UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Radio Resource Management Strategies in Virtual Networks

Sina Khatibi

Supervisor: Doctor Luís Manuel de Jesus Sousa Correia

Thesis approved in public session to obtain PhD degree in Electrical and Computer Engineering.

Jury final classification: **Pass with Distinction**

Jury

Chairperson: Chairman of the IST Scientific Board

Members of the Committee:

Doctor Raymond Knopp,
Doctor Susana Isabel Barreto de Miranda Sargento
Doctor Luis Manuel de Jesus Sousa Correia
Doctor Rui Manuel Rodrigues Rocha
Doctor Antonio José Castelo Branco Rodrigues

2016

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Radio Resource Management Strategies in Virtual Networks

Sina Khatibi

Supervisor: Doctor Luís Manuel de Jesus Sousa Correia

Thesis approved in public session to obtain PhD degree in Electrical and Computer Engineering.

Jury final classification: **Pass with Distinction**

Jury

Chairperson: Chairman of the IST Scientific Board

Members of the Committee:

Doctor Raymond Knopp, Professor, EURECOM – Graduate School and Research Center in Communication Systems, France

Doctor Susana Isabel Barreto de Miranda Sargento, Associate Professor (with habilitation), Universidade de Aveiro

Doctor Luis Manuel de Jesus Sousa Correia, Associate Professor, (with habilitation), Instituto Superior Técnico, Universidade de Lisboa

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

Doctor Antonio José Castelo Branco Rodrigues, Assistant Professor, Instituto Superior Técnico, Universidade de Lisboa

2016

To my beloved family for all their support.

Acknowledgements

Above all, I wish to express my deep and sincere gratitude to Prof. Luis M. Correia for his guidance, support, motivation, corrections, productive ideas and discussions throughout this work.

Several people have contributed to this work by inspiring discussions and fruitful collaboration. I want to mention namely our Group for Research on Wireless (GROW), in particular Lúcio Ferreira, Mónica Antunes Branco, and Luísa Caeiro, who worked close to me in the European projects, within which this work was partially developed.

Special thanks go to my family and friends, who have spared my company on many occasions when I has committed to this work. Above all, I thank my parents and my sister for their patience, love and support.

Finally, I would also like to acknowledge IST, INOV – INESC, as well as the European Commission through the Mobile Cloud Networking, NEWCOM#, COST 2100 and COST IC 1004 projects, for partially funding this research.

Abstract

This thesis addresses the concept of virtualisation of radio resources to achieve virtual wireless links, by aggregating all the physical resources from different Radio Access Techniques and offering Capacity-as-a-Service to Virtual Network Operators (VNOs). The proposal of an analytical model for the management of virtual radio resources is the key goal of this thesis, the estimation of network capacity and data rate allocation being its two main components. Based on the probability distribution of the Signal-to-Interference-plus-Noise Ratio observed at the user level, the model leads to the probability distribution for the total network throughput. It considers different approaches for the estimation of network global data rate, based on different channel qualities. The second component uses the outcome of the first one in order to maximise the weighted throughput subject to the total network capacity, the Service Level Agreements of VNOs, and fairness. The weights for services in the objective function of the resource allocation component enable the model to have prioritisation among services. The model performance is evaluated in a practical heterogeneous access network. Results show an increase of 2.5 times in network capacity by implementing an access point at the centre of each cell of a cellular network. One shows that the cellular network capacity itself can vary from 0.9 Gbps in the pessimistic approach up to 5.5 Gbps in the optimistic one. Finally, the isolation of service classes and VNOs by means of virtualisation of radio resources is clearly demonstrated.

Keywords – Virtualisation of Radio Resources. Virtual Radio Resource Management. Radio Access Networks. Network Function Virtualisation.

Resumo

Esta tese aborda o conceito de virtualização de recursos rádio para obter ligações sem fios virtuais, através da agregação dos recursos físicos de diferentes técnicas de cesso rádio, oferecendo capacidade como um serviço a operadores virtuais de redes. A proposta de um modelo analítico para a gestão de recursos de rádio virtuais é o objetivo principal desta tese, sendo a estimação da capacidade da rede e a atribuição de ritmos de transmissão as duas componentes principais deste modelo. Baseado na distribuição de probabilidade para a relação sinal ruído mais interferência observada no utilizador, o modelo fornece a distribuição de probabilidade para o ritmo de transmissão global da rede. O modelo considera abordagens diferentes para a estimação do ritmo de transmissão global da rede, baseadas em qualidades de canal diferentes. A segunda componente usa o resultado da primeira para maximizar o peso do ritmo de transmissão, sujeito à capacidade total da rede, aos acordos de prestação de serviço, e equidade. Os pesos dos serviços na função de custo da atribuição de recursos permitem que o modelo atribua prioridades aos serviços. O desempenho do modelo é avaliado num cenário prático de redes heterogéneas. Os resultados mostram um aumento de 2.5 vezes na capacidade de rede resultante da implementação de um ponto de acesso em cada estação base da rede celular. Mostra-se que a capacidade da rede celular pode variar entre 0.9 Gbps na perspetiva pessimista e 5.5 Gbps na otimista. Por último, é demonstrado o isolamento entre classes de serviços e operadores virtuais de rede através da virtualização de recursos rádio.

Palavras-chave: Virtualização de Recursos Rádio. Gestão de Recursos Rádio virtuais. Redes de Acesso Radio. Virtualização de Funcionalidades em Redes.

چکیده

این پایان نامه به مفهوم مجازی سازی منابع رادیویی برای رسیدن به لینک های بی سیم مجازی، با تجمع تمام منابع فیزیکی از تکنیک های رادیو دسترسی های مختلف و ارائه ظرفیت به عنوان یک سرویس به اپراتور های شبکه های مجازی می پردازد. هدف اصلی این پایان نامه پیشنهاد یک مدل تحلیلی برای مدیریت منابع رادیویی مجازی است که از دو بخش اصلی برآورد ظرفیت شبکه و تخصیص نرخ داده بودن تشکیل شده است. بر اساس توزیع احتمال سیگنال به نویز دریافتی ترمینال، تابع توزیع احتمال برای توان عملیاتی کل شبکه می باشد. در مدل پیشنهادی رویکردهای مختلفی برای برآورد نرخ داده کل شبکه بر اساس کیفیت کانال های مختلف مدنظر قرار میدهد. در بخش دوم بر اساس نتیجه مرحله قبل، بهینه سازی نرخ داده وزنی با در نظر گرفتن قیدهایی چون ظرفیت کل شبکه، موافقتنامه سطح سرویس اپراتورهای مجازی و عدالت است. وزن های تعریف شده برای سرویس های مختلف در تابع هدف در بخش تخصیص منابع، قابلیت اولویت بندی در میان سرویسهای مختلف را به مدل اضافه می نماید. عملکرد مدل در شبکه واقعی ناهمگن دسترسی بی سیم مورد بررسی قرار گرفت. نتایج عددی افزایش ۲/۵ برابری ظرفیت شبکه را با قرار دادن یک نقطه دسترسی وای فای در مرکز هر سلول نشان می دهد. همچنین بر اساس نتایج عددی می توان گفت که برآورد ظرفیت کل شبکه تلفن همراه در رویکردهای مختلف، متفاوت است و می توانید از ۰/۹گیگا بیت در ثانیه در رویکرد بدبینانه تا ۵/۵ گیگا بیت در ثانیه در خوش بینانه افزایش یابد. در نهایت، تفکیک کلاس های سرویس و اپراتورهای مجازی از یکدیگر با استفاده از مجازی سازی از منابع رادیویی است که به وضوح نشان داده است.

کلمات کلیدی - مجازی سازی منابع رادیو، مدیریت منابع مجازی رادیویی، شبکه های دسترسی رادیویی، شبکه مجازی سازی عملکرد.

Table of Contents

List of Figures

List of Tables

List of Acronyms

2G	2^{nd} Generation
3G	3^{rd} Generation
3GPP	3G Partnership Project
4G	4^{th} Generation
5G	5^{th} Generation
ACID	Atomicity, Consistency, Isolation, Durability
AP	Access Point
AuC	Authentication Centre
BBU	Base Band Unit
BE	Best-Effort
BG	Best-Effort with minimum Guaranteed
BS	Base Station
BSC	Base Station Controller
BSS	Basic Service Set
BTS	Base Transceiver Station
BWA	Broadband Wireless Access
CC	Cloud Controller
CDMA	Code Division Multiple Access
CN	Core Network
CQI	Channel Quality Indicator
CAPEX	Capital Expenditure
C-RAN	Cloud-Based Radio Access Network
CRRM	Common Radio Resource Management
CS	Circuit Switch
CSI	Channel State Information
DFT	Discrete Fourier Transfer
EDGE	Enhanced Data rates fir GSM Evolution
EIR	Equipment Identity Register
eNodeB	E-UTRAN NodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
ETSI	European Telecommunication Standards Institute
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
FDD	Frequency Duplexing Division
FDMA	Frequency Division Multiple Access

FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
GERAN	GSM/Edge Radio Access Network
GMSC	Gateway MSC
GPP	General Purpose Processor
GPRS	General Packet Radio Service
GSM	Global System for Mobile
GUTI	Globally Unique Temporary Identity
HAN	Heterogeneous Access Network
HLR	Home Location Register
HSPA	High Speed Packet Access
HSS	Home Subscription Server
IaaS	Infrastructure-as-a-Service
ICI	Inter-Carrier Interference
IDFT	Inversed Discrete Fourier Transfer
IMS	IP Multimedia Subsystem
IMSI	International Mobile Subscriber Identity
IP	Internet Protocol
ISDN	Integrated Service Digital Network
KPI	Key Performance Indicators
KQI	Key Quality Indicators
LAN	Local Area Network
LTE	Long Term Evolution
MAC	Medium Access Control
MCNSP	Mobile Cloud Network Service Provider
MME	Mobility Management Entity
MME/GW	Mobility Management Entity Gate Way
MIMO	Multiple-Input Multiple-Output
MOCN	Multi-Core Network
MORAN	Multi-Operator RAN
MSC	Mobile Switching Centre
MT	Mobile Terminal
MU-MIMO	Multi-User Multiple-Input Multiple-Output
MS	Mobile Station
NAS	Non-Access Stratum
NFV	Network Functions Virtualisation
OAI	Open Air Interface
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	OPerational Expenditure

OS	Operating System
OVSF	Orthogonal Variable Spreading Factor
PaaS	Platform-as-a-Service
PCEF	Policy Control Enforcement Function
PCRF	Policy and Charging Rule Function
PDF	Probability Density Function
PDN	Packet Data Network
PF	Proportional Fairness
P-GW	Packet Gate Way
PLMN	Public Land Mobile Network
PS	Packet Switch
PSTN	Public Switch Telephone Network
QCI	QoS Class Indicator
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RANaaS	RAN-as-a-Service
RANP	RAN Provider
RAT	Radio Access Technology
RNC	Radio Network Controller
RRH	Remote Radio Head
RRM	Radio Resource Management
RRU	Radio Resource Unit
SAE	System Architecture Evolution
SDR	Software Defined Radio
SGSN	Serving GPRS Support Node
S-GW	Switch Gate Way
SLA	Service Level Agreement
SM	Service Manager
SINR	Signal to Interference plus Noise Ratio
SO	Service Orchestrator
SU-MIMO	Single-User Multiple-Input Multiple-Output
TA	Tracking Area
TDD	Time Duplexing Division
TDMA	Time Division Multiple Access
TLB	Translation Look-aside Buffer
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Terrestrial Radio Access Network
VAF	Voice Activity Factor

VHO	Vertical Hand Over
VLR	Visitor Location Register
VM	Virtual Machine
VNO	Virtual Network Operator
VoIP	Voice Over IP
V-RAN	Virtual Radio Access Network
VRRM	Virtual Radio Resources Management
WCDMA	Wideband Code Division Multiple Access
Wi-Fi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network

List of Symbols

α	Spatial efficiency
α_p	Path loss exponent
α_s	Spreading factor
Δf	Channel bandwidth
Δt	Time interval
λ_{ji}	Arrival rate for service j of the VNO i
μ_{ji}	Serving rate for service j of the VNO i
ρ_0	An arbitrary SINR level
ρ_{AC}	Accumulated signal to noise ratio over the transmission channel.
ρ_{in}	SINR
$\rho^{R_{b_{min}}}$	Relative input SINR to achieve data rate equal or higher than R_{bmin}
$\bar{\tau}_{b_{ji}}^{k}$	Served average access delay for service j of the VNO i by from RAT k
$\bar{\tau}_{b_{ji}}^{Req}$	Requested average access delay for service j of the VNO i
$\bar{\tau}_{b_{ji}[s]}^{Srv}$	Served average access delay for service j of the VNO i
$\bar{\tau}_{b_{ji}}^{VRRM}$	Requested average access delay for service j of the VNO i from VRRM
a_k	Factor of polynomial approximation of SINR as function of data
f_c	Objective function of CRRM
f_v	Objective function of VRRM
f_v^{AP}	Objective function for access points
f_v^{Cell}	Objective function for cellular RATs
g_{RAT_i}	Transform function of i-th RAT data rate based on SINR
$n_{\Delta t}$	Number of time interval
$N_{b/sym}$	Number of bits per symbol
N_c	Number of subcarrier
N_{FFT}	FFT Size
N_{RAT}	Number of RATs
$N_{RRU}^{RAT_i}$	Total number of RRU in i-th RAT
$N_{SRRU}^{RAT_i}$	Number of spare RRU in i-th RAT
N_{srv}	Number of service for each VNO
N_{VNO}	Number of VNOs serving by this VRRM
p_x	Probability distribution function of random variable x
P_x	Cumulative distribution function of random variable x
R_b	Data Rate

$R_b^{c_i}$	Data rate of i-th subcarrier
$R_b^{Contrat}$	Contracted data rate
$R_{b_{ji}}^{AP}$	Allocated capacity for service j of the VNO i on CB RAT
$R_{b_{ji}}^{Cell}$	Allocated capacity for service j of the VNO i on CF RAT
R_b^{CRRM}	Total applicable data rate from multiple access technologies
R_b^{MCS}	Data rate estimation function
$R_{b_{ji}}^{Avg}$	Nominal data rate for service j of the VNO i
$R_{b_{ji}}^{min}$	Minimum guaranteed data rate service j of the VNO i
$\bar{R}_{b_{ji}}^{min}$	Minimum average data rate for service j of the VNO i
$R_{b_{ji}}^{Srv}$	Serving (allocated) capacity for service j of the VNO i
$R_{b_{RAT_i}}$	Applicable data rate from a single RRU of i-th RAT
$R_{b_{RAT_i}}^{max}$	Maximum applicable data rate from a single RRU of i-th RAT
$R_{b_n}^{RAT_i}$	Applicable data rate from n-th RRU of i-th RAT
$R_{b_{total}}^{RAT_i}$	Applicable data rate from a i-th RAT pool
$R_{b_{ji}}^{Srv}$	Served data rate for service j of the VNO
$\overline{R}_{b_{ji}}^{Srv}$	Average serving data rate for service j of the VNO i
$\overline{R}_b^{EU}$	End-user average data rate
R_c	Chip rate
s_t^{RRU}	State of radio resources at t
t^{out}	The amount of time that VNO is out of contract
t^{total}	Total time of measuring
$t_{Rsp_{ji}}$	Response time for service j of the VNO i
$t_{Rsp_{ji}}^{Nom}$	Nominal response time for service j of the VNO i
$t_{srv_{ji}}$:	Serving time for service j of the VNO i
$t_{Srv_{ji}}^{Nom}$	Nominal serving time service j of the VNO i
$\bar{t}_{s_{ji}}^{k}$ [s]	Average serving time for service j of the VNO i from RAT k
$\bar{t}_{s_{ji}}^{VRRM}$	Requested average serving time for service j of the VNO i from VRRM
W^{WLAN}	Weight for allocating capacity from AP RATs
W^{cell}	Weight for allocating capacity from cellular RATs
W_{ji}^{Srv}	Weight of serving unit of data rate for service j of the VNO i by VRRM
W_c^{Srv}	Weight of serving unit of data rate by CRRM

List of Software

MATLAB	MATLAB is used for optimisation, plotting and carve fitting in addition to create optimisation component for VRRMSim by its .Net Deployment tool.
MS Visio 2013	Visio is used to edit several figures present in this thesis.
MS Word 2013	Word is used to edit this thesis and all associated document such as publications.
MS-SQL Server 2012	MS SQL Server is used to store the intermediate results used for next step or next simulation with the aim of reducing the simulation time.
Open Air Interface	Linux Based Open-source software based eNodeB.
SQL Server Management Studio	SQL Server Management is used to manage the SQL databases created by VRRMSim.
VRRMSim	VRRMSim is the thesis simulator, developed in Visual C# and used to evaluate the majority of the proposed models.

Chapter 1

Introduction

This chapter presents an introduction to the thesis by presenting a brief history of wireless networks in Section 1.1. Section 1.2 addresses the motivation and objectives for 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 European projects contributions and published work. Finally, this Section 1.5 contains the dissertation contents in detail.

1.1 Brief History

Recent studies foresee future mobile networks supporting a significant number of users requesting for different services and applications, each one with different Quality of Service (QoS) requirements [AkHN11]. This predication is in line with the extraordinary evolution in wireless networks, which led to the emergence of new applications and services in addition to mobile user increment. The evolution of the wireless standards and technologies never ends, therefore, studies on the next generation of wireless networks starts once a given generation is standardised. Figure 1.1 briefly presents this evolution, a brief review being presented in what follows.

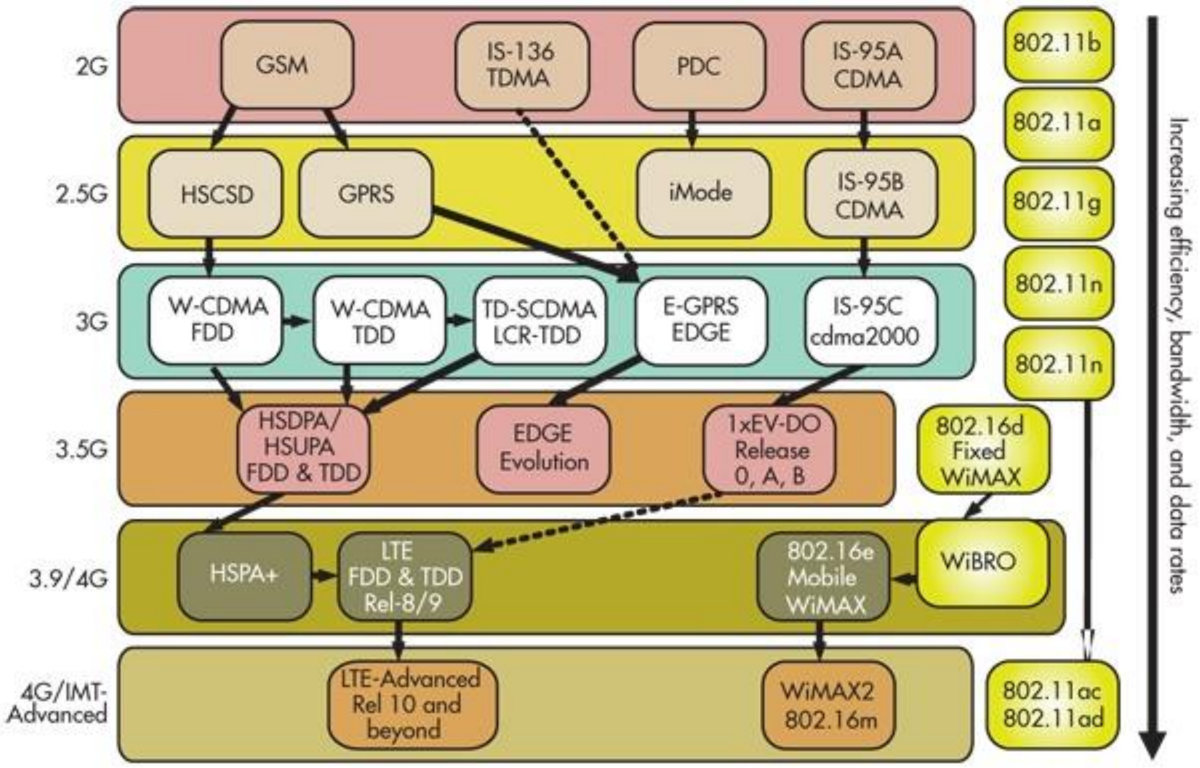


Figure 1.1 – Wireless and mobile system systems evolution (extracted from [Fren11]).

European countries in the 1980s introduced the first generation of mobile communications based on analogue cellular technologies. The standardising of the Global System for Mobile Communications (GSM), the digital 2nd Generation (2G), by European Telecommunications Standard Institute (ETSI) avoided the fragmentation and inefficiency of the 1st generation. The combination of Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA) were the fundamentals of GSM, packet switching services being integrated later by General Packet Radio Service (GPRS), followed by Enhanced Data rates for GSM Evolution (EDGE).

The 3rd Generation (3G) of mobile communications systems was the answer to two key challenges, i.e., providing higher-speed data services and increasing radio network capacity. Universal Mobile Telecommunications System (UMTS) started as the joint European and Japanese system for 3G, and the 3G Partnership Projects (3GPP) was created for its standardisation. Wideband Code Division

Multiple Access (WCDMA) was adopted for UMTS as the key multiple access technique. Although circuit switching was maintained in UMTS, the High Speed Packet Access (HSPA) schemes in Release 5 (2002) [HoTo04] brought the focus to packet switching.

Long Term Evolution (LTE, 3GPP Release 8) [3GPP09a] presented a new highly simplified flatter network architecture to address seamless mobility with minimum delay. In the new architecture, all interfaces are based on the Internet Protocol (IP), and there are only two types of nodes: eNodeB (evolved NodeB) and MME/GW (Mobility Management Entity/ Gateway). In contrast to the former mobile networks, LTE has only packet switching, the radio access being Orthogonal FDMA (OFDMA) and Single Carrier FDMA (SC-FDMA).

In addition, 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 system capacity and contributing to the optimisation of the diverse wireless resources utilisation.

Finally, the 5th Generation (5G) is a term used in the resent research papers and projects referring to the next major step in mobile and wireless communications beyond LTE – Advance. The key goal in 5G projects is to rethink the infrastructure, in order to provide ubiquitous super-fast connectivity and seamless service delivery in all circumstances. As summarised in Figure 1.2, the candidate infrastructure for 5G has to be able to meet the new requirements for the next generation of mobile communications networks, such as virtualised network functions.

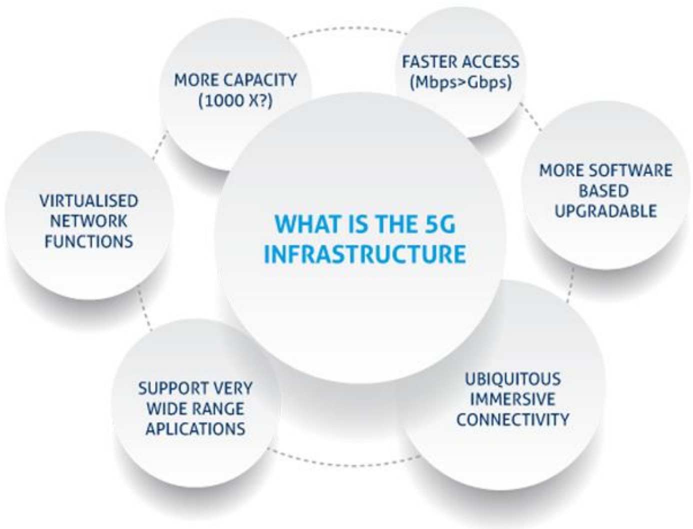


Figure 1.2 – The fundamentals of candidate infrastructure for 5G (extracted from [5GPP15]).

1.2 Motivation and objectives

Recent studies foresee the monthly global data traffic to surpass 10 Exabyte in 2017, as the result of the proliferation of smart devices and traffic hungry applications [Cisc13a]. Hence, operators using scarce radio resources have to find a practical, flexible, and cost-efficient solution for their networks. The increment of cellular networks' capacity by means of deploying dense base stations (BSs) is the groundwork in any candidate solution. In addition, traffic offloading, e.g., to Wi-Fi Access Points (APs), has recently proven to be a valuable complementary approach. According to [ArRA10, KJYI13], an acceptable portion of traffic can be offloaded to APs, just by deferring delay-tolerant services for a pre-specified maximum interval, until reaching an AP. Offloading approaches are generally based on using other connectivity capabilities of mobile terminals whenever possible, instead of using further expensive cellular bands. The authors in [JYSY14] discussed the economics of traffic offloading, and in [KDZH13] an energy-saving analysis is addressed.

Nevertheless, drastic temporal and geographical variations of traffic, in addition to the shortage of the network capacity, makes the situation for operators even worse [GuKM10]. The increase of traffic demand over mobile networks leads to growth of traffic imbalance among different locations and times, e.g., traffic load peak loads at noon over business areas and at night in residential ones [NaWK12]. The usual provisioning of Radio Access Networks (RANs) for busy hours leads to an inefficient resource usage, with relatively high CApital and OPerational Expenditure (CAPEX and OPEX) costs, which is not acceptable anymore. Instead, operators are more interested in flexible and elastic solutions, where they can also share their physical infrastructure.

It is well accepted by operators that infrastructures should be shared, in order to reduce operational costs. Based on [ViWC02] and [CSTM13], infrastructure sharing solutions can be categorised into three main types: geographical, passive, and active. In geographical sharing or national roaming, a federation of operators can achieve full coverage in a short time, by diving the service area into several regions, over which each of the operators provides coverage [SAMS11]. Passive sharing refers to the sharing agreement of fundamental infrastructures, such as tower masts, equipment houses and power supplies in order to reduce operational costs. Active sharing, however, is the sharing of transport infrastructures, radio spectrum, and baseband processing resources. Two types of active sharing are introduced in [FTAL08], i.e., Multi-Operator RAN and Multi-Core Network: in the former, operators maintain a maximum level of independent control over their traffic quality and capacity, by splitting BSs and their control nodes into logically independent units over a single physical infrastructure; however, in the latter, operators give up their independent control by sharing the aforementioned entities in conjunction with the pooling of radio resources. Although the cost items in Multi-Core Network are identical to Multi-Operator RAN, radio resources pooling leads to further savings in extremely low-traffic areas over equipment related costs. Moreover, a network-wide radio resource management framework is proposed [MKHR13], in order to achieve isolation in addition to the optimal distribution of resources across the network.

Despite of the RAN sharing benefits, surprisingly, few sharing agreements have been made, especially in mature markets. The reasons offered by operators for not engaging in sharing deals are often the up-front transformation costs, the potential loss of control over their networks, and the challenge of operational complexity [FPTG12, KKKZ11]. Sharing deals may be too expensive, and the initial cost of a network-sharing deal can be daunting, hence, operators without a comfortable margin of funds to make the necessary investment are likely to assume that they simply cannot afford to participate in such operation. The 3GPP standard also limits the shared RAN to serve only four operators [3GPP13]. Moreover, many operators, particularly incumbent ones, whose early entrance into their markets has given them advanced coverage and network quality, assume that sharing their network with rivals would dilute their competitive advantage. Some of them may feel that they would not be able to control the direction for the development of their network in future rollout strategies, and choices about hardware and vendors. Last but not the least, having a shared RAN running properly is an elaborate and complex task. Some operators believe that having a shared network operation puts many technological and operational challenges, which may lead to little financial benefits and great potential of chaos. However, from some studies [SoZC14], it is concluded that the pros of sharing are larger than the cons, and this general approach can really be seen as a very promising solution for the future, namely in a broader perspective, i.e., looking at RAN virtualisation instead of RAN sharing as the candidate solution.

Lately, the sharing of network infrastructure using Network Function Virtualisation (NFV) has become an active research topic. NFV is considered as a new approach to transform the way operators architect their networks [MDPA12], and some studies have also considered RAN as well. Virtualisation of wireless networks means setting up a network of individual virtualised network components, such as nodes, links and routers [ZLGT10]. It should not be mistaken with any of the aforementioned sharing techniques [ZZGT11]. The proper exploited isolation can reduce system downtime significantly. Better differentiation of these two concepts, however, can be achieved by the comparison of network sharing/virtualisation with relative well-known concepts on computers, known as multi-tasking (sharing)/virtualisation. Virtualisation is the creation of a virtual version of something, such as a hardware platform. Consequently, a virtual machine is a software implementation of a machine that executes programs like a physical machine. Although virtual machines may run over the same physical hardware, they can have different operating systems. No change is needed to run the processes and Operating Systems (OSs), since the virtualisation is not transparent to the software or OS. It is worth nothing that using an OS supporting multi-tasking on virtualisation leads to achieving multi-tasking and virtualisation on the same machine, as shown in Figure 1.3. The same concept of virtualisation can be proposed for networks. Multiple instances of V-RAN (Virtual – Radio Access Network) can co-exist in the same physical infrastructure, each one with different settings and requirements. The Virtual Network Operator (VNO), which does not own any RAN physical infrastructure, simply asks for the connectivity service and the virtual manager takes care of rest.

The key historical advantages of virtualisation can be counted as an efficient usage of resources, besides isolation. The authors in [KZJK12] present the advantages of network virtualisation, which can be also adapted to virtual RANs:

- **Coexistence of networks**: Applying virtualisation techniques to the network enables the creation of multiple virtual/logical networks on the same physical infrastructure. It is also possible to keep isolation among these networks, in such a way that multiple generations of cellular networks are accommodated in the same physical infrastructure.

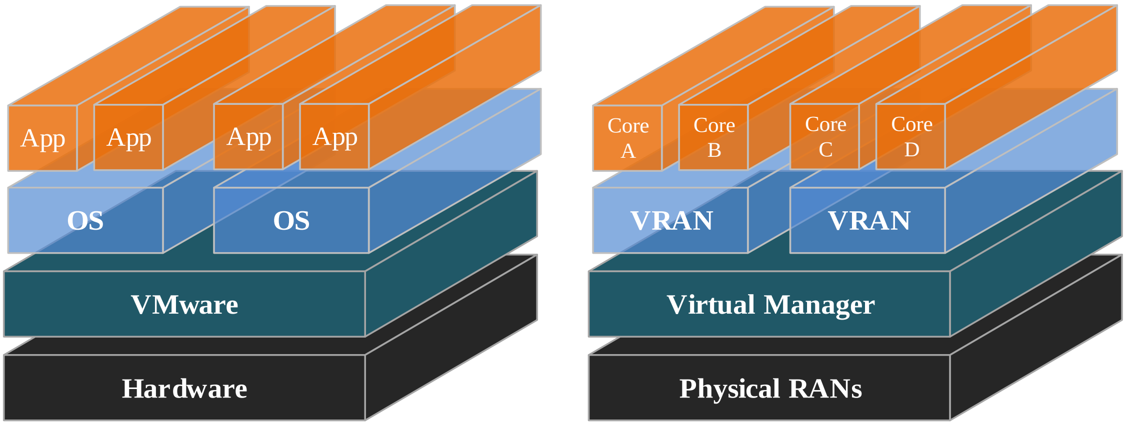


Figure 1.3 – Comparison of Virtual Machine (VM) and Virtual RAN (VRAN).

- **Protection**: There is no other way to protect an entity from its surroundings than isolation. Different protection features, such as access control or mandatory traffic encryption, may be used by different virtual networks. Network-wide protection is envisioned in network virtualisation rather than point-to-point protection offered by VPNs.
- **Optimisation of hardware utilisation**: By dynamically allocating virtual network nodes to the physical substrate, the hardware can be used for multiple virtual networks up to maximum capacity, thus, minimising the cost for the infrastructure provider.
- **Easy deployment of bundled network services**: Services are the equivalent of an application in virtual machines. In order to implement a network service, a specific network architecture (e.g. WLAN, sensor network, corporate internet) being needed. The simultaneous deployment of various network services with various required architectures can be achieved by mean of realising virtual networks, which do not rely on each other.
- **Separation of different agreement and licenses**: In addition to the separation of different classes of quality of service agreements, network virtualisation offers different traffic route based on different peering agreements.

Flexibility [SRYP13] and cost reduction [YrRu11] in RANs became the motivation for their implementation in cloud data centres, in order to achieve centralised processing, collaborative radio, real-time cloud computing [KMSS10], and clean RAN systems [CMRI11], also known as Cloud-based RAN (C-RAN). Although computer companies found it an interesting area, mobile operators have taken a more conservative approach to this phenomenon [YrRu11]. It also seems to be a promising solution for cost reduction in wireless communications [MCN15]. Mobile network entities and functionalities can be converted from dedicate hardware in current networks to cloud-base application operating over data centres. This evolution in mobile networks not only can reduce the operators' costs, but can offer also elastic, on-demand, and Pay-As-You-Go services. It also enables new sets of services, such as RAN-as-a-Service (RANaaS). These new proposed services seams very promising for network operators dealing with traffic growth explosion.

RANaaS, as the key topic of this thesis, is a service offered over cloud-based RAN. The novel architecture, as proposed in [CMRI11, NaWK12], contains three major parts as follows: Baseband Unit

(BBU), composed of high performance programmable processors and real-time virtualisation; high bandwidth low-latency optical transport networks connecting BBUs to Remote Radio Head (RRH); RRHs, which are distributed radio units and antennas, located in the remote site. In this new architecture, signalling, traffic data and Channel State Information (CSI) of active Mobile Terminals (MTs) can be easily shared. Therefore, capacity improvement can be achieved by the implementation of joint processing and scheduling. Adaptability to non-uniform traffic and smart internet traffic offload, in addition to energy efficiency and CAPX/OPEX cost saving, can be addressed as C-RAN advantages.

However, there are challenges in the design of wireless resources virtualisation, which can be enumerated as wireless channel fading, slice isolation, and slice resource management [LYLX12]. Wireless channels are subject to random fading and channel noise, consequently, radio links performance are variable unlike wired networks. The crucial part of wireless virtualisation, therefore, is dynamic radio resources sharing among various virtual networks.

Moreover, multiple slices with different requirements co-exist in the same physical infrastructure, which have to be isolated from each other. In other words, any change in one slice due to flow load change, variation of channel condition, and so on, should not lead to a reduction in resources allocated to other virtual instances. The isolation guarantees the right of all operators to run various applications in virtual sub-networks (slices). The realisation of isolation has also many levels: it can be done at the hardware level, which is the lowest one, or at the higher level of time-slots (flow level), traffic (application level), etc. Virtualisation and isolation in lower level offer more efficient resource usage at the cost of increased computational and operational complexity.

The last but not the least challenge in designing virtual networks is related to slice resource objectives and constraints, [LYLX12]. Various slice requirements need to be met simultaneously, therefore, a hypervisor has to run different sub-algorithms for each slice with special constraints and objectives.

Mobile communications systems virtualisation brings benefits to all involved parties:

- **Infrastructure providers**: with virtualisation, large companies, playing both the role of system operator and infrastructure provider, can concentrate on the maintenance of physical equipment, consequently, they can save the work force of running the networks.
- **Virtual Mobile operators**: while deploying and operating a mobile communications network requires enormous investments, sharing infrastructure is very attractive, since the huge investment on hardware and civil construction can be saved; it also enables small companies to step into the market without bulky investments, and virtual mobile communications networks deployment, maintenance, migration and upgrade will be flexible, and even on the fly.
- **End-users**: the diversity of services for end-users can be achieved by increasing the number of operators, and more options are offered to satisfy users' personal demand and subsequently to enjoy services with reasonable prices.

Developing new radio resource management strategies in virtual environments for efficiently sharing physical radio resources among various VNOs is the aim of this thesis. The solution puts emphasis on the optimisation of the usage of radio resources and requirements of VNOs, besides compensating for the reduction of capacity in a heterogeneous access network. A set of optimisation methods is studied, and their usage as a tool is evaluated. Various scenarios are established, and a series of simulations in there scenarios validates the proposed strategy.

In conclusion, the advantages of virtual network are high resource usage, improved system performance, and lower investment capitals [MLXX12]. Network virtualisation will allow VNOs to share the same physical infrastructure and have network co-existence in a flexible, dynamic manner, using the available resources even more efficiently. As in the virtualisation and multi-tasking over a machine, a virtual RAN can also be shared.

1.3 Novelty

The novelty of this work is the new approach for managing radio resources in wireless and mobile networks by means of the virtualisation of the wireless access, allowing one to extend the classical virtualisation of networks. In this thesis, the concept of virtualisation of radio resources is proposed. In order to have end-to-end virtual networks, one suggests the aggregation of all the physical resources from different RATs, creating virtual wireless links. In this novel methodology, VNOs ask for wireless capacity from a set of physical network providers to serve their subscribers, and not having to deal at all with the physical infrastructure. The RAN Provider (RANP) owns the physical infrastructure, offering Capacity-as-a-Service to VNOs on-demand. The advantages of RAN virtualisation, compared to RAN sharing (where each operator is allocated a portion of spectrum), comes from network element abstraction, isolation among virtual instances, and the ability to support multi-RATs.

In the same research path, a model for the management of virtual radio resources that is capable of supporting both non-cellular (e.g., Wi-Fi) and cellular (e.g., GSM, UMTS, LTE, and whatever comes next in 5G) networks is proposed. The model has two key parts: estimation of available radio resources and their allocation to various services from VNOs. In the first part, one obtains the estimation of network capacity based on the available radio resources from different RATs. In addition to a general approach (i.e., no presumption on MT channel quality), three more approaches are also considered, i.e., optimistic, realistic, and pessimistic. By having a practical estimation of available resources, it allocates to each service from each VNO a portion of the available capacity based on the VNOs' Service Level Agreements (SLAs). The presented model satisfies SLAs when there are enough resources available, and minimises SLAs violations in resource shortage cases; in both cases, the fairness of resource allocation is considered. Model performance is evaluated from different perspectives, by considering a set of practical scenarios.

Finally, in addition to the model, the proposed concept of virtualisation of radio resources was implemented in the Open Air Interface (OAI) [Open15], which is a software-based LTE eNodeB developed in Linux. OAI is an outstanding candidate for C-RAN, and it can act as the physical infrastructure. In order to implement the concept of virtualisation of radio resources, the structures and algorithms of OAI, especially inside the eNodeB scheduler, were modified, the changes enabling OAI to support multiple groups of subscribers in addition to serving them with different policies and priorities; these policies are fed to OAI through an external link from the output of the proposed model. Gathering the statistics and reports for each of these user groups was another modification introduced to OAI. Having a cloud environment with OAI server and the model, practical scenarios were considered for a practical evaluation of the concept and model. Based on the development done in OAI for supporting this concept, it can be claimed that the proposed concept and model are eventually put into practice.

1.4 Research Strategy and Impact

The work developed in this thesis was done within the scope of different European research frameworks and projects, such as the Seventh Framework Programme (FP7-ICT) and Cooperation in the field of Scientific and Technical Research (COST), namely, ICT-MCN [MCN15], ICT-NEWCOM# [NEWC15], and COST Action IC1004 [COST15]. Although all these projects had some work overhead beyond this thesis, they enabled sharing knowledge, visions and experience with multiple researchers of international institutions, namely networks' manufactures, cellular operators, research centres and universities, resulting in multiple cooperative activities and publications.

In the development of this thesis, these projects naturally had a considerable influence over many decisions taken. Reciprocally, the impact of the research activity carried in this thesis had impact on these projects. In this thesis, an architecture for the next generation of V-RAN is proposed, based on the C-RAN architecture, providing multi-tenancy of the RAN infrastructure by means of virtualisation. The implementation of virtualisation of radio resources is the milestone in the creation of end-to-end virtual networks over heterogeneous access infrastructures. The proposed model for Virtual Radio Resource Management (VRRM) was based on the final architecture and use cases of the MCN project: in addition to this project, the model was part of the NEWCOM# joint research, which also considers the proposed optimisation approaches for heterogeneous networks.

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

- Book Section:
    - L.M. Correia, L. Ferreira, "Virtualised and Cloud based Architectures", in N. Cardona (ed.), Cooperative Radio Communications for Green Smart Environments, to be published in 2015.
- International Journals:
    - S. Khatibi, L. Caeiro, L. Ferreira, L.M. Correia, N. Nikaein, "Implementation of Virtual Radio Resource Management for 5G Cloud RAN", *submitted to EURASIP Journal on Wireless Communications and Networking*.

- S. Khatibi, L.M. Correia, “Modelling Virtual Radio Resource Management in Full Heterogeneous Networks”, submitted to *IEEE Transactions on Networks and Service Management*, July 2015
- S. Khatibi, L.M. Correia, “A Model for Virtual Radio Resource Management in Virtual RAN”, *EURASIP Journal on Wireless Communications and Networking*, Vol. 68, Mar. 2015, pp. 01-12.

- International Conferences:
  - S. Khatibi, L.M. Correia, “The Effect of Channel Quality on Virtual Radio Resource Management”, in *Proc. of VTC2015-Fall – 82nd IEEE Vehicular Technology Conference,* Boston, MA, USA, Sep. 2015.
  - S. Khatibi, L.M. Correia, “Modelling Virtual Radio Resource Management with Traffic Offloading Support”, in *Proc. of EuCNC’15 – IEEE 24th European Conference on Networks and Communications*, Paris, France, June 2015.
  - S. Khatibi, L.M. Correia, “Modelling of Virtual Radio Resource Management for Cellular Heterogeneous Access Networks”, in *Proc. of PIMRC’14 – 25th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications,* Washington, DC, USA, Sep. 2014.

The main contributions made within the European research projects were the following ones:

- Mobile Cloud Networking (MCN) [MCN15], an FP7-ICT large-scale integrating project (2012-2015):
  - A. Georgiev (ed.), “*D6.4 Final Report on Testbeds Experimentation and Evaluation*”, Oct. 2015.
  - T. Metsch (ed.), “*D3.4 Infrastructure Management Foundations – Final Report on Component Design and Implementation”*, Apr. 2015.
  - T. Metsch (ed.), “*D3.3 Infrastructure Management Foundations – Components Final Release”*, Dec. 2014.
  - C. Prada (ed.), “*D4.3 Algorithms and Mechanisms for the Mobile Network Cloud”,* Nov. 2014.
  - T. Metsch (ed.), “*D3.2 Infrastructure Management Foundations – Components First Release”*, Apr. 2014.
  - T. Metsch, Peter Gray (eds.), “*D3.1 Infrastructure Management Foundations – Specifications & Design for Mobile Cloud framework”*, Nov. 2013.
  - J. Carapinha, Cc Parada (eds.), “*D2.1 Reference Scenarios and Technical System Requirements Definition”*, Apr. 2013.
- Network of Excellence in Wireless Communications (NEWCOM#) [NEWC15], an FP7-ICT network of excellence project (2012-2015):
  - M. Kieffer (Ed.), "*Final Report on Performance Limits of Wireless Communication"* Deliverable WP1.1 - NEWCOM#, 2015
  - P. Duhamel (Ed.), "*Consolidated Results on the performance limits of wireless communications"* Deliverable D11.2, WP1.1 - NECOM#, 2014.

    - M. Debbah and K. Hamidouche (Eds.), "*Fundamental Issues and related results on the performance of Wireless Networks"* Deliverable D11.1, WP1.1 - NEWCOM#, 2014.
- COST Action IC1004 [COST15] (2011-2015)
    - S. Khatibi, L.M. Correia, "*The Effect of Channel Quality on Virtual Radio Resource Management"*, Technical Report TD- 15-12034, Dublin, Ireland, Jan. 2015.
    - S. Khatibi, L.M. Correia, "*Modelling Virtual Radio Resource Management with Traffic Offloading Support"*, Technical Report TD- 14-11026, Krakow, Poland, Sep. 2014.
    - S. Khatibi, L.M. Correia, "*Virtualisation of Radio Resources – Next Step in Virtualisation of RAN*", Technical Report TD- 14-09044, Ferrara, Italy, Feb. 2014.

1.5 Content

This thesis is structured into eight chapters and two annexes. The current chapter presents an introduction to the thesis by presenting a brief history of wireless networks in Section 1.1. Section 1.2 addresses the motivation and objectives for 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 European projects contributions and published work. Finally, this Section 1.5 contains the dissertation contents in detail.

Chapter 2 provides an overview on RATs and their specifications, followed by a very brief review on common wireless networks. Section 2.1 presents a brief review on the various radio interfaces in wireless communications, followed by the description of the common wireless network architectures in Section 0. A short introduction of QoS and Quality of Experience (QoE), in addition to QoS traffic classes in different networks are the contents of Section 2.3.

An overview on the state of the art, algorithms, and strategies in virtualisation of RAN are the focus of Chapter 3. Section 3.1 addresses the concept, advantages, and drawbacks of RAN sharing. Section 3.2 presents the motivation, concept and related work in RAN virtualisation. The new architectures for RANs are the key topic in Section 3.3. A very brief overview on radio resource management techniques in Section 3.4 is the final part of this chapter.

Chapter 4 presents the concept of virtualisation of radio resources in addition to a management model by a brief description for the chosen architecture for virtual RAN, in addition to different resource management strategies and their relations to each other in Section 4.1. Section 4.2 summarises the design choices and Section 4.3 comprehensively describes the analytical model for managing virtual radio resources. Furthermore, Section 4.4 presents the partial VRRM, an approach for solve the modelled problem, and finally, Section 4.5 introduces the evaluation metrics.

Chapter 5 deals with the implementation of the algorithms and the management model. Section 5.1 provides an overview on the implementation of the VRRM Server and Open Air Interface (OAI). Section 5.2 discusses the details and algorithms of the VRRM server and its application. The changes in OAI to support multiple VNOs are described in Section 5.3. In the end, Section 5.4 reviews the assessment procedure of the model implementation.

Chapter 6 aims at presenting the used reference scenarios in addition to numeric results achieved by using the developed model in this thesis. Section 6.1 presents the reference scenario, followed by the scenarios’ road map in Section 6.2. The numeric results are presented in Section 5.4 to Section 6.11.

Chapter 7 is dedicated to the analysis of simulation results. Section 7.1 addresses the simulation scenario and the road map. Section 7.2 addresses the assessment procedure of the simulator. Finally, the results are presented in Section 7.3.

The main conclusions of the thesis are presented in Chapter 8. Section 8.1 presents a summary of the thesis. Section 8.2 presents the main results including the novel results of the work developed within this thesis. Section 8.3 points out aspects to be addressed in future work.

At last, there are two annexes: Annex A presents the system SINR and data rate model for different mobile network; Annex B presents the modelling of the traffic sources used in the emulations.

Chapter 2

Overview of Systems

This provides an overview on RATs and their specifications, followed by a very brief review on common wireless networks. Section 2.1 presents a brief review on the various radio interfaces in wireless communications, followed by the description of the common wireless network architectures in Section 0. A short introduction of QoS and Quality of Experience (QoE), in addition to QoS traffic classes in different networks are the contents of Section 2.3.

2.1 Radio Interfaces

The various multiple radio access techniques, presented in Figure 2.1 [BuBu06], can be divided in two main categories: conflict-free (or polling) methods and contention-based protocols. Conflict-free radio access methods have the advantage over contention-based ones, by using the available resources efficiently during high load periods. However, for the realisation of on-demand channel assignment, a random access uplink channel is always needed. It can be claimed that using the contention-based protocols besides the conflict-free techniques is unavoidable.

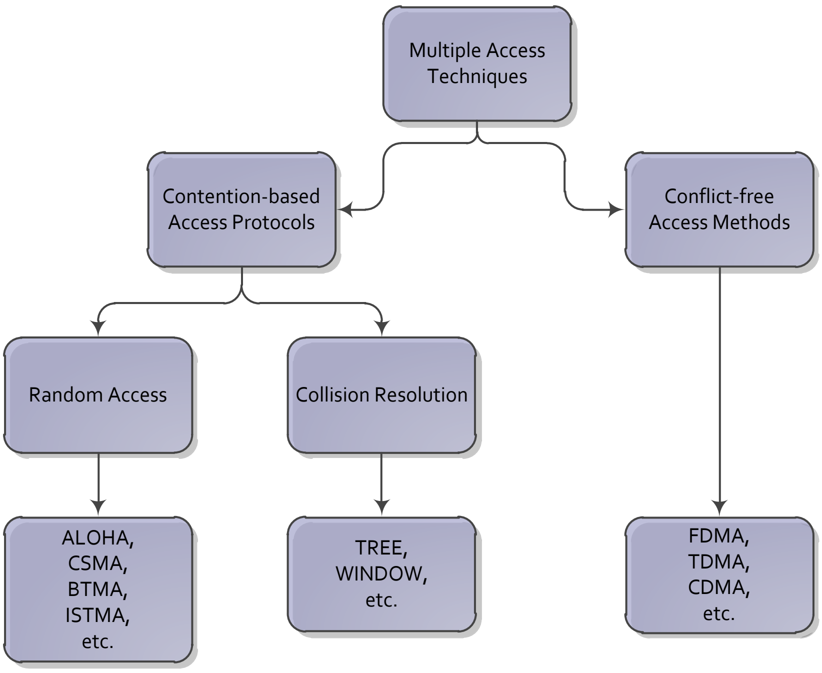


Figure 2.1 – Classification of multiple access technique (based on [BuBu06]).

The orthogonal channels, in which terminals can communicate without imposing any interference on each other, are the fundamental perspective of conflict-free access methods. In FDMA, the continuous time frequency channels are the orthogonal channels, the data rate of each channel, R_b, being [Sing10]:

$$R_{b[\text{bps}]} = \alpha_{[\text{bps/Hz}]} \Delta f_{[\text{Hz}]} \tag{2.1}$$

where:

- α :spectral efficiency,
- Δf: channel bandwidth.

In TDMA, non-overlapping time intervals are allocated to users on the same frequency band. It is a common practice to combine FDMA with TDMA, as it is done in GSM. The bandwidth of channels in GSM is 200 kHz, each carrier is divided into 8 time-slots, [Saut10], and the frame is 4.615 ms long, with 8 bursts of 577 μs. The channel data rate is:

$$R_{b[\mathrm{bps}]} = \frac{N_t \;\; N_{b/sym}}{\Delta t_{[\mathrm{s}]}} \tag{2.2}$$

where:

- $N_{b/sym}$: number of bits per symbol,
- N_t: number of time-slots,
- Δt: duration of each time-slot.

In CDMA systems, orthogonal channels are not separated in the time or frequency domains, but rather being based on pseudo-random waveforms using spreading codes to create noise-like transmissions [Kaar05]. The rate of the spreading code is referred to as chip rate, and the Spreading Factor is the ratio of the chips to baseband information rates

$$\alpha_s = \frac{R_{c[\mathrm{cps}]}}{R_{b[\mathrm{bps}]}} \tag{2.3}$$

where:

- α_s: spreading factor;
- R_c: chip rate;

Table 2.1 briefly presents the specification of UMTS Terrestrial Radio Access Network (UTRAN), as an example for a CDMA-based system.

Table 2.1 – UTRAN Specification (based on [3GPP11]).

Parameter	Value
Frame length [ms]	2
Number of time-slots in a frame	15
Chip rate [Mcps]	3.84
Normal Carrier Spacing [MHz]	5

Orthogonal Frequency Division Multiplexing (OFDM) uses multiple subcarriers for transmit high data rates, but it takes advantage of the Inversed Discrete Fourier Transfer (IDFT) to generate the subcarriers and of the Discrete Fourier Transfer (DFT) to extract them at the receiver [PrVe10]. The elimination of frequency guard band leads to a more efficient spectrum usage. In OFDM, all subcarriers are dedicated to a single user, an example being WiFi. In OFDMA, not all subcarriers are assigned to a single user, rather being divided into groups that are assigned to different users. The data rate of the OFDMA based LTE, in each subcarrier, is given by:

$$R_{b[bps]} = \sum_{i=1}^{N_c} R_{b[\mathrm{bps}]}^{c_i} \tag{2.4}$$

where:

- N_c: Number of subcarriers;
- $R_b^{c_i}$: data rate of *i*-th subcarrier.

Table 2.2 presents the physical layer of LTE, while the physical layer of WiFi based on 802.11ac is described in Table 2.3.

Table 2.2 – Physical Layer of LTE (based on [SeTB11]).

Parameter	Value
Channel Bandwidth [MHz]	1.4, 3, 5, 10, 15, 20
Subcarrier Spacing [kHz]	15
CP Length [μs]	5
FFT order (for 20MHz)	2048

Table 2.3 – Physical layer of WLAN 802.11ac (based on [BeKM13]).

Parameter	Value
Channel Bandwidth [MHz]	20, 40, 80, 160, 80 +80
Subcarrier Spacing [kHz]	312.5
Symbol Duration[ns]	3.6
FFT order (for 160MHz)	512

Table 2.4 presents the maximum data rates of different radio interfaces used in different networks with different radio interfaces.

Table 2.4 – Maximum data rates for different radio interfaces (based on [Cisc12], [Jaci09]).

System	Radio Interface	Max. Data Rate [Mbps]
WLAN (802.11ac)	OFDM	1300
GSM	TDMA / FDMA	0.058
UMTS	CDMA	43
LTE	OFDMA	1

Finally, the standard frequency bands for the aforementioned systems (i.e, GSM, UMTS, and LTE) in different geographical regions are summarised in Table 2.5.

Table 2.5 – Different systems frequency bands (extracted from [Saut10], [DaPS11], and [Kaar05]).

Band	Uplink [MHz]	Downlink [MHz]
GSM 900 (primary)	890 –915	935 – 900
GSM 900 (extended)	880 – 915	925 – 960
GSM 1800	1710 – 1785	1825 – 1880
GSM 1900 (North America)	1850 – 1910	1830 – 1990
GSM 850 (North America)	824 – 849	869 – 894
GSM-R	876 – 915	921 – 960
UMTS (core band)	1920 – 1980	2110 – 2170
UMTS 1900	1850 – 1910	1930 – 1990
UMTS 1800	1710 – 1785	1805 – 1880
UMTS 1.7/2.1 GHz (USA)	1710 – 1770	2110 – 2170
UMTS 850	824 – 849	869 – 894
UMTS 800 (Japan)	830 – 840	875 – 885
LTE Band 1 (Europe, Asia)	1920 – 1980	2110 – 2170
LTE Band 2 (Americas , Asia)	1850 – 1910	1930 – 1990
LTE Band 3 (Europe, Asia, Americas)	1710 – 1785	1805 – 1880
LTE Band 4 (Americas)	1710 – 1755	2110 – 2155
LTE Band 5 (America)	824 – 849	869 – 894
LTE Band 6 (Japan – only for UTRA)	830 – 840	875 – 885
LTE Band 7 (Europe, Asia)	2500 – 2570	2620 – 2690
LTE Band 8 (Europe, Asia)	880 – 915	925 – 960
LTE Band 9 (Japan)	1749.9 – 1784.9	1844.9 – 1879.9

Table 2.5 – Different systems frequency bands (extracted from [Saut10], [DaPS11], and [Kaar05]).

Band	Uplink [MHz]	Downlink [MHz]
LTE Band 10 (Americas)	1710 – 1770	2110 – 2170
LTE Band 11 (Japan)	1427.9 – 1447.9	1475.9 – 1495.9
LTE Band 12 (USA)	698 – 716	728 – 746
LTE Band 13 (USA)	777 – 798	746 – 756
LTE Band 14 (USA)	788 – 798	758 – 768
LTE Band 17 (USA))	704 – 716	734 – 746
LTE Band 18 (Japan)	815 – 830	860 – 875
LTE Band 19 (Japan)	830 – 845	875 – 890
LTE Band 20 (Europe)	832 – 862	791 – 821
LTE Band 21 (Japan)	1447.9 – 1462.9	1495.9 – 1510.9

2.2 Network Architecture

2.2.1 GSM and UMTS

The GSM network is a successful cellular phone technology for a variety of reasons, including the ability to roam worldwide. The architecture of UMTS is also based on GSM's. Figure 2.2 presents the architecture of these two networks, the two key blocks being the RAN and the Core Network (CN). Radio coverage and signal processing over a given geographical region are the roles of RAN, while CN manages the calls between its users and other users, such as other mobile users, ISDN (Integrated Service Digital Network) users, PSTN (Public Switch Telephone Networks) users, etc.

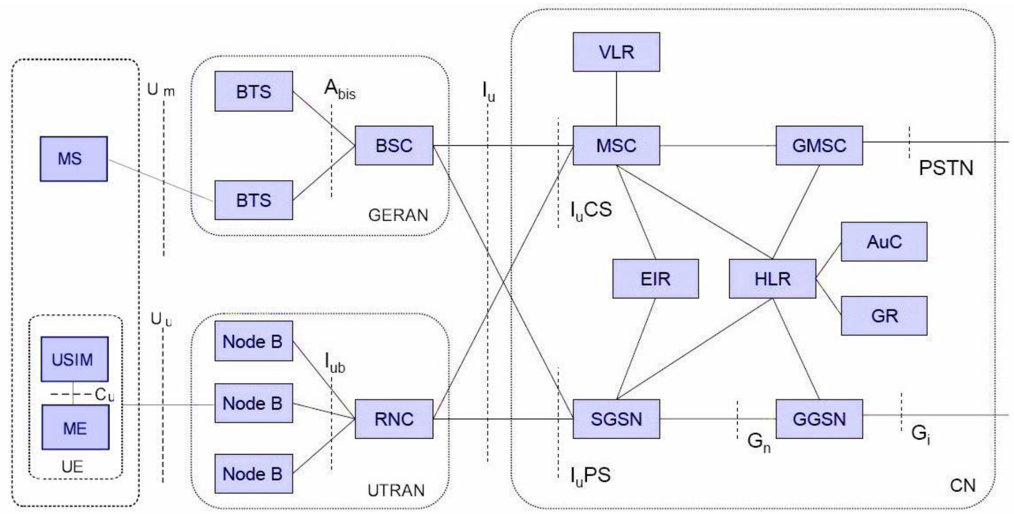


Figure 2.2 – GSM/EDGE and UMTS architecture (based on[Sing10]).

The key elements in GSM's network are:

- MSC (Mobile Switching Centre): it is the heart of the network, which is responsible for call setup and basic switching. MSC also manages communication between GSM and other networks.
- HLR (Home Location Register): a database located in the user's home system that stores a master copy of the user's service profile.

- VLR (Visitor Location Register): it is the database that serves MTs in their current location for circuit switched services. VLR contains information from a subscriber's HLR necessary to provide the subscribed services to visiting users.
- GMSC (Gateway MSC): it is the switch that connects the core to external circuit-switched networks.
- SGSN (Serving GPRS Support Node): its functionality is similar to the one of HLR, but it is used for packet switch services.
- AuC (Authentication Centre): it provides the parameters needed for authentication and encryption functions.
- EIR (Equipment Identity Register): it stores security-sensitive information about the mobile equipment.

UTRAN RAN, the main difference between GSM and UMTS, is divided in between Node B and RNC (Radio Network Controller). Node B converts data flows between UTRAN and core network in addition to MT and UTRAN; it also participates in radio resource management [GoSa05]. The functionality of RNC is radio resource management in the UMTS network, owning and controlling radio resources in Nodes B. The RNC is a service access point for all services (e.g., management of connections to the MT) that UTRAN provides to the CN.

2.2.2 LTE – The Long Term Evolution

In contrast to the circuit–switched model of previous cellular systems, the primary aim of LTE is to provide seamless IP connectivity between MT and the Packet Data Network (PDN) without any disruption to the end-users' application during mobility [SeTB11]. While the term 'LTE' encompasses the evolution of the radio access through the Evolved-UTRAN (E-UTRAN), it is accompanied by an evolution of the non-radio aspects under the term System Architecture Evolution (SAE), which includes the Evolved Packet Core (EPC).

Figure 2.3 illustrates the overall network architecture including network elements and the standardised interfaces; the access network contains only one essential node, which is the eNodeB, connecting MTs to the network. Each of the network components is connected through standardised interfaces in order to allow multivendor interoperability, which gives operators the possibility to purchase different network elements from different vendors. CN is responsible for the overall control of the MT and establishment of bearers. The main logical nodes of EPC are:

- MME: it is the control node, which processes the signalling between the MT and the CN. The protocols running between the MT and the CN are known as the non-access Stratum protocols. The main functions supported by the MME are related to bearer and connection management. The former includes the establishment, maintenance, and release of bearers, being handled by the session management layer in the Non-Access Stratum (NAS).
- Authentication and Security: on the first registry request of an MT, the MME searches Home Subscription Server (HSS) for related authentication vector based permanent identity of the MT.

The MT's permanent identity, known as International Mobile Subscriber Identity (IMSI), can be retrieve either from a former visited network or the MT itself. The MME may repeat the authentication procedure when needed or periodically. To protect the subscriber privacy and minimise sending IMSI over the air, the MME also allocates each MT a temporary identity called Globally Unique Temporary Identity (GUTI), which may be reallocated to prevent unauthorised MT tracking.

- Mobility Management: by registration of an MT in the network, the MME signals the HSS with current location of subscriber. It also requests resources to be set up on the eNodeB besides the Serving Gateway (S-GW) serving the MT. The MME also creates an entry and keeps track of MTs in the serving area on eNodeB levels for subscribers in active communication and in the Tracking Area (TA) for idle users. The MME also participates in the control signalling for handover of an active MT among eNodeBs, S-GWs or MMEs. The MT informs the MME as it moves in between different TAs. When data for an idle MT are received from an external network, the MME is notified and requests all eNodeBs in the TA which MT is stored to page the subscriber.
- Managing Subscription Profile and Service Connectivity: the MME is in charge of retrieving subscribing profile from the home network. This information is used for determining which Packet Data Network connection should be allocated to serve the user in addition to storing MT serving duration.
- Policy and Charging Rule Function (PCRF): it is in charge of policy control decision-making, as well as, for controlling the flow-based charging functionalities in the Policy Control Enforcement Function (PCEF), which resides in the P-GW. The PCRF provides the QoS authorisation (QoS class identifier and bitrates) that decides how a certain data flow will be treated in the PCEF and ensures that this is in accordance with the user's subscription profile.
- PDN Gateway (P-GW): it is responsible for IP address allocation for the MT, as well as QoS enforcement and flow-based charging according to rules from the PCRF.
- S-GW: all user IP packets are transferred through the S-GW, which serves as the local mobility anchor for the data bearers when the MT moves in between eNodeBs. It also retains the information about bearers when the MT is in idle state. In addition, the S-GW performs some administrative functions in the visited network, such as collecting information for charging (e.g. the volume of data sent to or received from the user), and legal interception. It also serves as the mobility anchor for inter-working with other 3GPP technologies, such as GSM and UMTS.
- eNodeB: the only node in E-UTRAN, is a BS in charge of all radio related functions in the fixed part of the network. From a functional perspective, the eNodeB can be counted as a layer 2 bridge between MT and EPC, relaying the radio connection and the corresponding IP base connectivity and being the termination point for all radio protocols. The eNodeB performs ciphering/deciphering of the User Plane (UP) in addition to IP header compression/decompression, and is responsible for Radio Resource Management (RRM).

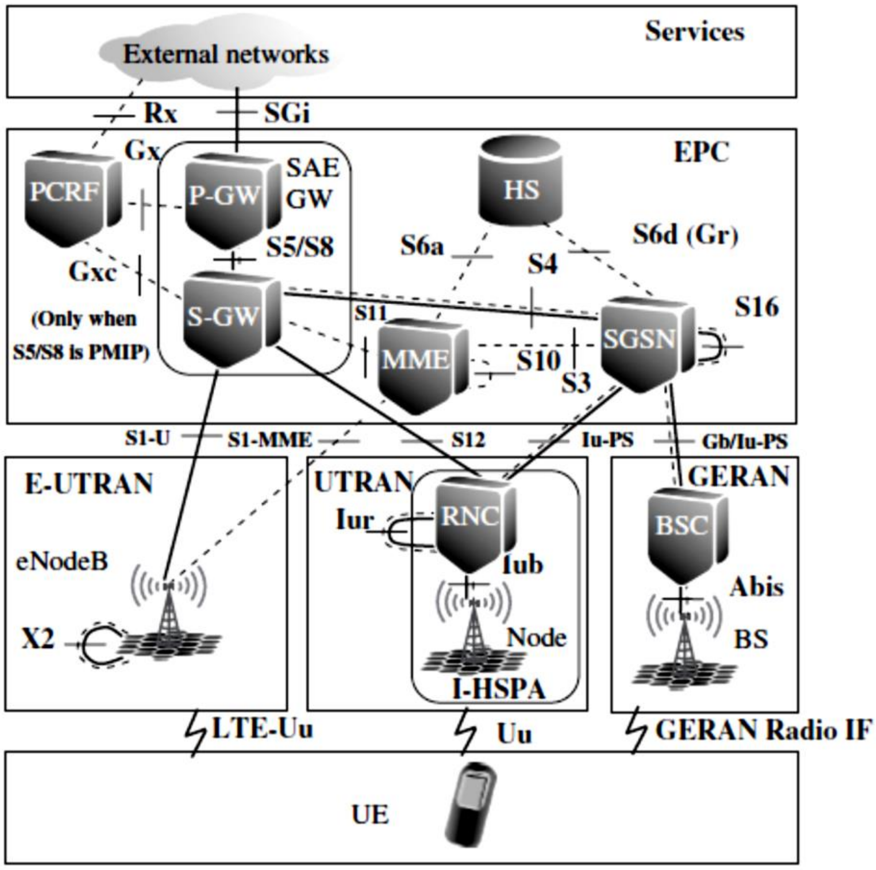


Figure 2.3 – System Architecture for 3GPP access networks (extracted from [HoTo10]).

Figure 2.3 also describes the network architecture and elements considering all 3GPP access networks (E-UTRAN, UTRAN, and GERAN) connected to EPC, known as 3GPP Inter-working System Architecture Configuration, which enables optimised inter-working between the mentioned accesses. From the end-user perspective, the connectivity services provided by these access networks are very similar, with exception of data rates and performance. However, from the architectural perspective, these access networks are very different, hence, new interfaces, namely S3, S4 and S12, and functions are needed in the EPC to connect and inter-work with UTRAN and GERAN, and the same is required for GERAN and UTRAN. Regarding mobility between access networks, in the case of handover from E-UTRAN to UTRAN/GERAN, the configuration into the eNodeB is required from the neighbouring cells from the other access networks. After analysing the measured signal levels of UTRAN/GERAN from the MT, if the eNodeB decides to start the handover procedure, a signal is sent to the MME, which sends to the eNodeB the information needed for the Handover Command from the target access system. This information is directly sent to the MT by the eNodeB without interpreting it.

Figure 2.4 describes the generic non-3GPP Inter-working System Architecture, which relies only on loose coupling with generic interfacing means, and without access network level interfaces. Since there are so many different kinds of access networks, they have been categorised into trusted and untrusted non-3GPP access networks, depending on the authentication method. If it can be safely assumed that 3GPP defined authentication can be run by the network, it is considered trusted, otherwise, it is considered untrusted. This characterisation influences the interface to which non-3GPP access networks connect to the P-GW, which maintains the role of mobility anchor.

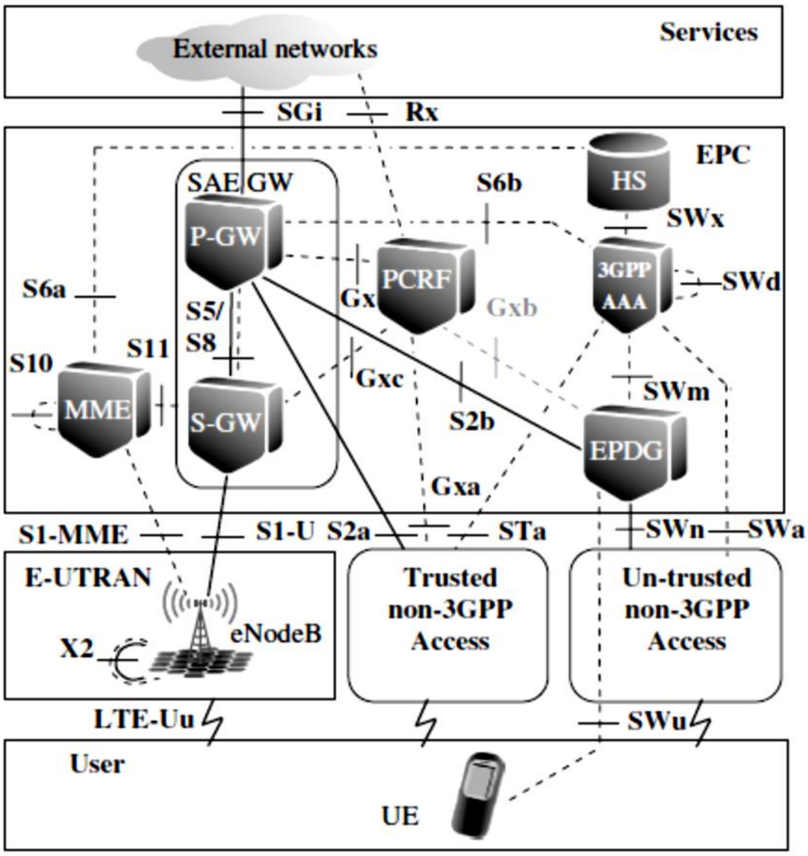


Figure 2.4 – Architecture for 3GPP/non-3GPP access networks (extracted from [HoTo10]).

2.2.3 Wireless LAN Architecture

There are two main architectures for Wireless LANs (WLANs), as infrastructure or ad hoc networks [Wise09]. In an infrastructure network, client stations communicate with each other through an AP and there are no direct communications, opposite to the ad hoc approach. The AP itself is connected to a wired LAN, thus, the client can also communicate with the wired LAN network. The set of an AP and its connected clients are referred to as Basic Service Set (BSS). Some time is needed to extend the wireless coverage by using a number of APs from a single subnet, the collection of BSSs in this situation being known as Extended Service.

IEEE 802.11ac, the most resent WLAN standard, has been proposed to enhance the former one, IEEE 802.11n, beyond Gbps rates. In contrast to all amendments of 802.11 standards, the goal in 802.11ac is to improve the total network throughput as well as individual link performance. The key features that allow 802.11ac to obtain Gbps transmission rates [BeKM13], are:

- Static and dynamic channel bonding,
- Multi-user multiple-input multiple-output (MU-MIMO).

Ad hoc networks consist of a group of randomly distribute nodes, used with wireless transceivers that enable them to send, received or repeat signals. These nodes can communicate directly to/from a wireless network without relying on any infrastructure, in contrast to infrastructure networks. Ad hoc is a kind of WLAN common for meeting temporary communication needs, but using infrastructure WLANs is more common.

2.3 Quality of Service and Quality of Experience

The main idea behind the next-generation of computer and communication networks is offering universal and easy access internet services over a single multi-service internet [ChFY04]. In other words, all form of communications are bounded into a single-service platform over internet technology. Although all services are based on transmitting and receiving bit streams, each application needs its own service characteristic. In order to offer an acceptable service to end-users, different requirements have to be met, hence, QoS and QoE are key topics in network design.

QoS is the ability of the network to provide a service at a guaranteed service level. It includes all functions, mechanisms, and procedures in the wireless networks and terminals that ensure the provision of the negotiated service quality between the MT and the network. QoE, on the other hand, is how satisfied the user is with a service. This satisfaction, for instance, can be in terms of usability, accessibility, retain-ability and integrity of the service [SoLC06]. To summarise, the key aim of the network and services is achieving maximum user's QoE, while QoS is the main building block for reaching that goal effectively. QoS and QoE are interdependent and they have to be studied together.

QoE essentially is a subjective evaluation, and measuring it accurately entails a careful mapping of QoS subsequently to a set of Key Performance Indicators (KPIs), which are derived from the discrete performance parameters of network sub-systems. KPIs are logically grouped to obtain Key Quality Indicators (KQIs). It is possible that one KPI be part of multiple KQIs as shown below. These KQIs enable the operation team to understand the high-level system performance and the effect on customer services. KQI may be mapped to various QoS categories, which gives the degree of QoE.

Generally timeliness, precision and accuracy are the three attributes to measure QoS in a process output [Nong02]: timeliness measures the time taken to produce the output of the process, precision the amount or quantity of the produce output, and accuracy the correctness of the produced output. However, work on QoS based on the aforementioned attributes uses the following measures:

- Response time expected by users is the time elapsed between sending a request and the reception of the first response by the user.
- Delay (network delay) is the time difference between the transmission of the first bit of a data block from the transmitter end system and its reception at the other end system.
- Jitter is referred to the variation of delay generated by the transmission equipment.
- Data Rate refers to the raw data rate of encoded multimedia data before transmission.

- Loss Rate is the number of bits lost between two points after transmission.
- Error Rate is the frequency of erroneous bits between two points.

Various applications and services in the network are classified according to their technological attributes. Each QoS-enabled network has its own QoS classes, which can easily be mapped based on their similarity. Generally, applications are divided into two categories: real time and non-real time, depending on their timing requirements. In real time applications, the latency of these applications has to be sufficiently small for the acceptable timeliness, and time-based information has to be delivered without changing its built-in time property, hence, for user satisfaction, it is needed to maintain stringent delay and jitter requirements. Non-real time applications, on the other hand, do not fail if timeliness metrics are not met, i.e., they do not required timing accuracy to be considerably acceptable. UMTS defines four major traffic classes: conversational, streaming, interactive, and background. Table 2.6 briefly discusses the key characteristics of each class.

Table 2.6 – UMTS traffic classes (based on [BaLu11]).

Traffic Class	Fundamental Characteristics	Example
Conversational	Preserves time relation (variation) between information entities of the stream. Conversational pattern (stringent and low delay).	VoIP
Streaming	Preserves time relation (variation) between information entities of stream.	Streaming video
Interactive	Requests response pattern. Preserves payload content.	Web browsing
Background	Destination is not expecting the data within a certain time. Preserves payload content.	Background non-real-time downloads

The first service class is conversational real-time. The services of this group are delay sensitive with interactive (or conversational) pattern, such as Voice over IP (VoIP). The traffic of these services has low delay tolerance, so RRM algorithms have to preserve time relation (variation) in their traffic flow. Because of their conversational pattern, the improvement of offered QoS by means of buffering is not an option. Although, the generated traffic rates for these services are dependent on the applied codec, for the sake of simplicity, it is assumed that user satisfaction can be achieved as long as user access with the service nominal data rate (or the maximum data rate in case of variable rate codecs) is granted. Therefore, the resource manager has to meet the minimum required data rate by allocating sufficient radio resources to these services. In addition to data rate, the response time (the time interval between receiving the user's request until granting access) also plays an important role in the offered QoS. The users' requests for a service have to be replayed in a time interval less than the expected response time.

Streaming services, services of the second group, likewise conversational ones are also delay sensitive; yet their traffic flow is highly asymmetric. The resource manager can serve streaming services easier by means of buffering. As long as the total offered data during any time interval from the starting of the service is larger or equal to the amount of data offered by the fixed nominal data rate during the same time interval, user's satisfaction can be obtained.

Services placed in the third class, interactive ones, are more tolerant to end-to-end delays, such as web browsing and FTP (File Transfer Protocol). Their delay tolerance allows the reduction of the assigned data rate or even pausing traffic flows (temporarily) to be an acceptable approach for RRM. QoS, though, forces the manager to serve the requesting user with at least a minimum data rate in a maximum time interval, as summarised in Appendix A. In order to have a common notation with other service classes, and for the sake of simplicity, the nominal data rate for interactive best effort services is referred to as minimum acceptable data rate, as presented in Appendix A. Therefore, it can be claimed that the offered data rate to each session of these services has to be higher than (or at least equal to) their nominal data rate. In addition, the serving time of each session should stay smaller than the maximum acceptable serving time for the specific service. In addition to all above, as in the previous service classes, response time is the last QoS metric for this service class.

Background traffic, the last service group, is the one with almost no delay or serving time limitation, having the lowest priority in a network. The source and destination of this kind of services do not require any minimum data rate or maximum serving time, hence, traffic being usually handled during low load periods of a network. For resource manager, these services almost do not impose any limitation or constraint, however, it is logical to assume an expiration time in the order of hour, day, or week for this kind of service requests.

LTE uses more QoS classes than UMTS, each one being defined by a QoS Class Indicator (QCI) and serving priority, Table 2.7; in this table, the mapping between UMTS' service classes and LTE ones is also presented.

Table 2.7 – LTE traffic classes (based on from [BaLu11]).

<table>
<tr><th>QCI</th><th>Resource Type</th><th>Priority</th><th>Delay Budget [ms]</th><th>Error Loss Ratio</th><th>Example Services</th><th>Equivalent UMTS Class</th></tr>
<tr><td>1</td><td rowspan="4">GBR</td><td>2</td><td>100</td><td>10^{-2}</td><td>VoIP</td><td rowspan="2">Conversational</td></tr>
<tr><td>2</td><td>4</td><td>150</td><td>10^{-3}</td><td>Video Calling</td></tr>
<tr><td>3</td><td>5</td><td>50</td><td>10^{-3}</td><td>Real-time gaming</td><td>Streaming / conversational</td></tr>
<tr><td>4</td><td>5</td><td>300</td><td>10^{-5}</td><td>Video streaming</td><td>Streaming / Interactive</td></tr>
<tr><td>5</td><td rowspan="5">Non-GBR</td><td>1</td><td>100</td><td>10^{-3}</td><td>IMS signalling</td><td rowspan="2">Interactive</td></tr>
<tr><td>6</td><td>6</td><td>300</td><td>10^{-3}</td><td>Video streaming TCP-based</td></tr>
<tr><td>7</td><td>7</td><td>100</td><td>10^{-5}</td><td>Voice, Video (live streaming interactive gaming)</td><td rowspan="3">Background</td></tr>
<tr><td>8</td><td>8</td><td>300</td><td>10^{-3}</td><td rowspan="2">Video (buffered, streaming), TCP-based</td></tr>
<tr><td>9</td><td>9</td><td>300</td><td>10^{-5}</td></tr>
</table>

Chapter 3

A Review on Virtual Radio Access Network State of the Art

This chapter provides an overview on the state of the art, algorithms, and strategies in the virtualisation of RAN. Section 3.1 addresses the concept, advantages, and drawbacks of RAN sharing. Section 3.2 presents the motivation, concept and related work in RAN virtualisation. The new architectures for RANs are the key topic in Section 3.3. A very brief overview on RRM techniques in Section 3.4 is the final part of this chapter.

3.1 RAN Sharing

Due to the increase of mobile data demand coming from the proliferation of smart devices and traffic-hungry applications, the monthly global data traffic is predicted to surpass 10 ExaByte in 2017 [Cisc13a]. The deployment of dense BSs to increase cellular networks capacity is fundamental in any possible approach. In order to decrease CAPEX and OPEX costs, operators have been looking into the possibility of sharing their network infrastructures, in order to reduce their costs. Network sharing is an agreement between operators and should be transparent to end-users. In other words, an MT needs to be able to discriminate between available network operators in a shared radio access network. Based on [ViWC02], infrastructure sharing can be categorised into three main areas: passive component sharing, active element sharing and geographical sharing, Figure 3.1. Passive sharing refers to the sharing agreement of fundamental establishments, such as tower frame, equipment houses and power supply, in order to reduce operation costs. Active sharing refers to sharing transport infrastructure, radio spectrum, and baseband processing resources. A service area can be divided into several regions, with each operator being in charge of each one, this federation of operation being known as geographical sharing or National Roaming [SAMS11], which leads to achieve a full coverage in a short time.

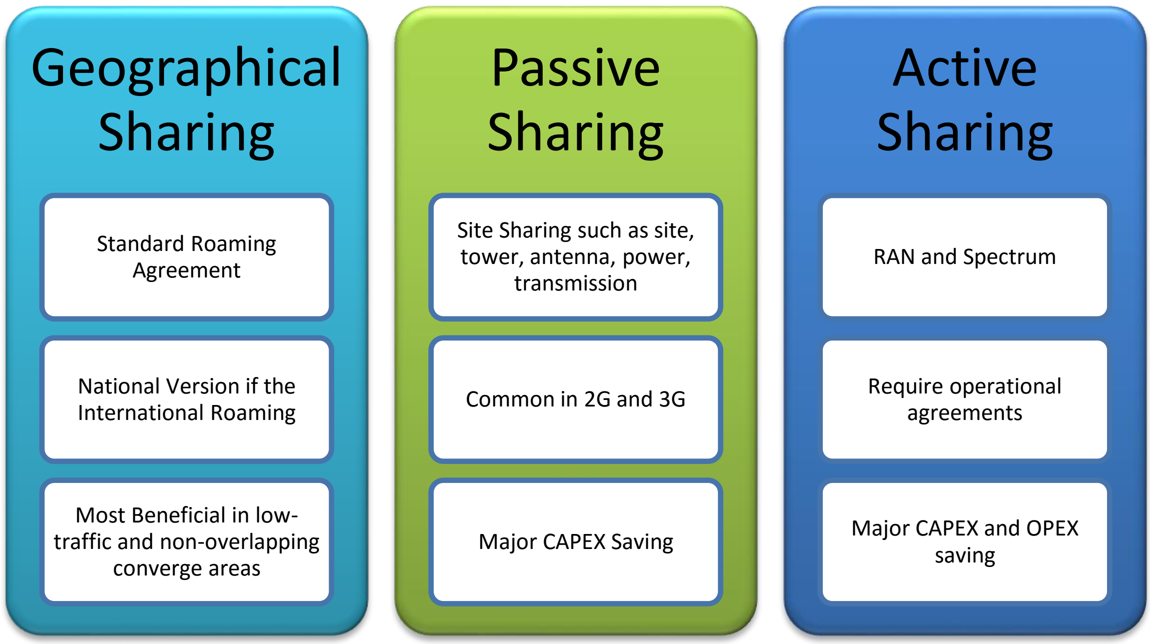


Figure 3.1 – Type of network sharing (based on [SAMS11]).

Two types of active sharing are introduced in [FTAL08], which are Multi-Operator RAN (MORAN) and Multi-Core Network (MOCN): in the former, operators maintain the maximum level of independent control over their traffic quality and capacity, by splitting eNodeB into logically independent units over a single physical infrastructure; in the latter, operators give up their independent control by sharing the aforementioned entities in conjunction with pooling frequencies. Although cost items in MOCN are identical to MORAN, frequency pooling leads to further marginal savings in extremely low-traffic areas over equipment related costs. Moreover, a network-wide RRM framework is proposed [MKHR13] in order to achieve isolation, in addition to an optimal distribution of resources across the network. From a user's perspective, he/she may realise that the network is shared according to [3GPP13].

However, surprisingly few RAN sharing agreements have been made, especially in mature markets, in contrast to its benefits. The reasons offered by operators for not engaging in sharing deals are often the up-front transformation costs, the potential loss of control over their destinies, and the challenge of operational complexity [FPTG12].

Sharing deals are too expensive. The initial cost of a network-sharing deal can be daunting, so the operators that do not have the funds in hand to make the necessary investment are likely to assume that they simply cannot afford to participate. Second, many operators, particularly incumbents whose early entrance into their markets has given them the best coverage and network quality, assume that sharing their network with rivals would dilute their competitive advantages, and they have the impression that they would not be able to control their network in future directions, their rollout strategies, or their choices about hardware and vendors. Last, but not the least, having a shared RAN running properly is an elaborated and complex task: according to some operators, a combined network has too many technological and operational perils, especially during the transition period. In conclusion, they are concerned that it may lead to little financial benefit and a potential for chaos.

3.2 RAN Virtualisation

3.2.1 Motivation

In addition to the shortage of radio resources to deal with the IP tsunami, operators are facing drastic variations of traffic, both geographically and temporally [GuKM10]. The common provisioning approach used in RANs, i.e., considering only busy hours, is no longer acceptable, since it leads to an inefficient resource usage with relatively high CAPEX and OPEX costs. In contrast, operators are looking for more flexible and elastic solutions, where they can also share their physical infrastructure.

The current direction of designing the Future Internet is favouring a multiple co-existing architecture, where each one is designed and customised to satisfy a specific type network requirement, rather than trying to achieve a global architecture that fits all [ZLGT10]. Recently, the sharing of network infrastructure using NFV has become an active research topic, being meant to transform the way operators architect their networks [MDPA12]. Wireless network virtualisation can shorten the process of research and innovation for new technologies, by supporting powerful and efficient systems [ChYu15]:

- The flexibility, programmability, and customisation of virtual networks enable an easier and faster evaluation of the new proposed networking technologies without considering the complicated interfaces and characteristics of physical infrastructures.
- By means of the isolation among virtual networks, various functions can be run and operated simultaneously, even in the real infrastructure, without disturbing the normal services.

3.2.2 The Concept of RAN Virtualisation

A network environment supports virtualisation by allowing the coexistence of multiple virtual networks in the same physical infrastructure [ChBo10]. Network virtualisation proposes the decoupling of the functionalities of the traditional network operator into physical infrastructure and service provider [ChBo10, ChYu15]. Addressing mobility in virtual network requires having virtual wireless links as well as other virtual components. While virtualisation for servers, routers, and wire line has been well studied, the wireless component has not yet received major consideration. The concept of wireless network virtualisation can address a very broad scope, including spectrum sharing and infrastructure virtualisation to air interface. Nevertheless, the innovation is to create end-to-end virtual networks, in which users can also have mobility.

Regardless of wired or wireless networks, virtualisation can be considered as the process of splitting the entire network [XiKT13]. On the ground of this definition, the wireless network virtualisation is considered as the technologies within which the physical network infrastructure, including physical radio resources, are abstracted and sliced into virtual wireless network resources, and shared among multiple parties, while holding isolation in addition to certain functionalities [ChYu15]. RAN virtualisation has also received attention, recent studies (e.g.,[Pere13]) considering virtualisation of radio resources. Despite the similarity between virtualisation of RAN and RAN sharing, it cannot be considered as any of the aforementioned sharing types [ZZGT11], the key differences being elements abstraction, instances isolation, and multi-RAT support.

In RAN virtualisation, in contrast to RAN sharing, the physical infrastructures are not transparent to VNOs. Operators have an abstract vision of the physical elements, and they do not deal with the management of the physical infrastructure, requesting from a RANP for a virtual RAN under a set of SLAs. Moreover, the isolation among instances ensures that traffic, and mobility in addition to the fluctuations in channel conditions of the MTs of a VNO do not affect the QoS offered to the other ones. It also enables multiple VNOs with different objectives and requirement to operate in the same physical infrastructure. Achieving a lower downtime is another advantage of having isolation among VNOs.

Last, but not the least, the ability to support multi-RATs is another difference between these two concepts. Wi-Fi traffic offloading, as an example for this matter, cannot be realised for multiple VNOs operating over the same physical infrastructure, since there are not enough available channels (i.e., only one 160 MHz channel for IEEE 802.11ac [BeKM13]) to allocate to each VNO separately. The only practical solution is to apply radio resource virtualisation to realise Wi-Fi virtual links for each VNO.

As pointed out briefly in Section 1.2, the better differentiation of these two concepts, however, can be achieved by comparing network sharing/virtualisation with relative well-known concepts in computers, relatively known as multi-tasking (sharing)/virtualisation. Based on the analogy presented in [KZJK12], a process in an Operating System (OS) is qualitative equivalent of a communication session in a network. In addition, both OS and network represent the intermediaries between the process and the required hardware resources for the former, where in the latter it is between the communication session and the end-to-end path. It is worth nothing that the logical and physical infrastructures, just like the OS and the hardware, are separated. From this perspective, Figure 3.2 illustrates the general envisioned architecture for the both computer and network.

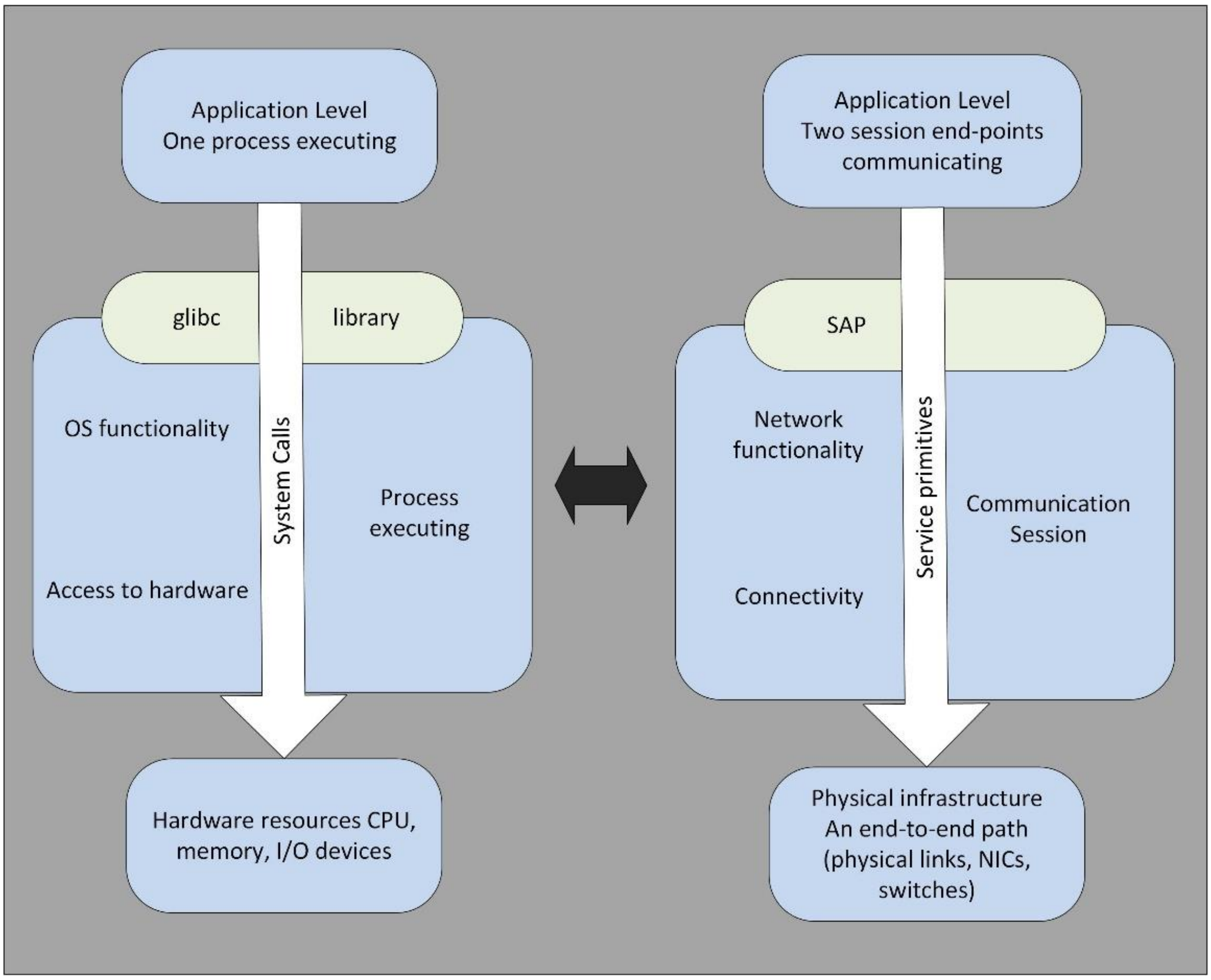


Figure 3.2 – Conceptual analogy between an computer and a network (extracted from [KZJK12]).

Based on this analogy, the multi-tasking or multi-threading concept on a computer can be seen as equivalent to network sharing. Multi-tasking is a method where multiple tasks, also known as processes, are performed during the same period. The tasks share common processing resources, such as a CPU, memory, and storage. By using multi-tasking, the unused resources of a machine can be assigned to another process to achieve a faster overall expectation. From the user's viewpoint, multi-tasking provides the ability to listen to music, copy files, and surf the web, all at the same time. Despite sharing the computer sources (e.g., processing, memory, and storage) advantages, there are some drawbacks as well. There is risk of interference among threads sharing hardware resources, such as caches or Translation Look-aside Buffers (TLBs). In addition, hardware support for multi-tasking is visible to software, thus more changes on applications and operating system are needed. Likewise, lack of infrastructure abstraction in a shared RAN makes modifications on MTs or RAN elements needed to support sharing the infrastructure among multiple users. On the ground that network sharing does not offer isolation, interference in between operators is unavoidable.

However, a VM is known as a software implementation of a machine that executes programs like a physical machine. Although VMs may run over the same physical hardware, they can have different operating systems. Since the virtualisation is not transparent to the applications and OSs, they need no changes to work on a VM. Even if it is possible to use an OS supporting multi-tasking on a VM, this leads to achieving multi-tasking and virtualisation at the same time. The same concept of virtualisation can be proposed for RAN, where multiple instances of a virtualised RAN can co-exist in the same physical infrastructure, each of them having different settings and requirements.

Despite the benefits and advantages of RAN virtualisation, its design and realisation are non-trivial and challenging tasks, which can generally be enumerated as slice isolation and slice resource management [LYLX12]. Multiple slices with different requirements, co-existing in the same physical infrastructure, have to be isolated from each other. This means that any change in one slice due to flow loads change, variation of channel condition, and so on, should not lead to the reduction of allocated resources to any other virtual instances. Isolation guarantees the right of all operators to run various applications in virtual sub-networks (slices). The realisation of isolation has many levels: it can be done at the lower level of the hardware, or the higher ones of time-slots (flow level), traffic (application level), etc. Virtualisation and isolation at lower levels offers more efficient resource usage at the cost of computational and operational complexity.

Concerning slice resource management, each slice has its own set of objectives and constraints, and they all have to be meet simultaneously, therefore, the resource manager (hypervisor) has to run different sub-algorithms for each slice, considering special constraints and objectives. A general architecture for virtual networks is presented in Figure 3.3. There is a virtual network provider aggregating the physical resources from different infrastructure providers to serve the VNOs. Hypervisor functionalities, such as virtual resources management, are the main part of the virtual network provider.

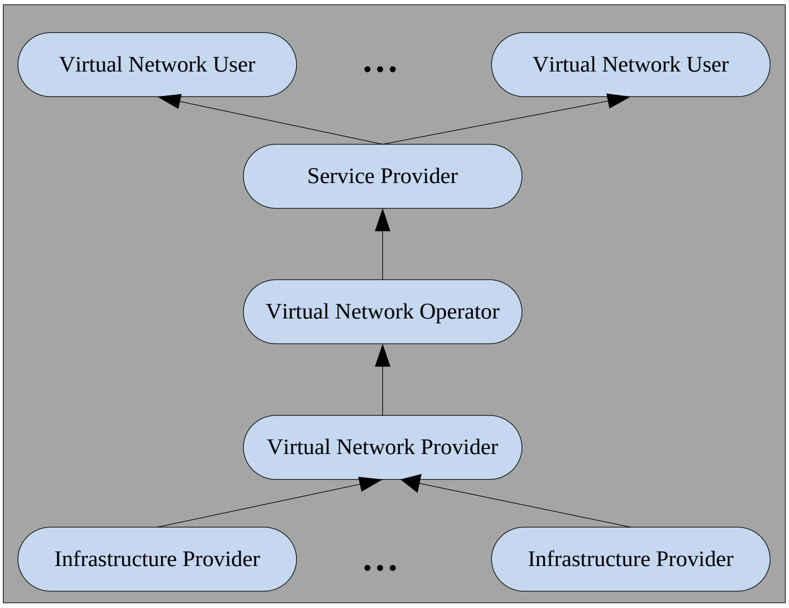


Figure 3.3 – Network virtualisation general architecture (extracted from [KZJK12]).

3.2.3 RAN Virtualisation Related Work

The concept of virtualised eNodeB is proposed in [ZLGT10], by introducing an entity called "Hypervisor" on the top of physical resources, which allocates these resources among various virtual instances. Using the concept of RAN sharing, the air interface resources, or the LTE spectrum is dynamically divided among various virtual eNodeBs. Figure 3.4 illustrates the LTE eNodeB virtualisation architecture. In [ZZGT11], eNodeB virtualisation progressed by considering resource allocation as static or dynamic spectrum sharing among virtual operators.

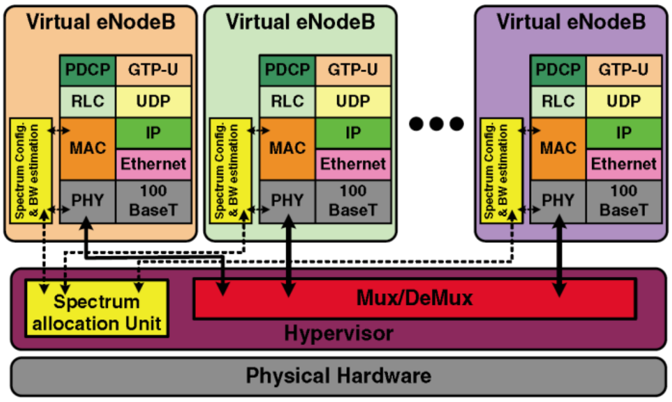


Figure 3.4 – Virtualised LTE eNodeB protocol stack (extracted from [ZZGT11]).

In [LMZT11], the advantage of virtualised LTE is investigated by an analytical model for FTP transmissions. Evaluations are done by considering realistic situations to present the multiplexing gain in addition to an analytical analysis.

Virtualisation of radio resources is a complex challenge. The radio resources at the bases station have to be shared and assigned to different virtual instances. This sharing needs to be fair, which can be defined differently, such as fairness in spectrum usage, in power distribution, or even in the QoS delivered to end-users. In addition, there are different criteria, like scheduling being based on bandwidth, data rate, power, interference, SLAs, channel conditions, traffic load, or a combination of them.

Based on the contract between virtual operator and infrastructure provider, VNOs can be divided mainly into four groups, as follows [ZLGT10]:

- Fixed guarantees: the operator asks for a fixed bandwidth at all time, hence, the bandwidth, regardless of being used or not, is always allocated to the operator.
- Dynamic guarantees: a guaranteed maximum bandwidth is offered to the operator upon request; contrary to the fixed guaranteed contract, the guaranteed bandwidth is not allocated at all time, and the operator may pay based on the used bandwidth.
- Best effort with minimum guarantees: the operator has a minimum guaranteed bandwidth, and a maximum best effort bandwidth.
- Best effort with no guarantees: there is no guaranteed bandwidth, and the operator receives only a portion of the available bandwidth, based on the existing load in a pure best effort manner.

In [BDPW11], the "Telco Cloud" model approach for hardware consolidation and management is put forward. In this LTE-based scenario, operators manage pools of hardware resources in key locations of their networks. These pools have been realised as small-managed clusters of networked hardware at both the edge and core of the telephone systems, Figure 3.5. The key advantages of the Telco Cloud over the existing Internet-based Cloud Computing lies on network QoS features, which normally are not available in the latter. Telco Cloud also can offer the lowest latency to computing resources, while exposing an unmatched level of control to third parties interested in high-performance hosting for their virtualised service infrastructures.

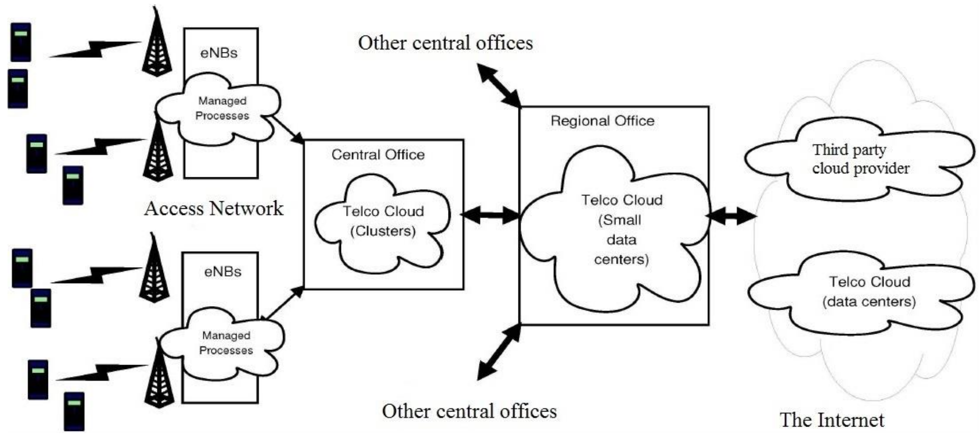


Figure 3.5 – Placement of resources in a Telco cloud (extracted from [BDPW11]).

The mobile cloud concept, [ZGWK11], developed as the combination of fibre-to-BSs and virtualisation via software radio, Figure 3.6. The author proposed to use fibre, metro-Ethernet capacity or a combination of these two, to connect BSs to the cloud. On the ground of these assumptions, the deployment physically unbundles BSs into Virtual Base Stations (VBSs) on IT platforms, pooling them at local "cloud" sites, and connecting them to shared RRHs attached to the BSs towers. The first working VBS is introduced in this paper.

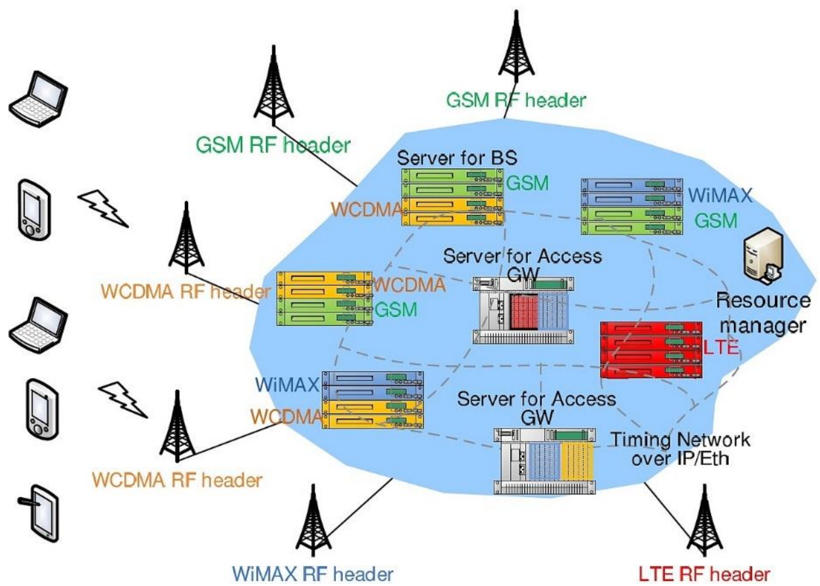


Figure 3.6 – Wireless Cloud Network (extracted from [ZGWK11]).

Wireless Network Cloud key benefits are counted as supporting multiple wireless standards over a low cost platform, enabling Mobile Virtual Network Operators (MVNOs) with network-sharing models, and integration of the service plane along with the wireless data- and control planes for load balancing in different districts in [LSZW10].

It is worth noting that the performance of cloud-deployed application has limits with increase of user demand because of three reasons. First, the centralised applications cannot scale beyond the hardware resources of the machine that runs the service. Secondly, stateless distributed applications can usually be scaled linearly by increasing the allocation of hardware and networking resources, and their scalability limits are usually determined by the presence of non-parallelisable sections. Last, but not at the least, distributed applications that state among instances are subject to additional trade-offs between consistency, availability, and partition tolerance. Applications in cloud systems based on their requirements can be categorised as:

- database storage with ACID (Atomicity, Consistency, Isolation, Durability) transaction processing;
- web service, dynamic generation of user content;
- distribution of stored data, live media streaming;
- interactive data processing and high speed messaging.

VANU, [VANU13], is one of the interesting commercial products of software radio, which is a wireless infrastructure solution that enables individual BSs to simultaneously operate as GSM, UMTS, and beyond. It is also claimed that virtualised base stations and RAN are also supported by their product.

In [ZGWK11], two deployment scenarios for VBSs over physical resources has been considered, as shown in Figure 3.7. First, the physical General Purpose Processor (GPP) node can be flexibly shared by multiple VBSs instances. Consequently, any VBS instance will be deployed on the same GPP node. Based on the second scenario, the processing resources from two or more GPP nodes can be abstracted and combined together to support a single BS instance. For example, the physical layer of a VBS (VBS-PHY) and the MAC (Medium Access Control) layer of a VBS (VBS-MAC) can be run on two different GPP nodes. In a hybrid resource pool, such mapping will provide better processing efficiency.

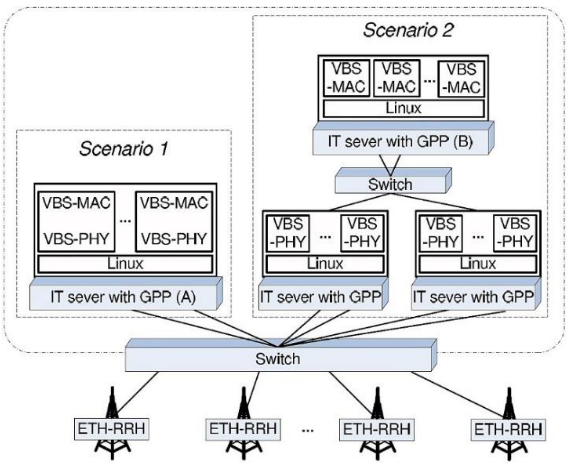


Figure 3.7 – Structure of VBS Pool (extracted from [ZGWK11]).

In[GoMG12], the SDR (Software Defined Radio) cloud is defined as a data centre that processes the signals transmitted and received by a set of distributed antennas. The SDR cloud uses the cloud computing technologies to share computing-resources. It is also suggested to divide the data centre into clusters of a few processors in each one. A high-level resource manager assigns users to clusters, or cluster groups as a function of the radio and computing conditions, and defines the resource-management strategy. Distributed low-level resource managers then allocate and de-allocate cluster-computing resources in real time.

In [KMSS10], the introduction of mobile cloud is found as a platform for the provisioning of context- and location-awareness enabling personalisation. The requirements on the wireless connectivity for Mobile Cloud Computing (MCC) differ in a number of important details from these classical heterogeneous access scenarios:

- MCC requires an “always-on” connectivity for a low data rate cloud control signalling channel;
- MCC requires an “on-demand” available wireless connectivity with a scalable link bandwidth;
- MCC requires a network selection and use that takes energy-efficiency and costs into account.

The solution presented in [KMSS10] assumes that one deploys MCC in a heterogeneous access scenario with a wide range of different available RATs, e.g., GSM, UMTS, LTE, WiMAX, and WLAN, and that context information, such as MTs locations and capabilities and user profiles, can be used by the mobile cloud controller to locally optimise access management, Figure 3.8.

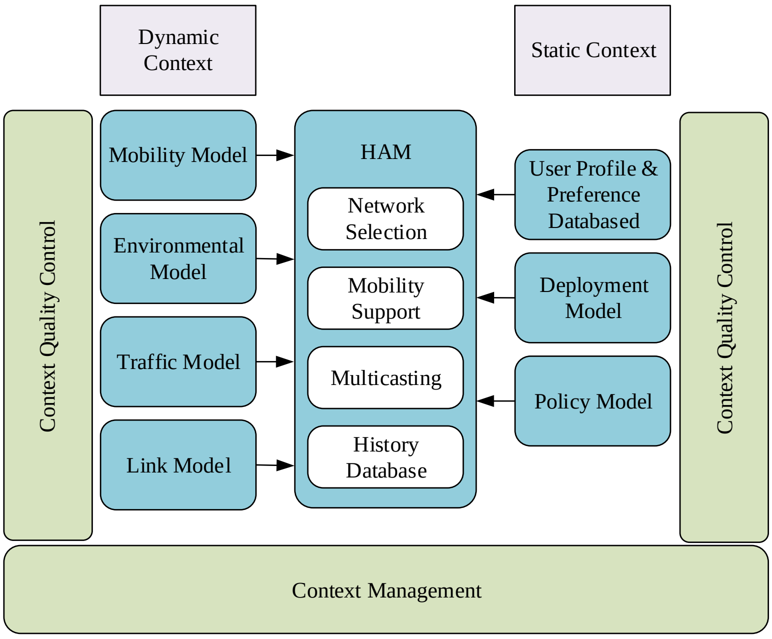


Figure 3.8 – Concept for Intelligent Radio Network Access, extracted from [KMSS10].

Moreover, add-on functionalities, such as charging, can be integrated using IMS. It is also assumed that the low data rate signalling cloud control channel is available, which might be provided by, e.g., GSM/GPRS, which is deployed in many countries with a very large coverage and already used for background low data rate machine-to-machine communications. Energy- and cost-efficiency can also be described in terms of the required radio resources. The “on-demand” request for scalable wireless connectivity is described by a statistical model derived from classical TCP/IP models for mobile users, where parameters were modified in order to model the expected mobile cloud traffic. This model is intended to describe the consumer and not the business use of MCCs, since it is expected that the majority of mobile cloud users belong to this group.

3.3 New Architectures for RAN

Interference scenarios become more complex in dense networks, due to multi-cell interference [RBDG14]. Central processing allows the coordination of radio resources across multiple cells, the implementation of efficient algorithms for the management of radio resources, and the optimisation of radio access performance at the signal level. In addition, it is required to orchestrate and optimise ultra-dense networks at the network level, by configuring the network for data traffic delivery. One of the novel architectures capable of solving today’s mobile networks challenges is Cloud RAN (C-RAN). This concept was first presented in [LSZW10] and the details are described in [CMRI11]. The architecture of C-RAN based on [CCYS14], Figure 3.9, can be summarised as follows:

- Backhaul transport network: a low latency optical transport network, which connects operators’ cores to the physical infrastructures of RANs.
- Base Band Unit pools data centres: a farm of VMs used for baseband processing of traffic among user terminal and network cores.
- Fronthaul transport network: transmit digitised radio signal between BBU-pools and RRHs using Common Public Radio Interface (CPRI) with high data rates.
- Radio Remote Heads: the transceivers in charge of exchanging data and control traffic toward/from MTs through the air interface, capable of supporting multiple RATs.

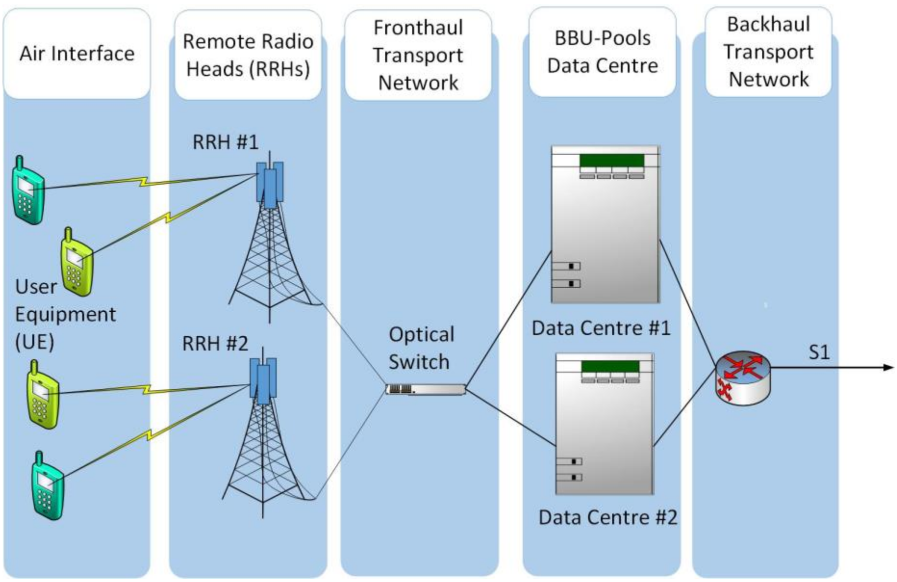


Figure 3.9 – C-RAN architecture.

By comparing the proposed architecture with the LTE RAN one, it can be seen that the functionalities of an eNodeB are now divided among RRH, fronthaul transport network, and BBU-pools.

Radio Access Network as a Service (RANaaS) addresses the trade-off between full centralisation, as in C-RAN, and decentralisation, as in today's network [RBDG14]. RANaaS, depending on actual needs as well as network characteristics, partially centralises the functionalities of RAN. In other words, considering that any kind of function may be delivered in the form of a service possibly centralised inside cloud, RANaaS can be stated as an application of XaaS paradigm. Figure 3.10 shows the architecture evolution towards a 5G network, presenting the usage of RANaaS to provide virtual eNodeBs. The mapping of physical links onto logical ones helps to understand the way virtual eNodeBs are managed.

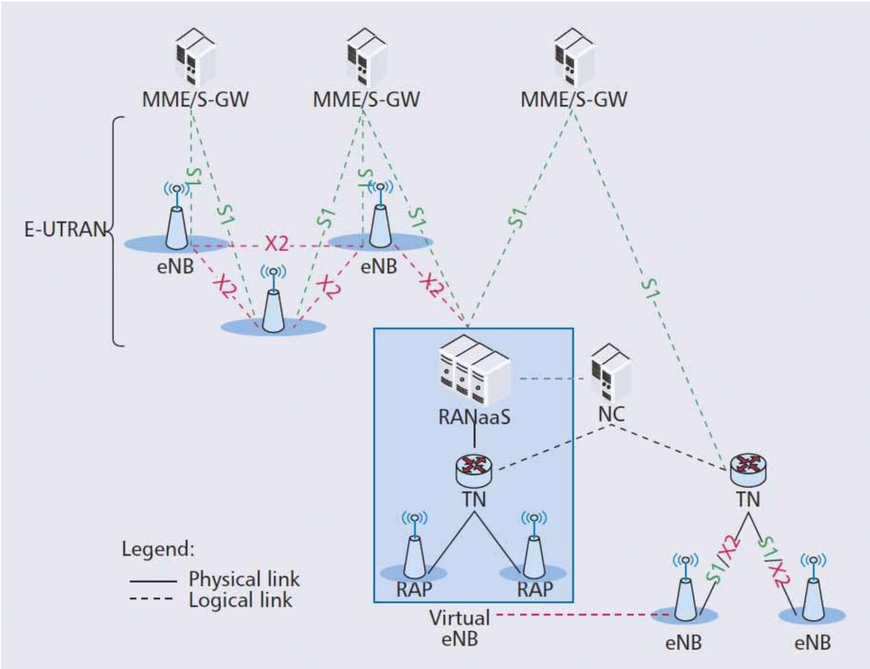


Figure 3.10 – Architecture evolution toward a 5G mobile network (extracted from [RBDG14]).

While the key aim of C-RAN is to have software-based and central RAN, NFV [ETSI13] aims to transform network architecture to consolidate many network equipment types into industry standard high volume servers, switches and storage. It involves the implementation of network functions in software that can run on a range of industry standard server hardware, and that can be moved to, or instantiated in, various locations in the network as required, without the need for the installation of new equipment. Figure 3.11 presents the proposed NFV architecture, the main constituents being:

- The NFVI (Network Functions Virtualisation Infrastructure), which provides the virtual resources required to support the execution of the Virtualised Network Functions (VNF).
- The VNF is the software implementation of a network function, which is capable of running over the NFVI. It can be accompanied by an Element Management System (EMS), as long as it is applicable to the particular function, which understands and manages an individual VNF and its peculiarities. The VNF is the entity corresponding to today's network nodes, which are expected to be delivered as pure software, free from hardware dependency.

- The NFV M&O (Management and Orchestration), which covers the orchestration and lifecycle management of physical and/or software resources that support the infrastructure virtualisation, and the lifecycle management of VNFs. NFV Management and Orchestration focuses on the virtualisation-specific management tasks necessary in the NFV framework. The NFV M&O also interacts with the (NFV external) OSS/BSS landscape, which allows NFV to be integrated into an already existing network-wide management landscape.
- The entire NFV system is driven by a set of metadata describing Service, VNFs and Infrastructure requirements, so that the NFV Management & Orchestration systems can act accordingly. These descriptions along with the Services, VNFs and Infrastructure can be provided by different industry players.

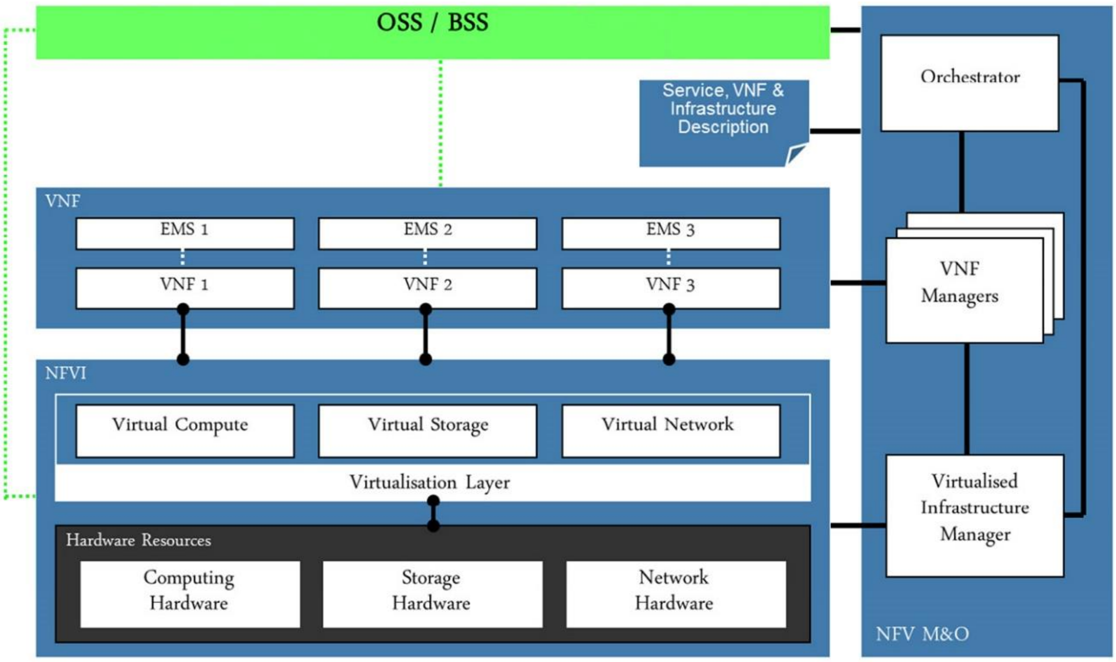


Figure 3.11 – NFV Architecture (extracted from [ETSI13]).

Furthermore, the reference architecture for RANaaS, presented in [MePe13], is illustrated in Figure 3.12. In the Mobile Cloud Network project [MCN15], RANaaS is a service offered to MCNSPs (Mobile Cloud Network Service Providers) through the interface of the RAN Provider's (RANP) Service Manager (SM), which is made up of the combination of the Service Orchestrator (SO) and the SM. The SO fully automatically manages the RANaaS service instance (e.g., making scaling decisions). The proposed architecture, as the C-RAN one, splits the base band functions of an eNodeB into a set of functional elements to make-up the per-RAT BBU. This architecture is using the combination of the proposed NFV architecture over the C-RAN infrastructure.

A BBU is composed by the scheduler, the Physical Cell, the Common Processing, the User Processing, the per-RAT control functions, and the RAN gateway. There is a one to one mapping between the RRH and the physical cell, and the interworking between cells is made by the collaboration between Control Functions, e.g., the X2 interface for LTE. These RANaaS functional elements are specific to the RAT.

They are instantiated by the SO through the Cloud Controller (CC) and then controlled, e.g., for providing new configuration files. This interface may also be used to enable a dialog between the RAN service instance and its service management, e.g., to enable smooth migration of running eNodeB between VMs by requesting running eNodeBs to not execute handover procedures during such critical phases.

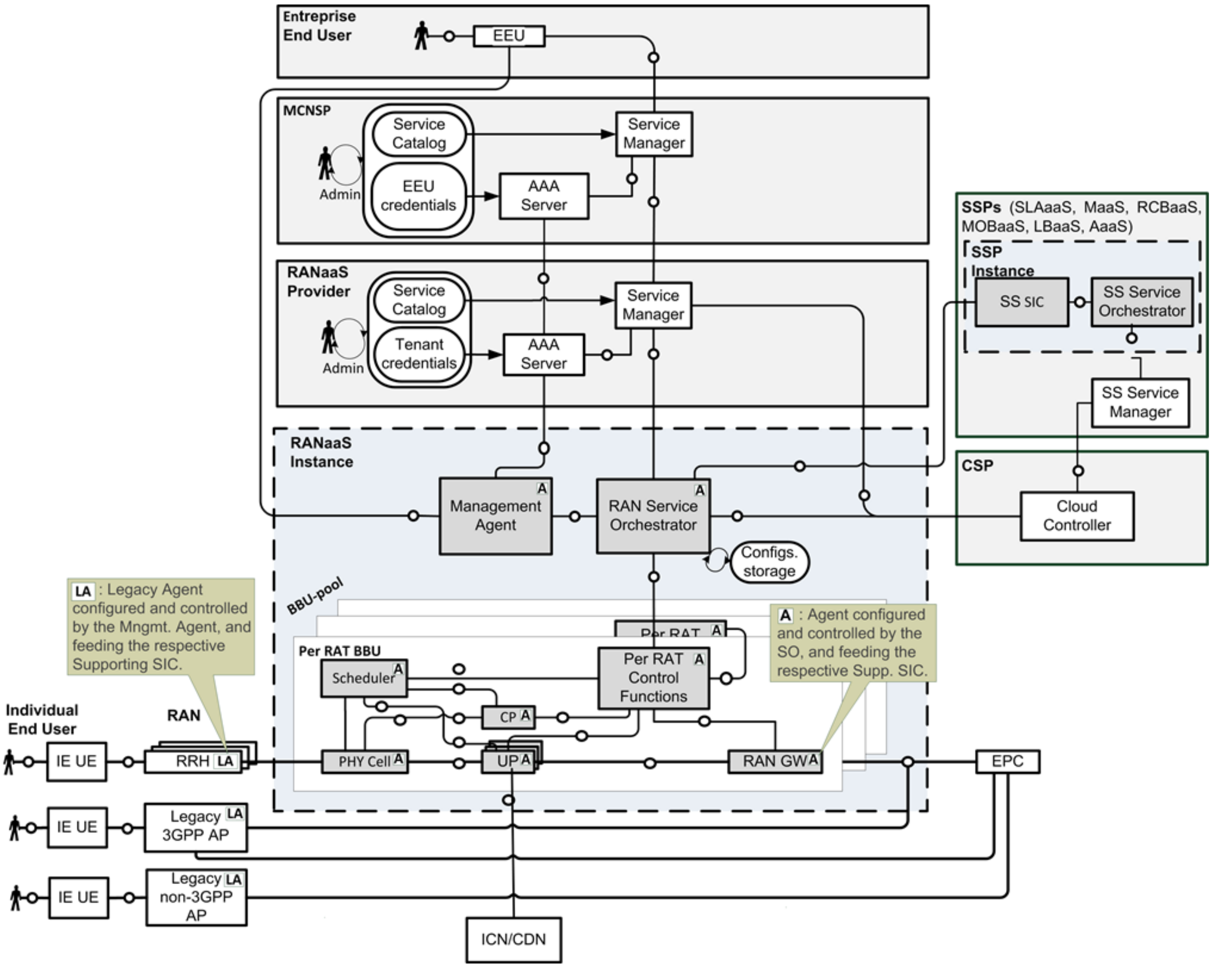


Figure 3.12 – RANaaS reference model.

3.4 Radio Resource Management

RRM is crucial for network operators to meet their subscribers QoS requirements. The role of RRM can be stated briefly as the adaptation of strategies and algorithms to ensure the efficient usage of the radio resources while serving subscribers according to their configured QoS parameters. RRM, however, concerns multi-user and multi-cell network capacity issues, rather than point-to-point channel capacity. Efficient dynamic RRM schemes may increase system capacity by an order of magnitude, which often is considerably more than what is possible with advanced channel and source coding schemes.

RRM techniques can generally be divided into static and dynamic approaches: static RRM contains manual or computer aided cell or network planning, such as deployment of BSs, antennas' setting, and channel frequency plans, while dynamic RRM, on the other hand, adjusts networks parameters adaptively regarding traffic load, users' position, QoS requirements, among others, to ensure that radio resources are efficiently used. It is more common to refer to static solutions as network planning, and use the term RRM for dynamic approaches, henceforth, the term RRM is used only referring to the dynamic perspective. Since RRM control various network parameters in different network layers, they are classified as different RRM functions working together, [Pere05]:

- admission control,
- packet scheduling and link adaptation,
- power control,
- handover,
- code management,
- RAT selection.

Admission control is the part of RRM that decides to admit or reject a new connection [Pere05]. The admission of new connections should not affect the guaranteed QoS of the active sessions, therefore, the admission control mechanism investigates the resource situation, the QoS requirements for the new bearer, as well as the priority levels and the currently provided QoS to the active sessions [HoTo10]. New connections are admitted up to the point where the packet scheduler in the cell can converge to a feasible solution. Admission control may also trigger handover procedures, transport channel type switching, etc.; the procedure may also be different, based on the function that has triggered it. Admitting a handover request, for instance, may need to have a higher priority comparing to set-up request. It is worth noting that the request of an MT for a new service connection may be either granted or denied, this denial of service is referred to as call blocking for voice calls in circuit-switch. The termination of a call based on a handover failure is referred to as call dropping, generally having more negative impact from a users' perspective than blocking a newly requested call [GhBo06]. In addition, admission control has to take user and/or service priorities into account. In the case that there are no more resources available to admit a new call or connection, admission control algorithms may decide to put more weight on some users or services, e.g., a user wanting to make voice call or who pays more money.

Packet scheduling and link adaptation have the goal to maximise cell capacity, making sure that the minimum QoS requirement for EPS bearers are fulfilled, and that there are adequate resources for best-effort bearers. Besides, link adaption provides information to the packet scheduler on the supported modulation and coding scheme for a user depending on the selected set of PBR. Link adaptation primarily bases its decisions on CQI feedback from users in the cell. Various techniques can be used for resource scheduling, where each of them has its own characteristics, advantages and drawbacks. Common scheduling algorithms can be counted as Maximal-Rate Scheduling, Round-Robin Scheduling, and Proportional Fairness Scheduling. In what follows, a brief description for these algorithms is presented.

Maximal-Rate Scheduling is used to maximise the total information theoretic capacity. The optimal strategy is to schedule at each time instant the user with the best channel. On the ground of the fact that in a system with many users whose channel varies very independently, it is likely to be a user whose channel is near its peak state at each time instant, raising diversity gain. The overall system throughput can be maximised by allocating the resource to the user who can best exploit it. Maximal-rate scheduling selects the user that has the largest value of the maximum available transmission rate.

Round-Robin Scheduling is one of the simplest scheduling schemes. Great fairness in radio resources assignment among users can be achieved by means of this algorithm, with the cost of lowering the whole system throughput. In this scheme, the first user is served with all the available resources (e.g., the whole frequency spectrum for a specific period), then the next user is served and the previous one is placed at the end of the queue, waiting to be served in the next round.

Proportional Fairness (PF) Scheduling is one of the most popular fairness criterion, which was first introduced in wired networks to share fixed bandwidth fairly. The key idea is to serve the user with a higher priority first, thus, the priority of each user should be calculated and the resources assigned to the user with maximum priority.

Power control is another function of RRM. A higher transmission power provides a better Signal-to-Interference-plus-Noise Ratio (SINR) at the receiver, consequently increasing the performance. On the other hand, a higher transmission power imposes more interference on the other cells or users, and it also increases the transmitter energy consumption. Taking the dynamic channel properties into account, including channel attenuation, noise and interference level at the receiver side, in addition to aforementioned trade-off, makes power control a crucial mechanism of radio networks [DaPS11].

Handover strategies in RRM maximise the number of active sessions over a set of cells [Pere05]. Although handover (hard handover) is known as an inherent procedure in cellular networks, with the introduction of CDMA new types of handover were provided, since MTs can be connected to more than one BS simultaneously (soft handover, softer handover being referred to the situation in which the MT is connected to more than one sector of the same BS). Regardless of the handover type, the handover mechanism has three main steps, i.e., measurements, decision and execution. The network controls the handover mechanism with the assistance of measurements reported by MTs. It is worth noting that the measurement reporting can be done periodically or it can be event trigger; in case there is no change in the network, the former option wastes network resources.

In addition, congestion control, also known as load control, is a mechanism to deal with situations in which QoS guarantees are at risk [Pere05]. Radio networks are taking advantage of the strict admission control, nevertheless there is always some probability of overload situations due to the evolution of system dynamics, such as mobility aspects, increase in interference, and traffic variability. Therefore, having congestion control mechanisms is essential, which can be supported by various functions, like:

- admission control, by refusing a new connection while the network is congested,
- handover, by transferring connections from a congested cell to neighbouring ones,
- packet scheduling, by reducing the transmission rates of a number of non-real time users.

Code management is devoted to managing the downlink Orthogonal Variable Spreading Factor (OVSF) code tree in CDMA. The efficient allocation and reallocation of codes is crucial in order to avoid code blocking, i.e., a situation in which a new connection is blocked although it could have been accepted based on interference analyses and remaining network capacity, the spare capacity not being available due to the inefficient usage of codes. In order to achieve an efficient code allocation, the code management function may have to force an MT using a given code to use another in the same layer, which is called code handover.

Last, but not the least, RAT selection is needed in order to realise RRM over heterogeneous access networks [Pere05]. RAT selection itself includes initial RAT selection and vertical handover (VHO). It is expected that the RAT selection algorithm, in case two or more RATs are co-located in the same coverage area, maximise system performance and QoS by allocating users to the most suitable RAT. RAT selection techniques can be classified as load balancing, service based, random selection, and priority based selection. In load balancing techniques, the RAT with the least load will be selected to admit the new users. Service based techniques select the RAT based on service type and network properties. The selection in random selection technique is based on availability and selecting probability of a RAT. Finally, priority based algorithm consider pre-defined selection priority of RATs for allocating new users. It is worth noting that there are more complex algorithms, such as fittingness factor based algorithm, fuzzy-neural based algorithm for RAT selection [AlBA12, AyAT14, Elba11, FaMu14, OAMI15, RaGa10, ROKB09].

Chapter 4

Models and Algorithms Development

This chapter focuses on presenting the concept of virtualisation of radio resources in addition to a management model by a brief description for the chosen architecture for virtual RAN, in addition to different resource management strategies and their relations to each other in Section 4.1. Section 4.2 summarises the design choices, and Section 4.3 comprehensively describes the analytical model for managing virtual radio resources. Furthermore, Section 4.4 presents the partial VRRM, an approach for solving the modelled problem, and finally, Section 4.5 introduces the evaluation metrics.

4.1 Virtual Radio Resource Management

According to the approaches described in Chapter 3 for RAN virtualisation, a Virtual RAN (V-RAN) defines a network with entities deployed on VMs with virtualised elements and functionalities. By adopting the C-RAN architecture, a new architecture is proposed in Figure 4.1, with the aim of serving multiple VNOs over the same physical infrastructure. VNOs are network operators that do not own radio access infrastructure and need wireless connectivity to serve their subscribers. In this new paradigm, they just have to ask from RAN Providers (RANPs) for wireless connectivity in term of capacities, and they do not have to deal with physical resources to serve their subscribers. Upon the request of VNOs, they are served by an instance of V-RAN, which is also known as RAN-as-a-Service.

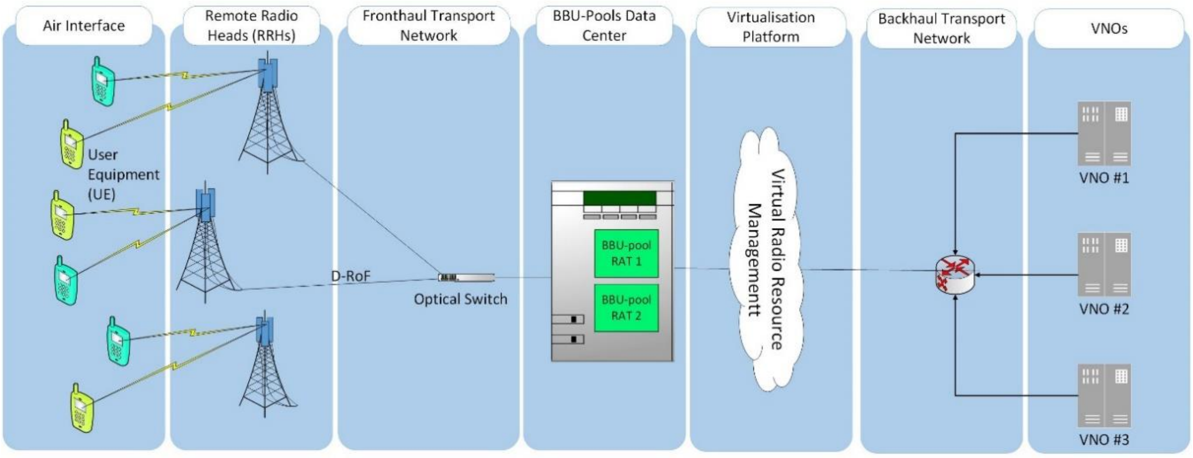


Figure 4.1 – V-RAN architecture.

By comparing the C-RAN architecture with the new one, the key difference is the virtualisation platform, placed between the BBU-pool data centres and backhaul transport network, which is in charge of abstracting the physical infrastructure for the VNOs, and handling their requests through the available physical resources. The most important functionality of the virtualisation platform is Virtual Radio Resource Management (VRRM), which is addressing the challenges of the co-existence of multiple instances of V-RAN in the same physical infrastructure from the air interface perspective. Optimising the radio resource usage, in addition to meeting the instances' objectives and any other additional ones (e.g., guaranteeing fairness among multiple instances), makes VRRM an elaborated procedure. The importance and complexity of VRRM is the reason to count it as the milestone in the realisation of end-to-end virtual radio networks.

The management of virtual radio resources, as shown in Figure 4.2, is a hierarchal management including VRRM and the common heterogeneous access networks RRM entities. VRRM is placed on the top, being in charge of translating VNOs' requirements and SLAs through sets of polices to lower levels. In other words, the key role of VRRM is to map virtual radio resources onto physical ones, in addition to issuing policies to manage these virtual resources. VRRM does not handles the physical radio resources, nevertheless, reports and monitoring information (e.g., estimated remained capacity) received from CRRM enable it to improve the policies.

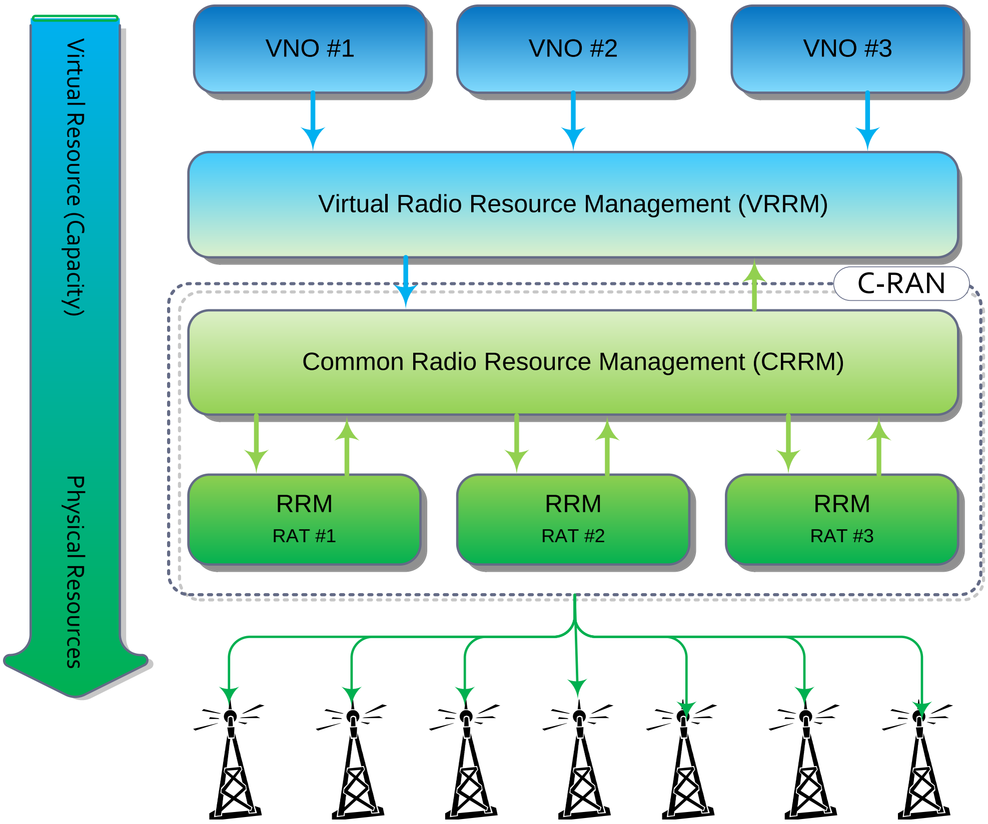


Figure 4.2 – Architecture of Virtual Radio Access Network.

In addition to VRRM, Common Radio Resource Management (CRRM), and local RRMs are placed in lower level. CRRM manages the resources of different RATs under its control to meet the requests of VRRM. It has to provide the requested capacity by demanding each RAT to deliver a portion of it. It also optimises its decision based on information from local RRMs.

Local RRMs, in the lower level, are liable for optimising radio resource usage in a single access technology, being in charge of assigning physical radio resource parameters, e.g., power, frequency bandwidth, and time-slots, to end-users upon receiving requests. The policies issued by CRRM to each local RRM are used as decision guidelines for that RRM; in addition to CRRM's policies, the resource allocation in each local RRM has to meet the QoS requirements of each service.

The proposed paradigm complies with ETSI proposal for the NFV architecture [ETSI13]. Based on the functionalities of the discussed entities, it is suggested to implement RRM management entities in NFV M&O block, and VRRM in the VNF managers block, as depicted in Figure 4.3. The same concept can be followed in the reference model of RANaaS proposed in the MCN project, Figure 3.12. In this model, the SO presents the same concept of NFV M&O. Conceptually, the best place for VRRM is inside the SO, as it is common to all instances and managing the require resources for them. The architecture design in this project considered RAN sharing as the primary solution for radio resources, therefore, granularity of SO in the MCN project [MePe13] was not designed for these elements, and SOs are created for each VNOs, where the RRMs entities are common to all.

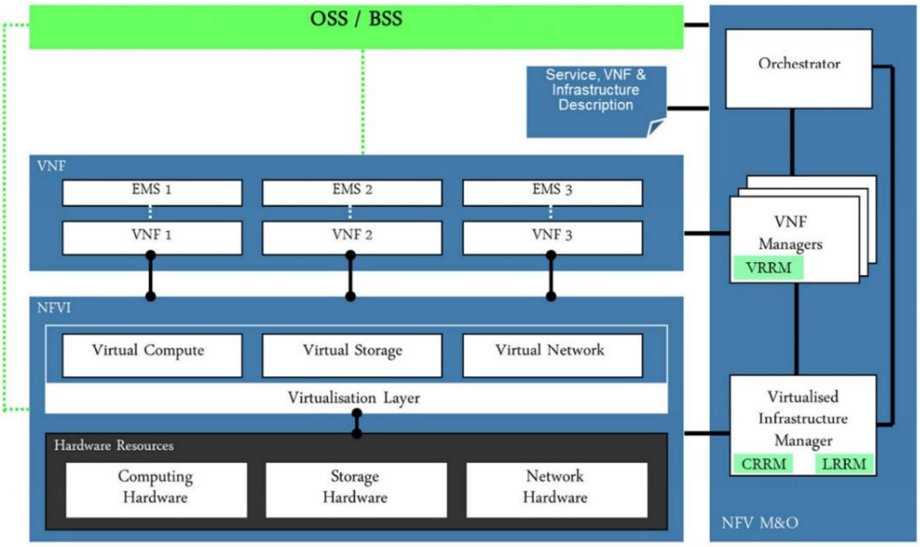


Figure 4.3 – The placement of RRM elements in NFV architecture.

To adopt the virtualisation of radio resources, some other changes needed to be done. Figure 4.4 presents the proposed corrections for the reference model for RANaaS. The new architecture considers another type of RANaaS service with support multi-tenancy. First, it was decided to have a new type of SO, which can serve multiple VNOs. The key difference, proposed here, is the "Common Control Functions" block that is going to have VRRM and CRRM.

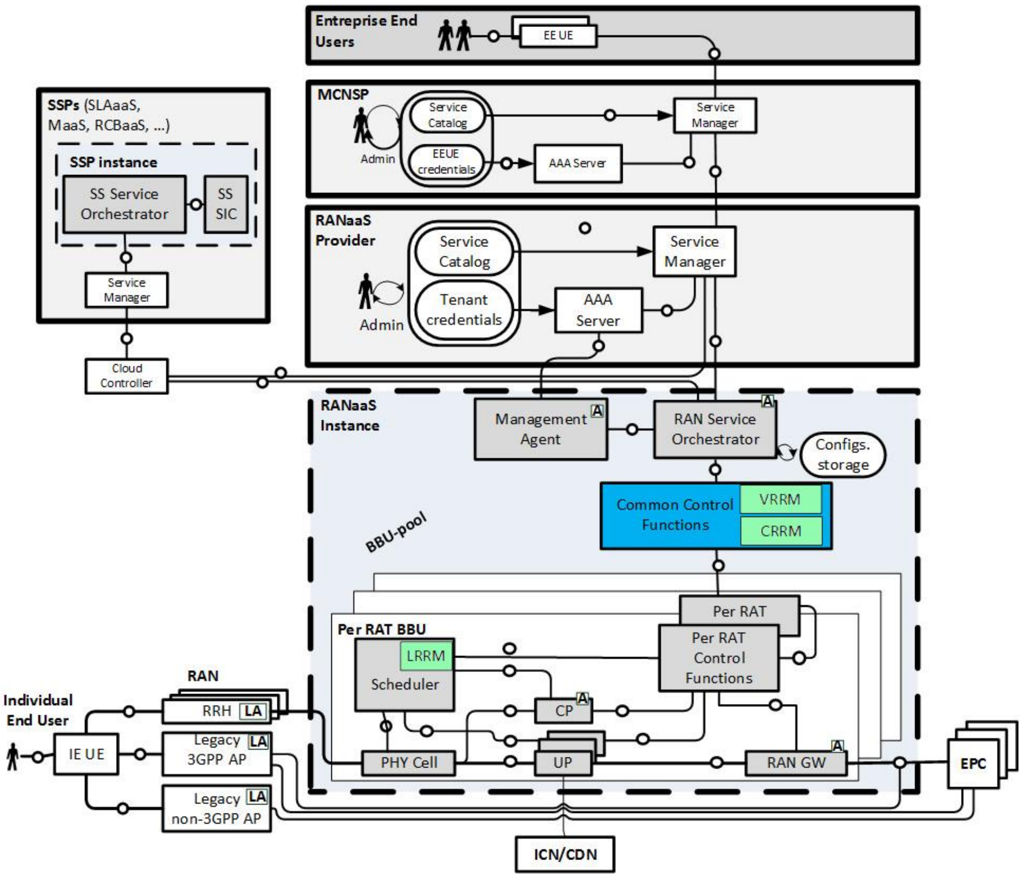


Figure 4.4 – Proposed correction for RANaaS model.

4.2 Design Choices

4.2.1 Service and User Requirements

The QoS offered to end-users and its requirements for various services in addition to the service classification in different systems (e.g., UMTS and LTE) is addressed in details in Section 2.3. Meeting services QoS requirements imposes unavoidable sets of constraints in various levels of VRRM. On the ground of this fact, the classification of various services based on their traffic characterisation is crucial for the VRRM problem. Despite the variety of offered services, it is possible to categorise all services based on their QoS constraints into four general classes: conversational, streaming, interactive, and background, as presented in Table 2.6.

The QoS requirement of conversational services can be summarised as follows.

$$\begin{cases} R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]} = R^{Nom}_{b_{ji}\,[\mathrm{Mbps}]} \\ t_{Rsp\,_{ji[\mathrm{s}]}} \leq t^{Nom}_{Rsp_{ji}[\mathrm{s}]} \end{cases} \quad \text{Conversational services} \qquad (4.1)$$

where:

- $R^{Nom}_{b_{ji}}$: nominal data rate for service j of VNO i,
- $R^{Srv}_{b_{ji}}$: served data rate for service j of VNO i,
- $t_{Rsp\,_{ji}}$: response time for service j of VNO i,
- $t^{Nom}_{Rsp_{ji}}$: nominal response time for service j of VNO i.

Equation (4.1) can be revised for streaming real time services as follows:

$$\begin{cases} \sum_{i=0}^{n_{\Delta t}} R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]}(t_i)\Delta t_{[\mathrm{s}]} \geq R^{Nom}_{b_{ji}\,[\mathrm{Mbps}]} n_{\Delta t}\Delta t_{[\mathrm{s}]} \\ t_{Rsp\,_{ji[\mathrm{s}]}} \leq t^{Nom}_{Rsp_{ji}[\mathrm{s}]} \end{cases} \quad \text{Streaming services} \qquad (4.2)$$

where:

- $n_{\Delta t}$: number of time intervals considered for evaluation,
- $R^{Srv}_{b_{ji}}(t_i)$: data rate at time interval t_i,
- Δt: size of each time interval.

The address requirements and constraint for the interactive service class is given by:

$$\begin{cases} R^{Srv}_{b_{ji}[\mathrm{Mbps}]} \geq R^{Nom}_{b_{ji}[\mathrm{Mbps}]} \\ t_{srv\,_{ji[\mathrm{s}]}} \leq t^{Nom}_{Srv_{ji}\,[\mathrm{s}]} \\ t_{Rsp\,_{ji[\mathrm{s}]}} \leq t^{Nom}_{Rsp_{ji}[\mathrm{s}]} \end{cases} \quad \text{Interactive services} \qquad (4.3)$$

where:

- $t_{srv_{ji}}$: serving time for service j of the VNO i,
- $t_{Srv_{ji}}^{Nom}$: nominal serving time for service j of the VNO i.

Finally, the imposed constraint by services of the background class is:

$$t_{srv_{ji[\mathrm{s}]}} \leq t_{Srv_{ji}[\mathrm{s}]}^{Nom} \qquad \text{Background Traffic} \tag{4.4}$$

Besides QoS requirements and their imposed constraints, service arrival and serving (holding) rates are also important for schedulers and resource managers. The assumed model for various services are presented in detail in Appendix B.

4.2.2 Virtual Network Operators SLAs

In addition to QoS requirement of each service session, VNOs request for a level of QoS from VRRM. Capacity per service is the term based on which SLAs can be defined. The RRM algorithms have also to consider the priority of the different services of different VNOs based on their SLAs. For instance, conversation (e.g., VoIP) and streaming (e.g., video streaming) service classes are delay sensitive, but they have almost constant data rates, hence, the assignment of data rates higher than the contracted capacities does not increase user's QoE; in contrast, in interactive (e.g., FTP) and background (e.g., email) service classes, the increase of data rates can indeed improve QoE. Therefore, operators offering the former services are not interested in allocating higher data rates. Based on the service set and requirements, VNOs may have different SLAs, but these SLAs can generally be categorised into three types of contract:

- Guaranteed Bitrate (GB), in which the VNO is guaranteed a minimum as well as a maximum level of data rates, regardless of the network status. In other words, the total satisfaction of the VNO is achieved when the maximum guaranteed data rate is allocated to it. It is expected that subscribers always experience a good QoE in return of relatively more expensive services.
- Best effort with minimum Guaranteed (BG), where the VNO is guaranteed with a minimum level of service, but the request for higher data rates than the guaranteed one is served in a best effort manner. In this case, although VNOs do not invest as much as former ones, they can still guarantee the minimum QoS to their subscribers. From the subscribers' viewpoint, the acceptable service (not as good as the other ones) is offered with a relatively lower cost.
- Best Effort (BE), in which the VNO is served in a pure best effort manner. Operators, and consequently their subscribers, in return, may suffer from low QoS and resource starvation during busy hours.

Figure 4.5 comperes the acceptable data rate to be allocated according to any of the three aforementioned SLAs. The lower band for GB and BG SLAs are the minimum guaranteed data rates. Likewise, the upper band for BG and BE is the total network capacity while the maximum guaranteed data rate limits the data rate provided to the VNOs with guaranteed data rates.

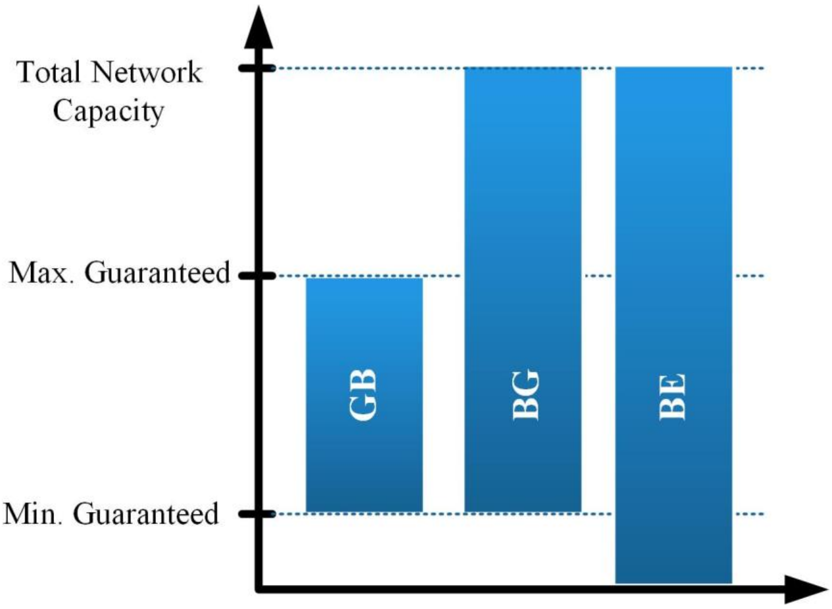


Figure 4.5 – Comparison of different VNOs' SLAs.

4.2.3 Canonical Scenario

In order to explain and analyse the proposed model, an urban hotspot environment is chosen as the conical scenario. The key parameters of the reference scenario are cell layout, services, and VNOs.

The scenario considers the full coverage of heterogeneous access cellular network with multiple RATs, which are OFDMA (based on LTE-Advance), CDMA (based on UMTS), and FDMA/TDMA (based on GSM). The cell layout of different RAT is as follows:

- The OFDMA cells with a radius of 0.4 km are the smallest ones. Based on the 100 MHz LTE-Advanced, each cell has 500 RRUs, which can be assigned to traffic bearers.
- The configurations of CDMA cells are chosen according to UMTS (HSPA+) working at 2.1 GHz. Each cell with the radius of 1.2 km has 3 carriers, each one with 16 codes. Only 45 codes out of all 48 codes in each cell can be assigned to users' traffic.
- The FDMA/TDMA cells are the biggest ones, with a radius of 1.6 km. Based on GSM900, each cell has 10 frequency channels and each channel has 8 time-slots. It is assumed that 75 time-slots out total 80 available time-slots in each cell can be used for users' traffic.

Furthermore, the canonical scenario contains two best effort VNOs. A VNO BG has the minimum guaranteed data rate of 0.4 Gbps while a VNO BE is served totally in the best effort manner. Each of these VNOs has only one service from the same service class. However, the priorities of the services in VNOs BG and BE are not the same, being assumed that the serving weight for the VNO BG is 0.6 while for VNO BE it is 0.4 (the concept of serving weight is described in detail in the following sections).

4.3 Analytical Model for VRRM

4.3.1 Guidelines for the model

RRM is known to be a non-trivial task, since radio interface performance is variable due to effect of fading, noise, and interference. In V-RAN, it becomes even more challenging, where multiple virtual networks (instances) with different requirements and SLAs co-exist in the same physical infrastructure. Hence, VRRM briefly can be considered as a decision-making problem under uncertainty for a dynamic environment: the decision is on the allocation of the resources to the different services of the VNOs by considering the set of available radio resources. In the following, the VRRM problem is discussed in more detail.

The first step is forming the state of the radio resources containing the number of available Radio Resource Unites (RRUs) per RAT (e.g. time-slot, radio block or code block), as follows:

$$s_t^{RRU} = \left\{ N_{SRRU}^{RAT_1}, N_{SRRU}^{RAT_2}, \dots, N_{SRRU}^{RAT_{N_{RAT}}} \right\} \tag{4.5}$$

where:

- s_t^{RRU}: the state of radio resources at t,
- $N_{SRRU}^{RAT_i}$: number of spare RRUs in i-th RAT,
- N_{RAT}: number of RATs.

Next, one needs to map the set of radio resources onto the total data rate of the network. Since radio resources' performance is not deterministic, it is not possible to have a precise prediction of capacity, but it is possible to estimate the total network data rate Probability Density Function (PDF) as a function of the available RRUs for further decisions.

$$s_t^{RRU} \xrightarrow{mapping} R_{b\,[\mathrm{Mbps}]}^{CRRM}(t) \tag{4.6}$$

where:

- R_b^{CRRM}: the total data rate from multiple access technologies.

The third step is the allocation of the available resources to the services of the VNOs. Sets of policies are designed in this step, in order to assign a portion of the total network capacity to each service of each VNO. Meeting the guaranteed service levels and increasing the resources' usage efficiency are the primary objectives, but other goals, such as fairness, may be considered. The resource allocation mapping of the total network capacity onto different services' data rates is given by:

$$R_{b\,[\mathrm{Mbps}]}^{CRRM}(t) \xrightarrow{mapping} \left\{ R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}(t) \middle| \, j = 1, \dots, N_{VNO}\,, i = 1, \dots, N_{srv} \right\} \tag{4.7}$$

where:

- $R_{b_{ji}}^{Srv}$: Serving (allocated) data rate for service j of the VNO i.

Finally, monitoring the used resources and updating the state of radio resources is the observation part of this decision-making problem. It enables the manager to evaluate the accuracy of its decisions and to modify the former decisions. Updating the changes in the set of radio resources, helps the manager to cope with dynamic changes of environment and VNOs requirements. In summary, it can be claimed that resource management solutions generally have two main parts, which are estimation of available resources and optimising their allocation, Figure 4.6. These two parts are addressed in detail in the following sections.

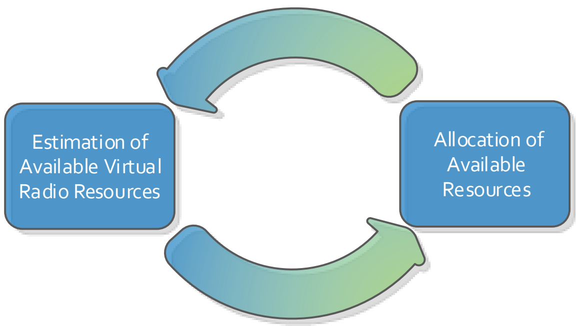


Figure 4.6 – Key concept of VRRM.

4.3.2 Estimation Available Virtual Radio Resource

The goal in the estimation of the available virtual radio resources to obtain a relationship in form of PDF from the set of available radio resources to the total network capacity. In general, the data rate of an RRU of an MT varies between a minimum and a maximum values, based on various parameters, such as RAT, modulation, coding schemes, and SINR. In a certain configuration (e.g., certain modulation), the data rate of a single RRU is a function of channel quality, SINR, varying as follows:

$$R_{b_{RAT_i[\mathrm{Mbps}]}}(\rho_{in}) \in [0, R_{b_{RAT_i[\mathrm{Mbps}]}}^{max}] \tag{4.8}$$

where:

- $R_{b_{RAT_i}}$: data rate of a single RRU of the i-th RAT,
- ρ_{in}: SINR,
- $R_{b_{RAT_i}}^{max}$: the maximum data rate of a single RRU of the i-th RAT.

The probability of having data rates higher than $R_{b_{min}}$ depends on the probability of having SINR higher than a given SINR corresponding to that data rate:

$$Prob\left(R_{b\,[\mathrm{Mbps}]} > R_{b_{min}\,[\mathrm{Mbps}]}\right) = Prob(\rho_{in_{[\mathrm{dB}]}} > \rho_{in_{[\mathrm{dB}]}}^{R_{b_{min}}}) \tag{4.9}$$

where:

- $\rho_{in}^{R_{b_{min}}}$: relative SINR to achieve data rate equal or higher than $R_{b_{min}}$.

In [DGBA12], considering an interference-limited heterogeneous cellular network, where each tier models the BSs of a particular class, it is assumed that the BSs in a given tier are spatially distributed as a Poisson Point Process (PPP) with a given density and transmission power. The received power is exponentially distributed (i.e., Rayleigh fading is assumed for the signal magnitude) and MTs are connected to the strongest BS (i.e., the SINR is larger than 1 dB), the probability of having SINR higher than an arbitrary level, ρ_{in}^0, being:

$$Prob(\rho_{in} \geq \rho_{in}^0) = \frac{\alpha_p}{2\pi \, csc\left(\frac{2\pi}{\alpha_p}\right)(\rho_{in}^0)^{2/\alpha_p}} \tag{4.10}$$

where:

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

Although (4.10) is proposed based on MTs connected to the strongest BS, the probability of SINRs lower than 1 is not zero, hence, (4.10) was modified as follows:

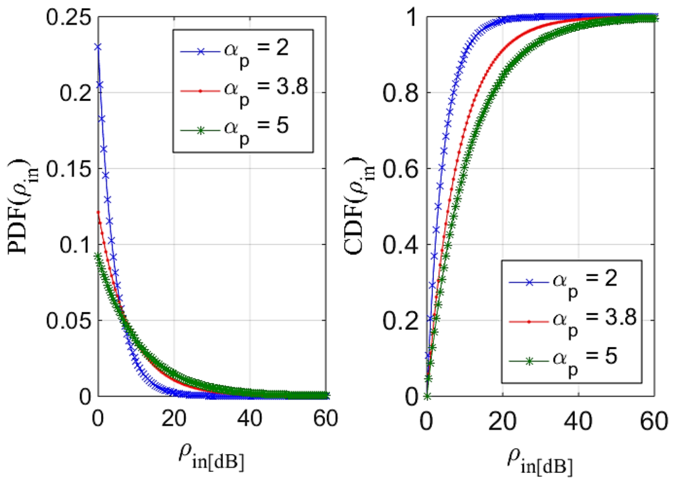
$$Prob(\rho_{in} \geq \rho_{in}^0 | \rho_{in} \geq 1) = \frac{Prob(\rho_{in} \geq \rho_{in}^0, \rho_{in} \geq 1)}{Prob(\rho_{in} \geq 1)} = \begin{cases} (\rho_{in}^0)^{-\frac{2}{\alpha_p}} & \rho_{in} \geq 1 \\ 0 & \rho_{in} < 1 \end{cases} \tag{4.11}$$

Converting the random variable ρ_{in} to an equivalent in logarithmic scale (for the use of dB) in (4.11), the CDF, $P_{\rho_{[\mathrm{dB}]}}$, and PDF, $p_{\rho_{[\mathrm{dB}]}}$, of SINR is obtained as follows:

$$P_{\rho_{[dB]}}(\rho_{in}) = 1 - e^{-\frac{0.2}{\alpha_p} ln(10) \rho_{in[\mathrm{dB}]}} \tag{4.12}$$

$$p_{\rho_{[dB]}}(\rho_{in}) = \frac{0.2}{\alpha_p} ln(10) e^{-\frac{0.2}{\alpha_p} ln(10) \rho_{in[\mathrm{dB}]}} \tag{4.13}$$

Equations (4.12) and (4.13) are plotted in Figure 4.7 for three models: free space, $\alpha_p = 2$, COST-Walfisch-Ikegami Model, $\alpha_p = 3.8$, and outdoor/indoor model, $\alpha_p = 5$, [DaCo99]. The higher value of path loss leads to a higher attenuation of interference, and consequently to a higher SINR.



(a) PDF (b) CDF

Figure 4.7 – PDF and CDF of SINR.

In [Jaci09], the data rate as a function of SINR, and vice versa, have been presented for various access technologies based on the real network logs. In next step, for the sake of simplicity, these functions have been approximated by the equivalent polynomial of degree 5, as described in Annex A. Therefore, the SINR as a function of data rate for the i-th RAT is given by:

$$\rho_{in_{[\mathrm{dB}]}}\left(R_{b_{RAT_i}}\right) = \sum_{k=0}^{5} a_{k\left[\frac{\mathrm{dB}}{\mathrm{Mbps^k}}\right]} R_{b_{RAT_i[\mathrm{Mbps}]}}^{\ k} \tag{4.14}$$

where

- a_k: coefficients of a polynomial approximation of SINR as a function of data in each RAT, from Table A.11.

Substituting (4.14) in (4.12), the CDF of the data rate for a single RRU can be written as:

$$P_{Rb}\left(R_{b_{RAT_{i[\mathrm{Mbps}]}}}\right) = 1 - e^{-\frac{0.2}{\alpha_p} ln(10) \sum_{i=0}^{5} a_k \left(R_{b_{RAT_i}}\right)^k} \tag{4.15}$$

On the ground of the data rates boundaries addressed in (4.8), the CDF of the data rate also has to consider these boundaries as follows:

$$\begin{aligned} P_{R_b}\left(R_{b_{RAT_{i[\mathrm{Mbps}]}}}\right) &= Prob\left(R_{b[\mathrm{Mbps}]} \leq R_{b_{RAT_i}} | 0 \leq R_{b_{RAT_i}} \leq R_{b_{RAT_i[\mathrm{Mbps}]}}^{max}\right) \\ &= \frac{e^{-\frac{0.46}{\alpha_p} a_0} - e^{-\frac{0.46}{\alpha_p} \sum_{k=0}^{5} a_k \left(R_{b_{RAT_i[\mathrm{Mbps}]}}\right)^k}}{e^{-\frac{0.46}{\alpha_p} a_0} - e^{-\frac{0.46}{\alpha_p} \sum_{k=0}^{5} a_k (R_{b_{RAT_i[\mathrm{Mbps}]}}^{max})^k}} \end{aligned} \tag{4.16}$$

Simply from (4.16), the PDF of R_b can be achieved as follows:

$$p_{R_b}\left(R_{b_{RAT_i}}\right) = \frac{\frac{0.46}{\alpha_p}\left(\sum_{k=1}^{5} k a_k \left(R_{b_{RAT_i[\mathrm{Mbps}]}}\right)^{k-1}\right) e^{-\frac{0.46}{\alpha_p} \sum_{k=0}^{5} a_k \left(R_{b_{RAT_i[\mathrm{Mbps}]}}\right)^k}}{e^{-\frac{0.46}{\alpha_p} a_0} - e^{-\frac{0.46}{\alpha_p} \sum_{k=0}^{5} a_k (R_{b_{RAT_i[\mathrm{Mbps}]}}^{max})^k}} \tag{4.17}$$

Equations (4.16) and (4.17) for an OFDMA (LTE) RRU are plotted in Figure 4.8, showing, in accordance with (4.6), that by increasing the path-loss exponent is more probable to obtain a higher throughput.

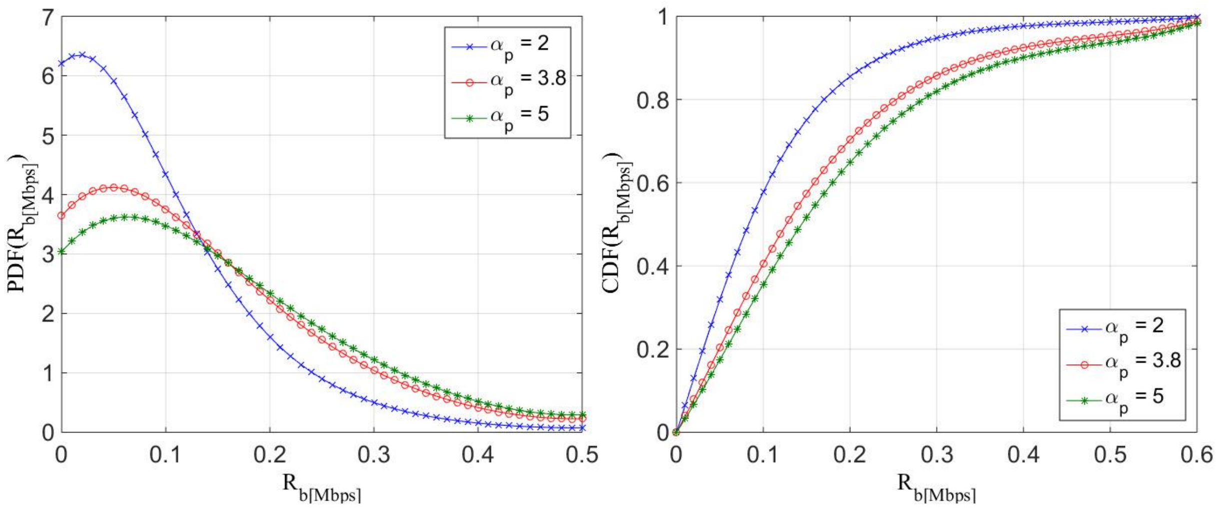
The next step is to extend the model to estimate the overall performance of a single RAT. The total data rate from a single RAT pool is:

$$R_{b_{tot}}^{RAT_i} = \sum_{n=1}^{N_{RRU}^{RAT_i}} R_{b_n}^{RAT_i} \tag{4.18}$$

where:

- $N_{RRU}^{RAT_i}$: number of RRUs in the i-th RAT,
- $R_{b_{tot}}^{RAT_i}$: data rate from the i-th RAT pool,

- $R_{b_n}^{RAT_i}$: data rate from the n-th RRU of the i-th RAT.



PDF(Rb). (b) CDF(Rb).

Figure 4.8 – PDF and CDF of an LTE RRU in downlink ($R_{b_{RAT}}^{max}$ =0.7 [Mbps]).

Based on the assumption that channels are independent, data rates random variables, $R_{b_n}^{RAT_i}$, are also independent, the PDF of the total data rate being equal to the convolution of all RRUs' PDFs, [PaPi02]:

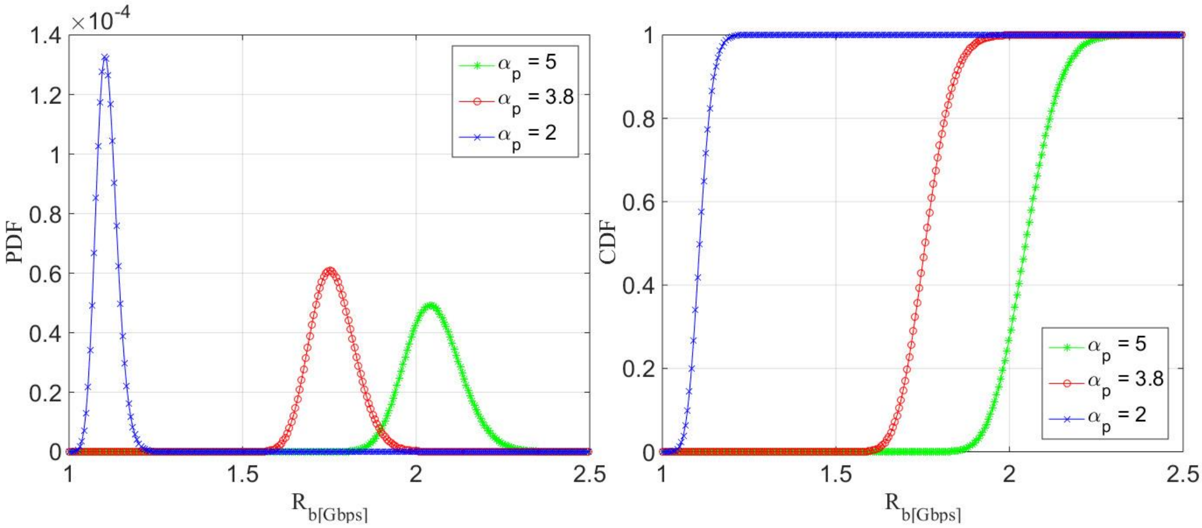
$$p_{R_b}\left(R_{b_{tot}}^{RAT_i}\right) = p_{R_{b1}}\left(R_{b_1}^{RAT_i}\right) * p_{R_{b2}}\left(R_{b_2}^{RAT_i}\right) * \dots * p_{R_{bn}}\left(R_{b_n}^{RAT_i}\right) \tag{4.19}$$

In heterogeneous access networks, the resource pools of RATs can be aggregated under the supervision of CRRM, the total data rate from all RATs being the sum of the individual ones:

$$R_{b\,[\mathrm{Mbps}]}^{CRRM} = \sum_{i=1}^{N_{RAT}} R_{b_{tot}[\mathrm{Mbps}]}^{RAT_i} \tag{4.20}$$

Using (4.19) and (4.20), the probability functions of the canonical scenario are obtained. Figure 4.9 plots these function for the three aforementioned path-loss exponents for the canonical scenario. As it is clearly shown, the median of the total network data rate varies from 1.1 Gbps for path-loss of free space up to 2 Gbps for path-loss exponent of 5.

As already discussed, the data rate of a single RRU, and consequently of the whole network, depends on the SINR. It means that serving the same data rate to an MT with low SINR requires assigning more RRUs comparing to one with a higher SINR, hence, the network throughput is higher when all users in the network have a higher SINR. On the other hand, the capacity of the network reduces when end-users with a low SINR have to be served. Furthermore, the lower the capacity of network is, the lower the offered QoS to the VNOs will be, and potentially also for their subscribers. On the very low capacity case, the VRRM cannot even meet the guaranteed requirements, so three main approaches are considered for allocation studies:



(a) PDF (b) CDF.

Figure 4.9 – Probability function of a practical HetNet.

a) Optimistic (OP) approach: all RRUs in the system are assigned to users with very good channel quality, being assumed that the data rate of each RRU satisfies:

$$0.5R_{b[\mathrm{Mbps}]}^{max} \leq R_{b[\mathrm{Mbps}]} \leq R_{b[\mathrm{Mbps}]}^{max} \tag{4.21}$$

b) Realistic (RL) Approach: each RATs RRUs is divided into two equal groups, being assumed that, depending on the SINR situation, one has:

$$0 \leq R_{b[\mathrm{Mbps}]} < 0.5R_{b[\mathrm{Mbps}]}^{max} \quad \text{1st group} \tag{4.22}$$

$$0.5R_{b[\mathrm{Mbps}]}^{max} \leq R_{b[\mathrm{Mbps}]} < R_{b[\mathrm{Mbps}]}^{max} \quad \text{2nd group} \tag{4.23}$$

c) Pessimistic (PE) Approach: all RRUs in the system are assigned to users with low SINR, so that the boundaries are:

$$0 \leq R_{b[\mathrm{Mbps}]} \leq 0.5R_{b[\mathrm{Mbps}]}^{max} \tag{4.24}$$

Equations (4.16) and (4.17) can be developed for these case studies, where the data rate is bounded between high and low values, the conditional CDF of a single RRU being calculated as [PaPi02]:

$$P_{Rb}\left(R_{b_{RAT_i}} | R_{bLow[\mathrm{Mbps}]} \leq R_{b_{RAT_i[\mathrm{Mbps}]}} \leq R_{bHigh[\mathrm{Mbps}]}\right) = \frac{e^{-\frac{0.46}{\alpha}\sum_{k=0}^{5} a_k R_{bLow\,[\mathrm{Mbps}]}^k} - e^{-\frac{0.46}{\alpha_p}\sum_{k=0}^{5} a_k \left(R_{b_{RAT_i[\mathrm{Mbps}]}}\right)^k}}{e^{-\frac{0.46}{\alpha}\sum_{k=0}^{5} a_k R_{bLow\,[\mathrm{Mbps}]}^k} - e^{-\frac{0.46}{\alpha}\sum_{k=0}^{5} a_k R_{bHigh\,[\mathrm{Mbps}]}^k}} \tag{4.25}$$

where:

- R_{bLow}: the low boundary for the RRU data rate;
- R_{bHigh}: the high boundary for the RRU data rate

$$0 \leq R_{bLow[\text{Mbps}]} \leq R_{bHigh[\text{Mbps}]} \leq R_{b[\text{Mbps}]}^{max} \tag{4.26}$$

The PDF of a RRU in this case, based on (4.25), is:

$$p_{Rb}\left(R_b \,|R_{bLow[\text{Mbps}]} \leq R_{b[\text{Mbps}]} \leq R_{b[\text{Mbps}]}^{max}\right) = \frac{\frac{0.46}{\alpha}\left(\sum_{\text{k}=1}^{5} k a_k R_{b\,[\text{Mbps}]}^{(k-1)}\right) e^{-\frac{0.46}{\alpha}\sum_{k=0}^{5} a_k R_{b\,[\text{Mbps}]}^{k}}}{e^{-\frac{0.46}{\alpha}\sum_{k=0}^{5} a_k R_{bLow\,[\text{Mbps}]}^{k}} - e^{-\frac{0.46}{\alpha}\sum_{k=0}^{5} a_k R_{bHigh\,[\text{Mbps}]}^{k}}} \tag{4.27}$$

The probability functions for the described approaches are plotted and analysed in Chapter 6.

4.3.3 Resource Allocation

After estimating the network throughput, VRRM can use the output in the allocation procedure, where each service of each VNO, based on their requirements and SLAs, receives a portion of the available resources.

In this framework, it is assumed that each VNO has a contracted data rate, based on the number of subscribers it serves:

$$R_{b[\text{Mbps}]}^{Con} = N^{EU}\overline{R_{b[\text{Mbps}]}^{EU}} \tag{4.28}$$

where:

- R_b^{Con}: the contracted data rate,
- N^{EU}: the number of end-users,
- $\overline{R_b^{EU}}$: the average data rate for end-user.

The required data rate for each service of the VNO is a portion of the contracted data rate, which is going to allocated according to the SLA and the availability of the resources.

The goal in the resource allocation procedure is to increase the total network throughput while considering the priority of different services of different VNOs and the other constraints. Hence, the objective function for VRRM is the total weighted network data rate, being written as:

$$f_{\mathbf{R_b}}^{v}(\mathbf{R_b}) = \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} W_{ji}^{Srv} R_{b_{ji}\,[\text{Mbps}]}^{Srv} \tag{4.29}$$

where:

- $f_{\boldsymbol{R_b}}^{v}$: the VRRM objective function,
- N_{VNO}: number of VNOs serving by this VRRM,
- N_{srv}: number of service for each VNO,
- W_{ji}^{Srv}: weight of serving unit of data rate for service *j* of the VNO *i* by VRRM where $W_{ji}^{Srv} \in [0,1]$,
- $\mathbf{R_b}$: vector of serving data rates, which is:

$$\mathbf{R}_{\mathrm{b}} = \left\{ R_{b_{ji}}^{Srv} \middle| j = 1, 2, \dots, N_{srv} \; and \; i = 1, 2, \dots, N_{VNO} \right\} \quad (4.30)$$

The weights in (4.30) are used to prioritise the allocation of data rates to the different services of different VNOs. The choice of these weights is done according to the SLAs between VNOs and VRRM, while their summation equals unity (i.e., they are normalised weights).

$$\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} W_{ji}^{Srv} = 1 \quad (4.31)$$

The weighted throughputs for different serving weights are plotted in Figure 4.10. It can be seen that higher weighted throughputs can be obtained by allocating the same data rate to services with higher data rates. For instance, the services of VNOs classified as background traffic have the lowest weight in (4.29), since it is desired that the service with the higher serving weight receives a higher data rate than the those with lower serving weights.

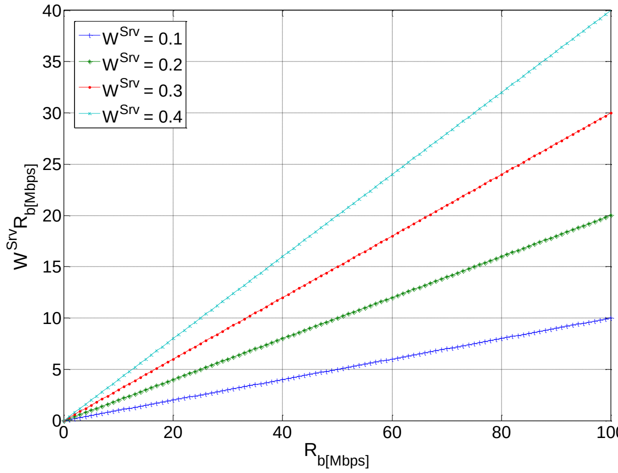


Figure 4.10 – Weighted throughput for different weights.

In addition, there are constraints for VRRM to allocate data rates to various services, which should not be violated. The very fundamental constraint is the total network capacity, estimated in the last section. The summation of all assigned data rates to all services should not exceed the total estimated capacity of the network:

$$\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv} \leq R_{b\,[\mathrm{Mbps}]}^{CRRM} \quad (4.32)$$

The GB and the BG SLAs impose the next set of constraints. The allocated data rate related to these services have to be higher than a minimum guaranteed level (for GB and BF), and lower than a maximum one (for GB only):

$$R_{b_{ji}[\mathrm{Mbps}]}^{Min} \leq R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv} \leq R_{b_{ji}[\mathrm{Mbps}]}^{Max} \quad (4.33)$$

where:

- $R_{b_{ji}}^{Min}$: minimum guaranteed data rate of service *j* of the VNO *i,*
- $R_{b_{ji}}^{Max}$: maximum guaranteed data rate of service *j* of the VNO *i.*

Hence, the allocation procedure can be formulated as a linear optimisation problem as follows:

$$\max_{\mathbf{R_b}} f_{\mathbf{R_b}}^{v}(\mathbf{R_b})$$
$$s.t.\begin{cases} \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} R_{b_{ji}\,[\text{Mbps}]}^{Srv} \leq R_{b\,[\text{Mbps}]}^{CRRM} \\ R_{b_{ji}[\text{Mbps}]}^{Min} \leq R_{b_{ji}\,[\text{Mbps}]}^{Srv} \leq R_{b_{ji}[\text{Mbps}]}^{Max} \end{cases} \tag{4.34}$$

Figure 4.11 visualises the optimisation problem addressed in (4.34). The region presented by the solid colour is the acceptable region for the data rate allocated to each of the VNOs. Any point in this region has data rate higher than 0.4 Gbps for VNO BG (i.e., R_{b1}) and the summation of the allocated data rates is smaller or equal to the total network capacity. The hyperplane of the objective function is also plotted in the figure: as the hyperplane moves to the right of the plot, the achieved value increases.

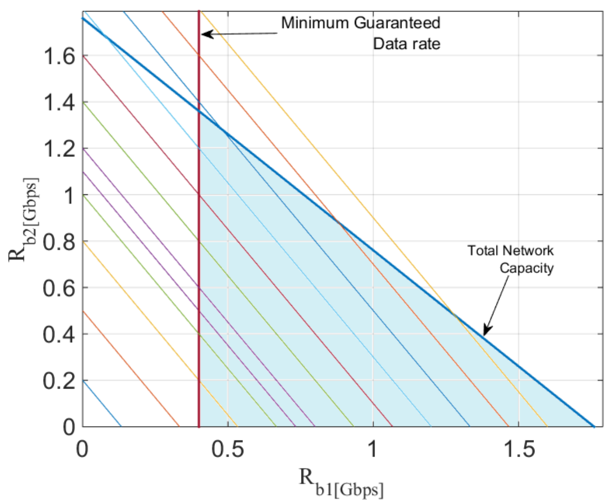


Figure 4.11 – Visualisation of resource allocation for canonical scenario.

However, solving the optimisation problem in the current situation causes the services with the highest serving weight to receive almost all the resources, while the other services are allocated a minimum possible data rate, which is neither fair nor desirable. In contrast, a fair allocation is when the normalised data rate (i.e., data rate divided by the serving weight) of the all services, and consequently the normalised average data rate, has the same value. This can be expressed as:

$$\frac{R_{b_{ji}\,[\text{Mbps}]}^{Srv}}{W_{ji}^{Srv}} - \frac{1}{N_{VNO}N_{srv}} \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \frac{R_{b_{ji}\,[\text{Mbps}]}^{Srv}}{W_{ji}^{Usg}} = 0 \tag{4.35}$$

Nevertheless, the resource efficiency and the fair allocations are two contradict goals: the increment in the one of them leads to the decrement of the other. Hence, instead of having the fairest allocation (i.e., the deviation of the all normalised data rate from the normalised average is zero), the minimisation of the total deviation from the normalised average is used:

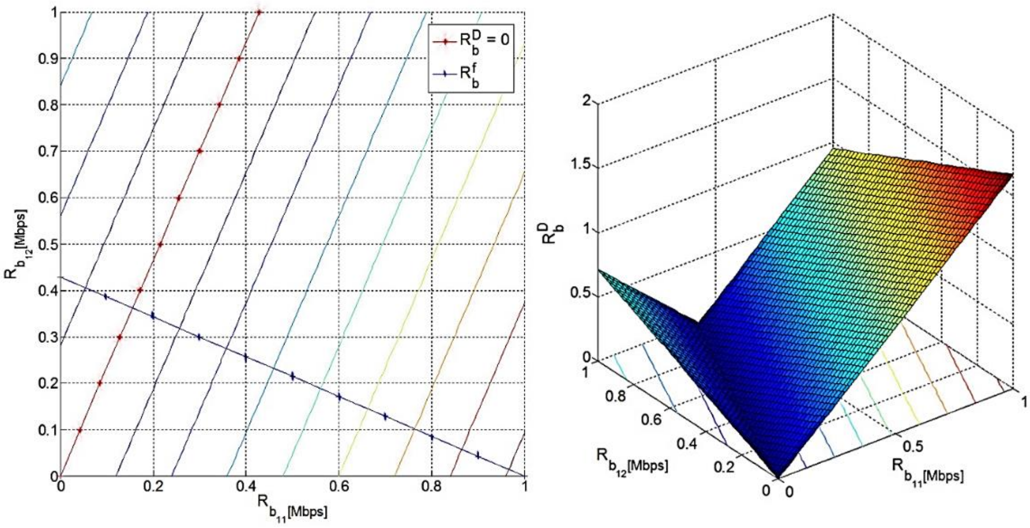
$$\min_{R_{b_{ji}}^{Srv}} \left\{ \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} R_{b_{ji}[\mathrm{Mbps}]}^{D} \right\} \tag{4.36}$$

where:

- $R_{b_{ji}}^{D}$: max. deviation from the normalised average for the service *j* of VNO *i,* given by:

$$R_{b_{ji}[\mathrm{Mbps}]}^{D} = \left| \frac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}}{W_{ji}^{Srv}} - \frac{1}{N_{VNO} N_{srv}} \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \frac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}}{W_{ji}^{Srv}} \right| \tag{4.37}$$

It is worth noting that the fairness for services with minimum guaranteed data rates, applies only to the amount exceeded to the minimum guaranteed level. Equation (4.37) for two service where $W_{11} = 0.3$ and $W_{12} = 0.7$ is illustrated in Figure 4.12.



a) Top View b) 3D view

Figure 4.12 – Maximum deviation from the normalised average as function of services data rate.

The geometrical representation of (4.36) is to minimise the total distances (i.e., shown by R_b^f) from the fair case.

In order to demonstrate the role of serving weights in the fair allocation, let us assume a study case where weighted average data rates of a network is 100 Mbps. Fixing the other services, (4.37) is plotted for a new service with three different serving weight in Figure 4.13. The maximum fairness is when the value for the equation is zero. Thus, allocating 20 Mbps in this scenario to a service with serving weight of 0.2 is a fair allocation. As the serving weight increases, the allocated data rate also has to increase to meet the fairness condition. As the matter of fact, the allocated data rate to the service with 0.4 and 0.6 has to be 40 Mbps and 60 Mbps in order to have fair allocation.

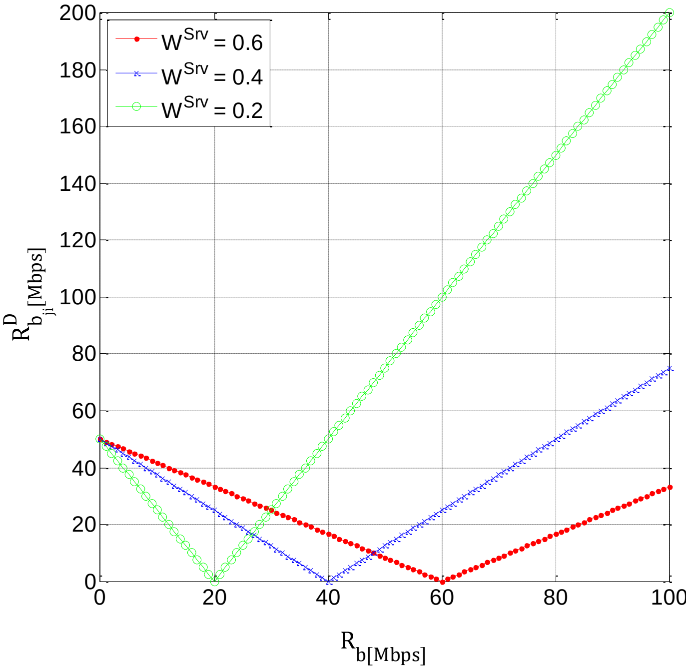


Figure 4.13 – R_b^D for a service with different serving weight.

Equation (4.36) can be also reformulate using the epigraph technique [BoVa09] as:

$$\begin{aligned}&\min_{R_{b_{ji}}^{Srv},R_{b_{ji}}^{f}}\left\{\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}}R_{b_{ji}[\text{Mbps}]}^{f}\right\}\\&s.t.\left|\frac{R_{b_{ji}\,[\text{Mbps}]}^{Srv}}{W_{ji}^{Srv}}-\frac{1}{N_{VNO}N_{srv}}\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}}\frac{R_{b_{ji}\,[\text{Mbps}]}^{Srv}}{W_{ji}^{Srv}}\right|\leq R_{b_{ji}[\text{Mbps}]}^{f}\end{aligned}\tag{4.38}$$

where $R_{b_{ji}}^{f}$ is an intermediate variable used to simplify the problem. In order to achieve Linear Programming (LP), (4.38) has to change as follows:

$$\begin{aligned}&\min_{R_{b_{ji}}^{Srv},R_{b_{ji}}^{f}}\left\{\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}}R_{b_{ji}[\text{Mbps}]}^{f}\right\}\\&s.t.\begin{cases}\dfrac{R_{b_{ji}\,[\text{Mbps}]}^{Srv}}{W_{ji}^{Srv}}-\dfrac{1}{N_{VNO}N_{srv}}\displaystyle\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}}\frac{R_{b_{ji}\,[\text{Mbps}]}^{Srv}}{W_{ji}^{Srv}}\leq R_{b_{ji}[\text{Mbps}]}^{f}\\\dfrac{-R_{b_{ji}\,[\text{Mbps}]}^{Srv}}{W_{ji}^{Srv}}+\dfrac{1}{N_{VNO}N_{srv}}\displaystyle\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}}\frac{R_{b_{ji}\,[\text{Mbps}]}^{Srv}}{W_{ji}^{Srv}}\leq R_{b_{ji}[\text{Mbps}]}^{f}\end{cases}\end{aligned}\tag{4.39}$$

In order to discuss better the balance of these two objectives, the boundaries of these two objective has to be compared. Obviously, the highest resource efficiency (i.e., the highest weighted throughput) is obtained when all the resources are allocated to the service(s) with the highest serving weight. Hence, the maximum of the first objective can be written as:

$$\max_{R_{b_{ji}}^{Srv}} \left\{ \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} W_{ji}^{Srv} \, R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv} \right\} = \max\left\{W_{ji}^{Srv}\right\} R_{b\,[\mathrm{Mbps}]}^{CRRM} \tag{4.40}$$

It means that as the network capacity increases, the summation of weighted throughput in (4.29) increases as well. In the same situation, the fairness objective function also reaches its maximum:

$$\max_{R_{b_{ji}}^{Srv}} \left\{ \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} R_{b_{ji}[\mathrm{Mbps}]}^{f} \right\} = \frac{R_{b\,[\mathrm{Mbps}]}^{CRRM}}{\max\left\{W_{ji}^{Srv}\right\}} \tag{4.41}$$

Therefore, the combination of (4.34) with (4.39) is done by means of the adoption of the fairness intermediate variable, R_{ji}^{f}, with the network's capacity, as follows:

$$f_{\mathbf{R_b}}^{v}(\mathbf{R_b}) \;= f_{\mathbf{R_b}}^{cell}\left(\mathbf{R_b^{\mathrm{cell}}}\right) - \;\alpha_f(W_f) f_{\mathbf{R_b}}^{f}\left(\mathbf{R_b^{f}}\right) \tag{4.42}$$

where:

- $f_{\mathbf{R_b}}^{cell}$: objective function for cellular RATs addressed in (4.29),
- $\mathbf{R_b^{\mathrm{cell}}}$: vector of serving data rates from cellular network, which is:

$$\mathbf{R_b^{\mathrm{cell}}} = \{R_{b_{ji}}^{cell} | j = 1, 2, \dots, N_{srv} \; and \; i = 1, 2, \dots, N_{VNO}\} \tag{4.43}$$

- $f_{\mathbf{R_b}}^{f}$: fairness function given by:

$$f_{\mathbf{R_b}}^{f}\left(\mathbf{R_b^{f}}\right) = \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \left(R_{b_{ji}[\mathrm{Mbps}]}^{f}\right) \tag{4.44}$$

- $\mathbf{R_b^{f}}$: vector of intermediate fairness variable,
- W_f: fairness weight,
- α_f: fairness coefficient as function of fairness weight.

Hence, the coefficient is proposed as follows:

$$\alpha_f\left(w_f\right) = \frac{w_f R_{b\,[\mathrm{Mbps}]}^{CRRM}}{(1 - w_f) R_{b\,[\mathrm{Mbps}]}^{CRRM} N^{ViS} + w_f \overline{R_{b[\mathrm{Mbps}]}^{max}}} \tag{4.45}$$

where:

- $\overline{R_b^{max}}$: maximum average data rate among all the network services (i.e., video streaming),
- N^{ViS}: number of subscriber using the video streaming service.

$$N^{Vis} = \frac{R_{b\,[\mathrm{Mbps}]}^{CRRM}}{R_{b[\mathrm{Mbps}]}^{max}} \quad (4.46)$$

The fairness coefficient for a practical network as a function of fairness is plotted in Figure 4.14: According to the figure, the objective function decreases as the fairness weight increases since the effect of not having the ideal fairness increases. However, the maximum possible fairness can be achieved when the fairness weight is set to unit.

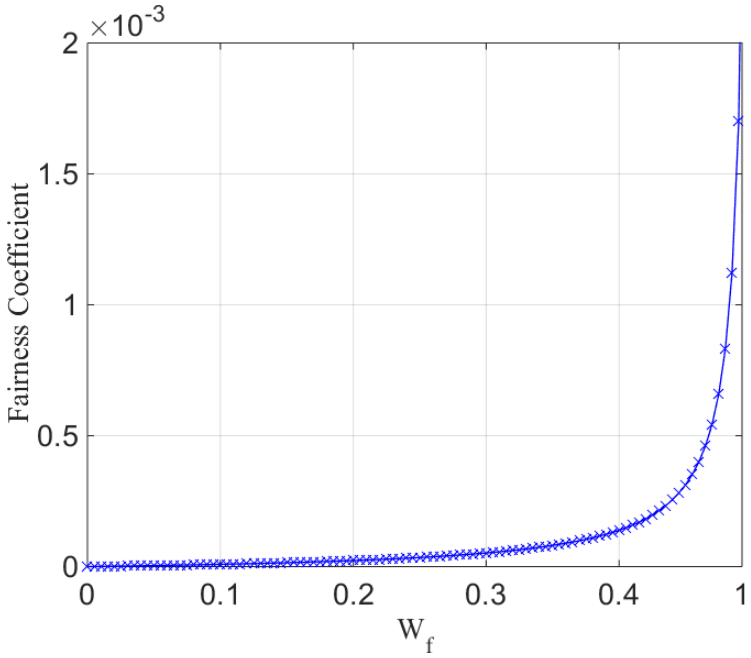


Figure 4.14 – Fairness coefficient vs fairness weight to the reference scenario.

4.3.4 Supporting Full Heterogeneous Access Network

Late studies, [CFHH05, TDGJ06], suggest that an acceptable portion of traffic can be offloaded to Wi-Fi APs just by allowing users to delay their delay-tolerant data for a maximum pre-specified interval, until reaching an AP. Since almost all MTs today have other connectivity capabilities, the offloading approaches are based on using them instead of expensive cellular bands whenever it is possible. In addition, the current cellular networks architecture, such as LTE, supports traffic offloads into Wi-Fi or any other access types too (Figure 2.4). Hence, in the next step the model for management of virtual radio resources is extended to support traffic offloading to Wi-Fi. Network architecture and throughput dependencies are two reasons for which cellular access network and Wi-Fi are considered separately in this thesis.

Referring to the network architecture depicted in Figure 2.4, the connection of APs to the network core is different from cellular RATs. The latter (i.e., GSM, UMTS, and LTE), which are considered as 3GPP trusted networks, are connected to S-GW, but the connection of the former, which may be considered as either trusted or untrusted non-3GPP networks, to the network core (i.e., MME) is done through P-GW instead of S-GW.

In addition, the throughput of a Wi-Fi network depends on collision rates and channel quality, while it can be claimed that the throughput in cellular networks is almost independent of the number of connected MTs. Since increasing the number of end-user leads to the increment of the interference level, the network throughput still depends basically on SINR. In contention-based networks, such as Wi-Fi, however, the increment of connected end-users not only decreases SINR but also increases conflict rates; whenever a collision occurs, MTs have to back-off for a while and then retransmit their packets. By not taking the back-off time interval after each conflict, the retransmission of packets, itself, leads to the lower network throughput, thus, the objective function has to minimise the number of connected users to APs in order to reduce the collision rate and increase the throughput. Hence, the number of connected MTs is given by:

$$N_u^{AP} = \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{Srv}} N_{u_{ji}}^{AP} \tag{4.47}$$

where:

- N_u^{AP}: total number users connect to APs,
- $N_{u_{ji}}^{AP}$: number of users of *i*-VNO using service *j* connected to APs.

By considering the average data rate of a service and the allocated capacity, the number of connected users is:

$$N_u^{AP} = \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{Srv}} \frac{R_{b_{ji}[\mathrm{Mbps}]}^{WLAN}}{\overline{R_{b_j[\mathrm{Mbps}]}}} \tag{4.48}$$

where:

- $R_{b_{ji}}^{WLAN}$: allocated capacity for service *j* of the VNO *i* from WLAN,
- $\overline{R_{b_j}}$: the average data rate for service *j*.

Based on (4.48), one can see that by that allocating a larger portion of network capacity to services with higher average data rates leads to having a smaller number of users in the network. Hence, there are two key objectives in modelling the traffic offloading capability: increment of network throughput and decrement of connected users. Although these two objectives are functions of the allocated data rate, they are not o the same type; in order to be able to optimise both objectives at the same time, the maximum summation of weighted data, an equivalent objective to the minimum number of connected users, is introduced.

The weighted data rate of a service in Wi-Fi network is defined by:

$$R_{b_{ji}[Mbps]}^{WAP} = \frac{\overline{R_{b_j[\mathrm{Mbps}]}}}{R_{b[\mathrm{Mbps}]}^{max}} R_{b_{ji}[\mathrm{Mbps}]}^{WLAN} \tag{4.49}$$

where:

- $R_{b_{ji}}^{WAP}$: weighted data rate for service *j* of the VNO *i,*
- $\overline{R_b^{max}}$: maximum average data rate among all the network services (i.e., video streaming).

By normalising the average data rate using the maximum average data rate, the weights in (4.49) are bound to unit:

$$\frac{\overline{R_{b_j[\text{Mbps}]}}}{\overline{R_{b[\text{Mbps}]}^{max}}} \leq 1 \tag{4.50}$$

For two services with the same allocated capacity from Wi-Fi, the one with a higher average data rate has the higher weighted data rate and the lower number of connected users. Maximising the summation of weighted data rates as a function of the allocated data rate for each service in the Wi-Fi network is equivalent to minimising the number of connected users:

$$min \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{Srv}} N_{u_{ji}}^{AP} \Leftrightarrow max \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{Srv}} R_{b_{ji}[\text{Mbps}]}^{WAP} \tag{4.51}$$

So, instead of minimising the number of connected users, the maximisation of the summation of weighted data rates for all network services can be used in the objective function. Hence, the objective function of VRRM for Wi-Fi has to be revises as follows:

$$f_{\mathbf{R_b}}^{WLAN}\left(\mathbf{R}_\mathbf{b}^{\text{WLAN}}\right) = \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} W_{ji}^{Srv} R_{b_{ji}[\text{Mbps}]}^{WLAN} + W^{SRb} \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \frac{\overline{R_{b_j[\text{Mbps}]}}}{\overline{R_{b[\text{Mbps}]}^{max}}} R_{b_{ji}[\text{Mbps}]}^{WLAN} \tag{4.52}$$

where:

- f_v^{AP}: objective function for APs,
- W^{SRb}: average session data rate weight, where $W^{SRb} \in [0, 1]$,
- $\mathbf{R}_b^{\text{WLAN}}$: vector of serving data rates from APs, which is:

$$\mathbf{R}_\mathbf{b}^{\text{WLAN}} = \{R_{b_{ji}}^{WLAN} | j = 1, 2, \dots, N_{srv} \; and \; i = 1, 2, \dots, N_{VNO}\} \tag{4.53}$$

In (4.52), W^{SRb} is used to control the weight of the average data rate per session in resources allocation. Obviously, assigning zero to this weight, completely eliminates the average data rate effect and converts the objective function of WLAN addressed in (4.52) to the cellular one presented in (4.29). Consequently, the objective function of a RAN with both cellular and Wi-Fi is:

$$f_{\mathbf{R_b}}^{v}(\mathbf{R_b}) = W^{WLAN} f_{\mathbf{R_b}}^{WLAN}\left(\mathbf{R}_\mathbf{b}^{\text{WLAN}}\right) + W^{cell} f_{\mathbf{R_b}}^{cell}\left(\mathbf{R}_\mathbf{b}^{\mathbf{cell}}\right) - \alpha_f\left(w_f\right) f_{\mathbf{R_b}}^{f}\left(\mathbf{R}_\mathbf{b}^{\mathbf{f}}\right) \tag{4.54}$$

where:

- W^{WLAN}: weight for allocating capacity from APs,
- f_v^{WLAN}: objective function presented in (4.52),
- W^{cell}: weight for allocating capacity from cellular RATs,

In (4.54), the allocated data rate for a specific service is defined as:

$$R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]} = R^{cell}_{b_{ji}\,[\mathrm{Mbps}]} + R^{WLAN}_{b_{ji}\,[\mathrm{Mbps}]} \tag{4.55}$$

4.3.5 Resource Allocation in Congest Situations

Due to changes in the physical infrastructure, users' channel, etc., in practice, there are situations where the resources are not enough to meet all guaranteed capacity, and allocation optimisation, as addressed in previous sections, is no longer feasible. A simple approach in these cases is to relax the constraints by the introduction of violation (also known as slack) variables. The new objective function contains the objective function of the original problem, plus the penalty for violations. In case of VRRM, the relaxed optimisation problem can be considered by adding a violation parameter to (4.33), given by:

$$\begin{aligned} &R^{Min}_{b_{ji}[\mathrm{Mbps}]} \leq R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]} + \Delta R^{v}_{b_{ji}\,[\mathrm{Mbps}]} \\ &\Delta R^{v}_{b_{ji}\,[Mbps]} \geq 0 \end{aligned} \tag{4.56}$$

where:

- $\Delta R^{v}_{b_{ji}}$: violation variable for the minimum guaranteed data rate service *j* of the VNO *i.*

By introducing the violation parameter, the former infeasible optimisation problem turns into a feasible one. The optimal solution maximises the objective function and minimises the average constraints violations. The average constraints violation is defined as follows:

$$\Delta \overline{R^{v}_{b}}_{[\mathrm{Mbps}]} = \frac{1}{N_{VNO} N_{srv}} \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} W^{v}_{ji}\, \Delta R^{v}_{b_{ji}\,[\mathrm{Mbps}]} \tag{4.57}$$

where

- $\Delta \overline{R^{v}_{b}}$: average constraint violation,
- W^{v}_{ji}: weight of violating minimum guaranteed data rate service *j* of the VNO *i* where $W^{v}_{ji} \in [0,1]$.

The objective function presented in (4.29) also has to be changed. The new objective function, the relaxed one, has to contain the minimisation of violations in addition to the maximisation of former objectives. Although the average constraint violation has a direct relation to the allocated data rate to services where the increment in one leads to decrement of the other, it does not have the same relation with fairness. It can be claimed that the maximisation of fairness and minimisation of constraints violations are independent. Therefore, the final objective function considering both issues has to consider the same approach for minimisation of violations as it does for fairness. In other words, the fairness variable is weighted as presented in (4.32) to compensate for the summation of weighted data rates of various services. The derivation from fair allocation, which is desired to be as minimum as possible, gains relatively a higher weight in the objective function and may confiscate the constraints violation strategies. Therefore, the average constraint violation also has two places in the objective function in a similar way:

$$f_{\mathbf{R_b}}^{v}(\mathbf{R_b}) = W^{WLAN} f_{\mathbf{R_b}}^{WLAN}\left(\mathbf{R_b^{WLAN}}\right) + W^{cell} f_{\mathbf{R_b}}^{cell}\left(\mathbf{R_b^{cell}}\right) - f_{R_b^v}^{vi}\left(\Delta\overline{R_b^v}\right) - f_{\mathbf{R_b}}^{f}\left(\mathbf{R_b^f}\right) \tag{4.58}$$

where:

- $f_{R_b^v}^{vi}$ is the constraint violation function:

$$f_{R_b^v}^{vi}\left(\Delta\overline{R_b^v}\right) = \frac{R_{b\,[\mathrm{Mbps}]}^{CRRM}}{R_{b[\mathrm{Mbps}]}^{min}} \Delta\overline{R_b^v}_{[\mathrm{Mbps}]} \tag{4.59}$$

The definition of fairness, however, in a congestion situation is not the same. The fairness objective in the normal case is to have the same normalised data rate for all services. As a reminder, when the network faces congestion, there are not enough resources to serve all services with the minimum acceptable data rates. Therefore, not only all best effort services are not allocated any capacity, but some violation is also introduced to the guaranteed data rates. Now, fairness is to make sure that the weighted violation of all services is the same. The ideal fairness with this definition is as follows:

$$W_{ji}^{v} \Delta R_{b_{ji}\,[\mathrm{Mbps}]}^{v} - \frac{1}{N_{VNO} N_{srv}} \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} W_{ji}^{v} \Delta R_{b_{ji}\,[\mathrm{Mbps}]}^{v} = 0 \tag{4.60}$$

Obviously, the violation data rates for the best effort services are always zero. Consequently, the fairness equation is changed as follows:

$$\min_{R_{b_{ji}}^{Srv}, R_{b_{ji}}^{f}} \left\{ \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} R_{b_{ji}[\mathrm{Mbps}]}^{f} \right\}$$
$$\begin{cases} W_{ji}^{v} \Delta R_{b_{ji}\,[\mathrm{Mbps}]}^{v} - \dfrac{1}{N_{VNO} N_{srv}} \displaystyle\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} W_{ji}^{v} \Delta R_{b_{ji}\,[\mathrm{Mbps}]}^{v} \le R_{b_{ji}[\mathrm{Mbps}]}^{f} \\ -W_{ji}^{v} \Delta R_{b_{ji}\,[\mathrm{Mbps}]}^{v} + \dfrac{1}{N_{VNO} N_{srv}} \displaystyle\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} W_{ji}^{v} \Delta R_{b_{ji}\,[\mathrm{Mbps}]}^{v} \le R_{b_{ji}[\mathrm{Mbps}]}^{f} \end{cases} \tag{4.61}$$

4.3.6 Fair Resource Allocation

Using serving weights, W_{ji}^{Srv}, beside violating weights, W_{ji}^{v}, enables the prioritisation of services in both normal and congestion situations. However, the fairness concept in VRRM also has to consider the traffic mixture of operators in addition to those aforementioned weights. Mobile video, for instance, exhausts significantly a larger portion of network capacity comparing to VoIP or music, according to [Cisc13a]. Allocating the same capacity to mobile video as to VoIP when they have the same priority is obviously not fair at all. On the ground of this discussion, the normalised data rate has to consider traffic usage in addition to serving/violation weight as follows:

$$R^N_{b_{ji}\,[\mathrm{Mbps}]} = \begin{cases} \dfrac{R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]}}{W^{Srv}_{ji} W^{Usg}_{ji}}, & \text{Normal Situation} \\ \dfrac{W^v_{ji}\Delta R^v_{b_{ji}\,[\mathrm{Mbps}]}}{W^{Usg}_{ji}}, & \text{Congestion Situation} \end{cases} \tag{4.62}$$

where:

- $R^N_{b_{ji}}$: normalised data rate,
- W^{Usg}_{ji}: traffic usage weight, where $W^{Usg}_{ji} \in [0,1]$,

Hence, the ideal fair allocation is when the following condition applies:

$$R^D_{b_{ji}[\mathrm{Mbps}]} = \begin{cases} \left| \dfrac{R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]}}{W^{Srv}_{ji} W^{Usg}_{ji}} - \overline{R^N_{b_{[\mathrm{Mbps}]}}} \right|, & \text{Normal Situations} \\ \left| \dfrac{W^v_{ji}\Delta R^v_{b_{ji}\,[\mathrm{Mbps}]}}{W^{Usg}_{ji}} - \overline{R^N_{b_{[\mathrm{Mbps}]}}} \right|, & \text{Congestion Situation} \end{cases} \tag{4.63}$$

where:

$$\begin{cases} \overline{R^N_{b_{[\mathrm{Mbps}]}}} = \dfrac{1}{N_{VNO} N_{srv}} \displaystyle\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \dfrac{R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]}}{W^{Srv}_{ji} W^v_{ji} W^{Usg}_{ji}}, & \text{Normal Situations} \\ \overline{R^N_{b_{[\mathrm{Mbps}]}}} = \dfrac{1}{N_{VNO} N_{srv}} \displaystyle\sum_{i=1}^{N_{VNO}} \dfrac{W^v_{ji}\Delta R^v_{b_{ji}\,[\mathrm{Mbps}]}}{W^{Usg}_{ji}}, & \text{Congestion Situation} \end{cases} \tag{4.64}$$

In order to show the role of serving weight in former equation, assume the case where the weighted average data rates of a network are 100 Mbps. Fixing the other services, it is desired to study (4.63) and (4.64) for a new service with different serving, violation, and usage weight; the results are plotted in Figure 4.15 and Figure 4.16.

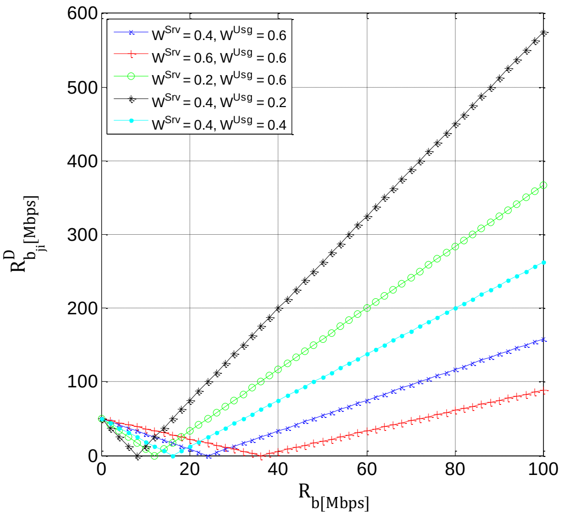


Figure 4.15 – Maximum deviation from weighted average (normal situations).

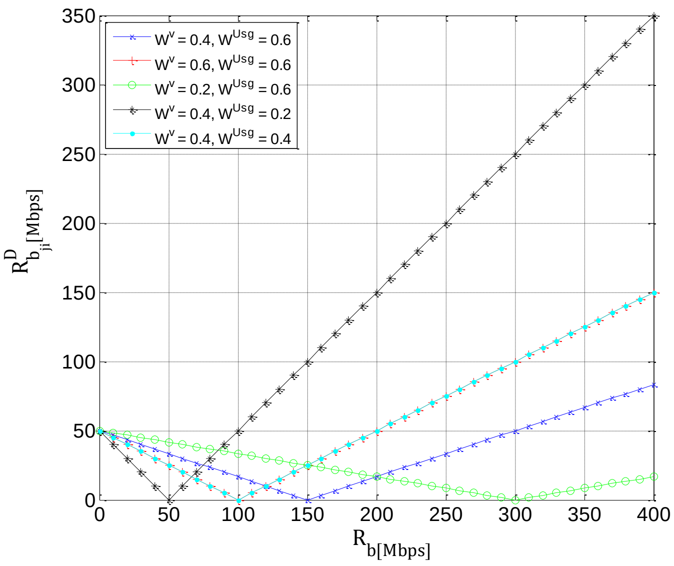


Figure 4.16 – Maximum deviation from weighted average (congestion situations).

4.3.7 VRRM Interaction with CRRM and Local RRMs

Although each local RRM works independent of the others, CRRM can coordinate them through a set of policies. The main objective of these policies is to consider the problem of allocating RRUs from an access technique viewpoint with various criteria, such as load balancing among resource pools. In brief, it can be said that CRRM is in charge of RAT selection and vertical handover. The key part of the objective function for CRRM can be written as follows:

$$f^{c}_{\mathbf{R}_\mathbf{b}}(\mathbf{R}_\mathbf{b}) = \sum_{i=1}^{N_{VNet}} \sum_{j=1}^{N_{srv}} \sum_{k=1}^{N_{RAT}} \left(W_c^{Srv} R^k_{b_{ji}\,[\mathrm{Mbps}]}\right) \tag{4.65}$$

where:

- N^K_{ji}: Number of RRU from RAT k allocated to service j of the VNO i,
- W_c^{Srv}: Weight of serving unit of data rate by CRRM,
- $R^k_{b_{ji}}$: Data rate served for service j of the VNO i from RAT k.

CRRM applies techniques, such as load balancing among different RAT resource pools, to maximise the objective function under some constraints, the main one being the limited RRUs. The summation of all data rates assigned to all services of all VNOs from a RAT resource pool should always be less than or equal to the expected total data rate for that RAT.

$$\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} R^k_{b_{ji}[\mathrm{Mbps}]} \leq \tilde{R}^k_{b_T\,[\mathrm{Mbps}]}, k = 1, \dots, N_{RAT} \tag{4.66}$$

where:

- $\tilde{R}^k_{b_T}$ is expected total data rate for the RAT *k*.

Furthermore, CRRM has to abide VRRMs' policy set, which imposes some more constraints. Starting with data rate, the total data rate offered to a service of a VNO from various RATs should be equal or higher than the requested rate asked by VRRM, $R_{b_{ji}}^{VRRM}$:

$$R_{b_{ji}[\mathrm{Mbps}]}^{VRRM} \leq \sum_{k=1}^{N_{RAT}} R_{ji\,[\mathrm{Mbps}]}^{k} \tag{4.67}$$

CRRM, like local RRMs, provides the higher level, VRRM, with periodic reports. The report of CRRM contains general information on QoS metrics and estimated remaining data rate. VRRM, receiving these reports, tries to allocate the available capacity to various services of various VNOs. The allocation should be done so that the guaranteed level of service is met.

4.3.8 Interior-Point Linear Programming Optimisation

Linear programming is the problem of finding a vector $\mathbf{x}$ that minimises a linear function subject to linear constraints. It generally can be written as [Math15]:

$$\begin{aligned} &\min_{\mathbf{x}} \mathbf{f}^{\mathrm{T}} \cdot \mathbf{x} \\ &s.t. \begin{cases} \mathbf{A} \cdot \mathbf{x} \leq b \\ \mathbf{A_{eq}} \cdot \mathbf{x} = \mathbf{B_{eq}} \\ l_i \leq x_i \leq u_i \end{cases} \end{aligned} \tag{4.68}$$

where:

- $\mathbf{f}$**:** linear objective function vector,
- $\mathbf{A}$**:** matrix for linear inequality constraints,
- $\mathbf{b}$**:** vector for linear inequality,
- $\mathbf{A_{eq}}$**:** matrix for linear equality constraints,
- $\mathbf{B_{eq}}$: vector for linear equality constraints,
- $\mathbf{l}$: vector of lower bounds,
- $\mathbf{U}$**:** vector of upper bounds.

The interior-point method is based on LIPSOL (Linear-programming Interior-Point SOLvers) [Zhan95], which is a variant of Mehrotra's predictor-corrector algorithm [Mehr92], a primal-dual interior-point method. The algorithm begins by applying a series of pre-processing steps before the actual iterative algorithm begins to achieve the transformed problem where:

- All variables are lower bounded by zero.
- All constraints are equalities.
- Fixed variables, those with equal upper and lower bounds, are removed.
- Rows of all zeros in the constraint matrix are removed.
- The constraint matrix has full structural rank.
- Columns of all zeros in the constraint matrix are removed.

- When a significant number of singleton rows exist in the constraint matrix, the associated variables are solved for and the rows removed.

After pre-processing, the problem can be presented in the following form:

$$\min_{\mathbf{x}} \mathbf{f}^T \cdot \mathbf{x} \ such\ that \begin{cases} \boldsymbol{A} \cdot \mathbf{x} = b \\ 0 \le x_i \le u_i \end{cases} \tag{4.69}$$

Lower bounds constraints are implicitly included in the constraint matrix. With the addition of primal slack variables, (4.69) becomes:

$$\min_{\mathbf{x}} \mathbf{f}^T \cdot \mathbf{x} \ such\ that \begin{cases} \mathbf{A} \cdot \mathbf{x} = b \\ \mathbf{x} + \mathbf{s} = \mathbf{u} \\ \mathbf{x} \ge 0, \mathbf{s} \ge 0 \end{cases} \tag{4.70}$$

where:

- **s:** the primal slack variable vector.

This equation is referred to as the primal problem: $\mathbf{x}$ consists of the primal variables and $\boldsymbol{s}$ consists of the primal slack variables. The dual problem is [BoVa09]:

$$\max_{x} \mathbf{b}^T \cdot \mathbf{y} - \mathbf{u}^T \cdot \mathbf{w} \ such\ that \begin{cases} \boldsymbol{A}^T \cdot \boldsymbol{y} - \boldsymbol{w} + \boldsymbol{z} = \boldsymbol{f} \\ \boldsymbol{z} \ge 0, \boldsymbol{w} \ge 0 \end{cases} \tag{4.71}$$

where:

- $\mathbf{y}, \mathbf{w}$: consist of the dual variables,
- $\mathbf{z}$: consists of the dual slacks.

The optimality conditions for this linear program, i.e., the primal (4.70) and the dual (4.71) are

$$\boldsymbol{F}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{s}, \mathbf{w}) = \begin{pmatrix} \mathbf{A} \cdot \mathbf{x} - \mathbf{b} \\ \mathbf{x} + \mathbf{s} - \mathbf{u} \\ \mathbf{A}^T \cdot \mathbf{y} - \mathbf{w} + \mathbf{z} - \mathbf{f} \\ x_i z_i \\ s_i w_i \end{pmatrix} = 0, \tag{4.72}$$
$$\mathbf{x} \ge 0, \mathbf{z} \ge 0, \mathbf{s} \ge 0, \mathbf{w} \ge 0,$$

where $x_i z_i$ and $s_i w_i$ denote component-wise multiplications.

The quadratic equations, presented in (4.72) are called the complementarity conditions for the linear program, which are:

$$\begin{cases} x_i z_i = 0 \\ s_i w_i = 0 \end{cases} \tag{4.73}$$

The other (linear) equations are called the feasibility conditions. The duality gap can be calculated using the following equation.

$$\mathbf{x}^{\mathrm{T}}\mathbf{z} + \mathbf{s}^{\mathrm{T}}\mathbf{w} \tag{4.74}$$

The algorithm is a primal-dual algorithm, meaning that both the primal and the dual programs are solved simultaneously. It can be considered a Newton-like method [BoVa09], applied to the linear-quadratic system in (4.72), while at the same time keeping the iterates $\mathbf{x}, \mathbf{z}, \mathbf{w},$ and $\mathbf{s}$ positive, thus the name interior-point method. The algorithm is a variant of the predictor-corrector algorithm proposed by Mehrotra. Consider an iterate as follows:

$$\mathbf{v} = [\mathbf{x} ; \mathbf{z} ; \mathbf{s} ; \mathbf{w}] > 0 \tag{4.75}$$

where $\mathbf{v}$ is vector of variables. The first compute called prediction direction is

$$\Delta\mathbf{v}_{\mathbf{p}} = -\left(\boldsymbol{F}^{T}(\mathbf{v})\right)^{-1}\boldsymbol{F}(\mathbf{v}) \tag{4.76}$$

where $\Delta\mathbf{v}_{\mathbf{p}}$ is the Newton direction; then

$$\Delta\mathbf{v}_{\mathbf{c}} = -\left(\boldsymbol{F}^{T}(\mathbf{v})\right)^{-1}\boldsymbol{F}\left(\mathbf{v} + \Delta\mathbf{v}_{\mathbf{p}}\right) - \mu\mathbf{e} \tag{4.77}$$

where:

- $\Delta\mathbf{v}_{\mathbf{c}}$**:** correction direction,
- $\mu > 0$: centering parameter,
- $\mathbf{e}$: zero-one vector with the ones corresponding to the quadratic equations in $\mathbf{F}(\mathbf{v})$.

The two directions are combined with a step length parameter α_{st} and update $\mathbf{v}$ to obtain the new iterate:

$$\mathbf{v}^{+} = \mathbf{v} + \alpha_{st}(\Delta\mathbf{v}_{\mathbf{p}} + \Delta\mathbf{v}_{\mathbf{c}}) \tag{4.78}$$

where

- $\mathbf{v}^{+}$: new iterate variable,
- α_{st}>0:the step length parameter is chosen so that:

$$\mathbf{v}^{+} = [\mathbf{x}^{+} ; \mathbf{z}^{+} ; \mathbf{s}^{+} ; \mathbf{w}^{+}] > 0 \tag{4.79}$$

In solving for the preceding predictor/corrector directions, the algorithm computes a (sparse) direct factorisation on a modification of the Cholesky factors [BoVa09] of $\mathbf{AA^T}$**.** If **A** has dense columns, it instead uses the Sherman-Morrison formula, see [Mehr92] for more details. If that solution is not adequate (the residual is too large), it performs an LDL factorisation of an augmented system form of the step equations to find a solution. The algorithm then loops until convergence. The main stopping criteria is a standard one:

$$\frac{||\mathbf{r_b}||}{max(1, ||\mathbf{b}||)} + \frac{||\mathbf{r_f}||}{max(1, ||\mathbf{f}||)} + \frac{||\mathbf{r_u}||}{max(1, ||\mathbf{u}||)} + \frac{\left|\mathbf{f^T} \cdot \mathbf{x} - \mathbf{b^T} \cdot \mathbf{y} + \mathbf{u^T} \cdot \mathbf{w}\right|}{max(1, |\mathbf{f^T} \cdot \mathbf{x}|, |\mathbf{b^T} \cdot \mathbf{y} - \mathbf{u^T} \cdot \mathbf{w}|)} \leq r_t \tag{4.80}$$

where:

- r_t: tolerance
- $\mathbf{r_b}$: primal residual, which is:

$$\mathbf{r_b} = \mathbf{A} \cdot \mathbf{x} - \mathbf{b} \tag{4.81}$$

- $\mathbf{r_f}$: dual residual, which is:

$$\mathbf{r_f} = \mathbf{A^T} \cdot \mathbf{y} - \mathbf{w} + \mathbf{z} - \mathbf{f} \tag{4.82}$$

- $\mathbf{r_u}$:upper-bound feasibility

$$\mathbf{r_u} = \mathbf{x} + \mathbf{s} - \mathbf{u} \tag{4.83}$$

- $\mathbf{f^T} \cdot \mathbf{x} - \mathbf{b^T} \cdot \mathbf{y} + \mathbf{u^T} \cdot \mathbf{w}$**:** the difference between the primal and dual objective values.

The sum in the stopping criteria measures the total relative errors in the optimality conditions in (4.72).

4.4 Partial VRRM

The former sections presented the modelling of VRRM addressing the objectives in the form of an optimisation problem. Since network status and constraints vary during time, this problem is an uncertain non-convex optimisation problem [WiGo04]. However, there are techniques to solve these kinds of problems, where "Partial VRRM" among various approaches is one of the simplest. In this approach, the optimisation problem is tackled by breaking the main problem into multiple sub-problems. In the case of VRRM, the time axis is divided into decision windows and the objective function is maximised in each of these intervals, independently. However, it is worth noting that decisions in each interval affect directly network state, and the outcome of a policy at a certain point depends on the decisions and states in former intervals. The optimal solution has to take this dependency of decisions also into consideration. On the ground of this discussion, the output of partial VRRM may be only a local minimum and not the global one. Nevertheless, the partial VRRM is a simple solution, which can be used as the starting step and reference point. The outcome of more sophisticated approaches can then be compared with these results.

Figure 4.17 illustrates the decision window of VRRM, CRRM, and LRRMs (Local RRMs). The VRRM decision window contains multiple CRRM ones, during which CRRM applies the decided policy set. In the next decision window of VRRM, after multiple network stages, the VRRM updates the network situation and makes a new decision for the next time interval.

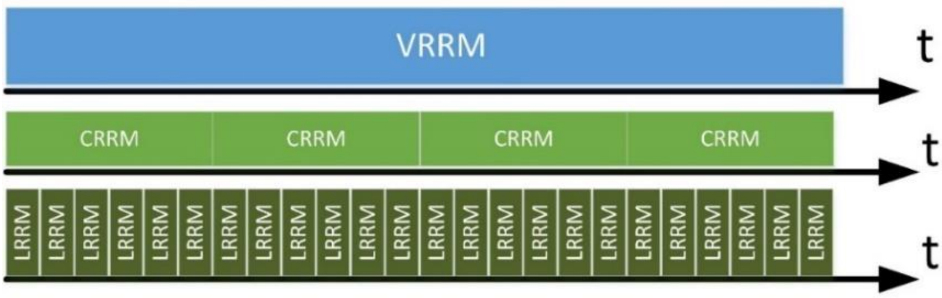


Figure 4.17 – Decision window of VRRM and CRRM.

Considering the objective function presented in previous sections, it can be considered as a summation of objective functions over time intervals in partial VRRM as follows:

$$f_{\mathbf{R_b}}^{v}(\mathbf{R_b}) = \sum_{t_i=0}^{n_t} f_{\mathbf{R_b}}^{v}(\mathbf{R_b}[t_k]) \tag{4.84}$$

where:

- n_t: number decision window,
- $\mathbf{R_b}[t_k]$: allocated data rate vector for at time interval t_k given by:

$$\mathbf{R_b}[t_i] = \{R_{b_{ji}}^{Srv}[t_k] | j = 1, 2, \ldots, N_{srv} \; and \; i = 1, 2, \ldots, N_{VNO}\} \tag{4.85}$$

- $R_{b_{ji}}^{Srv}[t_k]$: Served (allocated) data rate for service *j* of the VNO *i* at the time frame t_k,

The goal is to maximise the objective function for all running time. Considering the objective function in each time interval independent of the other time intervals, it can be written as:

$$\max_{\mathbf{R_b}} f_{\mathbf{R_b}}^{v}(\mathbf{R_b}) \Leftrightarrow \sum_{t_k=0}^{n} \max_{\mathbf{R_b}[\boldsymbol{t_k}]} f_{\mathbf{R_b}}^{v}(\mathbf{R_b}[t_k]) \tag{4.86}$$

As the VRRM starts to work, it has to gather information about serving VNOs plus their services and SLAs. This stage is referred to as “Network Initiation” in the flowchart. Estimating total systems (V-RAN) capacity is the next step. VRRM uses the information about the number of available RRUs in different RATs, and estimate the probability function of network capacity (Section 4.3.2). The allocation of the available capacity to VNOs services is done by solving the Linear Programming problem. Based on this discussion, the optimisation problem in each VRRM’s decision window for the case where only cellular RATs are considered is as follows:

$$\max_{\mathbf{R_b}[t_k], \mathbf{R_b^f}[t_k]} \left\{ \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \left(W_{ji}^{Srv} R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k] - \frac{R_{b\,[Mbps]}^{CRRM}}{R_{b[Mbps]}^{min}} R_{b_{ji}[\mathrm{Mbps}]}^{f}[t_k] \right) \right\}$$

$$s.t.: \begin{cases} \mathbf{1}^T \cdot \mathbf{R_b}[t_k] \le R_{b\,[\mathrm{Mbps}]}^{CRRM}[t_k] \\ R_{b_{ji}[\mathrm{Mbps}]}^{Min}[t_k] \le R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k] \\ R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k] \le R_{b_{ji}[\mathrm{Mbps}]}^{Max}[t_k] \\ \dfrac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k]}{W_{ji}^{Srv}} - \dfrac{1}{N_{VNO}N_{srv}} \displaystyle\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \dfrac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k]}{W_{ji}^{Srv}} \le R_{b_{ji}[\mathrm{Mbps}]}^{f}[t_k] \\ -\dfrac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k]}{W_{ji}^{Srv}} + \dfrac{1}{N_{VNO}N_{srv}} \displaystyle\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \dfrac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k]}{W_{ji}^{Srv}} \le R_{b_{ji}[Mbps]}^{f}[t_k] \end{cases} \tag{4.87}$$

In each decision window, the network status has to be updated. The network status at time interval t_i contains the following items:

- Remained network capacity, $R_b^{CRRM}[t_k]$, which is the data rate achievable from unassigned RRUs.
- Freed service data rate, $R_{b_{ji}}^{fr}[t_k]$, which is referred to RRUs freed by termination of service *j* of VNO *i* in time interval t_k. It should be reminded that this value is considered from the estimation of data rate based on number of freed RRUs. The actual data rate offered to the subscribers may not be the same value.
- On use service data rate, $R_{b_{ji}}^{Us}[t_k]$, which is referred to RRUs exhausted by service *j* of VNO *i* in time interval t_k. Like the freed service data rate, this value also is achieved based on the estimation of data rate of used RRUs.
- Actual freed service data rate, $R_{b_{ji}}^{Afr}[t_k]$, which is referred to actual data rate of service *j* of VNO *i* terminated in this time interval.
- Actual on used service data rate, $R_{b_{ji}}^{AUs}[t_k]$, which is the actual data rate exhausted by ongoing service *j* of VNO *i.*

Using the updated information of network status, the new network capacity in time interval *i*+1 is defined as follows:

$$R_{b\,[\text{Mbps}]}^{CRRM}[t_{k+1}] = R_{b\,[\text{Mbps}]}^{CRRM}[t_k] + \sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \left(R_{b_{ji}\,[\text{Mbps}]}^{fr}[t_k] - R_{b_{ji}\,[\text{Mbps}]}^{Us}[t_k] \right) \tag{4.88}$$

Relatively, the minimum and maximum guaranteed data rate on update as follows:

$$R_{b_{ji}[\text{Mbps}]}^{Min}[t_{k+1}] = R_{b_{ji}[\text{Mbps}]}^{Min}[t_k] + R_{b_{ji}\,[\text{Mbps}]}^{Afr}[t_k] - R_{b_{ji}\,[\text{Mbps}]}^{AUs}[t_k] \tag{4.89}$$

$$R_{b_{ji}[\text{Mbps}]}^{Max}[t_{k+1}] = R_{b_{ji}[\text{Mbps}]}^{Max}[t_k] + R_{b_{ji}\,[\text{Mbps}]}^{Afr}[t_k] - R_{b_{ji}\,[\text{Mbps}]}^{AUs}[t_k] \tag{4.90}$$

The maximum and minimum guaranteed data rates, obviously, cannot be less than zero. A service with zero (or negative) minimum guaranteed data is going to be served in best effort manner and the one with zero maximum guaranteed data is no longer going to be served. Using this technique, the VRRM algorithm repeatedly solves the (4.86) optimisation problem. In the provisioning phase, the system is in initial status and VRRM outputs lead to a set of policies for system runtime. As requests arrive and the network starts to work, system states (e.g., remained RRUs and served data rate) will change. The algorithms update network capacity estimation and then solve again the aforementioned optimisation. Figure 4.18 presents the partial VRRM flowchart in provisioning and runtime of virtual RAN.

Based on this framework, for different cases, the optimisation problem support traffic offloading based on (4.54) is:

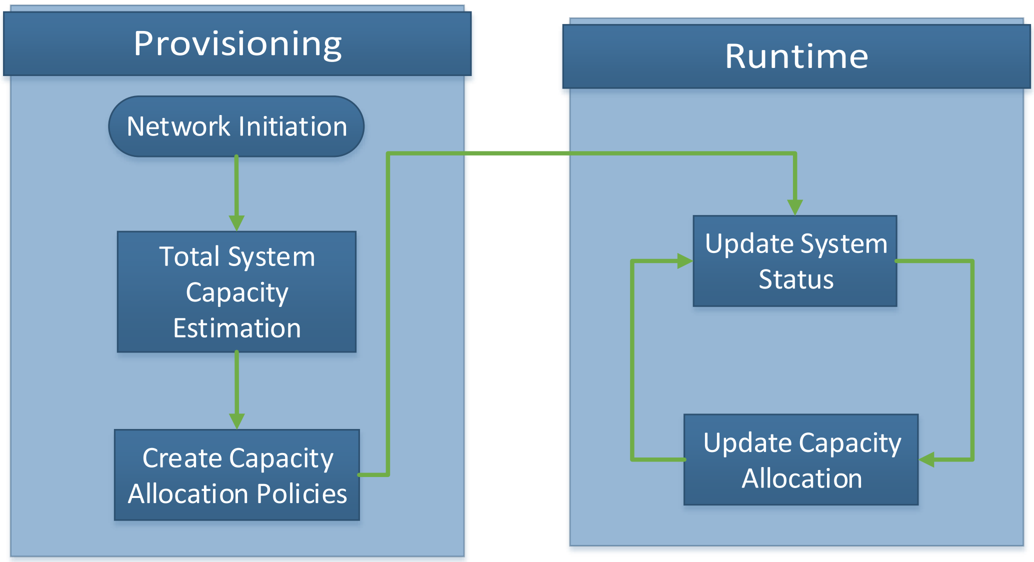


Figure 4.18 – Partial VRRM flowchart.

$$\max_{\mathbf{R}_{\mathrm{b}}^{\mathrm{cell}}[t_k],\mathbf{R}_{\mathrm{b}}^{\mathrm{WLAN}}[t_k],\boldsymbol{R}_b^{f}[t_k]} \Big\{ W^{WLAN} f_{\mathbf{R}_{\mathrm{b}}}^{WLAN}\big(\mathbf{R}_{\mathrm{b}}^{\mathrm{WALN}}[t_k]\big) + W^{cell} f_{\mathbf{R}_{\mathrm{b}}}^{cell}\big(\mathbf{R}_{\mathrm{b}}^{\mathbf{cell}}[t_k]\big) - f_{\mathbf{R}_{\mathrm{b}}}^{f}\big(\mathbf{R}_{\mathrm{b}}^{\mathrm{f}}[t_k]\big)\Big\}$$

$$s.t.: \begin{cases} R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k] = R_{b_{ji}\,[\mathrm{Mbps}]}^{cell}[t_k] + R_{b_{ji}\,[\mathrm{Mbps}]}^{WLAN}[t_k] \\ 1^T \cdot \mathbf{R}_{\mathrm{b}}[t_k] \le R_{b\,[\mathrm{Mbps}]}^{CRRM}[t_k] \\ R_{b_{ji}[\mathrm{Mbps}]}^{Min}[t_k] \le R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k] \\ R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k] \le R_{b_{ji}[\mathrm{Mbps}]}^{Max}[t_k] \\ \dfrac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k]}{W_{ji}^{Srv}} - \dfrac{1}{N_{VNO}N_{srv}} \displaystyle\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \dfrac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k]}{W_{ji}^{Srv}} \le R_{b_{ji}[\mathrm{Mbps}]}^{f}[t_k] \\ \dfrac{-R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k]}{W_{ji}^{Srv}} + \dfrac{1}{N_{VNO}N_{srv}} \displaystyle\sum_{i=1}^{N_{VNO}} \sum_{j=1}^{N_{srv}} \dfrac{R_{b_{ji}\,[\mathrm{Mbps}]}^{Srv}[t_k]}{W_{ji}^{Srv}} \le R_{b_{ji}[\mathrm{Mbps}]}^{f}[t_k] \end{cases} \tag{4.91}$$

For the congestion situation based on (4.57) and (4.58), the optimisation problem is:

$$\max_{\mathbf{R}_{\mathrm{b}}^{\mathrm{cell}}[t_k],\mathbf{R}_{\mathrm{b}}^{\mathrm{WLAN}}[t_k],\mathbf{R}_{\mathrm{b}}^{\mathrm{f}}[t_k],\Delta R_b^{v}[t_k]} \Big\{ W^{WLAN} f_{\mathbf{R}_{\mathrm{b}}}^{WLAN}\big(\mathbf{R}_{\mathrm{b}}^{\mathrm{WLAN}}[t_k]\big) + W^{cell} f_{\mathbf{R}_{\mathrm{b}}}^{cell}\big(\mathbf{R}_{\mathrm{b}}^{\mathrm{cell}}[t_k]\big) - f_{R_b^{v}}^{vi}\big(\Delta \overline{R_b^{v}}\big) - f_{R_b}^{f}\big(\mathbf{R}_{\mathrm{b}}^{\mathrm{f}}[t_k]\big)\Big\} \tag{4.92}$$

$$
s.t:\begin{cases}
R^{Srv}_{b_{ji}\,[\text{Mbps}]}[t_k] = R^{cell}_{b_{ji}\,[\text{Mbps}]}[t_k] + R^{WLAN}_{b_{ji}\,[\text{Mbps}]}[t_k] \\
1^T \cdot \mathbf{R_b}[t_k] \le R^{CRRM}_{b\,[\text{Mbps}]}[t_k] \\
R^{Min}_{b_{ji}[\text{Mbps}]}[t_k] \le R^{Srv}_{b_{ji}\,[\text{Mbps}]}[t_k] \\
R^{Srv}_{b_{ji}\,[\text{Mbps}]}[t_k] \le R^{Max}_{b_{ji}[\text{Mbps}]}[t_k] \\
0 \le \Delta R^{v}_{b_{ji}} \\
W^{v}_{ji}\Delta R^{v}_{b_{ji}\,[\text{Mbps}]}[t_k] - \frac{1}{N_{VNO}N_{srv}}\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}} W^{v}_{ji}\Delta R^{v}_{b_{ji}\,[\text{Mbps}]}[t_k] \le R^{f}_{b_{ji}[\text{Mbps}]}[t_k] \\
-W^{v}_{ji}\Delta R^{v}_{b_{ji}\,[\text{Mbps}]}[t_k] + \frac{1}{N_{VNO}N_{srv}}\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}} W^{v}_{ji}\Delta R^{v}_{b_{ji}\,[\text{Mbps}]}[t_k] \le R^{f}_{b_{ji}[\text{Mbps}]}[t_k]
\end{cases}
$$

where $\Delta\mathbf{R^v_b}[t_k]$ is vector of violation data rates given by:

$$
\Delta\mathbf{R^v_b}[t_k] = \{\Delta R^{v}_{b_{ji}}[t_k] \mid j = 1,2,\dots,N_{srv} \; and \; i = 1,2,\dots,N_{VNO}\} \tag{4.93}
$$

Obviously, the solutions for the VRRM problem better than the partial VRRM can be achieved by considering the whole problem at once and achieving an overall policy.

4.5 Evaluation Metrics

In order to evaluate the performance of the applied RRM algorithms and strategies, multiple evaluation metrics can be used. In fact, satisfaction besides the definition of a good performance for VRRM and VNOs may be very different. VRRM considers all networks, VNOs, and services, thus, a good performance from the VRRM viewpoint is a performance with which the highest level of resource usage, network benefit, and overall satisfaction can be achieved. VNOs, on the other hand, put value on only the level of offered services to them. Therefore, the evaluation metrics can be divided into two key categorises.

The first metrics set, the metrics used for evaluation of the VRRM perspective, are as follows:

- **VRRM Objective function Value –** obviously is the most crucial evolution metric. The goal of models presented in former sections is to maximise the value of objective functions. Therefore, the higher the objective function achieves; the better is the performance of the management algorithm. The general equation for the VRRM objective function is:

$$
\begin{aligned}
f_v\left(R^{Srv}_{b_{ji}\,[\text{Mbps}]}\right) = W^{cell} f^{cell}_v\left(R^{cell}_{b_{ji}\,[\text{Mbps}]}\right) + W^{WLAN} f^{WLAN}_v\left(R^{WLAN}_{b_{ji}\,[\text{Mbps}]}\right) \\
- \Delta\overline{R^{v}_{b}}_{[Mbps]} - \sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}} R^{f}_{b_{ji}[\text{Mbps}]}
\end{aligned} \tag{4.94}
$$

- **Average Constraint Violation –** The management algorithm has met constraints or at least minimises the constraint violations. The average constraint violation, thus, represents how good the VRRM algorithm is meeting the SLA constraints as follows:

$$\Delta\overline{R^{v}_{b}}_{[\mathrm{Mbps}]} = \frac{1}{N_{VNO}N_{srv}}\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}} W^{v}_{ji}\,\Delta R^{v}_{b_{ji}\,[\mathrm{Mbps}]} \tag{4.95}$$

The metrics of the second group, the VNOs perspective, are as follows:

- **Allocated data rate –** the date rate allocated to each service of each VNO is the key metric, based on which the performance of VRRM can be evaluated. Comparing the allocated data rate to minimum and maximum guaranteed ones clearly demonstrates how satisfied the VNO is.

$$R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]} = R^{cell}_{b_{ji}\,[\mathrm{Mbps}]} + R^{WLAN}_{b_{ji}\,[\mathrm{Mbps}]} \tag{4.96}$$

- **Service fairness** – The fairness in serving a service of a VNO is also an evaluation metric. In the ideal case, total fair serving is desired, though it may not be possible to achieve it in practice. A fair data rate allocation is done when the assigned normalised data rate to a service is equal to the average normalised one.

$$R^{f}_{b[\mathrm{Mbps}]} = \left|\frac{R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]}}{W^{Srv}_{ji}W^{v}_{ji}W^{Usg}_{ji}} - \frac{1}{N_{VNO}N_{srv}}\sum_{i=1}^{N_{VNO}}\sum_{j=1}^{N_{srv}}\frac{R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]}}{W^{Srv}_{ji}W^{v}_{ji}W^{Usg}_{ji}}\right| \tag{4.97}$$

- **Service data rate violation** – is a key metric to evaluate the QoS offered to each service of each VNO. The desired situation is when service data rate requests are all met without any violation. The violation of service data rate is given by:

$$\begin{aligned} & R^{Min}_{b_{ji}[\mathrm{Mbps}]} \leq R^{Srv}_{b_{ji}\,[\mathrm{Mbps}]} + \Delta R^{v}_{b_{ji}\,[\mathrm{Mbps}]} \\ & \Delta R^{v}_{b_{ji}\,[\mathrm{Mbps}]} \geq 0 \end{aligned} \tag{4.98}$$

Chapter 5

Models and Algorithms Implementation

This chapter aims at presenting the implementation of the algorithms and the management model. Section 5.1 provides an overview on the implementation of the VRRM Server and Open Air Interface (OAI). Section 5.2 discusses the details and algorithms of the VRRM server and its application. The changes in OAI to support multiple VNOs are described in Section 5.3. In the end, Section 5.4 reviews the assessment procedure of the model implementation.

5.1 Implementation Overview

The concept of virtualisation of radio resources and the comprehensive management models are presented in Chapter 4. This chapter describes the implementation of VRRM at Open Air Interface (OAI) [Open15]. The aim is to demonstrate the download operation of groups of users, belonging to different Virtual Network Operators (VNOs) in the same infrastructure, while providing them different type of services with pre-defined requirements. The goal has persuaded by implementing the VRRM model in a server called as the VRRM server. The implemented model in the server estimates the total network capacity and solve the optimisation problem of VRRM for various network configurations and scenarios. The VRRM server alone is used to obtain theoretical results presented in Chapter 6.

Furthermore, the VRRM server can also receive configurations and scenarios details from the infrastructure and perform management of virtual resources. The used infrastructure emulator for implementing the concept of virtualisation of radio resources is a cloud host environment with OAI servers

OAI wireless technology platform is a flexible platform towards an open LTE ecosystem. The more detailed version of this document can be found in [Open15].. The platform offers an open-source software-based implementation of the LTE system spanning the full protocol stack of 3GPP standard in both E-UTRAN and EPC. It can be used to build and customised an LTE base station and core network on a PC and connect a commercial UEs to test different configurations and network setups and monitor the network and mobile device in real-time. OAI is based on a PC hosted software radio frontend architecture. With OAI, the transceiver functionality is realised via a software radio front end connected to a host computer for processing. OAI is written in standard C for several real-time Linux variants optimised for Intel™ x86 and ARM™ processors and released as free software under the OAI License Model. The software radio access network, OAI, in addition to the VRRM server can be hosted by cloud host, as shown in Figure 5.1. The VRRM server issues policies based on information and statics received from the infrastructure emulator.

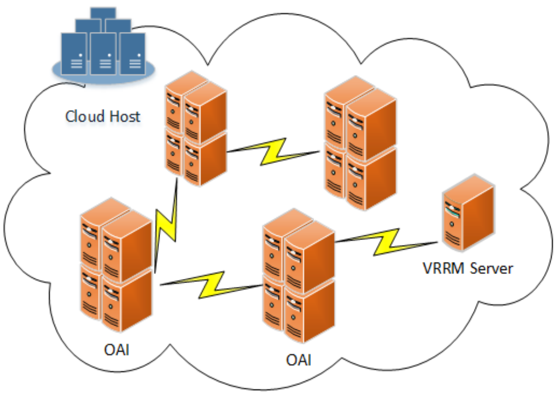


Figure 5.1 – VRRM and OAI servers on the cloud host environment.

These servers are connected through an internal network managed by the cloud provider, where the links can carry information and control signals among servers. Figure 5.2 shows the chosen approach.

The elements and modules coloured by green are the ones that are implemented or modified in this work in order to define a bidirectional interface between OAI and VRRM server to:

- transfer the scenarios and configurations from OAI to VRRM,
- send calculated policies from VRRM to OAI,
- retrieve real-time statics and information of VNOs from OAI for VRRM,
- send updated policies from VRRM to OAI.

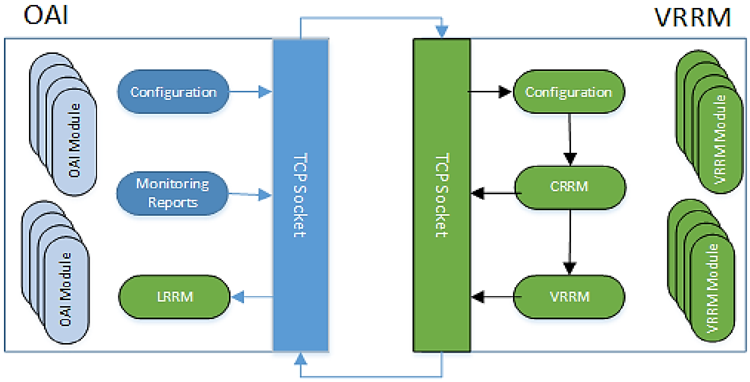


Figure 5.2 – Simplified block diagram of VRRM and OAI interaction.

In addition, the scheduler of OAI requires modifications in order to support the VRRM policies for multi-tenancy. These changes enable OAI to have multiple groups of subscribers, each one representing the subscribers of a VNO. With this approach, it is possible to impose different policies per VNO/group, offering different QoS to MTs according to the VNO/group they belong to, and collecting statistics of operation separately in order to adapt the scheduler to meet the diverse requirements. The following modifications on OAI were defined for the implementation:

- adding group-based statics to eNodeB;
- introducing new set of long-time statics using the already implemented listing concept;
- changing the codes to initialise, fill, and use the aforementioned statics;
- changing the algorithm of MAC scheduler in order to support the groups policies;
- adding support for bidirectional connection and the required protocols;
- changing the codes to add the groups' information into the XML file.

In the following, the implementation is described.

5.2 VRRM Server

5.2.1 Module Structure

Figure 5.3 presents the flowchart of the VRRM server. Before the appearance of the Graphical User Interface (GUI) on the screen, there are multiple initialisations to be done. First, the server makes a connection to the SQL Server [MSQL15]. The database is used to speed up the procedure by not recalculating intermediate steps or results. Considering the VRRM procedure, calculating the PDF of Radio Resource Units (RRUs), Radio Access Technologies (RATs), and the network, are the most time consuming tasks during evaluations and simulations. By saving the intermediate results during the calculation of probability functions, it is possible to significantly decrease the processing time. The details of the database implementation are comprehensively described in Section 5.2.3.

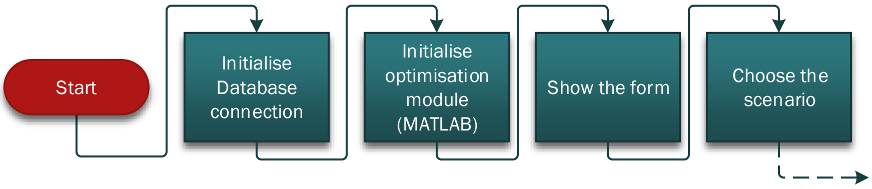


Figure 5.3 – VRRM module flow chart.

After having the database connected to the network, the optimisation module is initialised. The function for optimisation was developed in MATLAB [Math15] and converted to a managed Dynamic Library Link (DLL). The optimisation is done based on the interior-point method using the linprog function of MATLAB [KhCo14]. The initialisation of MATLAB-based modules is more time consuming than other elements, hence, the code is foreseen to initialise the module once at the beginning and to be used multiple times later. After all pre-requisite modules are loaded; the control panel appears, as shown in Figure 5.4.

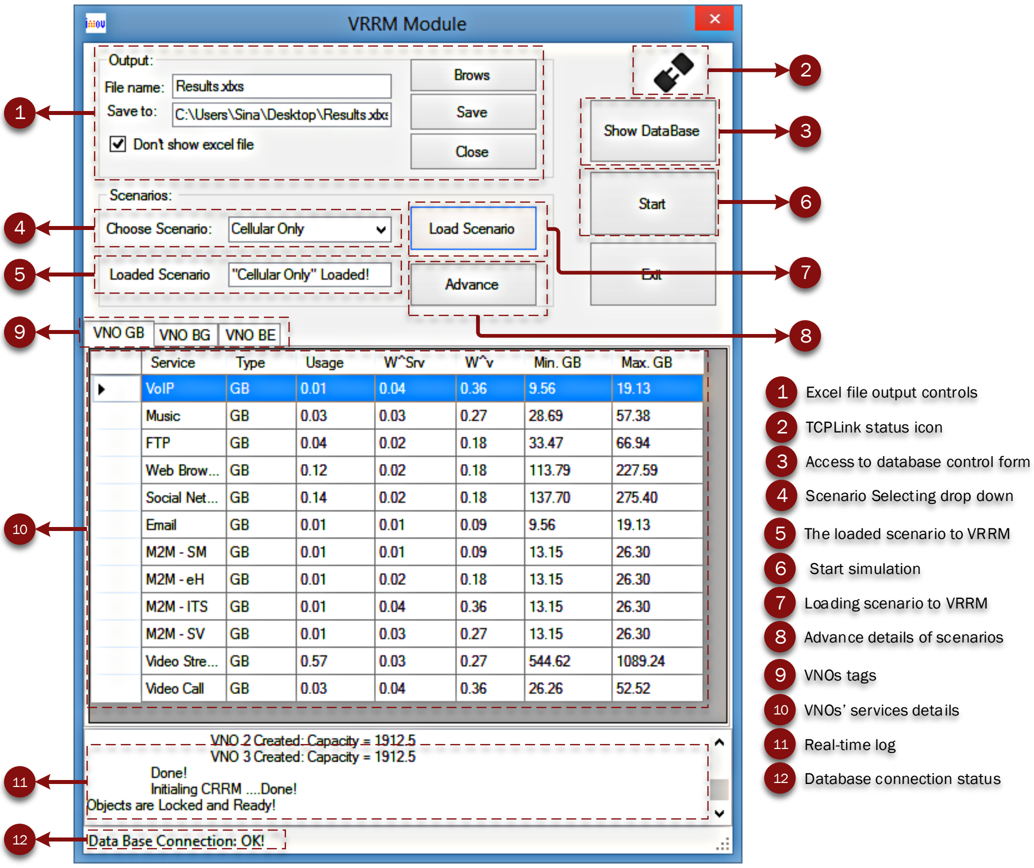


Figure 5.4 – VRRM Server control panel.

At the top, one has the controls for the excel file output, (1). In addition, an icon showing the status of the TCP link (connected or disconnected) was also placed on the upper part of the form, (2). The button “Show Database”, (3), opens a new form, in which the records in the database are shown.

At the middle, there is an area related to scenarios. The user can choose among predefined scenarios through a drop-down list, (4), and check if the scenario is loaded, (5). The buttons “Load Scenario”, (7), and “Advance”, (8), cause the server to read the scenario and to show advanced details of scenarios, respectively. The “Start”, (6), button is also in the middle of the form to start the simulations.

Finally, at the bottom of the form a dynamic grid view control is placed showing the details of the loaded scenarios, (10), including the name of the VNOs, their services, serving and violation weights, minimum and maximum guaranteed data rates, and the type of VNO. One can switch from the details of one VNO to another by using the VNO tags, (9). There are two additional areas in the bottom for real-time logging, (11), and database connection status, (12).

5.2.2 Communication with OAI

To enable the integration of VRRM in OAI, a pre-defined scenario, designated by “OAI Integration”, has been included. By choosing the “OAI Integration” scenario, and clicking the “Load scenario” button, the VRRM server follows the flowchart illustrated in Figure 5.5.

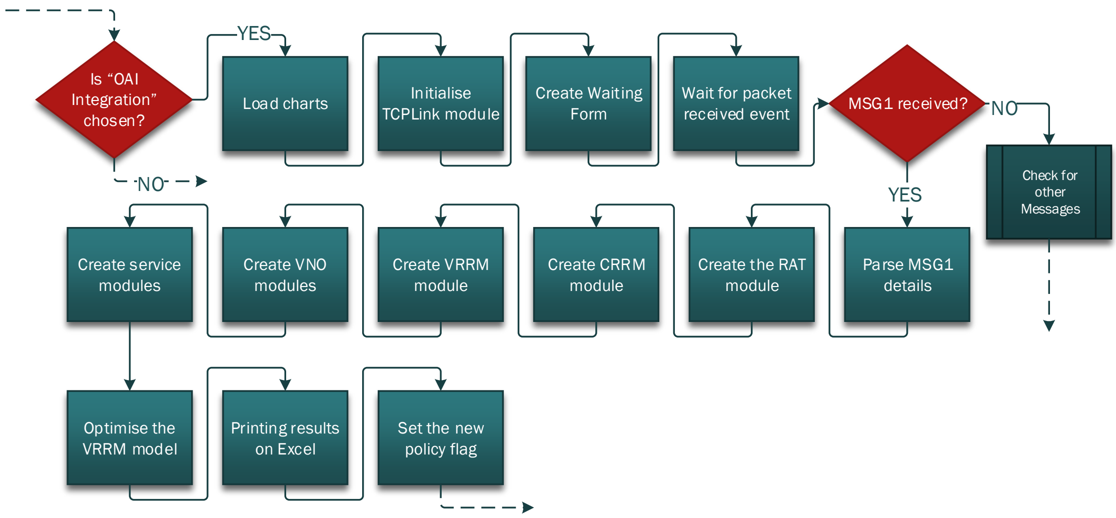


Figure 5.5 – VRRM module flowchart for OAI integration.

After loading the real-time charts, the TCPLink module starts, and the user is asked to choose the right interface for communication with the OAI server. Then, a waiting form is shown to inform the user that the VRRM server is waiting for receiving a packet event from OAI, meaning that the connection has been established. Upon reception of a new message, it is checked for the “Scenario Description” message, MSG#1Table 5.2; then, the message is parsed, and the scenario details are extracted. In the next step, the RAT module that contains the properties of the radio resources is created and initialised.

After initialising the CRRM and VRRM modules, the VNOs based on the received scenario are formed. Finally, the services are added to the VNOs.

In the next stage, the VRRM model optimisation is solved, results are outputted into the excel file, and the new policy flag is set.

Figure 5.6 illustrates the VRRM response to the policy update requests of OAI, MSG#2, Table 5.2. The new set of policies is sent to OAI when the new policy flag is set through MSG#4, Table 5.2. In other cases, the server just responds with MSG#3, Table 5.2, to inform the OAI that there is no update available.

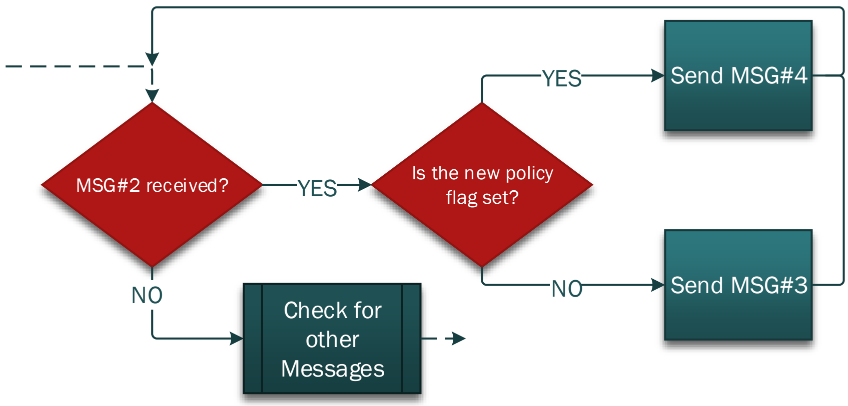


Figure 5.6 – The policy update sequence from VRRM module.

5.2.3 Implementation of the Database

As discussed in Section 5.2.1, for the sake of shorter simulation and development time, the probability functions used in the estimation procedure are save in an SQL-server database. In order to improve the performance of the SQL-Server, the PDFs are saved in binary files and a pointer to a relative file is placed in the SQL table; this technique enables the running time to be reduced to a couple of minutes. By clicking on the "show database" button on the VRRM server control panel marked by (4) in Figure 5.4, the database control panel appears. As presented in Figure 5.8, there is a dropdown box (marked by 1) to choose the target table from one of the following tables:

- GuidDLTB, holding downlink records (general approach [KhCo15]),
- GuidULTB, holding uplink records,
- APP1DLTB, holding downlink records for optimistic approach,
- APP1ULTB, holding uplink records for optimistic approach,
- APP2DLTB, holding downlink records for pessimistic approach,
- APP2ULTB, holding uplink records for pessimistic approach,
- APP3DLTB, holding downlink records for realistic approach,
- APP3ULTB, holding uplink records for realistic approach.

Figure 5.7 presents the relation among these tables.

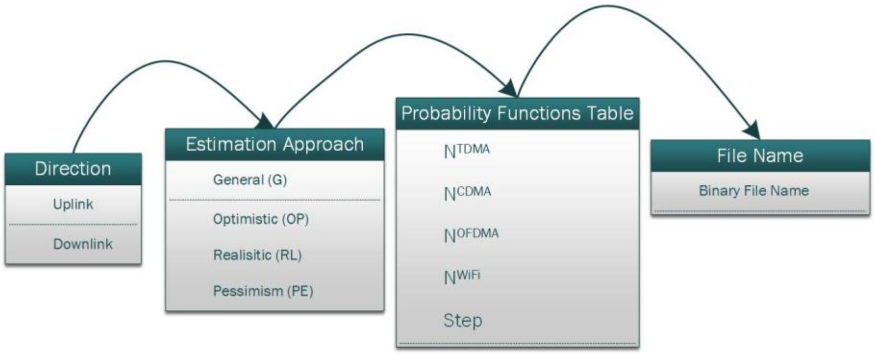


Figure 5.7 – Database tables relations.

By double clicking on any of the records on the table, the selected function in the drop down box, (2), in Figure 5.8 is plotted in a MATLAB figure. Checking the check box "Hold", (3), allows the user to plot multiple graphs on the same figure panel. Although the right path for the binary files is already set but, it is possible to change it by the "Path" button, (5). Regarding the information in the tables, it can be seen that the binary files are identified by 4 parameters, which are propagation exponent (alpha), number of RRUs (i.e., TDMA, CDMA, OFDMA, Wi-Fi), the step in Mbps, and the file name.

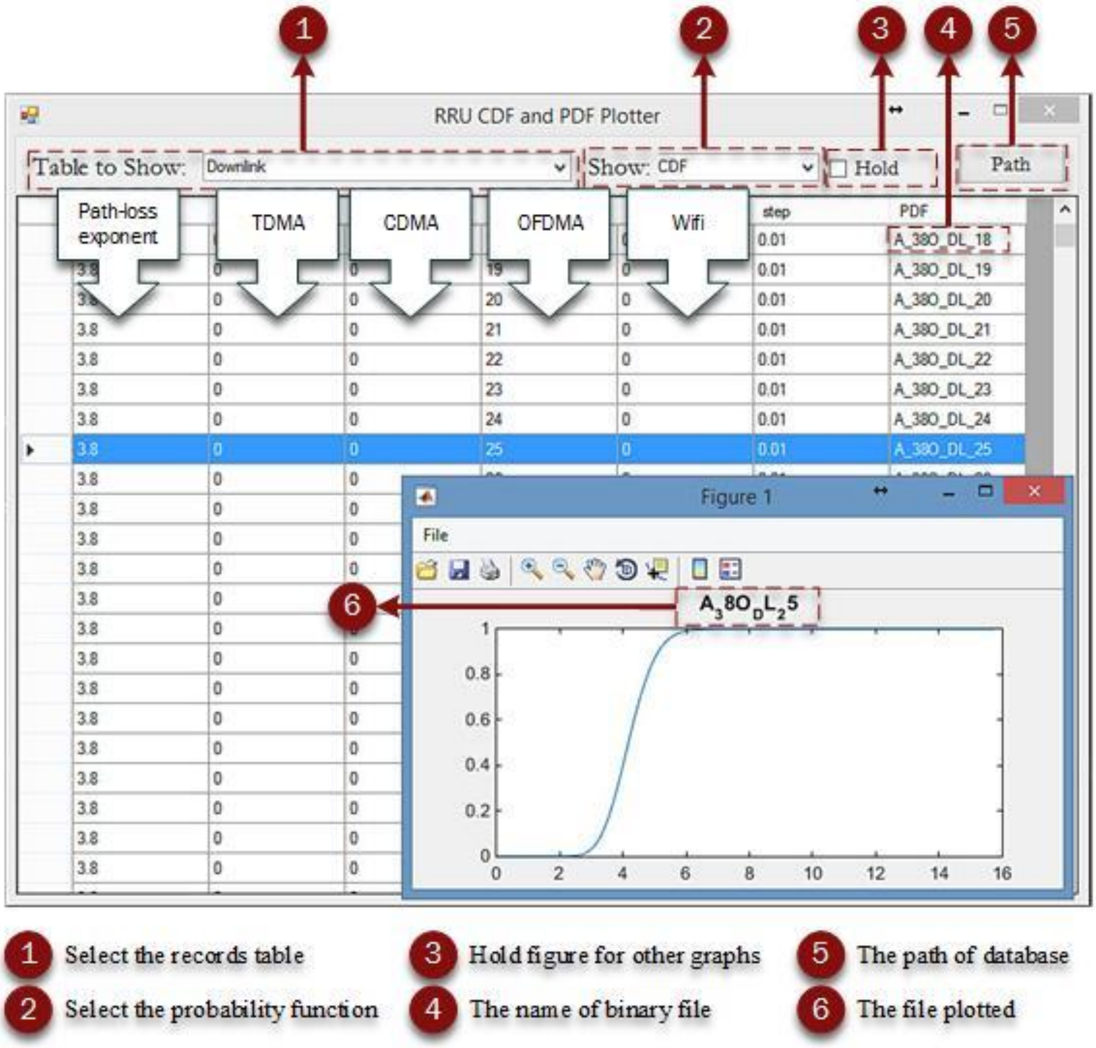


Figure 5.8 – Database control form.

5.3 OAI Extension to Support Multiple VNOs/Groups

5.3.1 Module Structure

In addition to the development of the VRRM server, OAI also needed some changes to support multiple VNOs or groups operating in the same infrastructure. One fundamental requirement was the gathering of statics per group/VNO instead of all UEs.

OAI stores UEs key performance indicators, e.g., total data rates offered to the UEs in downlinks. Based on the codes that gather these statistics, a new set of codes has been added to gather the same statistics for each group. The group-based statistics that have been added to OAI are:

- Group MAC PDUs (Protocol Data Units) per sub-frame (`dlsch_group_pdus_tx`),
- Group transferred size per sub-frame (`dlsch_group_bytes_tx`),
- Group download data rate per sub-frame (`dlsch_group_bitrate`),
- Group MCS (Modulation and Coding Scheme) per sub-frame (`dlsch_group_mcs`),
- Group active UEs per sub-frame (`dlsch_group_active_ue`).

These statistics are used in the scheduler for enforcing VRRM's policies and the management of radio resources. However, the reports for the VRRM server contain statistics over an observation window, which may be as long as the decision window discussed in VRRM modelling [KhCo15]. In order to calculate this long-term view, the statistics obtained in each sub-frame are collected. This task is fulfilled by using the already implemented concept of listing in OAI. The following set of list variable was added:

- List for group bitrate (`dlsch_listg_bitrate`),
- List for group Modulation and Coding Scheme (MCS) (`dlsch_listg_mcs`),
- List for group active UEs (`dlsch_listg_active_ue`).

First, these variables have to be initialised during the eNodeB MAC initialisation phase. Figure 5.9 shows the flowchart of collecting, calculating, and reporting the long-term statistics.

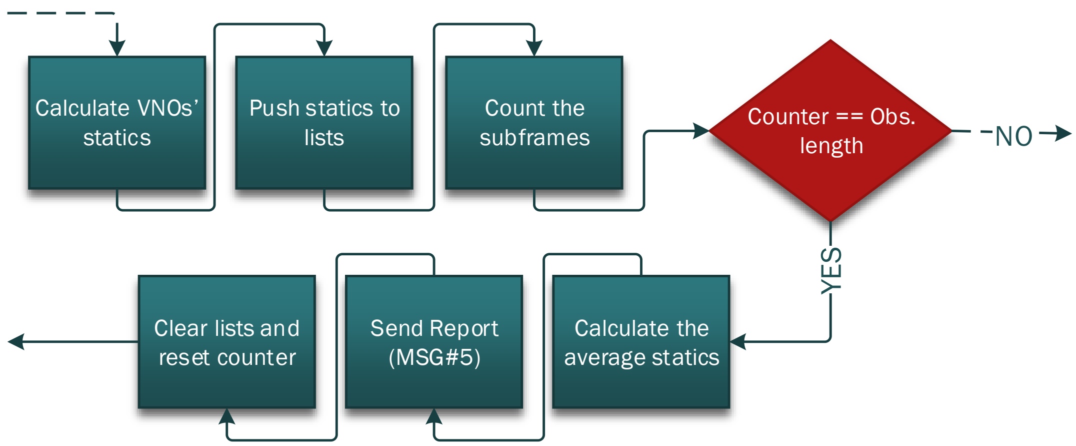


Figure 5.9 – Flow chart of calculating and reporting long-term statics.

As shown, as the VNOs statistics become available in each sub-frame, they are pushed into relative lists. At the end of each sub-frame, a counter is placed to count the number of sub-frames. When the value of counter becomes equal to the length of the observation window, the average of the aforementioned statistics pushed into lists are calculated for the observation interval and reported to the VRRM server using the “Real-Time Report” message, MSG#5, Table 5.2. Finally, lists are cleared and the sub-frame counter is set to zero.

5.3.2 Changing the algorithm of the MAC scheduler in order to support group policies

The key methods implemented in the downlink scheduler of OAI are storing the UEs’ DL-SCH (Downlink – Shared Chanel) buffer, calculating the number of required RBs by each MT, pre-allocation, and allocation of the RBs to the UEs, as shown in Figure 5.10. In what follows, the changes introduced to these functions to support multiple VNOs or groups are described.

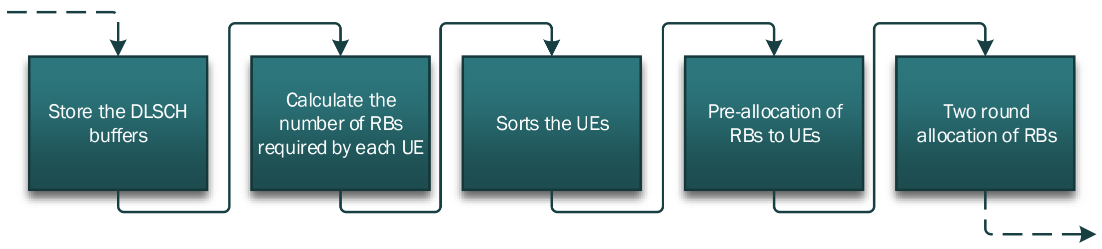


Figure 5.10 – Flowchart of downlink scheduler.

First, some modification was introduced in the Sorting function of UEs. Originally, the UEs were sorted based on the following aspects:

- *HARQ round*: the users that are in their second round of HARQ are given higher priority,
- *Bytes in the buffer*: the users that have more bytes to receive are moved to the beginning of the user list.
- *Maximum time of SDU creation*: the users with more tolerance of delay pushed to the back of list opening up space for other users.
- *UE ID*: in the case all the other criteria match, the UE ID is the last sorting criterion.

This function was changed in order to put the users of VNOs with a higher priority on the top of the list before checking the other criteria. Based on the policies received from VRRM, each VNO receives a portion of available RBs. The goal is to start the user list with the users of the VNO with the biggest share of RBs. The flowchart of sorting UEs is depicted in

Figure 5.11. The changes is presented in blue, while the other boxes in green represent the components that were not modified.

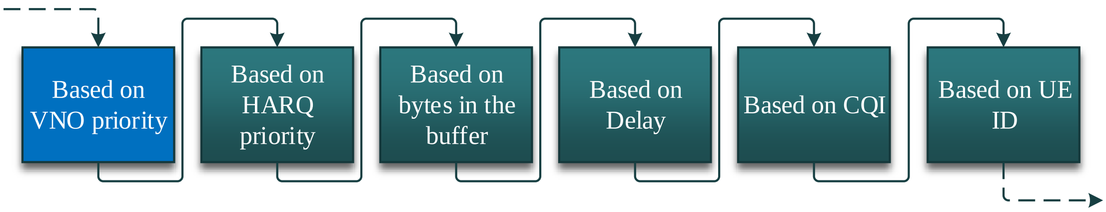


Figure 5.11 – MT sorting algorithms.

In the next step, the pre-allocation sequence was modified to support the possibility of having multiple VNOs (or groups) with different priorities. The pre-allocation initially was implemented as shown in Figure 5.12. When there are active UEs, the required RBs to fully serve each of them is calculated. Then, the possibility of allocating at least a minimum number of RB units (i.e., 2 RBs in the implementation of OAI) to all the active UEs is checked. The minimum RB units are pre-allocated to all active UEs when there are not enough resources. Otherwise, the average RB per UE obtained in the last step is pre-allocated. No pre-allocation is done when there are no active UEs.

Based on the sorted UE list, in the next step, each MT that has to receive the control information is granted with the required RBs. Regarding data traffic, all UEs are pre-allocated the average RBs, unless the required number of RBs for a user is smaller than the average. In this situation, a UE is pre-assigned only the required RBs.

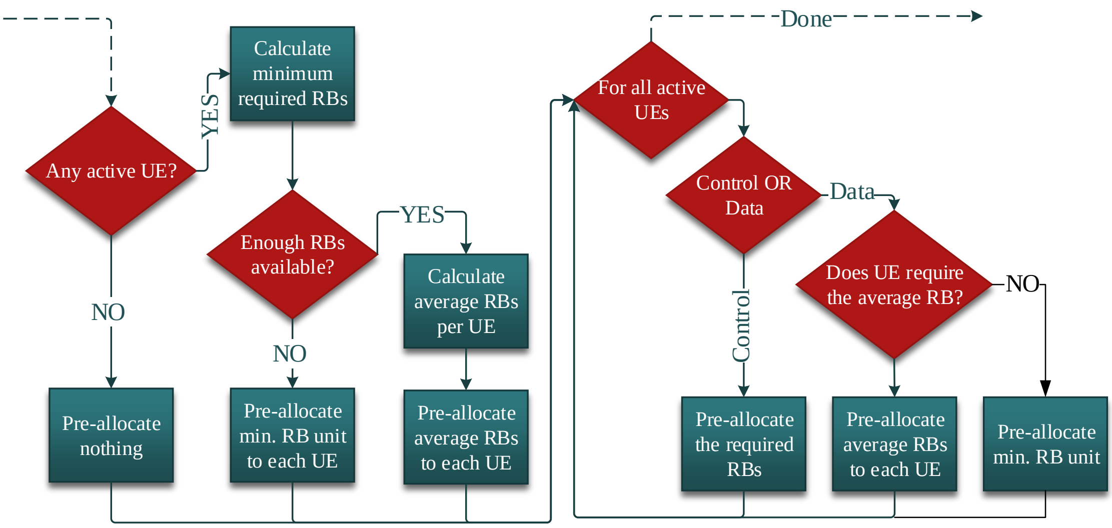


Figure 5.12 – Pre-allocation procedure in OAI.

In order to add multi-tenancy, the pre-allocation procedure is changed by calling a new function that maps the assigned data rate of each VNO to its share of available resources. For each frame, the sharing of available resources is updated, based on the policies received from VRRM, minimum data rate for each VNO, and key performance indicators, namely, the average MCS per VNO. In the first step, the number of RBs per VNO is calculated by dividing the minimum data rate assigned by average MCS achieved. The adaptation to the maximum number of RBs available for the eNode B is considered in a second step, in order to avoid overbooking.

Then, the same pre-allocation procedure is done, but this time only per VNO. In other words, the average RBs per VNO is calculated and used instead of the average number of RB in the former approach. The flowchart of the pre-allocation procedure with the new algorithms is illustrated in Figure 5.13, showing how the changes are implemented. It is worth noting that the primary goal was to implement the support for virtualisation of radio resources with minimum changes in OAI.

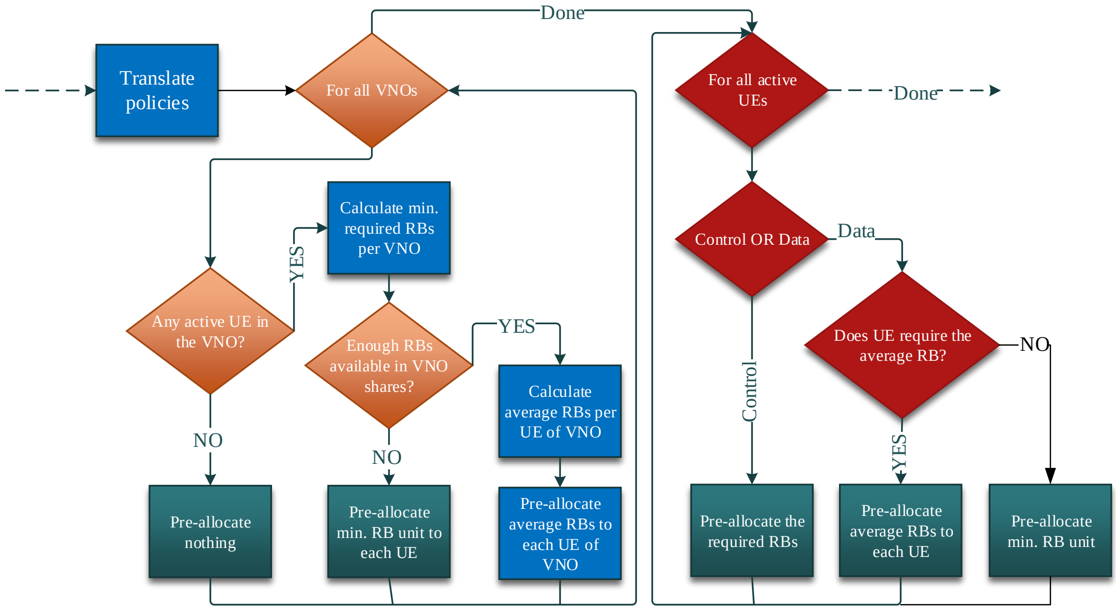


Figure 5.13 – Changed pre-allocation procedure.

5.3.3 Changing the codes to add groups information into the XML file

To add the group information into the XML file a new set of parameters identified by <PROTOCOL><MAC> tags was added. An example of the XML code for the configuration of OAI-VRRM integration is presented in Figure 5.14

In the <VRRM> section the parameters for the TCP/IP communication are defined, namely, the IP address of the machine in which VRRM is running (VRRM_IP) and the TCP port (VRRM_PORT). The VRRM_LINK parameter allows running OAI without the connection to VRRM, but maintaining the virtualisation of radio resources for serving different VNOs.

In the following lines, the number of groups/VNOs (NUM_GROUPS) and the quantity of frames that are considering for each report sent to VRRM (OBS_WIN_LENGTH) are set.

Finally, in the <GROUPS> section, the static information for each group is stated. Three type of groups are considered: BE, BE with minimum guaranteed data rate, and Guaranteed. The differentiation among the types of groups is made according to the values in Gbit/s assigned to GROUP_MIN_GB and GROUP_MAX_GB, Table 5.1. SERVICE_WEIGHT and VIOLATION_WEIGHT are parameters used for the optimisation performed by the VRRM module, allowing to distinguish among VNOs of the same type.

Table 5.1 – Group type settings.

Group Type	GROUP_MIN_GB [GB]	GROUP_MAX_GB [GB]
BE	0	0
BE with min Guaranteed	>0	0
Guaranteed	>0	>0

```
<PROTOCOL>¶
        <MAC>¶
          <!--Parameters·definition·for·OAI-VRRM·communication-->¶
          <VRRM>¶
          <VRRM_LINK>1</VRRM_LINK>¶
          <VRRM_IP>10.154.3.22</VRRM_IP>¶
          <VRRM_PORT>5000</VRRM_PORT>¶
  </VRRM>¶
          <!--Number·of·groups·(VNOs)·requesting·service-->¶
  <NUM_GROUPS>2</NUM_GROUPS>¶
          <!--Statistics·report·interval·(in·number·of·frames)·from·OAI·to·VRRM-->¶
  <OBS_WIN_LENGTH>20</OBS_WIN_LENGTH>¶
          <!—Groups·static·information-->¶
  <GROUPS>¶
          <SOURCE_ID>0</SOURCE_ID>¶
          <GROUP_MIN_GB>0</GROUP_MIN_GB>¶
        <GROUP_MAX_GB>0</GROUP_MAX_GB>¶
        <SERVING_WEIGHT>0.04</SERVING_WEIGHT>¶
        <VIOLATION_WEIGHT>0.36</VIOLATION_WEIGHT>¶
  </GROUPS>¶
  <GROUPS>¶
          <SOURCE_ID>1</SOURCE_ID>¶
          <GROUP_MIN_GB>3</GROUP_MIN_GB>¶
          <GROUP_MAX_GB>5</GROUP_MAX_GB>¶
          <SERVING_WEIGHT>0.06</SERVING_WEIGHT>¶
          <VIOLATION_WEIGHT>0.54</VIOLATION_WEIGHT>¶
          </GROUPS>¶
          </MAC>¶
</PROTOCOL>¶
```

Figure 5.14 – Example of XML code for configuration of OAI-VRRM communication and groups configuration.

5.3.4 Adding support for bidirectional connection and the required protocols

The communication protocol implemented between OAI and the VRRM module has the same structure for all messages, as shown in Figure 5.15.

- Header and trailer: are placed to determine the beginning and ending of the packets,
- Message number: defines the structure of the data field
- Packet size: 2-byte long field indicates the size of the packet.

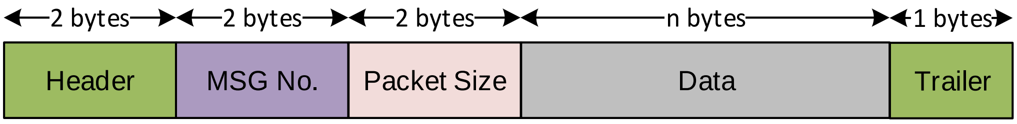


Figure 5.15 – Packet structure.

Table 5.2 – Messages Summary.

Message	Message Designation	Communication	Comments
MSG#1	Scenario Description	OAI → VRRM	Contains the scenarios parameters.
MSG#2	Update Request	OAI → VRRM	Request the policies update
MSG#3	No Updated Available	VRRM → OAI	Policies are not ready to be sent
MSG#4	Policies Update	VRRM → OAI	Contains the updated policies
MSG#5	Real-time Report	OAI → VRRM	Contains the operational report

- MSG1. Scenario Description

This message is used to send the description of the scenarios from OAI to VRRM module. As it is shown in Figure 5.16, it has the following fields:

- Ncc: number of carrier components (1 byte),
- N_{RBi}: number of available Radio Blocks (RBs) in carrier component *i* (1 byte),
- N_{UE}: number of total UEs (1 byte),
- N_{VNO}: number of VNOs (1 byte),
- *VNO info*: 9-byte long field containing the information for each of VNOs. It has the following subfields,
  - VNO ID (1 byte),
  - minimum Guaranteed (2 bytes),
  - maximum Guaranteed (2 bytes),
  - serving weight (2 bytes),
  - violation weight (2 bytes).

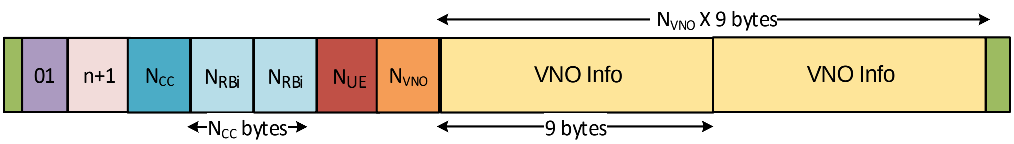


Figure 5.16 – MSG# 1 packet structure.

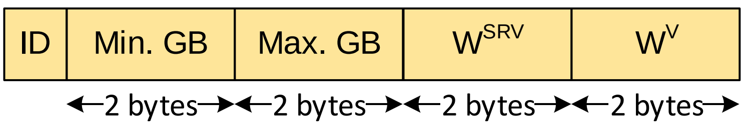


Figure 5.17 – VNO info field details.

- MSG2. Update Request

In order to receive the updated policies, OAI sends an update request using MSG#2. This message is very simple and only has a 2-byte long request number field. Figure 5.18 depicts the structure of MSG#2 packet.

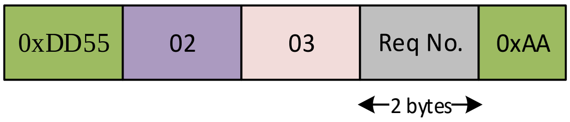


Figure 5.18 – MSG# 2 packet structure.

- MSG3. No Update Available

VRRM module receiving the MSG#2, replays with MSG#3 to inform the OAI that there is no update available. According to Figure 5.10, MSG#3 has very simple structure with no data field.

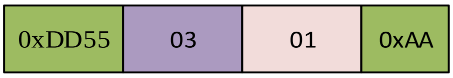


Figure 5.19 – MSG#3 packet structure.

- MSG4. Policies Update

VRRM responds with MSG#4 when there are new policies available. Figure 5.20 illustrates the fields of MSG#4 packet, which are:

- Req No.: the number of update request, to which this packet is transmitting,
- R_b^{Srv}: the allocated data rate to each of the VNOs.

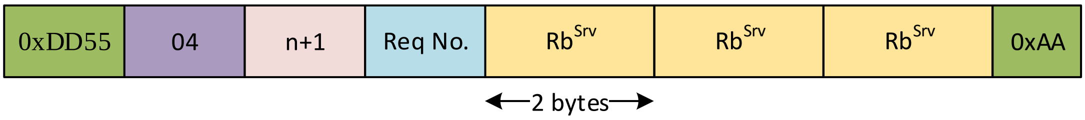


Figure 5.20 – MSG#4 packet structure.

- MSG5. Real-time Report

As the OAI works, it keeps updating the VRRM modules by means of set of real time measurements. These measurements are collected during the observational period and they are exchanged using MSG#5. Figure 5.21 illustrates the structure of MSG#5 packets. The fields are:

- Update number: a two-byte long sequential number,
- $R_b[CC_i][VNO_j]$: the average data rate of the VNO *j* on carrier component *i*,
- $MCS[CC_i][VNO_j]$: the average MCS of the VNO *j* on carrier component *i*
- $R_b[CC_i][VNO_j]$: the average number active UEs of the VNO *j* on carrier component *i.*

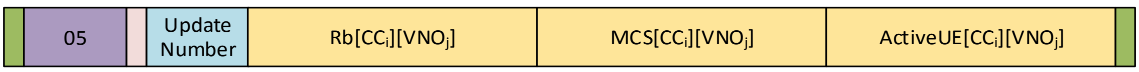


Figure 5.21 – MSG#5 packet structure.

5.4 Assessment of Model Implementation

Two very simple scenarios were considered to check the implementation of the model. In these scenarios, there is only a best effort VNO, which is allocated one third of resources (i.e., equivalent of RAN sharing). This VNO is offering only two services of VoIP and video steaming in the best effort manner. In the first step, the serving weight of these two service is assumed to be 0.5. It is expected that both services are served by equal data rates. However, since they have different average data rates, their share from the cellular and WLAN networks is not going to be the same. The output is presented in Table 5.3: both services have the same share of the serving data rates, but the biggest share of capacity from the cellular network is allocated to VoIP, since it has the lower average data rate, whether video streaming is the dominant service in the WLAN.

In the next step, the serving weight of VoIP is increased to 0.8. Consequently, the serving weight of video streaming is decreased to 0.2. According to the outcomes, presented in Table 5.3, as expected, the total served data rate for VoIP is considerably increased. The data rate allocated to VoIP is 4 times higher than to video streaming, nevertheless, the latter is still the dominant service in WLAN.

Table 5.3 – The data rate allocation in assessment scenarios.

Service	Scenario 1			Scenario 2		
	$R_{b[\mathrm{Mbps}]}^{cell}$	$R_{b[\mathrm{Gbps}]}^{WLAN}$	$R_{b[\mathrm{Gbps}]}^{Srv}$	$R_{b[\mathrm{Mbps}]}^{cell}$	$R_{b[\mathrm{Gbps}]}^{WLAN}$	$R_{b[\mathrm{Gbps}]}^{Srv}$
VoIP	589.40	3.78	4.36	592.07	6.39	6.97
Video Streaming	4.46	4.35	4.36	1.78	1.74	1.74

In order to assess the implemented traffic profiles, either the predefined or the customised ones, a VNO without any guaranteed data rate (i.e., a best effort VNO) with only one subscriber was considered. The traffic profiles can be categorised into Constant Bit Rate (CBR) and Service-based profiles. The CBR profiles are:

- Small packet CBR(SCBR): 32-byte packets with mean packet inter-arrival time of 20ms,
- Medium packet CBR (MCBR): 64-byte packets with mean packet inter-arrival time of 20ms,
- Big packet CBR (BCBR); 128-byte packets and mean packet inter-arrival time of 20ms,
- Full buffer: the buffers always have unlimited amount of packet to transmit.

The network throughput when the subscriber is doing one of the CBR profiles (i.e., full buffer, BCBR, MCBR, and SCBR) is presented in Figure 5.22. The highest average data rate belongs to the full buffer traffic profile.

In addition to CBR profile, four more traffic profile based on the common service in the network is considered. These services are video streaming, file sharing, web browsing, and VoIP. Figure 5.23 presents the average data rate for these traffic profiles during observation interval. According to the figure, the average data rate of FTP profile is about 1 Mbps while the VoIP has the data rate with less than 60 kbps.

In the next step, the MTs of the aforementioned VNOs are increased from 1 to 15 (i.e., the maximum number of MTs supported by OAI) with full buffer traffic profile in downlink. Figure 5.24 presents the maximum network data rate as a function of the number MTs in both FDD and TDD modes: network data rate increases until there are at least 5 MTs in the network; by increasing more the MTs, the data rate decreases to 2.8 Mbps due to increase of signalling. When the VNO has 5 MTs, the data rate of the network reaches the maximum of 7.15 Mbps. These results are compatible with the theoretical estimations, which are presented in Section 6.

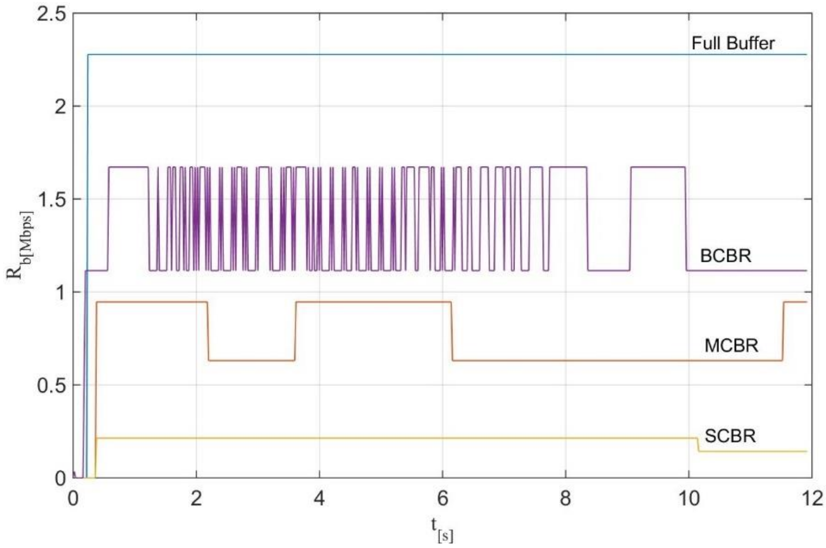


Figure 5.22 – OAI built-in traffic profiles in TDD mode.

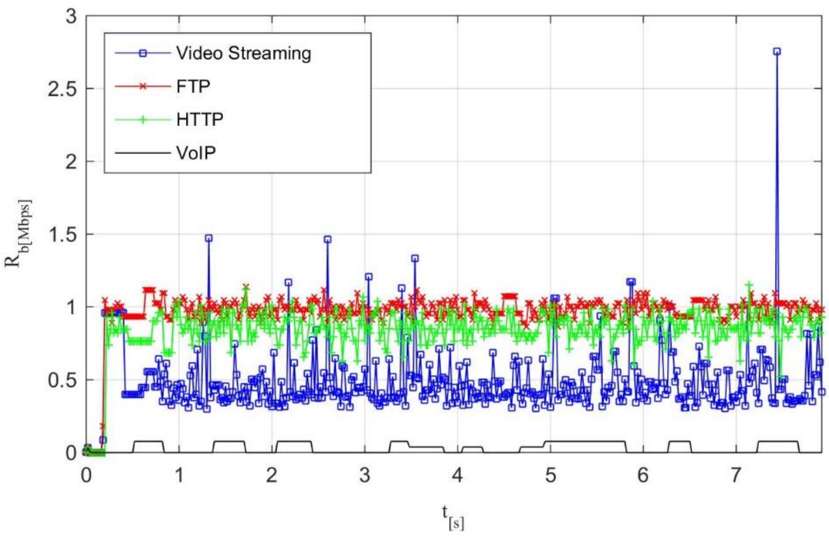


Figure 5.23 – Customised traffic profiles.

Figure 5.25 presents the number of active MTs, the MTs that are allocated resources and have traffic in downlink. It is obvious that the network operating in FDD or TDD modes can serve up to 5 MTs with full buffer traffic. It is worth noting that the number of active MTs is collected during the observation window (i.e., 20 sub-frames), their average being considered.

Finally, the average data rate allocated to each of the active MTs is plotted in Figure 5.26. The maximum data rate per MT reaches the maximum when there are 3 MTs in the network. However, the performance of both TDD and FDD modes in saturation, i.e., having more than 5 MTs, is almost the same.

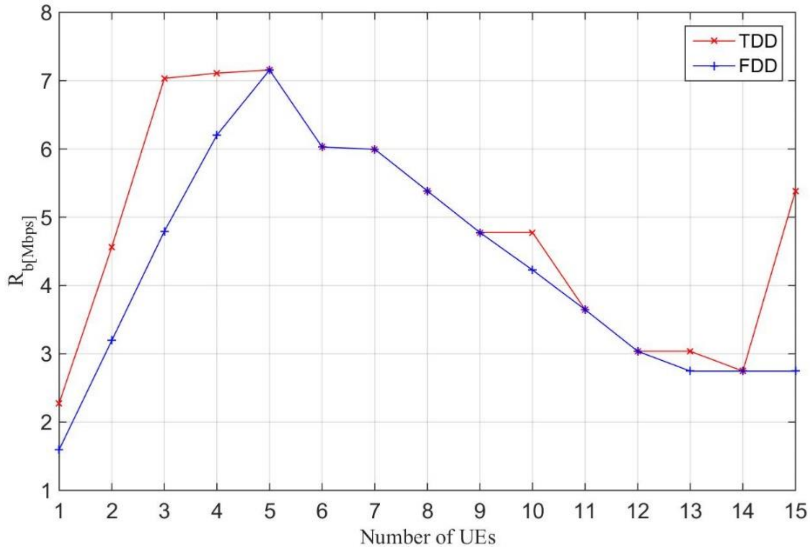


Figure 5.24 – The total network throughput.

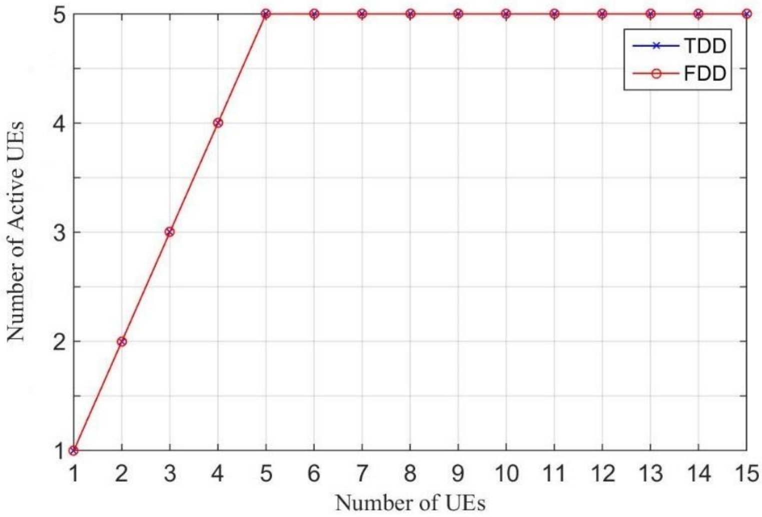


Figure 5.25 – Number of active users.

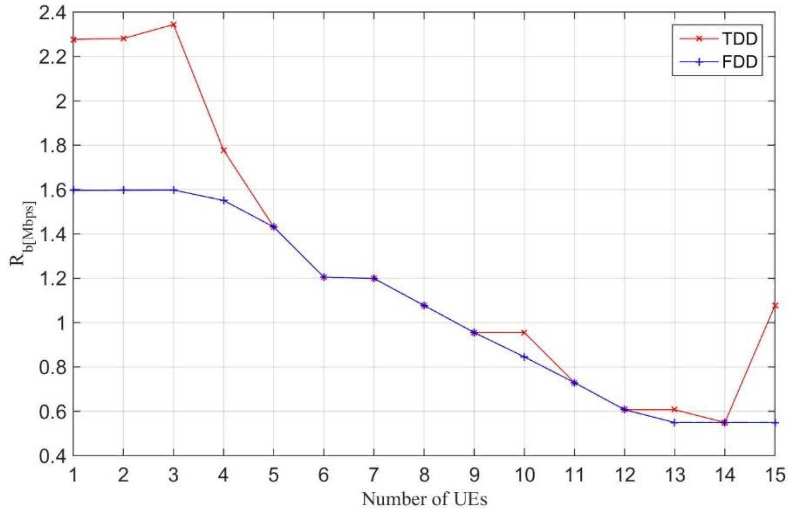


Figure 5.26 – Allocate data rate per active MT.

Chapter 6

Scenarios and Theoretical Results

This chapter describes the used reference scenarios in addition to numeric results achieved by using the developed model in this thesis. Section 6.1 presents the reference scenario, followed by the scenarios' road map in Section 6.2. The numeric results are presented in Section 5.4 to Section 6.11.

6.1 Reference Scenarios – Urban Hotspot

An urban hotspot scenario is chosen as the reference scenario to evaluate the performance of the proposed model for management of virtual radio resources. In addition, variations of this scenario (e.g., different number of subscribers) are also considered to study various aspect of the model. The key parameters of the reference scenario are the cell layout, subscribers, services, and VNOs.

It is assumed that the deployed RRHs are offering full coverage of the cellular networks. These RRHs are capable of supporting multiple RATs, which are OFDMA (based on LTE-Advance), CDMA (based on UMTS), and FDMA/TDMA (based on GSM). Although their flexibility enables various cell layouts for these RATs, the considered layout, as illustrated in Figure 6.1, is as follows:

- The OFDMA cells with a radius of 0.4 km are the smallest ones. Based on the 100 MHz LTE-Advanced bandwidth, each cell has 500 RRUs, which can be assigned to traffic bearers [DaPS11].
- The configuration of CDMA cells is chosen according to UMTS (HSPA+) working at 2.1 GHz. Each cell with the radius of 1.2 km has 3 carriers, each carrier with 16 codes. Only 45 codes out of all 48 codes in each cell can be assigned to users' traffic [SeTB11].
- The FDMA/TDMA cells are the biggest ones, with radius of 1.6 km. Based on GSM900, each cell has 10 frequency channels, each one with 8 time-slots. It is assumed that 75 time-slots out of a total 80 available ones in each cell can be used for users' traffic [Saut10].

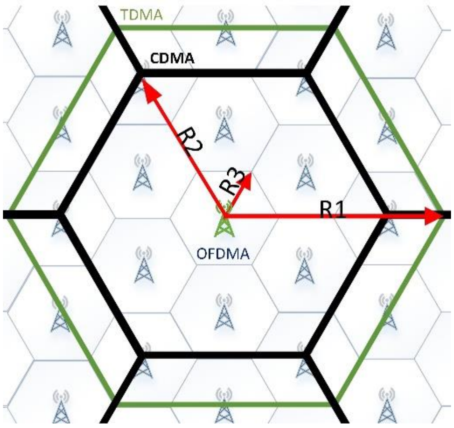


Figure 6.1 – Network Cell Layout for reference scenario (R1=1.6 km, R2=1.2 km, R3=0.4 km)

In addition to cellular networks, Wi-Fi (OFDM) coverage is provided by means of the IEEE802.11ac standard APs, configured to work in an 80 MHz channel bandwidth. It is assumed that each access point covers a cell with a radius 80 m, and that they are facilitated with beamforming and MU-MIMO to support up to 8 spatial streaming channels. Due to European Union rules and regulations, there are only 5 available channels for 80 MHz APs [BeKM13, Cisc12]. In contrast to cellular RATs, Wi-Fi uses the same set of links for up- and download streams. To achieve coherency among the various RATs, the total throughput of APs is equally divided between up- and downlinks. Therefore, in Table 6.1, where coverage information is summarised, the number of RRUs in each Wi-Fi cell is indicated as half of the total number of available channels, together with the maximum data rate for each RAT in downlink.

It is assumed that the APs are only deployed at OFDM BSs, and that there is no full coverage (which is referred to as partial coverage), but in this case, the entire Wi-Fi capacity can be used for traffic offloading. Regarding the number APs, a derivation of this reference scenario is also considered, in which full coverage of Wi-Fi is considered, but then only 5% of the provided capacity can be used for mobile subscribers and the rest belongs to private fixed users (which is referred to as full coverage).

Table 6.1 – Different RAT cell radius and capacity (based on [Cisc12]).

RAT	Number Cells	Cell Radius [km]	System	Downlink			
				Number of RRU/Cell	$N_{RRU}^{RAT_i}$	$R_{b_{RAT_i}}^{max}$ [Mbps]	$R_{b_{tot}}^{RAT_i}$ [Gbps]
OFDMA	16	0.4	LTE	500	8000	0.7	5.47
CDMA	~1.7	1.2	UMTS	45	80	43.0	3.36
TDMA	1	1.6	GSM	75	75	0.05833	4.37
OFDM	16	0.08	Wi-Fi	2.5	40	1300.0	50.78

Subscribers' terminals can be smartphones, tablets, laptops, and M2M devices. According to [ERIC13] and [Cisc13b], the average subscribers' traffic is related to the type of terminal they use, as presented in Table 6.2. By using this information, the VNOs' contracted capacity can be estimated according to (4.28). In this reference scenario, it is assumed that all MTs are smartphones, while in another variation, other mixtures of user terminals are considered.

Table 6.2 – Average Mobile Network Connection Speed (extracted from [Cisc13a])

Terminal	Average Connection Speed [Mbps]
Smartphones	6.375
Tablets	11.39
Laptops	14.0
M2M terminals	1.0

The service set in this scenario is based on Figure 6.2. The volume share of services, which is presented in the outer circle of this figure, is the percentage of the specific service traffic volume from the total operator's traffic. The service penetration, however, is the percentage of active subscribers using that service. This chart clearly shows that video streaming holds the highest volume, while VoIP is the most requested service. These charts may be changed for different types of terminals mixture (and consequently service profile).

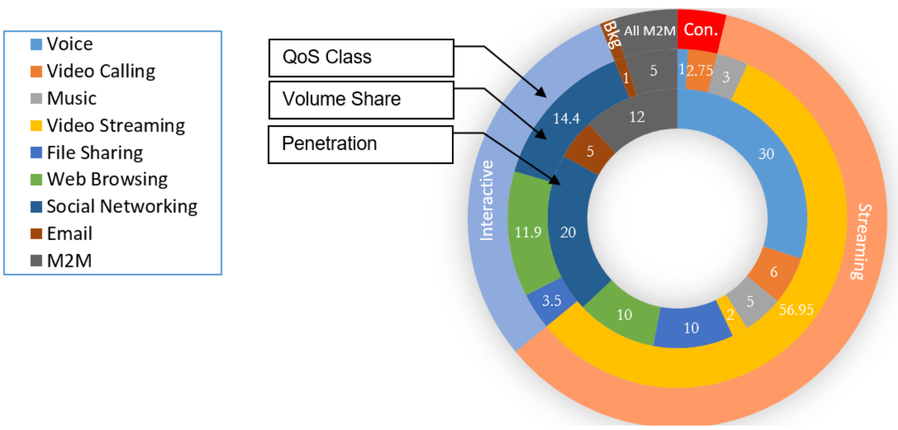


Figure 6.2 – Various service volume share and penetration in reference scenario.

Serving weights for VNOs' services are chosen based on the service class. Conversational services have the highest weight and best effort (also known as background) services have the lowest one. The services weights in addition to the guaranteed data rates for each service are summarised in Table 6.3.

Table 6.3 – Service mixture, their weights, and guaranteed capacities.

Service		W_{ji}^{Srv}	W_{ji}^{v}	VNO GB		VNO BG		VNO BE	
				$R_{b[\text{Mbps}]}^{Min}$	$R_{b[\text{Mbps}]}^{Max}$	$R_{b[\text{Mbps}]}^{Min}$	$R_{b[\text{Mbps}]}^{Max}$	$R_{b[\text{Mbps}]}^{Min}$	$R_{b[\text{Mbps}]}^{Max}$
VoIP		0.04	0.36	9.56	19.12	4.78	N/A	N/A	N/A
Music		0.03	0.27	28.68	57.37	14.34	N/A	N/A	N/A
File Sharing		0.02	0.18	33.47	66.94	16.73	N/A	N/A	N/A
Web Browsing		0.02	0.18	113.79	227.58	56.89	N/A	N/A	N/A
Social Networking		0.02	0.18	137.70	275.40	85.58	N/A	N/A	N/A
Email		0.01	0.09	9.56	19.12	4.78	N/A	N/A	N/A
M2M	Smart Metres	0.01	0.09	13.15	26.30	6.57	N/A	N/A	N/A
	e-Health	0.02	0.18	13.15	26.30	6.57	N/A	N/A	N/A
	ITS	0.04	0.36	13.15	26.30	6.57	N/A	N/A	N/A
	Surveillance	0.03	0.27	13.15	26.30	6.57	N/A	N/A	N/A
Video	Calling	0.04	0.36	26.26	52.52	13.13	N/A	N/A	N/A
	Streaming	0.03	0.27	554.62	1089.24	272.31	N/A	N/A	N/A

VNOs can also be categorised into "Multi-service" or "Mono-service", based on the services they offer to their subscribers: the former is a VNO that offers a variety of services (e.g. VoIP, Video, and FTP), while the latter only offers one or a small set of services, such as machine to machine communication. The VNOs assumed in this scenario are multi-service ones. Table 6.4 present more details on these services and their QoS requirements.

Table 6.4 – Services characteristics for multi-service VNOs.

Service		Service Class	Max Data Rate [kbps]			Duration [s]	Size[kB]
			Min.	Average	Max.		
Voice		Conversational	5.3	12.2	64	60	–
Music		Streaming	16	64	160	90	–
File Sharing		Interactive	384	1024	–	–	2042
Web Browsing		Interactive	30.5	500	–	–	180
Social Networking		Interactive	24	384	–	–	45
Email		Background	10	100	–	–	300
M2M	Smart Meters	Background	–	200	–	–	2.5
	e-Health	Interactive	–	200	–	–	5611.52
	ITS	Conversational	–	200	–	–	0.06
	Surveillance	Streaming	64	200	384	–	5.5
Video	Calling	Conversational	64	384	2048	60	–
	streaming	Streaming	500	5120	13000	3600	–

6.2 The Scenarios' Road Map

In addition to the reference scenario, a series of scenarios is taken into account to present the characteristics and advantages of virtualisation of radio resources. Although all of them are based on the reference scenario, each one has a difference in a parameter or assumption. In contrast to the variety of the considered scenarios, they can be classified into two main groups: one group aims to present the isolation property, while the other is focused on flexibility.

Furthermore, the isolation group itself consists of a set of scenarios, in which either the capacity of network or throughput demands may change. The scenario with diverse configurations (from the serving weights' perspective) and fairness expectation is placed in this group. The scenarios in the flexibility group are designed to demonstrate this property by changing the VNOs' type, SLAs, and services, besides the subscribers' terminals. This classification of scenarios is presented in Figure 6.3. In what follows, each of these scenarios is discussed briefly.

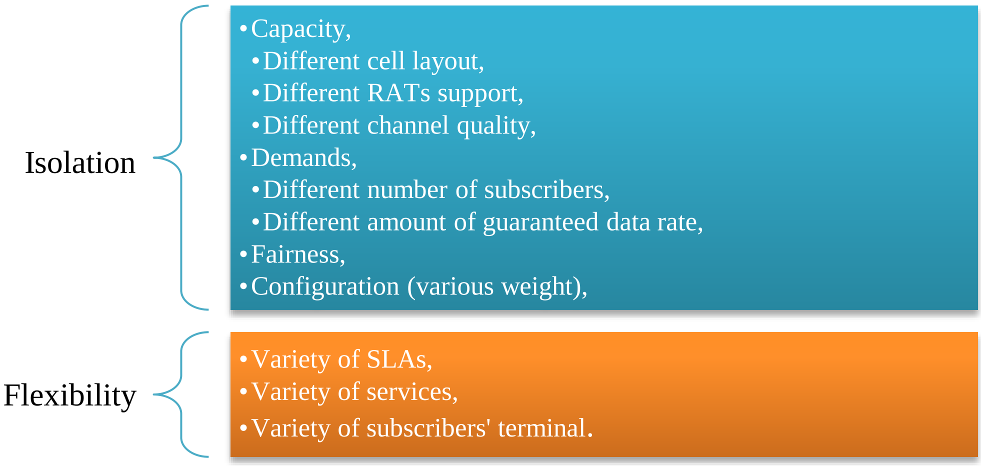


Figure 6.3 – Classification of scenarios.

It can be claimed that the network's capacity is the most important parameter in VRRM, since it is the main constraint. A higher capacity offer is the easiest scenario for the realisation of isolation, hence, the first scenario set is centred on different capacities by changing the cell layout, channel quality and supporting RATs.

Considering the cell layout, the reference scenario takes the serving area as 16 OFDM, 1.7 UMTS, and 1 GSM cells. Since the capacity offered by FDMA/TDMA (i.e., GSM) and CDMA (i.e., UMTS) cells is not that high compared to OFDMA (i.e., LTE) and OFDM (i.e., Wi-Fi), the variation on their cell layout is skipped. In these scenarios, one considers also the variation from 16 OFDMA cells down to 5, which drastically decreases capacity down to 45% of the initial value. Regarding Wi-Fi, since APs are places at the centre of OFDMA cells, the number of APs also varies with the number of the OFDMA cells. In the scenarios with the full Wi-Fi coverage, the portion of capacity that can be used for traffic offloading can be the floating variable.

Supporting different RATs is the focus of the second scenario set. In these scenarios, the physical infrastructure, over which the virtualisation of radio resources is happening, may or may not support some of the RATs. Similar to the reasons presented in cell layout, it is more interesting to concentrate on the presence or absence of Wi-Fi coverage. These scenarios are designed to demonstrate the effect of traffic offloading in network capacity and the offered isolation.

Moreover, the total network throughput changes as the distribution of SINR varies. In the reference scenario, the estimation of network capacity is done according to the general approach; however, as presented in Chapter 4, three other approaches can be used for further evaluations, i.e., optimistic, realistic, and pessimistic. The results for these cases are not only useful for evaluation, but they can help also the lower management of the systems (i.e., LRRM and CRRM) to make better decisions.

Regarding the network capacity effect, traffic demand variation is the next study topic. Depending on the number of subscribers and their request for the guaranteed level, each VNO demands a different throughput. The scenarios in this group take first a different number of subscribers into account (between 300 and 1 400), and then deal with the variety of their terminals. In the next step, VNOs with the same number of subscribers may ask for different guaranteed data rates.

Analysing the effect of the fairness weight on model performance is the key objective of the third scenario set. The fairness weight is swept in its acceptable range (i.e., between zero and one) and the weighted throughput, i.e., the objective function of model, is observed.

Finally, the ultimate presentation of isolation in VRRM can be addressed by serving multiple VNOs with various configurations (e.g., different weights) in the same physical infrastructure. Assuming that each VNO may be optimised for a different set of services, they may have dissimilar serving weights for the same service. The virtualisation of radio resources is expected to meet their unalike objectives simultaneously in the same infrastructure. The set of scenarios in this step is designed to demonstrate this situation, hence, unlike the reference scenario, the serving weight for the QoS classes (i.e., conversational, streaming, interactive, and background) varies from a VNO to another.

In addition to isolation, the flexibility property of the proposed concept is addressed by changing the following VNOs' parameters:

- Service set: in general two major types are considered, i.e., the multi-service and specific service ones.
- SLA: the same services set and serving weight are assumed for all VNOs, but their SLAs are different. For instance, a VNO may have a guaranteed SLA for its conversational services and request a best effort serving for its interactive ones, where another requests guaranteed SLA for its streaming and interactive services.
- Subscribers' terminal: the type of terminals used by subscribers are smartphones, laptops, and M2M. The assumption of the reference scenario is only smartphones, while the usage of other types and their different traffic demands is taken into consideration in this scenario set.

6.3 The Network Capacity Analysis

Referring to Chapter 4, the first step in the management of virtual radio resources is to estimate the available network capacity. The procedure starts by calculating the PDF and CDF of each RAT based on (4.16) and (4.17); for OFDMA (LTE). By means of implementing a numeric convolution, and based on (4.20), the PDF and CDF of data rates for the total network are achieved. In order to have a comparison between the virtualisation of radio resources and RAN sharing, the CDFs of the total network data rate for RAN sharing and the V-RAN approach are illustrated in Figure 6.4. In V-RAN, all resources are aggregated, where as in RAN sharing each network has a third of the resources and the total network data rate is three times the one of a single operator. It can be seen that it is more probable to serve all three operators when resources are aggregated rather than when they are divided. For instance, for 50% of the time, the total V-RAN network capacity is 1 800. Mbps, whereas RAN sharing offers 1 782. Mbps. The largest difference can be seen when the CDF is equal to 0.1, in which case, the relative data rate for the V-RAN is 1 725. Mbps, while RAN sharing offers only 1 656. Mbps.

However, when there is a prior knowledge about channel quality, (4.25) and (4.27) are used. As addressed in Section 4.3.2, three study approaches are considered, i.e., Pessimistic (PE), Optimistic (OP), and Realistic (RL). Table 6.5 presents the minimum and maximum data rates for each RAT, considering different approaches.

Table 6.5 – Minimum and maximum data rate of each RAT in different approaches.

RAT	PE		RL		OP	
	Min. $R_{b_{tot}[Gbps]}^{RAT}$	Max. $R_{b_{tot}[Gbps]}^{RAT}$	Min. $R_{b_{tot}[Gbps]}^{RAT}$	Max. $R_{b_{tot}[Gbps]}^{RAT}$	Min. $R_{b_{tot}[Gbps]}^{RAT}$	Max. $R_{b_{tot}[Gbps]}^{RAT}$
OFDMA	0	2.73	1.37	4.10	2.73	5.47
CDMA	0	1.68	0.84	2.52	1.68	3.36
TDMA	0	0.002	0.001	0.003	0.002	0.004
OFDM	0	25.35	12.67	38.08	25.35	50.78

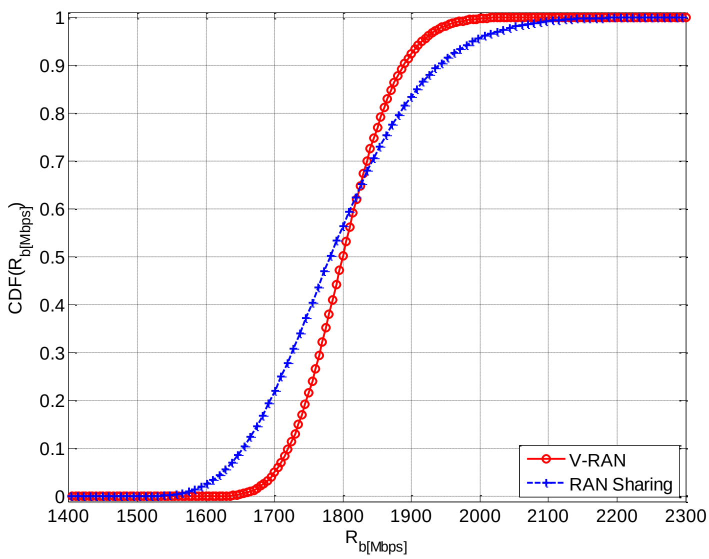


Figure 6.4 – CDF of network capacity for V-RAN and RAN sharing.

In order to compare the differences of these approaches in conjunction with the General (G) case, the CDF of the cellular networks is plotted for the three different path loss exponent cases in Figure 6.5: the lowest data rate is achieved in the PE case, since the data rate of each RRU is less than $0.5\ R_b^{max}$; the capacity of the network increases in other cases, the maximum being achieved in OP, with G and RL being in between.

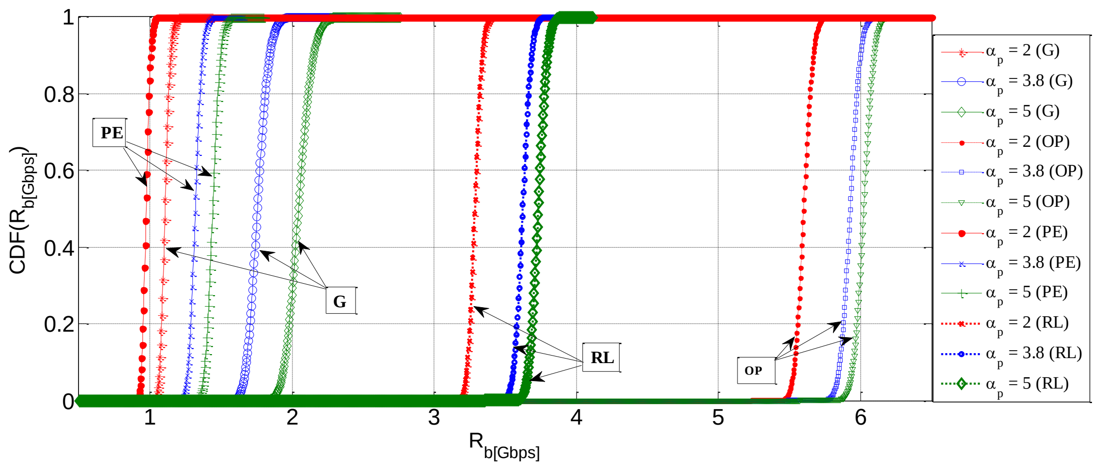
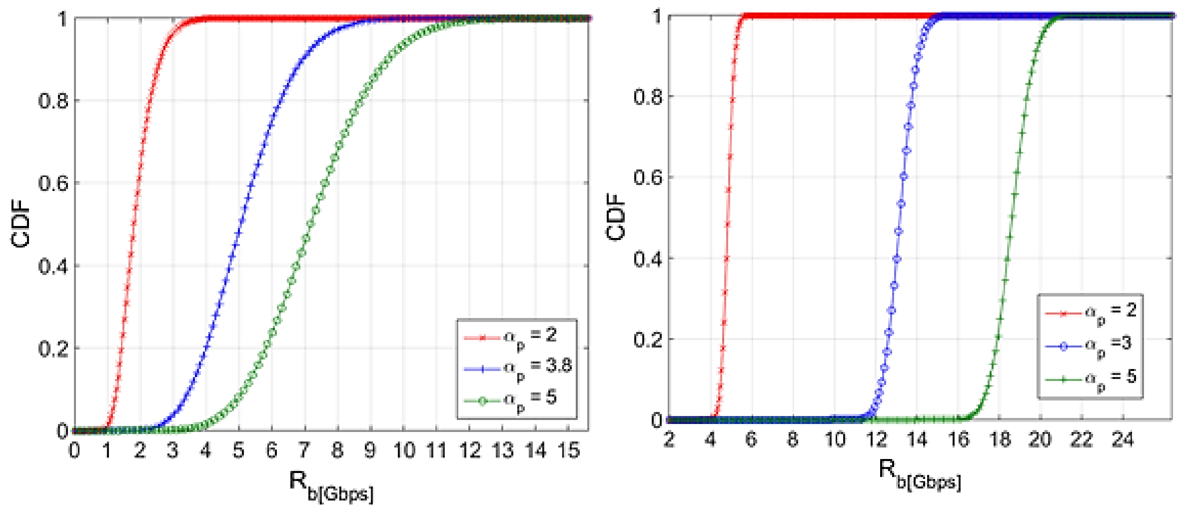


Figure 6.5 – CDF of the cellular network data rate for different approaches.

In addition to cellular networks, the reference scenario considers two configurations for Wi-Fi coverage: one corresponds to have a part of an OFDMA cell covered by Wi-Fi and the other to have full coverage. Figure 6.6 illustrates the CDF for capacity offered by Wi-Fi APs in both configurations. From numeric results, the median capacity offered by APs in partial coverage is 1.76 Gbps for free space, 5 Gbps when α_p = 3.8 (average urban propagation), increasing up to 6.8 Gbps when α_p = 5 (heavily obstructed propagation).



a) Partial coverage (40 channels) b) Full coverage (1024 channels)

Figure 6.6 – CDF of WLAN.

Adding the capacity offered by traffic offloading to Wi-Fi (i.e., partial coverage), the CDFs of the full heterogeneous access network capacity estimated with the aforementioned approaches are plotted in Figure 6.7. Based on these plots, the median capacity of the network estimated by the general approach is 6.8 Gbps, varying from 3.6 Gbps (in the PE approach) up to 35 Gbps (in the OP approach).

In conclusion, results show that by increasing the path-loss exponent the total network capacity increases, due to the increase of SINR. In addition, the channel quality has a considerable effect on the network data rate. The total network capacity may in the OP approach increase up to 10 times of its value in the PE one.

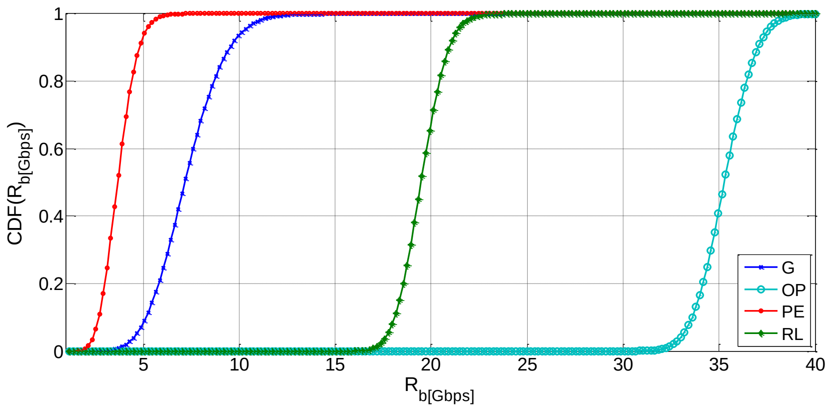


Figure 6.7 – CDF of the WLAN data rate for different approaches$(\alpha_p = 3.8)$.

6.4 Resource Allocation in Reference Scenario

In order to evaluate the performance of the model proposed in Section 4, three VNOs are assumed to operate in this area, i.e., GB, BG, and BE. Each of these VNOs has 300 subscribers with a smartphone, hence, the contracted data rate based on Table 6.2 is:

$$R_{b[\mathrm{Mbps}]}^{Con} = 300 \times 6.375 = 1\,912.5 \quad (6.1)$$

The VNOs offer the same service set addressed in the urban reference scenario, but they have different SLAs as follows:

- Services of VNO GB are guaranteed to be allocated in a data rate between 50% to 100% of the required service data rate.
- VNO BG has best effort with a minimum 25% of the required service data rate; the maximum guaranteed data rate in not applicable (N/A), being served with best effort.
- Services of VNO BE are all served in a best effort manner.

As shown in the previous section, the median of the total network capacity is 1 800. Mbps. Having the total capacity of the network estimated, the allocation of resources for the reference scenario is studied in this section. Regarding the allocation among VNOs, VNO GB receives the biggest portion of resources (i.e., 59%, which is 1 065. Mbps), since it is the VNO with the guaranteed SLA. VNO BE, the VNO in the best effort manner, is allocated only 7% of capacity (120. Mbps), and the rest of the resources are assigned to VNO BG in order to meet the minimum guaranteed.

The results for the different services of the different VNOs are listed in Table 6.6. It can be seen that VRRM manages to meet the minimum guaranteed data rates of the services. Based on the serving weight, as expected, conversational services (e.g., VoIP and Video call) have the highest data rates, in contrast to background services (e.g., Email and M2M – SM). M2M services represent this issue better, since they all have the same volume percentage, but different serving weights. Services of VNO GB and BG have a minimum guaranteed level, so the assigned data rates cannot go below a certain level. This is the reason for having a high data rate assigned to video streaming in these two VNOs. In general, it can easily be concluded that the same services in VNOs GB and BG receive more capacity comparing to VNO BE. It is also worth noting that VoIP of VNO GB is allocated with the maximum guaranteed data rate defined in the SLA, this being the reason that it has the lower capacity compared with other VNOs.

Table 6.6 – Resource allocation to various services of different VNOs in reference scenario.

Services	$R_{b_{ji}}^{Srv}$ [Mbps]		
	VNO GB	VNO BG	VNO BE
VoIP	19.12	21.78	16.94
Music	40.47	26.13	11.78
File Sharing	41.32	24.59	7.85
Web Browsing	121.65	64.75	7.85
Social Networking	145.55	76.70	7.85
Email	13.49	8.71	3.93
M2M-SM	17.08	10.50	3.93
M2M-eH	21.00	14.43	7.85
M2M-ITS	26.30	23.57	16.94
M2M-SV	24.93	18.36	11.78
Video Streaming	556.40	284.09	11.78
Video Call	43.14	30.12	16.94

In conclusion, the results demonstrate the effect of serving weights in addition to the minimum and maximum guaranteed data rates in the reference scenario. These results suggest that, when different services have the same serving weights but different minimum and maximum guaranteed data rates, such as video streaming in the different VNOs, the service with higher minimum guaranteed data rate receives a higher data rate. When the services have the same guaranteed data rates, the service with higher serving weight is allocated with higher data rates, as shown in the M2M services. Finally, the maximum guaranteed data rates, when there are sufficient resources, can also affect the allocation procedure, based on the allocated data rates to VoIP services of VNO GB and BG.

6.5 Resource Allocation with Different Cell Layout

6.5.1 Scenario Rationale

According to the scenarios road map, a different cell layout is the first derivation from the reference scenario in order to study the effect of cell layout on the total network capacity. The changes in cell layout may lead to resource shortage in the network. The considered scenario in this section is chosen for evaluation of the model under the congestion situation. The rationale of the assumed scenario is based on the milestone in realisation of V-RAN, which is meeting the tight timing constraints of RAN [3GPP09]. This constraint makes deployment of low delay equipment (e.g., non-smart optical switches) to be the only choice and the BBU-pools have to be managed (i.e., scale up and down) in run-time to guarantee the timing constraints satisfaction. The management of BBU-pools may contain migration of the pool from one data centre to another. Having non-smart optical switches forces this migration procedure to be done in the order of microseconds, which may not be possible.

The candidate scheme for migrating the data centres of virtual pools, proposed here, is to starve the pool at first, which means to push all connected MTs to other pools. The pool and its RRHs have to be marked temporarily as unavailable, then the migration procedure can start. However, practising this scheme may lead to radio resource shortage during the migration procedure; hence, the VRRM model has to be able to support these changes during the migration procedure.

The scenario's storyline is assumed to be as follows: The processing resource manager (load-balancer) of the network decides to migrate a BBU-pool from a data centre to another, therefore, it has to have the related RRH(s) temporarily unavailable. The effect of having unavailable up to 11 cells out 16 cells, in addition to the normal situation (i.e., all cells are available) has been studied.

6.5.2 Total Network and VNOs Capacity

Figure 6.8 illustrates the total network capacity when different numbers of OFDMA cells are used to cover the service area. The total network capacity with the 16 cells (i.e., the reference scenario) is 1.76 Gbps, which reduces to 48.5% of its initial value (i.e., 872. Mbps) when the full coverage is obtained by only 5 cells. According to the scenario definition, the total guaranteed data rate is 1.4 Gbps, which means that initially there is enough capacity to serve the minimum guaranteed data rate plus the best effort service. The layout with 12 cells is the marginal point where the network capacity and the total minimum guaranteed data rate are almost equal; the use of only 5 cells provides a very low capacity.

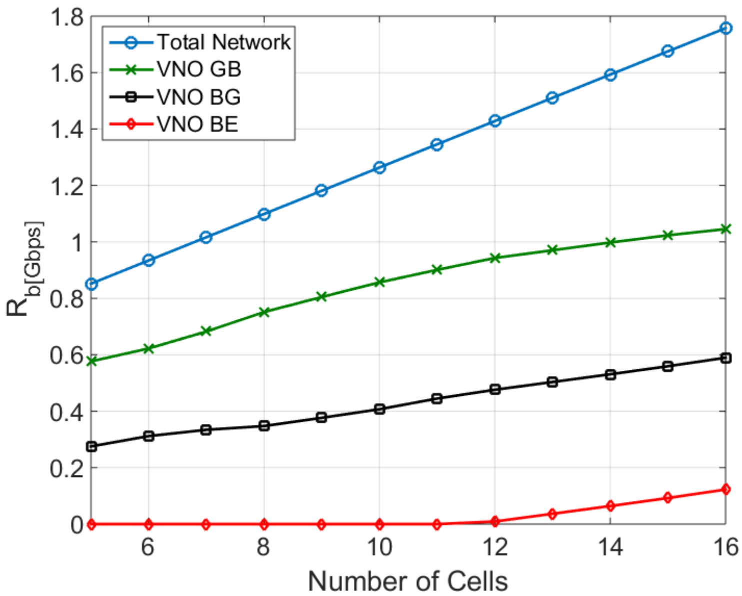


Figure 6.8 – Total network and VNOs capacity by changing number of cells.

Considering the allocated data rate to VNOs, as expected, all the capacity allocated to VNOs decreases by reducing the number of cells. Capacity reduction has a higher impact on VNO BE (the best-effort operator) comparing to VNOs GE and BG, since the network tries to meet these latter VNOs' guaranteed capacity before serving the best effort one. When there are 12 cells, VNO BE gets almost no data rate, but the other two VNOs still have a relatively acceptable data rate. In this situation, the total network capacity is still higher than the total guaranteed capacity.

The network capacity shrinks to 1.378 Gbps when another cell is reduced, i.e., 11 cells, which is lower than the guaranteed data rate. The violation is inevitable for the cell layout with less than 12 cells. In these situations, the main objective function becomes infeasible and VRRM switches to the objective function with violation, presented in (4.58). While no capacity is allocated to VNO BE, the other two VNOs share the violation between them. Since the model tries to minimise the weighted average violation, it can be seen that VNO GE always receives a relatively larger portion of the network capacity, since it has a higher guarantee rate.

6.5.3 Allocated Data Rate to each Service Class in VNO GB

In order to study how different services are affected by the changes of the total network capacity from the VNO GB, the four services classes are considered, with different serving and violation weights. Figure 6.9 illustrates the percentage of violation for each of these services by shutting down cells. Obviously, there is no violation of guaranteed capacity as long as there are more than 12 cells. However, VRRM has no other choice than start violating the level of data rate guaranteed when there are fewer cells. In fact, violations have to start by the service with lowest violation weight. According to the weights presented Table 6.3, background services are the first candidate for violation, since they have the lowest serving and violation weights. When there are only 11 cells, the background traffic violation reaches to 100%, which means that no capacity at all is allocated. Since these services (i.e., Email and M2M-SV) are low volume services, even the total violation of their capacity cannot cover the shortage of network capacity; therefore, interactive services, the ones with the second lowest violation weights, have also to be subject of capacity violations. Finally, when there are only 5 cells (the worst network situation considered in this case), background and interactive services have to be totally shut down (i.e., 100% violation), video streaming suffering a violating up to 8%. Nevertheless, conversational services, the ones with the highest weights, are served without any violations.

The violation of services of VNO BG is presented in Figure 6.10; the violation of the guaranteed data rate of VNO BG is very similar to VNO GB. The violation starts when the number of available cells is less than 12. While the services of background and interactive classes face extreme violations (i.e., up to 100%), the services of conversational and streaming experience no or small violations.

In conclusion, cell layout can have an extreme effect on the total network capacity and consequently on the allocation of virtual radio resources to different services of different VNOs. The results demonstrate the effect of these changes on network capacity, which can be used in the provisioning of virtual RANs. The violation to the minimum guaranteed data rates defined in the SLAs of the VNOs starts from the services with lower violation weight in order to minimise the effect of the resource shortage in the network. Based on the results, the proposed model demonstrates an acceptable level of isolation among VNOs facing extreme network capacity variation.

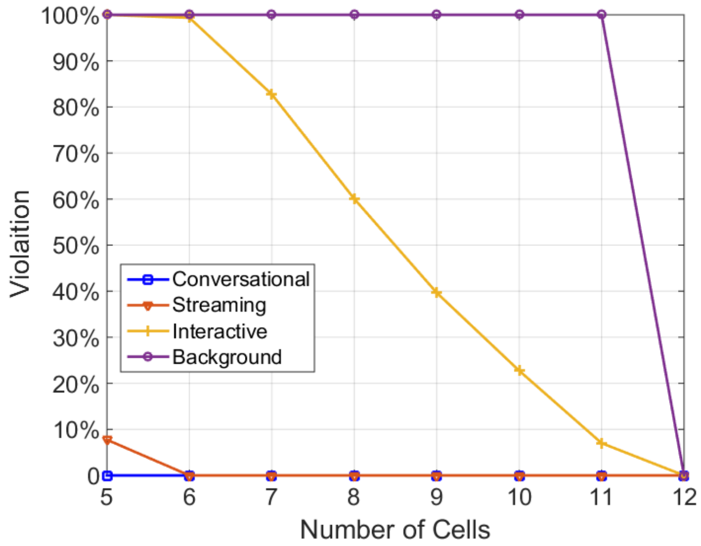


Figure 6.9 – Percentage of violation of services of VNO GB by changing number of cells.

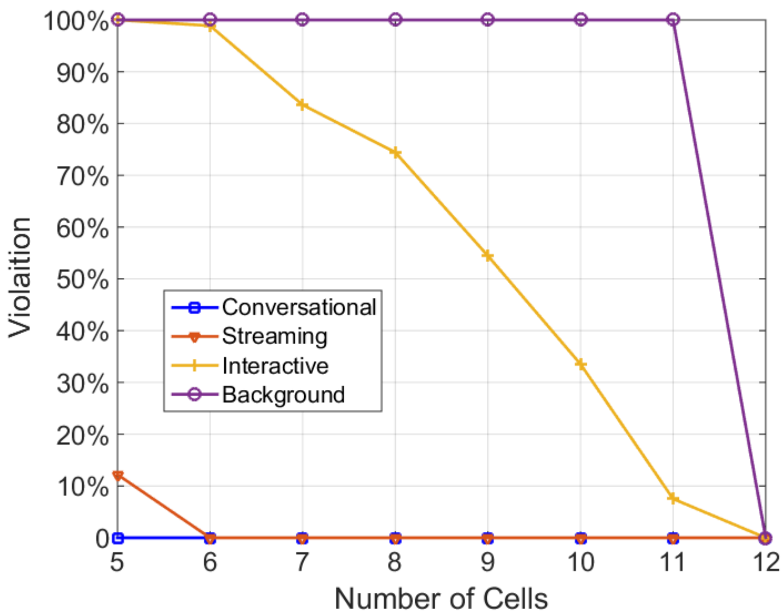


Figure 6.10 – Percentage of violation of services of VNO BG by changing number of cells.

6.6 Resource Allocation in Full HAN

Referring to the reference scenario, the main assumption is to have Wi-Fi APs deployed in the centre of the OFDMA cells. Subscribers placed in the Wi-Fi coverage area can connect to both Wi-Fi and cellular networks, while the cellular network is the only option for the rest of users. To serve users in both regions, the resource allocation in this situation has to supply capacity for each service from both cellular and Wi-Fi networks. However, it is worth noting that VRRM is allocating data rate to various services based on the entire physical radio resources. While it considers the differences between Wi-Fi and cellular networks in VRRM, the load balancing among different RATs as well as between Wi-Fi and cellular networks is done by CRRM.

Figure 6.11 presents the data rates allocated to each of the services from cellular networks and WLANs in the general approach, when there are 500 subscribers per VNO. As expected, Conversational services (VoIP, Video Call, and M2M – ITS), which are the ones with the highest serving weights, receive the highest data rates. Services of the Streaming class (Music, M2M – SV, and Video Streaming) are in the second place. The services of the Background class (email and M2M – SM) are the ones that are allocated the smallest portion of the available capacity. The effect of the weight for session average data rate in WLAN, W^{SRb}, can be observed in the data rates balance between WLANs and cellular networks. For instance, the data rate allocated to Video Streaming (ViS) from WLANs is 6.5 times higher than the one from cellular networks, while, in contrast, Email, a service with a low average data rate, is allocated a higher data rate in cellular networks than in WLANs. However, VoIP is not following the same rule since it has a relatively high serving weight, which overcomes the effect of the average data rate in (4.52), hence, being allocated a comparatively high data rate from both type of networks; the same phenomenon can be observed among M2M services, i.e., the ones with high serving weights, e.g., M2M – ITS, receives a relatively high capacity from WLANs compared to the other ones.

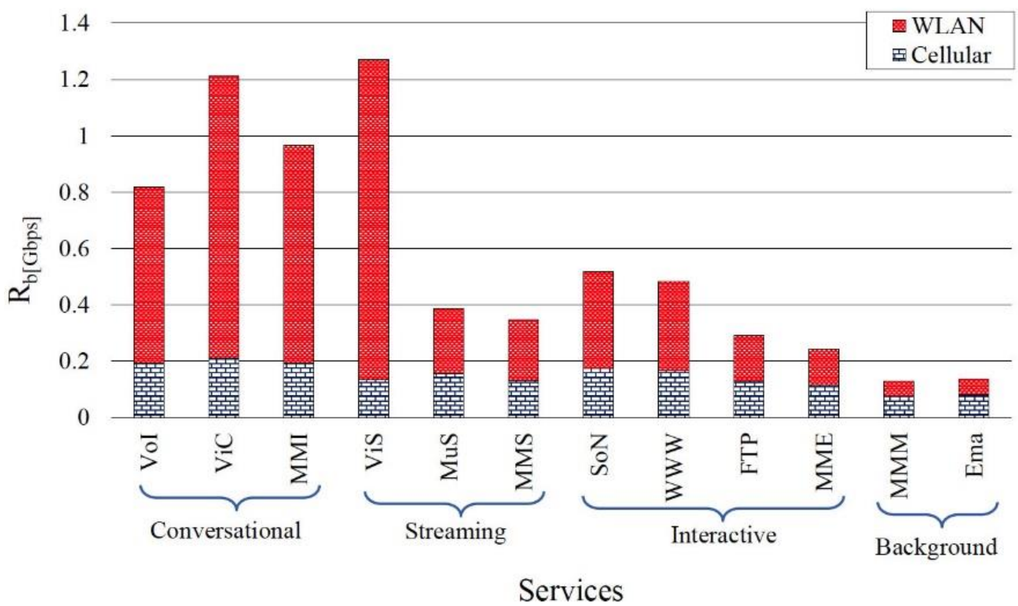


Figure 6.11 – Allocated data rate to different services from Cellular and WLAN (partial coverage).

Figure 6.12 illustrates the data rates that are allocated to each service of each VNO. It is apparent that most of the services of VNO GB are served with the maximum guaranteed data rate, e.g., VoIP and Music are guaranteed to be served with a relatively global maximum of 31.87 Mbps and 95.62 Mbps, and they have been allocated the same amount. The main difference between VNO GB and VNO BG is the service maximum data rate; since the services of VNO BG have no high limit while the VNO GB ones have, the total allocated data rate to the former is higher than the latter. Obviously, the offered capacity to VNO BG is a resource shortage situation that is going to be smaller than VNO GB.

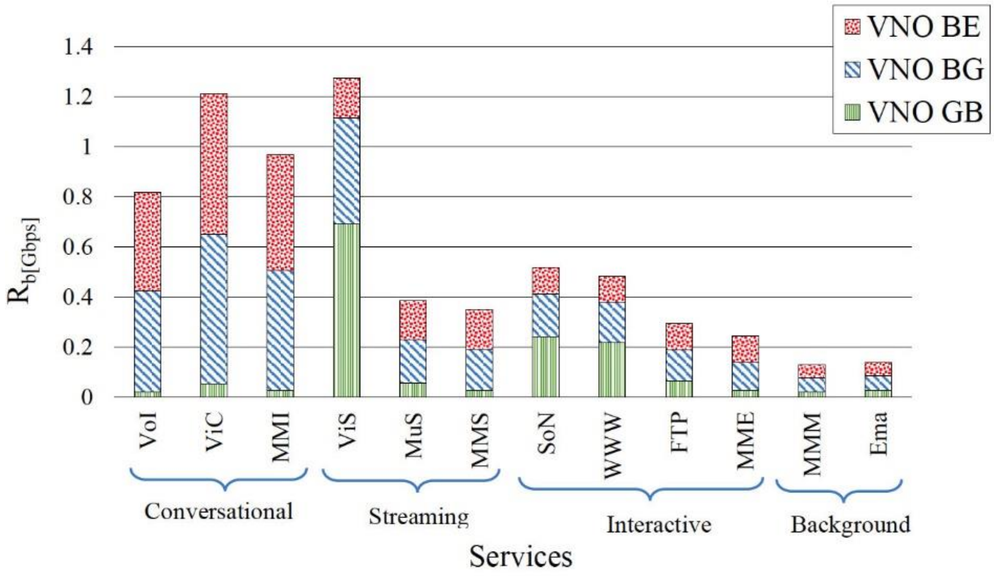


Figure 6.12 – Allocated data rate to different services of the VNOs in Full HAN (partial coverage).

The other assumption in the reference scenario is to have the APs offering the full coverage over the serving area, but only 5% of their capacity can be used for offloading. The capacity that the end-user can use from the Wi-Fi network is estimated to be 6.6 Gbps. Figure 6.13 shows the allocated data rate to each service from WLAN and cellular networks. By comparing this figure with Figure 6.11, one can see that the allocated data rates to services with low average data rates, such as Email, do not have any considerable change despite having 1.6 Gbps increase in the network capacity.

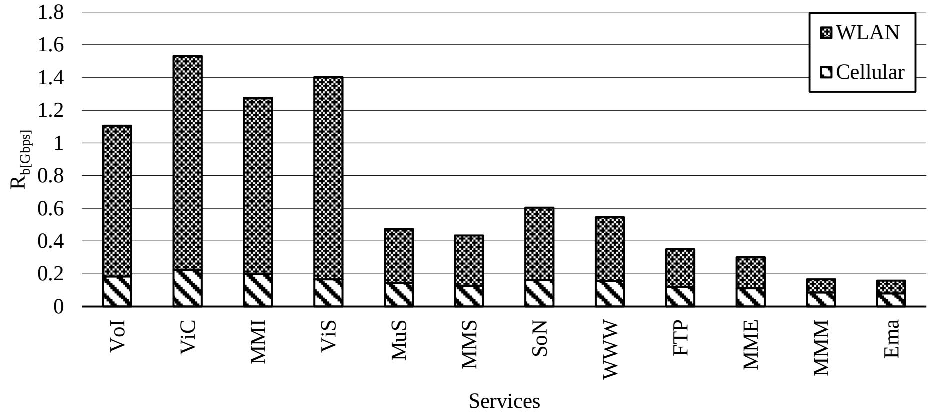


Figure 6.13 – Allocated data rate to different services from Cellular and WLAN (full coverage).

The allocated data rates to the services of the VNOs are shown in Figure 6.14. It is apparent that the extra capacity of the network is mostly allocated to the services of VNO BG and BE, since the services of VNO GB are already served by the maximum guaranteed data rates.

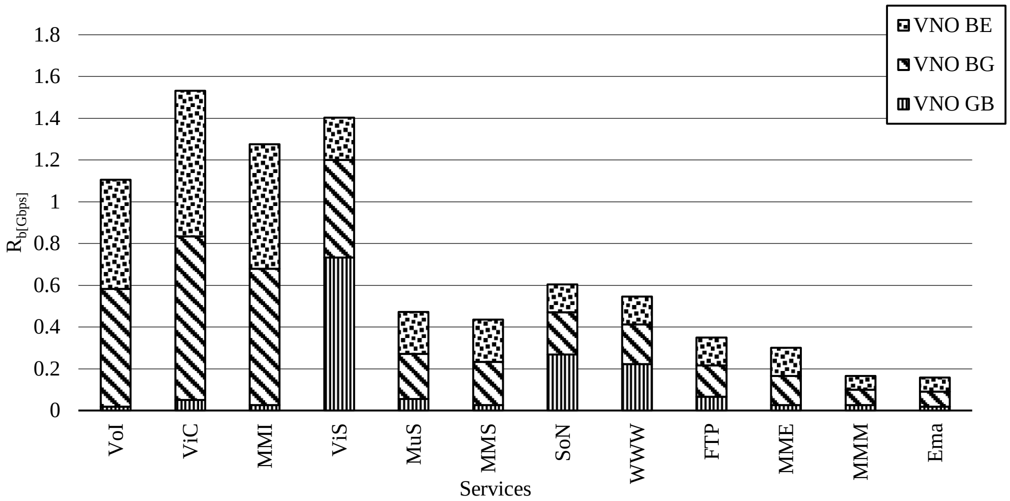


Figure 6.14 – Allocated data rate to different services of the VNOs (full coverage).

In conclusion, the results clearly show that the capacity added by WLAN enables VRRM to serve better the best effort VNOs and their services, while serving the guaranteed ones.

6.7 Resource Allocation for Different Number of Subscribers

In this section, the number of the subscribers of each one of the VNOs is swept from 300 (i.e., low load) up to 1 500 (i.e., high load). The goal is to study the performance of the model and allocated data rate to the VNOs and their services. Figure 6.15 illustrates the distribution of the available virtual resources among VNOs, in addition to the total network capacity (R_b^{CRRM}), the total minimum guaranteed, and the contracted data rates (R_b^{Con}). The contracted data rate for each VNO increases from 1.91 Gbps (low load) to 8.92 Gbps (high load). The acceptable region for VNO GB and BG in the plot are shown by solid blue and light yellow colours. The total minimum guaranteed data rate, i.e., the summation of minimum guaranteed data rates of VNO GB and BG, in the low load is 20.6% of the network capacity (i.e., 1.4 Gbps).

Since the network capacity is considerably higher than the minimum guaranteed data rates, best effort services are also served well; the allocation of 2.39 Gbps to VNO BE is the evidence to this claim. It is worth noting that the share of VNO BE is 35.1% of the whole network capacity, which is 1.6 times higher than VNO GB. The reason behind this observation is the maximum guaranteed data rate of the guaranteed services. Although the assigned portion of available resources to VNO GB is not as big as the other two VNOs, as it is served up to it maximum satisfaction. In contrast, VNO BG, which has a minimum guaranteed data rate but no maximum data rate, receives 43.3% of the network capacity (2.94 Gbps). The guarantee data rates grows up to 6.53 Gbps (i.e., 96.1% of the total available capacity) as the load increases. Obviously, the share of best effort services in this situation considerably decreases. The allocated capacity to VNO BE reduces to only 65.6 Mbps, which is 0.9% of the total available capacity and 97.3% of its initial value. As shown in Figure 6.15, the total network capacity is only enough for serving the contracted data rate of only one of the VNOs. In addition, the increase of subscribers to 1 400 makes the total minimum guaranteed data rate of the three VNOs equal to the total network capacity, which means that the data rates allocated to the services of VNO BE reaches zero.

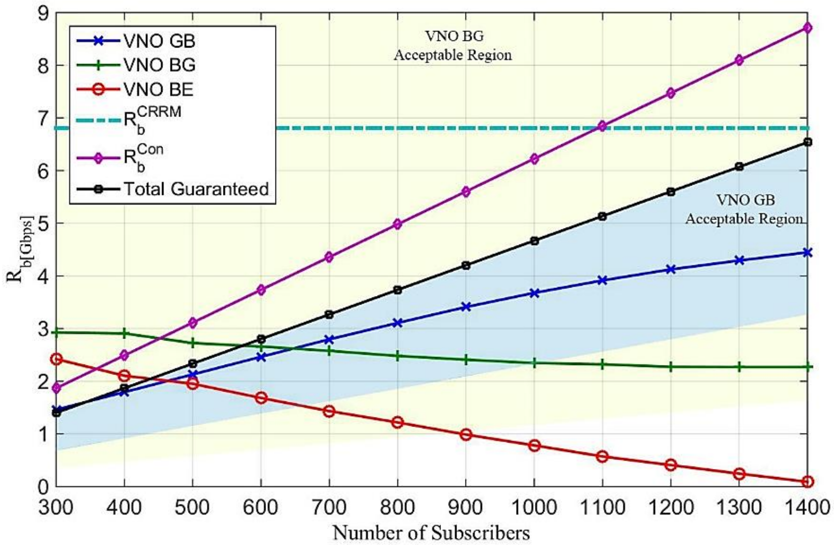


Figure 6.15 – Variation of allocated data to each VNO for different loads.

As the number of subscriber increases, the guarantee data rates grows to 6.54 Gbps, which is 96.05% of the whole available capacity. Obviously, the share of best effort services in this situation is considerably low. As it is presented in the figure, the increase in traffic demand leads to decremented of the best effort segment.

Furthermore, Figure 6.16 illustrates the effect of demand variation on the allocation of data rates to the service classes of VNO GB: this VNO is a guaranteed one, therefore, each service class has a minimum and maximum guaranteed data rate, presented in the figure with the solid colour. By increasing the number of subscribers, demand increases 4.7 times.

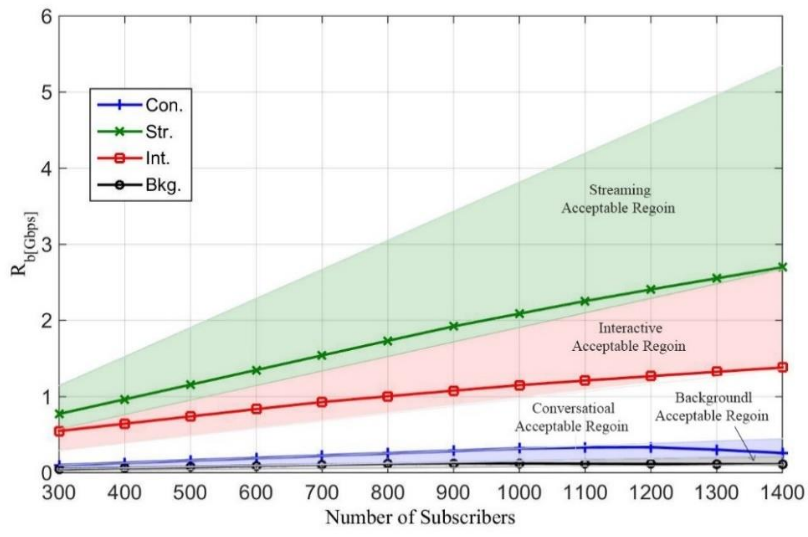


Figure 6.16 – Allocated data rate to service classes of VNO GB for different loads.

It can be seen that the streaming services are the services with the highest volume, having the highest data rate; the minimum guaranteed data rate varies between 0.58 Gbps and 2.71 Gbps. The allocated data rate to this class in low load (when there are only 300 subscribers) is 67% of the maximum guaranteed data rate, but it reduces to the minimum one (i.e., 50% of the contracted data rate) for the maximum load case. The other service classes are served according to capacity needs, i.e., interactive, conversational, and background. It can be seen that, in the low load situation, the maximum guaranteed data rates are assigned, but as the demand increases the data rates move towards the lower boundary. The interactive service class is a very good example for this behaviour: while it receives the maximum guaranteed data rate of 0.54 Gbps in low load, the allocated capacity for the high one is reduced from the maximum to 1.45 Gbps, the minimum guaranteed data rate. Considering the slope of allocated data rates in various services, the effect of serving weights and service volume can be seen. Since the interactive class has a lower serving weight compared to conversational, it receives almost the minimum acceptable data rate with 1 100 subscribers; in the same situation, conversational services are still provided by the highest acceptable data rate.

Figure 6.17 presents the allocation data rate in VNO BG. The services are this VNO are guaranteed to be served with a minimum data rate. There is no upper boundary for these services and the allocation is done in best effort manner, after providing them the minimum guaranteed capacity. Hence, as it is shown in the figure, the acceptable capacity for each of them is a value bigger than the minimum guaranteed data rate. Conversational services receive the highest data rate since they have the highest serving weight; however, as traffic demand increases, the allocated data rates to the services of this class have to reduce in order to meet all the guaranteed levels. Although the figure shows a drastic reduction for this service set, it still has a comparatively better serving situation.

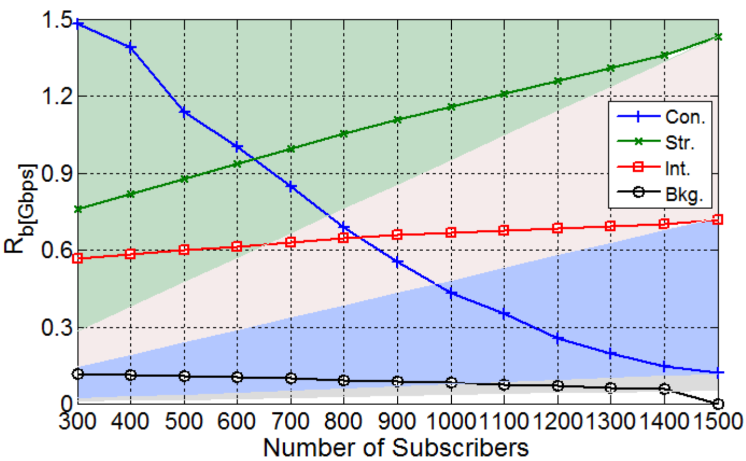


Figure 6.17 – Allocated data rate to service classes of VNO BG for different loads.

Figure 6.18 demonstrates the allocation of data rates for service classes in VNO BE. In this VNO, there is no lower boundary as well as upper one. The allocation is done totally in a best effort manner. In addition, a fairer allocation can be observed in this VNO, because of the absence of a minimum guaranteed data rate. The effect of serving weight on the allocation procedure can also be seen,

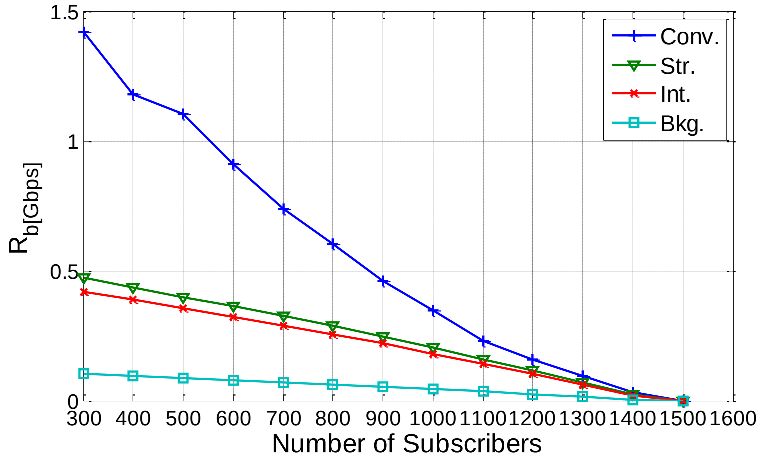


Figure 6.18 – Allocated data rate to service classes of VNO BE for different loads.

In conclusion, the results show that by increasing the networks' load, the services of the best effort VNOs are allocated less resources, in contrast to the guaranteed VNOs.

6.8 The Effect of Channel Quality Effect on VRRM

As previously comprehensively discussed, the channel quality of the link in between user's terminal and network has a considerable effect on the total network capacity, and obviously on the virtual resources allocation. This effect on the management of virtual radio resources by considering the three approaches (i.e. OP, RL, and PE) is studied in this section. Based on the reference scenario, the number of terminals per VNOs is swept between 300 (i.e., low load) up to 1 400 (i.e., high low). In addition, the total capacity of the network is estimated using the OP, G, RL, and PE approachs (see Chapter 4 for more details).

Figure 6.19 presents the allocated data rate to VNO GB in conjunction with minimum and maximum guaranteed data rates. As long as one has the data rates within the acceptable region (i.e., shown by the solid colour), there is no violation to the SLA and guaranteed data rate. In the OP approach, the maximum guaranteed data rate is offered to this VNO up to 600 subscribers. For the other approaches, as the number of the subscribers increases the allocated data rate moves towards the minimum level of the guaranteed data rate. Finally, in the PE approach, as the number of subscribers passes 1 100, violations to the minimum guaranteed data rate can be observed; this means that the network capacity is too low, so that the model faces a resources shortage, and it has to violate the minimum guaranteed data rate of this VNO.

Figure 6.20 shows that the maximum guaranteed data rate reaches to 3.735 Gbps when there are 600 subscribers, being possible to allocate all of it in the OP approach, and that for 1 400 subscribers it reaches 8.7 Gbps, but with only 83.3% of it being allocated to the OP approach. However, VNO GB faces the violation on the minimum guaranteed data rate in the PE approach, as the number of subscribers passes 1 100. While at least 3.42 Gbps is required, only 97.3% of it is allocated to the VNO; this means that the network capacity in this approach is lower than the total minimum guaranteed data rates, and the resource management has to violate some of the minimum guaranteed data rates. The VNO requires at least 4.36 Gbps for the heavy load, the allocated data rate being 6.28 Gbps in RL and 4.42 Gbps in G, which are still enough to fulfil the SLA, but it goes down to 3.3 Gbps, i.e., 76.8% of the minimum required data rate, in PE, in clear violation of the SLA. This clearly shows the effect of SINR on resources usage efficiency and quality of offered service to VNOs.

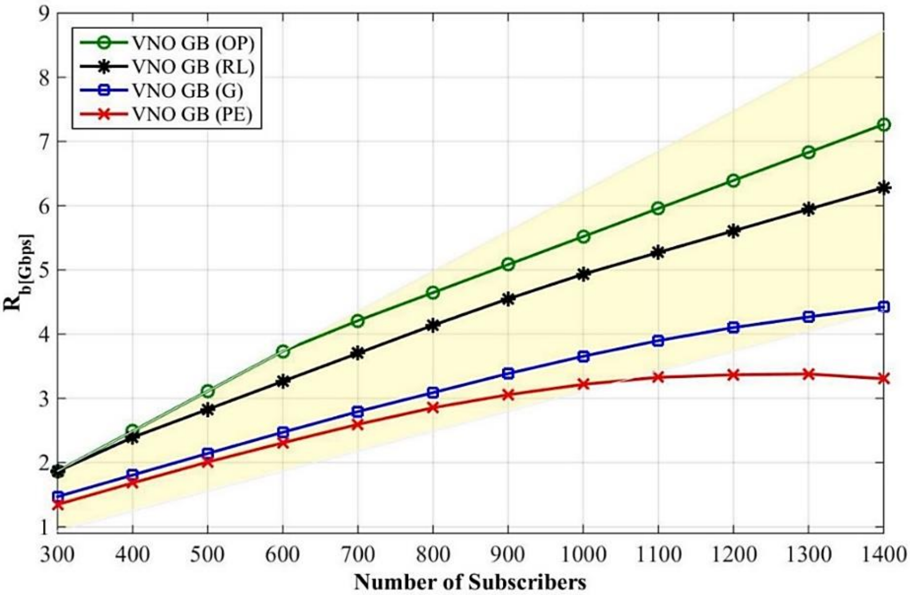
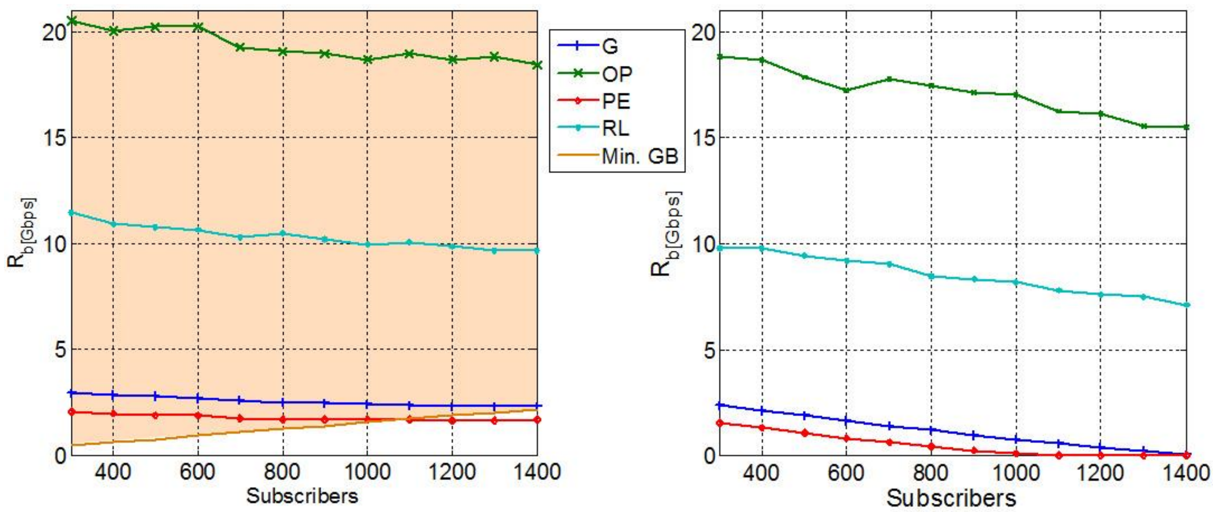


Figure 6.19 – The allocated data rate to VNO GB in different approaches.



a) VNO BG b) VNO BE

Figure 6.20 – The allocated data rate to VNO BG and VNO BE in different approaches.

Regarding the distribution of allocated data rates to VNO GB, Figure 6.21 illustrates the variation of the capacity assigned to each one in different approaches. It can see that the conversational class (i.e., the class with the highest service weight) for the OP and RL approaches receives the maximum guaranteed data rate. The allocated data rate for the G and PE approaches is more than 50% of the contracted data rate. Although for high-density situations in the PE case the data rate decreases to a minimum guaranteed data rate, services of this class never experience violation of guaranteed data rate. The streaming class, likewise, in this VNO is always served with the data rate higher than minimum guaranteed. The maximum guaranteed data rate in heavy load reaches 5.34 Gbps and 72.9% in OP, 63.2% in RL, 50.41 in G and 50% in PE is allocated. For interactive and background classes, it is shown that they face violation of minimum guaranteed data rate in the PE approach. The violation situation in the background class is to such an extent that no capacity is allocated to its services when there are more than 1 100 subscribers per VNO. The allocated data rate to interactive class reaches to 15.1% of the contracted data rate in heavy load, while the minimum guaranteed is 50%.

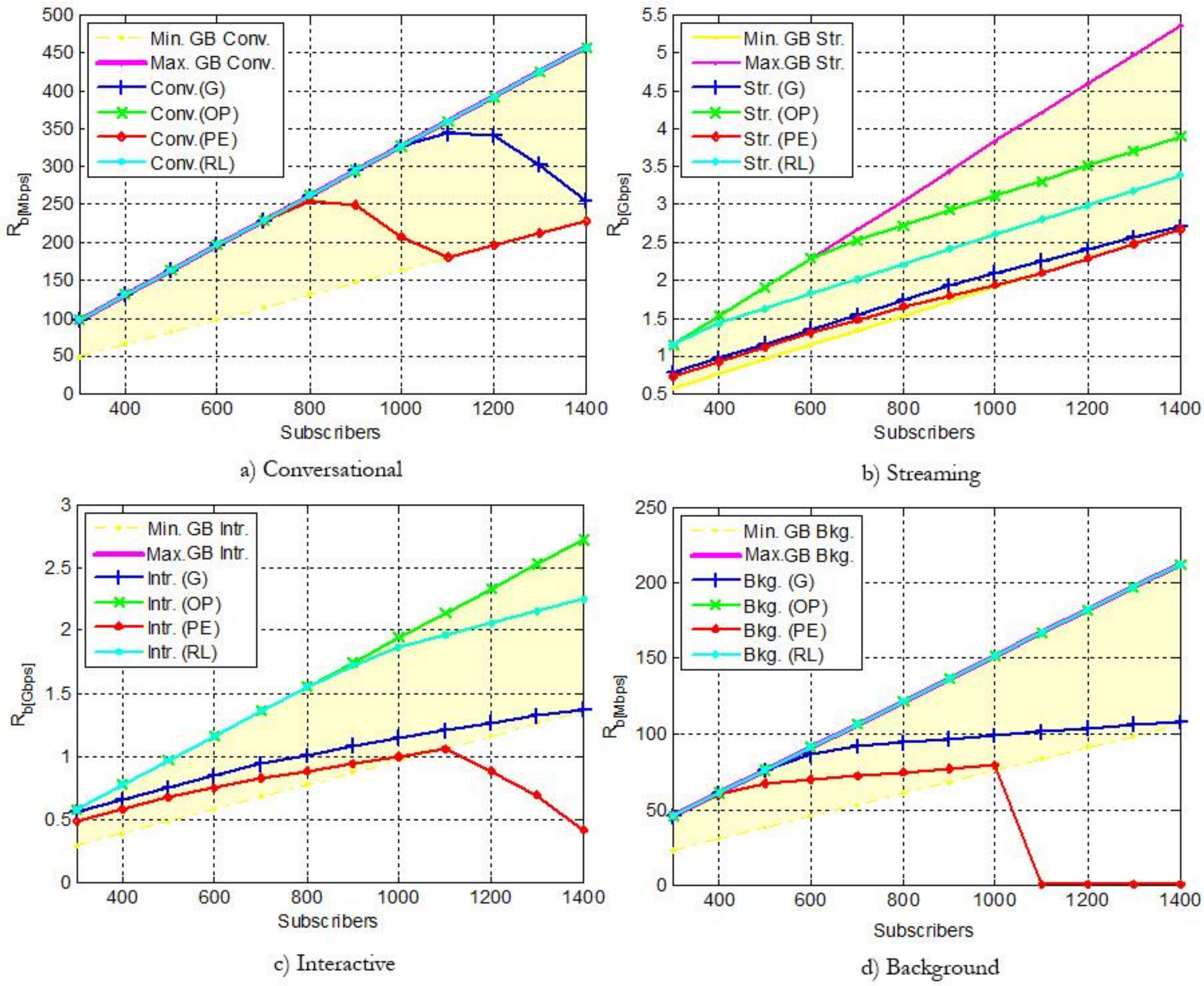


Figure 6.21 – The allocated data rate to service classes of VNO GB in different approaches.

Moreover, the allocated data rate to the service classes of VNO BG and BE is shown in Figure 6.22. It is can be seen that for the VNO BG the situation is very much similar to VNO GB. The main difference is the high boundary of allocated data rate. VNO BG does not have a maximum guaranteed data rate or high boundary for allocated data. Consequently, when a high network capacity is available, e.g., in the OP situation, the services of this VNO are served by comparatively higher data rates comparing to VNO GB. As an example, consider the conversational class in both VNOs. For OP approach with 400 subscribers, VNO GB is granted with 0.1 Gbps while VNO BG receives 1.3 Gbps; on the other hand, in the case of resources shortage, VNO BG receives data rates lower than VNO GB. For instance, the share of interactive class of VNO BG when there are 1200 subscribers in the PE approach is only 386.63 Mbps while VNO GB is allocated 907.77 Mbps.

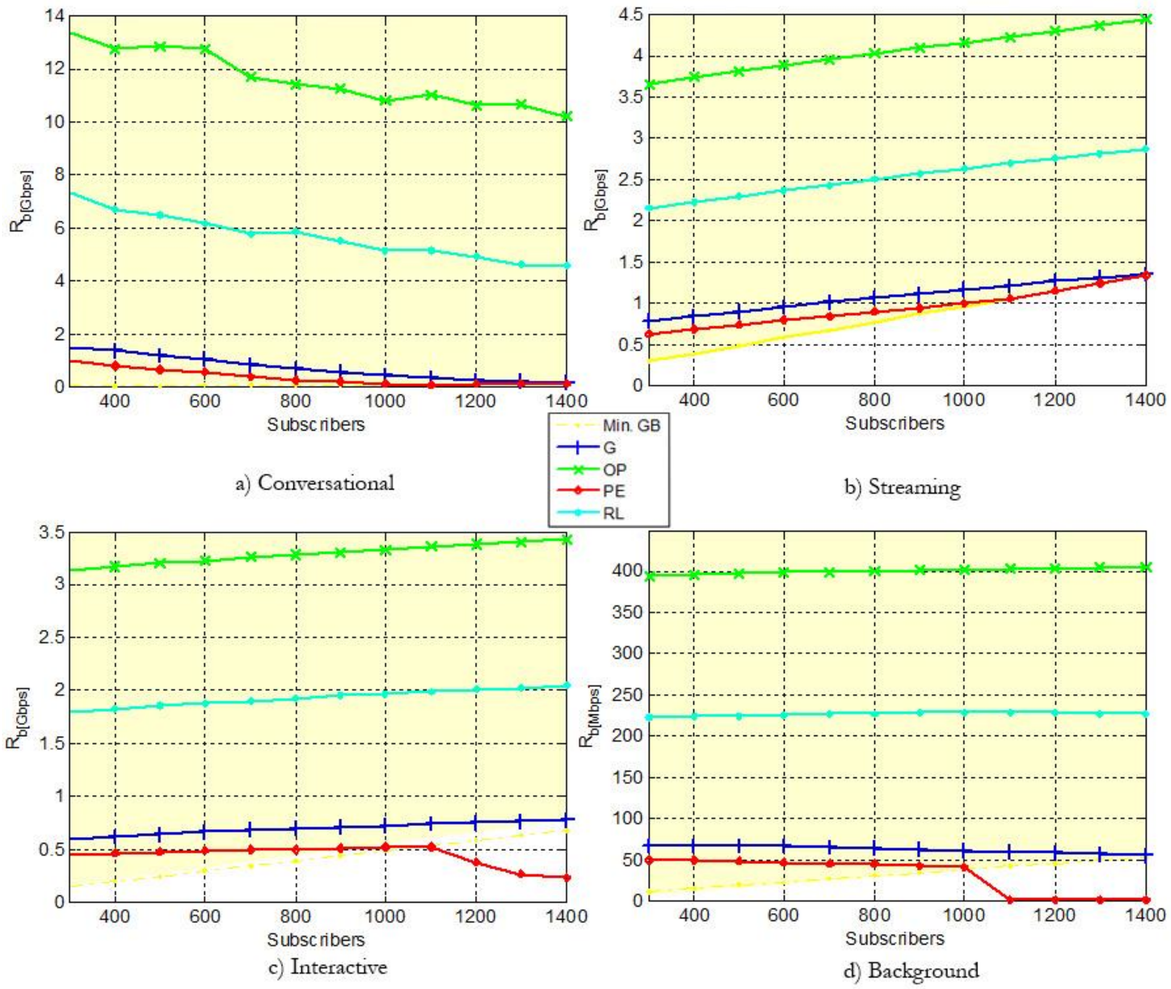


Figure 6.22 – The allocated data rate to service classes of VNO BG.

Furthermore, the allocated procedure to the service classes in VNO BE is very similar to the ones in VNO BG, according to Figure 6.23. By comparing the results, it can be concluded that services of VNO BE, such as conversational class, receive lower data rates than VNO BG, but higher than VNO GB, when there are more than the required resources (e.g., 400 subscribers in OP approach). However, the services of this VNO are on the top list to stop being served.

In conclusion, the effect of these capacity changes on the allocated data rates for different VNOs and their services classes is presented. It is shown that when there is enough capacity, not only the guaranteed VNO is satisfied, but also the best effort VNOs are also well served. However, as the network capacity decreases due the channel quality (G and PE approaches), the best effort VNOs are effected more than the guaranteed one. The same situation is shown in the service class level too. The conversational and the streaming classes are the classes with the highest serving priority (i.e., serving weights). These service classes are allocated with data rates higher generally than the other two service classes. When there are resources shortage, i.e., in the G and PE approaches, violations start by the background and interactive classes.

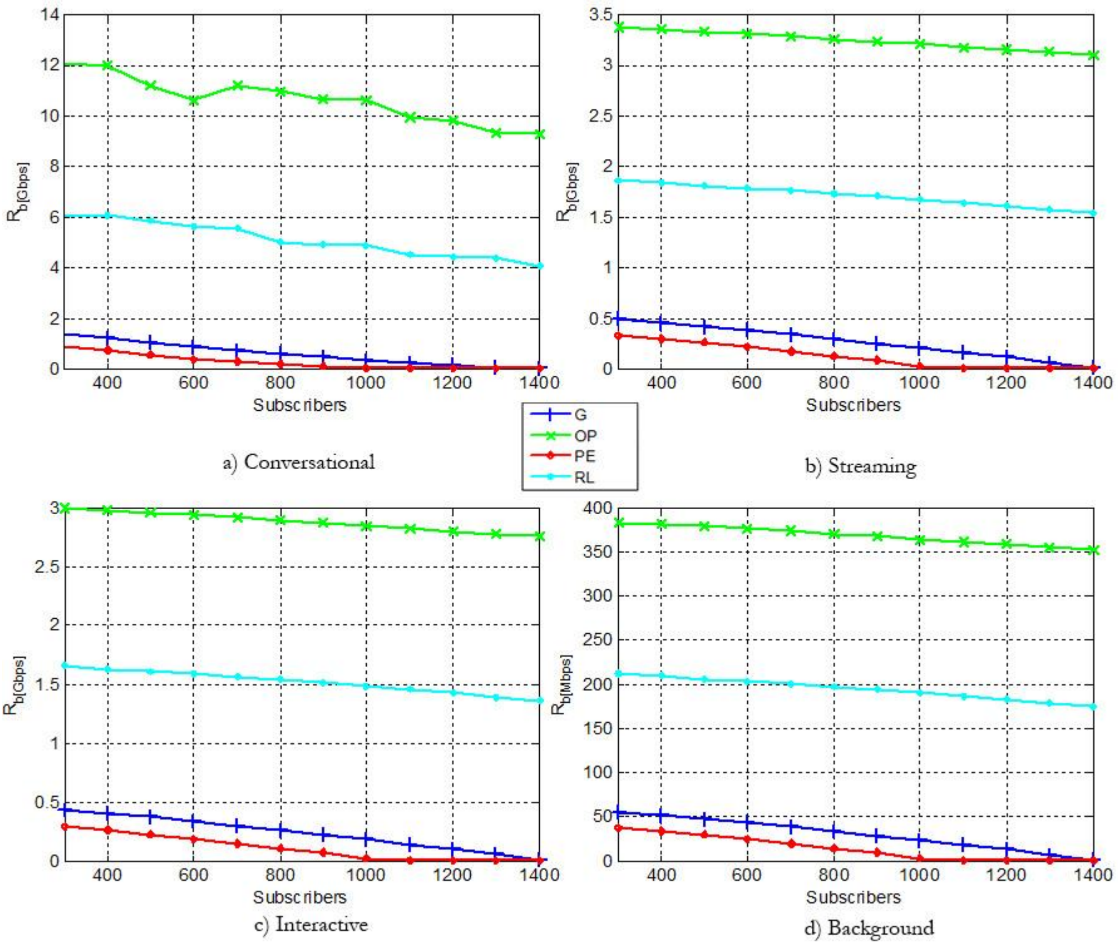


Figure 6.23 – The allocated data rate to service classes of VNO BE in different approaches.

6.9 Resource Allocation with Different Demands

In this section, the aim is to study the effect of different VNOs' demand in terms of minimum guaranteed data rates on the management of virtual radio resources. The performance of the proposed model is evaluated using two series of scenarios: one set with focus on guaranteed VNOs, and the other with best effort with minimum guaranteed VNOs. The number of subscribers per VNO is swept from 100 up to 1400, in both scenario sets.

In the first set of scenarios, three GB VNOs and a BE VNO are assumed. The guaranteed data rates for each of these VNOs are as follows:

- VNO GB75: the guaranteed data rate is 75% up 100% of contracted data rate,
- VNO GB50: the guaranteed data rate is 50% up 100% of contracted data rate,
- VNO GB25: the guaranteed data rate is 25% up 100% of contracted data rate,
- VNO BE: served totally in best effort manner.

The allocated data rates to each of these VNOs are plotted in Figure 6.24. In addition to the allocated data rate, the minimum guaranteed data rate for each VNO is also presented by dashed line. It can be seen that the VNO with higher guaranteed data rate is allocated with higher data rate too. The resource shortage occurs in this scenario set when there are more than 700 subscribers per VNO. In this situation, serving the VNO BE is stopped, while the other VNOs are also allocated with a fixed among of data rate.

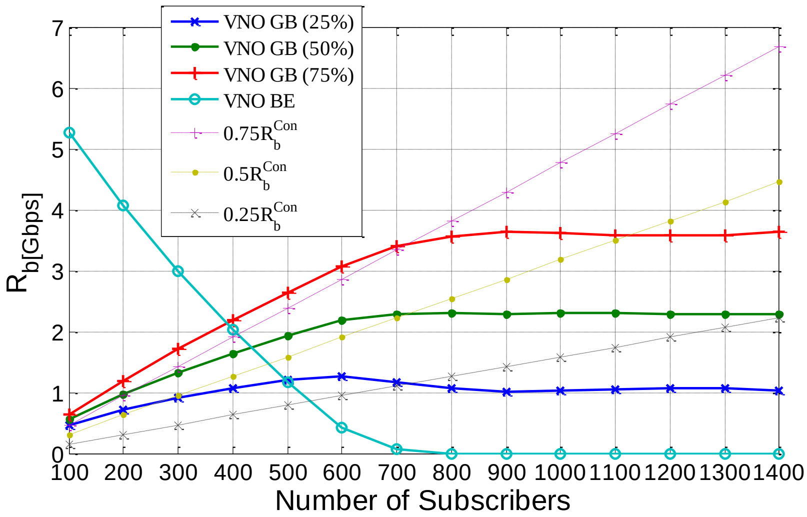


Figure 6.24 – Allocated data rate to various guaranteed VNOs.

Regarding VNO BE, it can see that acceptable data rates are assigned to it when the requested guaranteed data rate is relatively low (i.e., when the number of subscribers per VNO is less than 600). However, as the demand by the guaranteed VNOs increases, the share of VNO BE from the available resources decrease. The extreme situation happens is when the minimum demanded guaranteed data rate and the available network capacity become equal. In this situation, there is not enough capacity to serve the best effort VNO as well as guaranteed VNOs. Hence, the allocation of data rate to this VNO is totally terminated.

In the next step and in the second scenario set, the aforementioned guaranteed VNOs are changed into best effort with minimum guaranteed VNOs. Hence, the VRRM now has to serve four VNOs as follows:

- VNO BG75, which allocation of at least 75% of contracted data rate is guaranteed,
- VNO BG50: which allocation of at least 50% of contracted data rate is guaranteed,
- VNO BG25: which allocation of at least 25% of contracted data rate is guaranteed,
- VNO BE: served totally in best effort manner.

The allocation of data rates to these VNOs is illustrated in Figure 6.25. Although the allocation procedure is very similar to the former scenario, some differences can be spotted. Considering the low load situation, in contrast to the former scenario set, VNO BG 75%, the VNO with highest guaranteed data rate, is allocated with highest data rate. The reason of this difference lays on the maximum guaranteed data rates. In contrast to VNO GBs, VNO BGs do not have any upper boundary for allocation of data rates, therefore, higher data rates are going to be allocated to them when there are available resources, and the allocation of capacity to VNO GBs stops after the maximum data rates are met.

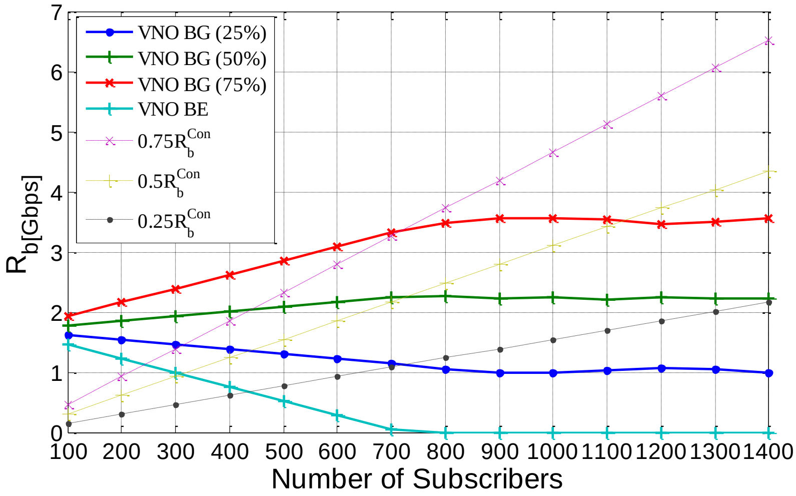


Figure 6.25 – Allocated data rate to various BG VNOs.

In conclusion, it can observe that VNO BE in these scenarios always receives the lowest capacity, since it has no guaranteed data rate. The behaviour for the VNO GBs and VNO BGs is almost similar. However, VNO BGs receive relatively higher data rates in the low load situation. When the network is under heavy load, the increase in the demand of VNOs does not affect the data rate allocated to them.

6.10 Influence of Different Parameters

6.10.1 Influence of SLAs

Recalling that the advantages of virtualisation of radio resources, as discussed in Chapter 4, are the flexibility and isolation of VNOs, by means of the virtualisation of radio resources and the proposed model, it is possible to have multiple configurations for serving the VNOs. In this section, these advantages are investigated in more detail, by considering two types of configurations for serving and violation weights. In the first configuration set, serving weights are kept as in the former sections (i.e., conversational class with highest serving rate and background class with the lowest one). However, serving weights for the interactive and conversational classes are exchanged in the second configuration set, thus, in this new scenario set, the most important service class is the interactive one.

Figure 6.26 summarises the results for three network situations, which are low load (i.e., 500 subscribers), resource shortage (i.e., 800 subscribers), and high load (i.e., 1200 subscribers). By comparing the conversational class of the two configurations, one can see that having configuration #1 leads to higher data rates for this class compared to configuration #2. As expected, the situation for the interactive class is in contrast to the conversational one: configuration #2 comparatively puts more weight on this class than conversational, hence, the allocated capacity to the interactive class is higher than the conversational one.

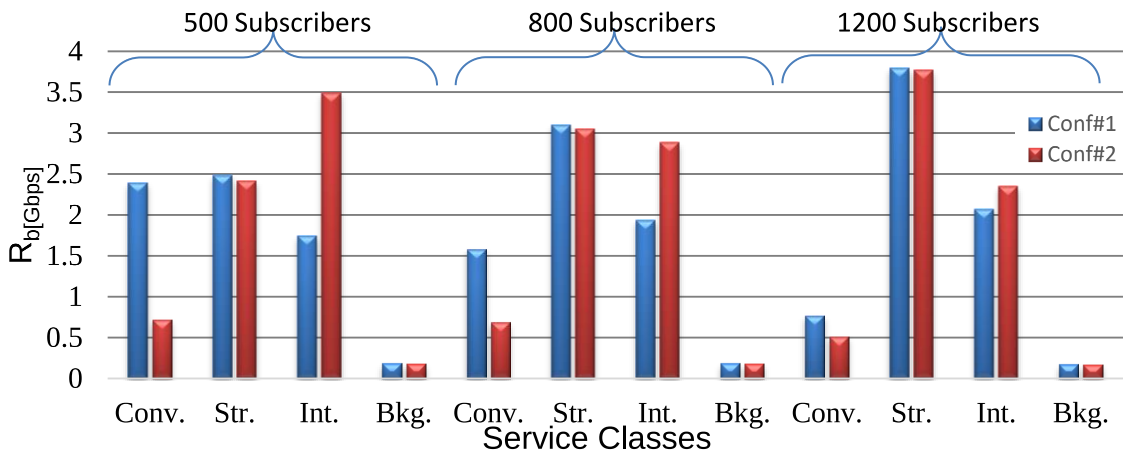


Figure 6.26 – Allocated data rates to service classes for different configuration.

6.10.2 Influence of subscribers' terminals

Regarding the terminal types, four types can be considered: smart phones, tablets, laptops, and M2M terminals. Due to different service profiles, each terminal type requires different average data rates. Hence, the VNOs serving subscribers with different types of terminals have to ask for different amounts of contracted data rates. The aim in this section is to demonstrate the effect of having different terminal types on the management of virtual radio resources. Therefore, considering the same service set and service classes, the average data rate required by each terminal is changed. It is worth noting that for the sake of simplicity and without loss of generality, it is assumed that the service set is independent of the terminal's type. The scenario in this section considers the same VNOs, but the contracted data rate is changed based on the terminal's type. In the next step, VNOs with different service sets are studied.

The data rates allocated to the aforementioned VNOs (i.e., VNO GB, BG, and BE) are plotted in Figure 6.27, when the terminal are laptops, tablets, and M2M terminals.

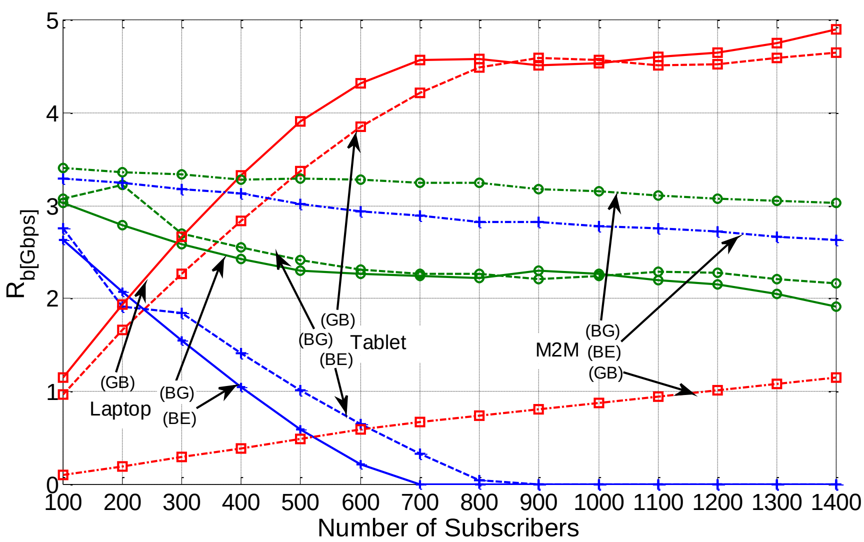


Figure 6.27 – Data rates allocated to the three VNOs with different terminal types.

The number of subscribers that can be served is highly dependent on the terminals' type (i.e., their service profile). As an evidence to this claim, consider that the M2M terminals need the lowest average data rates (i.e., 1 Mbps). The network serving M2M terminals is not only capable to serve 1 400 terminals per VNO but the VNO BE is also receiving an acceptable capacity. When changing the terminal to laptops and tables, it can be seen that resource shortage is inevitable; in these cases, VNO BE stops being served when there are more than 700 laptops or 900 tables per VNO in the network.

In conclusion, different users using different terminals require different average data rates. The VNOs based on the diversity of their subscribers' terminals have to contract different data rates. Hence, the allocated virtual radio resources to different VNOs change. Also, the number of subscribers to be served in the same physical infrastructure may change in relation to the terminals used by subscribers.

6.10.3 Influence of Service Set

In this section, the assumption is to have the four VNOs (with 1 400 subscribers) served in the best effort manner, each one of them have a different service set. The used scenario in this section contains different service sets for each of the VNOs. Each of these service sets considers services from the same service class. Based on the offered services, the four VNOs are:

- VNO Con: conversational services, i.e., VoIP, Video Calling and M2M-ITS.
- VNO Str: streaming services, including Music and Video Streamings, and M2M-SV.
- VNO Int: interactive services, i.e., FTP, Web browsing and Social Network.
- VNO Bck: background services, with email and M2M-SM, in addition to M2M-eH (the only remaining service from the interactive class).

The allocation of data rates to different services of the VNOs are depicted in Figure 6.28.

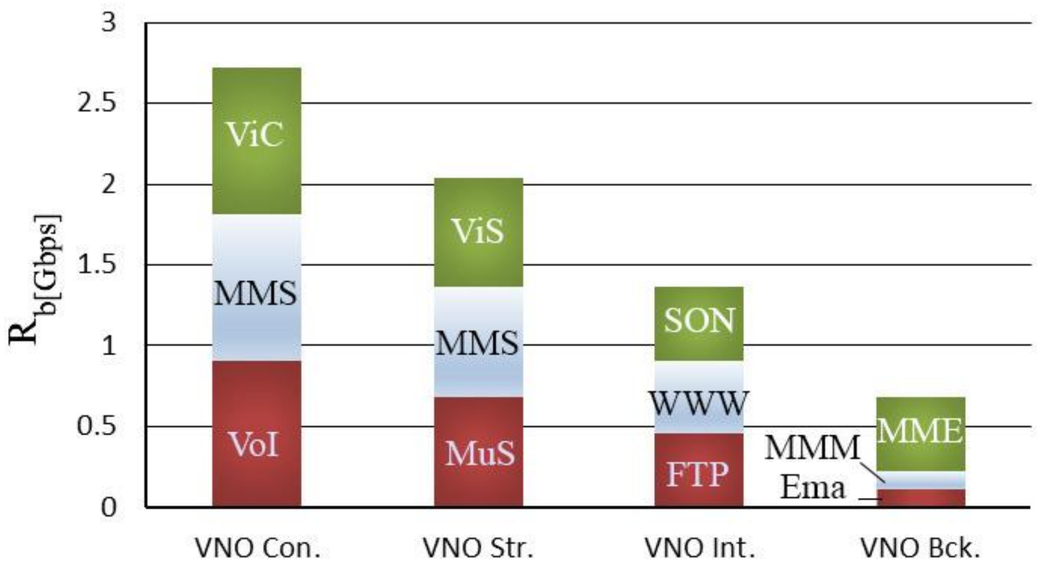


Figure 6.28 – Allocation of data rates to different types of VNOs.

As it can be observed in the figure, VNO Con. receives the biggest portion of network capacity (i.e., 2.72 Gbps) since it contains the conversational services with the highest serving weight. VNO Str. (2.04 Gbps), VNO Int. (1.36 Gbps), and VNO Bkg. (0.68 Gbps) are relatively placed in the next orders. Regarding the amount of allocation of capacity to the services of a VNO, all the services of a VNO receive the same amount of data rate, since they all have the same serving weight. However, in VNO Bkg., which has services from both background and interactive classes, M2M-eH is allocated with a higher data rate compared with the other two services, since it has a higher serving weight.

In conclusion, 40% of the network capacity is allocated to the VNO with highest serving weight, i.e., conversational services. The VNO Bck. with the lowest serving only receives 10% of the available network capacity.

6.11 Fairness in VRRM Modelling

Regarding fairness, there are two issues to be considered in the modelling of VRRM: the effect of fairness weight, W^f, and the definition of fairness in virtual radio resources management. In this section, the performance of the proposed model regarding these two issues is studied.

The effect of the fairness weight on different service classes is shown in Figure 6.29. As it is apparent from the plot, the conversational class receives 5.48 Gbps (i.e., 80.52% of the network capacity), the highest data rate when the fairness weight is comparatively small; the background class, the one with the lowest serving weight, at the same time receives only 34.07Mbps. As the fairness weight increases, the data rate allocated to the conversational class decreases to 3 Gbps, and in contrast, the background class data rate increases to 0.26 Gbps. The total allocated data rates to the service classes almost do not change when the fairness weight is chosen higher than 0.4.

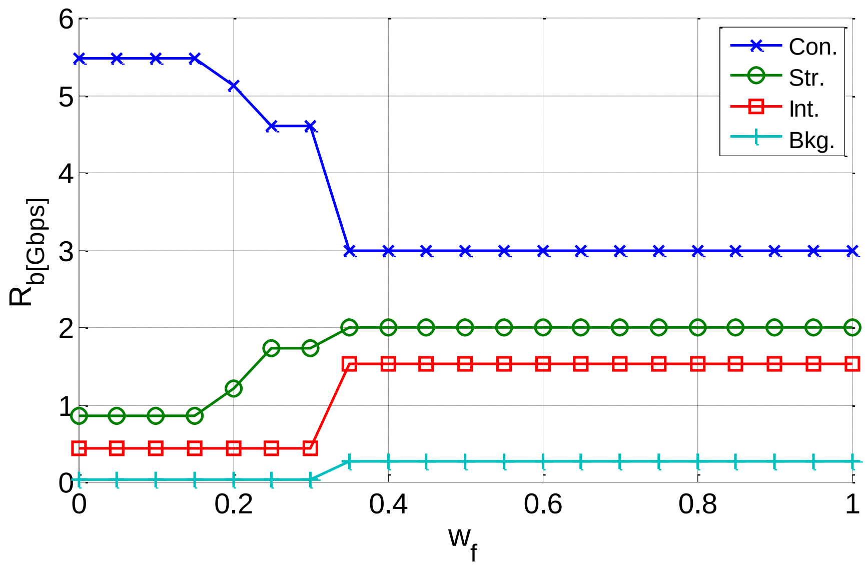


Figure 6.29 – The effect of fairness on the service classes.

Finally, the two different definitions for the fairness are studied. In Section 4.3.6, the definition of fairness is slightly changed by affecting the traffic mixture in the definition. In the new definition, the usage weight of the service is also considered in fairness measurement. In the reference scenario, VNO BE is the best case to observe the effect of the considering usage weight, since the services of the other two VNOs have a minimum guaranteed capacity. By not having any minimum guaranteed capacity, the allocation of data rates to services is done only based on the serving weights and fairness, so the service with the higher serving weights receives the bigger portion of capacity, as it can be seen in the results presented in previous sections.

Figure 6.30 compares the allocation of data rates to different services of VNO BE using the two definitions of fairness. Considering the case that the usage weight is not considered, it is obvious that services of the conversational class, the class with highest serving weight, are served with a higher data rate relative to the other service classes. Streaming, interactive, and background classes are relatively in the next positions. By adding the usage weight to the fairness equation, it can be seen that the allocation procedure is changed. Video streaming, web browsing, and social networking are services that received relatively high data rates, since they have high usage weights. On the other hand, services of the conversational class, which have lower usage weights, in this allocation procedure, are served by lower data rates. Comparing these two set of results, it can be seen that fairness with usage weight makes the allocation meaningful, since it takes the service traffic volume into account.

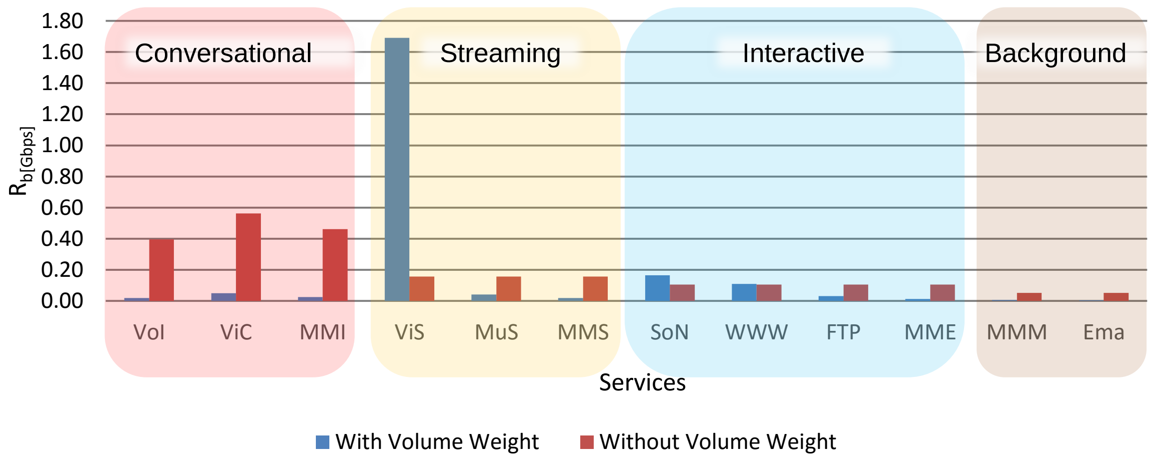


Figure 6.30 – Comparison of data rate allocation using two definitions of fairness.

In conclusion, the fairness definition and its weight can affect the distribution of available resources among the various services. Achieving the maximum level of fairness and the maximum weighted throughput are two contradicting objectives. The fairness weight has to be chosen carefully in order to achieve the best trade-off between the aforementioned objectives.

Chapter 7

Analysis of Simulation Results

This chapter is dedicated to the analysis of simulation results. Section 7.1 addresses the simulation scenario and the road map. Section 7.2 addresses the assessment procedure of the simulator. Finally, results are presented in Section 7.3.

7.1 Simulation Scenarios and Road Map

Similar to the previous chapter, an urban scenario and its variations were considered for simulations. The details of the reference scenario were provided to the OpenAir Interface Scenario Descriptor (OSD) through the XML file. Each OSD XML file is composed of five key parts, each of them defining a subset of components as follows:

- environment/system configuration, including fading and antenna,
- topology configuration, including area, MT/eNB distribution and mobility,
- application configuration, including predefined and customised traffic profiles,
- emulation configuration, including emulation time and performance metrics,
- VRRM configuration, including the number of VNOs, their requirements, etc.

The environment configuration is set for free-space propagation, considering an AWGN channel. The system's operational frequency is set to 1.9 GHz. Concerning the topology configuration, only one cell is considered, with the eNodeB located in the cell centre. MTs are randomly distributed, considering the static mobility type. The detailed specification of environment and system configuration parameters are described in Table 7.1 and Table 7.2.

Table 7.1 – Environment and system configuration parameters.

Parameter			Value
FADING	LARGE_SCALE		Urban
	FREE_SPACE_MODEL	PATHLOSS_EXPONENT	2
		PATHLOSS_0_dB	-100
	SMALL_SCALE		AWGN
WALL_PENETRATION_LOSS_dB			5
SYSTEM_BANDWIDTH_MB			7.68
SYSTEM_FREQMTNCY_GHz			1.9
ANTENNA	eNB_ANTENNA	RX_NOISE_LEVEL_dB	5
		NUMBER_OF_SECTORS	1
		BEAM_WIDTH_dB	1.13
		ANTENNA_GAIN_dBi	16
		TX_POWER_dBm	40
	MT_ANTENNA	RX_NOISE_LEVEL_dB	1
		ANTENNA_GAIN_dBi	5
		TX_POWER_dBm	20

Table 7.2 – Topology configuration parameters.

Parameter				Value
AREA	X_m			800
	Y_m			800
MOBILITY	UE_MOBILITY	RANDOM_MT_DISTRIBUTION	NUMBER_OF_NODES	10
		MT_MOBILITY_TYPE		STATIC
	eNB_MOBILITY	eNB_INITIAL_DISTRIBUTION		Fixed
		eNB_COORDINATES	POS_x	400
			POS_y	400
		RANDOM_eNB_DISTRIBUTION	NUMBER_OF_CELLS	1
		eNB_MOBILITY_TYPE		STATIC

The application configuration contains the traffic profiles to be used in the scenarios in downlink. The CBR traffic patterns were already implemented in OAI. Note that SCBR and MCBR traffic patterns are more representative of the machine-type communication, whereas BCBR captures conventional human-type communication. In addition, four traffic profiles for VoIP, web browsing, video streaming, and FTP based on Annex B were also considered. A summary of their characterisation is presented in Annex B.

Finally, the reference scenario considers two VNOs with 4 MTs each. VNO BG has 4 Mbps of guaranteed data rate, while VNO BE is served totally in a best effort manner. In addition, VNO BG has a higher serving priority, by means of having higher serving and violation weights. The serving (W^{Srv}) and violation (W^{v}) weights in VNO BG are 0.06 and 0.54, while in VNO BE they are 0.04 and 0.36.

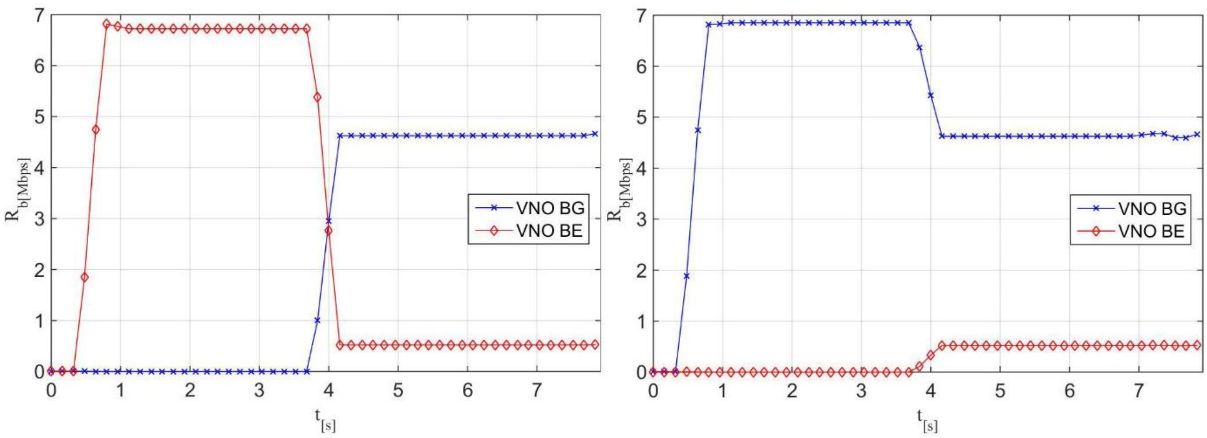
7.2 Analysis of the Results

7.2.1 Reference Scenario

The VRRM estimation of the network capacity using the general approach described in Chapter 4 is 4.17 Mbps. The calculated policy by VRRM considers 4.1 Mbps to VNO GB and the rest to VNO BE. By means of delaying the activities of one VNO, it is possible to study the effect of demand changes on VNOs with different SLAs. The scenario considers two situations:

a) Delayed VNO BG, where subscribers of VNO BG join the network with a delay, while the subscribers of VNO BE start at the beginning of the simulation,
b) Delayed VNO BE is the opposite of the former situation, where the activities of VNO BE's MTs are delayed.

The calculated policies are sent to OAI, and Figure 7.1 illustrates the data rate allocated to each of the VNOs during simulation.

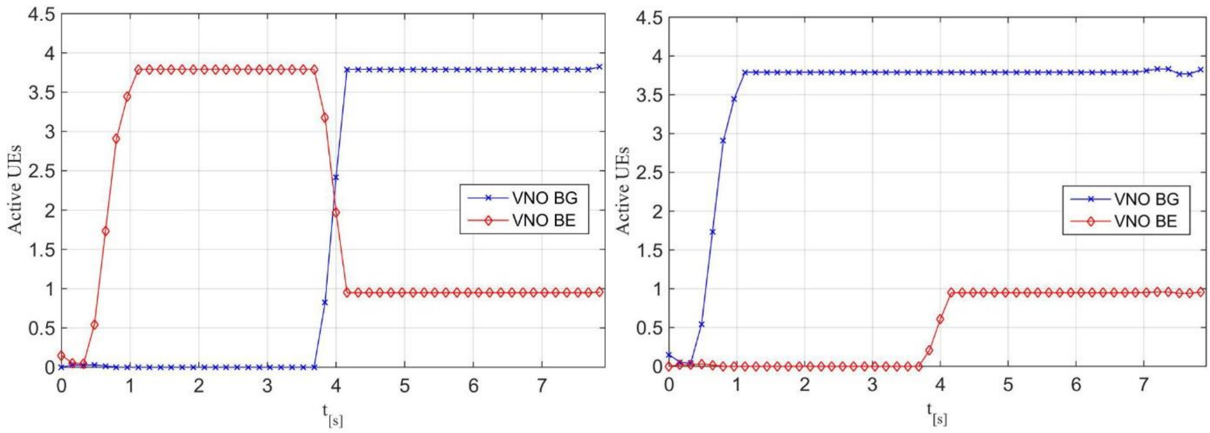


a) Delayed VNO BG b) Delayed VNO BE

Figure 7.1 – The allocated data rates to the VNOs in the reference scenario.

At the beginning of simulation, all of the network capacity is assigned to the active VNOs, since the other users are not requesting any. However, the allocated data rate reduces as the other VNOs start their activity. For delayed VNO BG, it can be seen that the data rate allocated to VNO BE reduces from 6.72 Mbps to 7.7% of its initial value (i.e., 0.52 Mbps) as VNO BG starts. VNO BG's data rate reaches 4.62 Mbps, however, when the delayed BE situation is studied, the reduction of data rate for VNO BG is 32.5% of the initial value.

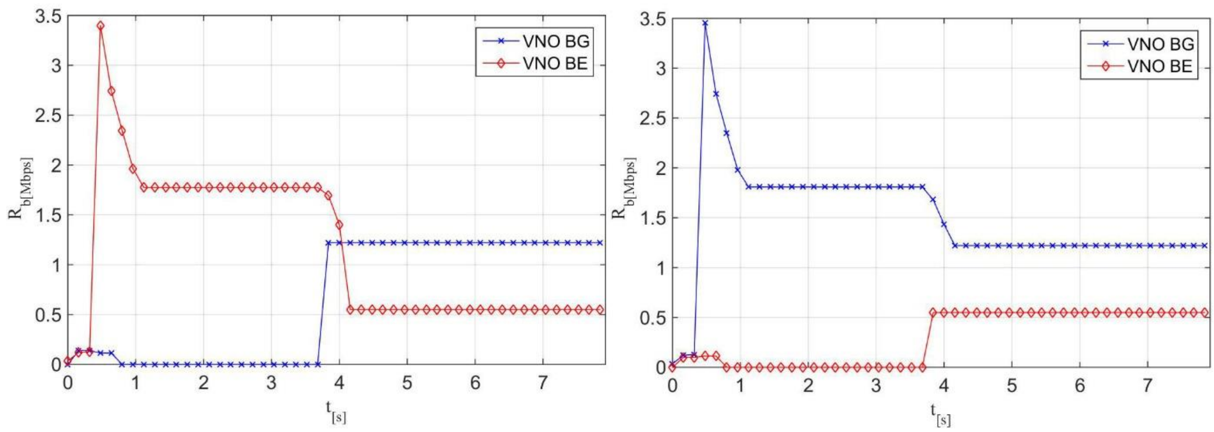
The number of active MTs for each of the VNOs is illustrated in Figure 7.2. On average, 3.7 MTs are being served when one of the VNOs is active. However, the number of active MTs of VNO BG is not affected by the activity of the other MTs. In contrast, the average number of active MTs of VNO BE decreases to 0.94 Mbps, equivalent to 25.4% of its original value.



a) Delayed VNO BG b) Delayed VNO BE

Figure 7.2 – The number of active MTs in the reference scenario.

Finally, the allocated data rate per active MT for each of the VNOs is shown in Figure 7.3. Each active MT of VNO BE when there is no activity from other VNOs gets 1.7 Mbps, but this value then decreases to 0.54 Mbps (31.7% of its initial value). As expected, the reduction of the same value in VNO BG is only 0.6 Mbps, roughly 33.3% of the original value. In addition, it is interesting to point out that the total network data rate reduces as the average number of active MTs increases.



a) Delayed VNO BG b) Delayed VNO BE

Figure 7.3 – The data rate per active MT in the reference scenario.

7.2.2 Different Serving Weight

In order to show the effect of the serving weights (i.e., the prioritising the VNOs), the minimum guaranteed data rate of VNO BG is reduced to zero. In other words, the scenario has two best effort VNOs with different serving weights. The model allocated 60% of network capacity to VNO BG with serving weight of 0.6, and the rest to VNO BE. Figure 7.4 shows the data rate of VNOs based on their serving weights. According to results, 4.58 Mbps (89.8% of the network capacity) is allocated to the VNO with the higher serving weight, while the other one receives only 0.52 Mbps.

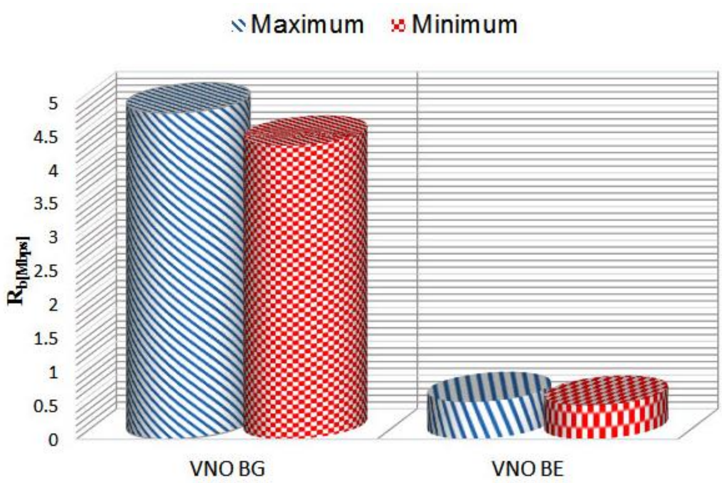


Figure 7.4 – The allocated data rates to the VNOs with different serving weights.

The number of active MTs per VNO for this scenario is shown in Figure 7.5. There are 5 active MTs, where 4 of them belong to the VNO with the higher serving weight, and the remaining one is from VNO BE with the serving weight of 0.4.

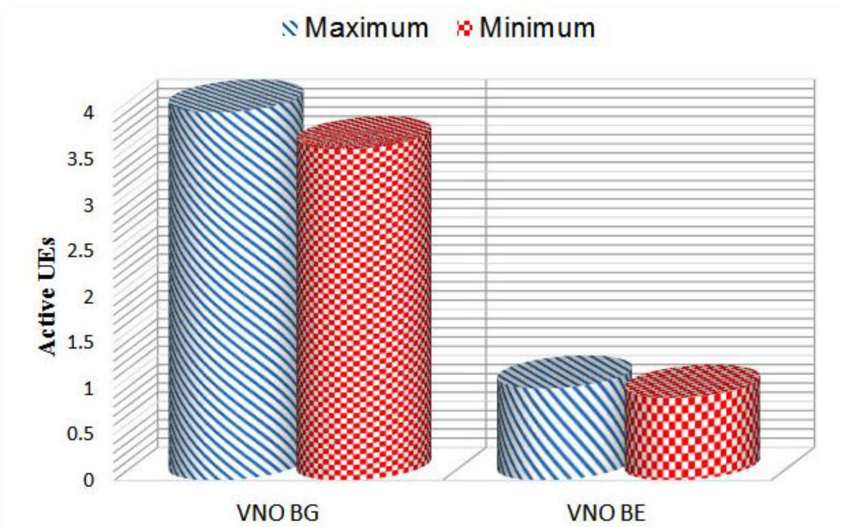


Figure 7.5 – The number of active MTs with different serving weights.

Regarding the data rate per MT, Figure 7.6 presents the average data rate per active MT for each VNO. The average data rate for active MTs in the VNO with higher serving weight is 1.2 Mbps, about 2.4 times higher than the VNO with low serving weight. These results clearly demonstrate the effect of the serving weight, which is prioritising the VNOs, in practice.

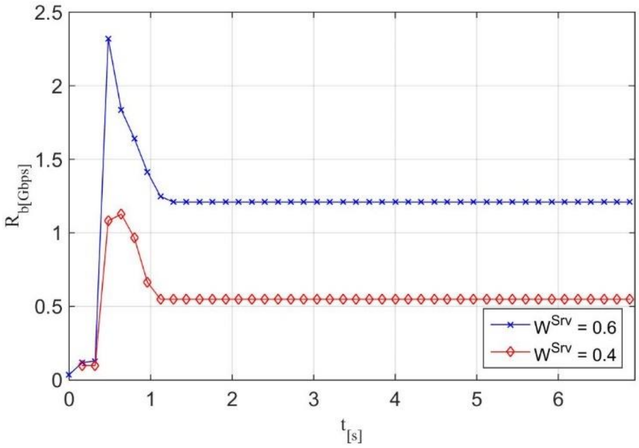
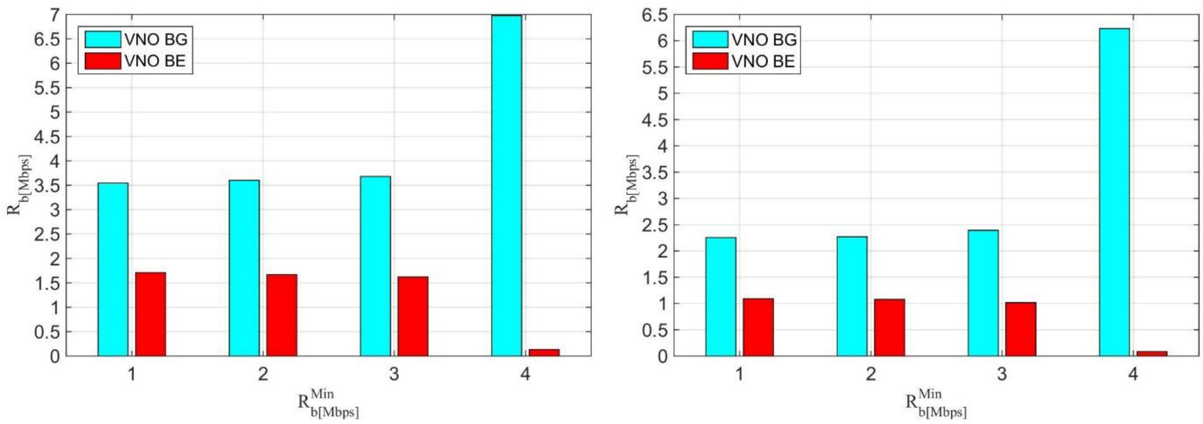


Figure 7.6 – The data rate per active MTs in VNOs with different serving weights.

7.2.3 Different Guaranteed Data Rates

Based on the reference scenario, two VNOs of BG and BE are considered. However, the minimum guaranteed data rate of VNO BG is chosen to be 1, 2, 3, and 4 Gbps. The aim is to study the effect of guaranteed data rate on VNOs and the network throughput. Figure 7.7 illustrates the data rate allocated to each of these VNOs during simulation. The data rates allocated to VNO BG, the VNO with the highest priority, are always higher than the ones of VNO BE. For a minimum guaranteed data rate of 4 Gbps, the allocated data rate of VNO BG reaches the maximum of 6.92 Gbps, while only 0.14 Gbps is allocated to VNO BE. The model estimation of network capacity using the realistic approach is 6.93 Gbps, hence, the model allocates 5.76 Gbps to VNO BG and 1.17 Gbps to VNO BE, therefore, almost all RBs are pre-allocated to VNO BG.

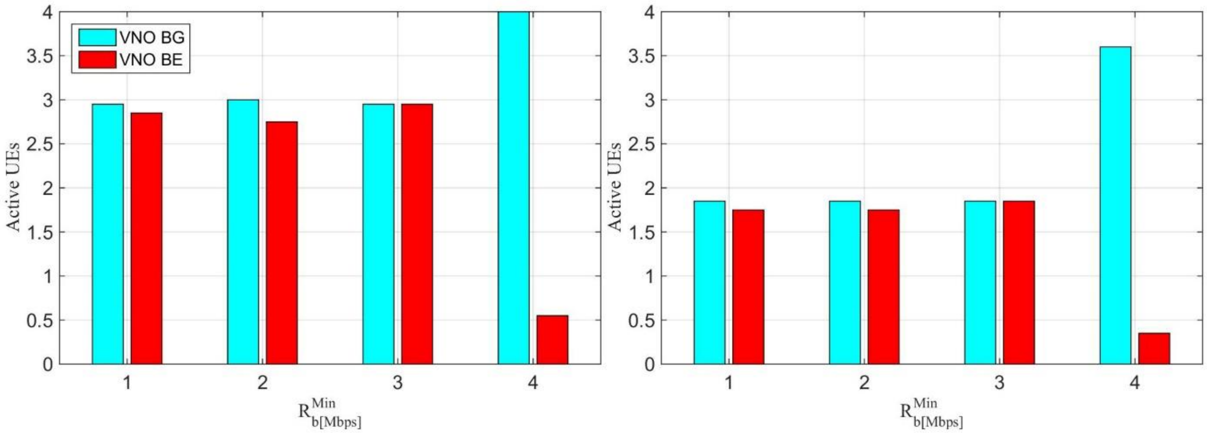


a) Maximum

b) Minimum

Figure 7.7 – Allocated data rates for different guaranteed data rates.

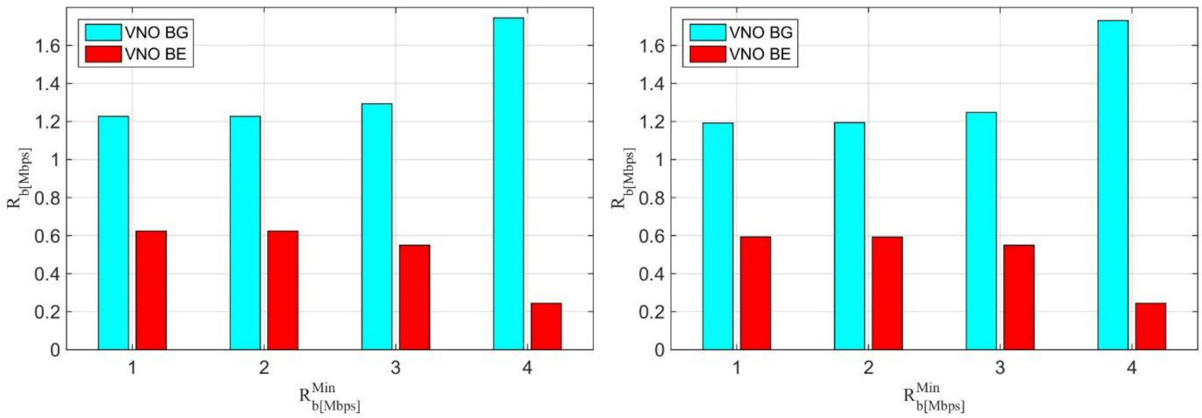
It is worth noting that the practical network data rate for 8 MTs, according to the results presented in the previous section, is 5.76 Gbps. However, the result present in Figure 7.8 indicates a higher value, which also varies by changing the guaranteed data rates. The reason for this behaviour is the number of active MTs presented in Figure 7.8. Although there are 8 terminals in this scenario, only 5 MTs are simultaneously active. The reason for having a higher value for the minimum guaranteed data rate of 4 Gbps is also the number of active MTs. While 5 MTs are served simultaneously in other cases, it decreases to 4 MTs as the guaranteed data rate increases to 4 Gbps.



a) Maximum b) Minimum

Figure 7.8 – Number of active MTs for different guaranteed data rates.

The data rates allocated to each active terminal in both VNOs are presented in Figure 7.9. As expected, the highest data rate per MT belongs to VNO BG, with 4 Gbps guaranteed data rate. As the average data rates of the terminals in VNO BG decrease, the subscribers of VNO BE receive higher data rates.



a) Maximum b) Minimum

Figure 7.9 – Allocate data rate per active MT for different guaranteed data rates.

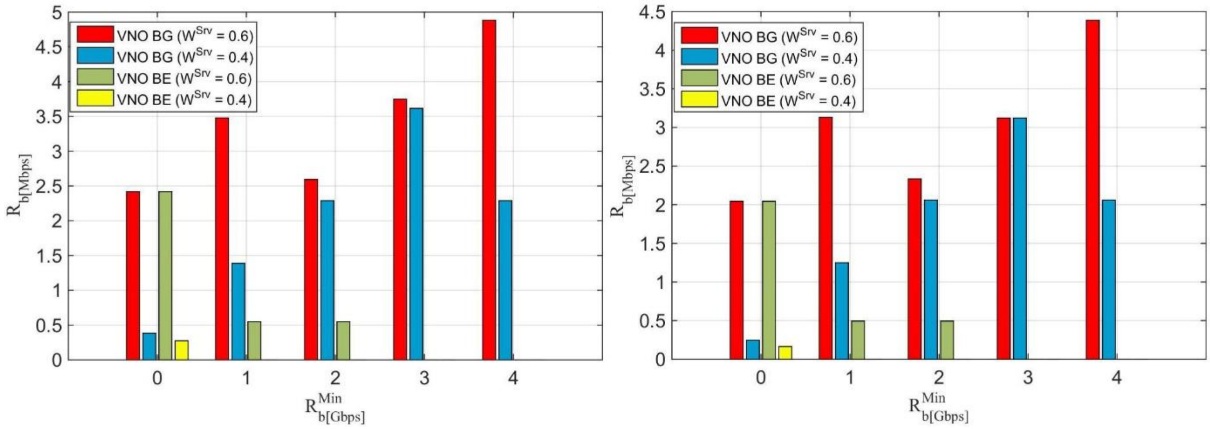
7.2.4 4 VNO with different SLAs

The reference scenario includes two VNOs with different SLAs. The effect of serving and violation weights is studied in Section 7.2.2. In the next step, the effect of minimum guaranteed data rate is considered in Section 7.2.3. The goal of this section is to study the effect of the serving weight and SLAs together. Hence, the number of VNOs is increased to 4 VNOs, where two of them are served totally in a best effort way and the other two have a best effort with minimum guaranteed data rate SLA. In addition, two of the VNOs (one BG and one BE) have serving weight of 0.6, while the other two have 0.4. The minimum guaranteed data rates for BG VNOs also changes from 4 Mbps down to 0 Mbps (i.e., 4 BE VNOs). The policies by the VRRM server for each SLA and VNO are presented in Table 7.3.

Table 7.3 – The VNOs' allocation policies issued by VRRM.

$R_{b[\text{Mbps}]}^{min}$	VNO BG (W^{Srv}=0.6)	VNO BG (W^{Srv}=0.4)	VNO BE (W^{Srv}=0.6)	VNO BE (W^{Srv}=0.4)
0	2.08	1.39	2.08	1.39
1	2.48	1.99	1.48	0.99
2	2.88	2.59	0.88	0.59
3	3.28	3.19	0.28	0.19
4	4.00	2.93	0.00	0.00

The aforementioned policies are applied to the scheduler and numeric results were generated. Figure 7.10 presents the minimum and maximum data rates allocated to each of the VNOs. It is apparent that the scheduler manages to follow the provided policies. When the minimum guaranteed data rates for the BG VNOs are set to zero, the VNOs with the higher serving weights (W^{Srv}=0.6), receive equal data rates, between 2 Mbps and 2.42 Mbps. The other two VNOs receive the remaining capacity, which is less than 0.5 Mbps. By increasing the guaranteed data rates to 1 Mbps, the VNO BE with a serving weight of 0.4, is not allocated any capacity in return for allocating minimum data rates to VNO BGs. Best effort VNOs stop receiving any data rates by increasing the guaranteed data rate to 3 Mbps. Network capacity in this situation is estimate to be about 6 Mbps, which means that all available resources have to be allocated to the BG VNOs to meet the guaranteed data rates. These VNOs are always allocated the minimum guaranteed data rates, but the other two VNOs are not getting any resources. Finally, the minimum guaranteed data rates increase to 4 Mbps per VNO. Obviously, there are not enough available resources in the network, hence, the VRRM model detects the situation and tries to minimise the violation to the SLAs. According to Table 7.3, the model allocates 4 Mbps, 57% of the available resources, to VNO BG with a serving weight of 0.6. The remaining resources, offering a data rate of 2.93 Mbps, are allocated to the other guaranteed VNO. Figure 7.10 confirms that the same policy is put into practice by the scheduler in OAI.



a) Maximum b) Minimum

Figure 7.10 – The allocated data rates to the VNOs with different SLAs.

Finally, the maximum number of active MTs for VNO in each of these case studies is presented in Figure 7.11. The VNOs with guaranteed SLAs have a maximum of 2 active MTs. The maximum number of active MTs for the best effort VNO with a higher serving weight reaches 1 active MT, while for the best effort VNO, the average active MTs in the observation period never passes 0.5. The subscribers of this VNO are not served when the guaranteed data rates for the other VNOs is requested.

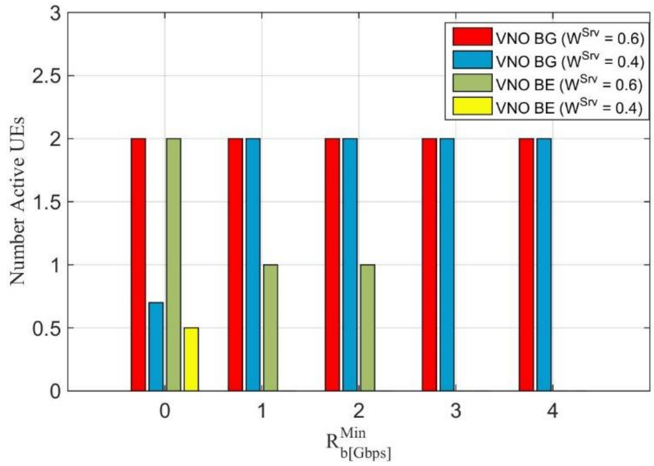


Figure 7.11 – The maximum active MTs in VNOs with different SLAs.

7.2.5 Different Traffic Profiles

In the final step, the traffic profiles, plotted in Figure 5.22 and Figure 5.23, are used for users in the reference scenario, when there are two VNOs and 8 MTs. Figure 7.12 presents the average and standard deviation of the allocated data rates for each VNO. The demanded data rates for the majority of the services are lower than network capacity. As evidence to this claim, VoIP traffic uses only 5.5% of the total network capacity. However, for traffic profiles like BCBR and HTTP, the allocated data rate for both VNOs reaches up to the total network capacity. Based on results, it can be claimed that the minimum guaranteed data rate of 4 Mbps for VNO BG is always met. However, there are situations that the demanded traffic is much lower than the guaranteed data rate.

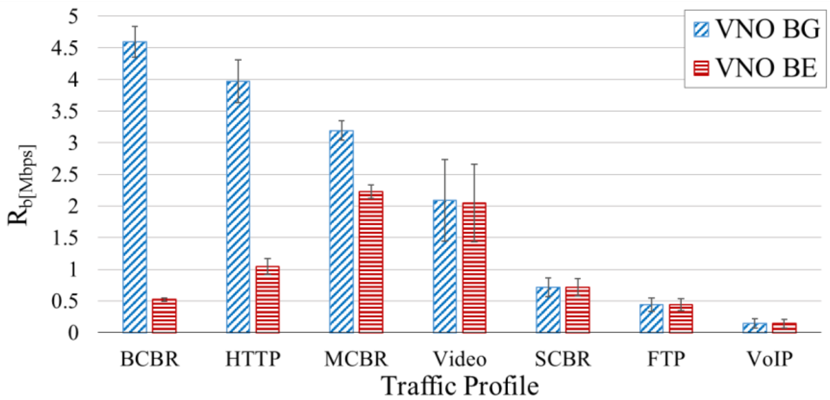


Figure 7.12 – Average data rates allocated to the VNOs with different traffic profiles.

Finally, Figure 7.13 presents the average and standard deviation of active MTs for each VNO and traffic profile. The results for active MTs also confirm the claim of low traffic demands. However, the priority of

VNO BG is apparent through the plot. As evidence to the claim, the MTs of VNO BG with the BCBR traffic profile are all served, while VNO BE is allocated enough resources for just 25% of its demand.

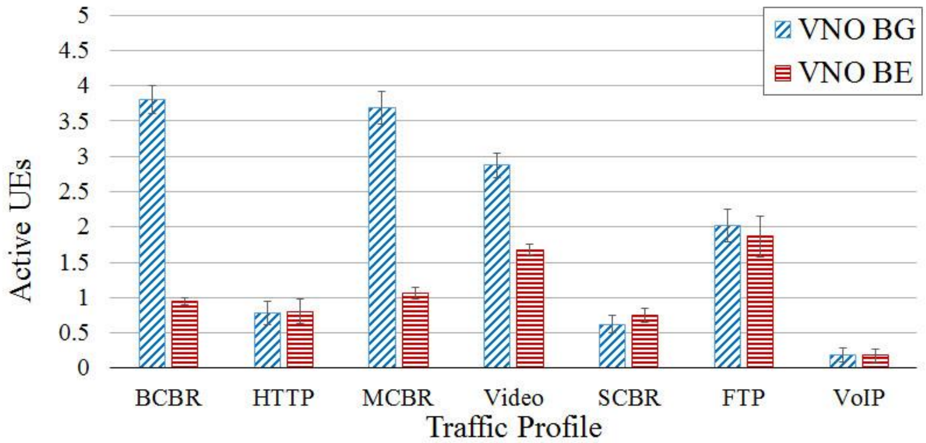


Figure 7.13 – Average active MTs of the VNOs with different traffic profiles.

Chapter 8

Conclusions

This chapter aims at presenting the main conclusions of this thesis, highlighting major results, and suggesting for further work in future. Section 8.1 presents a summary of the thesis. Section 8.2 presents the main results including the novelty of the work developed within this thesis. Section 8.3 points out aspects to be addressed in future work.

8.1 Overview of Thesis

This thesis is organised in 8 chapters, starting with the introduction to the thesis in Chapter 1. This chapter provides the motivation and objectives in addition to the novel aspects and concepts. The final section of this first chapter concerns the research strategy, European projects contributions, and publications.

Chapter 2 addresses the fundamentals of the thesis, with an overview on different RATs and their specifications, in addition to a brief discussion on common wireless network architectures. The chapter concludes by a short introduction to QoS and QoE, in addition to QoS traffic classes in different networks.

Chapter 3 presents the state of the art and related topics. This chapter starts with RAN sharing, by presenting its concept, advantages, and drawbacks. It continues by presenting the virtualisation of wireless networks, discussing about motivation and concept in addition to presenting related works in RAN virtualisation. The chapter concludes with a very brief review on resource management algorithms.

Chapter 4 presents the models and algorithms for the management of virtual radio resources. In the first step, the architecture of a V-RAN, based on C-RAN, is proposed, as well as the adaptation to other architectures, such as NFV and MCN architectures. An analytical model for the management of virtual radio resources is developed, with two main components: estimation of the available radio resources and allocation of virtual resources. In the first part, the model maps the number of available RRUs in all systems onto global network data rate by using probabilistic functions. In addition to the general approach for the estimation of the network capacity, three other approaches are also developed, i.e., pessimistic, realistic, and optimistic ones; each of these approaches estimates the total network capacity with different assumptions on terminals' channel quality. After having a realistic estimation of network capacity, the allocation of virtual resources is formed into an LP-optimisation problem. The model aims at maximising the weighted throughput of the network, subject to constraints: the total network capacity, minimum/maximum guaranteed data rates, and fairness. When there are not enough resources to meet the minimum guaranteed data rates (i.e., the optimisation problem becomes infeasible), a new optimisation problem is formed, in which violations of minimum guaranteed data rates are added. The goal is to maximise the weighted network throughput, while minimising the weighted summation of violations. Fairness in the allocation of resources to different services of the VNOs is also adopted in the proposed model. Finally, the partial VRRM algorithm is described in addition to evaluation metrics.

Chapter 5 discusses the implementation of the model and algorithms developed in the thesis. The implementation procedure has two main parts, which are the implementation of the VRRM server and the modification of OAI. The VRRM model was developed in a Windows-based environment, called VRRM server, while OAI is used as the realistic infrastructure emulator. There is a TCP bidirectional link between OAI and VRRM server, which is used for exchange of scenario parameters, policies, and reports.

Chapter 6 presents the used reference scenarios in addition to theoretical numeric results achieved by using the model developed in this thesis. Numeric results shows that using the concept of virtualisation and the proposed model, multiple VNOs with different SLAs and requirements are served on the same physical infrastructure. Through numeric studies, the performance of the VRRM model is evaluated.

Chapter 7 presents the results and analysis from the developed simulators. Based on the studies conducted in the former chapter, a set of practical scenarios (with respect to simulator’s limitation) was established. The realisation of the concept and objectives of virtualisation of radio resources are demonstrated through numeric results.

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

8.2 Main Results

This thesis presents the concept of virtualisation of radio resources as the final step towards an end-to-end virtual network by realising a virtual wireless link. A virtual RAN, as it is presented in this thesis, is a RAN that operates over cloud data centres (known as C-RAN) as the physical infrastructure. All physical radio resources are aggregated and there should be a central management to offer VNOs Capacity-as-a-Service. In this solution, VNOs no longer have to deal with the management of physical infrastructure, but rather to ask for capacity in order to serve their subscribers. By means of virtualisation of radio resources, it is possible to share the physical infrastructure of both cellular networks and WLANs among multiple VNOs. In this approach, VNOs are provided on-demand wireless capacity, while they are offered isolation, element abstraction, and multi-RAT support. The service offered to VNOs is referred to as Connectivity-as-a-Service.

In addition, a comprehensive analytical model for the management of virtual radio resources. with the capability of supporting various situations, is proposed. The model takes a number of available RRUs in different RATs as input, and maps them onto the total network capacity. The model is also able to consider multiple channel quality assumptions for the terminals through different estimation approaches. The allocation of resources to maximise the weighted throughput of the network based on the estimated network capacity is the next step. The serving weights in the objective function make possible to prioritise services. The model considers the ability of offloading traffic to WLANs as well as to cellular networks. In contrast to cellular networks, increasing the number of terminals in WLANs increases the collision rates and reduces network throughput, hence, the guideline of the model is to serve services with low average data rate per session from cellular networks and the others from WLANs, in order to reduce collision rates. Resource allocation in the case of shortage of resources (i.e., when there are not enough resources to meet the minimum guaranteed data rates) tries to minimise the violation to the guaranteed data rates.

In addition, the model also considers fairness among services. Although allocating all the resources to services with the highest serving weights leads to the weighted throughput (i.e., the primary goal), it does not satisfy fairness requirements. A fair resource allocation is when the summation of the deviation from the weighted average for all services is as small as possible. The ideal case is zero, but it may compromise the main objective, hence, there is always a trade-off between fairness and network throughput. When the model fails to find any feasible solution to serve all guaranteed data rates, due to the shortage of resources, it introduces violations to guaranteed data rates. This approach changes the former infeasible solution to a feasible one, and the model aims at minimising the summation of weighted violations. This way, the services with less importance are facing the violation of guaranteed data rates, while the more important ones are served properly.

Based on the outputs, it can be seen that services are served relatively to their input serving weights. Although there are services with different weights, due to the fairness technique, all of them receive a portion of the resources. A V-RAN is capable of offering up to 6% more data rate than RAN sharing, with the same probability conditions. Results indicate that, to cover the serving area with the mentioned number of VNOs, subscribers, and SLAs, at least 11 OFDMA cells are required. The reference scenario assumes 16 cells, where the total network capacity is estimate as 1.76 Gbps. By reducing the number of cells to 5, the total network capacity shrinks to almost half of its initial value (i.e., 48% of the reference case). It is shown that by adding an AP to each OFDMA cell, the network capacity increases up to 2.8 times. As the result traffic offloading, the VRRM model is able to properly serve not only guaranteed services adequately, but best effort ones are allocated with a relatively high data rate. Furthermore, the results show that the network capacity increases from 0.9 Gbps in the pessimistic approach to 5.5 Gbps in the optimistic one. The effect of these capacity changes on the allocated data rates to different VNOs and their services classes are presented through a series of plots.

Results also show that the changes in network capacity mostly influence the VNO with best effort services, while the other types of VNOs suffer relatively less from the reduction of resources. Among guaranteed services, the violation starts from the service(s) with the lowest violation weight, which in our case study are background ones. The numeric results justify that the model is able to prioritise the service according to their serving and violation weights. In the worst case studied in this thesis, background and interactive services are totally shutdown, while conversational ones experience no violation, as expected. It is shown that when there is enough capacity, not only the guaranteed VNO is satisfied, but also best effort VNOs are also well served. However, as the network capacity decreases due the channel quality (general and pessimistic approaches), best effort VNOs are effected more than the guaranteed one. The same situation is shown at the service class level too. Conversational and streaming classes are the ones with the highest serving priority (i.e., serving weights), being generally allocated with data rates higher than the other two service classes. When there is a resources shortage, i.e., in the general and pessimistic approaches, violations start relatively by background and interactive classes.

In addition, the model performance under different loads is evaluated. Results confirm that the model is able to realise an acceptable level of isolation. When there is a shortage of radio resources, the model for resource management starts violating the guaranteed levels of services with lower serving weights, which are the background and interactive services. As evidence to this claim, the guaranteed data rate VNO, and particularly its conversational class, can be considered where the requested service quality regardless of network situation is always offered.

The flexibility of the applied model makes the equilibrium among different services and different VNOs possible. In conclusion, it is shown that our model achieves the desired goals:

- on-demand wireless capacity offering, by serving three VNOs with different SLAs;
- isolation, by considering changes in networks into almost one tenth of its original value (from optimistic to pessimistic), in addition to changing the demand from 300 users per VNO to 1400;
- element abstraction and multi-RAT support, by providing wireless connectivity using both cellular networks and WLANs, while VNOs do not have to deal with the details.

Furthermore, the aforementioned concept and model were implement in OAI, a software-based LTE eNodeB, as the physical emulator. Various changes have been introduced to OAI, including new sets of variables and structure. The implementation guideline is to create multiple groups of subscribers, add support for different policies for each group in the scheduler, and create the communication links to/from the VRRM server.

Results show the proposed approaches offer almost the same capacity to guaranteed VNOs regardless of the other VNOs (i.e., it experiences only 32.5% of its initial allocated data rate at worst without violation to guaranteed data rate), while the allocated data rate to best effort VNOs may decrease to 7% of its initial value. The performance of the network in different situations, such as different guaranteed level, different serving weight, and different services is studied.

8.3 Novelty and Key Contributions

This thesis claim novelty in proposing the concept of virtualisation of radio resources to achieve virtual wireless links and to have end-to-end virtual networks, by aggregating all physical resources from different RATs, in order to offer VNOs with a more efficient wireless connectivity. In this novel methodology, VNOs ask for wireless capacity from a set of physical network providers to serve their subscribers, not having to deal with the physical infrastructure at all. The RAN Provider, owning the physical infrastructure, is capable of offering Capacity-as-a-Service to VNOs. The advantages of RAN virtualisation compared to RAN sharing (where each operator is allocated a portion of spectrum) comes from network element abstraction, isolation among virtual instances, and the ability to support multi-RATs.

In addition to the proposition of the aforementioned concept, the proposal of an analytical model for the management of virtual radio resources is another novelty of this thesis. For a network with multiple RATs, such as GSM, UMTS, and LTE, the model is capable of estimating the overall network capacity based on a given number of the available RRUs from each RAT. It also shares the available capacity among the different services of the VNOs, the allocation being based on the VNOs' SLAs, in which VNOs may be guaranteed with a minimum as well as a maximum capacity per service, or simply served in a best effort approach. The presented VRRM model satisfies SLAs when there are enough RRUs, and minimises SLAs violations in resource shortage cases; in both cases, fairness of resource allocation is considered. In addition to the proposition of the novel VRRM model, an architecture for a V-RAN based on a C-RAN infrastructure, with its required modifications to support virtualisation of radio resources, is briefly address.

Another novelty of the current work is a model for the management of virtual radio resources that is capable of supporting both non-cellular (e.g., Wi-Fi) and cellular (e.g., GSM, UMTS, LTE, and whatever comes next in 5G) networks. Moreover, the network capacity estimation technique is considered to have three approaches, i.e., optimistic, realistic, and pessimistic ones, which approximates the model to a real network operation. It also shares the available capacity among the different services of the VNOs, the allocation being based on the VNOs' SLAs, in which VNOs may be guaranteed with a minimum as well as a maximum capacity per service, or simply served in a best effort approach. The presented VRRM model satisfies SLAs when there are enough RRUs, and minimises SLAs violations in resource shortage cases; in both cases, fairness of resource allocation is considered.

Finally, the proposed concept of virtualisation of radio resources was implemented in the Open Air Interface (OAI), a software-based LTE eNodeB developed in Linux used as infrastructure emulator. Based on the development done in OAI for supporting this concept, it can be claimed that the proposed concept and model are eventually put into practice. It is shown through a realistic emulation that it is possible to serve multiple VNOs over the same physical infrastructure.

8.4 Future Work

The work presented in this thesis can be followed on, by exploring several other topics that can be investigated in the future. Examples of these topics are proposed below.

Considering the recent research towards 5G, the virtualisation of radio resources has to be added to the new architecture of 5G mobile network. In addition, the algorithms and approaches conducted for VRRM have to revisited for the 5G usage.

One possibility for extension is considering the short-term traffic demands changes. In this thesis, operators have an estimation of the average data rate per terminal and they know the traffic mixture. Hence, they set their provisioning policies and define/update their policies with VRRM and infrastructure provider. However, knowing a realistic prediction of traffic demands in a heterogeneous access environment and the effect of serving a service through different RATs on the traffic demands enables the VRRM model to issue more accurate and practical policies. The extension of the model first has to estimate the probability of demanding different amounts of data rate per service of VNO in addition to the costs of over- and understocking. Then, it can optimise the allocation of data rates to each of the services.

In addition, considering the management of virtual radio resources with a variable cost is an interesting and practical research topic. The development of an economical model for charging VNOs, based on the available resources and demands, is very valuable. Obviously, handling traffic in the busy hours of the network costs more comparing to its low load period. The VRRM model can be extended in order to take cost into consideration, as well as the VNOs requirement. New scenarios and used cases can be considered for the aforementioned problem.

Finally, the optimisation of the CRRM and LRRMs algorithm to work in a virtualised environment could be an interesting research topic. Although, these entities were the focus of recent studies, they are not optimised in the presence of VRRM and multiple group of MTs. An interesting research topic is to redesign these algorithms to have more efficient virtual network.

Annex A

System SINR and Data Rate Model

This annex presents the data rate models for EDGE, HSPA+ and LTE. These models are referred to the throughput over the physical layer standardised on OSI reference model. In addition, the model for the different branches of all functions do not take into consideration all the necessary throughput reductions, such as overheads loads and BLER to obtain the throughput at a higher level/layer, or up into the application level.

A.1 GSM/EDGE

Authors in [Kris02] proposed a formula for data rates in GSM/EDGE as function of SINR for downlink and uplink as follows:

$$R_{b_{EDGE}[\text{kbps}]} = \frac{A}{1 + e^{-\lambda(\rho_{[\text{dB}]} - \delta)}} \tag{A.1}$$

where:

- $R_{b_{EDGE}}$: data rate for a single time-slot in EDGE;
- ρ: SINR;
- λ, A, δ: parameters for sigmoid modelling related to coding schemes and is selected from Table A.1.

Table A.1 – Parameters for sigmoid modelling of EDGE (extracted from [Kris02])

MCS	A	δ	λ
MCS-9A	59.2	20	0.338
MCS-8A	54.4	17	0.367
MCS-7B	44.8	15	0.447
MCS-6A	29.6	10	0.451
MCS-5B	22.4	8	0.337
MCS-4C	17.6	11	0.379
MCS-3A	14.8	8	0.337
MCS-2B	11.2	4	0.357
MCS-1C	8.8	3	0.454

In addition to data rate as function of SINR, SINR as the function of data rate in EDGE is also required. The SINR as the function of data rate and also reverse function of (A.1) is as follows:

$$\rho_{[\text{dB}]} = -{}^{1}/_{\lambda} \ln\left(\frac{A}{R_{b_{EDGE}[\text{kbps}]}} - 1\right) + \delta \tag{A.2}$$

where:

- $R_{b_{EDGE}}$: data rate for a single time-slot in EDGE;
- ρ: SINR;
- λ, A, δ: parameters for sigmoid modelling related to coding schemes and is selected from Table A.1.

For the sake of simplicity, the coding mode of MSC-9A is chosen to present the performance of GSM/EDGE.

A.2 UMTS/HSPA+

A.2.1 Downlink

UMTS/HSPA+ has multiple configuration of MIMO and modulation, each has different performance in different channel quality. Figure A.1 shows throughput as function of SINR for these configuration. For the sake of simplicity, however, MIMO 2×2 configuration with 64-QAM is considered as UMTS configuration in models and simulations.

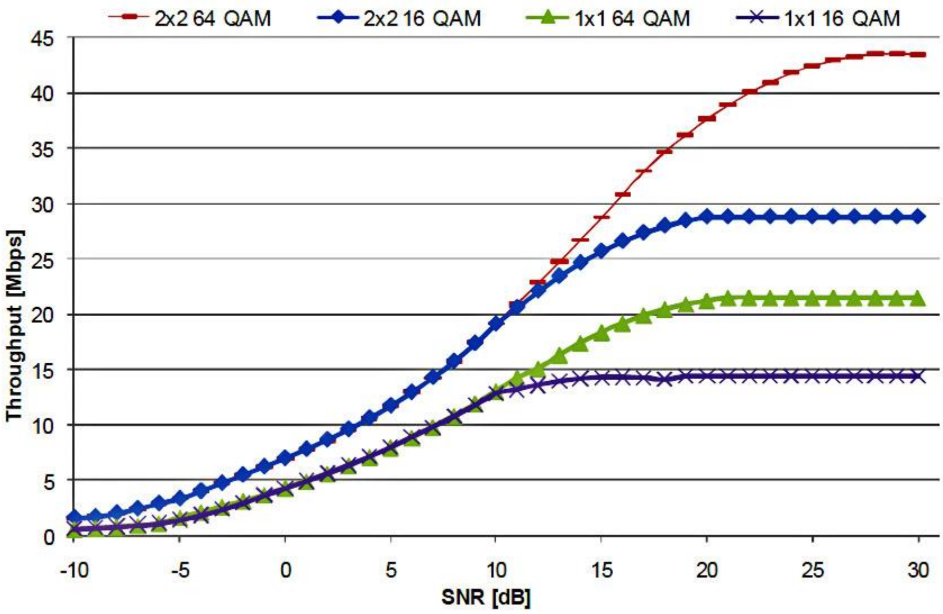


Figure A.1 – HSPA+ downlink throughput as function of SINR (extracted from [Jaci09]).

In [Jaci09], auther presents the relation of throughput as function of SINR in form of polynomial as follows:

$$R_{b_{UMTS}[\mathrm{Mbps}]} = 10^{-3} \times \sum_{i=0}^{4} a_i \rho_{[\mathrm{dB}]}^{i} \tag{A.3}$$

where:

- $R_{b_{UMTS}}$: data rate for a single code in downlink;
- ρ: SINR;
- a_i: factors of polynomial which is selected from Table A.2.

Table A.2 – Polynomial factors of UMTS downlink throughput.

a_4	a_3	a_2	a_1	a_0	Condition
0	-8.3	-135.7	-213.1	4805.7	$-10 \leq \rho_{[\mathrm{dB}]} < -6$
0.5	1.8	8.9	781.2	7078.4	$-6 \leq \rho_{[\mathrm{dB}]} < 1$
0	-0.1	65.7	579.2	7211	$1 \leq \rho_{[\mathrm{dB}]} < 6$
0	-0.8	59.3	804.6	6047.2	$6 \leq \rho_{[\mathrm{dB}]} < 17$
0	0	-75.7	4366.1	-19392	$17 \leq \rho_{[\mathrm{dB}]} < 30$

In addition to throughput as function of SINR, [Jaci09] describes the inverse function, SINR function of throughput, for MIMO 2×2 configuration with 64-QAM

$$\rho_{[\mathrm{dB}]} = \sum_{i=1}^{6} b_i R^{i}_{b_{UMTS}[\mathrm{Mbps}]} \tag{A.4}$$

where

- $R_{b_{UMTS}}$: data rate for a single code in downlink;
- ρ: SINR;
- b_i: selected base the data rate from the Table A.3.

The SINR as function of throughput for various configuration of HSPA+ in downlink is plotted in Figure A.2.

Table A.3 – Polynomial factors of UMTS downlink SINR.

b_6	b_5	b_4	b_3	b_2	b_1	b_0	Condition
-0.08	1.54	-14.34	69.41	-184.00	255.38	-154.75	$1.7 \leq R_{b[\mathrm{Mbps}]} < 3.5$
0	0	-0.02	0.52	-4.79	20.22	-37.28	$3.5 \leq R_{b[\mathrm{Mbps}]} < 6.4$
0	0	-0.02	0.52	-0.06	2.11	-14.02	$6.4 \leq R_{b[\mathrm{Mbps}]} < 7$
0	0	0	0	-0.08	2.46	-13.21	$7 \leq R_{b[\mathrm{Mbps}]} < 7.8$
0	0	0	-0.09	2.50	-21.19	58.00	$7.8 \leq R_{b[\mathrm{Mbps}]} < 9.5$
0	0	0	0	0	0.86	-5.18	$9.5 \leq R_{b[\mathrm{Mbps}]} < 14.1$
0	0	0	0	0	0.73	-2.43	$14.1 \leq R_{b[\mathrm{Mbps}]} < 34.5$
0	0	0	0	0.05	-2.88	60.00	$34.5 \leq R_{b[\mathrm{Mbps}]} < 42.5$
0	0	0	0	0.30	-21.91	417.40	$42.5 \leq R_{b[\mathrm{Mbps}]} < 43.2$

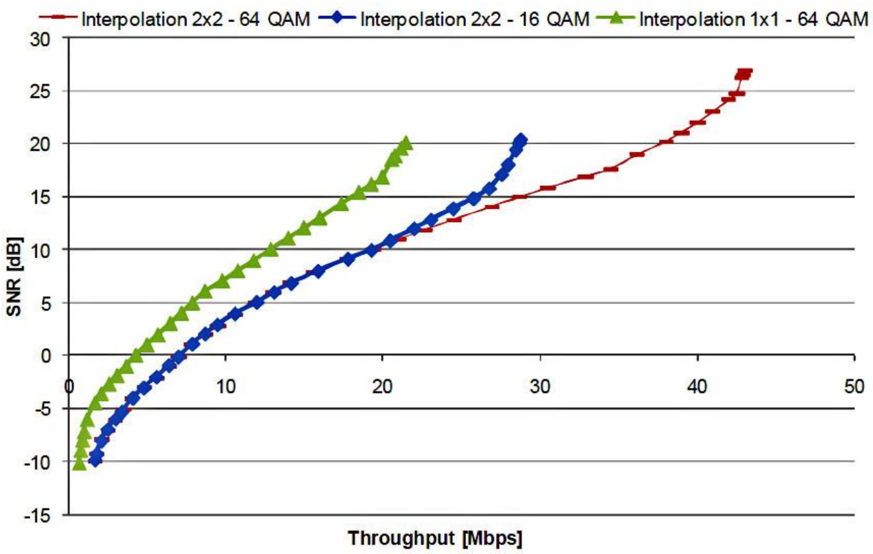


Figure A.2 – HSPA+ downlink SINR as function of physical throughput (extracted from [Jaci09])

A.2.2 Uplink

Uplink in UMTS as well as downlink has multiple configuration. In [Jaci09], functions for various configuration is proposed.

In this thesis, like downlink, 2×2 MIMO configuration with 64-QAM is selected. The equation for data rate as function of SINR is:

$$R_{b_{UMTS}[\text{Mbps}]} = \sum_{i=0}^{4} a_i \rho_{[\text{dB}]}^{i} \tag{A.5}$$

where:

- $R_{b_{UMTS}}$: data rate for a single code in downlink;
- ρ: SINR;
- a_i: factors of polynomial which is selected from Table A.4.

Table A.4 – Polynomial factors of UMTS uplink throughput.

a_4	a_3	a_2	a_1	a_0	Condition
-0.01	-0.07	-0.09	0.60	3.05	$-5 \leq \rho_{[\text{dB}]} < -3$
0.03	0.1	0.03	0.4	3	$-3 \leq \rho_{[\text{dB}]} < 1$
0	0.06	-0.57	2.37	1.66	$1 \leq \rho_{[\text{dB}]} < 5$
0	0	-0.02	0.96	2.19	$5 \leq \rho_{[\text{dB}]} < 15$
0	0	0	0	11	$15 \leq \rho_{[\text{dB}]} < 20$

The SINR as function of data rate in UMTS uplink, in [Jaci09], the following equation is proposed:

$$\rho_{[\text{dB}]} = \sum_{i=1}^{4} b_i R_{b_{UMTS}[\text{Mbps}]}^{i} \tag{A.6}$$

where:

- $R_{b_{UMTS}}$: data rate for a single code in downlink;
- ρ: SINR;
- b_i: factors of polynomial which is selected from Table A.5.

Table A.5 – Polynomial factors of UMTS uplink SINR.

b_4	b_3	b_2	b_1	b_0	Condition
0	-1.54	6.94	-6.94	-3	$1.8 \leq R_{b[\text{Mbps}]} < 3$
0	0	0.68	-4.61	8.79	$3 \leq R_{b[\text{Mbps}]} < 4.6$
0.13	-3.04	26.05	-95.83	129.02	$4.6 \leq R_{b[\text{Mbps}]} < 7.7$
0	0.14	-3.50	30.98	-87.22	$7.7 \leq R_{b[\text{Mbps}]} < 11$

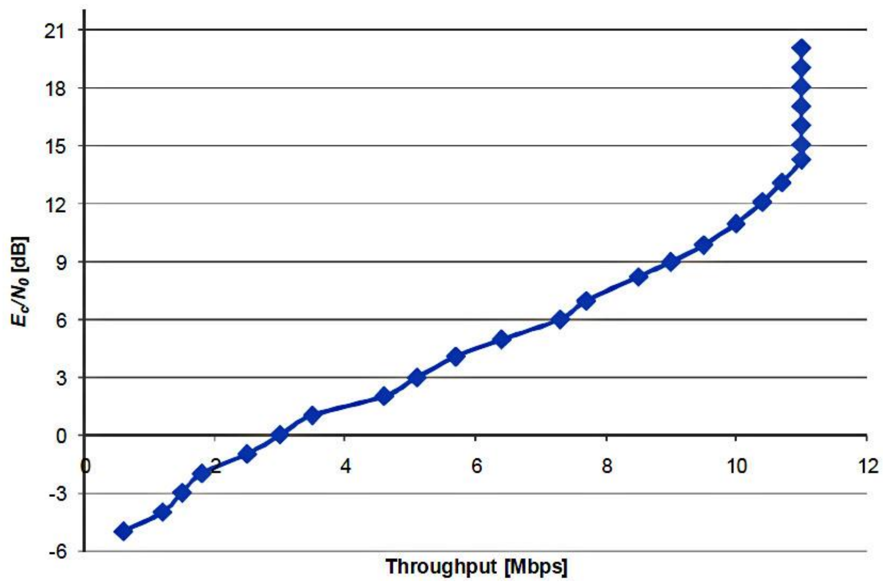


Figure A.3 – HSPA+ uplink SINR as a function of throughput (extracted from [Perg08]).

A.3 LTE

A.3.1 Downlink:

Three channel outline for LTE, as it is presented in Table A.6, is addressed by 3GPP and [Duar08].

Table A.6 – Overview of channel models (extracted from [AROM09]).

Channel Model	Doppler Frequency [Hz]	Delay Spread [ns]
EPA 5Hz	5	43
EVA 5Hz	5	357
EVA 70Hz	70	357
ETU 70Hz	70	991
ETU 300Hz	300	991

In [Jaci09], equations for data rates as function of SINR and SINR as function of data rate for various configuration and channel outline are presented. In this thesis for downlink, 64-QAM and performance for EPA 5Hz channel model is considered to represent LTE. The throughput as a function of SINR in this configuration is given by:

$$R_{b_{LTE}[\mathrm{Mbps}]} = \sum_{i=0}^{3} a_i \rho_{[\mathrm{dB}]}^{i} \tag{A.7}$$

where:

- $R_{b_{LTE}}$: data rate for a single code in downlink;
- ρ: SINR;
- a_i: factors of polynomial which is selected from Table A.7.

Table A.7 – Polynomial factors of LTE downlink throughput.

a_3	a_2	a_1	a_0	Condition
-0.001232	0.012683	-0.004081	0.002895	$0 \leq \rho_{[\mathrm{dB}]} < 6$
0	-0.000971	0.027847	0.036711	$6 \leq \rho_{[\mathrm{dB}]} < 10$
0	0.005579	-0.095673	0.615892	$10 \leq \rho_{[\mathrm{dB}]} < 16$
0	0	0.041523	-0.153522	$16 \leq \rho_{[\mathrm{dB}]} < 18$
0	-0.0022983	0.0974083	-0.413470	$18 \leq \rho_{[\mathrm{dB}]} < 22$
0	0	0	0.617108	$22 \leq \rho_{[\mathrm{dB}]} < 26$

The reverse function, SINR as function of data rate, for the same configuration and channel outline is given by (Figure A.4):

$$\rho_{[\mathrm{dB}]} = \sum_{i=1}^{3} b_i R^i_{b_{LTE}[\mathrm{Mbps}]} \tag{A.8}$$

where:

- $R_{b_{LTE}}$: data rate for a single radio block in downlink;
- ρ: SINR;
- b_i: factors of polynomial which is selected from Table A.8.

Table A.8 – Polynomial factors of LTE downlink SINR.

b_3	b_2	b_1	b_0	Condition
0	-291.09	70.06	-0.21	$0.003 \leq R_{b[\mathrm{Mbps}]} < 0.037$
0	0	25.42	1.05	$0.037 \leq R_{b[\mathrm{Mbps}]} < 0.116$
966.09	-108.09	1.30	3.79	$0.116 \leq R_{b[\mathrm{Mbps}]} < 0.228$
225.05	-299.69	144.27	-10.03	$0.228 \leq R_{b[\mathrm{Mbps}]} < 0.622$

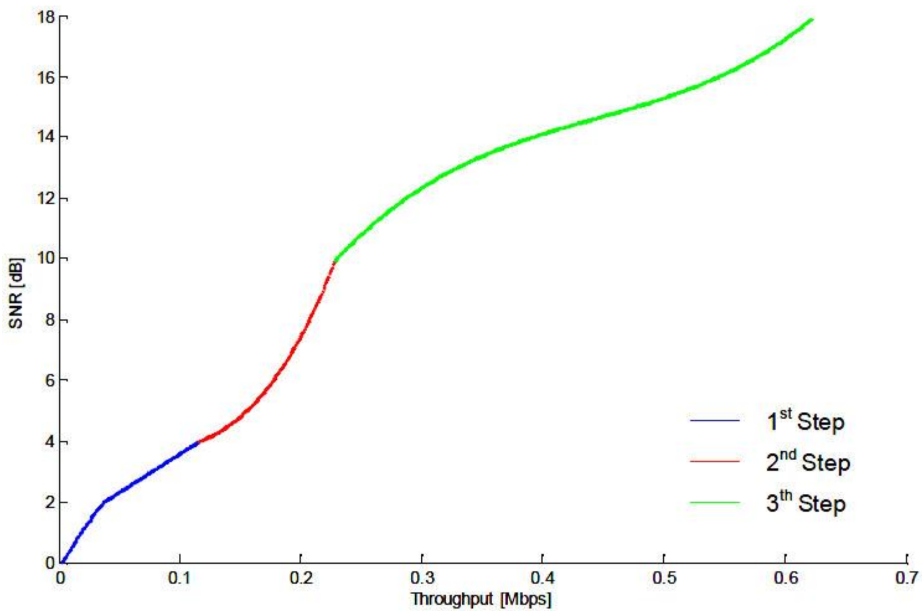


Figure A.4 – Downlink throughput interpolation for 64QAM, coding rate ¾, on EPA 5Hz over a bandwidth of 10MHz (extracted from [Jaci09]).

A.3.2 Uplink:

For LTE uplink, like downlink, 64-QAM and MIMO configuration of 2×2 is selected. In [Jaci09], data rate as function of SINR is given by:

$$R_{b_{LTE}[\mathrm{Mbps}]} = \sum_{i=0}^{4} a_i \rho_{[\mathrm{dB}]}^{i} \tag{A.9}$$

where:

- $R_{b_{LTE}}$: data rate for a single radio block in downlink;
- ρ: SINR;
- a_i: factors of polynomial which is selected from Table A.9.

Table A.9 – Polynomial factors of LTE uplink throughput.

$\boldsymbol{a_4}$	$\boldsymbol{a_3}$	$\boldsymbol{a_2}$	$\boldsymbol{a_1}$	$\boldsymbol{a_0}$	**Condition**
-4.9	416	9514	1.0962E5	196616	$4 \leq \rho_{[\mathrm{dB}]} \leq 18$
0	0	-2980	143480	-1010580	$18 \leq \rho_{[\mathrm{dB}]} \leq 24$

The reverse function, SINR as function of data rate, is given by:

$$\rho_{[\mathrm{dB}]} = \sum_{i=1}^{4} b_i R_{b_{LTE}[\mathrm{Mbps}]}^{i} \tag{A.10}$$

where:

- $R_{b_{LTE}}$: data rate for a single radio block in downlink;
- ρ: SINR;
- b_i: factors of polynomial which is selected from Table A.10.

Table A.10 – Polynomial factors of LTE uplink SINR.

$\boldsymbol{b_4}$	$\boldsymbol{b_3}$	$\boldsymbol{b_2}$	$\boldsymbol{b_1}$	$\boldsymbol{b_0}$	**Condition**
-644.15	581.30	-137.50	33.39	1.19	$0.12 \leq R_{b[\mathrm{Mbps}]} < 0.51$
0	0	104.21	-95.72	37.73	$0.51 \leq R_{b[\mathrm{Mbps}]} < 0.70$

A.4 Polynomial Approximation:

In (A.4), (A.8) and (A.10), SINR as function of data rate is presented. These functions have been approximated to equivalent polynomial of degree 5 with lasso technique.

$$\rho_{[db]} = f\left(R_{b[Mb]}\right) = \sum_{i=0}^{5} a_i \times R_{b[Mb]}^{i} \quad \text{(A.11)}$$

where:

- ρ: SINR;
- R_b: data rate;
- a_i: are presented in Table A.11.

Table A.11 – Factors of polynomial approximation of SINR as function of data rate.

RAT		a_5	a_4	a_3	a_2	a_1	a_0
GSM		-8713096.19	0	360715.89	-29325.88	964.66	7.96
UMTS	UL	0	0	0	-0.20	2.83	-6.85
	DL	0	0	0	-0.04	1.90	-12.15
LTE	UL	131.84	-78.49	-50.46	34.97	26.56	0.08
	DL	309.99	-52.78	-274.42	119.30	26.48	0.24

Annex B

Traffic Source Models

This annex presents the traffic source models for various services on the networks. These models are used in implementation of different service profile in the simulations and evaluations.

B.1 Introduction

Network traffic, as it is shown in Table B., can be considered at different levels; therefore, it is necessary to start with some ground terminology. A session, which a client host establishes to another one, is associated with human activity. An FTP session, for instance, starts by opening of a TCP connection to port 21 on a server. The TCP connection is going to be closed only at the end of session. A session, for connection-oriented networks, is a call established and terminated by signalling messages [Chen07]. Modelling network traffic at the session (or call) level only consist of characterising the start times and duration of each session.

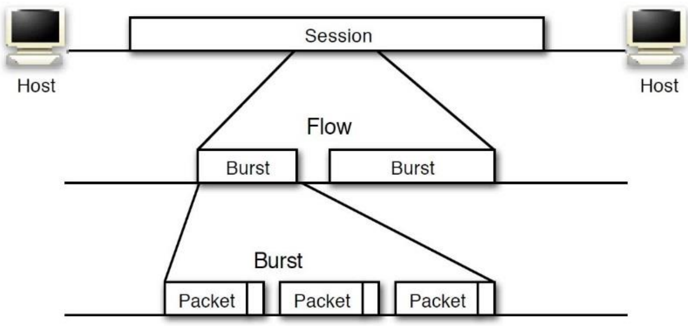


Figure B.1 – Level of Traffic (extracted form [Chen07])

During a session one or more packet flows may happen between two hosts [Robe04]. Despite inconsistency in usage of term “flow” in the literature, it is defined commonly as a series of closely spaced packets in one direction between a specific pair of hosts. Packets of a flow usually have common packet header fields such as protocol ID and port plus source and destination addresses [Robe04]. Modelling traffic at flow level is about characterising the random start times and duration of each flow.

In a more detailed level, intermittent bursts form a flow. Each of these bursts, itself, is consisting of consecutively transmitted packets. Bursts may arise in window-based protocols where a host is allowed to send a window of packets, and then it has to wait to receive credit to send another window

Finally, traffic can be viewed at the level of individual packets. This level is concerned only on the arrival process of packets and ignores any higher structure in the traffic (e.g., bursts, flows, and sessions). The majority of research (and this section) address traffic models mainly at the packet level. Studies at the packet level are relatively straightforward because packets can be easily captured for minutes or hours. Where studies of traffic flows and sessions require collection and analysis of greater traffic

B.2 Voice-Over-IP (VoIP) Traffic Model

[GCPA09] considered a single two-state activity model for VoIP, as it is presented in Figure B.2. The two states of this model are silence or inactive state (state 0) and talking or active state (state 1). The transition probability from state 0 to state 1 is α while the probability of staying in state 0 is $(1-\alpha)$. Moreover, transition from state 1 back to state 0 is done by probability of βwhile with $(1-\beta)$ is the probability of staying in state 1. The probabilities of being in state "0 and state 1, P_0 and P_1, are given as:

$$P_0 = \frac{\beta}{\alpha+\beta} \tag{B.1}$$

$$P_1 = \frac{\alpha}{\alpha+\beta} \tag{B.2}$$

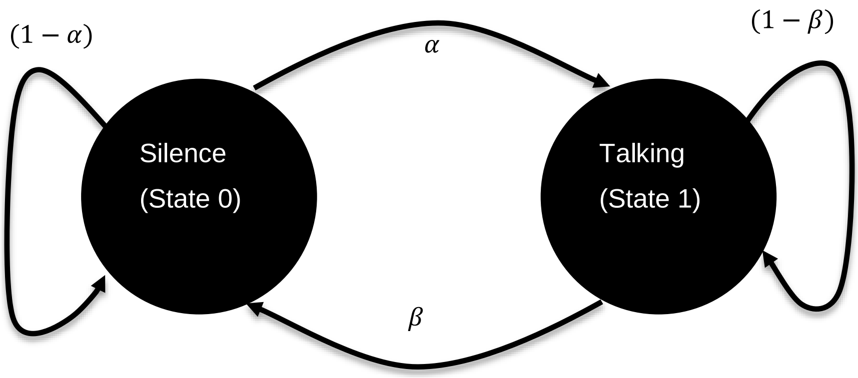


Figure B.2 – Two state voice activity model (extracted from [GCPA09]).

The probability of being in the talking state, state 1, is referred to as Voice Activity Factor (VAF):

$$VAF = P_1 = \frac{\alpha}{\alpha+\beta} \tag{B.3}$$

The mean silence duration and mean talking duration in terms of number of voice frame can be written as:

$$E[\tau_s] = \frac{1}{\alpha} \quad \text{The mean silence duration} \tag{B.4}$$

$$E[\tau_t] = \frac{1}{\beta} \quad \text{The mean talking duration} \tag{B.5}$$

The probability of n-frame long silence or talking is given by:

$$p_{\tau_s=n} = \alpha(1-\alpha)^{n-1} \quad n\text{-frame silence duration} \tag{B.6}$$

$$p_{\tau_t=n} = \beta(1-\beta)^{n-1} \quad n\text{-frame talking duration} \tag{B.7}$$

The distribution of the time period τ_{AE} (in voice frame) between successive transitions into the talking state is the convolution of the distribution of τ_s and τ_t. The probability that this duration is *n*-voice frame long is:

$$p_{\tau_{AE}=n} = \frac{\alpha}{\alpha-\beta}\beta(\alpha-\beta)^{n-1} + \frac{\beta}{\beta-\alpha}\alpha(1-\alpha)^{n-1} \tag{B.8}$$

Since the state transitions from state 0 to state 1 and vice versa are independent, the mean time between successive transitions into the talking state is simply the sum of the mean time in each state.

$$E[\tau_{AE}] = E[\tau_s] + E[\tau_t] = \frac{1}{\alpha} + \frac{1}{\beta} \tag{B.9}$$

The rate of arrivals into the active state can serve as a guide on the number of resource requests needed for persistent allocation or resources for VoIP traffic. In general, it is expected that a single resource request and/or scheduling grant will be required for persistent allocation of resources when a VoIP users move from the inactive to the talking state. The VoIP traffic model parameters are given in Table B.1.

Table B.1 – VoIP traffic model parameter in LTE (extracted from [Khan09])

Parameter	Value
Voice codec	RTP AMR 12.2, source rate 12.2 kbps
Encoder frame length	20 ms
Voice Activity Factor (VAF)	50% $\alpha = \beta = 0.01$
SID payload	SID packet every 160 ms during silence 15 bytes (5bytes + header)
Protocol overhead with header compression	10 bit + padding (RTP pre-header) + 4 byte (RTP/UDP/IP) + 2 byte (RLC/security) + 16 bits (CRC)
Total voice payload on air interface	40 bytes

B.3 Video Streaming Traffic Model

[Khan09] assumed that each frame of video data arrives at a regular interval T determined by the number of frames per second. Each video frame is decomposed into a fixed number of slices, each transmitted as a single packet. The size of these packets/slices is modelled as a truncated Pareto distribution. The video encoder introduces encoding delay intervals between the packets of a frame. These intervals are also modelled by a truncated Pareto distribution. The video streaming traffic model parameters are given in Table B.2. In this model, the video source rate is assumed at 64 kbps.

Table B.2 – Video Streaming traffic model parameters (extracted from [Khan09])

Parameter	Statistical Characterisation
Inter-arrival time between the beginning of each frame	Deterministic at 100 ms (10 frames per second)
Number of packets (slices) in a frame	Deterministic, 8 packets per frame
Packet (slice) size	Truncate Pareto Distribution Mean = m = 10 bytes, Maximum = 250 bytes (before truncation) PDF: $f_x = \frac{\alpha_k^{\alpha}}{\alpha+1}, k \leq x \leq m, f_x = \left(\frac{k}{m}\right)^{\alpha}$ $x = m, \alpha = 1.2, k = 20\ bytes$

Inter-arrival time between packets (slices) in a frame	Truncate Pareto Distribution Mean = m = 10 bytes, Maximum = 250 bytes (before truncation) PDF: $f_x = \frac{\alpha_k^\alpha}{\alpha+1}, k \leq x \leq m, f_x = \left(\frac{k}{m}\right)^\alpha$ $x = m, \alpha = 1.2, k = 2.5\ ms$

B.4 Best Effort FTP Traffic Model

In case of FTP, the same traffic model for uplink and downlink can be used. In general, An FTP session is a sequence of file transfers separated by reading times. The two key FTP session parameters are *S*, the size of the file to be transferred, and *D*, reading time (i.e. the time interval between end of former file download and the next user request for the next file.). The details of FTP traffic model is presented in Table B.3.

Table B.3 – FTP traffic model parameters (extracted from [Khan09])

Parameter	Statistical Characterisation
File Size (S)	Truncated Lognormal Distribution Mean = 2 MB Standard deviation = 0.7222 MB Maximum = 5 MB (before truncation) PDF: $f_x = \frac{1}{\sqrt{2pi}\sigma x} e^{\frac{-(\ln z - \mu)^2}{2\sigma^2}}\ x > 0, \sigma = 0.35, \mu = 14.45$
Reading Time (D)	Exponential Distribution Mean = 180 s PDF: $f_x = \lambda e^{-\lambda x}, x \geq 0, \lambda = 0.006$

B.5 Interactive Real-Time Services, Gaming

The model parameter of interactive gaming traffic for the uplink are given in Table B.4. An initial packet arrival time is uniformly distributed between 0 and 40 ms. This initial time was considered to model the random timing relationship between client traffic packet arrival and uplink frame boundary in cdma2000 systems [Khan09]. In the LTE systems with sub-frame duration of only 1.0 ms, this initial time to account for the resource request and scheduling grant is expected to be very small. The packet inter-arrival time is deterministic with a packet appearing every 40 ms. The packet size is assumed to follow the largest extreme value distribution, which is also known as the Fisher–Tippett distribution or the log–Weibull distribution. The values for this distribution can be generated by the following procedure:

$$x = a - b\ln(-\ln y) \tag{8.10}$$

where y is drawn from a uniform distribution in the range [0, 1]. Since the packet size needs to be an integer number of bytes, the largest integer less than or equal to x is used as the actual packet size. A compressed UDP header consisting of 2 bytes is added to each packet.

A maximum delay of 160 ms is applied to all uplink packets, i.e. a packet is dropped by the MT if any part of the packet has not started physical layer transmission, 160 ms after entering the MT buffer. The packet delay of a dropped packet is counted as 180 ms. A mobile network gaming user is in outage if the average packet delay is greater than 60 ms. The average delay is the average of the delays of all packets, including the delay of packets delivered and the delay of packets dropped.

Table B.4 -- Interactive gaming traffic model parameters for the uplink (extracted from [Khan09])

Parameter	Statistical Characterisation
Initial packet arrival	Uniform Distribution PDF: $f_x = \frac{1}{b-a}, a \le x \le b, a = 0, b = 40ms$
Packet arrival	Deterministic, 40 ms
Packet Size	Largest Extreme Value Distribution (also known as Fisher-Tippett distribution) $f_x = \frac{1}{b} e^{-\frac{x-a}{b}} e^{-e^{\frac{x-a}{b}}}, a = 45\ B, b = 5.7$
UDP header	Deterministic (2 Bytes). This is added to the packet size accounting for the UDO header after header compression.

The interactive gaming traffic model parameters for the downlink are given in Table B.5. An initial packet arrival time is uniformly distributed between 0 and 40 ms. The packet inter- arrival times as well as the packet size on the downlink are modelled using the largest extreme value distribution.

Table B.5 – Interactive gaming traffic model parameters for the uplink (extracted from [Khan09])

Parameter	Statistical Characterisation
Initial packet arrival	Uniform Distribution PDF: $f_x = \frac{1}{b-a}, a \le x \le b, a = 0, b = 40ms$
Packet arrival	Largest Extreme Value Distribution (also known as Fisher-Tippett distribution) $f_x = \frac{1}{b} e^{-\frac{x-a}{b}} e^{-e^{\frac{x-a}{b}}}, a = 55\ B, b = 6$
Packet Size	Largest Extreme Value Distribution (also known as Fisher-Tippett distribution) $f_x = \frac{1}{b} e^{-\frac{x-a}{b}} e^{-e^{\frac{x-a}{b}}}, a = 120\ B, b = 36$

B.6 Web Browsing HTTP Traffic Model

The modelling of an HTTP (hypertext transfer protocol) web browsing session can be done by modelling of web browsing session [ETSI98]. Web based services are typically classified as Long Constrained Delay (LCD) services beside they are unidirectional e.g. in downlink. In a web session, a packet call corresponds to the downloading for a web document such as a web page with images, text, applets, etc.

Figure B.3 illustrates packet trace of a typical web browsing session. As it is depicted, each session of this kind is formed by active and inactive periods [Khan09]. These periods are result of human interactions. A web user's request for information is interpreted by active period to which packet calls are also referred. . After the document is entirely retrieved to the terminal, the user consumes some time for studying the information. This time interval is referred as reading interval represented by inactive periods.

The web traffic is claimed to exhibits self-similar behaviour in [CrBe97, LTWW02]. In other words, the traffic statics on different timescales are similar. Hence, packet calls, as well as packet session, are also consists of active and inactive periods. In contrast to packet sessions, these periods rather than human interaction are attribute to hosts interaction. Depending on the application, it may have several packet call in a packet service.

A web-browser will begin serving a user's request by fetching the initial HTML page using an HTTP GET request. The retrieval of the initial page and each of the embedded objects (e.g. pictures, advertisements, etc.) is represented by the active period within the packet call while the parsing time and protocol overhead are represented by the inactive periods within a packet call. The parsing time refers to the time the browser spends in parsing for the embedded objects in the packet call or the web page.

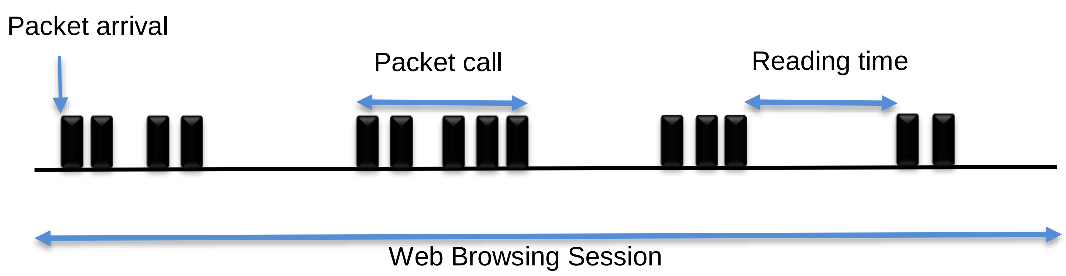


Figure B.3 - Web session (adopted from [Duar08] and [Khan09]).

The key characterising parameters for the web browsing traffic are the main object size, S_M, the size of an embedded object in a web page, S_E, the number of embedded objects, N_D, reading time, D and parsing time T_p. These parameters presented in Table B.6.

Table B.6 – HTP traffic model parameters (extracted from [Khan09])

Parameter	Statistical characterisation
Main object size, S_M	Truncated Lognormal distribution Mean = 10710 bytes, Standard deviation = 25032 bytes, Minimum = 100 bytes, Maximum = 2 Mbytes (before truncation)

	PDF: $f_x = \frac{1}{\sqrt{2pi}\sigma x} e^{\frac{-(\ln z-\mu)^2}{2\sigma^2}} \; x > 0, \sigma = 1.37, \mu = 8.37$
Embedded object size, S_E	Truncated Lognormal distribution Mean = 7758 bytes, Standard deviation = 126168 bytes, Minimum = 50 bytes, Maximum = 2 Mbytes (before truncation) PDF: $f_x = \frac{1}{\sqrt{2pi}\sigma x} e^{\frac{-(\ln z-\mu)^2}{2\sigma^2}} \; x > 0, \sigma = 2.36, \mu = 6.17$
Number of embedded objects per page, N_D	Truncate Pareto Distribution Mean = 5.64 Maximum = 53 PDF: $f_x = \frac{\alpha_k^{\alpha}}{\alpha+1}, k \leq x \leq m, f_x = \left(\frac{k}{m}\right)^{\alpha}$ $x = m, \alpha = 1.1, k = 2, m = 55$
Reading time, D	Exponential distribution with a mean of 30 s PDF: $f_x = \lambda e^{-\lambda x}, x \geq 0, \lambda = 0.033$
Parsing time, T_p	Exponential distribution with a mean of 30 s PDF: $f_x = \lambda e^{-\lambda x}, x \geq 0, \lambda = 7.69$

References

[3GPP09] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures*, Report No. TS 36.213 version 8.8.0 European Telecommunications Standards Institute, Sophia-Antipolis, France, Oct. 2009 (http://www.etsi.org/deliver/etsi_ts/136200_136299/136213/08.08.00_60/ts_136213v080800p.pdf).

[3GPP11] 3GPP, *Universal Mobile Telecommunications System (UMTS); User Equipment (UE) radio transmission and reception (FDD)* Report No. TS 25.101 version 10.1.0 European Telecommunications Standards Institute (ETSI), Sophia-Antipolis, France, May 2011 (http://www.etsi.org/deliver/etsi_ts/125100_125199/125101/10.01.00_60/ts_125101v100100p.pdf).

[3GPP13] 3GPP, *Universal Mobile Telecommunications System (UMTS); LTE; Network sharing; Architecture and functional description*, Report No. TS 23.251 Version 11.4.0, European Telecommunications Standards Institute (ETSI), France, Jan. 2013 (http://www.etsi.org/deliver/etsi_ts/123200_123299/123251/11.04.00_60/ts_123251v110400p.pdf).

[5GPP15] https://5g-ppp.eu/, Last Access: Oct., 2015.

[AkHN11] K. Akkarajitsakul, E. Hossain and D. Niyato, "Distributed resource allocation in wireless networks under uncertainty and application of Bayesian game", *IEEE Communications Magazine,* Vol. 49, No. 8, Aug. 2011, pp. 120-127.

[AlBA12] A. Al Sabbagh, R. Braun and M. Abolhasan, "A power efficient RAT selection algorithm for heterogeneous wireless networks", in *Proc. of ISCIT'12 – International Symposium on Communications and Information Technologies*, Queensland, Australia, Oct. 2012.

[AROM09] *Advanced Resource Management Solutions for Future All IP Heterogeneous Mobile Radio Environments, EC Project IST-AROMA*, Nov. 2009 (http://www.aroma-ist.upc.edu/).

[ArRA10] A. Balasubramanian, R. Mahajan and A. Venkataramani, "Augmenting mobile 3G using Wi-Fi", in *Proc. of 8th international conference on Mobile, Systems, Applications, and Services*, San Francisco, CA, USA, June 2010.

[AyAT14] B. Z. Aymen, M. Ayadi and S. Tabbane, "A fuzzy logic algorithm for RATs selection procedures", in *Proc. of ISNCC'14 – International Symposium on Networks, Computers and Communications*, Hammamet, Tunisia, June 2014.

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

[BDPW11] P. Bosch, A. Duminuco, F. Pianese and T. L. Wood, "Telco clouds and Virtual Telco: Consolidation, convergence, and beyond", in *Proc. of IFIP'11 – 6th IEEE International Symposium on Integrated Network Management*, Dublin, Ireland, May 2011.

[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.

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

[BuBu06] M. Buehrer and R. M. Buehrer, *Code Division Multiple Access (CDMA)* Morgan & Claypool Publishers, San Rafael, CA, USA, 2006.

[CCYS14] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger and L. Dittmann, "Cloud RAN for Mobile Networks - a Technology Overview", *IEEE Communications Surveys & Tutorials,* Vol. 17, No. 1, Mar. 2014, pp. 405 - 426.

[CFHH05] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt and A. Warfield, "Live migration of virtual machines", in *Proc. of 2nd conference on Symposium on Networked Systems Design & Implementation* Boston, MA, USA, May 2005.

[ChBo10] N. M. M. K. Chowdhury and R. Boutaba, "A survey of network virtualization", *Computer Networks,* Vol. 54, No. 5, Apr. 2010, pp. 862-876.

[Chen07] T. M. Chen, "Network Traffic Modelling," in H. Bidgoli, (Ed.) *Handbook of Computer Networks.* vol. 3, John Wiley & Sons, Hoboken, NJ, USA, 2007.

[ChFY04] Y. Chen, T. Farley and N. Ye, "QoS Requirements of Network Applications on the Internet", *Information-Knowledge-Systems Management,* Vol. 4, No. 1, Jan. 2004, pp. 55-76.

[ChYu15] L. Chengchao and F. R. Yu, "Wireless Network Virtualization: A Survey, Some Research Issues and Challenges", *IEEE Communications Surveys & Tutorials,* Vol. 17, No. 1, 2015, pp. 358-380.

[Cisc12] Cisco Systems, *802.11ac: The Fifth Generation of Wi-Fi*, Technical White Paper, Cisco, San Jose, CA, USA, Aug. 2012 (http://www.cisco.com/en/US/prod/collateral/wireless/ps5678/ps11983/white_paper_c11-713103.pdf).

[Cisc13a] Cisco Systems, *Global Mobile Data Traffic Forecast Update, 2012 - 2017*, from Visual Network Index (VNI) White Paper, Cisco Systems, CA, USA, Feb. 2013 (http://www.cisco.com/).

[Cisc13b] Cisco Systems, *The Zettabyte Era - Trends and Analysis*, from Visual Network Index (VNI) White Paper, Cisco Systems, CA, USA, May 2013 (http://www.cisco.com/).

[CMRI11] China Mobile Research Institute, *C-RAN – Road Towards Green Radio Access Network*, Shanghai, China, Oct. 2011 (http://labs.chinamobile.com/cran/wp-content/uploads/CRAN_white_paper_v2_5_EN.pdf).

[COST15] http://www.ic1004.org/, Last Access: Nov., 2015.

[CrBe97] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes", *IEEE/ACM Transactions on Networking,* Vol. 5, No. 6, Dec. 1997, pp. 835-846.

[CSTM13] X. Costa-Perez, J. Swetina, G. Tao, R. Mahindra and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks", *IEEE Communications Magazine,* Vol. 51, No. 7, July 2013, pp. 27-35.

[DaCo99] E. Damosso and L. M. Correia (Eds.), *COST 231 Final Report – Digital Mobile Radio: Evolution Towards Future Generation Systems* COST Secretariat, European Commission, Brussels, Belgium, 1999.

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

[Duar08] S. Duarte, *Analysis of Technologies for Long Term Evolution in UMTS*, M.Sc. Thesis, Instituto Superior Técnico,, Lisbon, Portugal, 2008

[Elba11] H. M. ElBadawy, "Optimal RAT selection algorithm through Common Radio Resource Management in heterogeneous wireless networks", in *Proc. of NRSC'11 – 28th National Radio Science Conference* Cairo, Egypt, Apr. 2011.

[ERIC13] Ericsson., *Mobility Report on The Pulse of The Networked Society*, ERICSSON, Stokholm, Sweden, Nov. 2013 (http://www.ericsson.com/res/docs/2013/ericsson-mobility-report-november-2013.pdf).

[ETSI13] ETSI, *Network Function Virtualisation (NFV) - Network Operator Prespectives on Industry Progress*, European Telecommunications Standards Institute (ETSI), Frankfurt - Germany, Oct. 2013 (http://portal.etsi.org/NFV/NFV_White_Paper2.pdf).

[ETSI98] ETSI, *Universal Mobile Telecommunications System (UMTS), Selection procedures for the choice of radio transmission technologies of the UMTS*, Report No. TR 101 112 v3.2.0, UMTS 30.03 version 3.2.0, France, Apr. 1998 (http://www.techstreet.com/products/953010).

[FaMu14] O. E. Falowo and A. Muponda, "Predictive RAT-selection algorithm for enhancing QoS in heterogeneous wireless networks", in *Proc. of LATINCOM'14 – IEEE Latin-America Conference on Communications*, Arequipa, Peru, Nov. 2014.

[FPTG12] R. Friedrich, S. Pattheeuws, D. Trimmel and H. Geerdes. *Sharing Mobile Networks - Why the Pros Outweigh the Cons*, (http://www.booz.com/media/uploads/BoozCo_Sharing-Mobile-Networks.pdf).

[Fren11] L. E. Frenzel, *Wireless Companies Follow The Roadmap Past 4G And on to 5G*, Electronic Designe, June 2011 (http://electronicdesign.com/print/4g/wireless-companies-follow-roadmap-past-4g-and-5g).

[FTAL08] T. Frisanco, P. Tafertshofer, P. Lurin and R. Ang, "Infrastructure Sharing and Shared Operations for Mobile Network Operators: From a Deployment and Operations View", in *Proc. of ICC '08 – IEEE International Conference on Communications*, Beijing, China, May 2008.

[GCPA09] G. Schaffrath, C. Werle, P. Papadimitriou, A. Feldmann, R. Bless, A. Greenhalgh, A. Wundsam, M. Kind, O. Maennel and L. Mathy, "Network virtualization architecture: proposal and initial prototype", in *Proc. of 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures*, Barcelona, Spain, Aug. 2009.

[GhBo06] M. Ghaderi and R. Boutaba, "Call admission control in mobile cellular networks: a comprehensive survey: Research Articles", *Wireless Communications and Mobile Computing,* Vol. 6, No. 1, Feb. 2006, pp. 69-93.

[GoMG12] I. Gomez Miguelez, V. Marojevic and A. Gelonch Bosch, "Resource Management for Software-Defined Radio Clouds", *IEEE Micro,* Vol. 32, No. 1, Jan. 2012, pp. 44-53.

[GoSa05] G. Gomez and R. Sanchez, *End-to-End Quality of Service over Cellular Networks: Data Services Performance Optimization in 2G/3G* John Wiley & Sons, Chichester, UK, 2005.

[GuKM10] H. Guan, T. Kolding and P. Merz, *Discovery of Cloud-RAN*, NSN Cloud-RAN Workshop, Beijing, China, Apr. 2010

[HoTo04] H. Holma and A. Toskala, *WCDMA for UMTS* John Wiley & Sons, Chichester, UK, 2004.

[HoTo10] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced* John Wiley & Sons, Chichester, UK, 2010.

[Jaci09] N. Jacinto, *Performance Gains Evaluation from UMTS/HSPA+ to LTE at the Radio Network Level*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2009

[JYSY14] L. Joohyun, Y. Yung, C. Song and J. Youngmi, "Economics of Wi-Fi Offloading: Trading Delay for Cellular Capacity", *IEEE Transactions on Wireless Communications,* Vol. 13, No. 3, Mar. 2014, pp. 1540-1554.

[Kaar05] H. Kaaranen, *UMTS networks: architecture, mobility, and services* John Wiley & Sons, Chichester, UK, 2005.

[KDZH13] A. Kliks, N. Dimitriou, A. Zalonis and O. Holland, "Wi-Fi traffic offloading for energy saving", in *Proc. of ICT'13 – 20th International Conference on Telecommunications*, Casablanca, Morocco, May 2013.

[Khan09] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance* Cambridge University Press, Cambridge, UK, 2009.

[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 Annual International Symposium on Personal, Indoor, and Mobile Radio Communications*, Washington, DC, USA, Sep. 2014.

[KhCo15] S. Khatibi and L. M. Correia, "The Effect of Channel Quality on Virtual Radio Resource Management", in *Proc. of VTC2015Fall – IEEE 82nd Vehicular Technology Conference*, Boston, MA, USA, Sep. 2015.

[KhCo15] S. Khatibi and L. M. Correia, "A model for virtual radio resource management in virtual RANs", *EURASIP Journal on Wireless Communications and Networking,* Vol. 2015, No. 1, 2015, p. 68.

[KJYI13] L. Kyunghan, L. Joohyun, Y. Yung, R. Injong and C. Song, "Mobile Data Offloading: How Much Can Wi-Fi Deliver?", *IEEE/ACM Transactions on Networking,* Vol. 21, No. 2, Apr. 2013, pp. 536-550.

[KKKZ11] A. Khan, W. Kellerer, K. Kozu and M. Yabusaki, "Network sharing in the next mobile network: TCO reduction, management flexibility, and operational independence", *IEEE Communications Magazine,* Vol. 49, No. 10, Oct. 2011, pp. 134-142.

[KMSS10] A. Klein, C. Mannweiler, J. Schneider and H. D. Schotten, "Access Schemes for Mobile Cloud Computing", in *Proc. of MDM'10 – 11th International Conference on Mobile Data Management*, Kansas City, Missouri, USA, May 2010.

[Kris02] D. Krishnaswamy, "Game theoretic formulations for network-assisted resource management in wireless networks", in *Proc. of VTC'02 Fall – 56th IEEE Vehicular Technology Conference*, Vancouver, Canada, Sep. 2002.

[KZJK12] A. Khan, A. Zugenmaier, D. Jurca and W. Kellerer, "Network virtualization: a hypervisor for the Internet?", *IEEE Communications Magazine,* Vol. 50, No. 1, Jan. 2012, pp. 136-143.

[LMZT11] Z. Liang, L. Ming, Y. Zaki, A. Timm-Giel and C. Gorg, "LTE virtualization: From theoretical gain to practical solution", in *Proc. of 23rd International Teletraffic Congress* San Francisco, CA, USA, Sep. 2011.

[LSZW10] Y. Lin, L. Shao, Z. Zhu, Q. Wang and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements", *IBM Journal of Research and Development,* Vol. 54, No. 1, Jan. - Feb. 2010, pp. 4:1-4:12.

[LTWW02] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)", *IEEE/ACM Transactions on Networking,* Vol. 2, No. 1, Aug. 1994, pp. 1-15.

[LYLX12] X. Lu, K. Yang, Y. Liu, D. Zhou and S. Liu, "An Elastic Resource Allocation Algorithm Enabling Wireless Network Virtualization", *Wireless Communications and Mobile Computing,* Vol. 75, No. 4, Apr. 2014, pp. 1827-1849.

[Math15] *MATLAB and Statistics Toolbox*, The MathWorks Inc., Natick, MA, USA, Feb. 2015 (http://www.mathworks.com/).

[MCN15] Mobile Cloud Networking Project, Nov. 2015 (http://mobile-cloud-networking.eu).

[MDPA12] M. Chiosi, D. Clarke, P. Willis, A. Reid, J. Feger, M. Bugenhagen, W. Khan, M. Fargano, C. Cui, H. Deng, J. Benitez, U. Michel, H. Damker, K. Ogaki and T. Matsuzaki, *Network Function Virtualisation: An Introduction, Benefits, Enabler, Challenges, and Call for Action*, European Telecommunications Standards Institute, Darmstadt, Germany, Oct. 2012 (http://portal.etsi.org/NFV/NFV_White_Paper.pdf).

[Mehr92] S. Mehrotra, "On the Implementation of a Primal-Dual Interior Point Method", *Society for Industrial and Applied Mathematics Journal on Optimization,* Vol. 2, No. 4, Nov. 1992, pp. 575–601.

[MePe13] T. Metsch and P. Gray (Eds.), *Infrastructure Management Foundations – Specifications & Design for Mobile Cloud framework* Deliverable D3.1, MobileCloud Networking, Nov. 2013 (http://www.mobile-cloud-networking.eu/).

[MKHR13] R. Mahindra, M. A. Khojastepour, Z. Honghai and S. Rangarajan, "Radio Access Network sharing in cellular networks", in *Proc. of ICNP'13 – 21st IEEE International Conference on Network Protocols*, Göttingen, Germany, Oct. 2013.

[MLXX12] L. Ming, Z. Liang, L. Xi, L. Xiaona, Y. Zaki, A. Timm-Giel and C. Gorg, "Investigation of Network Virtualization and Load Balancing Techniques in LTE Networks", in *Proc. of VTC'12 – 75th IEEE Vehicular Technology Conference*, Yokohama, Japan, May 2012.

[MSQL15] http://www.microsoft.com/en-us/server-cloud/products/sql-server/default.aspx, Last Access: July, 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.

[NEWC15] Network of Excellence in Wireless Communications, Nov. 2015 (http://www.newcom-project.eu/).

[Nong02] N. Ye, "QoS-Centric Stateful Resource Management in Information Systems", *Information Systems Frontiers,* Vol. 4, No. 2, July 2002, pp. 149-160.

[OAMI15] A. Orsino, G. Araniti, A. Molinaro and A. Iera, "Effective RAT Selection Approach for 5G Dense Wireless Networks", in *Proc. of VTC2015Spring – 81st IEEE Vehicular Technology Conference*, Glasgow, Scotland, May 2015.

[Open15] http://openairinterface.eurecom.fr/, Last Access: July, 2015.

[PaPi02] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes* McGraw-Hill, NY, USA, 2002.

[Pere05] J. Pérez-Romero, *Radio resource management strategies in UMTS* John Wiley & Sons, Chichester, UK, 2005.

[Pere13] X. C. Perez, *Radio Access Network Virtualization and sharing for Future Mobile Carrier Networks*, RAN evolution to the Cloud Workshop, June 2013

[Perg08] R. Perguiça, *Comparison Between UMTS/HSPA+ and WiMAX/IEEE 802.16e in Mobility Scenarios*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2008

[PrVe10] R. Prasad and F. J. Velez, *WiMAX Networks: Techno-Economic Vision and Challenges* Springer, London, UK, 2010.

[RaGa10] V. Rakovic and L. Gavrilovska, "Novel RAT selection mechanism based on Hopfield neural networks", in *Proc. of ICUMT'10 – International Congress on Ultra Modern Telecommunications and Control Systems*, Brno, Czech Republic, Oct. 2010.

[RBDG14] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella and D. Wubben, "Cloud technologies for flexible 5G radio access networks", *IEEE Communications Magazine,* Vol. 52, No. 5, May 2014, pp. 68-76.

[Robe04] J. W. Roberts, "Internet traffic, QoS, and pricing", *Proceedings of the IEEE,* Vol. 92, No. 9, Sept. 2004, pp. 1389-1399.

[ROKB09] V. Rakovic, O. Ognenoski, A. Kaloxylos, A. Boucouvalas, V. Atanasovski and L. Gavrilovska, "A RAT selection mechanism in emergency scenarios", in *Proc. of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology* Aalborg, Denmark, May 2009.

[SAMS11] L. Samson, *LTE Network Sharing: Some Operational and Management Aspects*, ITU Cross Regional Seminar for CIS, ASP and EUR Regions on Broadband Access (Fixed, Wireless including Mobile), Chisinau, Moldova, Oct. 2011 (http://www.itu.int/ITU-D/tech/events/2011/CrossReg_BWA_Chisinau_October11/Presentations/CrossReg_Broadband_2011_Presentation_P21.pdf).

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

[SeTB11] S. Sesia, I. Toufik and M. Baker, *LTE - The UMTS Long Term Evolution: From Theory to Practice* John Wiley & Sons, Chichester, UK, 2011.

[Sing10] T. L. Singal, *Wireless Communications* McGraw-Hill, New Delhi, India, 2010.

[SoLC06] D. Soldani, M. Li and R. Cuny, *QoS and QoE management in UMTS cellular systems* John Wiley and Sons, Chichester, UK, 2006.

[SoZC14] Y. K. Song, H. Zo and A. P. Ciganek, "Multi-criteria evaluation of mobile network sharing policies in Korea", *ETRI Journal,* Vol. 36, No. 4, Aug. 2014, pp. 572-580.

[SRYP13] D. Sabella, P. Rost, S. Yingli, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. Di Girolamo and G. Giuliani, "RAN as a service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network", in *Proc. of Future Network and Mobile Summit*, Lisbon, Portugal, July 2013.

[TDGJ06] F. Travostino, P. Daspit, L. Gommans, C. Jog, C. de Laat, J. Mambretti, I. Monga, B. van Oudenaarde, S. Raghunath and P. Yonghui Wang, "Seamless live migration of virtual machines over the MAN/WAN", *Future Generation Computer Systems,* Vol. 22, No. 8, Oct. 2006, pp. 901-907.

[VANU13] http://www.vanu.com/, Last Access: Jan., 2013.

[ViWC02] J. A. Village, K. P. Worrall and D. I. Crawford, "3G shared infrastructure", in *Proc. of 3rd International Conference on 3G Mobile Communication Technologies*, London, UK, May 2002.

[WiGo04] W. L. Winston and J. B. Goldberg, *Operations Research: Applications and Algorithms* Thomson/Brooks/Cole, Belmont, CA USA, 2004.

[Wise09] D. Wisely, *IP for 4G* John Wiley & Sons, Chichester, UK, 2009.

[XiKT13] W. Xin, P. Krishnamurthy and D. Tipper, "Wireless network virtualization", in *Proc. of ICNC'13 – International Conference on Computing, Networking and Communications*, San Diego, CA, USA, Jan. 2013.

[YrRu11] R. Yrjo and D. Rushil, "Cloud computing in mobile networks - Case MVNO", in *Proc. of ICIN'11 – 15th International Conference on Intelligence in Next Generation Networks*, Berlin, Germany, Oct. 2011.

[ZGWK11] Z. Zhu, P. Gupta, Q. Wang, S. Kalyanaraman, Y. Lin, H. Franke and S. Sarangi, "Virtual base station pool: towards a wireless network cloud for radio access networks", in *Proc. of CF'11 – 8th ACM International Conference on Computing Frontiers*, New York, NY, USA, May 2011.

[Zhan95] Y. Zhang, *"Solving Large-Scale Linear Programs by Interior-Point Methods Under the MATLAB Environment*, Technical Report TR96-01, Department of Mathematics and Statistics, University of Maryland, Baltimore MD, USA, July 1995

[ZLGT10] Y. Zaki, Z. Liang, C. Goerg and A. Timm-Giel, "LTE wireless virtualization and spectrum management", in *Proc. of WMNC'2010 – 3rd Joint IFIP Wireless and Mobile Networking Conference*, Budapest, Hungary, Oct. 2010.

[ZZGT11] Y. Zaki, L. Zhao, C. Goerg and A. Timm-Giel, "LTE mobile network virtualization", *Mobile Network Application,* Vol. 16, No. 4, Aug. 2011, pp. 424-432.