

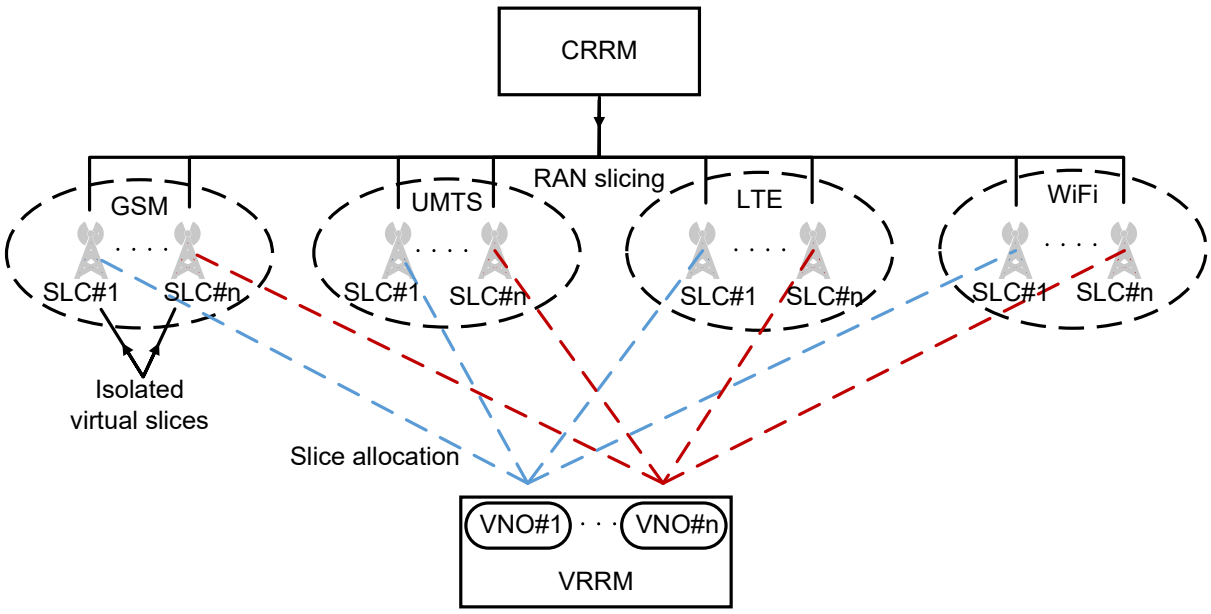


Figure 4.5 – Relation between the output of CRRM and the demand of VRRM.

4.4 Assessment of Simulator Development in CVX

In order to evaluate the convergence and accuracy of CVX in handling the proposed optimisation problems, a simple test scenario with 2 variables and the logarithmic objective function was defined and solved, both analytically and with the CVX solver, in order to make a comparison on the outcomes. The reason why just 2 variables were considered is that all variations can be shown in 2D or 3D plots. It is assumed that R_T is the total network capacity of 1 Gbps, to be shared between the two services, R_1 and R_2, which are the variables of the problem, in order to satisfy the following optimisation problem:

$$f(R_1, R_2) = \underset{R_1,R_2}{\text{Max}} \log\left(\frac{R_{1[\text{Mbps}]}}{R_{T[\text{Mbps}]}}\right) + 3\log\left(\frac{R_{2[\text{Mbps}]}}{R_{T[\text{Mbps}]}}\right) \quad s.t.\ R_{1[\text{Mbps}]} + R_{2[\text{Mbps}]} = R_{T[\text{Mbps}]} \tag{4.1}$$

By substituting R_2 as $R_T - R_1$, one can reformulate (4.1) to a single variable problem as:

$$f(R_1) = \underset{R_1}{\text{Max}} \log\left(\frac{R_{1[\text{Mbps}]}}{R_{T[\text{Mbps}]}}\right) + 3\log\left(\frac{R_{T[\text{Mbps}]} - R_{1[\text{Mbps}]}}{R_{T[\text{Mbps}]}}\right) \quad s.t.\ 0 \le R_{1[\text{Mbps}]} \le R_{T[\text{Mbps}]} \tag{4.2}$$

By taking the derivate of (4.2) with respect to R_1, the final solution is:

$$\frac{df}{dR_1} = 0 \ \rightarrow\ R_{1[\text{Mbps}]} = \frac{1}{4}R_{T[\text{Mbps}]}, \qquad R_{2[\text{Mbps}]} = \frac{3}{4}R_{T[\text{Mbps}]} \tag{4.3}$$

Since the total capacity is assumed to be 1 Gbps, the values of $R_1$and R_2 are 250 Mbps and 750 Mbps, respectively. By plotting (4.1) using MATLAB, the relation between the objective, constraint and variables are shown in Figure 4.6: the objective is indeed concave and the constraint is linear, therefore the optimisation problem can be solved in CVX. The maximiser point has to be located in the intersection of the objective function and the constraint, which in this case can be visually found as 250 Mbps and 750 Mbps for $R_1$and R_2 respectively, to confirm the analytical solution of (4.1), and the equivalent optimisation of (4.2) has the same solution. Figure 4.7 is the graphical expression of (4.2), which is in 2D, since the objective is reformed with respect to a single variable, the two subplots representing the cases when the objective is written with regard to R_1 or R_2, respectively, and in either case the results are exactly the same as before.

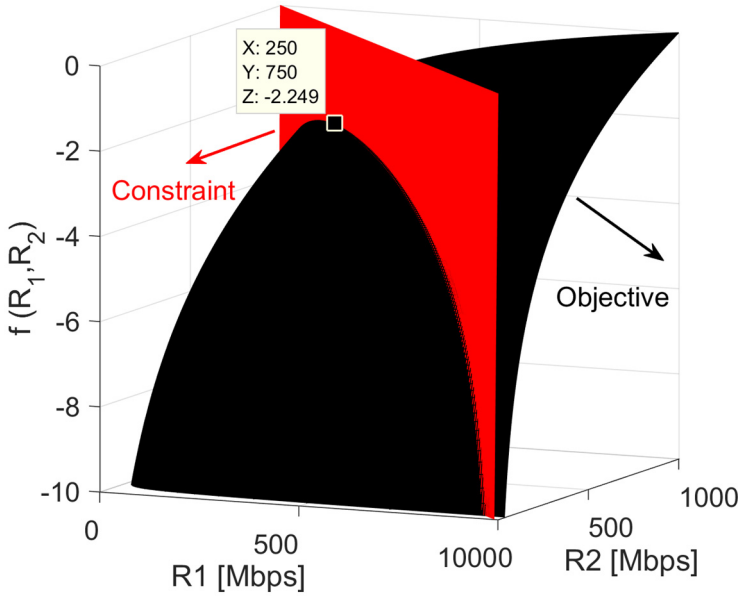


Figure 4.6 – The relation between the variables, objective function and constraint.

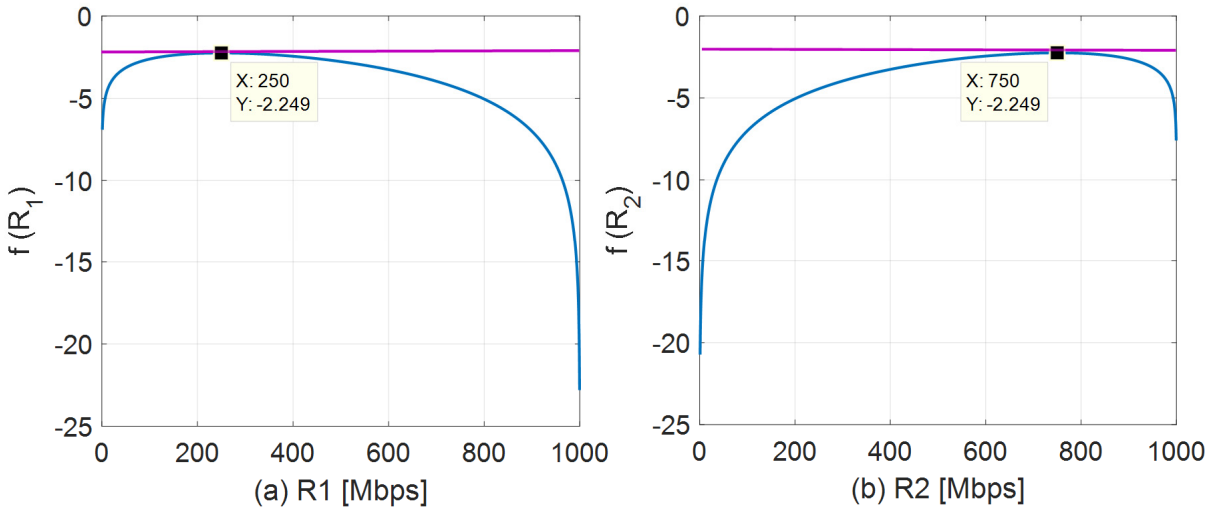


Figure 4.7 – Graphical expression of the problem in 2D.

The last step was to validate the performance of the developed simulator in CVX, by implementing the same example in the numerical solver. The results are given in Figure 4.8, having been obtained after 6 iterations. By moving from the top to the bottom of the table, it is noticeable that the errors are getting smaller in each step, meaning that the optimisation algorithm is improving and getting closer to the solution. In the last step, none of the variables evolved, therefore, the solution vector, as well as the objective value (optimal value) is returned, which are also exactly the same as the analytical results, as well as the graphically obtained maximum values that were plotted in MATLAB.

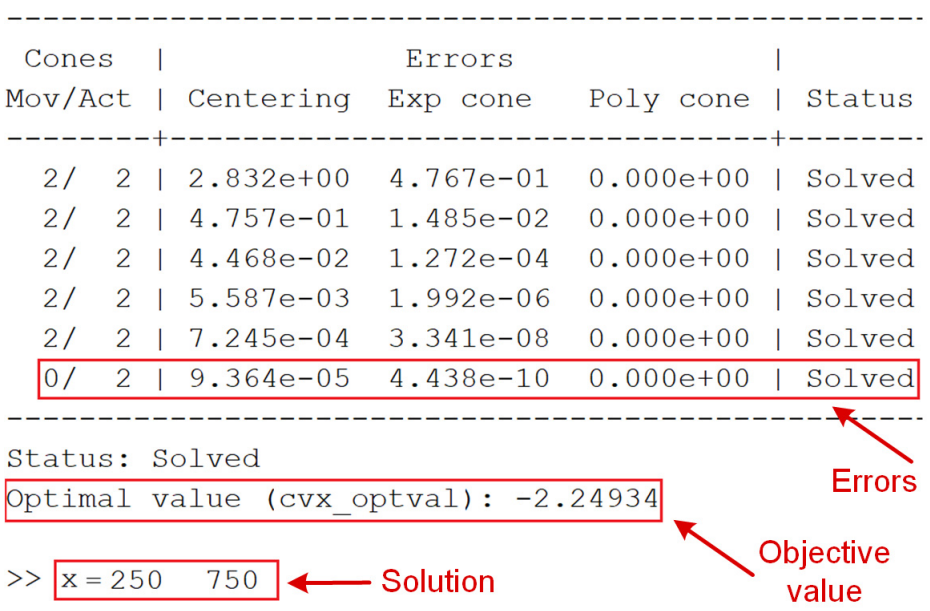


Figure 4.8 – Solution of the optimisation problem in CVX.

Chapter 5

Scenarios and Theoretical Results

This chapter presents the scenarios and theoretical results that are intended to confirm the performance efficiency and flexibility of the model in terms of satisfying the envisioned goals. In this regard, Section 5.1 introduces the reference scenario and effect of variation in key parameters such as traffic load, SLAs, active RATs and serving weights. Then in the following, by applying some changes to the reference scenario, other comparable scenarios are provided in Sections 5.2 and 5.3, in order to evaluate other aspects of the model's performance, such as the effect of variation in channel quality and achieving performance isolation among the VNOs.

5.1 Description of Reference Scenario

In order to evaluate various aspects of the model, the reference scenario is defined in this section. The key network parameters are then evaluated by some assessment metrics in the following sections through variations of different network parameters in this scenario, making other comparable versions of the reference scenario. The set of assumptions that are defined in each scenario are aimed to evaluate model performance in terms of fulfilling the specific goals provisioned for it. The key parameters of the reference scenario are cell layout, the profile of users, different services and SLAs, as well as the assigned service weights.

It is assumed that the area under analysis is 1 km^2 of an urban hotspot that is uniformly covered by all available RATs (cellular and Wi-Fi), with specifications summarised in Table 5.1, where $R_j^{RRU_{max}}$ is the maximum data rate of each RRU from different RATs, and $R_j^{RAT_{max}}$ the maximum aggregated capacity of each BS or AP in the covered area. However, since the quality of the physical channel is greatly influenced by SINR, these values are not achievable in practice. Using the proposed convolution PDFs to calculate the aggregated capacity of each RAT, by choosing $\alpha_p = 3.8$ as a common value for power decay in urban outdoor environments [BlCh14], and then randomly selecting from the associated convolution PDF of RATs, R_j^{RAT} as the value for the total available capacity of each BS or AP from different access technologies can be obtained. Summing up over all these values, R^{VRRM} is obtained as 510 Mbps.

Table 5.1 – RATs specifications for the reference scenario.

RAT	BSs, APs	RRU	$R_{j[\mathrm{Mbps}]}^{RRU_{max}}$	$R_{j[\mathrm{Mbps}]}^{RAT_{max}}$	$R_{j[\mathrm{Mbps}]}^{RAT}$
OFDM (Wi-Fi)	16	Carrier	0.97	336	252
OFDMA (LTE)	16	Res. block	0.75	240	192
CDMA (UMTS)	~ 1.7	Code	1.4	50	36
TDMA (GSM)	1	Time-slot	0.059	1	0.8

The main assumptions for network parameters together with their specifications are presented in Table 5.2, given three VNOs, each one with a different SLA contract (i.e., GB, BG and BE), providing a range of individual customised services from the four service classes: *Conversational*, *Streaming*, *Interactive* and *Background*. R_v^{vno} represents the contracted data rate of each VNO with InP, which varies according to the type of SLA, while the values of $R_{v_s}^{srv}$ indicate the range of average users' acceptable data rate, which is set according to the internal policies of different VNOs, and U^{srv} is the user mix of each service. VNO GB with the highest priority provides *Voice* (Voi), *Video calling* (Vic), *Video* (Vis) and *Music Streaming* (Mus), VNO BG offers *File sharing* (Fil), *Web* (Web), and *Social networking* (Soc), and finally VNO BE with the least priority provides *Email* (Ema) and narrowband IoT.

Table 5.2 – Network parameters for reference scenario.

VNO	Service	Class	$R^{srv}_{v_s\,[\mathrm{Mbps}]}$	$U^{srv}_{[\%]}$	δ_s	γ_v	SLA	$R^{vno}_{v\,[\mathrm{Mbps}]}$
1	Voice	Conversational	[0.032, 0.064]	25	5	10	GB	[0.4 R^{VRRM}, 0.8 R^{VRRM}]
	Video calling		[1, 4]	15	4			
	Video	Streaming	[2, 13]	45	3			
	Music Streaming		[0.064, 0.32]	15	1			
2	File sharing	Interactive	[1, R^{VRRM}]	50	4	2	BG	[0.3 R^{VRRM}, R^{VRRM}]
	Web		[0.5, R^{VRRM}]	15	3			
	Social networking		[0.4, R^{VRRM}]	35	2			
3	IoT	Background	[0, 0.1]	50	1	1	BE	[0, R^{VRRM}]
	Email		[0, R^{VRRM}]	50	1			

Moreover, InP's policy in RAT selection and service demand to RAT assignment is presented in Table 5.3. GSM is just dedicated to Voice and narrow-band IoT applications, since these two services have the lowest demand among all, which can be served by GSM, while WiFi is not accessible to these two services. In each horizontal line, the higher values of the $\mu^{RAT}_{v_s}$ weight represents the higher priority of a RAT for a specific service, compared to the other RATs. The weights are positive integer values, which vary between 1 and 10. It is further assumed that the capacity provided to VRRM according to the InP's policy is 70% of the whole capacity that can be obtained from the underlying RATs, with the possibility that each RAT can provide up to 80% of its available capacity to meet the assigned service demands.

Table 5.3 – CRRM's policy for service demand to RAT mapping.

Services	$\mu^{RAT}_{v_s}$			
	TDM (GSM)	CDMA (UMTS)	OFDMA (LTE)	OFDM (Wi-Fi)
Voice	10	5	2	NA
Video calling	NA	7	10	8
Video	NA	5	8	6
Music Streaming	NA	3	6	4
File sharing	NA	4	5	7
Web	NA	4	4	7
Social networking	NA	3	3	6
Email	NA	2	3	5
IoT	5	2	1	NA

5.2 Analysis of Different Parameters in the Reference Scenario

5.2.1 Influence of the Variation in Number of Offered Users

The performance of the proposed VRRM model under the assumptions of the reference scenario is presented and analysed in this section, in terms of some evaluation metrics to study the effect of traffic load on the capacity share of VNOs and their associated services. Figure 5.1 shows the effect of the number of offered users on the total capacity share of the three VNOs. From this perspective, two parameters play decisive roles in decision making: the acceptable ranges of data rate variation for each VNO, which are indicated in the figure as one type of constraint, representing the *internal policy* of each VNO for service management considering the values of $R_{v_s}^{srv}$. Another factor is the SLA type and contracted capacity between the VNO and InP, according to the values of R_v^{vno}, which are fixed and do not change with the variation of traffic. The coloured areas in Figure 5.1 represent the case when all offered users in each VNO are being served. However, it is notable that when the number of offered users is very low (roughly less than 50), all users in VNO GB are served with the highest acceptable data rates, which are in fact lower than the values of the SLA contract. In this special situation, the capacity allocated to VNO GB according to the VRRM algorithm is set to the maximum acceptable data rate that can be offered to GB users.

After the intersection points, when the minimum acceptable threshold of data rate surpasses the allocated capacity of each VNO, VRRM starts to delay the users from the suffered VNO based on the service priorities. As an example, when the delay in VNO GB starts, since this VNO has the highest priority among the three, all users of VNO BE have already been delayed (as there is no minimum SLA guarantees for this VNO), and the allocated capacity of VNO BG has reached down to the minimum contracted SLA threshold.

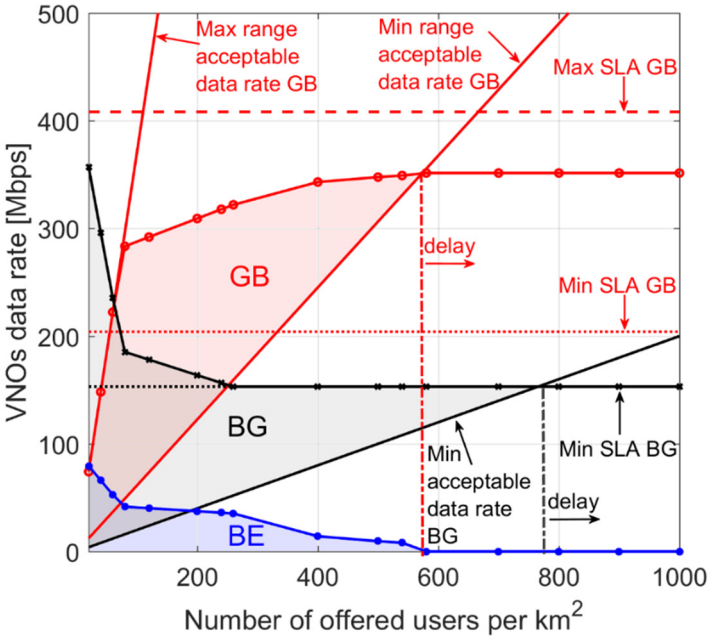
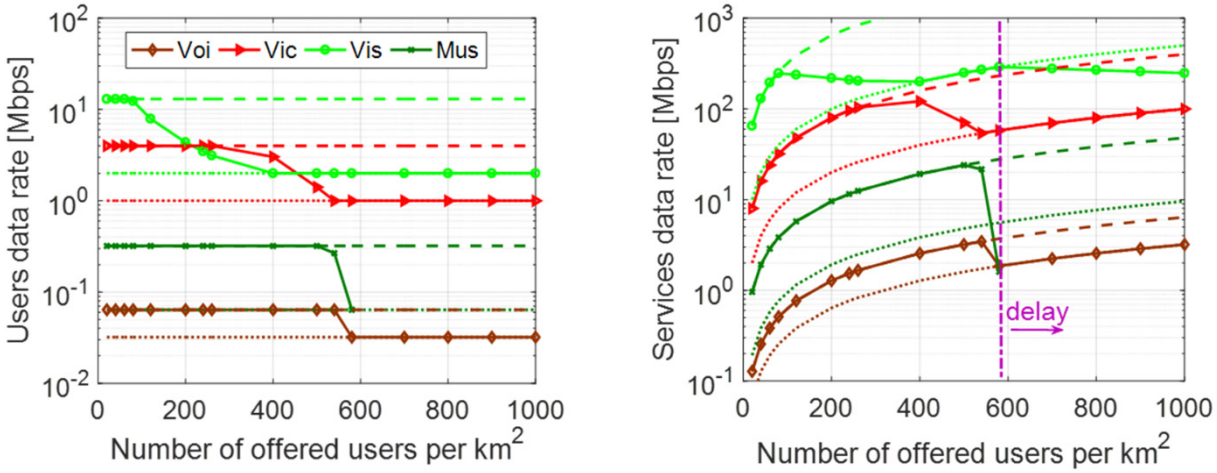


Figure 5.1 – The effect of number of users on the VNOs data rates.

As another set of evaluation metrics, the influence of traffic load on the average data rates of served users, as well as the capacity share of each service, are studied and the results are presented in Figure 5.2, Figure 5.3 and Figure 5.4, for VNOs GB, BG and BE, respectively. The *dotted* and *dashed* lines in all subplots represent the *lower* and *higher* thresholds for capacity, which can be assigned to each user or service, according to the customised range of serving data rates in Table 5.2. As the number of users increases, the average data rate of users, and consequently the total allocated capacity to each service, reaches down the minimum acceptable level. In the beginning, when the traffic load is lower, all GB users are well served with the maximum achievable data rates (defined in Table 5.2), therefore, the rest of the capacity is shared among BG and BE users with much higher data rates compared to GB ones. However, as the number of offered users increases, the data rates of BG and BE users suffer from a sharp decrease, since the services provided by these VNOs have lower priority compared to GB ones. Regarding the ratio of capacity share among the internal services of each VNO, one can see from all sub-plots that when there is no limitation in data rate assignment (such as Figure 5.3 (b)), the whole capacity of each VNO is divided among the services *proportional* to their corresponding values for serving weights.

Afterwards, when the total aggregated capacity is not enough to satisfy all users with at least the proposed minimum data rates, VRRM starts the process of *delaying* users based on their service priority, as a logical solution to handle the extreme situations of network traffic load: the whole process is shown after the vertical dashed lines. For example, in VNO GB, first *Mus* users are delayed, and then some of *Vis* users, in order to release enough capacity for the rest of higher priority ones (i.e., Conversational, which includes *Voi* and *Vic,* while the two latter services are not affected). It is also notable that the process of delaying BG and GB users starts just after the point when there are no BE users left to be delayed, since there are no service guarantees for VNO BE and consequently under extreme situations of capacity shortage VNO BE is the one that suffers first. It is also evident that the starting points of delay in all VNOs are exactly compatible with the ones of Figure 5.1.



(a) VNO1 (GB) – Users data rate

(b) VNO1 (GB) – Total data rate

Figure 5.2 – Average users and services data rates of VNO1 (GB).

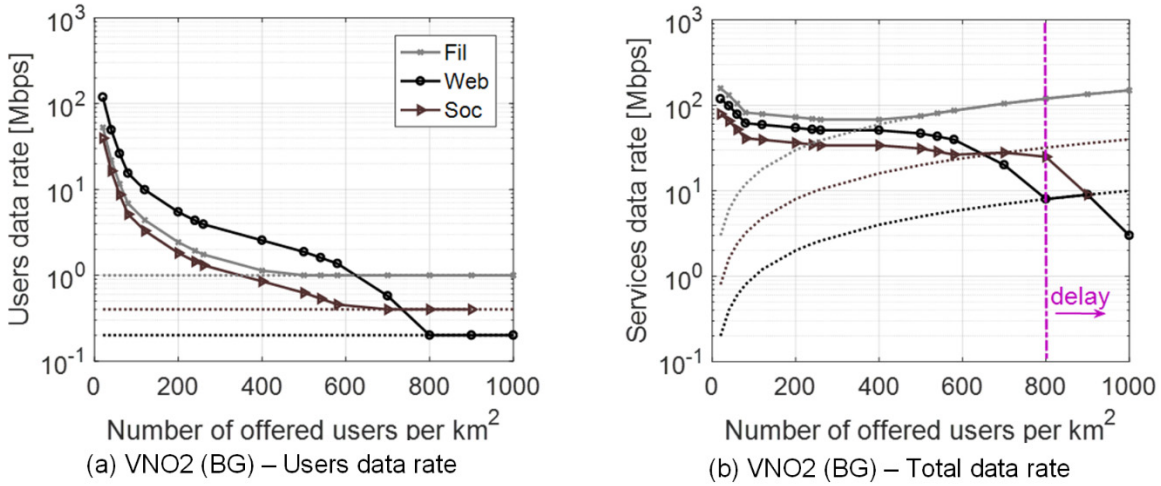


Figure 5.3 – Average users and services data rates of VNO2 (BG).

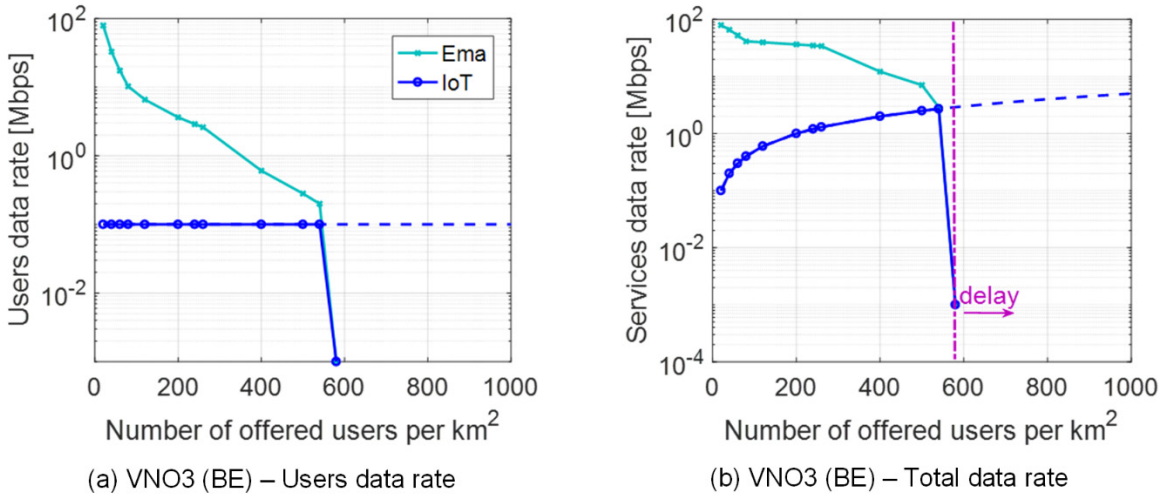


Figure 5.4 – Average users and services data rates of VNO3 (BE).

5.2.2 Evaluation of CRRM Performance

This section analyses the performance of CRRM with the logarithmic objective, in terms of RAT selection and mapping the service demands requested by VRRM (which are obtained in the previous sub-section), through CRRM, in order to be provided by the underlying RATs. Figure 5.5 represents the percentage of available capacity that is assigned from different RATs to satisfy the required service slice demanded data rates. Since GSM is just accessed by Voi and IoT users, and those service data rates are upper bounded by predefined thresholds, the demands grow linearly by increasing the number of users, therefore the assigned capacity from GSM also grows linearly up to the point when VNO GB starts delaying users (roughly around 600). At this point, all IoT users are delayed and Voi users continue their service with the minimum acceptable data rate, which is the reason why there is a sharp drop when the number of offered users reaches to 600. The overall trend of capacity allocation in UMTS is comparable to GSM, meaning that it is increasing up to 600, and then there is a drop afterwards, however, the total allocated capacity is much more balanced compared to GSM, since much more service demands are to be addressed by UMTS.

By comparing the general trend of the total allocated capacity from LTE and Wi-Fi, one realises that the variation is almost always around 70% of the whole capacity of those RATs, except for the very beginning, when the traffic load is very light. The reason is that LTE has a higher weight for Vis and this service is the most demanding one compared to the others, therefore, when the traffic load is very low, Vis data rate is bounded by the maximum demanded data rate and consequently decreases the total capacity share of LTE, while Wi-Fi has higher weight for unbounded, lower priority services such as Fil and Web, which have higher demanded capacity in the beginning but drop faster by increasing the number of users. Furthermore, by comparing the values of assigned data rates from different RATs, it is also noticeable that when there is no limitation in capacity allocation, the shares of capacity for each service from different RATs are proportional to the values of $\mu_{v_s}^{RAT}$ from Table 5.3. As an example, the share of capacity that is assigned to Vis from UMTS, LTE and Wi-Fi, as shown in Figure 5.5, is exactly proportional to the predefined values of $\mu_{v_s}^{RAT}$, which are 5, 8 and 6, respectively. Finally, the total assigned capacity of all the RATs remain less than 80% of the aggregated capacity of each RAT, which is in accordance with the policy of InP, while in total 70% of the available aggregated capacity (in case of existing service demands) is utilised independent of the variation of traffic load and different network parameters to satisfy the VRRM demands.

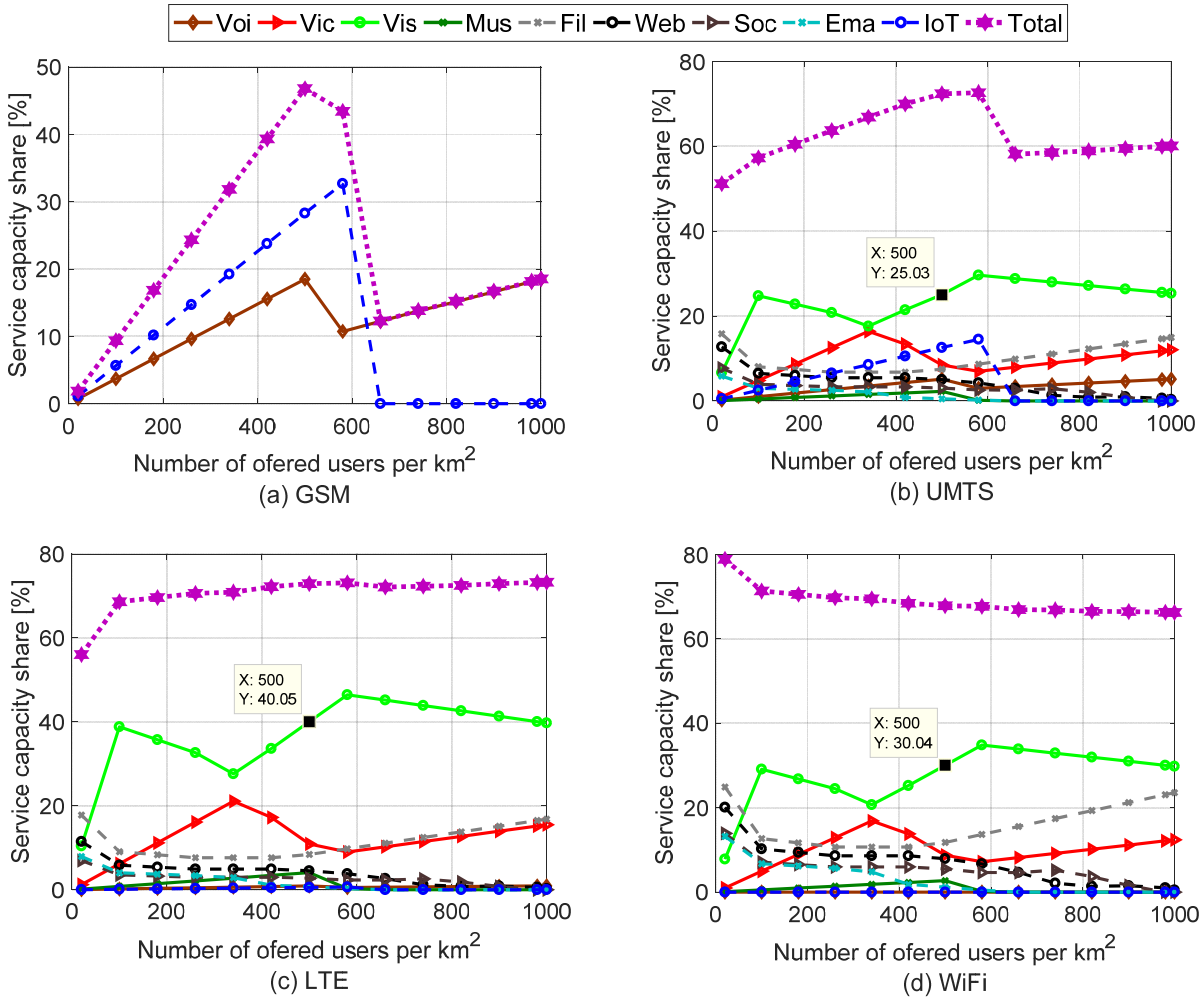


Figure 5.5 – CRRM mechanism of mapping service demands onto the suitable available RATs.

5.2.3 Influence of the Variation in SLAs and Serving Weight Coefficients

Another evaluation metric is the influence of SLA contracts on the percentage of capacity share among VNOs, which is shown in Figure 5.6, assuming that the minimum SLA threshold of BG VNO varies from 10% to 60% of the total available capacity (it is not supposed to be higher, since the minimum SLA of VNO GB is set to 40%) and the number of offered users, in this case, is 300. When the minimum BG SLA is roughly less than 25%, capacity shares are constant, the reason being that the *minimum SLA* values remain less than the *minimum acceptable* data rates of VNO BG, therefore, the variation of BG SLA up to this point is not effective on the capacity shares (the dominant constraint is the minimum acceptable data rate, which just varies with the number of users, since this number is fixed to 300, this value remains constant). After this point, the values of minimum SLA become dominant, therefore the BG area starts to expand (with the width of at least the minimum BG SLA), while the GB and BE shares shrink. When this value reaches 60%, since the minimum GB SLA threshold is 40%, all BE users are delayed. Accordingly, the BE area disappears at this point.

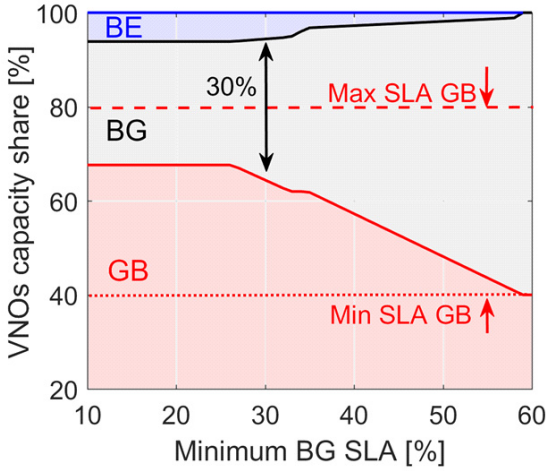


Figure 5.6 – The effect of SLA contracts on the VNOs' capacity share.

The effect of variation in serving weights on the capacity share of VNOs is studied in Figure 5.7. It is assumed that all serving weights of VNO BG (related to different services) are multiplied by a value between 1 and 12, when the number of offered users is 300. In the beginning, when the value of the multiplier is 1, the share of VNO BG is 30%, the reason being that (from Figure 5.1) the data rate of the VNO BG has just reached down the minimum SLA level, when the number of offered users is 300. However, by increasing the serving weights compared to the other two VNOs, the capacity share of VNO BG also increases compared to the other two VNOs, up to the point that the share of VNO GB reaches down the minimum level of the contracted SLA, which is 40%. After this point, by increasing the serving weights, there is not a considerable increase in the capacity share of VNO BG, therefore, one can conclude that, although by variation of serving weights the capacity share of the VNOs changes, these values always vary within the range of the contracted SLAs.

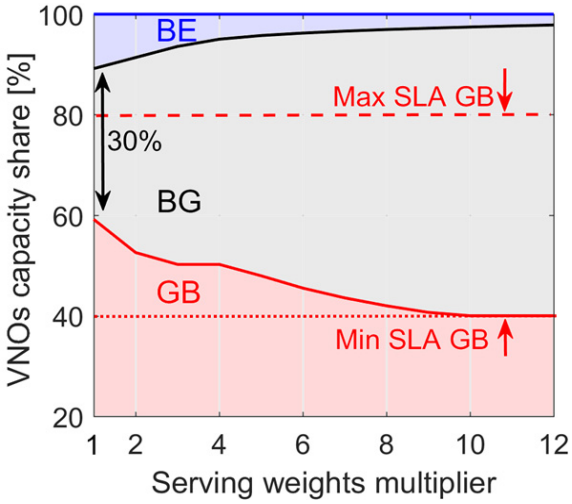


Figure 5.7 – The effect of serving weights on the VNOs' capacity share.

After evaluating the global effect of variation in *SLA* and *serving weights* on the capacity share of the existing VNOs, in the following, the effect of SLA variation on some internal key parameters of each VNO are studied. One may further analyse the influence of serving weights in the same way, however, due to the similarity of the procedure and the general trend of results (as shown in Figure 5.7), the effect of varying serving weights inside each VNO is not discussed here. Figure 5.8(a), represents the effect of SLA variation on average users' data rate of VNO GB, when the total offered number of users is 600. It is noticeable that by increasing the minimum BG SLA, since the capacity share of VNO GB decreases (according to Figure 5.6), the data rate of higher priority services, i.e., Voi and Vic, reaches the minimum acceptable, whereas all Mus users are delayed after 40%, and therefore some of the Vis users are also delayed after this point to provide enough capacity for the reset of higher priority users in order to continue their services with the minimum acceptable data rates. Figure 5.8(b) is the evidence that independent from the choice of number of users, the data rate of services decreases as the minimum BG SLA increases, however, this variation is always between the predefined acceptable data rate of each service, which in this case for Vic varies between 1 and 4 Mbps.

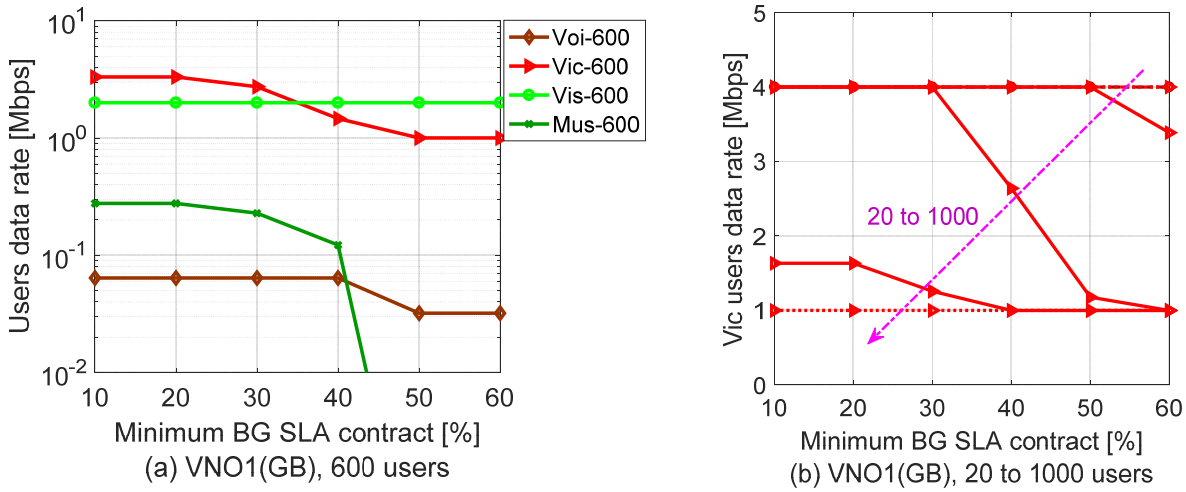


Figure 5.8 – Effects of variation of SLA on the average users' data rates of VNO GB.

The percentage of served users from VNO GB is shown in Figure 5.9(a), which is completely compatible with the results and thresholds of Figure 5.8(a). As one can notice, all the users are being served before 40%, after this point all Mus users as the lowest priority ones are delayed, and since the released capacity is not enough, a portion of Vis users are also delayed. Figure 5.9(b) represents the total percentage of served users by VNO (GB), for different number of users and under different SLA agreements. By increasing the number of users (from the top to the bottom of the figure) it is obvious that the total number of served users decreases, and then by increasing the minimum BG SLA (from the left to the right of the figure) the percentage of served users also decreases.

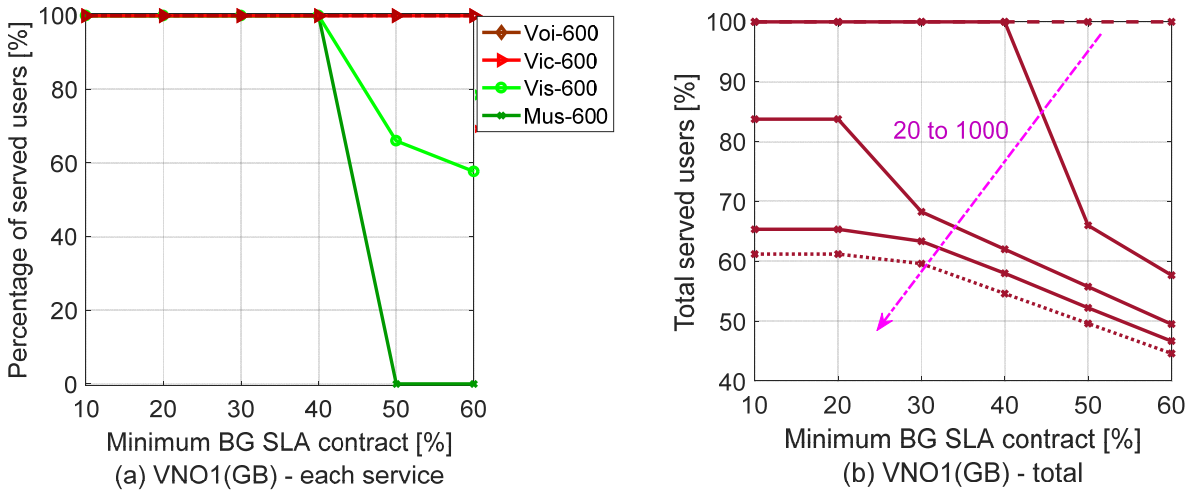


Figure 5.9 – Percentage of served users in VNO GB.

The average users' data rate of VNO BG is shown in Figure 5.10(a) for 800 users. The trend of variation in the values of data rate in VNO BG is opposite to the ones from VNO GB, meaning that by increasing the SLA values the users' data rates in VNO BG also increases, starting from the minimum acceptable values due to an increase in the capacity share of VNO BG, however, this is in contrast to variation of the values in VNO GB (Figure 5.8), since the share of VNO GB decreases. One can also notice that the data rate of Web users grows faster than the other two services, which is due to the fact that (according to Table 5.2) the traffic of Web users is less than the other two services. Figure 5.10(b) represents the variation of data rates for Fil users, when the SLA and number of offered users change. By increasing the number of users from 20 to 1000, since the traffic load increases the values of data rates for a specific SLA decreases, however, by increasing the SLA values for a particular number of users, as the capacity share of VNO BG increases, the data rate of all the services including Fil increases as well.

The percentage of served users in each service of VNO BG is shown in Figure 5.11(a), when the number of users is 800. For the lower values of SLA, a portion of Soc users as the lowest priority ones are delayed to provide enough capacity for the rest of higher priority services, i.e., Fil and Web. As the values of SLA increases, all users are served due to an increase in the capacity assigned to VNO BG. Figure 5.11(b) represents the overall percentage of served users in VNO BG when varying the SLA, under different traffic loads. It is obvious that by increasing the number of users a higher percentage of users are delayed, whereas by increasing the SLA values, a higher percentage are served.

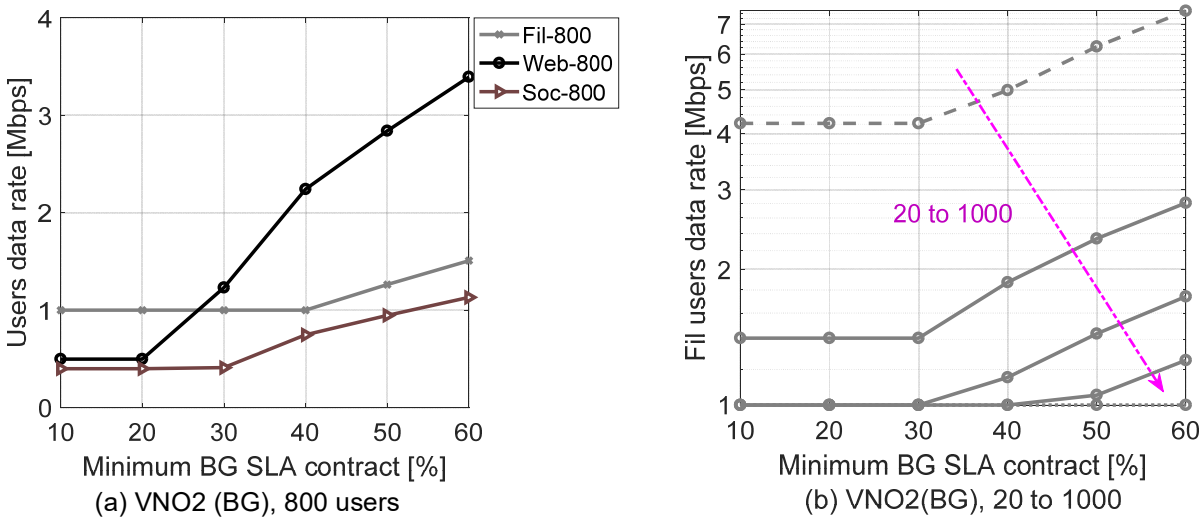


Figure 5.10 – Effects of variation of SLA on the average users' data rates of VNO BG.

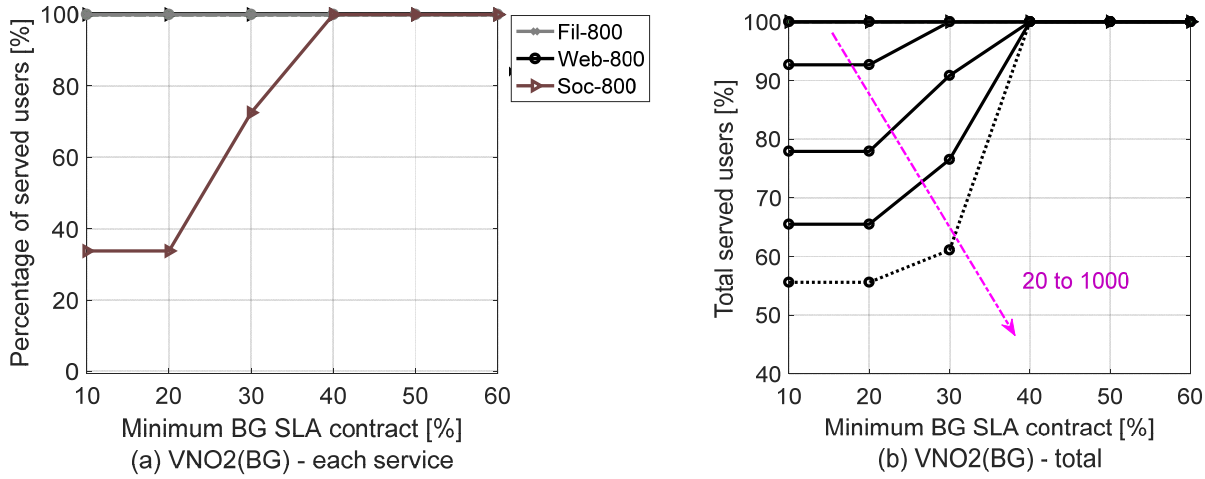


Figure 5.11 – Percentage of served users in VNO BG.

Finally, the effect of SLA variation on users' data rate of VNO BE is shown in Figure 5.12(a) for 600 users. For the lower values of SLA, since the share of VNO BE capacity is higher, IoT users are served with the maximum achievable data rates. As SLA values increase, the data rate of both Ema and IoT services decrease and drop to zero when the minimum BG SLA reaches to 50%, the reason being that at this point VRRM starts delaying GB users, and therefore all BE users have to be delayed, since they have a lower priority compared to GB ones. The variation of IoT data rate with SLA is presented in Figure 5.12(b); under the very low traffic loads, IoT is served with 0.1 Mbps as the highest achievable data rate and under high traffic load none of IoT users is served.

The percentage of served users of the two services of VNO BE is shown in Figure 5.13(a). Since the minimum data rate of all BE users is set to zero, these users can be served for any positive data rate value, which is the reason why the percentage of served users suddenly drops to zero. The results for different users in Figure 5.13(b) also confirm that by increasing the traffic load and minimum BG SLA, the total number of served users decreases as well.

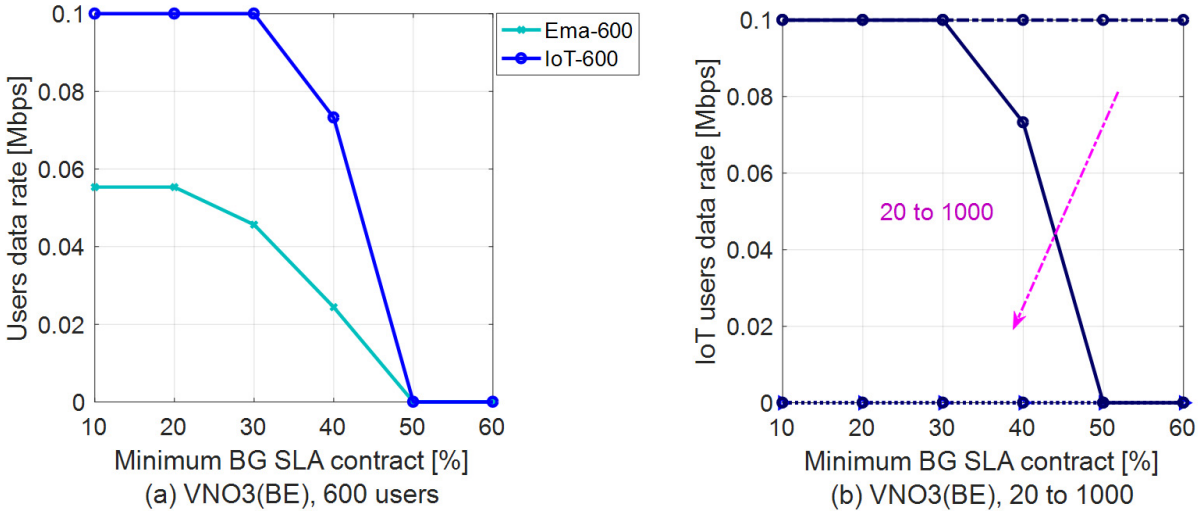


Figure 5.12 – Effects of variation of SLA on the average users' data rates of VNO BE.

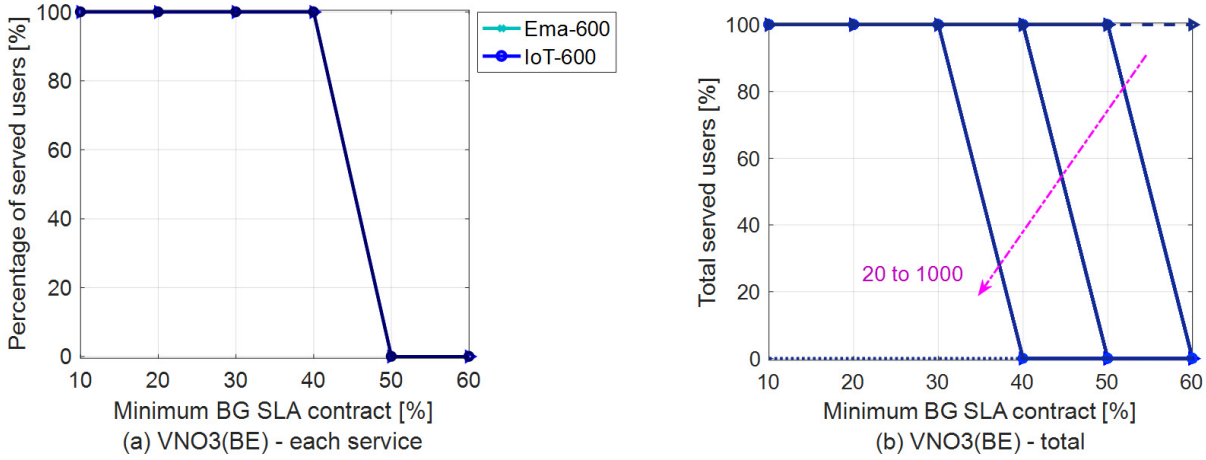


Figure 5.13 – Percentage of served users in VNO BE.

5.2.4 Effect of Varying the Number of Available RATs for Energy Efficiency

Coverage and capacity planning are essential parts of cellular network design. This subsection discusses the effect of varying the number of RATs, which can be considered as a metric for inter-cell coordination in 5G network design, namely cell switch on/off for energy saving. Since small cells have much less coverage compared to macro cells, it is likely that certain small cells in the network have no served users or traffic in long periods and can be switched off or go to a low power mode for energy saving. The cell switch-off approach is a system level design, which works in an area covered by multiple cells with the same or different RATs [3GPP10b], has been introduced in LTE networks. When the traffic load in a certain area is low, some cells could be shut down and the served users can be handed over to neighbouring ones. In this approach, there is no need to modify the lower-layer components in BSs. The popular cell switch-off structure is called the hierarchical cell structure [3GPP09], in which always macro-cells are deployed for basic coverage and micro-cells are planned for capacity boost.

Therefore, it is assumed that the InP's policy in this subsection is to activate a number of RATs according to the traffic load, in order to address the service demands and SLAs with the minimum acceptable level.

By varying the number of LTE and Wi-Fi RATs in the reference scenario, the effect of this parameter on satisfying the SLAs of different VNOs and service requirements is analysed as follows. If the capacity is not enough to service capacity requirements, 4 pairs of LTE BSs and Wi-Fi APs are further activated to satisfy the demands and SLAs, until the full capacity of the network is reached (16 pairs of LTE and Wi-Fi). The performance of the proposed energy efficient SLA-based slicing strategy is presented and analysed in this section, in terms of some evaluation metrics. Figure 5.14 shows the effect of the number of offered users on the total capacity share of VRRM among the three VNOs.

In this regard, as mentioned before, two parameters play significant roles in decision making: the acceptable ranges of data rate variation for each VNO, which are indicated in Figure 5.14 as one type of constraint, representing the *internal policy* of each VNO for service management considering the values of $R_{v_s}^{srv}$. Another factor is the SLA type and contracted capacity between the VNO and InP, according to the values of R_v^{vno}, which change with the number of active RATs. This way, when the traffic load is very low (roughly less than 150 users), InP activates 4 RATs and as the number of users increases, when the minimum acceptable thresholds of GB and BG capacity shares (considering both SLAs and demands), reaches to the available capacity of VRRM, InP activates 4 more RATs and in each step continues the same routine until the full capacity of 16 RATs is achieved. As one can notice, the lowest acceptable capacity limit for VNO GB is mostly set by the minimum demands, while for VNO BG is set by the minimum SLA threshold. Furthermore, considering the energy saving policy of InP in satisfying the service slice demands, the capacity share of VNO BE always remains very low and close to zero. After the intersection points, since all available RATs are already activated, when the minimum acceptable threshold of data rate surpasses the allocated capacity of each VNO, VRRM starts delaying the users from the suffered VNO based on the service priorities, in the same procedure of Figure 5.1.

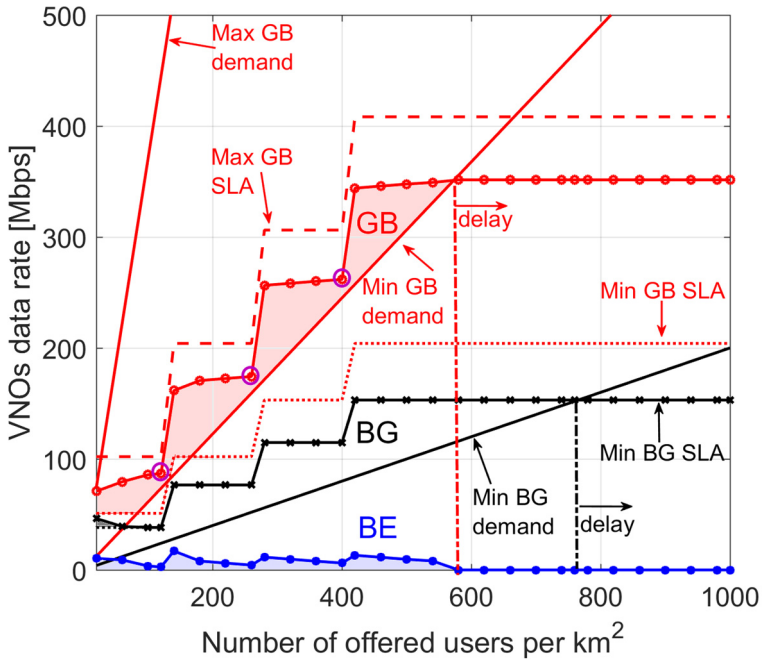


Figure 5.14 – The effect of traffic load on VNOs capacity share under the energy-efficient technique.

As another set of evaluation metrics, the influence of traffic load on the average data rates of served users, as well as the capacity share of each service slice, are studied, the results being presented in

Figure 5.15, Figure 5.16 and Figure 5.17, for VNOs GB, BG and BE, respectively. As the number of users increases, in general the allocated data rate to service slices, and accordingly the average users' data rates decrease and then by activating more RATs experience a relatively sudden increase due to the higher available capacity, except for the services that are constrained by the lower or higher thresholds. For instance, *Vis* in VNO GB, which has the highest data rate requirement among all other service slices, is mostly served by its minimum acceptable data rate, while, on the other hand, *Mus* and *Voi*, which need lower data rates, are fully satisfied before delay starts.

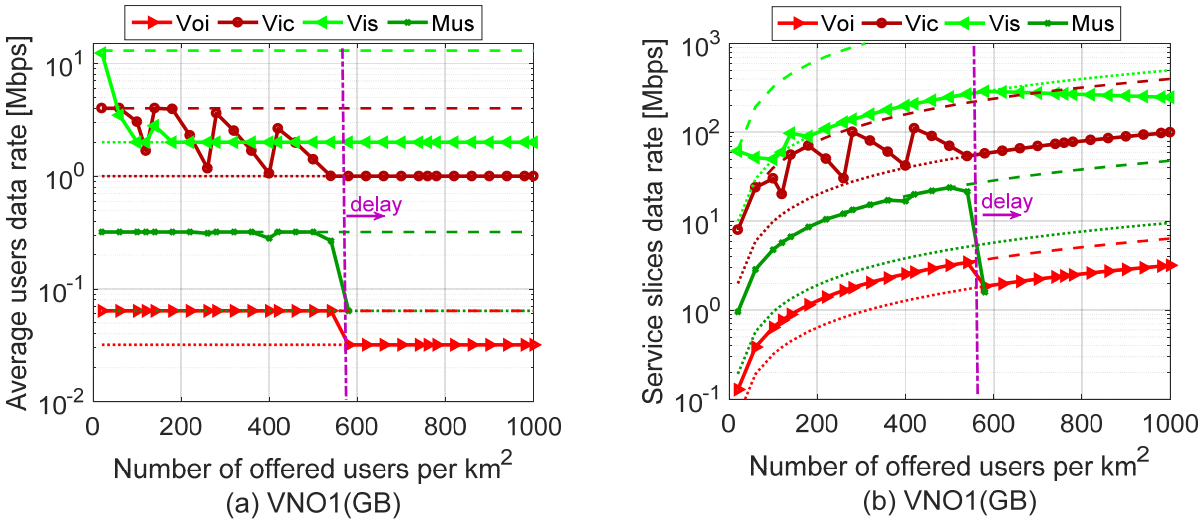


Figure 5.15 – Users and service slices data rates of VNO GB, under the energy efficient approach.

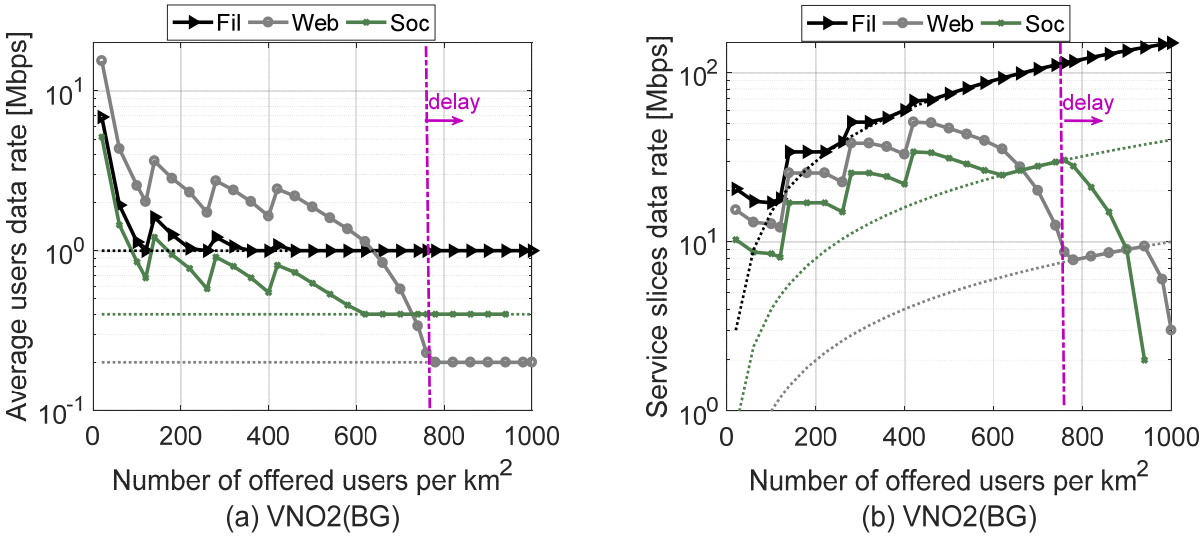


Figure 5.16 – Users and service slices data rates of VNO BG, under the energy efficient approach.

It is also notable that when the capacity share among service slices is not constrained by the thresholds, the data rates allocated to different service slices are *proportional* to the customised values defined by VNOs. As an example, in VNO BG, when the number of offered users is less than 700, the data rate allocated to *Web* slice, is exactly 1.5 times higher than that of *Soc*. Furthermore, although the proposed approach is energy efficient, the number of active RATs does not affect the number of delayed users, meaning that when the number of active LTE and Wi-Fi RATs is less than 16 pairs, all users are satisfied

with at least the minimum acceptable data rates defined in the reference scenario. Delay starts only after the *full* available aggregated capacity of VRRM is used.

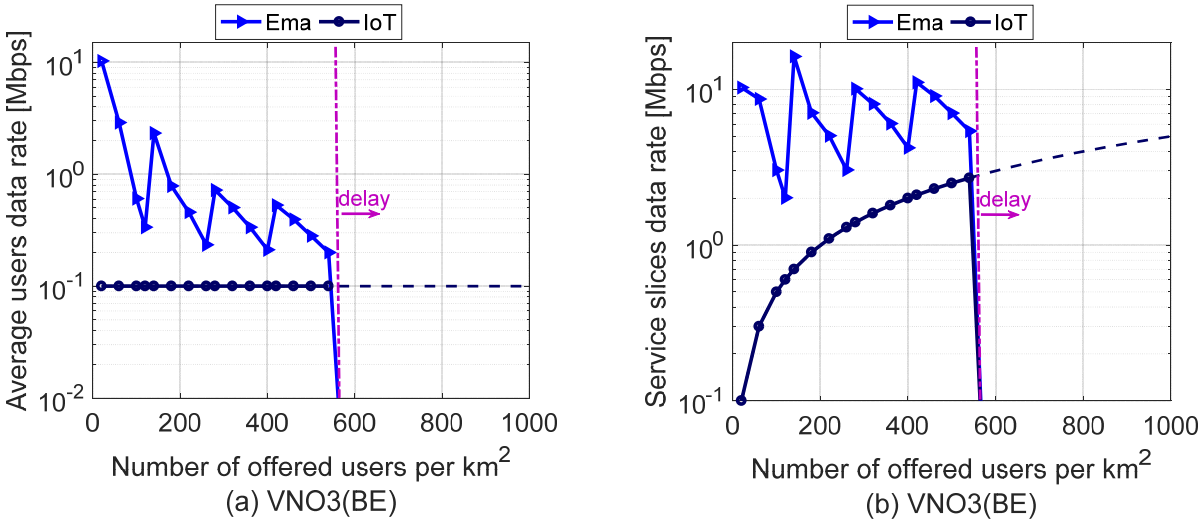


Figure 5.17 – Users and service slices data rates of VNO BE, under the energy efficient approach.

Another evaluation metric is related to the efficiency of CRRM in mapping service demands onto the underlying RATs, according to the InP's policies in RAT selection and load balancing, which is shown in Figure 5.18. When the traffic load is very low, less than 20% of the RRUs from both LTE and Wi-Fi are enough to address service demand; as the traffic load increases, in each step 4 more RATs are activated. Concerning the effect of the load balancing factor, δ_j^{RAT}, if it was assumed that the InP does not have any preference in RAT suitability (i.e., all the values of $\mu_{v_s}^{RAT}$ are equal), since the values of δ_j^{RAT} are *proportional* to the capacity of each RAT, it would have been expected to experience exactly the same level of load (in percentage) for both LTE and Wi-Fi RATs. However, as the values of $\mu_{v_s}^{RAT}$ change according to the suitability in service-to-RAT matching, the load of LTE and Wi-Fi RATs is not exactly the same.

For instance, when 100% of the VRRMs' capacity is used (with 16 pairs of active RATs), both LTE and Wi-Fi should have been equally loaded with 70%, since according to the InP's policy, the maximum capacity provided to VRRM is defined as 70% of the aggregated available capacity of CRRM. However, since $\mu_{v_s}^{RAT}$ is different for each service, the loaded level of each RAT varies accordingly. As an example, considering the values of $\mu_{v_s}^{RAT}$ from the reference scenario for mapping of *Vis* and *Vic* demands to Wi-Fi and LTE, the ratio of the suitability weights is proposed as 0.75 and 0.8, respectively. By comparing the associated values from Figure 6(a) and Figure 6(b), one can notice that the share of provided demands between the two RATs is proportional to the ratio of $\mu_{v_s}^{RAT}$ factors for both *Vis* and *Vic*. It is also notable that the load of both RATs remains less than 80%, which is the maximum nominal value set by the InP.

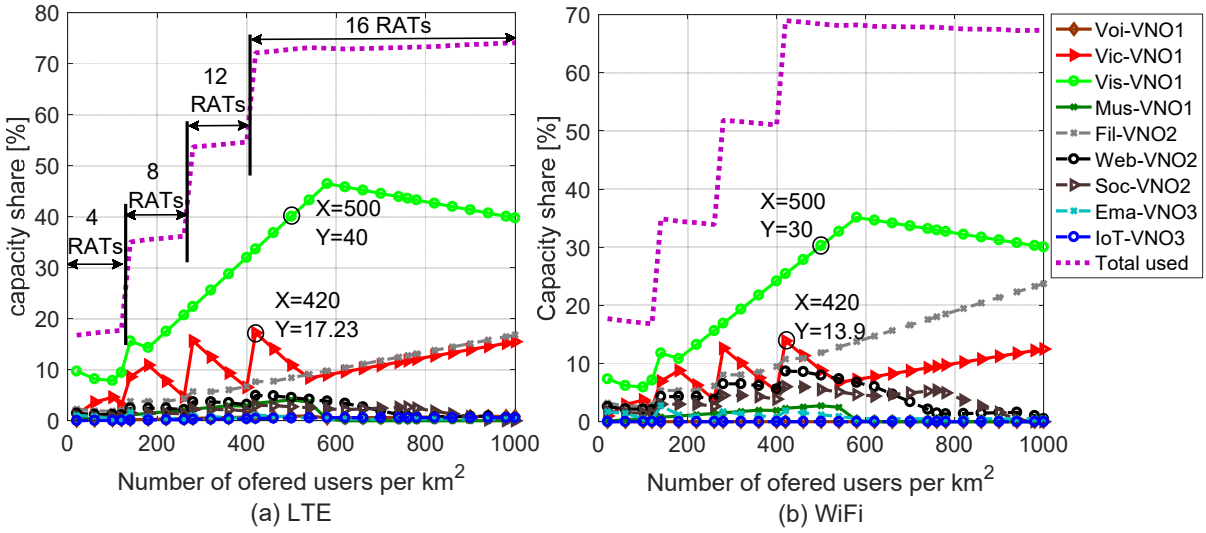


Figure 5.18 – Distribution of load between the active LTE and Wi-Fi RATs.

5.3 Assessment of the Effect of Variation in Channel Quality

The aggregated capacity of the network can be obtained by considering different channel conditions, as the link quality between user terminals and BS/AP has a considerable effect on the total available capacity. Therefore, this section studies how *good* and *bad* channel conditions affect the data rate that has to be guaranteed for different services. Obviously, in the case of *good* channel the better radio channel condition allows obtaining a higher channel quality indicator and consequently a higher data rate. The definition of *good* and *bad* channel quality, in terms of the achievable data rate of each RRU is defined as follows:

- *Good* channel quality, in which RRUs satisfy:

$$0.5\, R_{j[\text{Mbps}]}^{RRU_{max}} \leq R_{j[\text{Mbps}]}^{RRU} \leq R_{j[\text{Mbps}]}^{RRU_{max}} \tag{5.1}$$

  where $R_{j[\text{Mbps}]}^{RRU_{max}}$ is the maximum achievable data rate of a single RRU from RAT j.

- *Bad* channel quality, in which all the RRUs' data rates fall within the following interval:

$$0 \leq R_{j[Mbps]}^{RRU} \leq 0.5\, R_{j[Mbps]}^{RRU_{max}} \tag{5.2}$$

Regarding the scenario assumptions, it is assumed that two VNOs are in charge of providing the four service classes, Table 5.4. In each service class, the first service is associated with VNO1, and the second one is assumed to be served by VNO2. Half of the services are categorised as GB, therefore, having higher serving weights, while this parameter for both low priority BE services is set to be 1, in accordance with the framework for serving weight assignment.

Table 5.4 – Network parameters for two VNOs.

Service	Class	Data rate [Mbps]	User mix [%]	λ_{v_s}	SLA
Voice	Conversational	[0.032, 0.064]	10	50	GB
Video call		[0.3, 5]	10	50	
Video stream	Streaming	[2, 13]	30	30	
Music stream		[0.064, 0.32]	15	16	
File sharing	Interactive	[1, R^{VRRM}]	15	6	BG
Web		[0.2, R^{VRRM}]	10	4	
Email	Background	[0, R^{VRRM}]	5	1	BE
Smart metering (Sma)		[0, R^{VRRM}]	5	1	

The numerical trend of VRRM data rate assignment is presented as follows. The average long-term data rate of users under good and bad channel conditions from both VNO1 and VNO2 are shown in Figure 5.19(a) and Figure 5.19(b), respectively. Under good channel conditions, all users from both VNOs are served with at least the minimum guaranteed data rates defined in Table 5.4, which is also the case under bad channel conditions up to Th_{BE}. However, after this point, under bad channel situations, by increasing the traffic load, capacity is not sufficient to serve all users, therefore, after delaying BE ones from both VNOs, VRRM starts discarding users from the lowest priority BG service according to Table 5.4, i.e., Web from VNO2, until there is no one left at Th_{BG}; subsequently, it is the turn for Fil users from VNO1 to be delayed. It is also notable that Th_{BE} in both VNOs is placed at the same point, the reason being that BE services from both VNOs have an identical range of data rate variations, as well as the same serving weights.

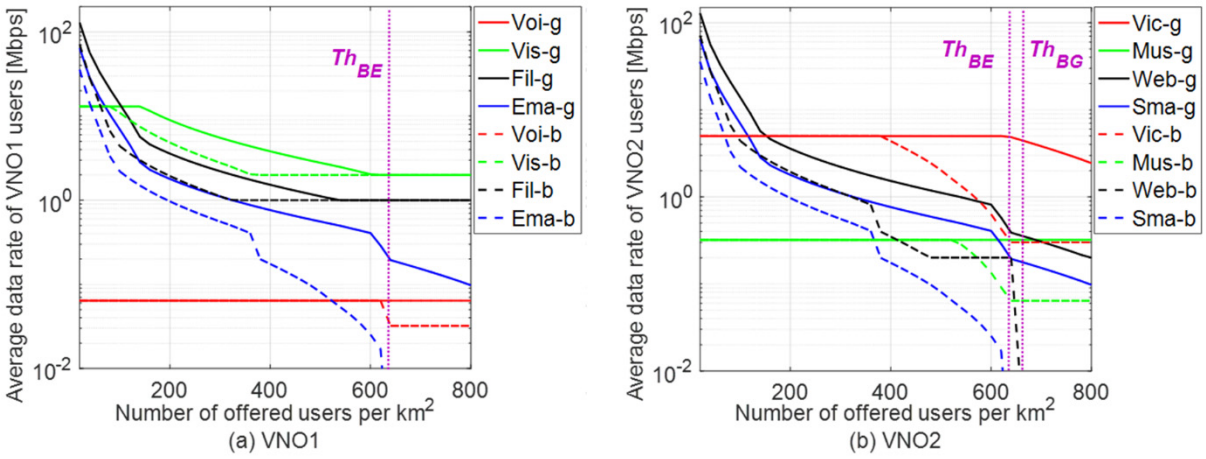


Figure 5.19 – Average data rate of users served by VNO1 and VNO2.

The total data rates of each service slice associated with VNO1 and VNO2 are presented in Figure 5.20(a) and Figure 5.20(b) respectively. First of all, it is obvious that the thresholds are compatible with the ones of Figure 5.19. Furthermore, one can see that the distribution of capacity between the two VNOs is in accordance with the serving weights of Table 5.4, when there is no constraint or capacity

shortage. For instance, when the number of offered users is around 200, in VNO1, the capacity share of Vis is 5 times higher than Fil and 30 times higher than Ema, independent of channel conditions; likewise, in VNO2, Web is 4 times higher than Sma under both good and bad channel conditions. This confirms a level of isolation in the performance of each VNO, according to the specific needs and SLA contracts of that VNO. Another important parameter to be evaluated is the total network throughput in terms of the aggregated utilised capacity. By summing over all values of the total data rates assigned to all the services from both VNOs, the resulting value under each channel condition is almost always equal to the ones of R^{VRRM}, which proves the maximum level of resource utilisation.

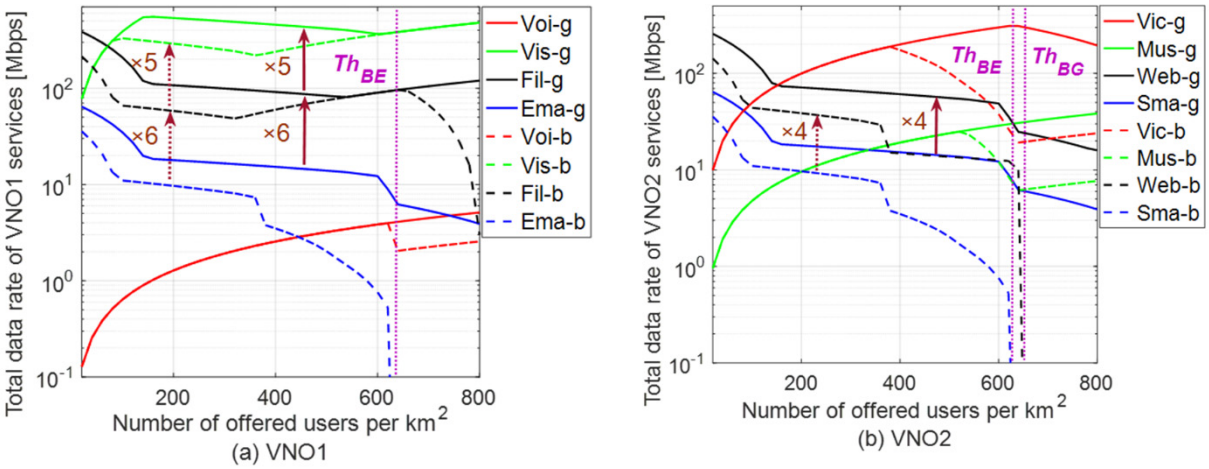


Figure 5.20 – Allocated data rates to each service slice.

The aggregated capacity share of the three SLA categories under the two channel conditions is presented in Figure 5.21. When traffic is low, since all GB users are well-served with the maximum guaranteed data rates, the rest of the capacity is shared among BG and BE ones, satisfying those users with comparatively higher data rates.

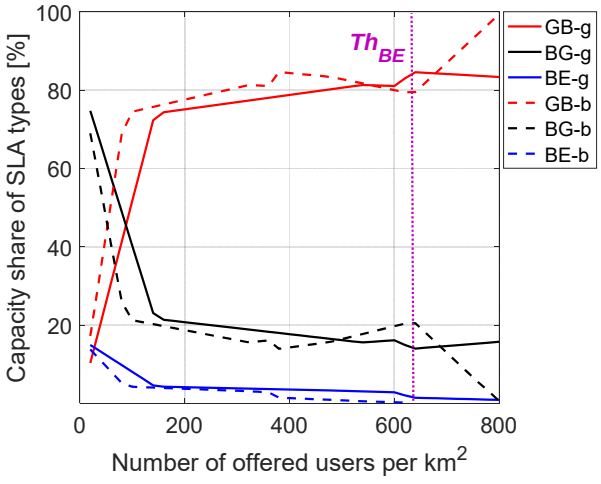


Figure 5.21 – Capacity share among the categorised SLA types.

As the number of users increases, the data rate of GB ones sharply increases, since they have the highest priority among all, which leads to a sudden decrease in both data rates of BG and BE ones. It is also notable that the share of data rate assigned to GB users increases under lower channel quality,

since it is more important to satisfy those users in first place, leading to a reduction in the portion of BG and BE ones. The only exception is around Th_{BE}, mostly because there is still enough capacity to serve *Fil* users with the minimum guaranteed data rates.

5.4 Evaluation of the VNOs' Performance Isolation

Network slicing, which is envisioned for the next generation of mobile networks, aims to enable the deployment of multiple logical networks as *independent* business operations on a common physical infrastructure. In order to analyse the independence and performance isolation of each VNO, a scenario is proposed in this section that defines the same set of services for 3 VNOs, the same user mix per service, and the same ratio of serving weights for all the services that each VNO provides, but with the three different types of SLA contracts, as presented in Table 5.5.

Table 5.5 – Network parameters for three types of VNO SLAs.

VNO#	Service	Class	User mix [%]	δ_s	γ_v	SLA	Contracted capacity [Mbps]
1	Voice	Conversational	30	5	10	GB	[0.4 R^{VRRM}, 0.7 R^{VRRM}]
	Video stream	Streaming	40	4			
	Web	Interactive	10	3			
	File transfer protocol	Background	20	1			
2	Voice	Conversational	30	5	5	BG	[0.4 R^{VRRM},R^{VRRM}]
	Video stream	Streaming	40	4			
	Web	Interactive	10	3			
	File transfer protocol	Background	20	1			
3	Voice	Conversational	30	5	1	BE	[0, R^{VRRM}]
	Video stream	Streaming	40	4			
	Web	Interactive	10	3			
	File transfer protocol	Background	20	1			

As a set of evaluation metrics, the influence of traffic load on the service data rates of each VNO, as well as on the percentage of served users, are studied and the results for the 3 VNOs are presented in Figure 5.22, Figure 5.23 and Figure 5.24, respectively. The *dotted* and *dashed* lines in all subplots represent the *lower* and *higher* thresholds for the total capacity, which can be assigned to each service, according to the customised range of serving data rates proposed in Table 5.5. By increasing the number of users, the total allocated capacity to each service drops to the minimum level. Afterwards, when capacity is not enough to satisfy all users with at least the minimum contracted data rates, VRRM starts delaying users based on their service priority as shown in subplot (b) of the 3 figures. For example, in VNO BG, first *Ftp* data rate is decreased, then *Web*, and in the end some of *Vis* users, while *Voi* is

not affected due to the highest priority among all. It is also notable that the process of delaying BG and GB users starts just after the point when there are no BE users left to be delayed. Since VNO BG has a lower priority compared to VNO GB, the delaying process of each BG service starts before the same GB service.

Regarding the ratio of capacity share among the internal services of each VNO, one can see from all the subplot (a) of the 3 figures that when there is no limitation in data rate assignment, the whole capacity of each VNO is divided among the services proportional to their corresponding values for serving weights (δ_s), while these data rates are also proportional to the values of the same services from the other VNOs, in accordance with the VNOs' weights (γ_v). For instance, when there is no saturation or limitation, the data rate of *Vis* in all three VNOs is 4 times higher than the one of *Ftp*, while at the same time the data rate of *Vis-GB* is exactly 10 times higher than the equivalent service in VNO BE, in order to satisfy the relation between the VNOs' weights. Therefore, the results confirm that not only the internal policy of each VNO is managed separately from the others, but also there is an isolation level in terms of satisfying the agreement between the InP and each VNO.

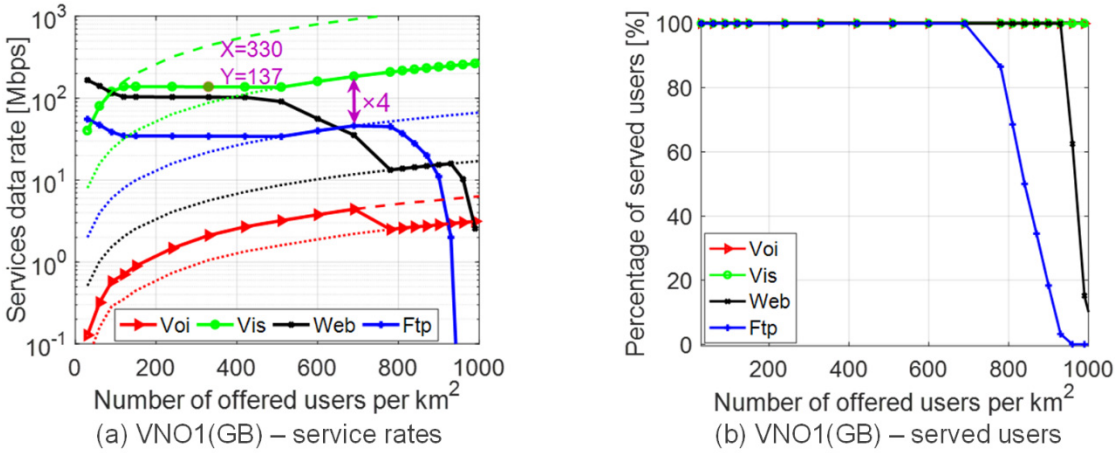


Figure 5.22 – VNO GB's service slices' data rate and percentage of served users.

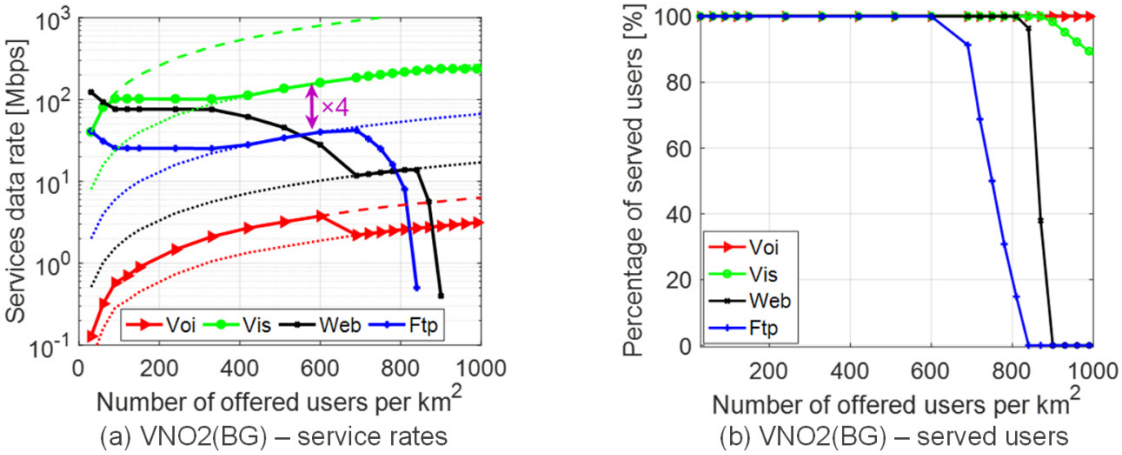


Figure 5.23 – VNO BG's service slices' data rate and percentage of served users.

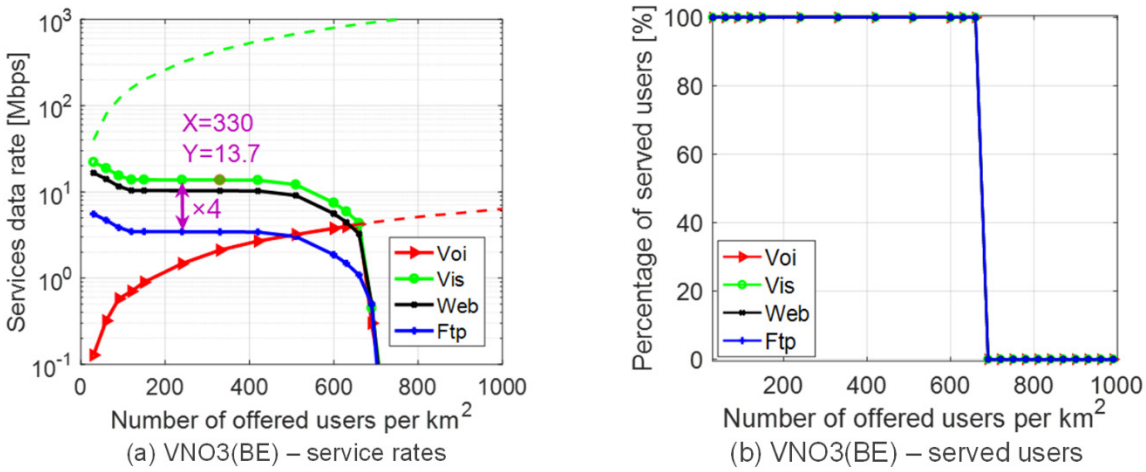


Figure 5.24 – VNO BE's service slices' data rate and percentage of served users.

To study the effect of the tuning weight on the total capacity share of the VNOs, it is assumed that the weight of VNO BG, γ_{BG} varies between the nominated values for VNO BE, γ_{BE} and VNO GB, γ_{GB}, which are stated in Table 5.5 (i.e., 1 to 10). Results for 500 users are shown in Figure 5.25. For the lower values of γ_{BG}, although the first expectation is to have comparable amounts of capacity share with VNO BE (since the weights of the two VNOs are similar), one can see that VRRM preserves the values of γ_{BG} at 40%. This is the minimum guaranteed threshold for VNO BG, which is much higher that the share of VNO BE. By increasing γ_{BG} the capacity share of VNO GB and VNO BG get closer and finally converge at 10, when the same values are assigned to γ_{BG} and γ_{GB}. The results confirm that all VNOs are independently satisfied according to their SLA contracts with InPs, and in respect to the VNOs pre-defined weights, while 100% of the available capacity is shared among VNOs.

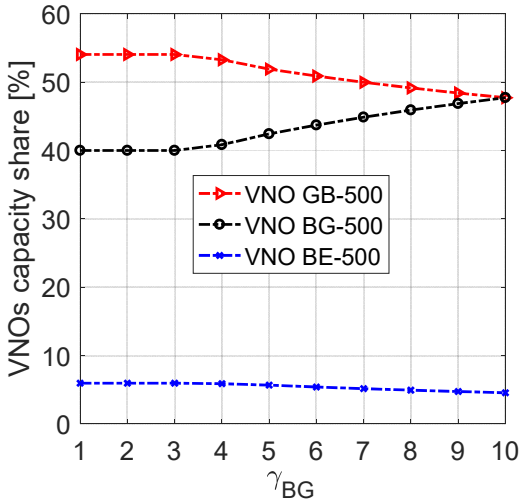


Figure 5.25 – Effect of tuning weights on the VNOs capacity share.

The results that are presented in this section confirm that 100% of the available capacity is used independently of the variation of traffic load and different network parameters. All SLAs are satisfied in an isolated manner, and users are served according to the specific customised range of serving data rates. When there is the possibility that the capacity is *internally* and *globally* shared among services

proportional to the *serving weights* and *VNOs' weights*, respectively, which is in accordance with the criterion of proportional fairness as well.

Chapter 6

Analysis of Results

This chapter aims to analyse the performance of the proposed VRRM and CRRM models in terms of some key aspects. In order to show the benefits of RAN slicing and virtualisation, especially regarding the increased multiplexing gain, first a scenario with non-uniform traffic load is developed in Section 6.1, then following in Section 6.2 the performance of VRRM in terms of addressing the SLAs is evaluated in comparison with the typical RRM deployed in traditional Het-Net MNOs to present the achieved multiplexing gain. Section 6.3 studies the performance of CRRM in terms of InP's policies for RAT selection and load balancing. The analysis of users' throughput and satisfaction are presented in Sections 6.4 and 6.5 respectively. Finally, the performance of the model in terms of satisfying the proportional fairness assumptions is discussed in Section 6.6.

6.1 Simulation Scenario and the Road Map

In order to evaluate the model under non-uniform traffic load and to show the increased multiplexing gain, a case study scenario is developed. Starting from the RATs specifications, it is assumed that $R_j^{RAT_{tot}}$ has a Normal Distribution, $R_j^{RAT_{tot}} \sim N\left(\overline{R_{b_j}}, \sigma_j\right)$, and that the interval for R_j^{RRU} is $R_j^{RRU} \in \left[R_j^{RRU_H}/2\,,\ R_j^{RRU_H}\right]$, then, the distribution parameters associated with each RAT, within a 95% confidence interval, are given in Table 6.1, where $R_j^{RAT_{tot}} \in \left[R_{b_j}^{min}, R_{b_j}^{max}\right]$.

Table 6.1 – Parameters of the Truncated Normal Distributions.

RATs	$R_{b_j[\mathrm{Mbps}]}^{min}$	$R_{b_j[\mathrm{Mbps}]}^{max}$	$\overline{R_{b_j[\mathrm{Mbps}]}}$	$\sigma_{j[\mathrm{Mbps}]}$
OFDM (Wi-Fi)	6704	15408	11056	87.1
OFDMA (LTE)	2400	4800	3655	61.2
CDMA (UMTS)	44.1	88.2	66.2	1.63
TDMA (GSM)	0.62	1.24	0.94	0.053

It is assumed that the area under analysis is uniformly covered by all available RATs and that users' terminals can have access to all RATs. A summary of RATs specifications can be found in Table 6.2, where $R_{RAT_j}^{RRU_{max}}$ is the maximum data rate that a single RRU from different RATs can provide. Nevertheless, these values are not achievable in reality, since the quality of the physical channel is greatly influenced by SINR. Using the proposed convolution PDF to calculate the aggregated capacity, by choosing α_p= 3.8 (the average power decay) as a typical value for urban wireless environments, and then randomly selecting from the convolution PDF, $R_j^{RAT_{tot}}$ as the numerical value for the total aggregated capacity of each RAT averaged over 1 km^2 is obtained. It is further assumed that InP has a policy to provide only 75% of its actual total aggregated capacity, as an upper limit to VRRM, to avoid RAT congestion, while at the same time the individual used capacity of each RAT cannot surpass 80% of the actual total capacity of that specific RAT.

Table 6.2 – RAT specifications (updated from [KhCo14]).

RAT	# BS, AP	RRU	$R_{j[\mathrm{Mbps}]}^{RRU_{max}}$	$R_{j[\mathrm{Mbps}]}^{RAT_{tot}}$
OFDM (Wi-Fi)	32	Carrier	0.97	1114
OFDMA (LTE)	16	Res. block	0.75	350.2
CDMA (UMTS)	~ 1.7	Code	1.4	6.54
TDMA (GSM)	1	Time-slot	0.059	0.10

Regarding the distribution of users, it is supposed that the area under analysis is divided into two regions: *residential* and *business* areas. Considering the real traffic pattern of mobile users during a period of 24 hours, which is studied in [XLWZ17] for the urban environment of Shanghai, the provided data regarding the normalised traffic profiles of residential and office areas are extracted and adapted to match the assumptions of this work. Users in each area are being served by the VNO associated with that region. Figure 6.1 represents the average number of users in each area, at different times of the day, together with the aggregated traffic in both areas.

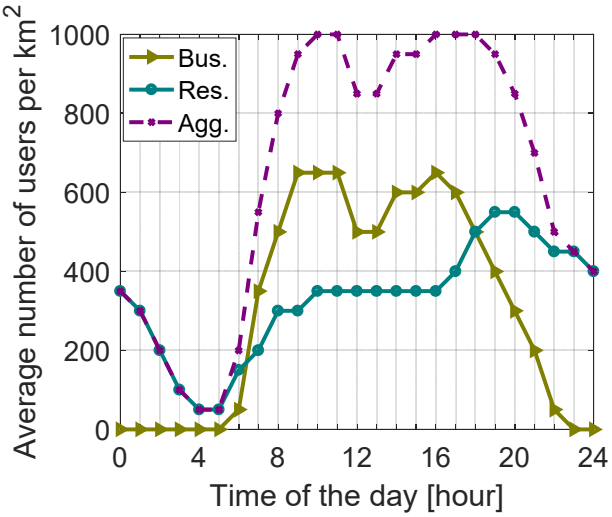


Figure 6.1 – Average traffic load of residential and business areas.

Concerning network parameters, the specifications are presented in Table 6.3. It is assumed that both VNOs are categorised as a GB SLA, being in charge of providing the four service classes: Conversational, Streaming, Interactive and Background. Service types and their assigned weights, as well as user mix of each service $U^{srv}_{[\%]}$, are given the same value for both VNOs, in order to provide an easier way of making a comparison between their performance efficiency. $R^{usr}_{v_s}$ represents the users' acceptable range of data rate variation, per service of each VNO. The provided services are assumed to be *Voi, Vis, Web* and *Ftp*.

Table 6.3 – Service parameters.

Service	**Class**	$R^{usr}_{v_s\,[\mathbf{Mbps}]}$	$U^{srv}_{[\%]}$	δ_s	$R^{vno}_{v\,[\mathbf{Mbps}]}$
Voice	Con.	[0.032, 0.064]	30	5	[200, 800]
Video	Str.	[2, 10]	40	4	
Web	Int.	[0.5, R^{VRRM}]	10	3	
Ftp	Bac	[1, R^{VRRM}]	20	1	

As mentioned before, the two VNOs do not own the physical RATs and are sharing a pool of aggregated radio resources, provided by the InP. In contrast, as a sake of comparison between the performance efficiency of the proposed VRRM model for virtual RANs and RRM in typical MNOs, it is also assumed that the same RATs specifications, traffic load and set of services are applied to a practical Het-Net

scenario with two typical operators (residential and business). Each MNO owns half of the available dedicated bandwidth, while serving the users in the two areas happens separately from each other.

Concerning the InP policies for service-to-RAT mapping, including RAT suitability and load balancing, it is assumed that for the InP it is desirable to have an *equally loaded* group of RATs (in percentage) as much as possible, therefore the load balancing factor of each RAT, δ_j^{RAT}, has to be *proportional* to its capacity. However, RATs suitability factor for each service, $\mu_{v_s}^{RAT}$, is another important parameter for decision making in CRRM (which can potentially reduce the signalling overheads). These prioritised weights are specified in Table 6.4, where the most preferred RAT for each service is given the highest weight and NA represents the case when it is not feasible to associate a particular service to a specific RAT. It is also assumed that InP has *shuttled down* the UMTS cells for more energy efficiency or system upgrade (therefore the proposed UMTS weights are not effective while this RAT is unavailable).

Table 6.4 – InP's service based policy for prioritised RAT selection.

Services	$\mu_{v_s}^{RAT}$			
	TDM (GSM)	**CDMA (UMTS)**	**OFDMA (LTE)**	**OFDM (Wi-Fi)**
Voice	10	8	2	NA
Video	NA	5	10	8
Web	NA	4	5	7
Ftp	NA	3	4	7

6.2 Comparison of VNOs Multiplexing Gain with MNOs

The performance of the model in terms of different evaluation metrics is presented and analysed in this section. Figure 6.2 represents the capacity share of VRRM between business and residential VNOs. Since the maximum SLA threshold of capacity allocation is 800 Mbps (according to Table 6.3), even when there is no user to be served in the *residential area*, VRRM will not assign more than that to the VNO residential, which is the reason why from 23h00 to 5h00, the bandwidth share of VNO residential is fixed to 800 Mbps (around 72% of the available VRRM capacity). This does not mean that the unallocated VRRM capacity is simply wasted, in contrast to the Het-Net MNOs, since the resources are not dedicated to VRRM and are provided *on-demand*. Therefore, the InP may use this capacity to serve other tenants. One can also realise that capacity is flexibly shared between VNOs: each VNO based on their traffic pattern receives a higher share during the traffic peak time, while in contrast, for the MNOs in this scenario the capacity is not set based on the demands, and therefore remains always as a dedicated fixed amount of 50% share.

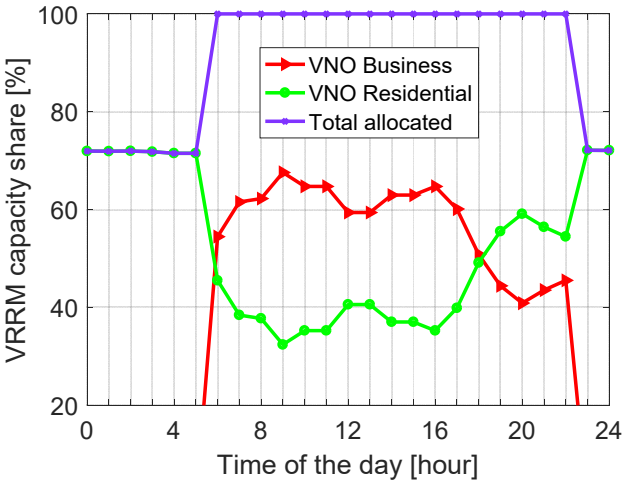


Figure 6.2 – Capacity share of VRRM among the two VNOs.

In order to study the performance of VRRM in more detail, the numerical results of resource allocation to different services of the residential and business VNOs are compared to the ones of MNOs and presented in Figure 6.3. The *dotted* and *dashed* lines in all subplots accordingly represent the *minimum* and *maximum* capacity that can be assigned to the services according to the customised range of data rates defined by the VNOs or Het-Net operators in Table 6.3. As one can see, from late night until early in the morning (roughly from 23h00 to 5h00), when there is no traffic in the business area, the serving data rates in the residential area are higher for the associated VNO, compared to the equivalent MNO. This is due to the fact that VRRM, as a central management entity, has control over the aggregated virtualised pool of radio resources, and is also aware of the demands that the traffic profile implies in the both areas, while each of the Het-Net MNOs have access to half of the physical resources, which are dedicated. This means that when there is no user in the business area the associated bandwidth of the business operator is *wasted*, while the residential operator cannot use this extra capacity.

Regarding the ratio of capacity share among the internal services of each VNO or MNO, it is notable that when there is no limitation in data rate assignment, the whole capacity of each one is divided among the services *proportional* to their corresponding values for serving weights (δ_s). For instance, from 0h00 until 6h00, the assigned data rate to Web in VNO residential is exactly three times higher than Ftp.

On the other hand, during the traffic *peak time* of the business area (roughly from 8h00 to 18h00), the available capacity of business MNO is not enough to serve all users with at least the minimum acceptable data rates, therefore, according to Figure 6.3(c) and Figure 6.4(a), some of the low priority users (Web and Ftp) are being delayed so that the rest of higher priority ones can just continue their services with the minimum satisfaction level. However, thanks to the flexibility of VRRM, business VNO can perfectly handle the extreme situation without any capacity shortage, since VRRM has set the data rate of residential VNO to the lower, still acceptable, level, compared to the residential MNO. The same procedure occurs for the peak traffic time of the residential area (around 19h00 to 20h00), while the associated MNO has to delay some of the Ftp users, the capacity is enough for the residential VNO to serve them all.

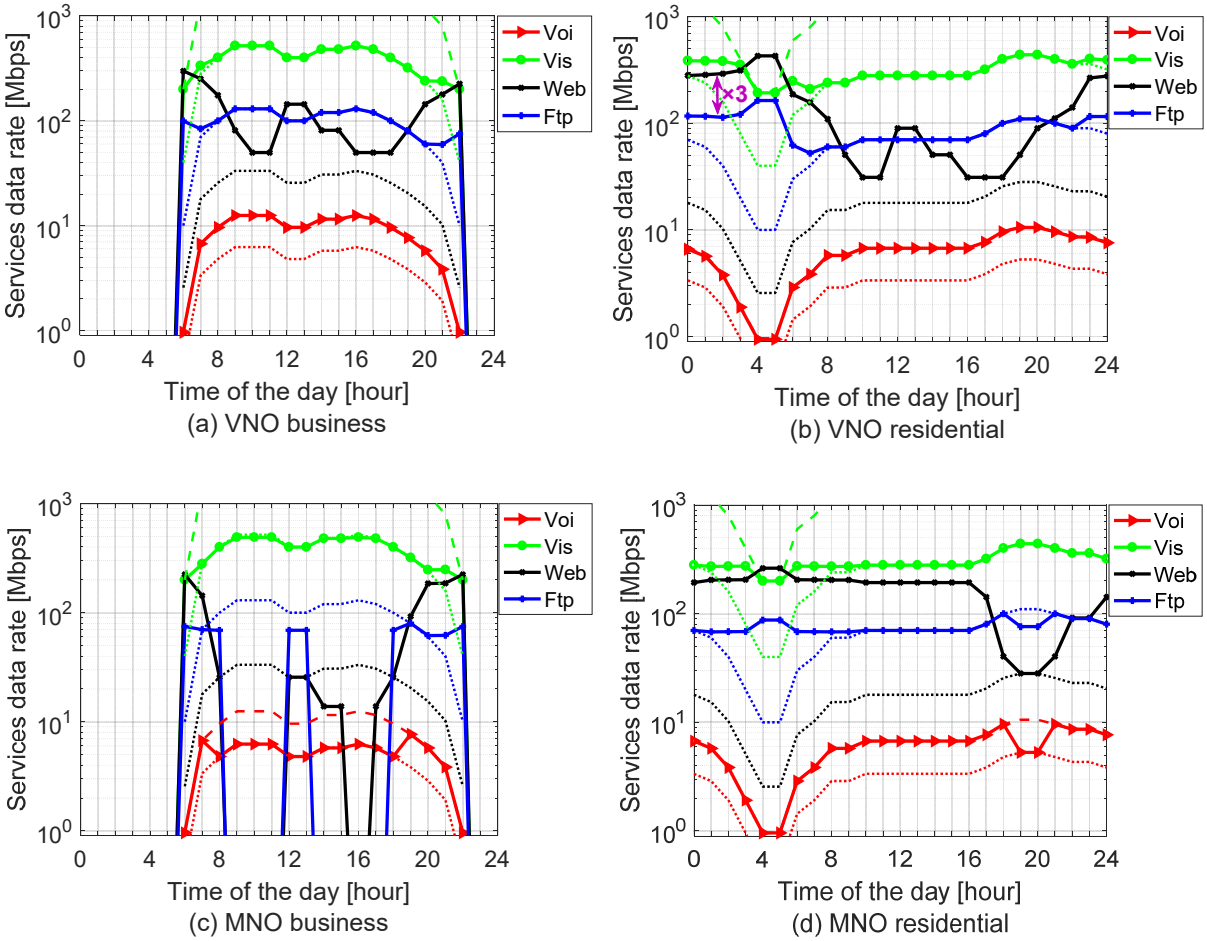


Figure 6.3 – Capacity share of the services provided by the residential and business VNOs/MNOs during 24 hours.

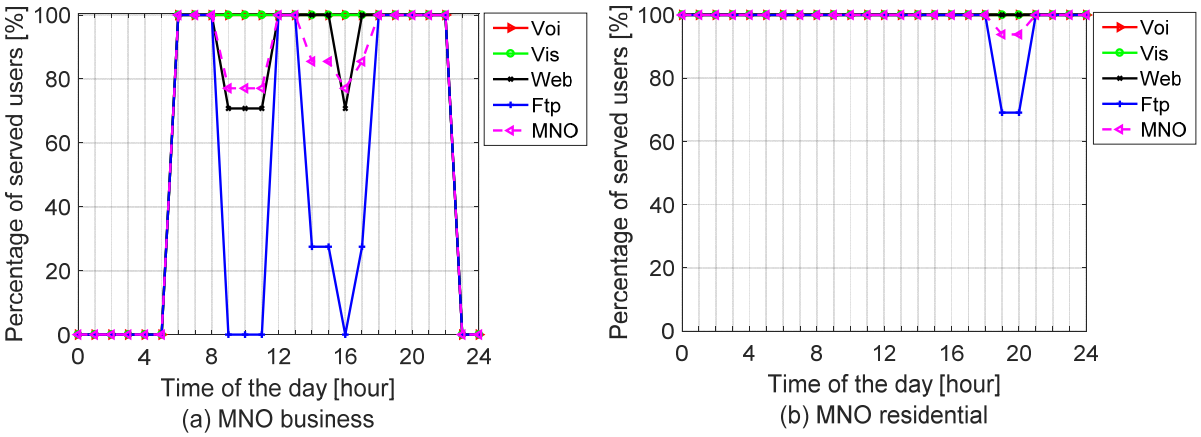


Figure 6.4 – Percentage of served users in each Het-Net MNO.

6.3 Evaluation of Service-to-RAT Assignment of CRRM

Regarding the performance efficiency of CRRM in the service-to-RAT assignment and load balancing, the procedure of allocating capacity from different RATs, to satisfy the demanded data rates of VRRM, is shown in Figure 6.5. First of all, since the InP has shut down UMTS, all the traffic load of this RAT is pushed to the other suitable RATs and according to the assumptions of Table 6.4, GSM is dedicated to Voice only, and this service is not supposed to be served by Wi-Fi.

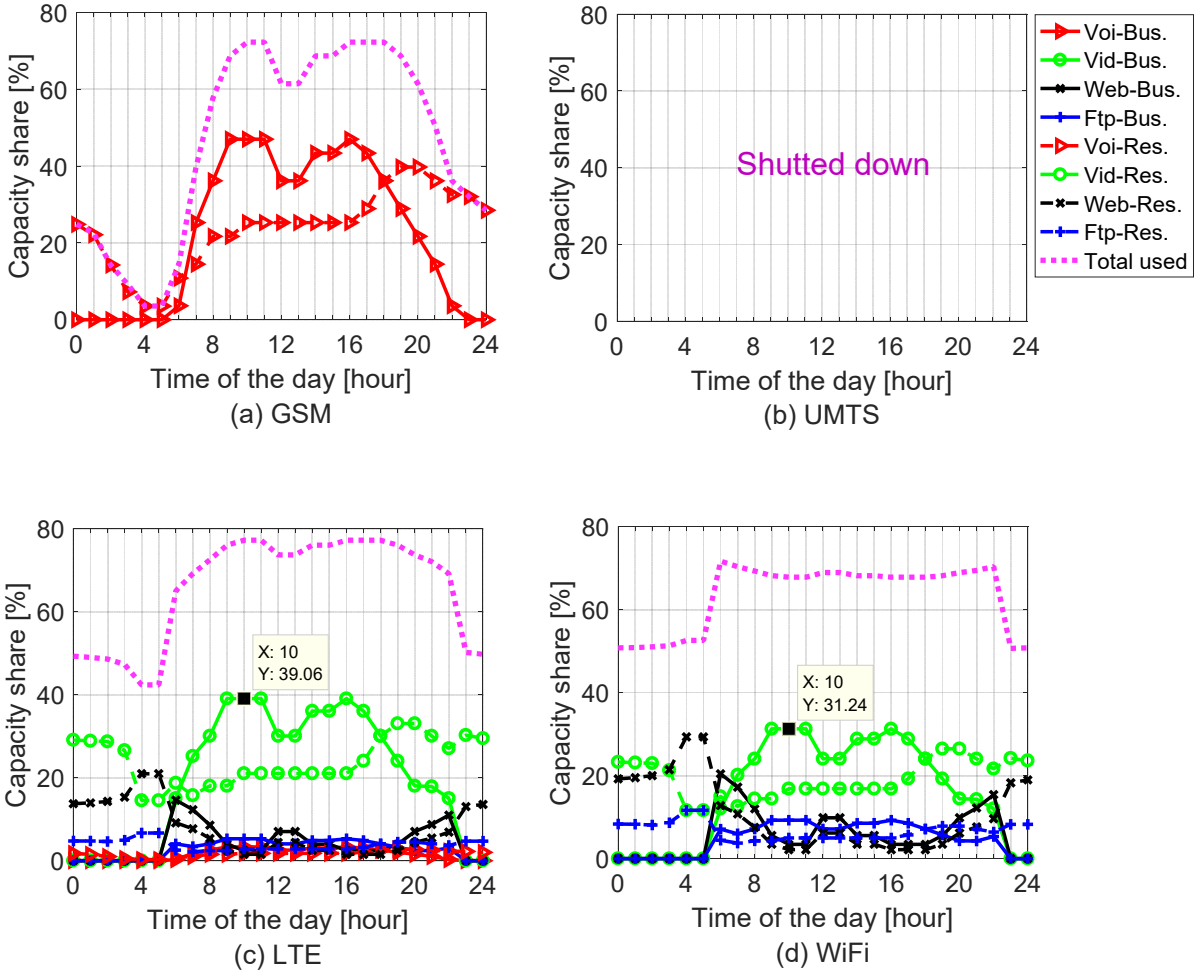


Figure 6.5 – Capacity assignment from different RATs to the service demands of the VNOs.

Concerning the effect of load balancing factor, δ_j^{RAT}, if the InP does not have any preference in RAT suitability (i.e., all the values of $\mu_{v_s}^{RAT}$ where equal), since the values of δ_j^{RAT} are *proportional* to the capacity of each RAT, it would have been expected to experience exactly the same level of load (in percentage) for each RAT. For instance, when 100% of the VRRMs' capacity is used (from 6h00 to 22h00), all RATs should have been equally loaded as 75%, since according to the InP's policy, the maximum capacity provided to VRRM is 75% of the aggregated available capacity of CRRM. However, as $\mu_{v_s}^{RAT}$ is different for each service, the loaded level of each RAT varies accordingly. As an example, considering the values of $\mu_{v_s}^{RAT}$ from Table 6.4 for the mapping of *Video* demands to Wi-Fi and LTE, the ratio of the suitability weights is 0.8. By comparing the associated values from Figure 6.5(c) and

Figure 6.5(d), one can clearly see that the share of provided demands between the two RATs is proportional to the ratio of $\mu_{v_s}^{RAT}$factors for Video, which is the case for both business and residential VNOs. It is also notable that the load of all the RATs remain less than 80%, which is the maximum nominal value set by the InP.

6.4 Analysis of Average Users' Throughput in VNOs and MNOs

The average data rate of users served by VNOs and MNOs are shown in Figure 6.6. By comparing the performance of VNO and MNO business, one can notice that when the traffic load is light (roughly from 19h00 to 7h00), the allocation of date rate follows almost the same pattern since both VNO and MNO are capable of addressing the whole demand. However, the data rates of lower priority services, i.e., Web and Ftp, are served with a higher rate in the VNO compared to the MNO. This is also the case for the VNO and MNO residential, especially from 23h00 to 5h00, since there is no user in the business area and more resources are available to be assigned to VNO residential.

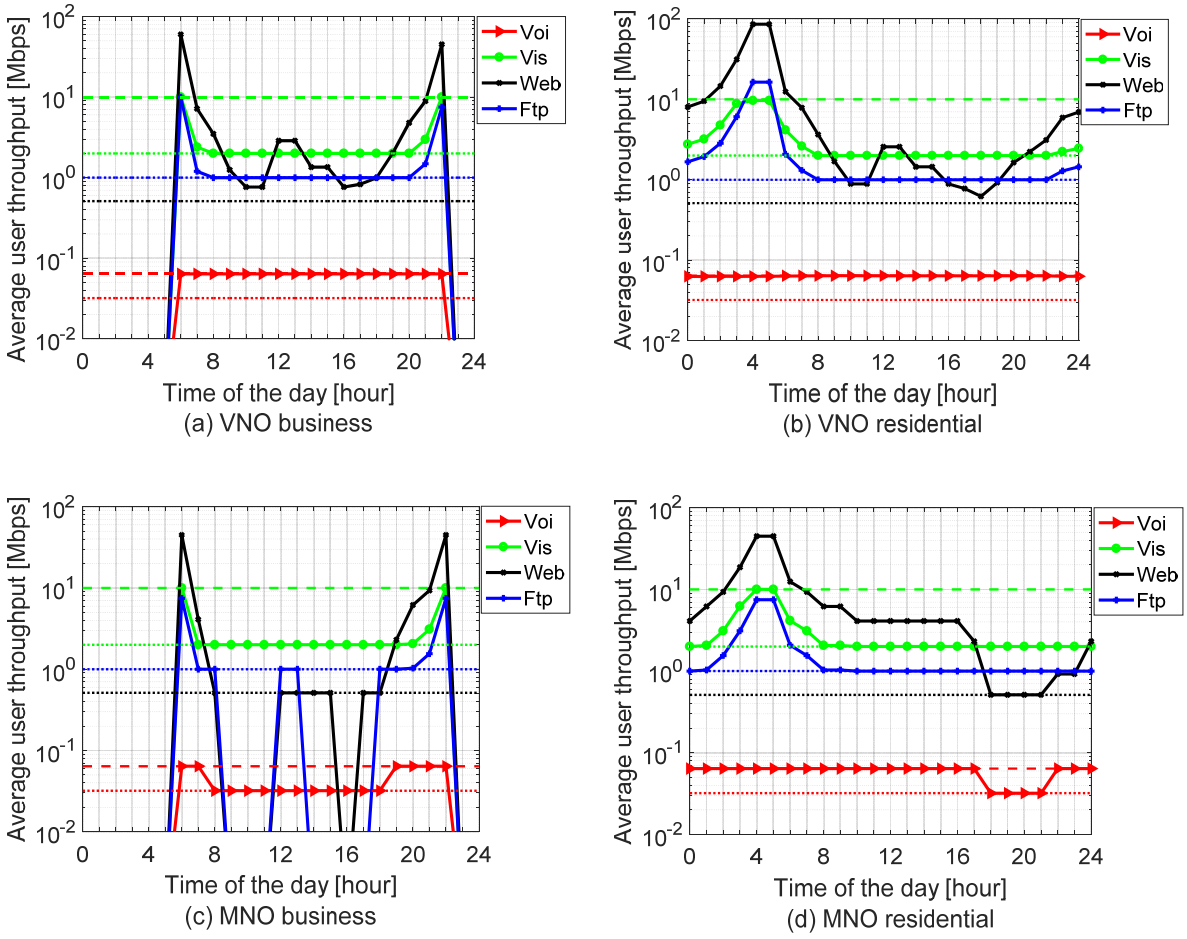


Figure 6.6 – Average users' data rate being served by business and residential VNOs/MNOs.

When it comes to the working hours of business areas (roughly from 8h00 to 19h00), the traffic load gets comparatively heavier. By comparing the performance of the VNO and MNO business, as explained

in Section 6.2.1, it is obvious that the MNO cannot satisfy all demands and it has to delay lower priority users in order to be able to serve the rest of higher priority ones with the minimum acceptable data rates. However, by looking at the performance of VNO and MNO residential in this period, one can notice that MNO residential allocates much higher data rate to Web users, while MNO business users are suffering from capacity shortage, whereas VNO residential keeps serving users with almost the same pattern and data rates to maintain an efficient way of resource sharing and achieving maximised multiplexing gain.

6.5 Comparison of Users' Satisfaction in VNOs and MNOs

Regarding the users' satisfaction ratio, the results are shown in Figure 6.7. First of all, since the satisfaction ratio has not been defined for background, Ftp users are not shown here (the Ftp service is performed automatically by an Ftp client in the background [RBR14]). The minimum acceptable level of satisfaction for each service is defined with a dashed line. The zero satisfaction in active hours represents the case when there is at least one service request at that specific hour that has not been served.

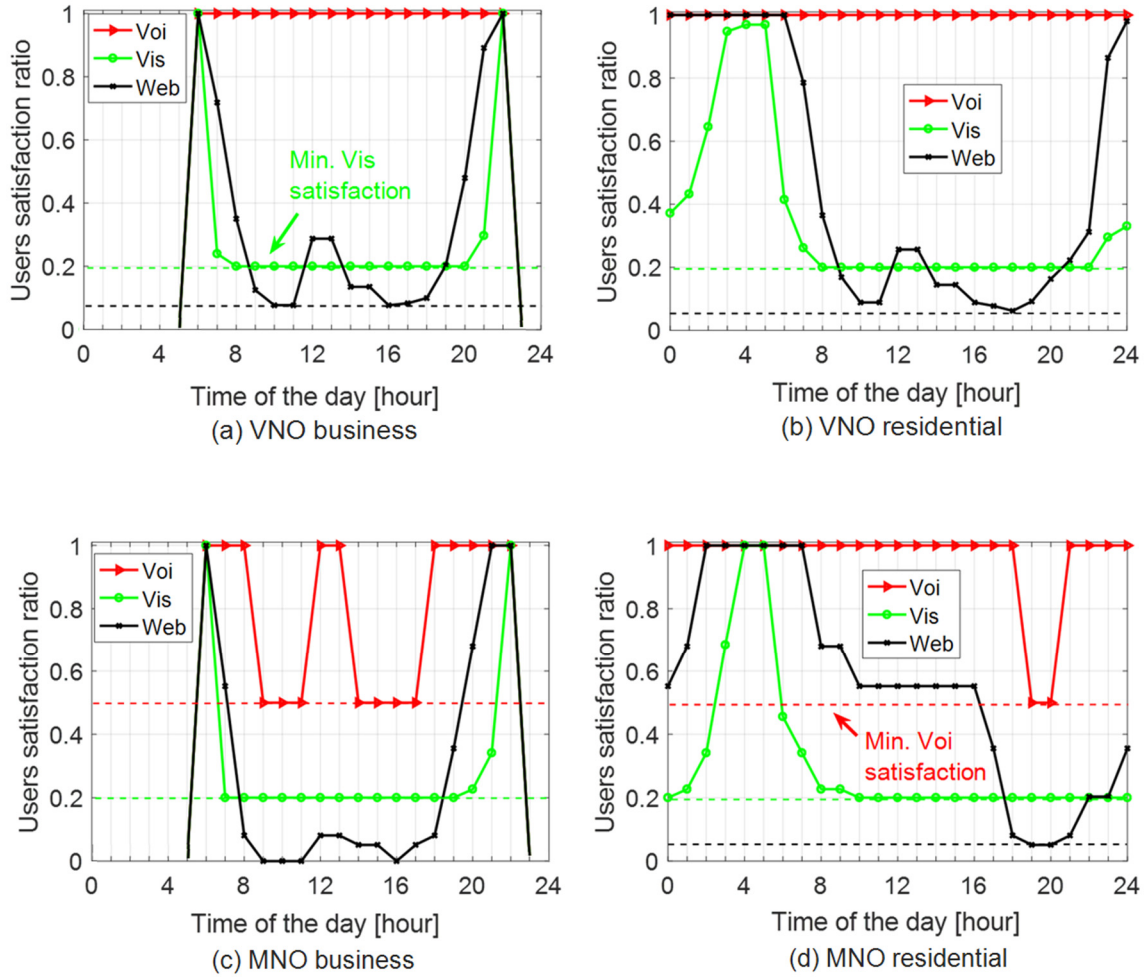


Figure 6.7 – Users satisfaction ratio in the business and residential areas.

By comparing the performance of the corresponding business VNO and MNO from Figure 6.7(a) and Figure 6.7(c), respectively, it is noticeable that some of the MNO Web users suffer and therefore their satisfaction is less than the minimum required and even sometimes in the peak hours drop to zero in order to keep serving the Voi users with their minimum satisfaction level, while the VNO business is capable of satisfying all served users with at least their minimum level of satisfaction and always keeping the satisfaction of Voi users as the highest priority one, on the maximum achievable level.

This is the case for VNO residential as well, meaning that all users are satisfied and Voi satisfaction level remains as 1 during a 24 hours period, while during the peak hours of MNO residential (19h00 to 20h00) Voi users drop to their minimum. This is a consequence of inefficient resource allocation, since in the peak hours of MNO business, when there is a capacity shortage, the satisfaction of MNO residential users is generally higher and vice versa, when there is a capacity shortage in MNO residential, on average the satisfaction level of MNO business users is higher. On the other hand, by comparing the overall pattern of users' satisfaction in both VNOs, it is apparent that the values are close to each other (except for the inactive period of VNO business), since VRRM is aware of both traffic loads and therefore decisions are made in a much more efficient way.

6.6 Analysis of Proportional Fairness in VNOs

As mentioned in Chapter 3, maximising the utilisation of the aggregated capacity and keeping a level of fairness are two contradictory goals. Since maximising the utilisation of limited radio resources is more important, VRRM always tries to assign 100% of the available capacity provided by CRRM on-demand, and if there is a possibility, it assigns the data rates proportional to the serving weights. In this case, the allocated data rates in terms of satisfying proportional fairness is of course highly dependent on the proposed scenario, including the values of serving weights and traffic loads, however, it is worth to compare the performance of VRRM in this regard, when the serving weights of VNOs are the same. Therefore, in this section, the proportional fairness index is used as the evaluation metric to compare the fairness of allocated capacity to services slices, in the business and residential VNOs and the results are presented in Figure 6.8. First of all, the values are compared with Ftp as the reference, and all vary between 0 and 1. The values that are closer to zero are fairer compared to the ones that are closer to 1, and having a zero value for fairness index means that the portion of capacity assigned to a specific service is exactly proportional to the values of Ftp service slice according to their serving weights.

In both subplots of Figure 6.8 it is noticeable that the highest unfair value is Voi. The reason is that this service has the highest priority and accordingly the highest serving weight, however, on the other hand the maximum acceptable data rate of Voi is 64 kbps, which is a very low value, compared to the other services. Therefore, VRRM keeps the average data rate of Voi users up to 64 kbps, since assigning higher data rates will be a waste of resources. The most extreme case occurs in VNO residential around 4h00 to 5h00, since in this period traffic is very light. As a result, Ftp and Web services receive higher data rates, while Voi and Vis that have a limitation on upper bound cannot receive more than the

maximum threshold values. The lowest overall fairness index values in VNO residential happens around 2h00 and roughly from 6h00 to 8h00, the reason being that except Voi, all the other 3 service slices in these periods are not constrained by the limitation on minimum or maximum thresholds of data rates and VRRM is assigned the bandwidth to the other 3 services proportional to their serving weights.

In VNO business, during the working hours, which is roughly from 8h00 to 19h00 all services except Web are already constrained, and therefore the most variation in allocated data rates comes from Web service slice. Since the data rate of Web in this period is always less than Ftp (although it has a higher priority), except from 12h00 to 13h00, the fairness index has a higher value during the working period, except around mid-day, when the network experience lower traffic in the business area. Overall, it can be concluded that, since VRRM has a global vision of both areas, the values of fairness index are also comparable, except for specific peak points when the traffic in one area is too light or intense.

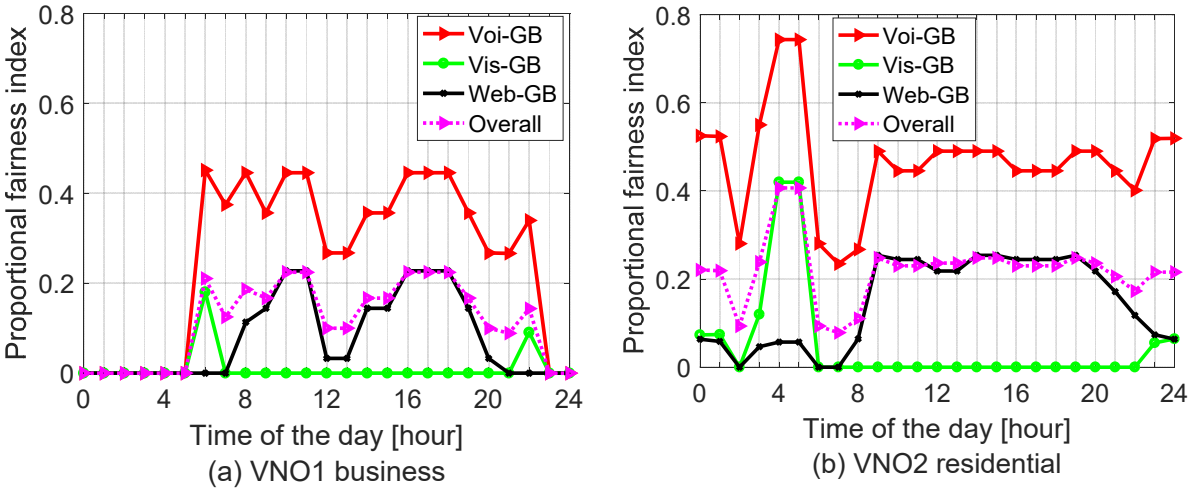


Figure 6.8 – Variation of proportional fairness index in VNO business and residential.

Chapter 7

Conclusions

This chapter aims at presenting the key outcomes and conclusions of this thesis, highlighting the main results, as well as suggesting some research directions for further work.

7.1 Overview of the Thesis

The current thesis is comprised of 7 chapters. It starts with Chapter 1, which gives a brief historical overview of the evolution of wireless technologies, as well as an introduction to the thesis, including the motivation and key objectives. The main goal is to realise a separation in the roles of VNOs and InPs. In this regard, the functionalities related to service orchestration such as satisfying the contracted SLAs and customising radio bearers are managed by VRRM, inside SDAP sublayer of 5G New Radio, while mapping the service demands to be addressed by the underlying RATs is performed by CRRM, which is an individual management entity and controlled by InPs. The key contributions of the author in terms of international publications and European project contributions are also provided in this chapter, which are summarised as follows:

- 1 journal paper has been published and 1 is currently under the second round of review,
- 4 conference papers have been presented and published,
- 7 contributions in the form of TD is presented in the technical meetings of the IRACON project.

Chapter 2 presents some basic concepts and topics, which form the fundamentals of this thesis. The main focus is on the key principles of network virtualisation and on-demand RAN slicing, including the framework of wireless network virtualisation, models and requirements and RRM among service slices. To conclude this chapter, the state of the art of recent strategies for service based RRM in virtual wireless networks is also given in the last part of this chapter.

Chapter 3 presents a model of RAN slicing and resource management for emerging virtual wireless networks, based on the interaction between two separated management entities in the proposed network architecture from a high-level perspective: CRRM is in charge of coordination of radio resources among the available RATs, while a centralised virtualisation platform on top of it, called VRRM, is responsible for service orchestration among VNOs, enabling the definition of various services and policies, separately from vendors and underlying RATs. The main objective of VRRM is to satisfy the SLAs associated with different service classes to the highest possible level, within the framework of proportional fairness. On the other hand, CRRM is in charge of mapping the demanded capacity of each service onto the most suitable RATs according to the policies of InPs for RAT selection, load balancing and energy efficiency. Both VRRM and CRRM analytical models are formulated based on solving constrained convex optimisation problems, considering various design parameters including human-based behaviours, network requirements and translation of policies for different network entities. The performance of the proposed model is also assessed in terms of key evaluation metrics, such as network throughput, capacity share of service slices, percentage of delayed users and distribution of service demands among the underlying RATs in a practical multi-RAT scenario.

Chapter 4 presents the details of the model and algorithms implementation in the simulator. In this regard, the processing and functional blocks of the model are categorised in 3 different parts. The first one is capacity aggregation module, which is implemented in MATLAB to obtain the total aggregated capacity of the network through the convolution of all the RRUs' PDFs. The second one is the VRRM unit which is defined and solved inside CVX solver, together with admission control and delay process

to ensure the feasibility of the VRRM optimisation problem. The last one is the CRRM component, which is also solved in CVX and is responsible for redirecting the demanded service traffics among the available RATs. CRRM optimisation provides the best solution of service-to-RAT mapping, based on InPs policies such as RAT suitability and energy efficiency.

A reference scenario with some assumptions is presented in Chapter 5 to evaluate the performance of the proposed model in terms of the provisioned goals. The numeric results obtained from the simulation shows that using the concept of service based RAN slicing, multiple VNOs with customised service requirements and agreements are effectively served on a common shared physical infrastructure, which is managed separately by InPs.

In order to prove the efficiency of the proposed model in comparison to typical Het-Net operators, a scenario with 2 types of VNOs and MNOs (business and residential), with different traffic patterns associated with each type is presented in Chapter 6. The analysis of results confirms the effectiveness of the proposed model specially in terms of achieving statistical multiplexing gain.

Finally, the current chapter gives a summary of the thesis, the main results and achievements, and the novelty of this work. Some directions for future work are also provided.

7.2 Main Results

This thesis proposes an SLA-based model for RRM in virtual RANs. While service management and resource slicing is performed in a centralised virtualisation platform called VRRM, the whole aggregated capacity is shared on-demand among the VNOs, considering their customised internal policies and the type of SLA contracts in order to realise the concept of performance isolation between the tenants, which share the same infrastructure and to minimise the effect of variation in different network parameters on the guaranteed level of bandwidth allocation to different VNOs. The algorithms of slicing and data rate allocation are based on prioritisation of services and network pricing, which is defined according to the concept of weighted proportional fairness definition and it also includes an admission control policy to guarantee the required SLAs, being a common approach for service-based RAN slicing. Moreover, in the extreme situation when network capacity is not enough to satisfy all users with at least the minimum guaranteed service level, delay is introduced to some of the low priority users, releasing the required capacity to serve the remaining ones with the minimum acceptable data rate.

Furthermore, in order to realise the separation in the role of InPs and VNOs, the task of mapping the demanded capacity from different virtual slices onto the underlying physical RATs is defined in CRRM, in order to promote the notion of end-to-end slicing. To accommodate this function, a mechanism of cooperation between the two entities (i.e., VRRM and CRRM) is also proposed to define which information has to be exchanged in order to achieve an efficient interaction, while keeping a desired level of isolation.

The model is capable of supporting different situations. The performance of VRRM is evaluated in terms of the effect of variation in traffic load, on the capacity share of different VNOs with 3 types of SLA agreements (GB, BG and BE). Results show that under lower traffic loads, all the service demands of GB slices are addressed with the highest achievable level. In this case, if the minimum GB SLA is higher than the maximum demanded data rate, the minimum threshold will be adapted to the maximum demand, in order to avoid wasting the resources, while the rest of capacity will be shared among BG and BE service slices. Under higher traffic loads, when there is a capacity shortage, all the BE services are being delayed and the capacity share of VNO BG will be reduced to the minimum contracted, which is 30% of the total available VRRM capacity, while the rest of capacity will be allocated to VNO GB since it has the highest priority. In addition, the capacity share of service slices inside each VNO is proportional to the serving weights if there is no constraint. For instance, the total assigned data rate to Fil, Web and Soc in VNO BG under low traffic loads are proportional to 4, 3 and 2. This shows that all the SLAs are satisfied and all the service demands are addressed to the highest achievable level, according to the internal policies of existing VNOs for capacity sharing.

The results also show the effect of variation in SLA and serving weights to verify the efficiency of the VRRM model in terms of fulfilling the service demands. By varying the minimum BG SLA from 10% to 60% of the total VRRM capacity, when the number of users is fixed as 300, one can see that the capacity share of VNO BG also increases up to 60%, which is the extreme case, since the minimum GB SLA is set to 40%. At this point, all the BE services are being delayed so that the higher priority VNOs are satisfied with the minimum contracted SLAs. A similar behaviour can be observed by scaling the serving weights of VNO BG from 1 to 12, meaning that by increasing the serving weights the capacity share of VNO BG increases compared to the other two VNOs up to the point that the capacity share of VNO GB reaches down to 40% and all the BE users are delayed. After this point, increasing the weights of VNO BG will not have any effect on the amount of assigned capacity, as well as on other parameters such as the percentage of served users or satisfaction level.

Another outcome of the model is the effect of number of active RATs on SLA assurance of VRRM to find out the minimum number of available RATs that provide enough aggregated capacity for a specific network traffic load to address all the service demands, since this parameter is important for energy saving, maintenance, or system upgrade. In this regard, when the traffic load is increased and the capacity share of VNO GB gets close to the minimum acceptable level, 4 pairs of LTE and Wi-Fi RATs are activated to avoid the unnecessary delay, until the full capacity (16 pairs of LTE and Wi-Fi RATs) is reached. Furthermore, from the CRRM viewpoint, both LTE and Wi-Fi RATs remain almost equally loaded under the different intensity of offered traffic load, and always less than 80%, which is the nominal value set by InP.

In addition, the influence of variation in channel quality conditions on the performance of the model is studied for 2 different types of good and bad channels. Obviously the data rates for each VNO under worse channel quality is less, however, still under both channel conditions when there is no constraint and the capacity is enough to serve all the users, independent from the channel quality, the capacity is shared among the services proportional to the customised serving weights of each VNO and separated

from each other, to prove that the channel quality does not affect the general policy of RRM in each VNO. Furthermore, when there is a shortage of capacity under bad channel condition, still the lowest priority services are affected so that the VNO continues satisfying the higher priority ones with the minimum satisfaction level. It is also notable that under bad channel conditions the percentage of capacity share for GB service slices generally increases compared to the good channel condition, while for the BG and BE ones decreases, since the former service type has a higher priority.

All the presented results prove a level of performance isolation in achieving the desired SLAs and addressing the individual policies of VNOs in a multi-tenant environment. This is performed by VRRM among the involved VNOs, while different network parameters change. In all cases when there is enough demand, 100% of the available VRRM capacity is allocated to satisfy the service demands to the highest achievable level and when there is a possibility the capacity shared among the service slices according to the concept of proportional fairness.

Regarding the results of CRRM performance evaluation for mapping the service demands to the underlying RATs according to the InP's policy for RAT selection and load balancing in the reference scenario, one can notice that although the load of available RATs changes by variation in traffic load, the aggregation of the total load of existing RATs remain almost always around 70%, which is the threshold of available capacity provided from CRRM to VRRM. It is also evident that the load of individual RATs will never surpass 80% of its total aggregated capacity, which is also an upper limit put by InP. Furthermore, the share of assigned capacity from different RATs to a specific service slice is proportional to the service-to-RAT suitability factor.

The performance of the proposed model is further evaluated in comparison with typical Het-Nets. In this regard, two groups of users who are located in residential and business areas are supposed to be served by the associated VNOs and typical MNOs during a period of 24 hours. Results from VRRM shows the efficiency of the model on achieving the so-called statistical multiplexing gain in comparison to traditional Het-Net operators with dedicated radio resources. Therefore, under higher traffic loads, MNOs have to delay some of their users, while the VNOs' users in both areas are being served in the same situation, thanks to the flexibility of VRRM. Results also confirm that when the VRRM demands for the full available capacity (75% of the total CRRM capacity) the variation of the load in all RATs is around 75% most of the times, while never passes 80%, which is the limitation put by InP.

7.3 Novelty and Key Contributions

The novelty of this work is in the proposal of a model for on-demand, RAN-as-a-Service concept, which deals with the high-level management of virtual RAN, considering the framework of network slicing according to the specific customised service requirements in a multi-tenant environment, which includes prioritisation of services and network pricing, in addition to an admission control policy to guarantee the required SLAs. This has been achieved by defining two separated, centralised management entities

based on a conceptual hierarchical architecture which has been introduced for virtual RAN, in order to address some of the key challenges associated with RRM.

The management entities, VRRM and CRRM, are analytically modelled as two constrained convex optimisation problems with logarithmic objective functions to cope with the framework of proportional fairness. The first impression comes from economic models for pricing based on optimising an objective function with the purpose of balancing network throughput and users' fairness as two competing interests. The internal VNOs' policies and the SLA contracts are defined as sets of constraints for the VRRM objective function, while the capacity limitations of the underlying RATs and InPs policies are specified in the set of CRRM constraints.

The main specifications which are considered in the model include the provision of customised service requirements in resource slicing by applying the individual policies of VNOs to realise a level of performance isolation among them, as well as independence in decision making for service orchestration in higher levels, from the network management perspective of InPs in the lower level physical parts, which includes their traditional role of RAT selection and load balancing. At the same time, there is an effort to improve and adapt some other parameters, such as developing a more precise definition and framework for fairness or reducing the complexity of the proposed model compared to the previous studies. Therefore, various key parameters are taken into account for designing the model, which has not been thoroughly addressed previously in the scale of this thesis work.

Moreover, to realise the idea of separation between the roles of InPs and VNOs, as well as independence in decision making for resource management, CRRM as the entity that is controlled by InP, is in charge of translating and mapping VRRMs' demands onto physical RRUs from different RATs according to InPs' policies and has to closely cooperate with VRRM. In this regard, CRRM maps the demanded capacity of VRRM to the available RATs, considering the suitability and loading factor of the entire underlying RATs, this way also promoting the notion of end-to-end slicing in virtual RANs. Since VNOs do not have any information about the InP's policies and InP is not aware of VNOs' policies neither, a level of isolation between the two is established. Additionally, it is worth to mention that by decoupling the VRRM functionalities, such as the ones related to service orchestration for different VNOs, from the underlying CRRM tasks associated with physical resource management among different RATs, the complexity of resource management has been reduced.

7.4 Future Works

In addition to what has been developed and achieved, there are several research lines which can be explored and carried out as future development to this thesis work. Some potential topics can be proposed as follows:

Although there is no indication of when network slicing might be commercially available, and the definition and use cases are still very much in the research stage, it is expected that virtualisation and

network slicing (both at network's core and RAN), play a critical role in the forthcoming 5G standard, because of the multitude of use cases and new services 5G will support. Therefore one research direction could be to adapt the proposed model in order to include the 5G specific protocols, network architecture and service requirements. It would also be an added value to joint this thesis work (which is focused on RAN slicing) with other available studies on network's core slicing and virtualisation of computational resources, in order to propose an end-to-end model of 5G network slicing.

One challenge of RAN slicing is that since each service has its own network requirement, the complexity of centralised models in highly dense areas will increase drastically, which will consequently affect the service slices with real-time QoS requirement. Although the current model of this thesis proposes two separated management entities to reduce the resource management complexity, still by shifting some functionalities from VRRM to the associated VNOs or even connected users, near optimal, low complex, distributed algorithms can be carried out based on this model to increase not only the network scalability, but also the performance isolation level among the existing VNOs.

This thesis work does not control the traffic coming from individual end-users, rather deals with challenges of network management from a higher layer perspective. Therefore, scheduling is not performed at the RRU level, but rather at the flows level for service slices, which takes place in VRRM. Consequently, additional requirements, such as delay, are not addressed in real-time implementations, instead performed in a snapshot view of the network, which corresponds to the decision window of VRRM. Therefore, another research direction could be to extend this work in order to include real-time user traffic by generating arrival-departure user processes according to specific service distributions, while defining delay requirements of each service in VRRM optimisation model as an additional constraint. One may also consider to add another level of optimisation in LRRMs for RRU scheduling among the end-users and then compare the results to the one which is obtained in flow level.

Annex A

SINR and Data Rate Model

This annex presents the proposed model of obtaining data rate based on SINR, for the underlying physical RATs. The model is referred to the throughput over the physical layer. It is notable that, the model does not take into consideration all the necessary throughput reductions, such as overhead loads to obtain the throughput at higher levels/layers

The goal for deriving the relation between the SINR and data rate, is to obtain a PDF function from a set of possible values taken by a RRU from specific RAT, to the capacity which can be delivered by that RRU, and ultimately the aggregated capacity that each access technology can provide. Based on the model suggested in [DGBA12], considering an interference limited Het-Net, and users experiencing a channel fading with a Rayleigh Distribution (hence, the received power being described by the Exponential Distribution) probability of having SINR higher than an arbitrary value is derived. The Author in [Khat16], further modified the model to propose a new assumption that the arbitrary SINR value is larger than 1 dB. CDF, $P_{\Gamma[dB]}$ and PDF, $p_{\Gamma[dB]}$ of SINR in this work are obtained as follows:

$$P_{\Gamma[\mathrm{dB}]}(\gamma) = 1 - e^{-\frac{0.2}{\alpha_p}\ln(10)\gamma_{[\mathrm{dB}]}} \tag{A.1}$$

$$p_{\Gamma[\mathrm{dB}]}(\gamma) = \frac{0.2}{\alpha_p}\ln(10)e^{-\frac{0.2}{\alpha_p}\ln(10)\gamma_{[\mathrm{dB}]}} \tag{A.2}$$

where:

- α_p: the path loss exponent, where $\alpha_p \geq 2$.

The CDF and PDF function of SINR are shown in Figure A.1 for different path loss exponents. One can see that higher path loss exponents indicate higher attenuation of interference, leading to a higher SINR.

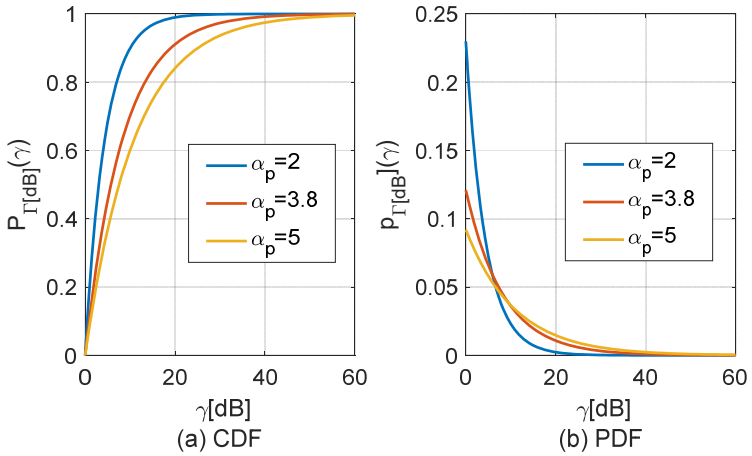


Figure A.1 – CDF and PDF functions of SINR.

By substituting (3.29) in (A.1), the variable of the CDF changes from SINR to the data rate of RRU, which can be expressed as follows:

$$P_{\mathrm{R[Mbps]}}(R_j^{RRU}) = 1 - e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_{m[\frac{\mathrm{dB}}{\mathrm{Mbps}^m}]}\left(R_{j[\mathrm{Mbps}]}^{RRU}\right)^m} \tag{A.3}$$

As the data rate of RRUs for different access technologies is bounded between a minimum and maximum value, the developed CDF and PDF functions are needed to be modified in a way to consider this limitation as well. According to [PaPi02], the conditional CDF and PDF, based on the bounded data rate variation of a single RRU can be obtained as:

$$P_{\mathrm{R[Mbps]}}(R_j^{RRU}|R_j^{RRU_L} \le R_j^{RRU} \le R_j^{RRU_H}) = \frac{e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU_L}\right)^m} - e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU}\right)^m}}{e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU_L}\right)^m} - e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU_H}\right)^m}} \quad \text{(A.4)}$$

where:

- $R_j^{RRU_L}$: the lower bound for the RRU's data rate,
- $R_j^{RRU_H}$: the higher bound for the RRU's data rate.

$$p_{\mathrm{R[Mbps]}}(R_j^{RRU}|R_j^{RRU_L} \le R_j^{RRU} \le R_j^{RRU_H}) = \frac{\frac{0.2}{\alpha_p}\left(\sum_{m=1}^{5} m a_m \left(R_{j[\mathrm{Mbps}]}^{RRU}\right)^{m-1}\right) e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU}\right)^m}}{e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU_L}\right)^m} - e^{-\frac{0.2}{\alpha_p}\ln(10)\sum_{m=0}^{5} a_m\left(R_{j[\mathrm{Mbps}]}^{RRU_H}\right)^m}} \quad \text{(A.5)}$$

It is also notable that $R_j^{RRU_L}$, is a nonnegative value and $R_j^{RRU_H}$, can reach to a maximum value corresponding to the highest MCS for each RAT:

$$0 \le R_j^{RRU_L} \le R_j^{RRU} \le R_j^{RRU_H} \le R_j^{RRU_{max}} \quad \text{(A.6)}$$

where:

- $R_j^{RRU_{max}}$: the maximum achievable data rate for a single RRU from each RAT.

As an example, Figure A.2 represents the CDF and PDF functions of a single LTE RRU obtained from (A.4) and (A.5) respectively. One can notice that as the path loss exponent increases, the probability of experiencing a higher throughput is greater.

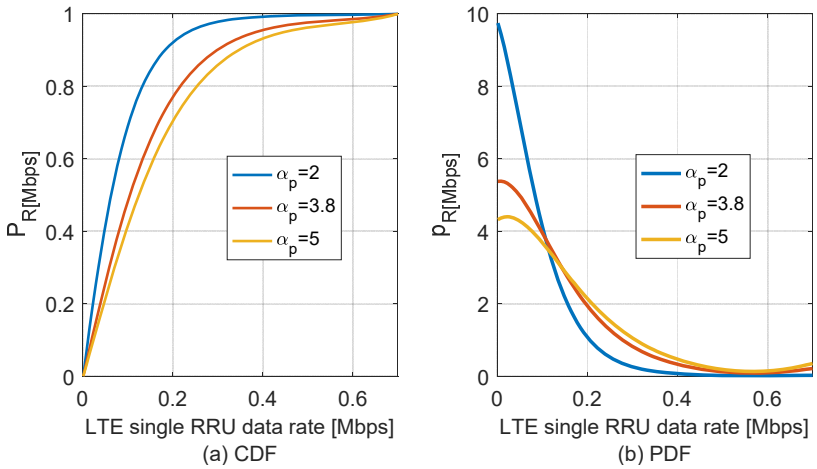


Figure A.2 – CDF and PDF of LTE single RRU data rate.

References

[3GPP09] 3GPP, *Considerations on Energy Saving Solutions in Heterogeneous Networks*, Report R3-092478, Oct. 2009 (http://www.3gpp.org/DynaReport/TDocExMtg--R3-65b--27587.htm).

[3GPP10] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Overall description Stage 2*, Report TS 136.300, V8.12, Apr. 2010 (http://www.3gpp.org).

[3GPP10b] 3GPP, *Overview to LTE energy saving solutions to cell switch off/on*, Report R3-100162, Jun. 2010 (http://www.3gpp.org/ftp/tsg_ran/WG3_Iu/TSGR3_66bis/docs/R3-100162.zip).

[3GPP13] 3GPP, *Technical Specification Group Services and System Aspects; Network Sharing; Architecture and functional description,* Report TS 23.251, V11.5.0, Mar. 2013 (http://www.3gpp.org /ftp/Specs/htmlinfo/23251.htm).

[3GPP13b] 3GPP, *Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Policy and charging control architecture*, ETSI TS123.203, V11.11.0, Sep. 2013 (http://www.etsi.org/deliver/etsi_ts/123200_123299/123203/11.11.00_60/ts_123203v111100p.pdf).

[3GPP14] 3GPP, *Main Specifications of HSPA*, http://www.3gpp.org/technologies/keywords-acronyms/99-hspa, Dec. 2014.

[3GPP16] 3GPP, *Technical Specification Group Services and System Aspects; Policy and charging control architecture*, Technical Specification TS 23.203 V14.1.0, Sep. 2016.

[3GPP17] 3GPP, *Study on New Radio Access Technology; Radio Access Architecture and Interfaces*, TR 38.801 V14.0.0, Mar. 2017.

[3GPP17b] 3GPP, *System Architecture for the 5G System; Stage 2 (Release 15)*, TS 23.501 V15.0.0, Dec. 2017.

[3GPP18] 3GPP, *NR and NG-RAN Overall Description*, TS 38.300 V1.2.0, Jan. 2018.

[4WAR10] 4WARD Project Report, *Architecture and Design for the Future Internet*, EC FP7-ICT Project 216041, June 2010 (www.4ward-project.eu).

[5GPP16] 5G PPP METIS-II project, Preliminary Views and Initial Considerations on 5G RAN Architecture and Functional Design, White paper, Mar. 2016 (https://5g-ppp.eu/white-papers/).

[AgZe15] D. Agrawal and Q. Zeng, *Introduction to Wireless and Mobile Systems*, Cengage Learning, Boston, USA, 2015.

[Aija16] A. Aijaz, "Towards 5G-enabled Tactile Internet: Radio Resource Allocation for Haptic Communications", in *Proc. of WCNC'16 – 17th IEEE Wireless Communications and Networking Conference,* Doha, Qatar, Apr. 2016.

[BAEK13] J. Burbank, J. Andrusenko, J. Everett and W. Kasch, *Wireless Networking: Understanding Internetworking Challenges,* John Wiley & Sons Ltd, New York City, NY, USA, 2013.

[BaLu11] M. Barreiros and P. Lundqvist, *QoS-Enabled Networks: Tools and Foundations* John Wiley & Sons Ltd, Chichester, UK, 2011.

[BBRR10] A. Bou, O. Bulakci, S. Redana, B. Raaf and J. Inen, "Enhancing LTE-advanced relay deployments via Biasing in cell selection and handover decision", in *Proc. of PIMRC'10 – The 21th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, Istanbul, Turkey, Sep. 2010.

[BeKM13] O. Bejarano, E. W. Knightly and P. Minyoung, "IEEE 802.11ac: from channelization to multi-user MIMO", *IEEE Communications Magazine,* Vol. 51, No. 10, Oct. 2013, pp. 84-90.

[Bidg12] H. Bidgoli, *Handbook of Computer Networks: Distributed Networks, Network Planning, Control, Management, and New Trends and Applications*, John Wiley & Sons, New Jersey, USA, 2012.

[BlCh14] N. Blaunstein and C. Christodoulou, *Radio Propagation and Adaptive Antennas for Wireless Communication Networks*, John Wiley & Sons, New Jersey, USA, 2014.

[BoVa09] S. Boyd and L. Vandenberghe *Convex Optimization*, Cambridge University Press, Edinburgh, UK, 2009.

[CMKR13] J. Chen, R. Mahindra, M.A. Khojastepour and S. Rangarajan, "A scheduling framework for adaptive Video delivery over cellular networks", in *Proc. of MobiCom'13 – ACM 19th annual international conference on Mobile computing &*

networking, Miami, Florida, USA, Sep. 2013.

[Corr07] L.M. Correia, *Mobile Communications Systems Lecture Notes,* Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, 2007 (https://fenix.ist.utl.pt/disciplinas/scm-1/2006-2007/2-semestre/pagina-inicial).

[Corr13] L.M. Correia, *Mobile Communications Systems Lecture Notes*, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, 2013.

[Cox12] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*, John Wiley & Sons Ltd, Sussex, UK, 2012.

[CSGM13] X. Costa-Perez, J. Swetina, T. Guo and R. Mahindra, "Radio access network virtualization for future mobile carrier networks", *IEEE Communications Magazine*, Vol. 51, No. 7, Jul. 2013, pp. 27–35.

[CVX17] CVX – Software for Disciplined Convex Programming, http://cvxr.com, Feb. 2017.

[DaPS11] E. Dahlman, S. Parkvall and J. Sköld, *4G: LTE/LTE-Advanced for Mobile Broadband* Elsevier/Academic Press, Oxford, UK, 2011.

[ErDa15] J. Erfanian and B. Daly, *5G White Paper*, NGMN Alliance White Paper, Mar. 2015 (https://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf).

[ETSI18] ETSI – Industry Specification Group (ISG) Network Functions Virtualization (NFV), http://www.etsi.org/technologies-clusters/technologies/nfv, Aug. 2018.

[FGR07] N. Feamster, L. Gao and J. Rexford, "How to Lease The Internet in Your Spare Time", *ACM SIGCOMM Computer Communication Review,* Vol. 37, No. 1, Jan. 2007, pp. 61-64.

[FGYK17] M. Farooq, H. Ghazzai, E. Yaacoub, A. Kadri and M. Alouini, "Green Virtualization for Multiple Collaborative Cellular Operators", *IEEE Transactions on Cognitive Communications and Networking,* Vol. 3, No. 3, Sep. 2017, pp. 420–434.

[FoMD11] T. Forde, I. Macaluso and L. Doyle, "Exclusive Sharing & Virtualisation of The Cellular Network", in *Proc. of DySPAN'11 – IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks,* Aachen, Germany, May 2011.

[DGBA12] H. S. Dhillon, R. K. Ganti, F. Baccelli and J. G. Andrews, "Modeling and Analysis of K-Tier Downlink Heterogeneous Cellular Networks", *IEEE Journal on Selected Areas in Communications,* Vol. 30, No. 3, Apr. 2012, pp. 550–560.

[GMF15] M. Gerasimenko, D. Moltchanov and R. Florea, “Cooperative Radio Resource Management in Heterogeneous Cloud Radio Access Networks”, *IEEE Access*, Vol. 3, Apr. 2015, pp. 397–406.

[GPLK13] A. Gudipati, D. Perry, L.E. Li and S. Katti, “Softran: Software defined radio access network,” in *Proc. of ACM SIGCOMM’13 – the 2nd ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking (HotSDN)*, Hong Kong, China, Aug. 2013.

[Hill02] F. Hillebrand, *GSM and UMTS: The Creation of Global Mobile Communication*, John Wiley & Sons Ltd, Sussex, UK, 2002.

[Hoss09] E. Hossain, *Heterogeneous Wireless Access Networks: Architectures and Protocols*, Springer, New York, NY, USA, 2009.

[HoTo07] H. Holma and A. Toskala, *HSDPA/HSUPA for UMTS*, John Wiley & Sons Ltd, Sussex, UK, 2007.

[IEEE13] IEEE 802.11 WG, “*IEEE Standard for Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements – Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications – Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz*”, IEEE Std 802.11ac – 2013, Dec. 2013 (http://standards.ieee.org).

[ITUR94] ITU-R, *Framework for The Radio Interfaces and Radio Sub-system Functionality for International Mobile Telecommunications-2000 (IMT-200)*, Rec. ITU-R M.1035, 1994 (http://www.itu.int).

[ITUR17] ITU-R, *Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)*, IMT-2020.Tech PERF REQ, Feb. 2017 (http://www.itu.int).

[Jacn14] General architecture of cellular networks: GSM/UMTS/LTE, http://www.joc182 .com.ve, Dec. 2014.

[Kell08] F. Kelly, “Charging and rate control for elastic traffic”, *European transactions on Telecommunications,* Vol. 8, No. 1, Sep 2008, pp. 33-37.

[KGJ13] E. Khloussy, X. Gelabert and Y. Jiang, “A Revenue-Maximizing Scheme for Radio Access Technology Selection in Heterogeneous Wireless Networks with User profile Differentiation”, in *Proc. of EUNICE’13 – 19th IFIP Workshop on Advances in Communication Networking,* Chemnitz, Germany, Aug. 2013.

[Khat16] S. Khatibi, *Radio Resource Management Strategies in Virtual Networks*, Ph.D.

Thesis, University of Lisbon, Lisbon, Portugal, 2016 (http://grow.tecnico.ulisboa.pt/wp-content/uploads/2016/08/Thesis_sina_khatibi_IST172360.pdf).

[KhCo14] S. Khatibi and L.M. Correia, “Modelling of Virtual Radio Resource Management for Cellular Heterogeneous Access Networks”, in *Proc. of PIMRC’14 – IEEE 25th International Symposium on Personal, Indoor and Mobile Radio Communications,* Washington, DC, USA, Sep. 2014.

[KhCo17] S. Khatibi and L.M. Correia, “Modelling Virtual Radio Resource Management in Full Heterogeneous Network,” *EURASIP Journal on Wireless Communications and Networking*, Vol. 2017, No. 73, Apr. 2017.

[KMZR12] R. Kokku, R. Mahindra, H. Zhang and S. Rangarajan, “Nvs: a substrate for virtualizing wireless resources in cellular networks,” *IEEE/ACM Transactions on Networking*, Vol. 20, No. 5, Apr. 2012, pp. 1333–1346.

[Koro11] L. Korowajczuk, *LTE, WiMAX and WLAN Network Design, Optimisation and Performance Analysis,* John Wiley & Sons Ltd, Sussex, UK, 2011.

[LAL17] J. Lucena, P. Ameigeiras and D. Lopez, “Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges”, *IEEE Communications Magazine,* Vol. 55, No. 5, May 2017, pp. 80–87.

[LiYu14] C. Liang and F. Yu, “Wireless Network Virtualisation: A Survey, Some Research Issues and Challenges”, *IEEE Communications Surveys & Tutorials,* Vol. 17, No. 1, Aug. 2014, pp. 358-380.

[LiYu15] K. Liang and L. Yu, “Enabling 5G mobile wireless technologies”, *EURASIP Journal on Wireless Communications and Networking,* Vol. 2015, No. 218, Sep. 2015.

[LPPA11] A. Leivadeas, C. Papagianni, E. Paraskevas, G. Androulidakis and S. Papavassiliou, “An Architecture for Virtual Network Embedding in Wireless Systems”, in *Proc. of NCCA’11 – 1st International Symposium on Network Cloud Computing and Applications,* Toulouse, France, Nov. 2011.

[LZLZ12] M. Li, L. Zhao, X. Li, Y. Zaki, A. Timm-Giel and C. Gorg, “Investigation of network virtualization and load balancing techniques in LTE networks,” in *Proc. of VTC Spring’12 – 75th IEEE Vehicular Technology Conference*, Yokohama, Japan, May 2012.

[MABP08] N. McKeown, N.T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker and J. Turner, “OpenFlow: Enabling Innovation in Campus

Networks", *ACM SIGCOMM Computer Communication Review*, Vol. 38, No. 2, Apr. 2008, pp. 69-74.

[MBQB18] P. Marsch, O. Bulakci, O. Queseth and M. Boldi, *5G System Design: Architectural and Functional Considerations and Long Term Research*, John Wiley & Sons, Hoboken, NJ, USA, 2018.

[MiSo15] G. Miao and G. Song, *Energy and Spectrum Efficient Wireless Network Design*, Cambridge University Press, Cambridge, United Kingdom, 2015.

[NaWK12] S. Namba, T. Warabino and S. Kaneko, "BBU-RRH switching schemes for centralized RAN", in *Proc. of ICST'12 – 7th International Conference on Communications and Networking in China*, Kunming, China, Aug. 2012.

[NGMN15] NGMN, *5G White Paper 2015*, White paper, Nov. 2015, (https://www.ngmn.org /uploads/media/NGMN_5G_White_Paper_V1_0.pdf).

[NGMN18] NGMN Alliance, "Description of Network Slicing Concept," https://www.ngmn.org/ uploads/media/160113_Network_Slicing_v1_0.pdf, Apr. 2018.

[OMM16] A. Osseiran, J.F. Monserrat and P. Marsch, *5G Mobile and Wireless Communications Technology*, Cambridge University Press, Cambridge, United Kingdom, 2016.

[PABC06] L. Peterson, T. Anderson, D. Blumenthal, D. Casey, D. Clark and D. Estrin, "Geni Design Principles," *IEEE Computer*, vol. 39, no. 9, Sep. 2006, pp. 102–105.

[Piao07] G. Piao, *Radio Resource Management for Integrated Services in Multi-radio Access Networks*, Springer, Kassel, Germany, 2007.

[PSAD05] J. Prez, O. Sallent, R. Agusti and M. Diaz, *Radio Resource Management Strategies in UMTS*, John Wiley & Sons Ltd, Sussex, UK, 2005.

[Rais04] V. Raisanen, *Implementing Service Quality in IP Networks*, John Wiley & Sons, West Sussex, UK, 2004.

[RBHR16] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed and T. Ahmed, "Scheduling Wireless Virtual Networks Functions", *IEEE Transactions on Network and Service Management,* Vol. 13, No. 2, Jun. 2016, pp. 240–252.

[RBS16] M. Richart, J. Baliosian and J. Serrat, "Resource Slicing in Virtual Wireless Networks: A Survey", *IEEE Transactions on Network and Service Management*, Vol. 13, No. 3, Sep. 2016, pp. 462–476.

[RBR14] R. Riggio, D. Boru and T. Rasheed, “Joule: Software-defined Energy Metering”, in *Proc. of NOMS'14 – 14th IEEE/IFIP Network Operations and Management Symposium*, Krakow, Poland, May 2014.

[RMM17] P. Rost, C. Mannweiler and D.S. Michalopoulos, “Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks”, *IEEE Communications Magazine,* Vol. 55, No. 5, May 2017, pp. 72–79.

[Rohd18] Rohde & Schwarz, https://www.rohde-schwarz.com/us/solutions/test-and measurement/wireless-communication/5g/5g-fundamentals/5g-fundamentals_ 229439.html, Aug. 2018.

[SAIL13] Scalable and Adaptive Internet Solutions (SAIL), EC FP7-ICT Project 257448, Jan. 2013 (www.sail-project.eu).

[SAR10] M. Salem, A. Adinoyi and M. Rahman, “Fairness-aware radio resource management in downlink OFDMA cellular relay networks”, *IEEE Transactions on Wireless Communications,* Vol. 9, No. 5, May 2010, pp. 1628–1639.

[Saut10] M. Sauter, *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*, John Wiley & Sons, Chichester, UK, 2010.

[SBT11] S. Sesia, M. Baker and I. Toufik, *LTE – The UMTS Long Term Evolution: From Theory to Practice*, John Wiley & Sons, Chichester, UK, 2011.

[Sing10] T. Singal, “Emerging Wireless Network Technologies”, in Choudhari, M. (ed.), *Wireless Communications*, McGraw Hill, New Delhi, India, 2010.

[Song99] D.J. Songhurst, *Charging Communication Networks: From Theory to Practice*, Elsevier, Amsterdam, The Netherlands, 1999.

[SPBP06] N. Spring, L. Peterson, A. Bavier and V. Pai, “Using PlanetLab for Network Research: Myths, Realities, and Best Practices”, *ACM SIGOPS Operating Systems Review,* Vol. 40, No. 1, Jan. 2006, pp. 17-24.

[SYHT16] S. Singh, S. Yeh, N. Himayat and S. Talwar, “Optimal Traffic Aggregation in Multi-RAT Heterogeneous Wireless Networks”, in *Proc. ICC’16 – 52th IEEE International Conference on Communications*, Kuala Lumpur, Malaysia, May 2016.

[TAV16] G. Tseliou, F. Adelantado and C. Verikoukis, “Scalable RAN Virtualization in Multitenant LTEA Heterogeneous Networks”, *IEEE Transactions on Vehicular Technology,* Vol. 65, No. 8, Aug. 2016, pp. 6651–6664.

[WCLM99] C.Y. Wong, R.S. Cheng, K.B. Letaief and R.D. Murch, “Multiuser OFDM with adaptive subcarrier, bit and power allocation,” *IEEE Journal on Selected Areas in Communications*, Vol. 17, No. 10, Oct. 1999, pp. 1747–1758.

[WHWX05] S. Wei, J. Hai, Z. Weihua and S. Xuemin, “Resource Management for QoS Support in Cellular/WLAN Interworking”, *IEEE Network Magazine*, Vol. 19, No. 5, Oct. 2005, pp. 12-18.

[WTL13] H. Wen, P. Tiwary and T. Le-Ngoc, *Wireless Virtualisation*, Springer International Publishing, Heidelberg, Germany, 2013.

[XLWZ17] F. Xu, Y. Li, H. Wang, P. Zhang and D. Jin, “Understanding mobile traffic patterns of large scale cellular towers in urban environment”, *IEEE/ACM Transactions on Networking*, Vol. 25, No. 2, Apr. 2017, pp. 1047–1061.

[Xu12] X. Xu, “From Cloud Computing to Cloud Manufacturing,” *Robotics and Computer-Integrated Manufacturing*, Vol. 28, No. 1, Feb.2012, pp. 75–86.

[XZS17] W. Xiang, K. Zheng and X. Shen, *5G Mobile Communications*, Springer, Zürich, Switzerland, 2017.

[YLJZ14] M. Yang, Y. Li, D. Jin, L. Zeng, X. Wu and A. Vasilakos, “Software-Defined and Virtualised Future Mobile and Wireless Networks: A Survey”, *Mobile Networks and Applications,* Vol. 19*,* No. 1, Sep. 2014, pp. 1-15.

[YuZh14] X. Yu and H. Zhu, “Optimal Resource Management with Delay Differentiated Traffic and Proportional Rate Constraint in Heterogeneous Networks”, *Journal of Communications*, Vol. 9, No. 9, Sep. 2014, pp. 714–722.

[ZZGT10] Y. Zaki, L. Zhao, C. Gorg and A. Timmen-Giel, "LTE wireless virtualization and spectrum management”, in *Proc. of WMNC’10 – 3rd International Conference on Wireless and Mobile Networking,* Oct. 2010, Budapest, Hungary.

[ZZGT11] Y. Zaki, L. Zhao, C. Gorg and A. Timmen-Giel, “LTE mobile network virtualization—exploiting multiplexing and multi-user diversity gain”, *Mobile Networks and Applications,* Vol. 16, No. 4, Aug. 2011, pp. 424-432.