Small Cells Deployment for Traffic Handling in Centralized Heterogeneous Network

A thesis submitted in partial fulfilment of the requirements for the degree

of Doctor of Philosophy

by

Raed Saadoon

Department of Electronic and Computer Engineering

College of Engineering, Design and Physical Sciences

Brunel University London

United Kingdom

August 2019

Abstract

As the next phase of mobile technology, 5G is coming with a new vision that is characterized by a connected society, in which everything will be effectively connected, providing a variety of services and diverse business models that require more than just higher data rates and more capacity to target new kinds of ultra-reliable and flexible connection. However, next generation of applications, services and use cases will have extreme variation in requirements which in turn amplified the demand on the network resources. Therefore, 5G will require a whole new design that take into consideration efficient resource management and utilisation. An observation that was made throughout this research refers to the demand for more capacity, reduced latency, and increased density as common factors of many of the next generation use cases. This inescapably implies that the use of small cells is an ideal solution for next generation applications requirements, provided that the necessary storage and computing resources need to be distributed closer to the actual user. In this context, this research proposed an architecture of a centralised heterogenous network, consisting of Macro and Small cells with storage and computing resources, all controlled by a centralized functionality embedded within a gateway at the edge of the network. Compared to the basic network, the proposed solutions have been proven to provide overall system performance enhancement. This involves extending the system by adding small cells to serve dedicated services for User Equipment (UE) with dual connectivity from local server which reduces the overall system delay while increasing the overall system throughput. The added centralized mobility management was proven to be capable of tracing the mobility of the UEs within the system coverage, by keeping one connection with the main cell while moving between small cells resulting in enhancement to the handover delay by 11% without service interruptions. Finally, the proposed slicing model demonstrated the system's ability to provide different levels of services to users based on different Quality of Service (QoS) requirements and to differentiate between various applications without affecting the performance of other services, benefiting from more flexible infrastructure than the traditional network. In addition, a 50% improvement in the performance was observed in terms of the CPU utilization. In such architecture, the required capacity can be added exactly where it is needed and when it is needed, coverage problems can be directly addressed, higher throughput, lower latency, and efficient mobility management can be achieved as a result of efficient resource management and distribution which is one of key factors in the deployment of next generation mobile network system.

Publications Based on this Research

- R. Saadoon, R. Sakat and M. Abbod, "Small cell deployment for data only transmission assisted by mobile edge computing functionality," *2017 Sixth International Conference on Future Generation Communication Technologies* (FGCT), Dublin, 2017, pp. 1-6.
- R. Sakat, R. Saadoon M. Abbod. (2019) "Small Cells Solution for Enhanced Traffic Handling in LTE-A Networks". Intelligent Computing. SAI 2018. Advances in Intelligent Systems and Computing, vol 857. Springer, Cham.
- R. Saadoon, R. Sakat, M. Abbod, H. Hasan, "Small Cells Handover performance in Centralized Heterogenous Network". FOURTH 4th International Congress on Information and Communication Technology (ICICT 2019), London, UK.
- Raid Sakat, R. Saadoon, Maysam Abbod, "Load Balancing Using Neural Networks Approach for Assisted Content delivery in Heterogeneous Network". Intelligent Systems Conference (IntelliSys) 2019. London, UK.
- N. Jawad, M. Salih, R. Saadoon, R. Sakat, K. Ali, J. Cosmas, M.A. Hadi and Y. Zhang. "Indoor Unicasting/Multicasting service based on 5G Internet of Radio Light network paradigm". European Conference on Networks and Communications EuCNC'2019. Valencia, Spain.

Declaration

It is hereby declared that the thesis in focus is the author's own work and is submitted for the first time to the Post Graduate Research Office. The study was originated, composed and reviewed by the mentioned author in the Department of Electronic and Computer Engineering, College of Engineering, Design and Physical Sciences, Brunel University London, UK. All the information derived from other works has been properly referenced and acknowledged.

Raed Saadoon
August 2019
London, UK

Acknowledgements

Firstly, I would like to express my sincere appreciation and gratitude to all those who made this thesis possible. This work would not have been possible without help, support and strong motivation of my supervisor, Dr Maysam Abbod, who also gave me the opportunity to work with him in his lab research module and his encouragement, guidance and support from the beginning to the end made my work much easier and more enjoyable.

I would also like to thank my colleagues for their fascinating discussions about the area and the work undertaken.

Lastly, I thank my family for their support, love and encouragement throughout the period. Also I express my sincere thanks to all my teachers, relatives, colleagues, elders and all those from whom I have learnt and gained knowledge.

Thesis Contents

List of Figures

List of Tables

List of Abbreviations

1G	First Generation
2G	Second Generation
3G	Third Generation
3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
6G	Not invented yet, only Trump think it is available
ABS	Almost Blank Subframe
APN	Access Point Name
AS	Access stratum
BSC	Base Station Controller
BTS	Base Transceiver Station
CA	Carrier Aggregation
CAPEX	Capital expenditure
CDMA	Code Division Multiple Access
CDNs	Content Delivery Networks
CN	Core Network
CoMP	Coordinated MultiPoint
CoS	Class of Service
C-RAN	Centralized Radio Access Network
DC	Dual Connectivity
DiffServ	Differentiated Services
DSL	Digital Subscriber Line
E2E	End to End
EDGE	Enhanced Data Rates for GSM Evolution
eICIC	enhanced Intercell Interference Coordination
eMBB	Enhanced mobile broadband
eNodeB	evolved NodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
ETSI	European Telecommunications Standards Institute
EV-DO	Evolution Data Optimized
GBR	Guaranteed Bit Rate
GPRS	General Packet Radio System
GSM	Global System for Mobile Communication
HeNodeB	Home eNodeB
HetNet	Heterogeneous Networks
HO	Handover

HSPA	High Speed Packet Access
HSS	Home Subscriber Server
ICIC	Intercell Interference Coordination
IETF	Internet Engineering Task Force
IMT-2000	International Mobile Telecommunications-2000
ISPN	Internet Service Provider Networks
ITU	International Telecommunication Union
KPI	Key Performance Indicator
LTE	Long Term Evolution
LTE-A	Long Term Evolution - Advanced
MEC	Mobile Edge Computing
MeNodeB	Main eNodeB
MIMO	Multi Input – Multi Output
mIoT	Massive Internet of Things
MME	Mobility Management Entity
MMS	Multimedia Messaging Service
MNO	Mobile Network Operators
NAS	Non-Access Stratum
NFV	Network Functions Virtualization
NS	Network Slicing
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operational expenditure
OSI	Open Systems Interconnection
PCEF	Policy Control Enforcement Function
PCell	primary cell
PCRF	Policy Control and Charging Rules Function
PDCP	Packet Data Convergence Protocol
PDN	packet data network
PGW	PDN Gateway
QoE	Quality of Experience
QoS	Quality of Service
RAT	Radio Access Technology
RLC	Radio link control
RRC	Radio Resource Control
RRHs	Remote Radio Heads
RRM	Radio resource management
RSRP	Reference Signal Receive Power
RSRQ	Reference Signal Received Quality
SC	Small Cell
SCell	Secondary Cell
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDN	Software-Defined Networking
SeNodeBs	Secondary eNodeB
SGW	Serving Gateway
SLA	Service Level Agreement

SMS	Short Messages Services
SNR	Signal to Noise Ratio
ToS	Type of Service
UDN	Ultra Dense Network
UE	User Equipment
UMB	Ultra Mobile Broadband
UMTS	Universal Terrestrial Mobile System
URLLC	Ultra-reliable and low latency communications
UTRAN	UMTS Radio Access Network
WAP	Wireless Access Protocol
WiMAX	Worldwide Interoperability for Microwave Access

Chapter 1

Introduction

1.1 In the Information Technology Era

Personalisation of the computer devices along with the revolution in computer and communication networks paved the way to the current information technology era; in which our world (people and organizations) became more connected. In a more connected and communicative society, the amount of information that are shared and exchanged throughout the network are way higher than ever before, encouraged by the integration of devices (computers, tablets, cameras …, etc.) with the mobile networks which represent an important factor that have a vital impact in the information technology revolution due to the primary role of the telecommunication network in facilitating the transfer of data. In this information era and due to the integration between the powerful computing capabilities and the telecommunication abilities, a vast range of applications and use cases were and are presented as a need, or as industry evolve, resulting in a massive amount of data to be generated and transmitted. From wire to wireless, from fixed to mobile, from analogue to digital, and from voice only service to the data and all IP services; with each step in the journey of the evolution of the telecommunication system, new opportunities and challenges are presented. And what starts as a need for good communication system ends up as a way of living and a trend, especially when the broadband service has become the main source of traffic in the mobile system providing an access to the internet benefiting from the integration with IT networks and supported by the powerful computational capabilities of the smart devices. As a result, the way the people access and use the internet on their mobile devices have been changed dramatically, and with the capabilities of the smart devices there was a wide adoption of the advanced multimedia applications leads to a huge increase in the mobile traffic. The evolution towards smarter mobile devices represent another major trend contributing to the growth of the mobile data traffic as indicated in the Cisco Visual Network Index (VNI). According to Cisco VNI [1], the global mobile data traffic will increase 7-folds between 2017 and 2022, at a compound annual growth rate (CAGR) of 46%, reaching

to 49 exabytes/month by 2022. This mobile data traffic will represent 20% of the total IP traffic, and video content will represent three-fourths of the world's mobile data traffic accounting for 78% of that traffic by the end of forecasted period. The domination of the smart devices is a major reason beyond this mobile traffic as nearly three-quarters of all the devices connected to the mobile network will be "smart" accounting for 86% of the mobile data traffic by 2022. The information era is the inevitable result of the maturity of both technologies, computer systems and communication systems. This is why not only the user's equipment has to be smart, but the next generation mobile system as well, has to be smart system capable of adopting to change in order to provide different levels of services in a timely manner based on the customer's demand.

1.2 Motivations

The integration between mobile networks and computer networks with the advanced capabilities of the devices connected to the mobile network, led to an exponential growth on the traffic being generated. In order to cope with this explosive growth of generated data, traditional macro-cell only networks have evolved to heterogeneous networks (HetNet) in order to improve the performance of the system. While 4G connectivity is predicted to have the majority share of the market for an interval, new structure is required with enough resources or optimized resource management to handle the traffic with an acceptable performance that will increase the end-user quality of experience which is a key factor in increasing the Average Revenue per User for the operators. Backhaul capacity and long latency are among the KPIs that are directly impact the performance and efficiency of the current mobile network.

Therefore, this research describes a common vision on small cells enhancement by dual connectivity for differentiated content delivery service. The proposed architecture is based on the current LTE-A system as a base model with the integration of the SCs and added computing functionality component of the IT technology to enhance the current 4G network. The expected outcome of such design includes:

- Low-latency edge computing being provided by network operators and improve resource efficiency.
- Increased use of small cells to achieve higher data rates.

- Support of multiple different uses over a single network by differentiating service characteristics through network slicing.

The next generation of mobile network or at least the next phase of the current generation will become promising as huge capabilities can results from the convergence of the two industries adding various capabilities and new opportunities.

1.3 Aim of the Research

The proposed architecture aims to enhance the current 4G network in terms of reducing transmission delay, providing higher data rate and backhaul traffic savings with good coverage and enhanced mobility, to provide reliable migration to the next generation infrastructure and to propose new scenarios for next generation mobile system that requires greater service portability, interactivity and interoperability for efficient content delivery to end devices.

1.4 Design Objective

Design of the system involves placement of low latency solutions such as small cells and functionality of the core network to be closer to the end user at the edge of the access network, and investigation of sharing the resources in the core and the ran for network slicing requirement and possibility for per-user services differentiation.
In previous generations the end user is far from the service provider network (i.e. the source of data), and the packets in this connection will pass through many points and links, and each one represents a delay in time and a degradation in throughput if one of the links provides a throughput that is less than the application consumption rate.
In the proposed infrastructure, the sources of the data will be located closer to the user (i.e. the edge of the network) and the delivery of the content will be through the use of the small cells that are in close proximity to the user than the normal cells.

Design of such model take in consideration the following objectives:

- Deployment of content server at the edge of the network to cache and serve content faster to the end user by localizing of the traffic in the network.

- Deployment of small cells to provide specific and differentiated services to the user.

- Benefit of such deployment include:
- Reduction of the latency and free out more bandwidth.
- Reduce the cost per bit resulting from sending any popular content many times over the same communication link.
- Give more flexibility to the network and efficient use of resources through the use of light weight, easy to install devices that can be installed where they are needed and when they are needed.

1.5 Contributions to Knowledge

The advantages of integrating the small cells into the next generation mobile system is the main scope of this thesis. And the contributions made in this demand are explained below:

- Proposal of an architecture of a HetNet system with a centralized controller capable of providing dedicated services from an edge server for UE's with DC. By which multiple small cells are controlled by a centralized functionality embedded within a gateway at the edge of the network for resource co-ordination that will enhance the network performance.
- A centralized mobility management system was proposed to deal with the handover of a mobile UE with dual connectivity in Heterogeneous network in order to improve the handover performance within the small cells coverage. The proposed architecture provides enhanced functionality compared to standardized approach. the Macro-cell (MeNodeB) is the anchor for UEs movement, while the other small cell will act as the SeNodeBs (SCs), which only provide service for the users under its coverage.
- Introduction to an approach of the NS based on the proposed architecture in which different application can be differentiated based on different QoS requirements and benefiting from more flexible network infrastructure than the traditional one size-fits-all approach. The proposed method can deliver assured service quality across the entire network and is flexible enough to accommodate business goals and subscriber needs.

1.6 Thesis Outline

This thesis in total consists of seven chapters, beginning with an introductory chapter to outline the motivation behind the research, with the aims and objectives and the contribution of the research. The remaining chapters are organised to start with an introduction and end with a summary for each chapter while information related to the research are included in each chapter based on the purpose of the chapter as below:

Chapter 2 "Background and outlooks": The first section of this chapter provides an overview to the Mobile Network Evolution, LTE Architecture, Trends and Drivers of the Next Generation System and some key technologies that could enable the implementation of the new system. The second section of this chapter is dedicated to address the current research activities and efforts.

Chapter 3" Implementation of the Key Transformation Technologies": includes an overview of the 5G performance requirements, Constraints and technology enablers in the first section, while the second section is dedicated for The Design Objective and the network Simulator.

Chapter 4 "Small Cells Deployment for Enhanced Traffic Handling in LTE-A Networks": This chapter contain the main design architecture of the proposed system, with the main components of the network, how was it deployed and then the implantation and results discussion.

Chapter 5 "Small Cells Handover performance in Centralized Heterogenous Network ": This chapter presents a mobility-enhanced scheme for improving the handover performance of a mobile UE with dual connectivity in Heterogeneous network taking into consideration decision factors, such as signal power and resource distribution with the performance evaluation and results discussion.

Chapter 6 "Differentiated Service and Network Slicing": This chapter discuss one of the most important aspects in the vision of the next generation wireless network demands; which is network slicing. The chapter describes how small cell networks can be used to meet the requirements of 5G network slicing use cases with provided results discussion.

Finally, Chapter 7 "Conclusion and Future Work": This chapter provides an overall summary of the research with conclusions and future research trends based on the challenges arises with the deployment of the next generation mobile system.

Chapter 2

Background and Outlooks

2.1 Introduction

The first section of this chapter provides an overview to the Mobile Network Evolution, LTE Architecture, Trends and Drivers of the Next Generation System and some Key technologies that could enable the implementation of the new system. The second section of this chapter is dedicated to address the current research activities and efforts.

2.1.1 Background

Communication is an ongoing process that has become indispensable today, especially in today's globalized world, where faster means of communication are changing everyday' s life and had led to the information's revolution. Communication development is largely driven by human needs in both personal and professional lives. It is essential to humanity and to the progress of the societies. Therefore, facilitation the way of communication is the main reason behind the development of communication systems.

2.1.1.1 Data Growth with Mobile Network Evolution

History of the mobile networks has shown that the mobile industry undergoes a major technology shift approximately once every 10 years [2]. It is the technical requirements that necessitate such generation shift. "1G" was analogue with almost no data communication, "2G" is digital with poor data rate, "3G" is faster in data rates, "4G" is an all IP packet network [3]. With each generation will come many applications and use cases which are presented even as a need or as industries evolve or will be introduce as an advantage of new network existing capabilities.

Transition between different generations has been driven by the greater capabilities of each generation over its predecessor. The 1st mobile systems that were introduced in the 80's; later known as the first generation (1G); were an analogue system that were driven by the

need for a communication system to provide voice services in the go and were way far from providing data services. The main limitation of such system was of poor capacity and lack of handover due to the different standardisations of different systems [4] [5].

Digitalization and common standardization represent the major shift towards the 2nd generation (2G), which is well known as Global System for Mobile Communication (GSM) that was introduced in the early 90's [6]. 2G became very popular system and expanded quickly all over the world due to the common standardization that created a single market for mobile phones and due to a better utilization of the frequency bands, enabling new features and services to be deployed and it is the 1st mobile system to introduce the use of data through enabling the short messages services (SMS) of at most 160 characters to be sent between handsets and other stations [7]. Hence, it could be said that data services were born with the existence of the 2G.

Although GSM was widely spread attracted by new services that we still use today such as, SMS, caller ID, call hold, conference calls, internal roaming, along with other network features. Yet still have some issues like, small coverage range, and very slow data transmission rates. Network congestion and drop calls were among the most important KPI's of the GSM system; and users used to hear announcement such as "Network Busy" due to that congestion, and "part of the text missing" due to network and earlier phone data limitations.

Later on, 2G has evolved to provide more features related to data services, these features enable packet data transport which provides relatively higher data rates than SMS services. Such evolutions that presented between 2000 and 2003 were referred to as 2.5G (GPRS) and 2.75G (EDGE). General Packet Radio System (GPRS) offered the first always-on data service with theoretical data rates up to about 144kbit/s, the Multimedia Messaging Service (MMS) was introduced, and internet access through the wireless access protocol (WAP) was first introduced as a basic web connectivity. WAP gave way to EDGE (Enhanced Data Rates for GSM Evolution) to be presented as a bolt on protocol using the same infrastructure as the GPRS and it managed to boost the data rate of the GSM network up to 237 kbps [6] [8].

Till this moment and despite of the introduction of the WAP enabled handsets but it was not so smooth to navigate the web with early presented WAP phones due to the snail like connection speed and the lack of the rich web applications. In spite of that, consumers

“mostly business users” start checking their emails on the go. This trend brought by the new internet ability was the first inkling of what phones would eventually turn into, that is: web-connected mini PCs. But still the low data rate services provided by 2G and its evolutions, 2.5G and 2.75G systems did not fulfil the need for mobile internet access, and didn’t cope with the abilities of the newly presented advanced mobile handset technology (to be known later as smart phones), in addition to the dramatic increase in the number of subscribers that exceeded the expectations and exposing the system to a new challenges such as customer’s satisfaction; This lead to a demand for new generation standards, in order to support high-speed data transfers and inter-communication of mobile devices with the internet.

Hence, it was time to move on again and as a result, a third generation of mobile networks (3G) appeared in 2001, which developed as a family of standards to provide fast data rate services. The most two competing and accepted systems were called UMTS (Universal Terrestrial Mobile System) in Europe and CDMA2000 in America [5]. CDMA stands for (Code Division Multiple Access) which is the access technique used in those two (3G) networks, and represents the fundamental change done to the systems as there is almost not much change have been done to the architecture. The CDMA technique enabled the 3G systems to provide pretty impressive data rate of 2Mbps and more capacity for mobile multimedia services. Improvements were also done to 3G systems, the HSPA (High Speed Packet Access) protocol and HSPA+ were optimized for data and implemented as new addition to the UMTS. This evolution enabled higher data rates of up to 21Mbps leads to the appearance of a number of new applications over 3G networks, including mobile Internet access, video calls, mobile TV, GPS etc [9]. As the second mobile system to support data; after (2G); things got faster with (3G), the huge step forward in data speeds compared to (2G) facilitate the establishment of mobile broadband allowing for fast browsing from the mobile devices on the go. Which in turn led to more data generation and transformation.

It seems that the data rate provided by 3G is sufficient to fulfil the bandwidth requirement for many applications, but the emergence of smartphone has made a major impact. The number of the connected devices increased in an enormous manner due to the high adoption rate of mobile devices and new bandwidth consuming applications arises due to the capabilities of the new devices and the high data rate provided by 3G networks. This increase in the scale of mobile networks affects the performance of the 3G due to the rapid increase in the data carried by system, revealing the need for significant improvements in the Quality

of Service (QoS) of users, which in turn lead to the need to introduce a new generation that could satisfy the new challenges. And this will be the Long-Term Evolution (LTE).

LTE was released commercially just before 2010 as a replacement to 3G (UMTS) to cope with the QoS and rate requirements set by upcoming applications like wireless broadband access, video chat, HDTV content, and many other services that utilize bandwidth. While LTE is much faster than 3G, and was labelled as 4G for marketing purposes, it wasn't considered as "true" 4G by the standardization bodies till further enhancement to the system to be known as LTE-Advanced which formed the fourth generation (or 4G) of mobile networks [10]. 4G aims to provide services for the users at anytime and anywhere with higher data rates of more than 100Mbps in mobility conditions as part of the technical requirement [11]. Among others, 4G was the first generation to introduce all-IP packet switched networks and greatly improved the spectral efficiency of the radio interface to support more simultaneous users per cell, but its labelled 'evolution' to show that the process is not fully completed.

2.1.2 Standardization of the 4G

Standards allow systems to work efficiently, locally and globally; for the ICT industry, standardization has become a key business process. A best example is the worldwide success of GSM as the 2G mobile system. The widespread acceptance of the system was in part due to the collaborative spirit in which it was developed although it wasn't the only system that was presented as 2G. This is in contrast with the early systems represent the 1G mobile system that comprise number of systems that are individually and nationally or regionally developed around the world. GSM among other competitors became a widely accepted, interoperable and robust standard as a result of the collaboration between number of companies harnessing their creative expertise and sponsored by the European Telecommunications Standards Institute (ETSI) to provide a single unified standard [12].

GSM system continue to evolve under ETSI role, and as a founding partner, ETSI transferred further development of the system to the 3rd Generation Partnership Project (3GPP) established in 1998 to produce Technical reports and Specifications for a 3G mobile system [13].

The creation of 3GPP that unites the efforts of seven telecommunications standard development organizations around the world, reflects the considerable devoted attention to the importance of standardization that was a main factor in the GSM success. Furthermore, there were a diverse of standards for developing the mobile networks all over the world. Hence, global decision has been made to have a standard network design to be able to provide services that are independent of the technology platform. Thus, 3G was born [14].

The International Telecommunication Union (ITU), which is the United Nations specialized agency for information and communication technologies – ICTs, and as the main organization to coordinate telecommunication operations and services throughout the world, and for the first time introduced the first family of standards to be a global standard for 3G under the name of International Mobile Telecommunications-2000 (IMT-2000) in order to allow for seamless service evolution from the various standards of 2G systems and to achieve full interoperability and interworking of mobile systems [15].

IMT-2000 standard was first released in 1999, until that time mobile network generations had been developed under different standards, creating a fragmented market. The aim of IMT-2000 is to harmonize worldwide 3G systems on the basis of a single standard of highly flexible system, capable of providing a wide range of services and applications [16].

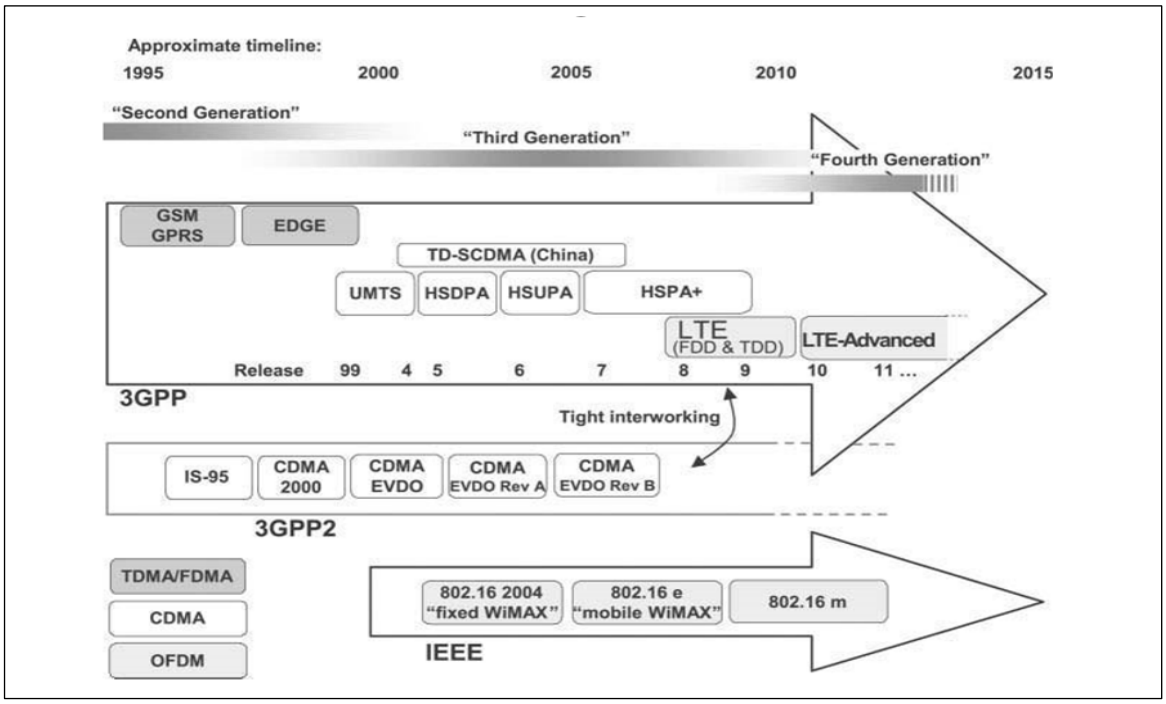


Figure 2.1: Approximate timeline of the mobile communications standards landscape [7].

Deployment of 3G system started quickly based on the IMT-2000 system requirement. Three main organizations were competing in developing standards relevant to that requirements in three evolution tracks, as shown in Figure 2.1.

The outcome was as the following:

The first evolution track of the 3GPP resulted in the development of the UMTS system based on the GSM network and its evolutions.

3GPP2, an organization equivalent to 3GPP and was formed to help development of 3G networks in North America and Asia based on the IS-95 mobile network and resulted in the CDMA2000 system as the 3G system variant in North America and some parts of Asia. 3GPP2 technologies evolved later towards Evolution Data Optimized (EV-DO) and Ultra Mobile Broadband (UMB).

The third track of evolution known as 802.16-2004, and referred to as WiMAX, which is an evolution that has emerged from the IEEE 802 LAN/MAN standards committee and promoted by the WiMAX Forum. The system was restricted to fixed access, then a basic support of mobility was included in the later version of 802.16e and is therefore referred to as ‘mobile WiMAX’.

The aim behind all these efforts is to come out with a mobile system that provides communications multiservice anytime and anywhere with minimum data rate of 2Mbit/s suitable for mobile broadband with seamless global roaming [7].

Rollout of 3G was delayed in many countries mainly due to the costs of further spectrum licensing and the overheads of upgrading equipment to the new systems. Nevertheless, IMT-2000 systems allow for growth in users, due to increased coverage areas, and new advanced services, with relatively low cost. Therefore, with the deployment of 3G and in parallel with continues evolution of 2G, The ITU immediately started developing the vision recommendation on the future development of IMT-2000 and systems beyond IMT-2000 on 2003 [17], and later on 2008 issued requirements for IMT-Advanced, which is initially considered by many entities as a definition of 4G.

Requirements involved operation in up to 40MHz radio channels and extremely high spectral efficiency. Peak spectral efficiency of 15bps/Hz is required and recommended

operation radio channels in up to 100MHz, resulting in a theoretical throughput rate of about 1.5Gbps [18].

2.1.3 LTE

Due to standardization robustness, fewer standards are being proposed for 4G than in previous generations, especially after the announcement of Qualcomm, the lead sponsor of the UMB of 3GPP2 in 2008 to abandon the development of the system and give full support to LTE (Long Term Evolution) of 3GPP as 4G candidate [19]. Leaving only two candidates to compete as 4G: LTE-Advanced and IEEE 802.16m, which is the evolution of the WiMAX standard known as Mobile WiMAX™ from the development track of IEEE [20].

Meeting the IMT-Advanced requirements has been the goal that 3GPP has to achieve and standardization in 3GPP has progressed well. According to studies published in 3GPP technical report [13], it was stated that 3GPP Release 8 (LTE) could meet most of the 4G requirements. Thus, LTE was commercially considered as 4G by operators, while in fact it is yet to meet the requirement of the IMT-Advanced. In 2009, Commercial LTE network deployment started with the term used as 4G LTE as the term "4G" is being linked with the other mobile broadband technologies' competitor being deployed at the time, which is WiMAX. In 3GPP Release 10, also known as LTE-Advanced, it was determined by 3GPP that LTE-Advanced applying technology components were satisfied to meet the ITU-R requirements for 4G. As a consequence, ITU-R has approved LTE-Advanced as IMT-Advanced Radio Interface Technologies (RIT) or "true 4G system" in November 2010 [20].

Figure 2.2: 3GPP Release Timing Chart by Qualcomm.

Figure 2.2 shows the timeline for LTE releases. The first commercial LTE-A network have been launched by SK Telecom, in Korea during summer 2013 with Nokia Networks as a supplier [21]. With the ITU announcement, the competition is settled in favour of the LTE, and with further enhancement for the LTE technology specified in 3GPP Release 9, 10, 11, 12, 13, 14 & 15 Many WiMAX manufacturers have been incorporating LTE into their

product, and many WiMAX operator starts considering migrating from mobile WiMAX to LTE [22].

2.1.3.1 Overall Network Architecture

LTE uses two different multiple access schemes on the air interface one for the downlink and the other is for the uplink. In the downlink, Orthogonal Frequency Division Multiple Access (OFDMA) is used and for the uplink Single Carrier Frequency Division Multiple Access (SC-FDMA) is used. The major technique used for obtaining the high data rates is MIMO (Multi Input – Multi Output) antenna schemes. In addition, LTE has two modes of operation, the Frequency Division Duplex (FDD) mode which requires couple of existing spectrums, one assigned to uplink and the other to be used in the downlink, or a Time Division Duplex (TDD) mode that operates in unpaired spectrum. LTE is developed for a number of frequency bands, ranging from 700 MHz up to 2.7 GHz and supports a subset of bandwidths containing 1.4, 3, 5, 10, 15 and 20 MHz or an aggregation of their integer multiples for each band [13].

LTE system has evolved from preceding mobile networks and the evolved system is referred to as (EPS) Evolved Packet System, which is a combination of the LTE access system and the Evolved Packet Core (EPC).

The EPC network represents the latest evolution of the 3GPP core network architecture based upon the GSM/WCDMA core networks in order to allow easy operations and simplified deployment. While the evolved radio access for LTE is built throughout the UMTS access network (UTRAN) and referred to as the E-UTRAN. In contrast to previous generations, LTE is an all IP network structure that supports both real time services and data services from end to end. The IP address is assigned to the mobile when it is switched on and will be released when the mobile is switched off. Services will be carried by the IP protocol [13].

EPS uses the concept of EPS bearers to provide the user with IP connectivity to route IP traffic from a gateway in the core to the UE for accessing the internet. A bearer is an IP packet flow associated with a defined quality of service (QoS) established between the gateway and the UE. Multiple bearers can be established for a user in order to provide different QoS streams so that a user might be engaged with different services with different

QoS at the same time. Bearers are set up and released by the system as required by applications [7].

Figure 2.3 shows the overall network architecture, including the network elements and the standardized interfaces. At a high level, the network is comprised of the EPC Core Network (CN) and the E-UTRAN access network. While the CN consists of many logical nodes, the access network is made up of essentially just one node, which is the evolved NodeB (eNodeB), which connects to the UEs. Standardized interfaces are used to interconnect each of these network elements allowing the multi-vendor interoperability. This means the network operators have been given the possibility to source different network elements from different vendors. In fact, and based on commercial considerations, network operators may choose in their physical implementations to split or merge these logical network elements [23].

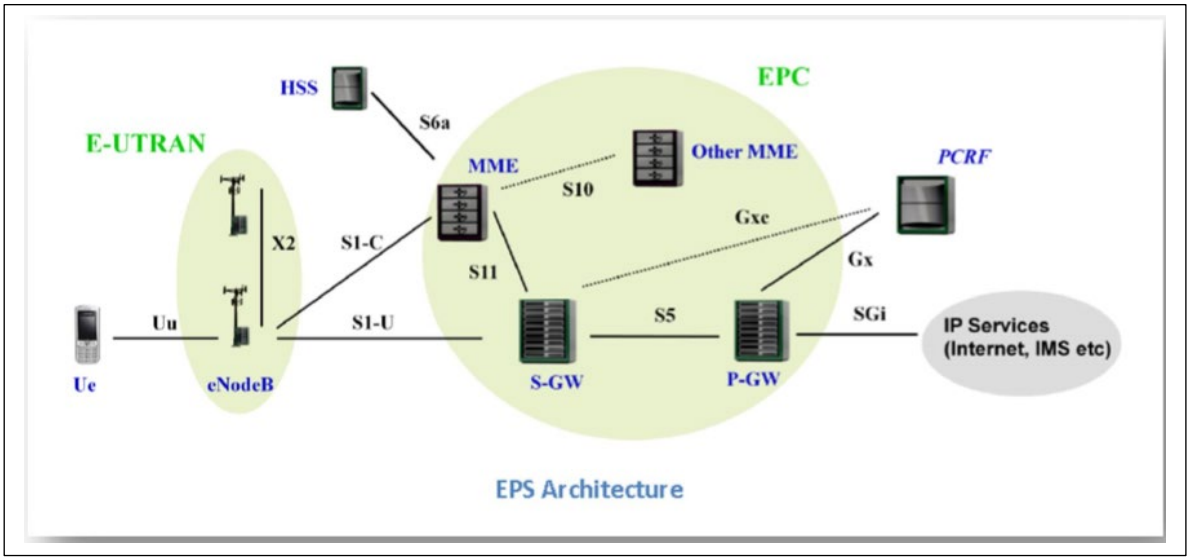


Figure 2.3: EPS Architecture [23].

2.1.3.2 The Core Network

The main elements of core network (EPC) with their functionality are: [13] [23] [24]

- Mobility Management Entity (MME): The MME is the main control node responsible for mobility management, identifies UE and security parameters. The protocols running between the UE and the EPC are known as the Non-Access Stratum (NAS) protocols.

- Serving Gateway, SGW: It is a data plane element, and it is the node that terminates the interface towards the access network. The SGW also sustains the data paths between the

eNodeBs and the PDN Gateways. In addition, when UEs move across areas served by different eNodeBs, the SGW serves as a mobility anchor ensuring that the data path is sustained. For each UE associated with the EPS, at a given point of time, there is one single Serving Gateway.

- PDN Gateway, PGW: Provides connectivity for the UE to external packet data networks, achieving the function of being the point of entry and exit for UE data. The UE may have connectivity with more than one PGW in case of accessing multiple PDNs. The PGW performs policy enforcement, packet filtering for each user, charging support, lawful Interception and packet screening.

Other network elements include:

- The Home Subscriber Server, (HSS): Is a database that contains user-related and subscriber related information. It also provides support functions in mobility management, call and session setup, user authentication and access authorization.

- The Policy Control and Charging Rules Function, (PCRF): Responsible for policy control decision making, as well as controlling the flow-based charging functionalities in the Policy Control Enforcement Function (PCEF), which resides in the P-GW. Provides the QoS authorization.

2.1.3.3 Radio Access Network Architecture

The access network of LTE is simply a network of eNodeB, i.e. base stations, generating a flat architecture as illustrated in Figure 2.4 [24].

There is no centralized intelligent controller, and the eNodeBs are normally inter-connected with each other via the X2-interface and towards the EPC network by means of the S1-interface. To the MME by means of the S1-MME interface and to the S-GW by means of the S1-U interface.

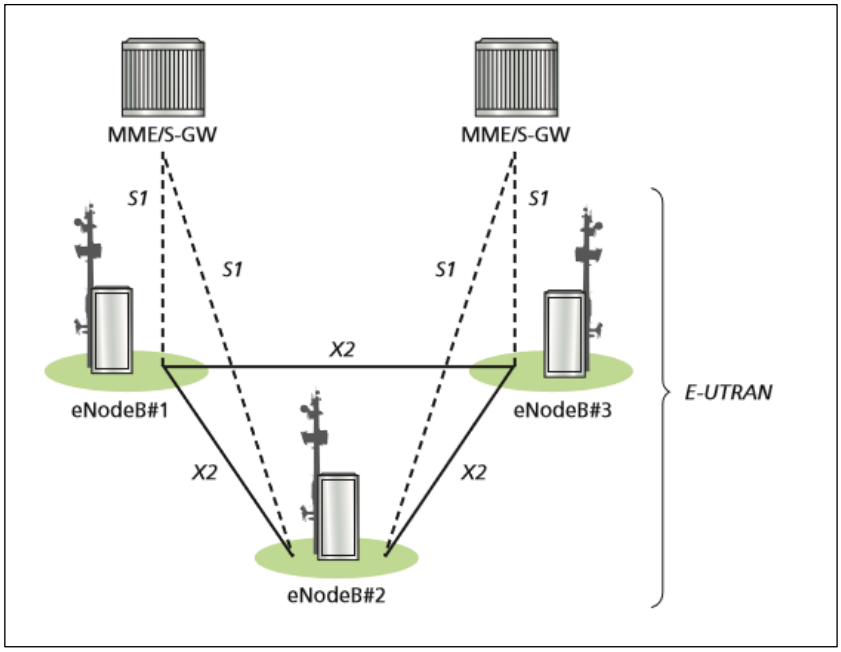


Figure 2.4: LTE access Network [24].

The purpose for distributing the intelligence between the eNodeBs in LTE is to speed up the connection setup and reduce the time required for a handover. As for the end-user a connection set-up time for a real time data session is in many cases crucial. The time for a handover is vital for real-time services such as on-line gaming.

The other advantage is that the MAC protocol layer, which is responsible for scheduling, is only presented in the UE and in the base station, leading to fast communication and decisions making between the eNodeB and the UE.

2.1.4 Next Generation Mobile Network's Key technologies

The mobile/Wireless technologies and due to many reasons such as the dependency on the mobile devices, remote applications, radio capabilities and the integration of different and many networks are becoming the foundation for other industries, including M2M, sensor networks, business-process optimization, and consumer electronics, at the same time, people and machines will have the ability to be connected together in the socio-technical trends as mobile communications has become part of the daily life and closely integrated within the society. Therefore, IMT could play a better role to serve the needs of the networked society in 2020 and beyond and the connected society will characterize the 5G Era. As it is foreseen that there are vast ranges of technology to be developed on the horizon and that most of the

development is based on the higher demand of data and low latency connected to the term of IoT in one way or another, therefore there will be a demand for diverse service requirements with a number of innovative solutions to cope with the varies many more devices seeking a better quality of experience than ever before.

“IMT for 2020 and beyond” or simply “IMT2020” programme development was embarked by ITU in 2012 in order to provide a framework for “5G” research and development, and as of September 2015 the “Vision” of the ITU for “5G” mobile broadband connected society has been finalized [25].

From the viewpoint of IMT2020, the key constraints are set by the following two submission deadlines [26]:

Initial technology submission by ITU-R WP5D meeting #32, June 2019

Detailed specification submission by ITU-R WP5D meeting #36, October 2020

5G will be built to integrate networking, computing and storage resources into one programmable and unified infrastructure. This new model of unified system will allow for an optimized usage of all distributed resources, and the convergence of fixed, mobile and broadcast services. In addition, 5G will enable prioritization of media content and will support multi tenancy models, enabling operators and other players to collaborate in new ways. 5G is a door opener for new opportunities and use cases, many of which are as yet unknown. Meanwhile, 4G is ever-increasing, and its capabilities will continue to develop to support many new use cases and applications, therefore many of the features that could be seen as 5G may in fact be implemented as LTE-Advanced extensions and enhancement before the full release of 5G. This helps to establish 4G as a very solid base for 5G [27] [28].

There are 6 “5G” performance requirements, and 3 usage scenarios associated with this requirement as identified by ITU are as below and shown in Figure 2.5:

Requirement [29]:

- Latency of 1 millisecond end-to-end round-trip delay.
- Peak data of 2 Gbps.
- Connection density of 1 million devices per km^2.

- Area traffic capacity of 10 Mbps/m^2.
- Spectrum efficiency of 3X higher than LTE-A. DL 30 bit/s/Hz; UL 15 bit/s/Hz.
- Mobility of 500 km/h.

Use scenarios [30]:

Enhanced mobile broadband (eMBB): examples include large file transfers, 3D video, work and play in the cloud, where user experience is the main constrain. This use case includes business video conferencing or video games "that require high data rates and medium latency".

Ultra-reliable and low latency communications (URLLC): examples include: Augmented Reality (AR), and eHealth related functions, mission critical applications like a surveillance, self-driving cars, industry automation controls manufacturing equipment. "where high data rates, high reliability and low latency are required".

Massive Internet of Things (mIoT): examples include: Smart cities, smart homes/buildings. Ranging from hundreds IoT devices/km^2 to 1 million IoT devices/km^2. "where low but chatty data rates and high connection density are of high importance".

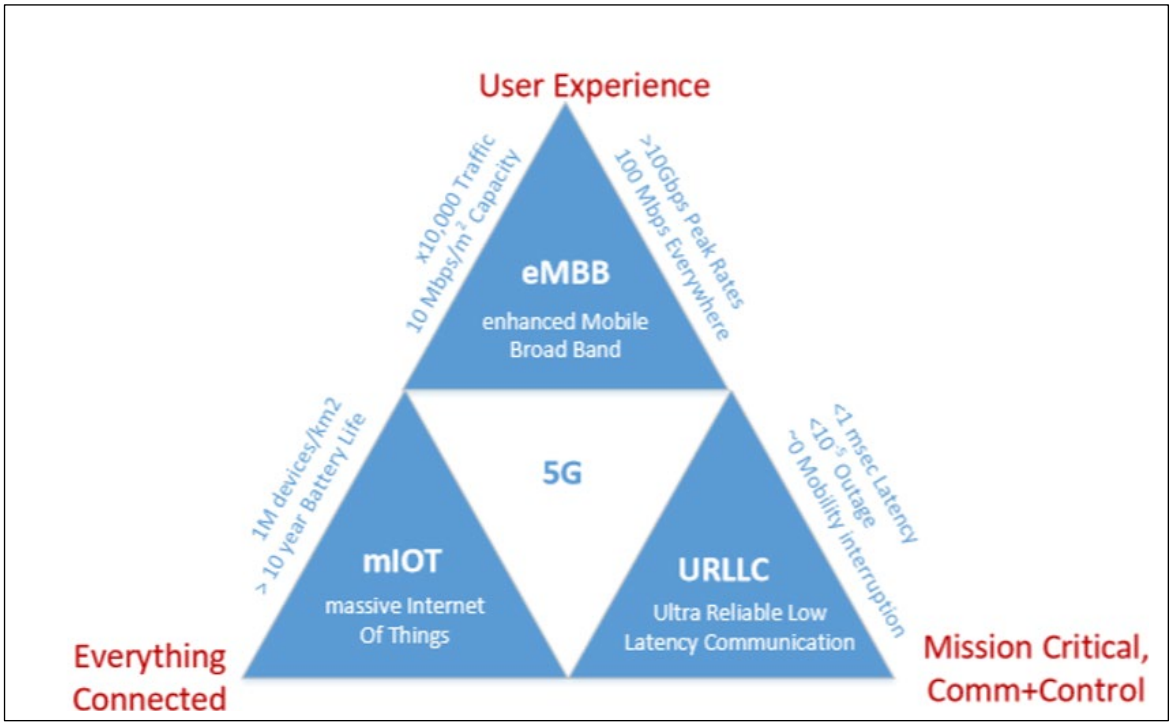


Figure 2.5: 5G use cases (IMT requirement) [30].

The minimum technical requirements for IMT-2020 could possibly be achieved by integrating new technology components and functionalities that will complement the existing IMT system and its enhancements taking in consideration the need to minimize further infrastructure investment and to allow time for customer to adapt to the new services provided by the new system. IMT systems, at the same time, will also closely interwork with other radio systems concerning optimal and cost-effective connection of the users.

Paying attention to the 5G use cases cited above, a best realization is that such use cases could be achieved by employment of radio access points that are close to the end users and supported by computing and storage resources capable to adapt to various services required by diverse type of devices and in this matter we consider the main key technologies that are essential for enabling such use cases to be:

The small cells deployment technique, the mobile edge computing and system functionality virtualization.

2.1.4.1 Heterogeneous Networks and Small Cell Deployments

When mobile network was initially deployed, the deployment scenario was basically based on cells with similar cell size, i.e. homogeneous networks. Later on, and in order to address various capacity conditions; different cell sizes were needed resulting in what is referred to as heterogeneous networks, or Het-Net in which multiple site types were involved.

Unlike the homogeneous network that consists of macrocells only, HetNet is typically comprised of a combination of different access technologies and different cell types of varying transmission power, i.e. macrocells, microcells, picocells, femtocells and either Wi-Fi hotspots. Where macrocells are used to provide coverage while other types of cells are used to enhance capacity in busy areas, such as city centres, stadiums, shopping centres, train stations, offices and at homes.

Data growth and cell size are the main drivers of HetNet. the Bigger the cell site, the less the capacity per person is provided, and as more users are attached to the site the limited capacity of the site need to be shared between the users based on different circumstances. While; If the cell coverage is reduced then fewer users will be served resulting in higher capacity levels and faster data speeds.

Although, HetNet technology is being talked about in conjunction with modern mobile communications networks, particularly with LTE, still the concepts of HetNets is being used with legacy systems (e.g. GSM end UMTS) at the time where small base stations like pico and femto cells were involved to improve the coverage. And the natural progression of the reduction of cell sizes has developed into what is termed small cells in today's 3G/4G networks where deployments of HetNets are widely applied. The small cell deployments are driven by the need for improved network quality, higher data rates and more capacity.

So far most of the base stations have been high power macro base stations while some small cells using low power nodes are being deployed in the advanced LTE networks. Small cells are considered promising solution to cope with mobile traffic explosion, especially for hotspot deployments in indoor and outdoor scenarios. According to 3GPP definition of small cells [31] is that, a node whose transmit power that is lower than the BS classes of macro node can be considered as small cell. In such situation pico and femto cells that have a transmit power of 0.25 W and 0.1 W respectively are both considered as small cells, which is not the case for microcells. 3GPP doesn't specify a separate power class for microcell. Instead a wide area BTS, i.e. macrocell with reduced transmit power could be designed and considered as microcell [32]. In some other scenarios, Wi-Fi hotspots, and Remote Radio Heads (RRHs) within a Centralized Radio Access Network (C-RAN) are also considered for small cells deployment options [33]. As in [31], the scenarios target the deployment of small cells with and without the coverage of macroells, indoor and outdoor hotspots, with both ideal and non-ideal backhaul, with the possibility of small cells to use the same or different frequency band when under the macro cell layer. The distribution technique is also an important factor, therefore the sparse and dense distribution is also considered within the 3GPP TR. Study on the impact of each architecture option will lead to better decision on how to deploy small cells and the service that could be delivered as investigated throughout a study item phase in 3GPP (see [34], [35] and [36]), taking into consideration the user's locations and dual connectivity in heterogeneous network.

In particular the following design goals were addressed in [34]:

- Improve mobility robustness
- Reduce signalling load

- Enhance per-user throughput

Throughout the analysis phase three scenarios were investigated as shown in figure 2.6 below:

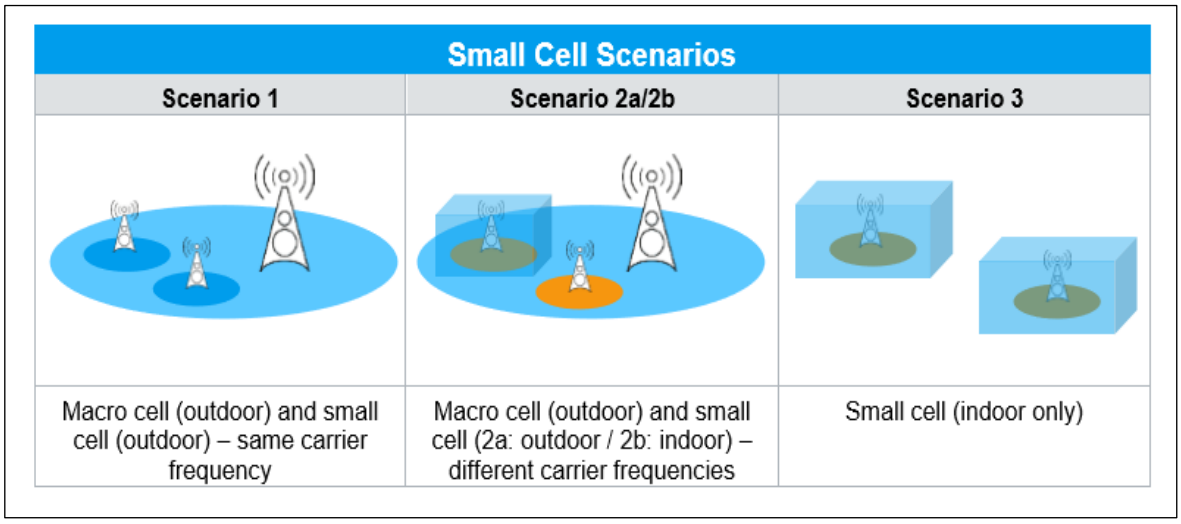


Figure 2.6: Small Cell Scenarios [37]

Regardless of their ability to optimize the system performance, particularly in terms of cell-edge and overall capacity improvement. The emergence of HetNets lead to more complex network architecture resulting in more network design challenges as there are more cells within a given area where a user can be located. Therefore, location management becomes harder in terms of co-channel interference and RF planning, as the various types of BTS typically sharing the same channel bandwidth [32]. One of the obstacles in the heterogenous network is the intercell interference mainly in the boundaries of the cells that can cause handover problems taking in consideration that the mobile users pass through small cells more quickly than a macro cell, hence; handovers must be processed more quickly [37]. Moreover, energy management is also an important issue, since activating small cells nodes at all times may be energy inefficient. Thus, small cell eNodeBs should have some kind of sleep strategy.

Solutions such as intercell interference coordination (ICIC) and enhanced intercell interference coordination (eICIC) techniques as specified by 3GPP specifications in release 8 and 10 respectively can be used to mitigate such issue. In ICIC techniques the signal to noise ratio (SNR) is improved at the cost of reducing the total number of resource blocks i.e. the bandwidth available for transmission, by avoiding the use of the resource blocks used

by neighbouring eNodeBs. eICIC also improves the SNR by benefiting from almost blank subframe (ABS), which allow neighbouring eNodeBs to reduce their transmission during certain time intervals.

2.1.4.2 Mobile Edge Computing

Edge Computing is a network architecture concept that enables cloud computing capabilities and an IT service environment at the edge of the network [38]. When this network architecture is at the edge of the mobile network, and is connected to users/devices using mobile links, then it is referred to as a mobile edge computing (MEC) [39]. The concept originated with the idea to develop mobile content delivery networks (CDNs) for mobile networks specifically after the widespread and popularity of smart phones and then evolved to support the 5G use cases to meet the needs of low latency and balance mobility [40]. MEC can carry out tasks that could not be achieved with traditionally centralized network architectures, in a way that can improve user experience, and minimize congestion in the other parts of the network with computing and storage resources to be distributed to more optimized locations in the edge of the mobile network (i.e. RAN). The correct deployment of MEC nodes in the RAN can offer low latency and rich bandwidth as well as direct access to real time radio network information (like subscriber location, cell load, etc.). This can be used to deploy applications that are capable of differentiating the mobile broadband experience [41]. This is important to operators who start to distribute PGW's for reasons of performance and economics.

Some of the most important aspects of MEC that could influence the architecture of future mobile network and making it a strategic option in the 5G network architecture is the tight correlation between small cells and edge computing, since both distribute resources to be as close as possible to the end user. This can create a new ecosystem when both are deployed together as they are natural development in the evolution of mobile base stations and the convergence of IT and telecommunications networking where new services are developed in and around the base station.

There are several deployment architectures of the MEC server that provides computing resources, storage capacity, connectivity, and access to user traffic which can be used in a mobile edge computing environment, The most popular and MNO-focused option is the ETSI MEC (Multi-access Edge Computing) which is specified and supported by an Industry

Specifications Group (ISG) within the standards body of ETSI focusing on the use of 3GPP standard interfaces to the largest possible [42] as shown in figure 2.7. Other MEC examples include technologies that are originally initiated by companies such as Cisco, Amazon, and Facebook [39].

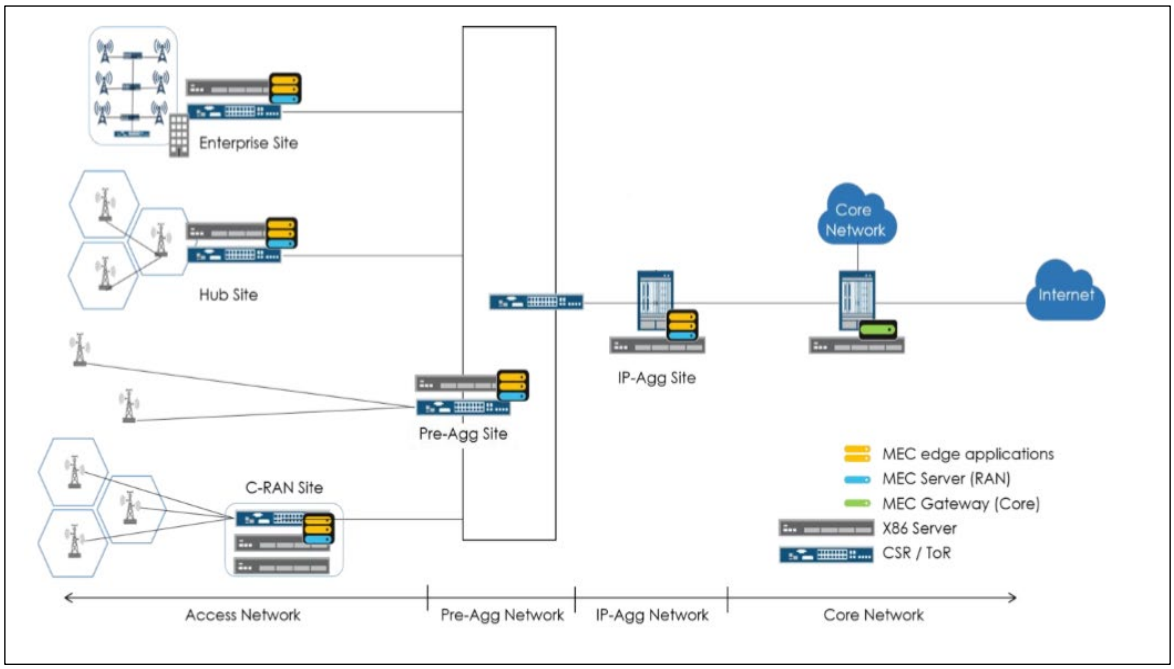


Figure 2.7: MEC deployment in Service Provider Network [Source: Juniper Networks].

2.1.4.3 Network Functions Virtualization and Software-Defined Networking

Virtualization refers to the ability of implementing a network functions of infrastructure nodes programmatically in software instead of by a physical piece of hardware on IT platforms. Modern virtualization technologies, such as Network Functions Virtualization (NFV), allows the separation of hardware from software or 'function'. This approach enables instant scalability and empower the flexibility to execute network functions independently of location [40].

The core network and the radio access network can both be virtualized. Core network virtualisation reinforces growth of existing systems targeting for more flexible network management, hardware consolidation (no need to change hardware for nodes' upgrades), easier multi-tenancy support and faster configuration of new services [41].

RAN virtualizing could provide the highest network efficiency gains, specifically for small-cell deployments where it can simplify coordination among cells and use of methods such as interference coordination and can accelerate RAN cloudification [41].

SDN in the other hand is an extension of NFV wherein the behaviour of the network through wide network programmability can be changed. Thus, SDN can simplify management of networks [27].

NFV and SDN are highly complementary but are not dependent on each other. NFV can be deployed without an SDN being required and vice-versa. although the two concepts and solutions can be combined and potentially greater value accrued such as reduce operator CAPEX, OPEX is also reduced through power savings as network capacity is only provided when and where it is needed, faster deployment of new services, energy savings, and improved network efficiency [40] [42].

2.2 Survey of Current 5G Research Activities

The ITU-R's road map for standardizing the 5G network is fixed, the global program to develop "IMT for 2020 and beyond" was embarked early in 2012 setting the stage for research activities on 5G around the world. The "Vision" of the 5G mobile broadband connected society has been finalized in September 2015. In February 2017, studies on 5G's key performance requirements was completed and ITU-R M.2410 report "Minimum requirements related to technical performance for IMT-2020 radio interface(s)" was adopted in November 2017. Initial candidate radio technologies already submitted to WP5D as of Meeting #31 October 2018 by different working groups with proposals from 3GPP, Korea, China, and DECT Forum & ETSI, final submissions due July 2019. The evaluation process has started and candidates meeting the ITU-R requirements will be an "IMT-2020" technology and ITU recommendation for IMT-2020 will be released by the end of 2020 [43].

A worth mention notice from the development of 4G is that the industry was less interested in developing multiple standards and was much faster to develop a single standard than previous generations. Cost scales and the need for global roaming were the most influencing factors in this manner. This gives an interesting sign for the development of the 5G if the

same factors remain true and a quick convergence toward a single set of 5G standards could be seen instead of several sets of different standards [44].

5G goals are ambitious, aiming to improve capacity, energy efficiency, and reliability, while reducing latency in massively increasing connection density in order to bring 5G KPIs into reality. This can be done by involving seamless operation capabilities of MEC/caching, NFV/SDN in new network topology that employ small cells in the edge of the network which is the main field of researchers in both academe and industry. Many of research groups and alliances have been formed for this purpose, researches from academia and industry will be explored in the following sections.

2.2.1 Current Academic Research

Researchers in the academic field are very active and many researches have been conducted aiming to develop their ideas and views in order to present their visions of the shape of next generation mobile network before the proposed scheduled commercial lunch in 2020. Researchers proposed various techniques, enhancements and architectures within their efforts to address the 6 "5G" performance requirement [29] of low latency, high data rates, more capacity, spectrum efficiency, in addition to connection density and mobility.

In their efforts, researches are considering solutions in different areas of the network, and a considerable amount of the researches are focusing on the main key enabling technologies of the "5G", i.e. UDN/C-RAN/SCs, NFV/SDN, caching/CDN/MEC and Cloud/Fog networks along with backhaul based solutions and carrier aggregation in mmWave [45] [46] [47].

As part of these efforts, many surveys on the 5G network are presented to brief and summarise the achieved works and studies on the architecture [46] [48], and some of the key emerging technologies such as NFV/SDN/MEC based core network [49] [50], data centric network [51] [52], caching [53] [54], backhaul [55], and resource management [56].

Heterogeneity that marks the architecture of the "5G" mobile network was addressed focusing on the use of the small cells concept as an integral part of the "5G" mobile network [57]. Coordinated operation and resources allocation in HetNet to introduce interference reduction and efficient frequency reuse in [58]. Various interference management challenges

in 5G hybrid networks are addressed in a research on multi-tier networks and interference upgradation for 5G communications by researchers at University of Manitoba, Canada [59]. Dense small cell deployment enabling by cooperative and distributed radio resource management were discussed in [60]. Mitigating the unnecessary signalling overhead generated by such small cell deployment for efficient handover decision was discussed in [61].

With the evolution of edge computing and the emergence of the new technologies such as NFV/SDN in wireless networking to provide cloud computing and caching capabilities at the edge of cellular networks, with this emergence, the implementation of the network can be more distributed [62] [63]. Review on the state-of-the-art research efforts on mobile edge networks, including the ETSI MEC, Cloudlet, Fog Computing and edge caching were made in [64] and [65]. Both surveys cover the issues on caching computing, and communication techniques, the applications and use cases. An overview of mobile edge networks including definition, architecture and advantages was given. The key enablers of mobile edge networks are such as cloud technology, SDN/NFV and smart devices are also included and discussed in the surveys.

Employment of such technologies can reduce the CAPEX and OPEX of the network considerably. Furthermore, significant reduction of the E2E delay can be achieved by bringing the elements of core network closer to the users. In [66], SDN and cache enabled heterogeneous 5G network is proposed where C-plane and U-plane are split. The caches of macro and small cells are overlaid and cooperated in a limited backhaul scenario while ensuring seamless user experiences.

2.2.2 Ongoing Organizations and Industry Activities

The globally leading telecommunication venders in cooperation with other mobile operators are working on the infrastructure of the next generation mobile network based on their visions.

3GPP as one of the most focal standardization body is well involved in the 5G definition. The study on the 5G specifications within 3GPP were started in earlier phases, with all the specifications being labelled “5G” from Release 15 onwards to reflect the alignment between IMT-2020 and the 3GPP 5G work plan. 3GPP aims to the definition of a full system

(Radio and Core Network), where the radio being defined as 5G-New Radio (5G-NR) and the core network being defined under the term 5G EPC or simply 5G Core (5GC). According to Chairman of 3GPP RAN, 3GPP is quite confident that 3GPP's submission to IMT-2020 will meet the ITU requirements [13].

Huawei as a major vendor in the telecommunication technology is collaborating with different partners including governments, universities and international associations, to establish crucial 5G innovations mainly focusing on Cloud RAN technology [67] [68]. China Mobile in the same path is strongly supporting C-RAN in its work towards 5G [69].

Ericsson Visual Technology unit is driving research, development and standardization of next generation video compression and media delivery technology [70]. Besides that, Ericsson vision for 5G development is that it should start in a backward compatible way with existing 4G LTE networks. This will help in continuing services using the same carrier frequency to traditional devices [71].

Nokia's recognition for 5G wireless include dense small cells and improved spectrum usage as key concepts of revolutionary 5G [72].

Docomo network has identified two important trends, the first is pervasive wireless connectivity and the second is extensive rich content delivery in real time. It believes the key to 5G deployment is to integrate the higher and lower frequency bands. The lower frequencies will be responsible for basic coverage and the higher frequencies will provide high data rates [73].

Qualcomm and in order to achieve the maximum potential, is driving the development of 4G and 5G in parallel for a unified platform that enables a vast range of new services while improving cost and energy efficiency [74].

Many other telecommunication vendors and working groups are working to bring network infrastructure and UE for 5G roll out by 2020, such as: 5GPP research project initiated and funded by the EU [75], The 5G Innovation Centre (5GIC) at the University of Surrey, which is the largest UK academic research centre dedicated to the development of the next generation of mobile and wireless communications, funded by HEFCE (the Higher Education Funding Council for England) and co-investment from the centre's industry and regional partners [76]. Collaborative research and development efforts between South

Korea, Japan and China have resulted in the formation of 5G forum [77]. “5G Training and Certification” initiated by IEEE, with the process of developing a 5G certification program [78].

2.3 Summary

Data growth is the main motivation behind the next generation mobile network, this growth started in the term of short text messages (SMS), then multimedia messages (MMS), then came the (WAP) protocol to allow access to internet content on the mobile devices which started their revolution to become “smart devices”. At this point, things go faster, as new services were developed, and new mobile applications start to spread benefiting from mobile broadband offered by the mobile networks. As a result, 4G was developed aiming to provide higher data rates and lower latency for the new emerging applications. However, the increased number of devices connected to the network resulting in denser network and capacity demand explosion that wasn’t addressed when 4G was first specified.

Therefore, new requirement in terms of latency, data rate, capacity, spectrum efficiency, connection density and mobility has been set to cover new application scenarios for 5G which is, eMBB, URLCC & mIoT in order to serve billion of devices throughout the world. The fact that there is and there will be billion of devices connected to the mobile network represent a success in the telecom world that wouldn’t be possible without standardization.

Standardization started with GSM as 2G represent a quantum leap in the telecom industry as it moves the system from country or regional based to the globally based system that can provide services for the users anywhere and anytime throughout the rooming service. Therefore, the industry is much keen to have a single standard that give the flexibility for the operators to choose different equipment from different venders in a unified market.

Under this implicit framework, developing the vision based on the ITU standardization and requirements of 5G has started by various groups and many researches have been made discussing the possibilities and giving the vision of what might 5G consist of, bringing the debate of whether 5G should be a revolutionary with completely new system, or to be deployed on the existing 4G system with revolutionary architecture designs that will enable quick and efficient deployment. Bearing in mind that the new key enabling technologies of

the 5G are already started to be integrated with 4G to be able to support many new use cases and applications.

Although the use case scenarios for 5G has been defined, but still many applications could be yet unpredicted, as till now, the mobile system was driven mostly by the human interaction, while in the case of 5G it is expected that billions of devices will join the network and will have the ability to communicate together in the term of IoT and MTC and that could expand the demands for higher levels that could not be predicted.

In this work, the concept of adding computing and storing capacity to the main eNodeB is taken into consideration, in which the content will be cached and stored in a server attached to the main eNodeB. Small cell nodes will be distributed with different frequency band under the coverage of the main eNodeB, and the small cells are connected to and controlled by the server attached to main eNodeB through fiber cables connection, as in the case of C-RAN architecture. In this architecture the system will be moving from the flat platform of the current Lte system that uses only one base station type to be a heterogeneous distributed platform with a centralised management system. In this way the resources of several cells can be pooled in one centralized entity. Centralized processing of the resources would result in efficient interference avoidance and will allow cancellation algorithms to be run across multiple cells in parallel with joint detection algorithms. In addition, the dense deployment of small cells under flexible centralization of the radio access network will allow for flexible functional split based on the virtualization functionality provided by the computing ability at the edge of the network, in this way, the main eNodeB could be used for the normal connection, handling most of the system control signalling, while the small cells could be seen as hot spots used for downloading the required content.

Chapter 3
Implementation of the Key Transformation Technologies

3.1 Introduction

In contrast to the previous generations, the next generation mobile network will be largely driven by data. Therefore, data delivery will signify a rising challenging area for the mobile networks. New methods, strategies, and techniques that provide interworking and harmonization between computer and telecom network in a unified heterogenous ecosystem that is capable of providing computing abilities to the edge should be considered for the implementation of the next generation mobile networks. The key transformation technologies that can enable such implementation like caching, small cells and dual connectivity will be explained in this chapter. Also, in this chapter the simulation platform used to implement the network which is OPNET will be described.

3.2 5G Performance Requirements

The three usage scenarios for 5G defined by the ITU which are (eMBB, URLLC, and mIoT), as well as the set of the performance requirement [29] are mainly focusing on developing a critical infrastructure that is capable of providing services in the dimension of high throughput, reduced latency and increased number of devices/high density [30].

Those requirements and the use scenarios impose challenges as the different type of users in different places will need to use the network for different applications, and all of those users

should be given the best quality of experience whatever the location, application or the type of user.

Achieving those requirements calls for advanced new flexible network architecture capable of providing highly efficient processing and transmission of data. This will require an integration of programmable IT component with computing and storage capabilities so that the required resources could be placed at the edge of the network closer to the users. Moreover, many of the core network capabilities could be placed at the edge of the network benefiting from the virtualization functionality provided by the programmable IT components. In the other hand, small cells in the Heterogeneous networks (HetNets) are idly suited to meet the 5G future requirements.

Indeed, there are several challenges need to be addressed in order to satisfy these requirements. The low latency objective, the high throughput objective and the required flexibility may require a redesign of the network infrastructures. For example, a core network which can be reconfigured and is able to serve users with different requirements, the means of caching at the edge of the network necessitate the use of the key enabler technologies such as (SDN/NFV, EMC/cloud computing, and HetNet) that when implemented together can add flexibility to the network and improve the performance in terms of latency and throughput by bringing the resources closer to the ned user.

3.3 Constraints and Technology Enablers

The target and ambitious goal for any infrastructure that provide services is to do so in an elegant manner with the best quality of service and without constraints. Unfortunately, this is not the case; as each system, including the mobile system has its own restrictions. In mobile and computer networks, constraints are governed by the physical laws that introduce delay and losses as well as throughput restrictions. These constraints are becoming more critical to the services provided by the mobile networks and with the emerging of new applications and use cases that will be supported by the 5G, the requirement for lower latency, higher reliability and throughput are becoming more stringent.

3.3.1 Constraints

3.3.1.1 Latency

The performance of many applications is greatly affected by delays in the network, making it a primary design target. The delay for a packet sent from a server to a client and back is called latency [79]. Latency is highly critical in some applications supported by the 5G, mainly the interactive services such as Intelligent Transportation Systems, control/robotics, tele-surgery, and Gaming [45].

In general, and recalling from the Packet Switching Networks, when a service is requested, the packet travels throughout the network from one node to another starting from the host (i.e. the source) and ends in another host (i.e. the destination). During this journey, the packet experiences several types of delays at each network node along its route. The most important types of delay include processing delay, queuing delay, transmission delay, and propagation delay; which altogether accumulate to give a total nodal delay that can be expressed as [80]:

$$d_{nodal} = d_{proc} + d_{queue} + d_{trans} + d_{prop} \quad (3.1)$$

where: d_{proc}, d_{queue}, d_{trans}, d_{prop} symbolize the processing, queuing, transmission, and propagation delays.

The contribution of these delay components can vary significantly depending on different factors such as the topology of the network. Hence, the end-to-end delay is the accumulation of the nodal delay of all the nodes participating in delivering the service; including propagation delays in links; and end-system processing delays [80].

Moreover, the overall service latency in the mobile networks (including the upcoming 5G system) depends on the delay accrued in different component of the network including the radio interface, data processing (mainly in the core network), transmission within the mobile system, and transmission to servers outside the mobile system [81].

In the LTE system architecture, there are 2 different planes, the control plane and the user (data plane), that connects the UE to the MME and S-GW respectively. Delay in each plane as defined by 3GPP is [81] [82]:

User plane latency: defined as the one-way time it takes to successfully deliver an application layer packet from the layer 2/3 SDU ingress point to the layer 2/3 SDU egress point of the radio interface, in either uplink or downlink in the network for a given service assuming unloaded conditions (i.e., a single user in active state).

Control plane latency: defined as the transition time from a most "battery efficient" state (e.g., idle state) to the start of continuous data transfer (e.g. active state).

The control plane is responsible for signalling and control purpose such as UE registration to the network throughout the attach procedure, to provide the appropriate configuration for the data plane and UE mobility management. While the user plane is responsible for transferring user data. These data will be transferred by the mean of bearers; the bearer concept is an abstraction used in LTE to refer to the flow of data within a certain quality of service. Although the design of the LTE is based on the separation of the control plane and the user plane but still some of the procedures of the control plane also introduce delays to the user data, as full independence of the layers is not feasible due to the interaction required between the two planes to enable some services. Such interaction include signalling for bearer session management [9] [32].

The Attach procedure is the most time-consuming procedure of the control plane, and it is important as it includes the negotiation process to establish the bearers. This procedure will normally have a low impact on the data plane, as it is done once at the beginning of the connection, therefore and since the application performance is dependent mainly on the U-plane latency, U-plane is the main focus of interest for low latency communication [83].

In this regard and taking the definition of the end to end latency from 3GPP in consideration as:

"***end-to-end latency:*** *the time that takes to transfer a given piece of information from a source to a destination, measured at the communication interface, from the moment it is transmitted by the source to the moment it is successfully received at the destination*" [81].

The end to end latency in the LTE system can be calculated as [45]:

$$T_{e2e} \approx T_{Radio} + T_{Backhaul} + T_{EPC} + T_{Transport} \tag{3.2}$$

and as explained in Figure 3.1:

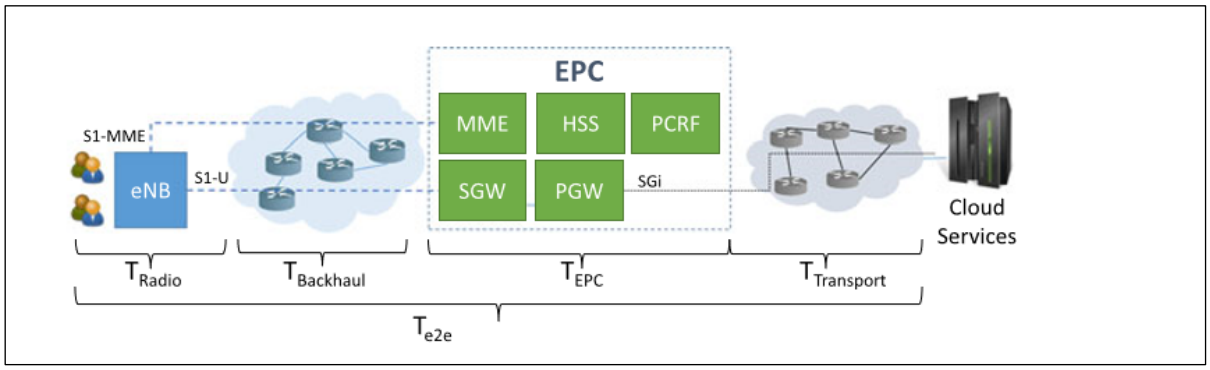


Figure 3.1: Latency contribution in E2E delay of a packet transmission in LTE System [83].

T_{Radio} : The time consumed for packet transmission between eNBs, and UEs. It is the summation of transmit time, propagation latency, processing time, and retransmission time (due to packet loss).

$T_{Backhaul}$: is the time for building connections between eNBs and the EPC (i.e. core network). It is highly depending on the network's configuration as it may include different types of communication technologies (fibre, copper, microwave links, etc).

T_{EPC} : is the processing time taken by the various core network entities.

$T_{Transport}$: is the delay to data communication between the core network and Internet/cloud services. Generally, distance between the core network and the server; bandwidth, and communication protocol affect this latency.

Enhancement in various areas of the architecture need to be carried out in order to reduce the latency, especially for the use cases which require extremely low latencies within the 5G use cases. Approaches of architectures that take the server closer to the base stations, cutting down the time consumed on the transport and backhaul such as caching/fog enabled networks, new core network consists of SDN, NFV, and various intelligent approaches that can provide reduced latency such as the employment of small cells are to be considered.

3.3.1.2 Throughput

Another critical performance measure is throughput. Which is the amount of data per second that can be transferred between end systems. Or as defined by ITU-T as [84]:

"*Throughput is the number of data bits that are transferred successfully in one direction between specified reference points per unit of time. And is a parameter describing service speed.*"

Throughput is fluctuating depending on the transmission rates of the links over which the data flows, and as more devices start using the network, they will be sharing the same bandwidth along the network route, and this may lead to throughput degradation. Consequently, leading to overall system degradation, in terms of spectral efficiency and overall QoS, as less throughput may mean higher transmission time (i.e. less speed) or less available bandwidth (i.e. capacity).

Bandwidth and speed are defined by Cisco as [85]:

"*Bandwidth is the capacity and speed are the transfer rate*"

Many applications have throughput requirements and are said to be bandwidth-sensitive applications, such application require high data rates and medium or low latency such as large file transfers, video conferencing or video games [80].

3.3.1.3 Reliability

Reliability is a critical design principle for 5G, as according to NGMN 5G White Paper [86], a reliability rates of 99.999% is required to be provided by 5G technology for the use cases that demand it.

In general, the probability of transmitting data of X bits size successfully between two nodes and within a certain time period is referred to as the reliability. Or as defined by 3GPP [81] to be," *the percentage value of the amount of sent network layer packets successfully delivered to a given destination node within the time constraint required by the targeted service, divided by the total number of sent network layer packets*".

Reliability is related to flexibility, anticipating that 5G network will operate in a heterogeneous atmosphere, the flexible combination of the different technology components will be used to provide the most reliable link based on the user's application needs, location and mobility. Moreover, the flexibility of the heterogeneous environment will provide efficient trade-off mechanisms between link reliability, throughput and latency in order to provide the required QoS for the 5G system [81] [87].

ITU referred to the reliability in relation to the ability to provide a given service with highest level of availability [29]. The availability is the probability that a given service is available, the relationship between the reliability and availability in the communication service can be depicted in figure 3.2 below:

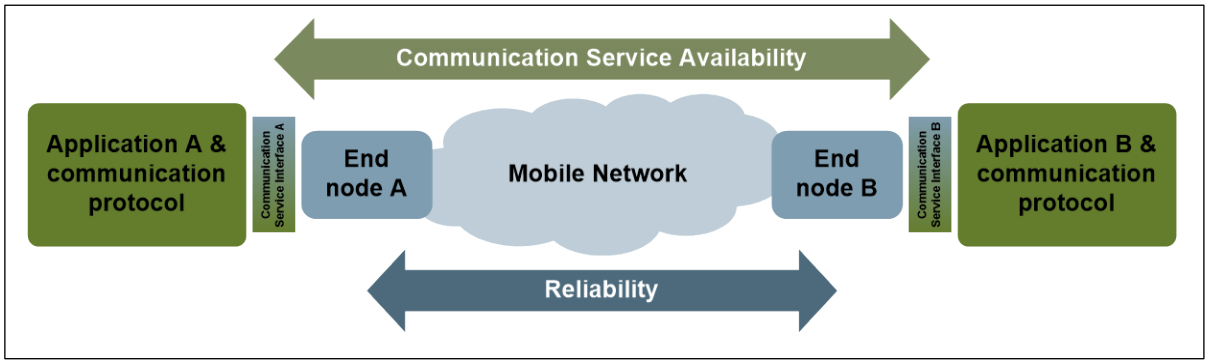


Figure 3.2: Illustration of the concepts communication service availability and reliability [81].

As depicted in the figure, reliability related to the availability of a communication link between two nodes (here: end nodes A & B) and covers the communication-related aspects between them; while communication service availability relates to the availability of a communication service between two communication service interfaces (here: Interfaces A & B) and covers the communication-related aspects between them. In other words, the "gap" between both terms is the communication interface. This might be look as a small difference, but it can lead to situations, where reliability and communication service availability have different values. However, based on the agreed QoS, reliability can be the same value or higher than the communication service availability [30] [81].

3.3.2 Technology Enablers Effects

In order to resolve the above-mentioned challenges, diverse technology solutions are being addressed, mainly focusing on the parameters that have critical impact on the overall network performance and the quality of services provided and hence the overall user experience. The QoE provided to the divers' number of users using the network should be pretty much the same whatever the application or the location and regardless of the type of the user.

Therefore, next generation mobile network architecture needs to address in the design the critical KPIs related to throughput, higher capacity, higher data rates, lower E2E latency, availability and reliability. A Key enabler solution for this task include the ability to provide network functionality and content delivery services at the edge of the network based on an ecosystem of adaptable components capable of accelerating the service delivery by introducing elegant integration of IT and Telecommunications Networks in order to bring new possibilities and capabilities.

A promising technique to meet the challenges of the new applications is to bring the communicating endpoints (i.e. small cells) in close proximity to the mobile subscriber and adding more intelligence at the edge of the network in order to provide IT and cloud computing capabilities to the edge system. The benefit of such approach is that: E2E latency can be reduced, efficient improvement of capacity and data rates at critical locations, network can perform a high degree of agility, adaptivity, and flexibility, and finally the vision of the next generation mobile network to be seen as "network of (virtual) functions" instead of a "network of entities", as in legacy systems could be achieved [88].

3.3.2.1 Small Cells

With the fast-growing data traffic, more capacity will be called for. Three main components for adding more capacity to the network as shown in Fig 3.3. Adding more spectrum and increasing the spectrum efficiency are two methods used to squeeze out additional capacity from the macro cell sites [33].

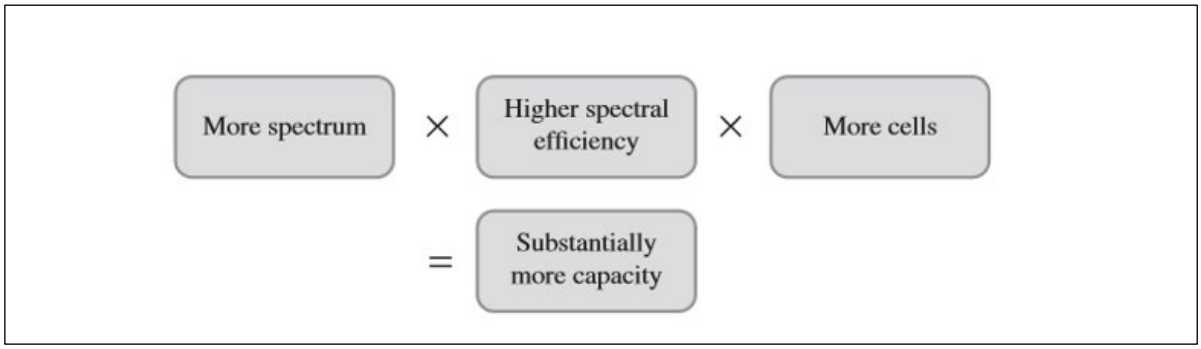


Figure 3.3: Three main components to increase network capacity [33].

Adding more spectrum like higher spectrum bands, and allowing more capacity from the same spectrum using innovative approaches such as Carrier Aggregation (CA) and advanced antenna techniques like MIMO and Coordinated MultiPoint (CoMP) as employed in the

LTE-A are effective solutions in dealing with the challenge of increasing data capacity of today's networks [89]. Although, each one of these solutions played a pivotal role in increasing the capacity and improving the user experience, SCs in HetNets will add a new dimension and will bring out the most of these enhancements.

Small cells will be the glue that bind these solutions, mainly when these solutions are exhausted, for example, higher frequency bands when added, will provide smaller coverage footprint hence lending itself very well to small cells. On the other hand, while the small device form factor represents a challenge of fitting all the antennas for the MIMO technology, small cells will be highly beneficial in enabling MIMO techniques by using smaller antennas that are in tight correlation with the shorter wavelength of the higher frequencies [89].

Therefore, it can be seen that the most gain of the three solution is coming from consolidating the SCs in HetNets for the capacity reason. Another important factor is the non-uniform distribution of the customers making the use of small cells a highly practical and efficient solution to improve the wireless capacity and the overall performance by placing the small cells in high traffic hotspot areas closer to the end user [33].

Due to these features, small cells gain more importance and development of small cells become critical in realizing the ITU requirement for the next generation network. An important feature within 3 GPP small cells improvement is the support of dual connectivity which is an important extension to the principle of carrier aggregation in which the UE will have simultaneous connection to both macro cell and small cell, allowing maximum data rate to be aggregated from both cells [31].

Moreover, with its close proximity to the users, small cells can effectively support network slicing by delivering capacity and throughput requirement for diverse RAN slices. This will make efficient slicing orchestration and will accelerate critical service deployment of next generation system [30].

3.3.2.1 MEC/CDN/Caching

Bringing the communicating endpoints closer to the end user by deployment of small cells will effectively address the requirement of capacity and data rates. In the same way, flexible network deployment and optimal use of network resources can be achieved by adding more

intelligence and storage capability to the edge of the network. Edge existence is viewed as absolutely necessary to enable certain defined 5G use case classes as all call for some compute/storage capacity and data processing to be allocated at the edge of the network [90].

One approach that can add IT and storage resources within the RAN while providing virtualization platforms to contribute to both reliability and latency constraints is the use of Mobile Edge Computing platform (MEC). MEC platform will empower the flexibility of the network to allow instant execution of various virtualized functions to be operated in a virtualized environment, and enable faster applications and content delivery through distributed service and local content caching technique to improve network efficiency and user experience [91].

The origination of the MEC concept arises after the widespread and popularity of 3G smart phones, with the idea to develop mobile content delivery networks (CDNs), and later took a step forward to support the use cases of 5G. Furthermore, the MEC concept was used to add compute capability to the base stations, and most recent evolution in the concept is to draw on network function virtualization (NFV) technologies to allow the network services to be operated in a virtualized environment [40] [91].

With a complementary approach to NFV, the MEC concept and based on the virtualization platform with the aims to place compute and storage resources in the RAN in order to boost the delivery of content and applications to the end user; all of this place MEC into a broader and more strategic consideration when discussing network architecture concept and distributed computing capabilities in the journey to 5G, specially as far as the demanding KPIs of 5G are concerned [91] [92].

As depicted in Figure 3.4 below, the proposed MEC architecture consists of a hosting infrastructure and an application platform. The hosting infrastructure include the connectivity to the radio network element (i.e. eNodeB) and consists of hardware resources and a virtualization layer. Multiple implementation options can be used to integrate the server within the RAN. While The MEC application platform provides the capabilities for hosting applications allowing 3rd party applications to register and get authenticated/authorized, in order to access a set of services [93] [94].

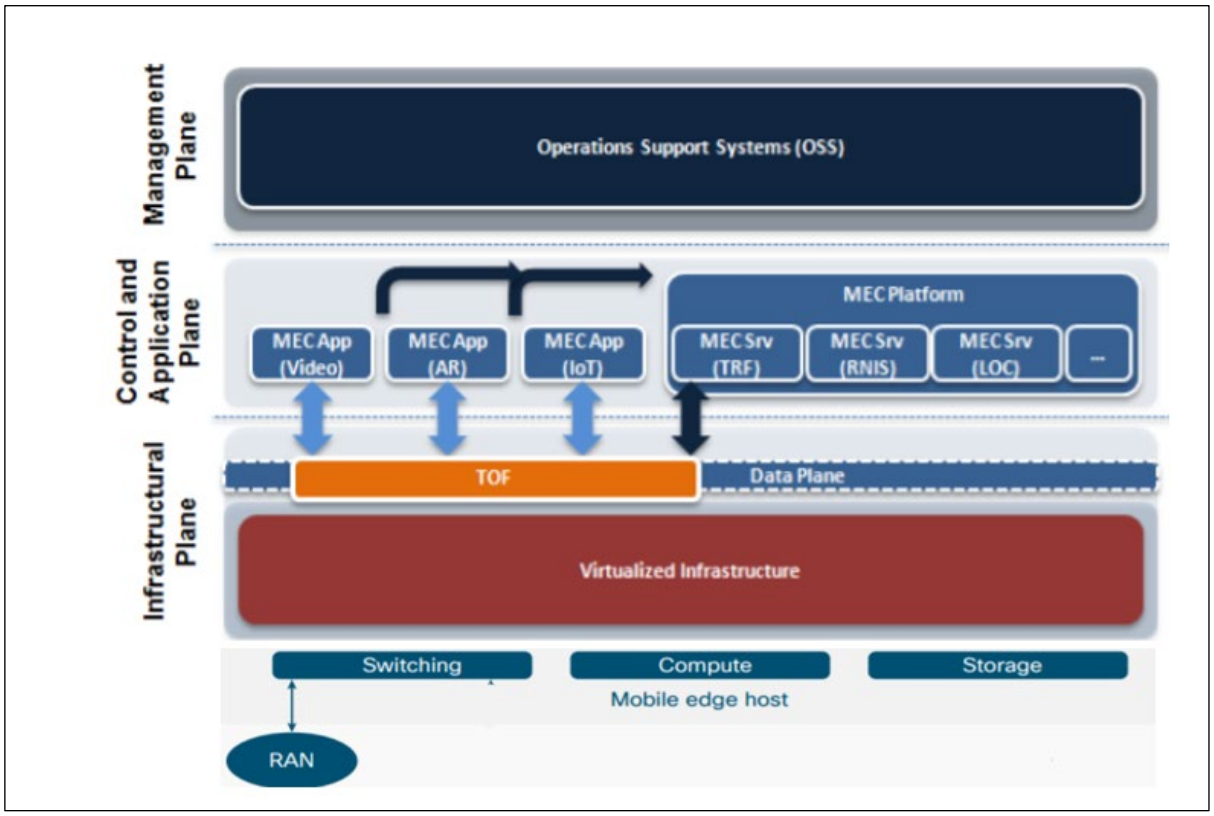


Figure 3.4: Architecture of MEC integration and management [93] [94].

3.4 The Design Objective and the Simulator

In this section a description of our design will be given, and the simulation platform used to implement the network which is OPNET will be described.

3.4.1 The Design Objective

The usage scenarios of the next generation mobile network (i.e. eMB, URLLC, & MIoT) with their associated use cases such as car self-driving, telemedicine and online gaming will require higher demands of network availability, very low latency and very high bandwidth. These requirements will be a key driver for 5G as users will always require more data due to the exponentially increase of the everyday new different applications. And thus, data delivery will represent a challenging area for the mobile networks, therefore it is necessary to find solutions for many of the use cases that require such needs.

In all generations of mobile system, the data delivery (i.e. internet applications) were served to the end user from an outside network (i.e. Internet service provider networks (ISPN)) which are connected to the core side of the mobile networks (ex. the PDN gateway in LTE

system). In other means, the mobile networks were used as a bridge between the end users and the data delivery networks. Although facilitating the exchange of data among the end users, the mobile system is without no concern with the applications that are the source or sink of data. Mobile networks are mainly acting as an infrastructure that provides services to applications. Importantly, many users today access the internet through this infrastructure, and most recently, many end users are wirelessly connected to the Internet via a mobile network [80].

In this infrastructure, the end user is far from the service provider network (i.e. the source of data), and the packets in this connection will pass through many points and links, and each one represents a delay in time and a degradation in throughput if one of the links provides a throughput that is less than the application consumption rate. Another drawback is that not only increase in the bandwidth consumption but will also increase the cost per bit for sending any popular content many times over the same communication link.

A main solution is to bring the data sources to locations closer to the end users (i.e. the edge of the network) through the use of Content Distribution Network or Content Delivery Network (CDN) thereby localizing much of the traffic in order to reduce the latency and free out more bandwidth.

Placing the CDN in the edge of the network making it highly complementary with the small cells and MEC technologies, as CDN was the main motivation behind the development of the MEC [40].

3.4.2 Content Distribution Network

A CDN is a system of globally positioned edge servers that delivering content on behalf of the original server [95]. CDNs, carry nearly half of the world's Internet traffic by managing servers in multiple geographically distributed locations (i.e. edge servers), cashes and stores copies of the content (mainly videos in addition to other types of documents, images, and audio) in its servers, and the end users are directed to use the nearby edge server that will provide the best user experience [70].

Two main strategies used in CDN to replicate contents based on the popularity, when the content is highly popular it will be pushed to the edge servers, when the content is rarely

used or only popular in some locations, then a simple pull strategy is used; Similar Web caching, if the end user requests a content from a server that is not storing it, then the content will be retrieved from an another server or a central repository and the first server will locally stores a copy while delivering the content to the end user. In the same way, the sever removes contents that are not frequently in demand to free up its storage [80]. This is depicted in Figure 3.5.

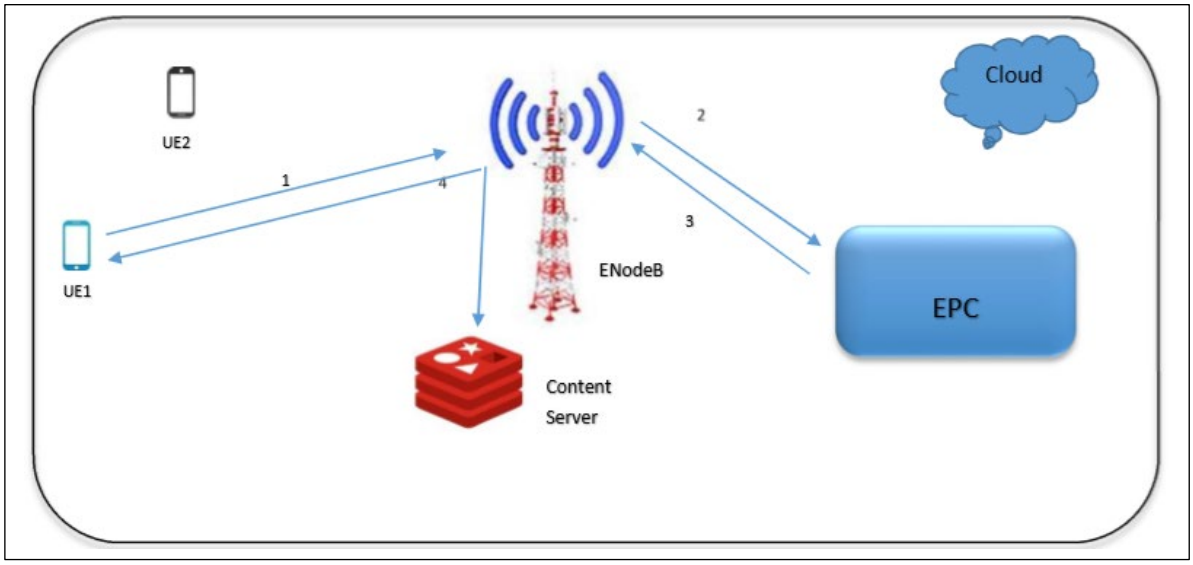


Figure 3.5: Content caching and delivery mechanism using edge server (A).

In figure 3.5: Content Caching (A), UE1 will request data from the network which will be provided by the core network from its connection with the outside network, and at the same time the information will be cached or stored in a content server attached to the eNodeB as Follows:

1- UE1 will request the data through the air Interface.
2- eNodeB will redirect the request to the EPC.
3- EPC will provide the required data after getting them from the outside network.
4- The data will be provided by the eNodeB and at the same time will be stored in the content server.

Next time when another device will request the same data, the data will be served from the content server rather than the external network as shown in Figure 3.6: Content Caching (B)

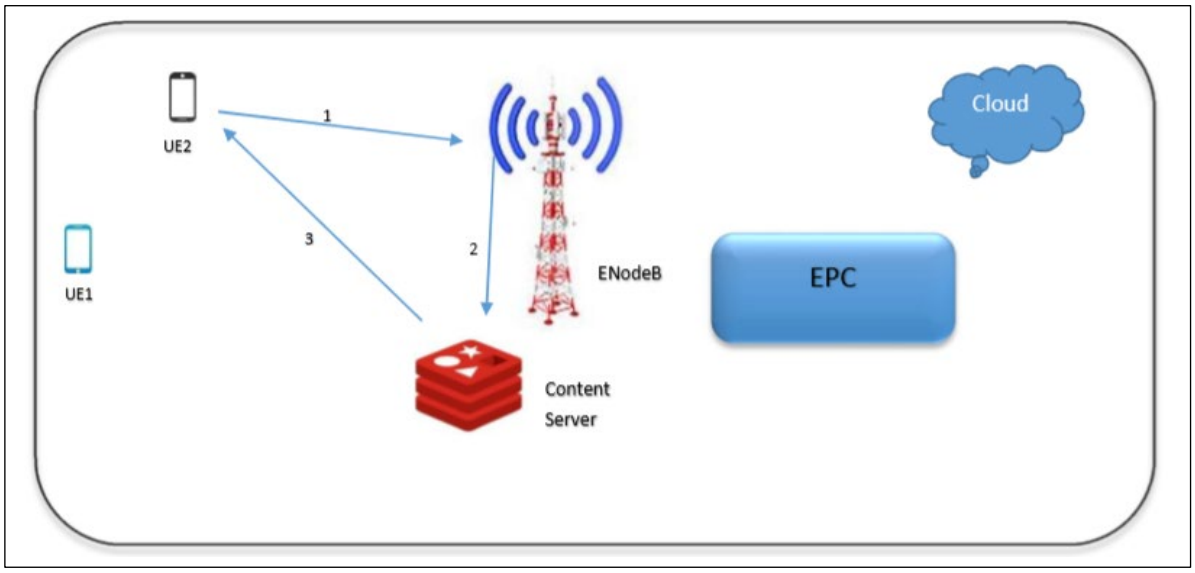


Figure 3.6: Content caching and delivery mechanism using edge server (B).

1- UE2 will request the data already requested by UE1 and stored in the content server.
2- eNodeB will check the content server for data availability.
3- If the data still available, UE2 will be served by the content server directly.

Content server could be expand to serve more than 1 eNodeB (e.g. small cells), or a cluster of eNodeBs (e.g. C-RAN), taking in consideration that this solution may require an SLA between the ISPs or data providers and the mobile operators in the real world cases if the solution will be successful.

The main benefits from using CDN are:

1- Faster content delivery - latency and packet loss are minimized due to a shorter distance travelled.
2- Decrease Server Load- more resources for other tasks.
3- Increased reliability.
4- Scalability - Scale up or down within a short amount of time.

Finally, CDN could be private, that is, owned by the content provider itself, or alternatively be a third-party CDN that distributes content on behalf of multiple content providers. The two main companies in this field are AKAMAI As the largest third-party CDNs distributed platform operating at the edge of the internet, with more than 240,000 servers in over 130 countries [95] and GOOGLE with its own private CDN that serves around 30 to 40% of the total internet traffic [96].

3.4.3 The Network Simulator

An important methodology in network research field is the use of network simulation, and in our research, we will be using OPNET Modeller network modelling and simulation.

OPNET (currently known as Riverbed) is a powerful high-level event-based network level simulation tool with wide variety of possibilities that enable the simulation of entire heterogeneous networks with various protocols. The friendly design of its graphical user interface (GUI) makes it nice and easy to start with.

OPNET modeller environment support the modelling of communication networks and distributed systems with a huge library offers more than 400 ‘out of the box’ protocols and vendor device models including TCP/UDP, IPv6, VoIP/Video/FTP/HTTP/Email, WiMAX, LTE, UMTS, WLAN (a/b/g/n) etc, to support accurate event driven simulation scenarios [97].

Despite the fact that the simulator incorporates tools for all phases of study, but there still exists typical lack of the up to date systems; therefore, much of the work considering recent modern topics must be done by oneself.

3.4.4 OPNET Modeller Architecture

OPNET Modeller presents its capabilities to support model specification (i.e. the task of developing a representation of the system to be studied) with a number of tools, called editors. The capabilities offered by these editors to handle the required modelling information mirror the types of structure found in actual network system.

Therefore, the model-specification editors are hierarchically structured, in a manner that naturally paralleling the structure of real communication systems.

The main editors in OPNET with the interaction between the editors' domains are shown in Figure 3.7.

- Project editor
- Node editor
- Process editor

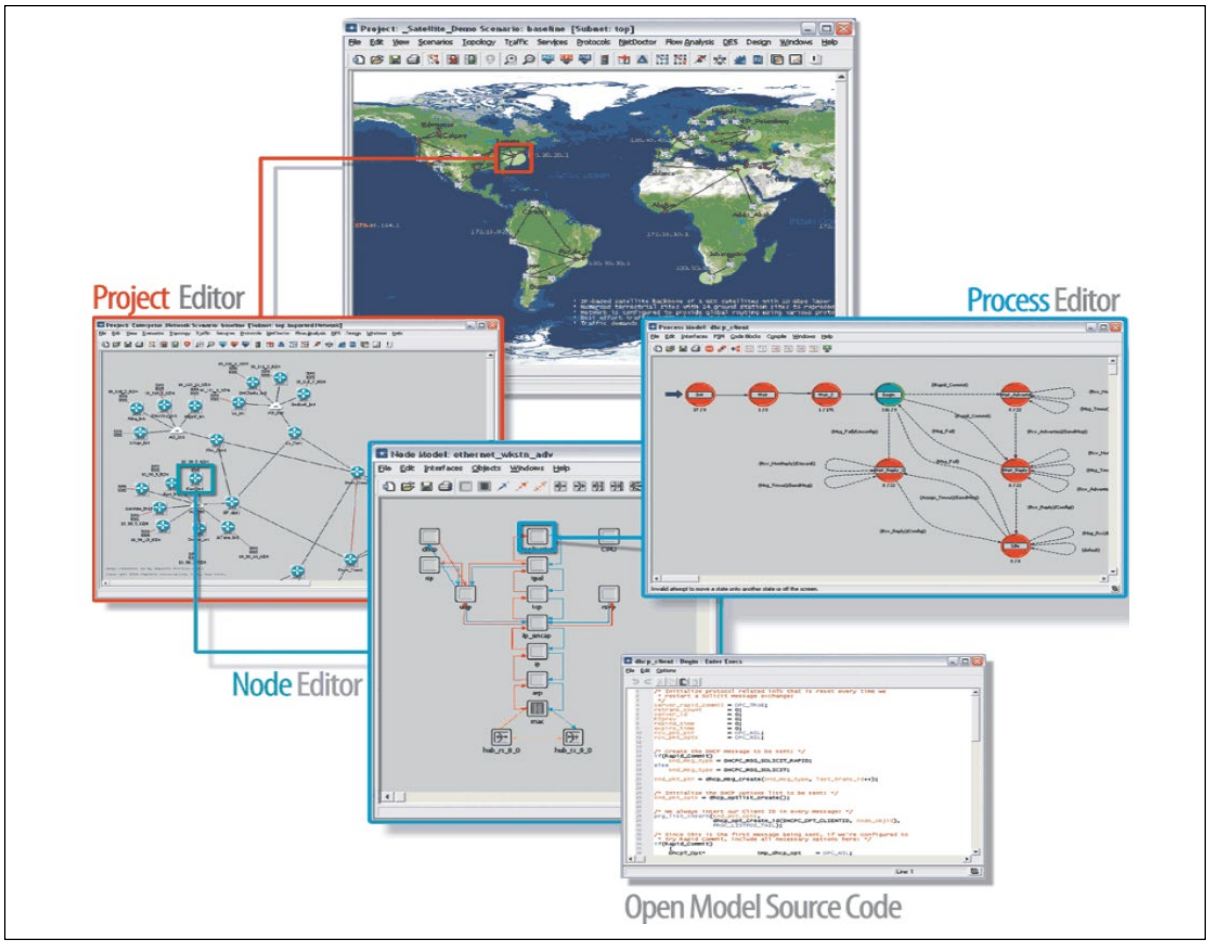


Figure 3.7: OPNET modeller components [97].

Each editor captures the characteristics of a modelled system's behaviour and allows users to set of related functions within a window that is contained in the overall graphical environment.

As the editors are hierarchically organized; therefore, models built in the Project Editor rely on elements specified in the Node Editor; in turn, when working in the Node Editor, we use models defined in the Process Editor and External System Editor. The remaining editors are used to define various data models, typically tables of values that are later referenced by process or node level models.

The project editor, which reflects the network domain, constitutes network topology and specifies the overall scope of the system to be simulated. It is a high-level description of the objects contained in the system. The node editor generates node models and specifies the internal structure of a network node by reflecting the OSI reference model. The process editor develops a decision-making process models representing protocols, and algorithms

used to specify the behaviour of processor or queue modules, which exists in the Node Domain. Basically, process editor is a tool used to expresses process models in language called Proto-C to ease the development of C or C++ source code of the desired model.

3.4.4.1 LTE Networks in OPNET Modeller

The Riverbed LTE Modeller 18.7 is based on the 3GPP Rel.8., with the following node models are included in this version of the OPNET Modeller and depicted in Figure 3.8:

- LTE Attribute Configuration Object
- LTE UE Router Model
- LTE Workstation Model
- LTE server Model
- LTE eNodeB Model
- LTE EPC Model

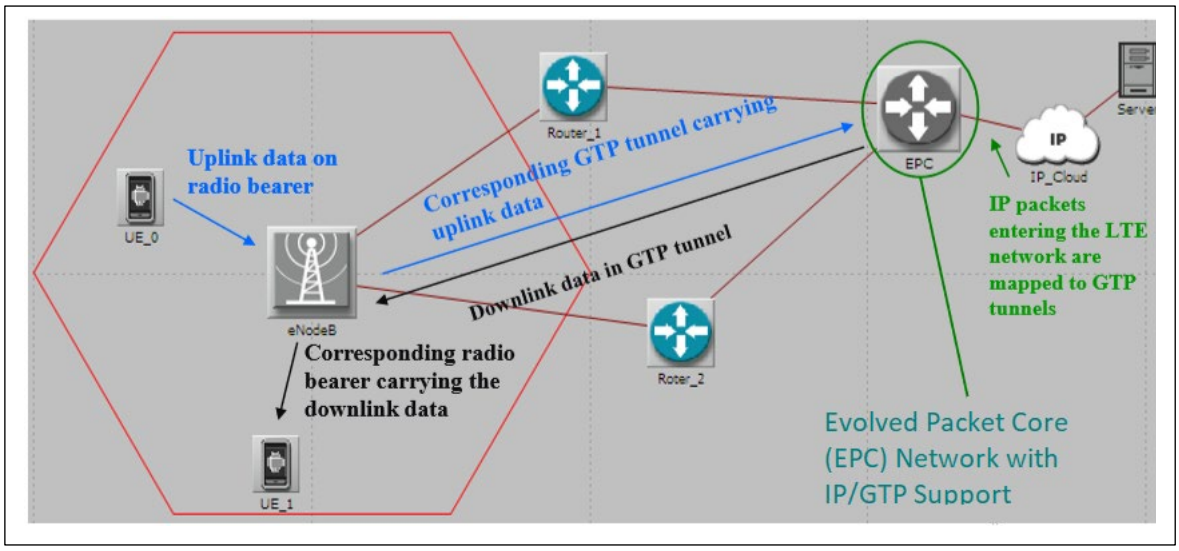


Figure 3.8: Typical LTE Network Architecture with OPNET Modeller.

The following topologies are supported in this model:

- An eNodeB can communicate with one or more UEs and with one or more EPCs.
- An EPC node can communicate with one or more eNodeBs.
- Applications can be configured to communicate UE to UE, UE to external network, and external network to UE.

With the consideration that a single EPC device models the MME, SGW, and PDN, an eNodeB can serve multiple EPCs, and EPC can be served by multiple eNodeBs. Also, only intra-EPC handover is supported.

3.5 Summary

The 5G use cases will not only require improved network solution such as small cells deployment but will call for more data processing at the RAN. A solution that integrate computing and storage infrastructure at the edge of the network enabling cloud and IT capabilities in close proximity to the user. Such environment can provide measurable benefits in different directions including cost, scalability, reliability, and simplified management.

In this environment, the integration of various enablers that bring the communication endpoint closer to the end user through the use of small cells, and adding more intelligence with storage and processing functions at the edge of the network such as the use of MEC, the deployment of new services and the expansion of the network will be more flexible as the hardware is becoming more generic and the functionality is more software dependent. Such architecture can improve the resource utilization efficiency by maximizing the associated sharing gain and can effectively address the challenges of both capacity and throughput.

A key transformation for such architecture, is the ability to provide network functionality and content delivery at the edge of the network, based on Mobile Edge Computing borrowed from the IT network while providing the service by deploying a heterogeneous Radio Access Network. In this architecture, the deployment of the macro and SCs is on different carrier frequencies (inter-frequency) where the SC will be distributed as hot spots covering specific areas under the coverage of the macro-cell layer. The proposed system architecture and its implementation on OPNET will be presented in chapter 4.

Chapter 4
Small Cells Deployment for Enhanced Traffic Handling in LTE-A Networks

4.1 Introduction

The current mobile network (i.e. 4G), has provided users with considerably faster data speed and lower latency, changing the way people accessing and using the internet on their mobile devices in a dramatic way. This leads to growth in mobile population and encourages the introduction of new applications and use cases which represents a major shift in the paradigm from voice to data especially with the adoption of the IP network from the internet to mobile networks.

This evolution of mobile networks from analogue providing voice only service to an all IP packet system providing multimedia services had reinforced the network capabilities leading to an explosion in the usage of mobile applications and resulting in a huge amount of data being generated which impose challenges to the current architecture of mobile networks and

represent additional burden of the already loaded network in terms of capacity and efficiency which in turn result in more delay and affect the user experience.

During the evolution of the mobile system, the network had examined different fundamental changes and challenges. The structure of the mobile network itself passed through different changes, in the two main parts of the structure: the core network and the radio network. Furthermore, the next generation of mobile network is not apart from that evolution specially when the mobile/Wireless technologies are becoming the foundation for other industries, including M2M, sensor networks, and many more.

Therefore, the deployment of the next generation mobile networks requires an ecosystem of changeable components that represent an integration of IT and Telecommunications networking in order to bring new possibilities and capabilities that are key sources to the Average Revenue per User for operators. A key transformation is the ability to provide network functionality and content delivery at the edge of the network based on Mobile Edge Computing borrowed from the IT network while providing the service by deploying a heterogeneous Radio Access Network.

The proposed system architecture and its implementation on OPNET will be presented in this chapter along with the results obtained from such architecture that employed small cells to be used in a HetNet to deliver specific services to the user in dual connectivity mode.

4.2 System Architecture

In this section the architecture of the HetNet system with a centralized controller capable of providing dedicated services for UE's with DC will be explained with the fundamental principles of parameters and measurement involved in the design of the system.

4.2.1 System Model

The architecture is based on the 3GPP LTE-A Evolved Packet System (EPS), with the same main components for radio and core networks with use of small cells as in Scenario #2 of Release 12 [34]. In this scenario the deployment of the macro and SCs is on different carrier frequencies (inter-frequency) where the SC will be distributed as hot spots covering specific areas under the coverage of the macro-cell layer. The SC layer with frequency (F2) will be

located at the centre of the hot spot, where the macro with frequency (F1) will be like an umbrella covering the SC layer.

In this work, we consider the concept of adding computing and storing capacity to the main eNodeB, in which the content could be cached and stored in a server attached to the main eNodeB. In such way, the main eNodeB can be used to provide coverage and basic services and handle most of the control signalling and it will be in full control of all the SCs within its coverage area by the mean of a centralized controller. The SCs that are operating in different frequency band could be used to deliver a dedicated data services such as big video content.

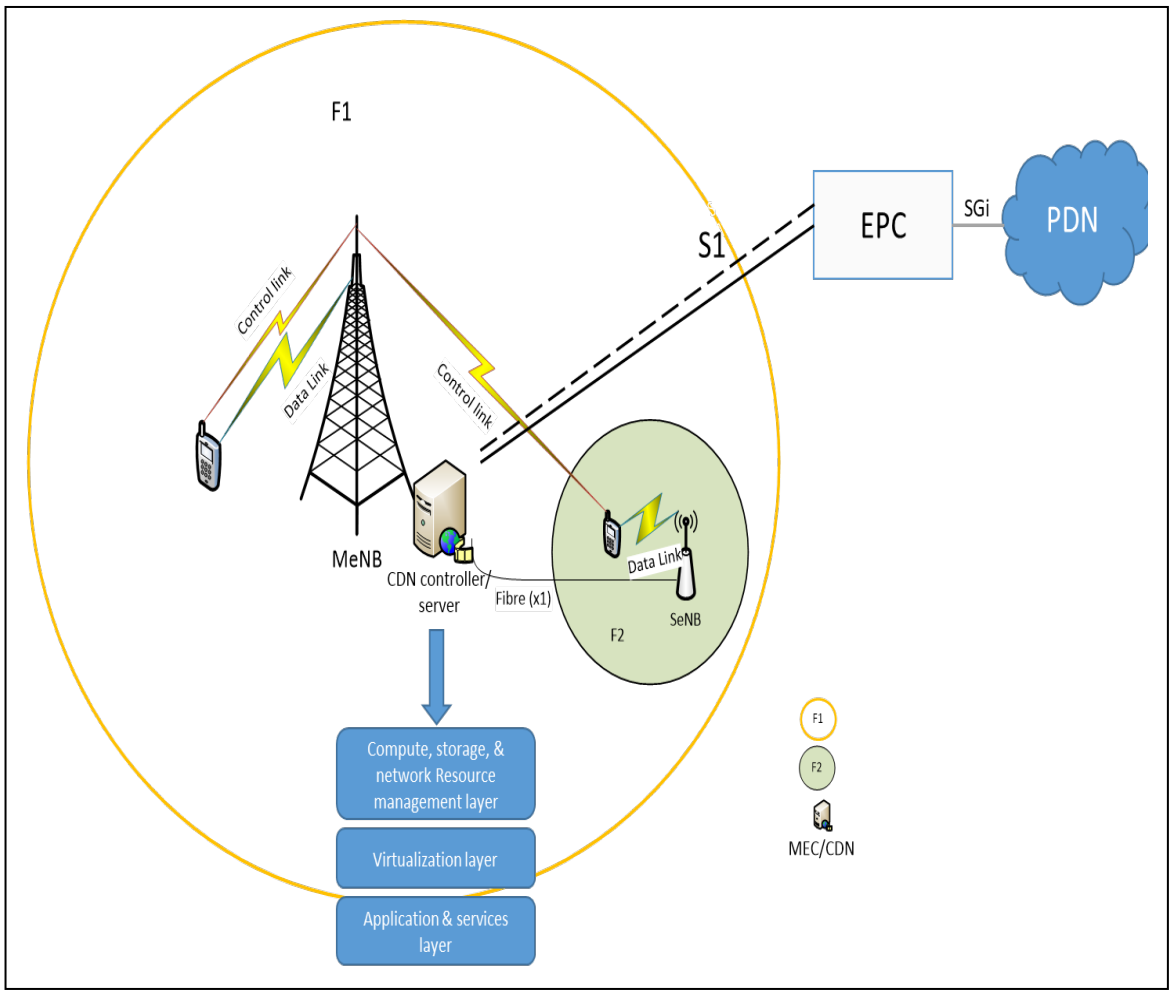


Figure 4.1: Proposed System Architecture.

Centralized processing would result in efficient use of the resources and the dense deployment of SCs under flexible centralization of the RAN will allow for flexible functional split based on the virtualization functionality provided by the computing ability at the edge of the network. The UE of DC maintain a normal connection with MeNodeB and will establish a U-Plane connection with a SC for the downlink of big data applications (i.e. videos) that could be saved in content delivery (CD) server located in or near the MeNodeB. An ETSI MEC server could be used for this purpose as suggested by 3GPP, which can add computing capabilities to the RAN, or could be used as an aggregation point in the IP transport layer [41]. Figure 4.1 shows the proposed system diagram.

In this architecture we will have a HetNet system of RAN components that are controlled by a central server. A UE in RRC – connected mode first obtain access to the MeNodeB and keep C-plane connection with this node, which is the only RAN element that is visible to the core network (EPC), measurement and statistics information related to the UE gathered by the mobile network element based on the 3GPP signalling messages and Performance Measurements (PM) defined by 3GPP can be aggregated and processed by the controller of the MeNodeB, a table of information will be generated that will also contain measurements considering the information coming from the SCs. As soon as a big size of content is requested by a UE, the MeNodeB will direct the UE (i.e. through the system information Block SIB) to connect to the best SC based on the parameters provided by the controller. Figure 4.2 shows the flow chart of this procedure.

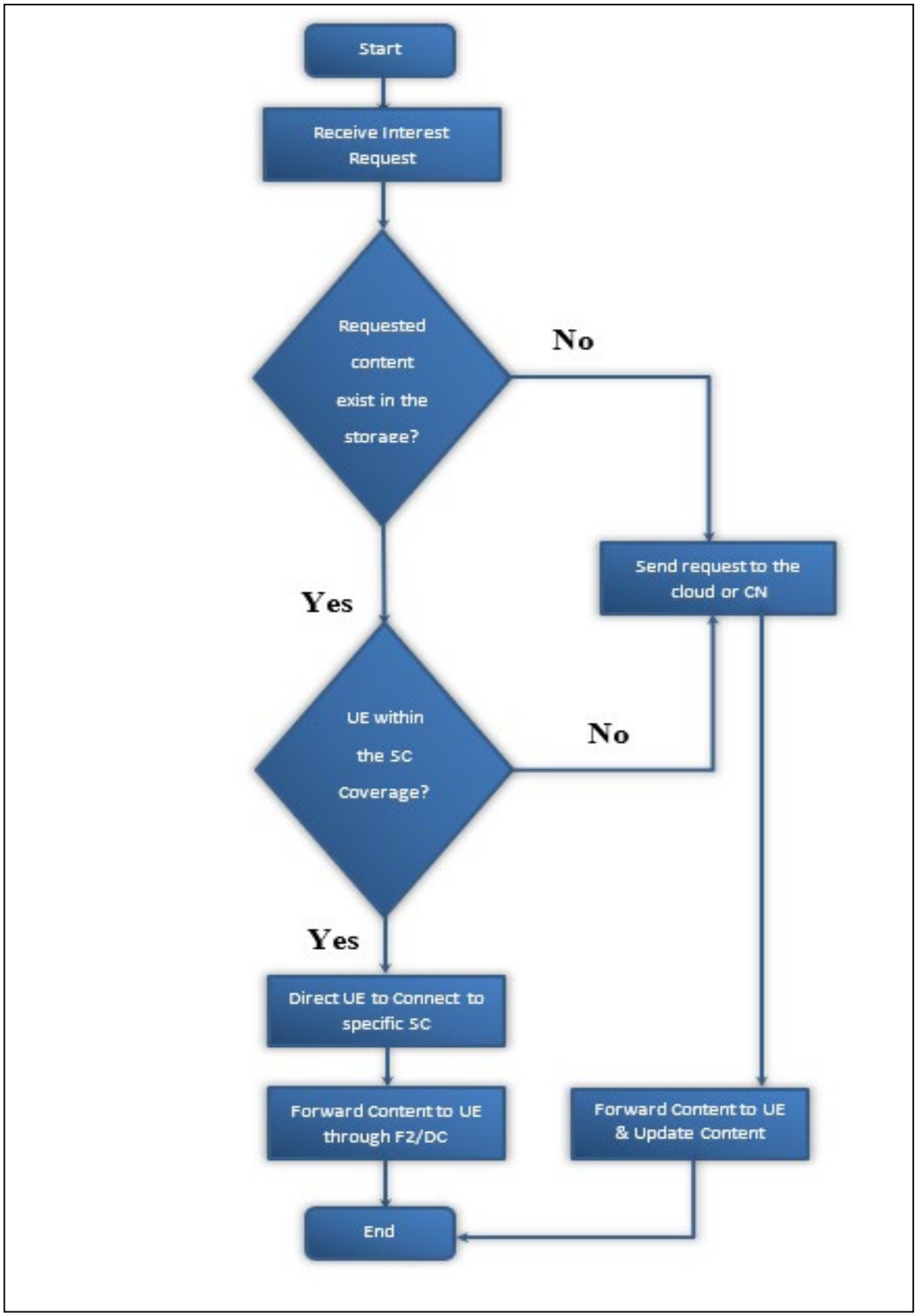


Figure 4.2: Content delivery procedure.

The following steps could explain such a procedure:

Step 1: The measurement report is made by the UE and sent to the controller of the MeNodeB to be added to the measurement table.

Step 2: A content is requested by a UE

Step 3: MeNodeB decided to which node of the SCs the UE is to be connected considering the measurement parameters available in the measurement table.

Step 4: MeNodeB send the decision to the UE (through dedicated RRC signalling, i.e. RRC Connection Reconfiguration)

Step 5: UE connect to the SC decided by the controller of the MeNodeB.

Such platform can provide functional split and could improve the adaptability and the manageability of the network. This architecture will be used throughout this work and in the following sections a brief description of the concepts used in this architecture will be provided.

4.2.2 Heterogenous Network (HetNet)

Simply, a network consists of only one site type (i.e. macrocells) is known as a homogeneous network. Conversely, a network consists of multiple site types (i.e. Macrocells, microcells, picocells, femtocells and either Wi-Fi hotspots) is known as a heterogeneous network. As explained previously in more details in Chapter 2 [32].

A HetNet not only increases the network capacity, but also provides better and even dedicated coverage to serve traffic hotspots or specific type of traffic and enhances the user's experience. These benefits are achieved by offloading data traffic dynamically from MeNodeBs to SeNodeBs using an algorithm based on several parameters such as the characteristics of the traffic, the required QoS and network congestion.

4.2.3 Small Cells

As stated earlier in Chapters 2 and 3, the enhancement of the radio network capacity can be achieved with more spectrum being added, with higher spectral efficiency and with more

cells being deployed. When the first two solutions being exhausted, then the need for small cells will be needed as shown in Figure 4.3 [33].

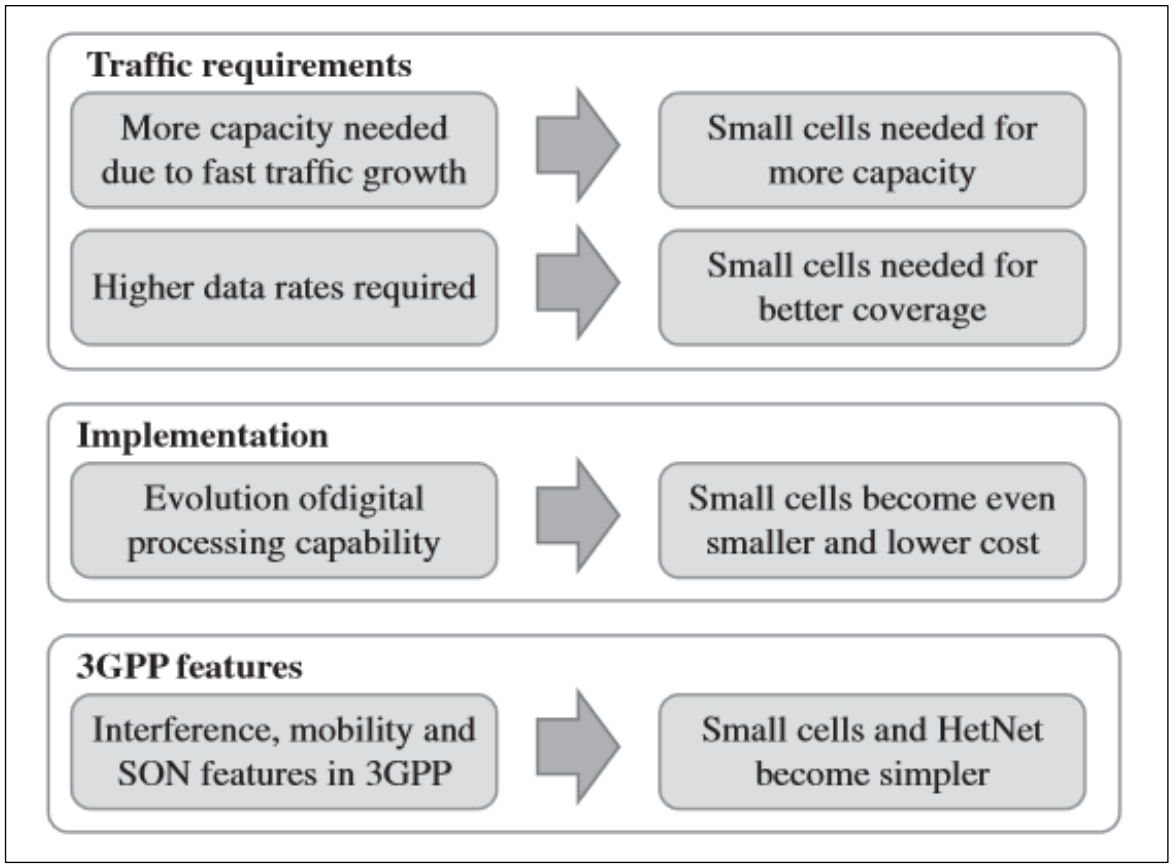


Figure 4.3: Drivers and enablers for small cell deployments [33].

There are many reasons why a HetNet and SCs can be beneficial as shown in Figure 4.3 above [32] [33], and the main reasons are listed below:

- Site acquisition: Site acquisition is often easier for smaller cells.
- Deployment: the small cell implementation has become simpler with small and compact products.
- Coverage: small cells can provide coverage at locations outside the reach of the main network.

In addition to that, SCs are considered a promising solution that can cope with mobile traffic explosion. In 3GPP TR 36.932 [31], any node with transmission power that is lower than the BS classes of macro node can be considered as SC. In such situation Pico and Femto cells that have a transmit power of 0.25W and 0.1W respectively are both considered as SCs.

This is not the case for microcells, as there is no separate power class classification in 3GPP for microcell. Instead, a wide area BS, i.e. macrocell with reduced transmit power could be designed and considered as microcell. Some other scenarios take Wi-Fi hotspots [33] and Remote Radio Heads (RRHs) within a Centralized Radio Access Network (C-RAN) in consideration for SCs deployment options.

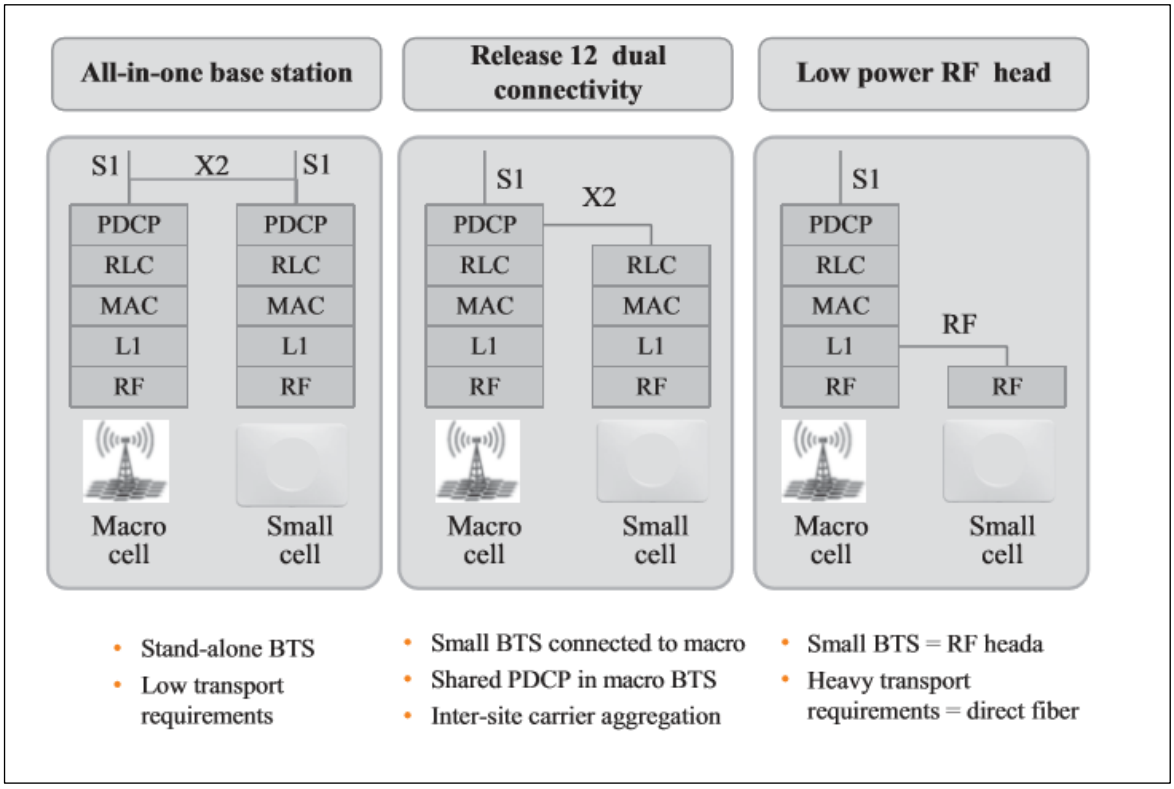


Figure 4.4: Small cell network architecture options [33].

The main three architecture options that are considered for SCs deployment are shown in Figure 4.4 [33]. The left-most side shows the option of a complete all-in-one small BS that is directly connected to the packet core with S1 interface and to the macro cell layer with X2 interface. All the protocol layers 1–3 are included in the small cell. The other option shown in the right-most side shows what is referred to as C-RAN or Cloud RAN since the baseband processing can be done in a centralized location and a low power radio frequency (RF) head without any layer 1–3 functions is connected to the microcell baseband through a tough requirements interface of high bandwidth and low latency. In this option, only the macro cell is connected to the core network and the SC is just an RF that appears to the core network as a new sector of a macro cell.

The solution adopted in 3GPP Release 12 target the deployment of SCs with and without the coverage of macrocells, where the deployment can be indoor and/or outdoor hotspots, with both ideal and non-ideal backhaul options. In addition, when under the macro cell layer, the SC has the possibility to use the same or different frequency band as of the macro cell. The distribution technique is also an important factor, therefore the sparse and dense distribution is also considered within the 3GPP TR. Specification also supports DC where UE has the ability to be connected to both macro and SC simultaneously, as will be discussed in more details in the next section. In this solution, as represented in the middle option, the macro cell layer is connected to the core network by the S1 interface while the X2 interface is used between macro and SCs. All the protocol layers are included in the macro cell, while layers located in the SC includes Layer 1, (MAC) and (RLC) [31].

4.2.4 Dual Connectivity

An important and interesting technology which is an extension to the CA and CoMP principles, as defined and adopted in LTE release 12 specifications of the 3GPP is Dual Connectivity (DC) [31].

CA is the most successful feature of the LTE-Advanced system that increases the channel bandwidth by combining multiple RF carriers coming from the same co-located eNodeB, this technique was introduced in 3GPP release 10 and gives the UE the ability to receive and transmit multiple signals in both directions i.e. downlink and uplink rather than only one signal. The operation of CoMP in the other hands was introduced in release 11 and enables the eNodeBs to operate more jointly. In previous releases, the UE was capable of communicating with only one eNodeB at the time, while with CoMP, a single UE has the possibility to receive data from even more multiple eNodeBs [32] [33].

Based on that, DC was introduced in release 12 specification as part of the study of SCs enhancement. With DC the UE will be able to have connection with both macro and SC simultaneously, in which the SCs are typically deployed as hotspots within macro cell coverage. This will allow the maximum data rate to be aggregated from the macro and SC layers. With such feature, SCs could be used in the most efficient way as the macro cell could be used to maintain the coverage keeping the UE connected all the time to the system even in the absence of the SC layer coverage. 3GPP TR 36.842 provide different architecture

options for the deployment of such HetNet in which the UE in DC is in connection with one MeNodeB and at least one SC. By doing so, three options for U-plane split can be recognized as depicted in Figure 4.5 [34].

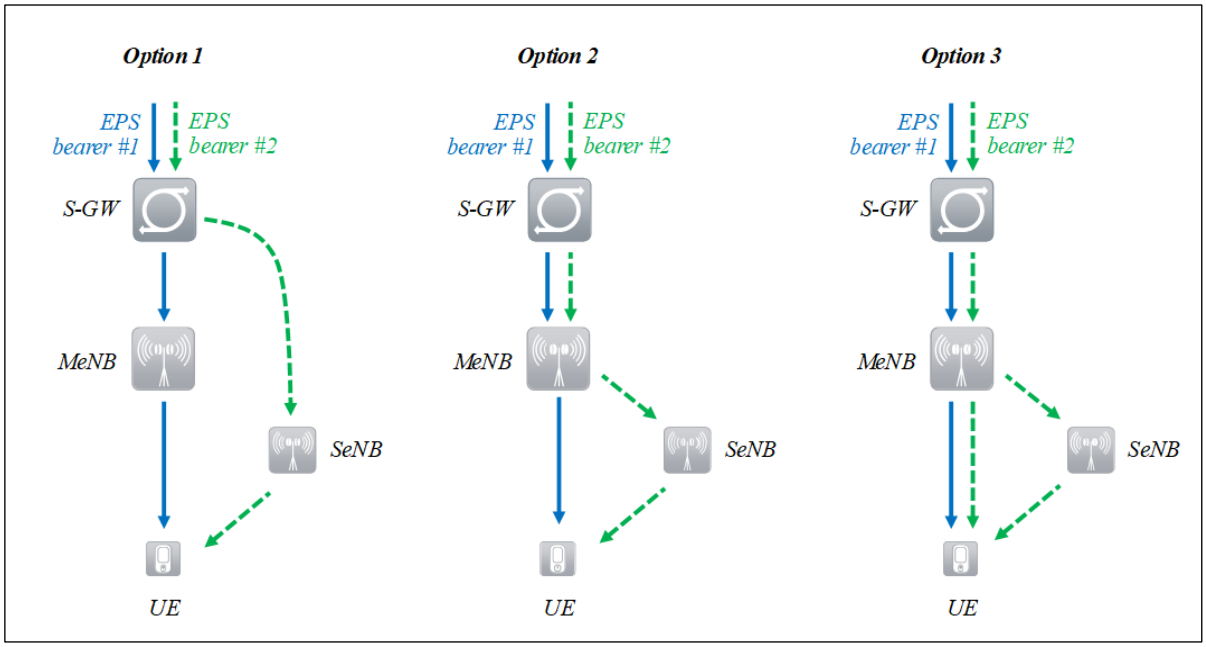


Figure 4.5: U-Plane split options in the downlink [34].

Based on the options for bearer split and U-plane protocol stack, nine alternatives can be obtained to support U-plane data split options of Option 1 and 3 and based on comparison evaluation by 3GPP, as in appendix A. Two alternatives are to be progressed that is alternative 1A and 3C as shown in Figures 4.6 and 4.7 [34].

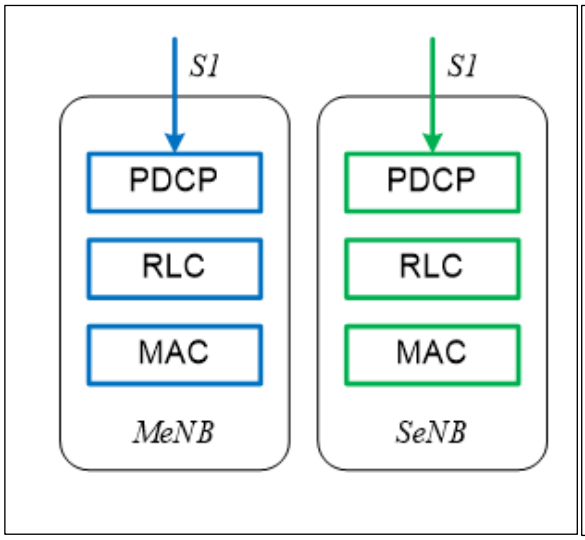


Figure 4.6: Alternative 1A [34].

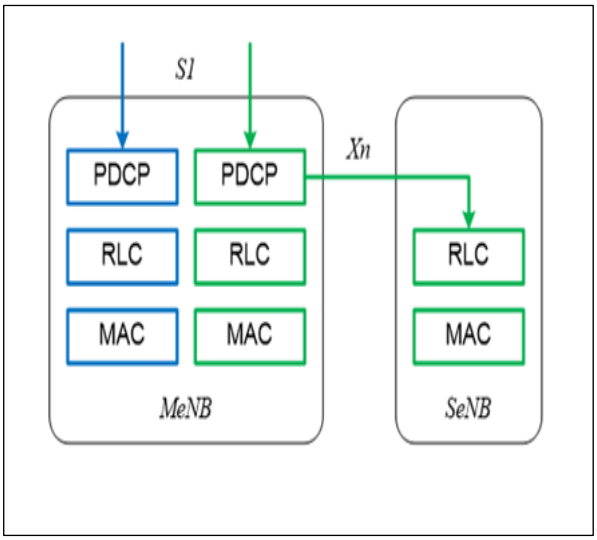


Figure 4.7: Alternative 3C [34].

Some of the expected benefits from such enhancement are:

- Continuous connectivity.
- Increased UE throughput for cell edge UEs in particular.
- Reduce the overhead occurred from signalling towards the core.
- The possibility to use the SCs layer for dedicated services such as content delivery.

Although such architecture brings many benefits, as it may enhance the system performance to higher levels by increasing the capacity and reducing the latency and system messaging overhead; yet it brings many challenges due to increasing the system complexity. Challenges may include: Efficient distribution of the SCs and the discovery scenario of such SCs by the UE that may exhaust the battery power; the mis measurement and exchange of the signals between the Macro and SC layers may lead to a ping pong situation and HO problems during the mobility.

4.2.5 In-Network Cache

The increasing consumption of web-based streaming video and Apps in the networks and the internet shows a clear shift to content consumption which are expected to seize the majority of 4G/5G bandwidth. Therefore, storing a copy of popular contents in cache servers placed at the edge of the network (i.e. Local cache) is a solution to solve the issues that may arise from long transportation distance between the server hosting the application and the RAN.

local cache is presented in 3GPP release 14 as part of the solution of Context Aware Service Delivery in RAN for LTE [41] as shown in Figure 4.8.

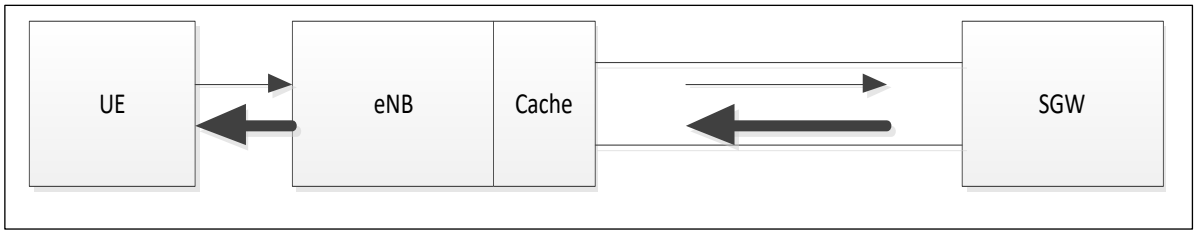


Figure 4.8: cache server collocated in the eNodeB [41].

4.3 Network Model Implementation

The proposed DC architecture that integrate the SCs to the LTE MeNodeB at the PDCP layer, will be implemented using Riverbed 18.5 Modeller based on the 3GPP technical requirement for SCs enhancement [31].

The LTE model features supported by this modeller are based on 3GPP release 8 & 9, that don't support DC. Therefore, a modification to the LTE node models is required. This modification can be done using the Device Creator to create custom model or modify the existing one [98]. Furthermore, virtualization functionality is not supported but the powerful program capabilities of creating different applications can be used instead to simulate the network to assign tasks and choose the statistics of each node. A measurement entity is created for each UE, which records values of RSRP and RSRQ, thus the UE continuously measures RSRP and RSRQ for all nodes within its range.

Finally, the system performance is evaluated using multiple scenarios using riverbed simulator, with the same LTE simulation parameters, that are set according to 3GPP TR 36.842 [34] and TR 36.872 [35].

4.3.1 LTE Node Models Implementation

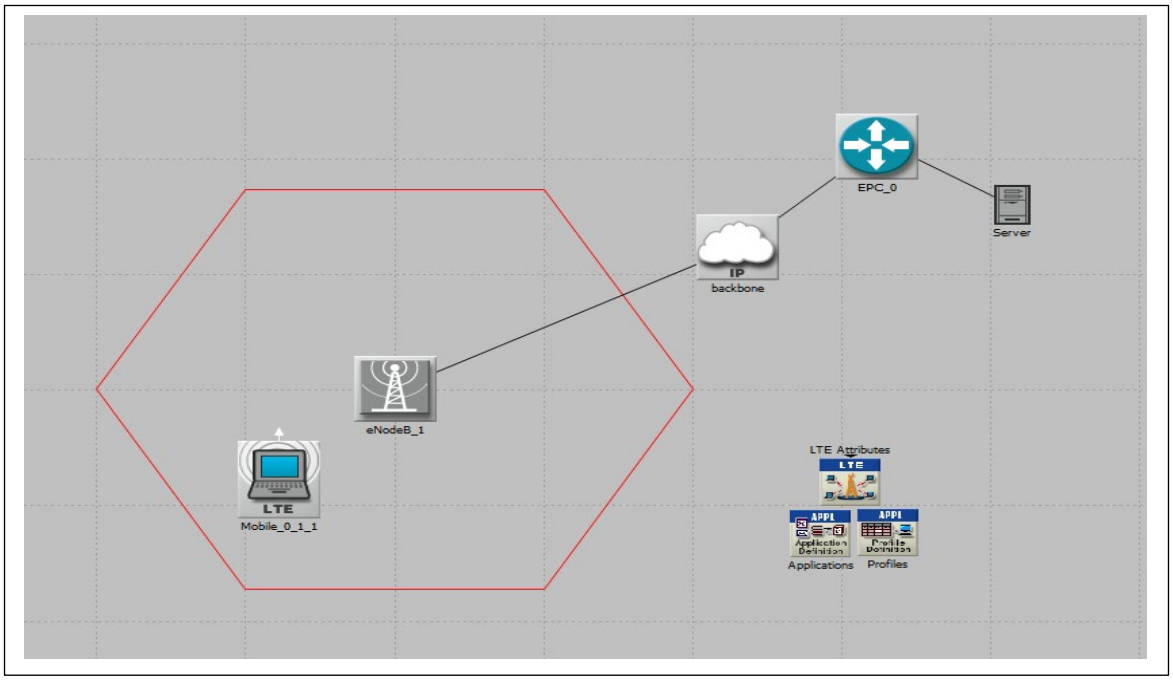


Figure 4.9: Basic LTE architecture.

The basic LTE architecture in OPNET contains three main nodes, specifically, UE, eNodeB and a single node performing all the EPC-related functionalities that is the Evolve Packet Core (EPC) node as shown in Figure 4.9.

The packet data flow through the system is by means of the EPS bearer and its radio and S1 bearers, where the S1 bearer is running between the EPC and eNodeB and carried out by the GTP tunnel, while the radio bearer is running through the air interface between the eNodeB and the UE. As shown in figure 4.10 which also depicts the U-plane protocol stack of LTE.

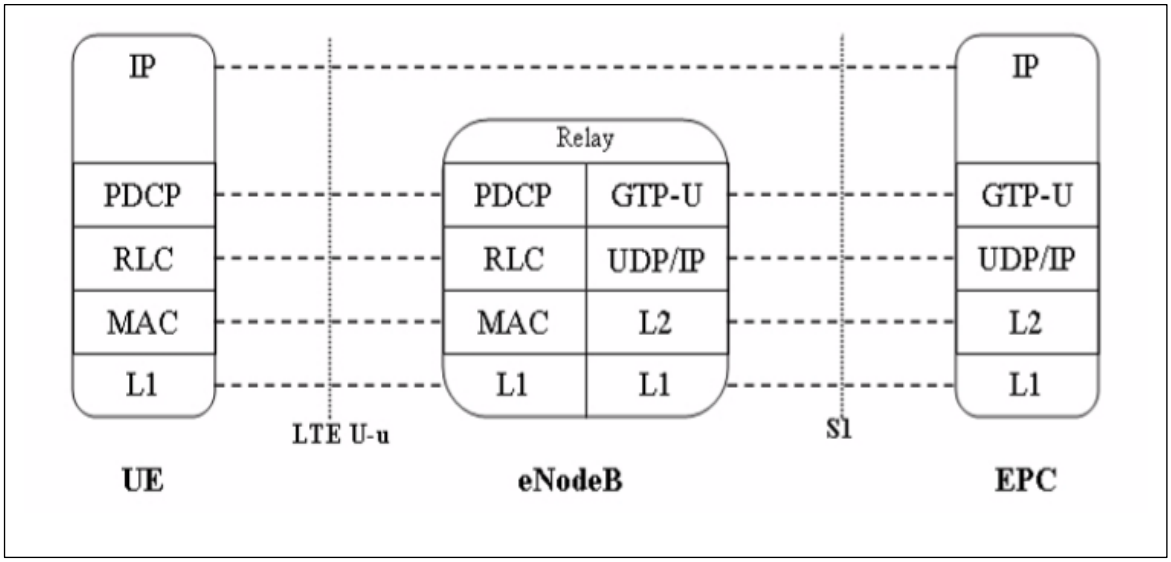


Figure 4.10: GTP Tunnelling Between eNodeB and EPC Nodes [98].

In the OPNET node domain, the node models include the full protocol stack from the physical layer up to the application layer represented by modules for the AS and NAS protocols while the layers representing the U-Plane protocol stack are embedded as process modules inside them as shown in Figure 4.11.

As the DC allows a UE to have two simultaneous connections to a main eNB (MeNodeB) of macro cells and a secondary eNB (SeNodeB) of SCs while exchange of the information between the MeNodeB and UE may take place on different layers, such as MAC, PDCP and RRC layers. Therefore, a modification to the node model is required so that a UE will have the protocol stack defined by 3GPP for DC as depicted in Figure 4.7.

A UE in RRC – connected mode first obtain access to the MeNodeB and keep C-plane connection with this node, which is the only RAN element that is visible to the core Network (EPC), measurement and statistics information related to the UE gathered by the mobile

network element based on the 3GPP signalling messages and Performance Measurements (PM) defined by 3GPP can be aggregated and processed by the controller module of the MeNodeB, a table of information will be generated that will also contain measurements considering the information coming from the SCs.

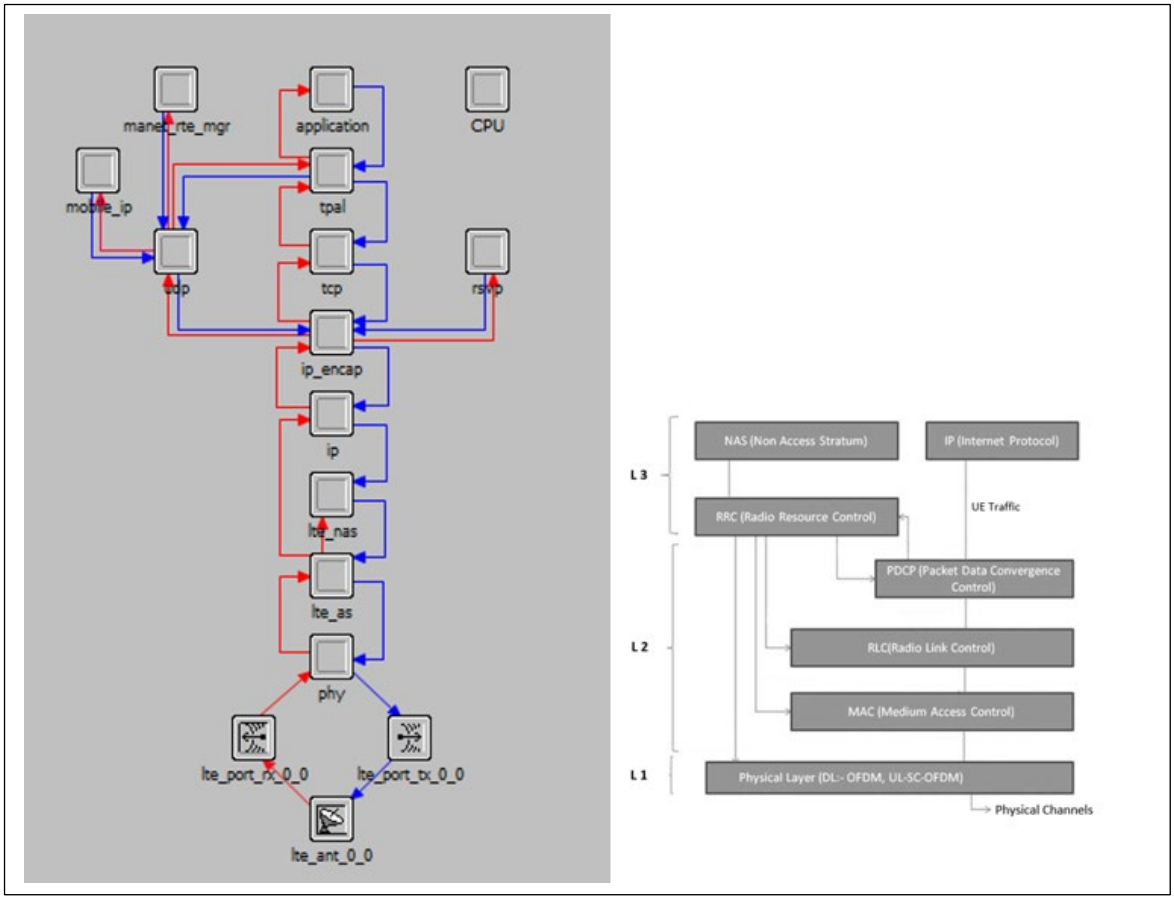


Figure 4.11: LTE UE node model and the equivalent protocol stack.

The node model for the modified UE with DC is shown in Figure 4.12. This modified model has the same layers of the original node model, except for the LTE-As DC which has limited functionality compared to the original one, as it has only the PDCP and RLC layers.

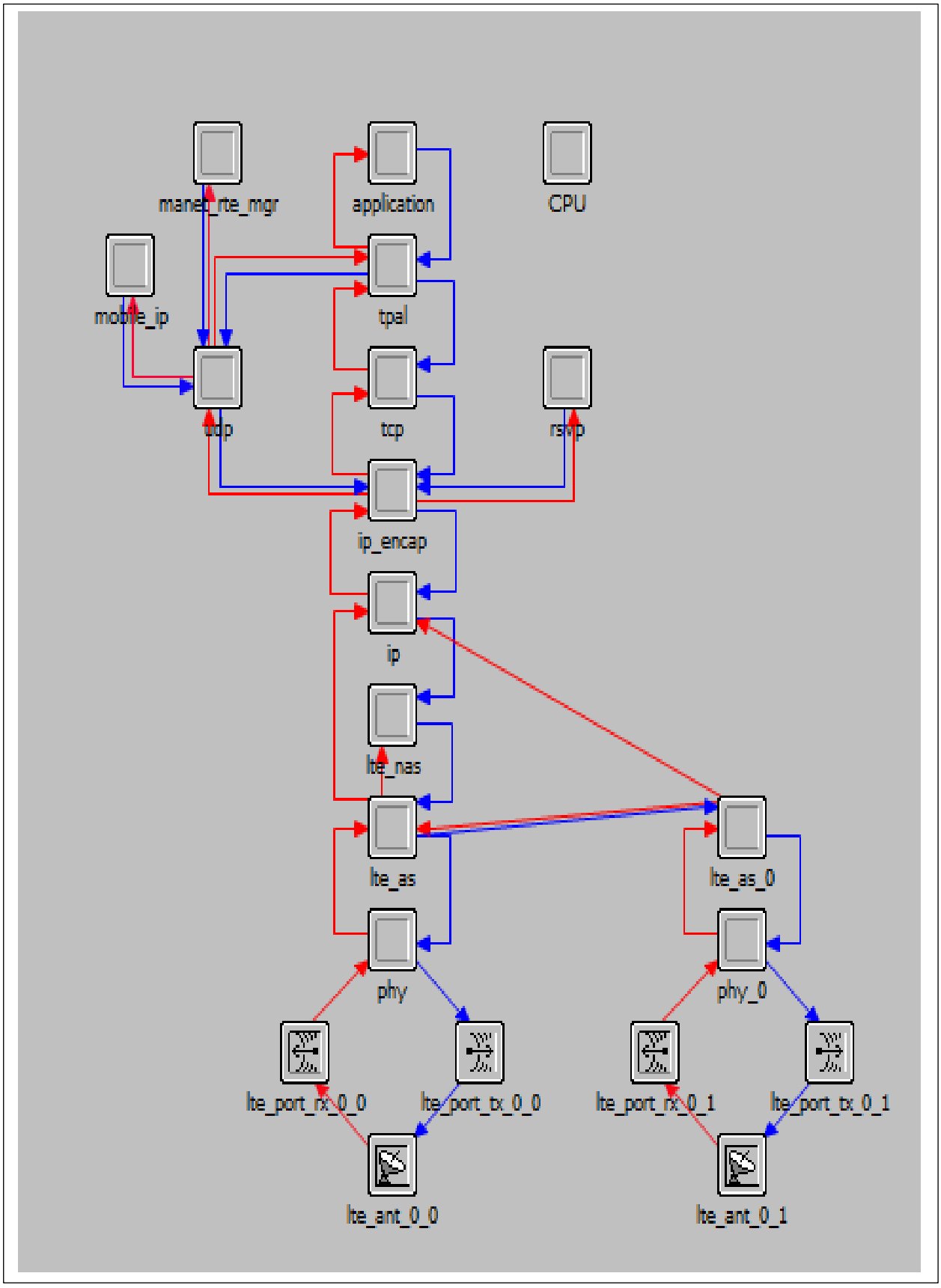


Figure 4.12: Node model for the modified UE with DC

The process of AS protocol will be done only by the original protocol through the attached procedure as explained below and shown in Figure 4.13.

1- The UE is first turned on and attached to the network.

   - A UE context is created.

2- The UE said to be in the EMM-deregistered state.

   - The UE cannot be paged, and the MME has no knowledge of the UE location.
   - The UE cannot have any user plane bearer while in this state.

3- The UE moves to the EMM-registered state after completing the attached procedure.

   - The UE is registered with the MME while in this state, and a default bearer is established.

4- When in EMM-Idle, the UE can:

   - Responded for paging messages.
   - Perform service request procedure.

5- UE and MME enter the ECM-Connected state after NAS signalling connection has been established.

   - UE View: RRC-Connection established between UE and eNodeB.
   - MME View: S1 Connection established between the eNodeB and MME.

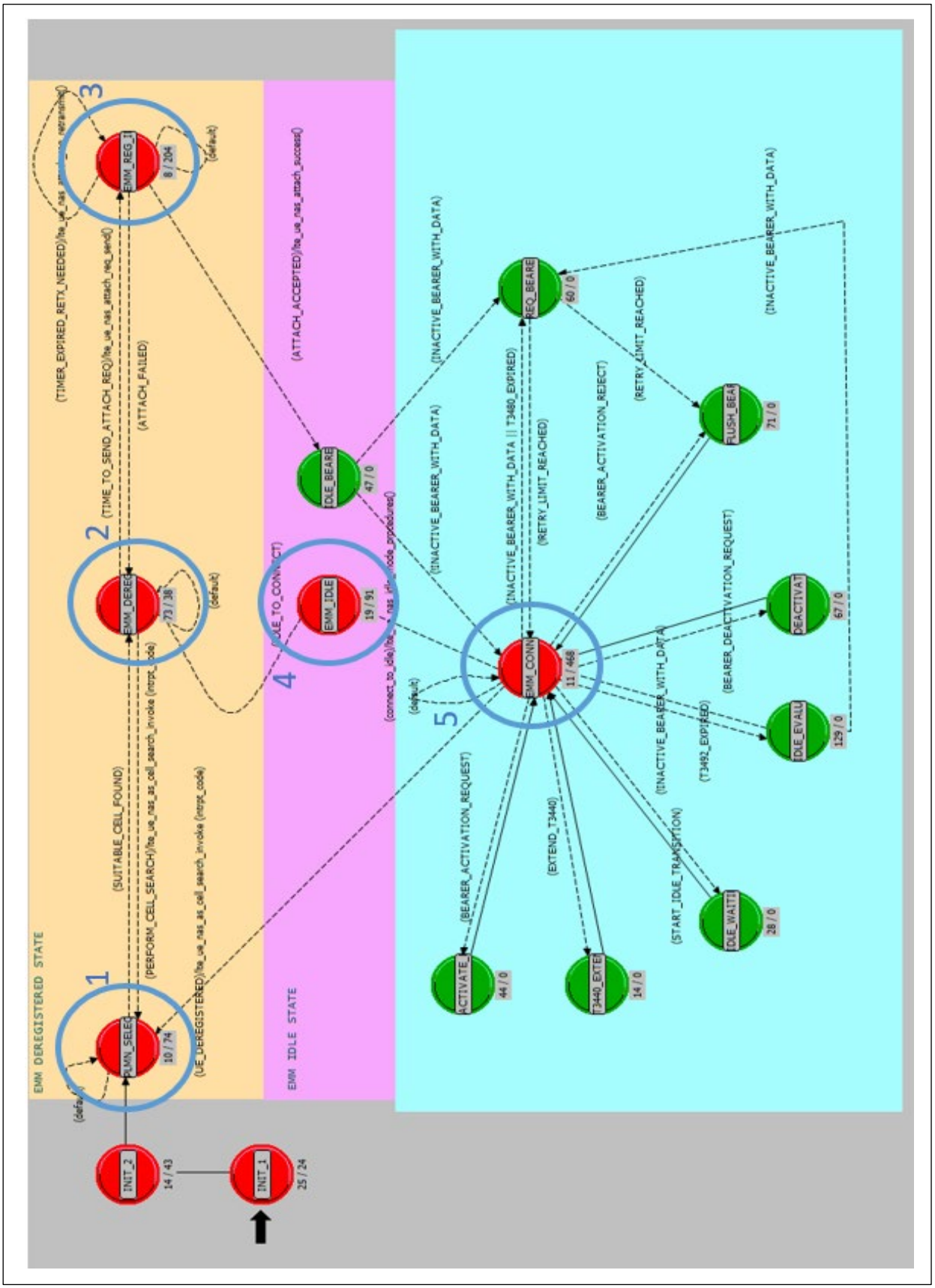


Figure 4.13: LTE UE NAS process model.

After this point the UE is said to be in the RRC connection mode and is successfully connected to the eNodeB and can start reading the system information of the cell and performs the PDCP status report procedure with the eNodeB. LTE modes are RRC-Connected and RRC-Idle mode, and in the Idle mode the UE is just paged for the downlink data while in the connected mode. The UE is in full operation for transmission and reception. The NAS, S1 and other RRC connections are active in the connected mode, while in the idle mode all the mentioned connections are removed.

Regarding the MeNodeB, this node will be attached to router via point to-point protocol (PPP) link to add routing capabilities and will be acting as a gateway unit linked to the EPC. In this case the MeNodeB GW serves as a concentrator for the C-Plane, specifically the S1-MME interface. The S1-U interface from the SCs may be terminated at the MeNodeB GW. The MeNodeB GW appears to the MME as a normal eNodeB while appears to the SCs as an MME. This is similar functionality to the HeNodeB GW [99] with some modification made for the support of DC.

The adjusted SC node model is shown in Figure 4.14. while, figure 4.15 shows the node model for MeNodeB. The designated eNodeB structure includes Ethernet and PPP ports in the physical layer to provide capability of communication to the servers by Ethernet and optical fibre links.

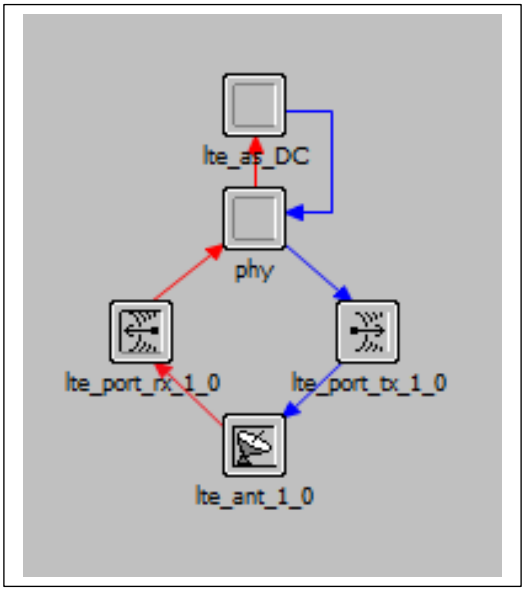


Figure 4.14: Lte SC Node Model.

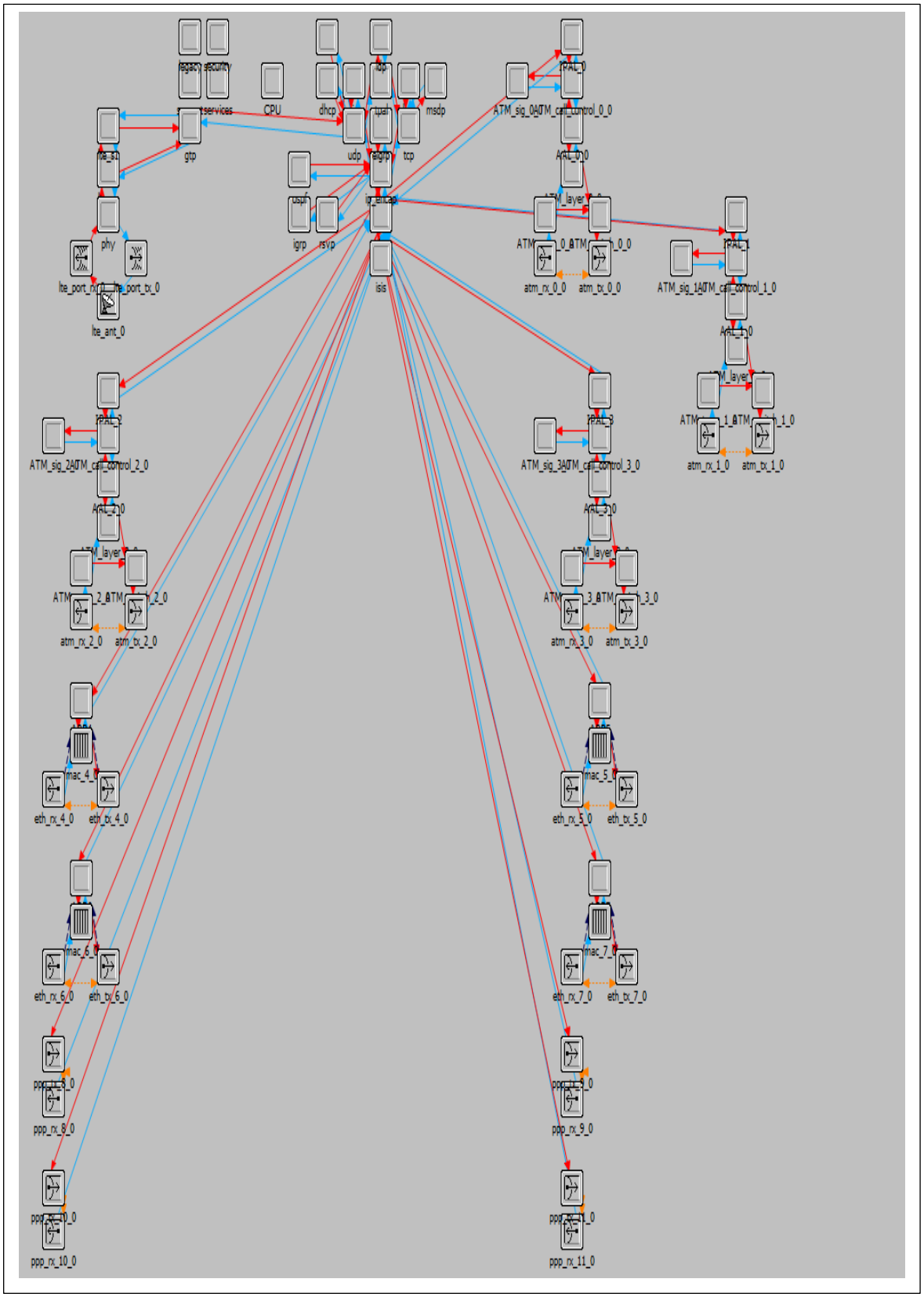


Figure 4.15: LTE eNodeB Node Model.

4.4 Overall System Implementation

After choosing and implementing the node models, the overall system will consist of the MeNodeB, SCs (SeNodeBs), the EPC, and the servers to deliver the required services. As shown in Figure 4.16.

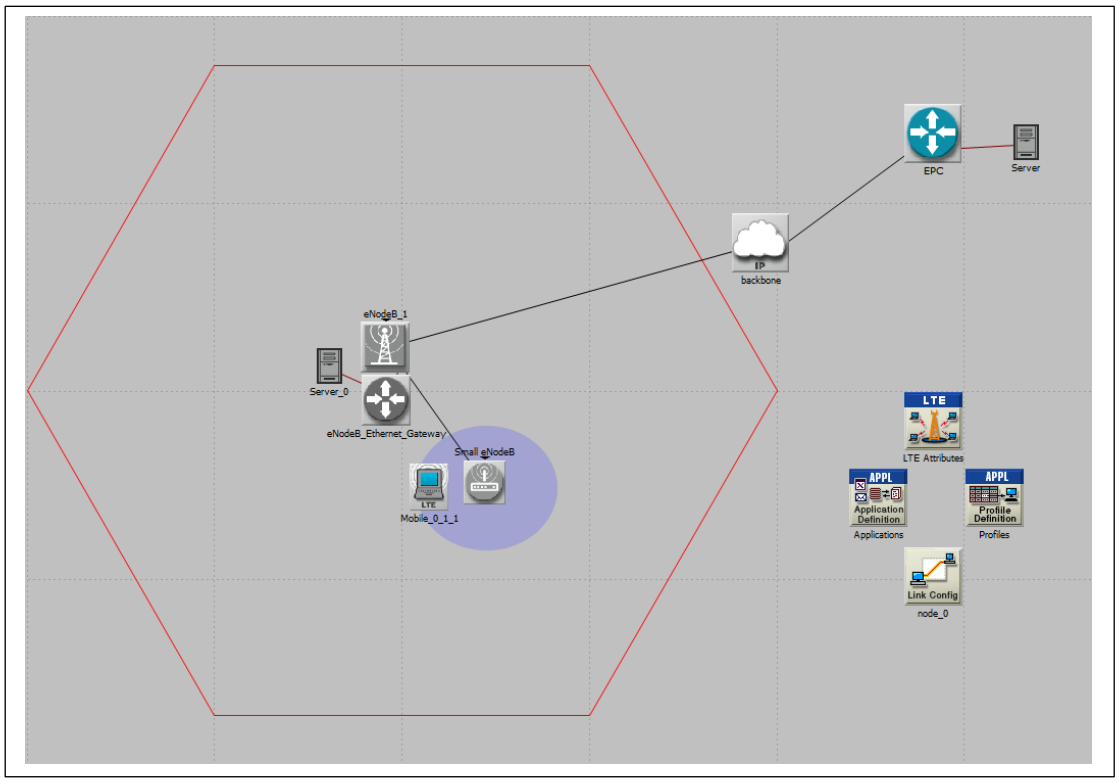


Figure 4.16: HetNet system architecture.

Based on the specification requirements of 3GPP in [34], the deployment scenario consider SC (SeNodeB) layer to be deployed under the main macro (MeNodeB) cell layer, where the UE could be under coverage of both the macro cell (MeNodeB) and the SC simultaneously or only under the coverage of the macro cell. The MeNodeB will be used to provide coverage and handle most of the control signalling and will have control of the SCs. The SCs that are operating in different frequency band will be used to provide capacity and deliver the required data services. The MeNodeB is connected to all SCs within its coverage area and a list of the attached SCs will be generated and modified every time new SC to be added. The controlling capability that is added to the MeNodeB is suggested to be implemented as an equivalent to imitate the MEC/CDN functionality like caching, and other specific operations such as synchronization, baseband processing, interference coordination, carrier

aggregation, inter-eNodeB CoMP and DC which is yet to be standardized and not yet available in the OPNET simulator.

In this system of DC, shown in figure 4.16, the UE attain access to the MeNodeB using the normal attached procedure and maintain RRC connection with the MeNodeB as long as it is under its coverage. After that, when a service is required by the UE, the MeNodeB can direct the UE to the best SC (i.e. through SIB) based on the specific measurements such as location of the UE and distance from the SC. The Server, Application and Profile configuration units are used to define the application specifications on the selected cells and users, and LTE Attributes and link configuration units are responsible for general LTE and SC configurations in the network respectively.

Parameter	**Value**
Type of Service	HTTP with Video
Video Type	On demand
File Size	200 Mbytes
File inter-arrival distribution	Exponential
Total MeNodeB Tx Power	46 dBm
SC Tx Power	30 dBm
Noise figure	9 dB in UE, 5 dB in eNodeBs
UE Tx Power	23 dBm
No of MeNodeB	1
MeNodeB Carrier Frequency (F1)	2 GHz
No of SC	1
SC Carrier Frequency (F2)	3.5 GHz
No of users	25
LTE Bandwidth/Duplexing	20 MHz/FDD
Sub-carrier spacing	15 kHz

Table 4.1: Simulation parameters.

The system performance is evaluated using multiple scenarios using riverbed simulator to investigate the optimal solution, with the same LTE simulation parameters, that are set according to 3GPP TR 36.842 [34] and TR 36.872 [35] summarised in Table 4.1.

4.5 Performance Evaluation

4.5.1 Simulation Scenarios

In this section the proposed architecture will be analysed based on different scenarios of the same parameters as in table I. And with three main cases are considered as:

1- Content served from outer sever (i.e. No Content is cached).

2- Content served from the MeNodeB server (i.e. Cache enabled).

3- The UE is connected to the MeNodeB in the UL and to the SC-eNodeB in the DL, with Content cached enabled.

The network will be examined for the e2e delay and throughput as the key performance factors for the network for all scenarios. Adaptive modulation and coding were enabled in order to enable the UE to communicate with the eNodeBs in variable channel conditions. The interference and multi-path are modelled. IP traffic is established between the UEs and HTTP server connected to the LTE network through internet backbone. The simulation time has the duration of 60 minutes, there is a warmup time of 90 seconds approximately, before the start of the simulation and results collection.

The first scenario is set to start with basic LTE network with 1 MeNodeB, and then we will examine the performance when adding the SC layer. After that the system will be tested with variable number of SCs and UEs that are distributed as hotspots under the MeNodeB coverage, then for the succeeding scenarios, the number of SCs and UEs will be increased to be as 1, 2, 3 for the SCs, and 5, 10, 20, and 25 for the users as shown in Table 4.2.

CSs	UEs
1	5
1	10
2	10
2	20
3	20
1	25

Table 4.2: Simulation Scenarios used to test the system with variable numbers of SCs and UEs.

4.5.2 Gain Analysis with Dual Connectivity

In addition to the expected enhancement in terms of delay to be achieved by SCs, the gain increasing mechanism when introducing DC to the network and provisioning the layers to Macro and Small, will be demonstrated. Theoretically, and according to Shannon-Hartely equation, the maximum rate at which information can be transmitted over a communications channel of a specified bandwidth in the presence of noise can be expressed as below [100]:

$$Ci = Bi \log2 (1+ SNRi) \ldots\ldots. (4.1)$$

where:

C is the capacity (hence throughput),

B is the bandwidth, and

SNR is the signal to noise ratio, all related to cell (i) which is assumed to have the best (maximum) throughput /user in both Macro and Small.

When there is no DC, the user is assumed to be served by a single cell, when introducing DC, the users are assumed to receive data from the small cells and the controls from the macro cell. The candidate cells characterized by the best estimated throughput in both the macro and small cell layers are selected as the serving cells. The Shannon capacity for the user with DC is expressed as:

$$C_{DC} = C_{im} + C_{is} \ldots\ldots\ldots\ldots (4.2)$$

The user throughput gain with DC will be

$$\text{Gain} = \frac{\text{CDC} - \text{CnoDC}}{\text{CnoDC}} \text{ X } 100\%$$

$$= \frac{\text{Bq log2 (1+ SNRq)}}{\text{Bp log2 (1+ SNRp)}} \text{ X } 100\% \ldots\ldots\ldots(4.3)$$

where q = arg min (Cq), and p= arg max (Cp)

If the same bandwidth is deployed in both two layers (i.e. Bq/Bq =1) it will explicitly show that DC is most beneficial for users exposed to similar channel conditions in both layers Plus 100 % DC gain when SNR difference is 0. Noting the DC gain cannot be larger than 100 %,

due to the reason that for cases without DC the selected serving cell is assumed to have the highest estimated throughput from the candidate cells in the two layers.

4.5.3 Simulation Results

The simulation of the network will be run twice for the three cases explained in subsection 4.5.1 In the first run, the network is configured with low load traffic to decrease the probability of packet loss due to either the buffer overflow or repeated retransmissions due to the traffic congestion. And in the second run the network will be configured with full load (i.e. all the UEs are requesting data from the server at the same time imposing full utilization of the offered capacity).

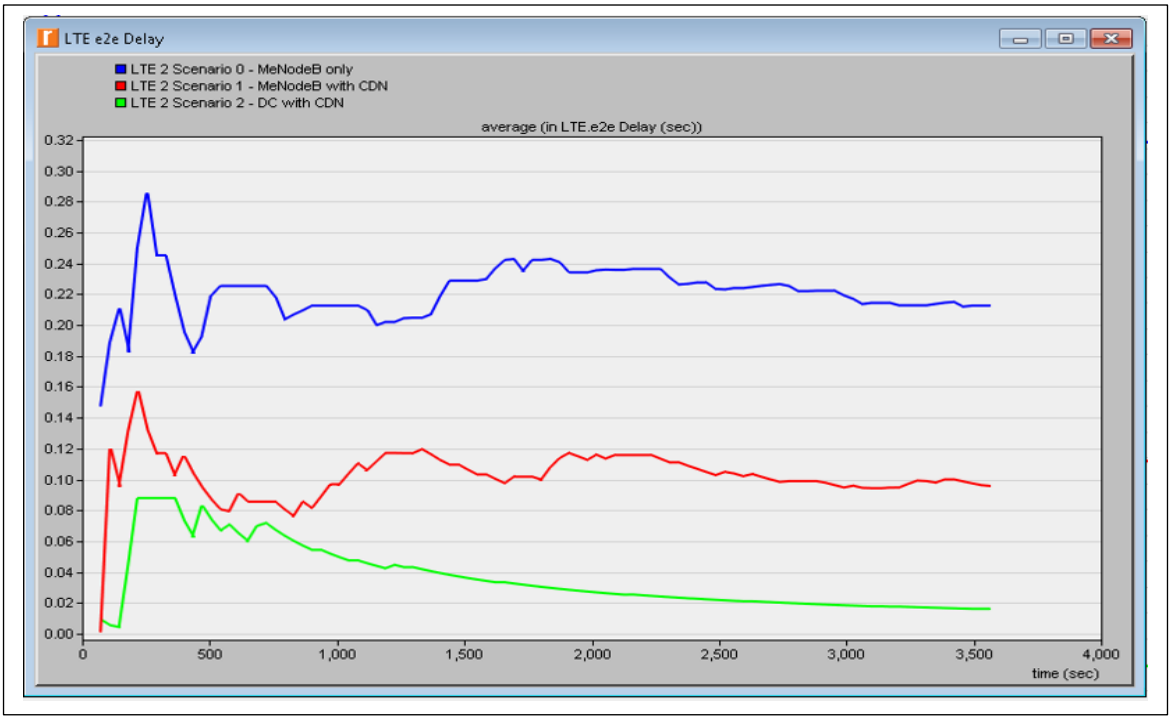


Figure 4.17: e2e network delay.

Figure 4.17 shows the response of the network in terms of e2e delay for the three scenarios. It can be observed that the delay is very high when the UE is connected to the MeNodeB provided that no data is cached in the network and it is acceptable when the contents are cached in the MeNodeB, while it has dropped significantly when the UEs are connected to the SC in the proposed scheme.

The explanation of drop in the delay is the fact that the distance to the MeNodeB is quite larger than the distance to the SC, provided that the network is running in different

frequencies for the UL and DL connections which helps to reduce the interference in the network in order to decrease the losses as proposed.

The next result of the simulation shows the throughput of the network delivered in bits/seconds. It can be observed that the throughput is increasing when the content server is getting closer to the UE, achieving its best when the UEs downlink is connected to the SC and using the proposed scheme as illustrated in Figure 4.18.

Figure 4.18: e2e network throughput.

Logically, the throughput (bits/sec) increases in 2 cases:

i. when the data traffic increases,

ii. when the elapsed time decreases.

Hence in this model, when the content is placed in the MeNodeB (i.e. closer to the UE) the performance of the network expected to improve. However, the results show that the system starts to perform even better in the third scenario when the DL is connected to the SC after

20% of the simulation time, this is due to SC initialization and time spent fetching the content from the main sever.

The second run of the simulation will examine the case when the network is configured and routes full load in its data plane. Figures 4.19 and 4.20 show the response of the network in terms of e2e delay and throughput for the three scenarios.

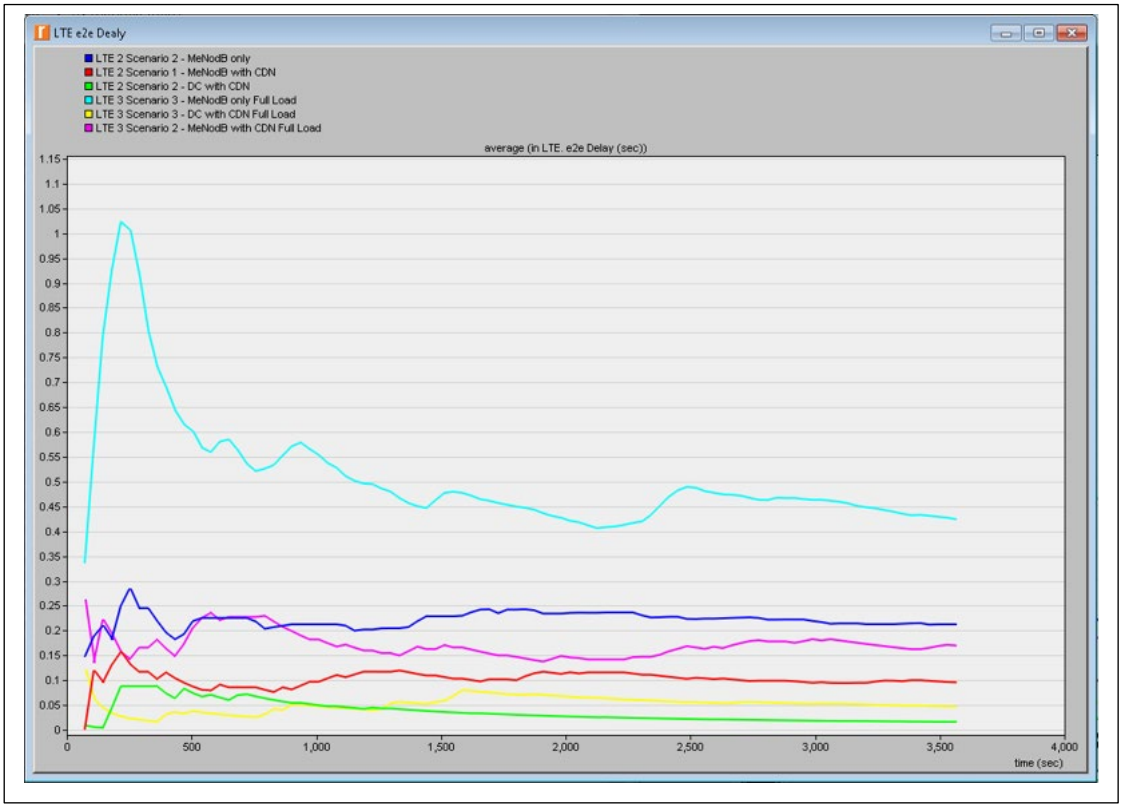


Figure 4.19: Full load network delay

It can be observed that the delay is very high when the UE is connected to the MeNodeB with no data server available in the RAN while it is acceptable when the content server is attached to the MeNodeB and has dropped remarkably when the UEs are connected to the SC and using the proposed scheme.

Figure 4.20: Full load network throughput.

The same thing can be said regarding the throughput delivered in bits/seconds. when the content server is getting closer to the UE, it can be observed that throughput is increasing achieving its best when the UE downlink is connected to the SC. Noting that when the content server is placed in the MeNodeB the throughput drops dramatically when the network is running full load, this is due to high traffic that is being processed and requests from the UEs to be fulfilled by one content server.

Finally, the system was run with multiple scenarios of different numbers of SCs and UEs that are distributed as hotspots under the MeNodeB coverage, the number of SCs and UEs will be increased to be as 1, 2, 3 for the SCs, and 5, 10, and 20 for the users as in Table 4.2. The users are randomly distributed.

Figure 4.21 shows the response of the network in terms of e2e delay for the multiple scenarios. The observation is that the delay is increasing with the increase numbers of UE, even when there are more than one SC to serve the same number of UEs equally or when the number of UEs under the coverage of the SCs are close. This is well noticed in the behaviour of the curves that represents the different cases, the Red curve is the case for 5

UEs per SC; the Blue curve is the case for 10 UEs and 2 SCs which is almost the same case as the Red curve; 5 UEs/SC; and it can be noticed that the two curves are close in values. The same thing can be said in the case of Green and Light Blue curves as both of them represent the results when there is 10 UEs/SC; the Green curve is for 10 UEs and 1 SC while the Light Blue is the case for 20 UEs and 2 SCs; the delay is higher than the previous case due to the increase number of users. This is the expected response as the burden increase on the MeNodeB assuming that same content is routed in the network in every scenario. The Yellow curve can be seen as the mean or average of the results as there are 20 UEs served by 3 SCs, which means approximately 7 UEs per SC. Whilst the incremental of SC number in the entire network significantly drops the delay as the time elapsed to fetch data from the cloud is narrowed or sub-zeroed. The difference between the distance to the SC and to the MeNodeB is major factor to the rise and drop in the delay. In other words, the drop in the delay is due to the distance to SC is being quite smaller than to the MeNodeB, provided that the access time is increasing when increasing the number of UEs to be attached to the network and requesting contents to be delivered from the server. As can be noticed in the result at the time between 100 and 500 sec.

Figure 4.21: e2e delay for multiple Scenarios.

4.6 Summary

The convergence between the IT and Telecommunication Technologies bring new capabilities that will enable the deployment of new services at the edge of the network. In this chapter, design and implementation of heterogeneous network based on the LTE system that supports dual connectivity and data delivery at the RAN was introduced. In this system, the processing of the data and the control of the SC is within a controller at the edge of the network, in which SCs are used to provide services to the users. The proposed system and resource management are implemented using the OPNET modeller and evaluated through multiple scenarios with and without full load. When comparing the data traffic received in the DL direction under many network circumstances, it can be noted that the proposed method has successfully enhanced the performance of the network and suppressed any performance degrading that may occur due to a sudden high network load or conditions. The outcomes demonstrate that the scheme proposed in this work can decrease the latency while the total network capacity is highly improved for the end users due to the throughput enhancements that occur when SCs are inserted into the system, especially when the resource management and content delivery servers are also applied in the solution.

Finally, as the conclusion to this chapter, if the heterogeneous LTE-A network architecture is accompanied with CDN and MEC network management, then the improved bandwidth utilization obtains higher network capacity for the growing number of mobile users within the network. In addition to this, the planning of such a cooperative system with higher number of SCs has undeniable consequences in overall users QoE and also the implementation of such low-cost/low-power cells will have huge impact to the expenditure for mobile network operators in their implementation and operational costs.

Chapter 5

Small Cells Handover performance in Centralized Heterogenous Network

5.1 Introduction

One of the most attractive attributes of the mobile system is mobility; the ability of the mobile user to benefit from the services without interruption while in the move through the network represents one of the charming characteristics that keeps the user attracted to the mobile system. At the same time this attribute represents one of the most challenging areas in the design of the system and for the mobile network operators. Yet, this feature is getting more complicated in the heterogeneous network and thus requires a variety of network planning considerations in such system that contains cells of different sizes.

This chapter presents a mobility-enhanced scheme for improving the handover performance of a mobile UE with dual connectivity in Heterogeneous network taking into consideration decision factors, such as signal power and resource distribution.

5.2 Overview and Related Work

Keeping high level of satisfaction for the mobile users is a main target of the mobile network operators, providing the mobile users with seamless connectivity and all-time services, especially as the mobile devices are passing through the network is an essential factor to

keep users satisfaction. Mobile devices are no longer a luxury and become an essential part of people's everyday life and business. People are using their mobile devices while on the move for different purpose and needs, they are surfing the internet, checking their e-mails, connecting with each other through social media, and streaming audio and video feeds. Many of these activities are done while the user is moving, and as the system is heading towards a HetNet system, therefore, a reliable HO technique for mobile UEs in HetNet is required.

In mobile networks, HO is the process that took place when the mobile node moves through networks from one cell to another and changing its point of attachment in a network by disconnecting from one point and reconnects to another point in the same or different network. The most common driver for such process is the deterioration in signal quality, or due to congestion when the cell has become overloaded [99]. Therefore, HO is generally classified into two categories based on the occurred procedure; the inter-cell and intra-cell HO, taking the change of cell in consideration; and soft and hard HO, taking the HO mechanism in consideration.

In inter-cell HO, the mobile moves from one cell to another but remains within the same cell site (i.e. the same site or point of attachment), such as HO of UE from one BTS to another in GSM network, where both BTSs are controlled by the same BSC. While in intra-cell HO, both the source and target nodes belong to the same cell (e.g. the same BTS in GSM network), and therefore the cell is not changed during the HO process, such as moving between the different sectors of one BTS in GSM network.

In regards of the hard and soft HO; with hard HO, the link is first terminated with the prior cell's node before or as the user is transferred to the new cell's node (break-before-make). Thus, the mobile is linked to no more than one cell at a given time. With soft HO the connection to the prior cell's node is still valid for a limited period of time and the mobile continues to receive and accept radio signals from both cells before the connection to the old cell is fully terminated. This is referred to as (make-before-break).

Furthermore, hard HO is considered as an "Event" during the ongoing communication and requires the least processing by the network, while soft HO is considered as a "State" of the call during the ongoing communication rather than an event.

5.3 Handover in LTE

LTE supports only hard HO and doesn't support soft HO. The HO in LTE is initiated and controlled by the eNodeB, with support from the UE, i.e. the UE provides measurements which help the eNodeB to make HO decisions. There are two types of HO procedure in LTE for UEs in active mode: The S1 HO procedure and the X2 HO procedure [32].

The S1 HO procedure is mainly used when the HO is towards another RAT (i.e. inter-system HO) or when there is no X2 connection between the eNodeBs. This type of HO involves exchange of several messages with the core network and could increase the latency and lead to service interruption.

In the other hand, the X2 HO procedure is normally used for intra-LTE HO, which involve a change of eNodeB (inter-eNodeB) and/or a change of SGW.

The key features of X2-HO for intra-LTE (inter-eNodeB) HO are [24]:

- The HO is directly performed between two eNodeBs, direct interaction between the source and the target eNodeBs, making the preparation phase and the release of the resources quicker.
- Data is forwarded from the source eNodeB to the target eNodeB in order to minimize data loss.
- The CN is only involved at the end of the HO procedure by informing the MME when the HO is successful, in order to update the path switch in the S-GW toward the target eNodeB.

Figure 5.1 illustrate the X2-HO procedure. The HO can be termed "seamless" or "lossless" according to its resilience to packet loss. The two modes use data forwarding of user plane downlink packets. Seamless mode minimizes the interruption time during the move of the UE, and is used to forward packets in RLC UM, which means that the packets that are already processed by the PDCP or the RLC layer buffers but not yet forwarded to the UE will be lost. While lossless mode tolerates no packets loss at all; this is done by forwarding the packets using the RLC AM, this includes PDCP PDUs that are not yet transmitted, and those transmitted but not acknowledged.

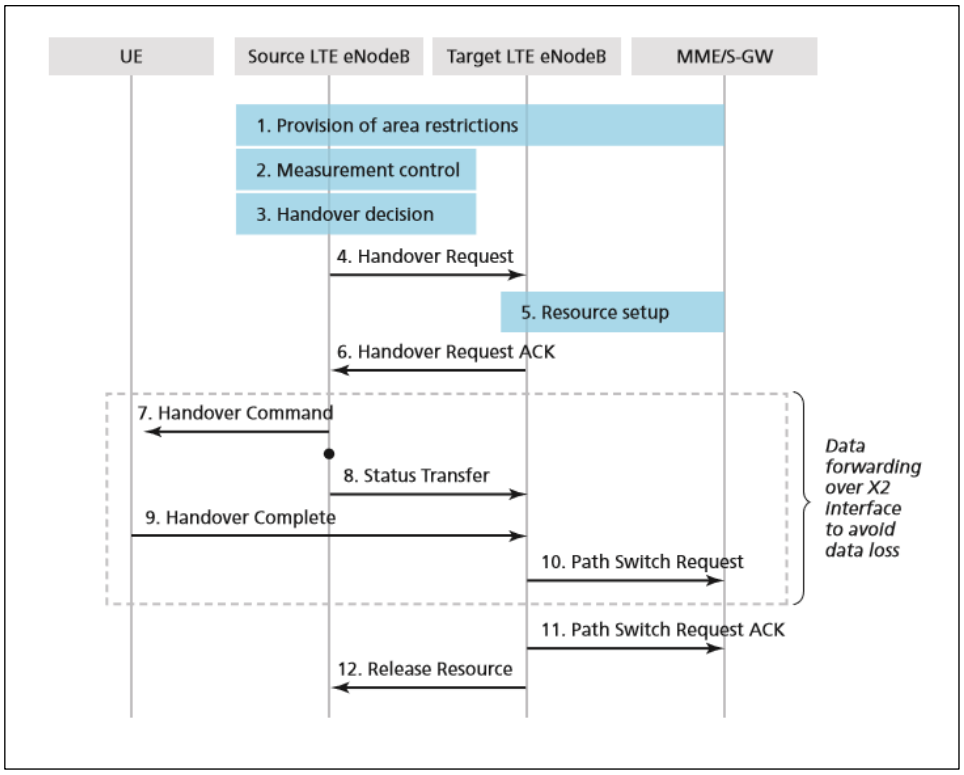


Figure 5.1: x2-based HO procedure [24].

In case of in-sequence delivery of packets is required, the source eNodeB will provide the target eNodeB through the Status Transfer message (i.e. step 8 in fig 5.1) with the sequence number (SN) that should be assigned to first packet to be delivered. With the Status Transfer message being sent, the source eNodeB stops assigning PDCP SNs to downlink packets and freezes it transmit/receive status [7]. The PDCP and RLC layers' structure are shown in Figure 5.2.

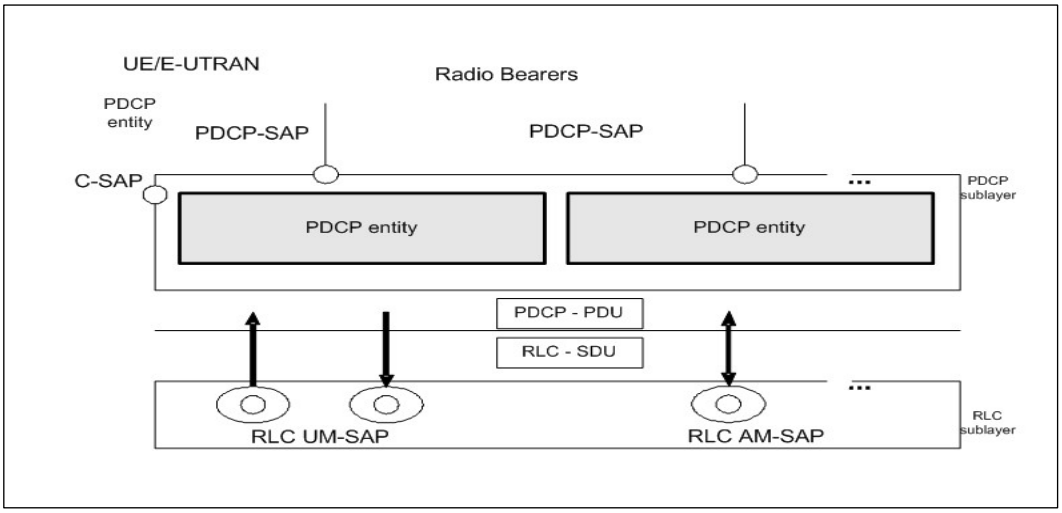


Figure 5.2: PDCP layers and structure view [102].

5.4 System Design and Implementation

In this section the architecture of the HetNet system for UE's with DC where the mobility is controlled by the centralized MeNodeB will be explained with the fundamental principles of parameters and measurement involved in the design of the system.

5.4.1 System Model

The architecture is based on the 3GPP LTE-A Evolved Packet System (EPS), with the same main components for radio and core networks (Release 12) [34]. As explained in the previous chapter, SCs will be distributed as hot spots covering specific areas under the coverage of the macro-cell layer. The SCs layer with frequency (F2) will be located at the centre of the hotspot, where the macro with frequency (F1) will be like an umbrella covering the SCs layer as shown in Figure 5.3.

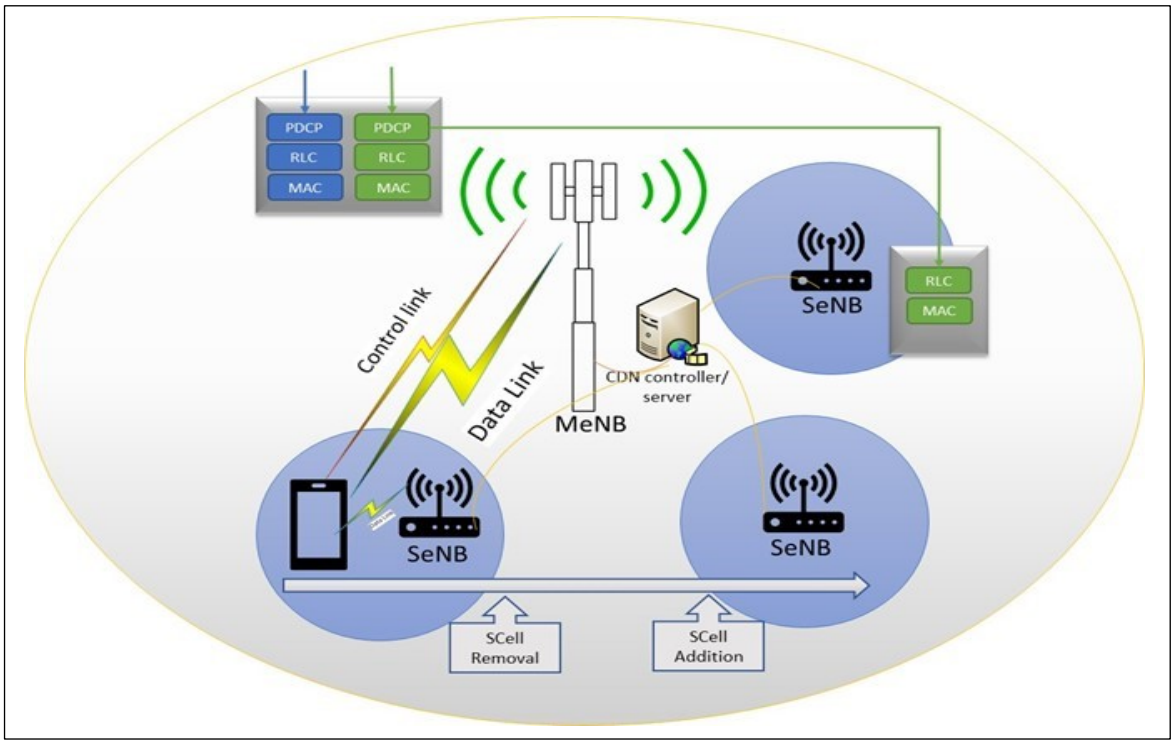


Figure 5.3: Proposed system architecture

A UE with DC will maintain an RRC CONNECTION with the MeNodeB at all times, while receiving U-plane data from the MeNodeB and SC simultaneously. Hence there will be only one S1-MME connection per UE. The MeNodeB will be used to provide coverage and the main services required by the UE will be the responsibility of the MeNodeB as it keeps all

time connection with the UE and it is the main point of attachment to the core network, on the other side, SCs could be used to provide capacity and some specific services such as content delivery. In such way, the mobility of the UE will be controlled by the MeNodeB, and there will be no need to move the RRC context of the UE between the SC's as the MeNodeB will be responsible of the RRC signalling with the MME.

Figure 5.4 shows the division of control between the MeNodeB and the SCs in DC.

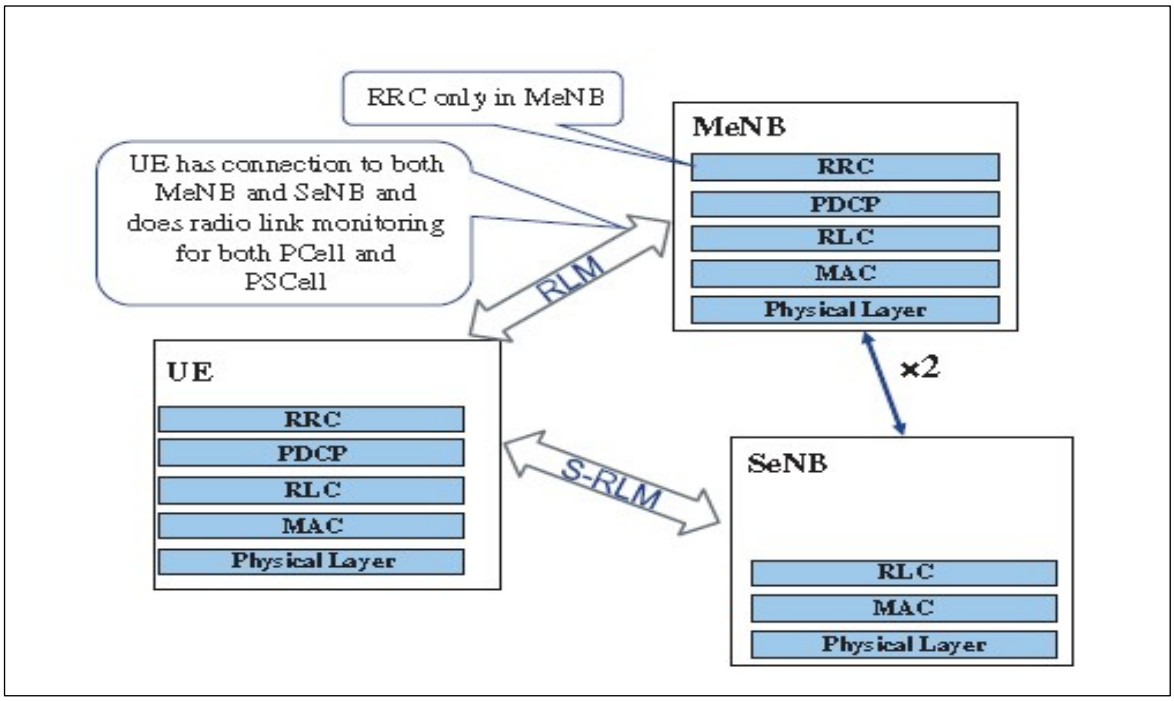


Figure 5.4: Division of control between MeNodeB and SeNodeB (i.e. SCs) [31].

In this method, HO between SCs will be like adding and removing new cells as all the information and resource management (RRM) are in the MeNodeB GW. The system will be a centralized radio access network similar to the C-RAN with RRH; but, instead of the RRH that just have a RF layer, the SCs will have a protocol stack that eliminate the need for the RRC and PDCP layers and have the RLC , MAC and PHY layers. The Service Data Units (SDUs) coming from the PDCP layer are then encapsulated and directed to the RLC layers in each eNodeB that will convert them to Protocol Data Units (PDUs). Passing through the rest of the protocol stack, UE will receive the data from one or more eNodeB, assuming the scheduler in the MeNodeB GW has shared the total set of the resource blocks, separate IFFT functionality could be used to separate between the signals. The (PDCP) layer (Fig.5.2) within its responsibility is to transfer the user plane and control plane data by means of

PDCP entities, several PDCP entities may be defined for each UE, and each PDCP entity is carrying data for one data bearer.

In terms of mobility, a UE in RRC connected mode relies on network-controlled UE assisted mobility, in which the network informs the UE via a dedicated RRC signals when it has to perform HO process to the required cell. When a UE follows a certain trajectory, while in DC, it can have its PCell on the macro layer, i.e. MeNodeB, while SCell can be configured within the small cell layer, i.e. SeNodeB.

As stated earlier in this chapter, the HO in LTE system supports only hard HO. hard HO is considered as an "Event" during the ongoing communication and requires the least processing by the network.

Event is a set of predefined measurement indicating the change of the signal strength above or below a threshold value. 3GPP defines several sets of these predefined measurement for the LTE system, mobility is triggered by an event, and each of these events can be based upon either RSRP or RSRQ, and Layer 3 filtering evaluation is applied as the criteria for whether or not any event is to be triggered [101].

Table 5.1 list these different events:

Event Type	Description
Event A1	Serving becomes better than threshold
Event A2	Serving becomes worse than threshold
Event A3	Neighbour becomes offset better than serving
Event A4	Neighbour becomes better than threshold
Event A5	Serving becomes worse than threshold1 and neighbour becomes better than threshold2
Event A6	Neighbour become offset better than SCell (This event is introduced in Release 10 for CA)
Event B1	Inter RAT neighbour becomes better than threshold
Event B2	Serving becomes worse than threshold1 and inter RAT neighbour becomes better than threshold2

Table 5.1: List of events.

In such architecture, two types of HO could occur, the first is between different MeNodeBs and the second is between different SeNodeBs (i.e. SCs). The assumption is that MeNodeB HO is based on the RSRP A3 event as the neighbouring cells in A3 reporting event can be either intra-frequency or inter-frequency which can be well served by the PCell in the MeNodeB, while addition and removal of SCells in the SeNodeBs (SCs) under the MeNodeB coverage, is based on the A6 event trigger for intra-frequency HO, as the neighbouring cell in A6 reporting event must be on the same frequency as the secondary cell, i.e. this reporting event can be used to trigger a change of the secondary cell on that frequency.

A network will keep control of the mobility of the UE as the MeNodeB is controlling the C-plane. All SCs are connected to the controller of the MeNodeB they are laying under its coverage area and a list of the attached SCs will be generated and modified every time new SC is to be added. Information about the network and its users will be gathered at the controller. Such information includes mobility, number of users per cell, list of candidate cells for under the MeNodeB coverage. This can provide significant impact in the UE data rate. When those SeNodeBs (i.e. SCs) are well known by the control server of the MeNodeB and sharing the same C-plane, all the information necessary to provide the required service can be started directly after the UE is under the SC coverage. The triggering criteria for SCell addition is configured by the MeNodeB controller by means of certain event (i.e. A3 and A6), in this case RRM measurement are not reported to the core network as it is already the MeNodeB's server responsibility. Similar architecture for controlled RRM and HO between SCs could be a promising method to reduce signalling overhead towards the core network, as well for increasing the data rate per UE.

The controlling capability that is added to the MeNodeB is suggested to be implemented as an equivalent to imitate the MEC/CDN functionality like caching, and other specific operations such as synchronization, baseband processing, interference coordination, carrier aggregation, inter-eNodeB CoMP and DC which is yet to be standardized and not yet available in the OPNET simulator.

5.4.2 Reporting Measurements

Reporting of the measurement required to perform the HO and event triggering are based on the RSRP and RSRQ made by the UE and then reported to the eNodeB as defined in the 3GPP specification in [TS 36.133 and TS 36.214].

The RSRP is the average power received by the UE from a single cell and can be measured as:

$$RSRP_{i,UE} = P_i - L_{UE} - L_f \quad (5.1)$$

where:

$RSRP_{i,UE}$ is the reference signal received power for a UE received from cell i

P_i : is the transmit power for eNodeBi

L_{UE} : Path loss gain from the UE to the source eNodeB

L_f : is the fast fading channel gain

RSRQ is the reference signal received quality and can be measured as:

$$RSRQ = N \times \frac{RSRP}{RSSI} \quad (5.2)$$

where: N is the number of resource blocks (RB)

RSSI is the carrier received signal strength indicator of the total received power from all sources around the UE, serving and non-serving cells, including interference and noise.

RSRI is measured as:

$$RSSI = RSRP_{s,UE} + RSRP_{int,noise} \quad (5.3)$$

Hence, assuming that the UE measures the reference signals from the neighbouring wireless channels on periodic basis. The collected measurement passes through a processing mechanism of averaging and filtering before any event to be triggered. After applying the layer 3 filtering the processed measurements are to be reported to the MeNodeB where the handover decision is to be made. The processing mechanism of layer 3 filtering eliminates the fading effect and provide more stable measurement values by estimation inaccuracies

from the reported measurements to ensure the accuracy of HO decision and can be calculated using the below equation:

$$R_{L3}^{n} = (1-\propto)R_{L3}^{n-1} + \propto . M_n \quad (5.4)$$

where:

R_{L3}^{n} is the updated filtered measurement result

R_{L3}^{n-1} is the older filtered measurement result

$\propto$ = 1/2^(K/4) where K is the appropriate filter coefficient that can configured independently for RSRP & RSRQ.

M_n = Latest received measurement result from physical layer.

layer-3 filtering is applied only in RRC-CONNECTED and is applied during HO to ensure no HO is triggered due to bad measurements as no event is to be triggered before evaluated by layer 3 filtering.

5.4.3 Speed Dependent Effect

UEs with different mobility pattern must have seamless and robust mobility between the different nodes in the network, otherwise the resulting frequent unnecessary HO will lead to degradation in the QoS. Therefore, good attention to the UE's mobility situation by estimating the UE's speed must be taken in consideration.

In a wireless communication environment, continuous measurements aim to enable the UE to discover and attach to the new cells quickly when the signal quality of the serving cell is lower than a defined threshold in order to avoid a radio link failure. The UE may pass through the different cells with different speeds and this affects the robustness of the HO process, the channel estimation and reporting mechanism is delay sensitive and the HO process is triggered based on reported values of this mechanism. Therefore, fast measurement with evaluation processes are needed to search for and switch to a new cell.

In addition to mobility, another important consideration is how long the UE will camp under the discovered cell, and this is referred to as time of stay (ToS). A HO from one cell to another and then back to first cell is referred to as ping-pong, this is normally happening when the (ToS) in the second cell is less than a predetermined minimum time of stay (MTS).

MTS is the time required for UE to establish reliable connection with the serving eNodeB and begin data transmission [102].

In HetNet, when the UE moves fast it passes across a large number of cells quickly, this may degrade its performance especially when the cells have small coverage areas which the UE can pass very quickly after connecting with very short ToS. In addition, the UE doesn't know the location of the SC, so it has to keep performing measurements searching for a new cell and restart the attach procedure with every cell the UE passes through which is not efficient for the UE's battery power consumption and the performance of HetNet that could undergo high potential degradation [103].

Deploying a large number of SCs increases the site density. And the presence of SCs within the coverage of the macrocell creates hotspots. In this situation, if the mobility patterns have not been taken in consideration in comparison to pure macrocell deployment, the UE HO rates will increase because the number of HO is inversely proportional to the cell size. High HO rates have a negative impact on the system when the measurement mechanism and reselection process is performed frequently, in addition to the result in high percentage of radio link failure due to failure in the HO process [104].

A very possible scenario, is when a high speed UE passes through multiple SCs coverage each for a short period of time; SC that have low power capabilities, provides short range and limited coverage for the UE; this could result in a notable number of unnecessary HOs which causes multiple and therefore a noticeable reduction in QoS.

Therefore, in order to ensure good HO performance (i.e. minimizing the number of unnecessary HOs, reduced HO Failure and Ping-pongs) accurate estimation of mobility state is considered as a dominant objective for the HO strategy.

5.5 Performance Evaluation

After introducing some basic terminologies and concepts, a study of the UE behaviour moving under the coverage of MeNodeB and receiving data from a content server attached to that MeNodeB while experiencing HO procedure between the SCs will be presented. The study is based on the system architecture explained in Section 5.4.1 of this work and will be implemented using Riverbed Modeler formally well known as OPNET Modeler.

5.5.1 Simulation Setup

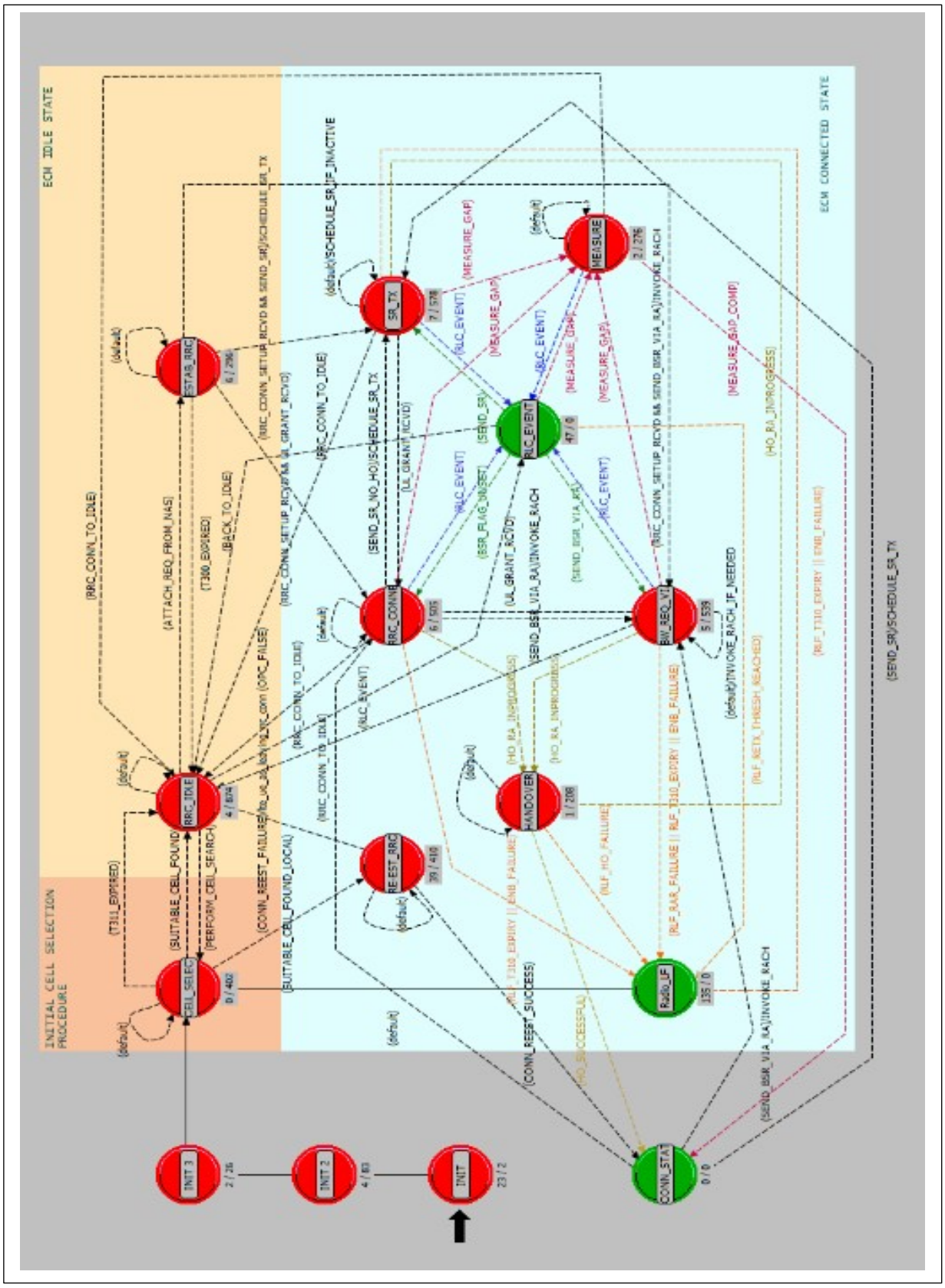


Figure 5.5: LTE-As process model.

The Riverbed LTE Modeller 18.7 is based on the 3GPP Rel.8., therefore, a modification to the system has to be made to support DC using the device creator operation to create custom LTE nodes for our network [OPNET documentation create Custom Device Model Operation]. The HO process state for the UE is reside in the (LTE-As) node mode which also contains the RRC-Connection states and the measurement states as shown in Figure 5.5. The modification to the (LTE-As) states will require adding new states as secondary link states in similar way to the existing states in order to serve the requirement of DC by the UE. In this case the UE will have only one RRC-Connection state within the MeNodeB that keeps it connected to the system and have the secondary states required for DC, as explained in Figure 5.4.

An LTE system simulation scenario that consist of 1 MeNodeB that will represent the coverage area where the UE will be moving, and 4 SCs that will represent the hotspots dedicated for specific service and distributed under the MeNodeB coverage using frequency band different from the one used by the MeNodeB is shown in Figure 5.6.

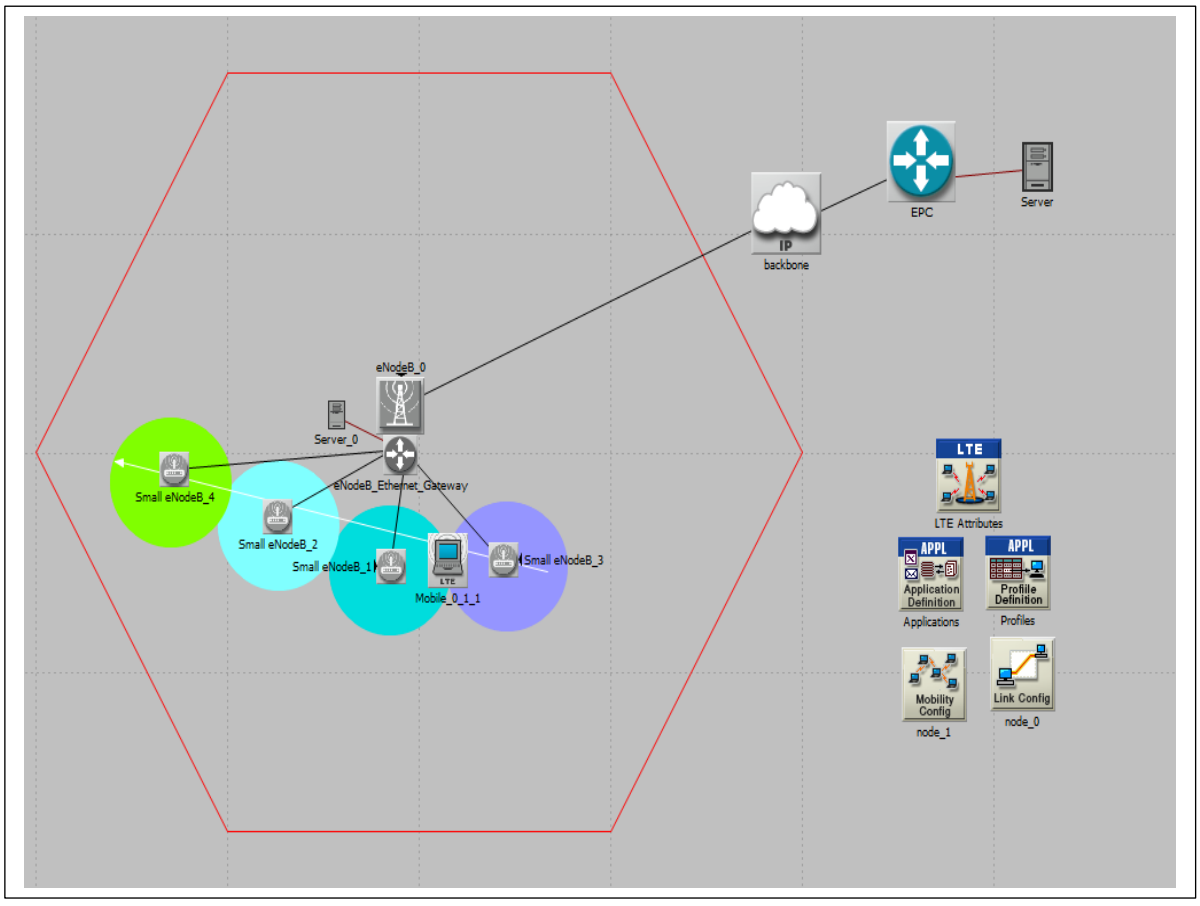


Figure 5.6: Deployment scenario.

The UE will keep RRC-Connection with MeNodeB at all time of the simulation, thus having 1 S1-C interface with the MME through the MeNodeB. Mobility management function tracks the location and activity of a UE in the network. The core network needs to know the location of a UE for traffic transmission. A measurement of RSRP and RSRQ are supported, thus the UE continuously measures RSRP and RSRQ for all cells within the range. The MeNodeB is configured to receive periodic measurement from the UE. In addition, it will be managing a table of neighbouring cells and UE information. By using this table, the MenodeB can manage the SCs for DC transmission to provide user centric service and maintain the C-plane for UE within its coverage area. HO is initiated and controlled by the MeNodeB, assisted by the UE measurement and triggered when the RSRP of the serving SC becomes lower of that of the target SC based on event A6, as expressed in Equation 5.5.

$$RSRPT > RSRPS + \Delta \tag{5.5}$$

where:

RSRPT: is the RSRP from the neighbouring SC.
RSRPS: is the RSRP from the serving SC.
Δ: is the A6 offset.
In addition, modification and release of the SC's SCells, the MeNodeB initiates the procedure to add or release SCells of the SC upon receiving the Measurement Report from the UE. The SC status may trigger the MeNodeB to add (e.g., SC underutilization) or to release SCells (e.g., SC overload). Then the SC decides its configuration within the restriction. After the SC decides the configuration, it forwards results to the MeNodeB by using *SCellToAddModList* or *SCellToReleaseList* for adding or releasing, respectively. The MeNodeB forwards the SCell configuration to the UE using a message container in the RRC Connection Reconfiguration without modifying message content sent by the SC. The UE then follows the instructions of the eNodeB.

Data forwarding, while HO is supported and the SC modification event is used by the MeNodeB to modify, establish or release bearer contexts. The report interval function and numerical reports attributes within the Riverbed modeller govern the number of reports sent to the MeNodeB. UE will perform cell search and generate reports, when the reported measurements violate the HO triggers, the MeNodeB decides to HO the UE to a different cell from a list of the SC in its attached server. The modification of SCs HO procedure is shown in Figure 5.7.

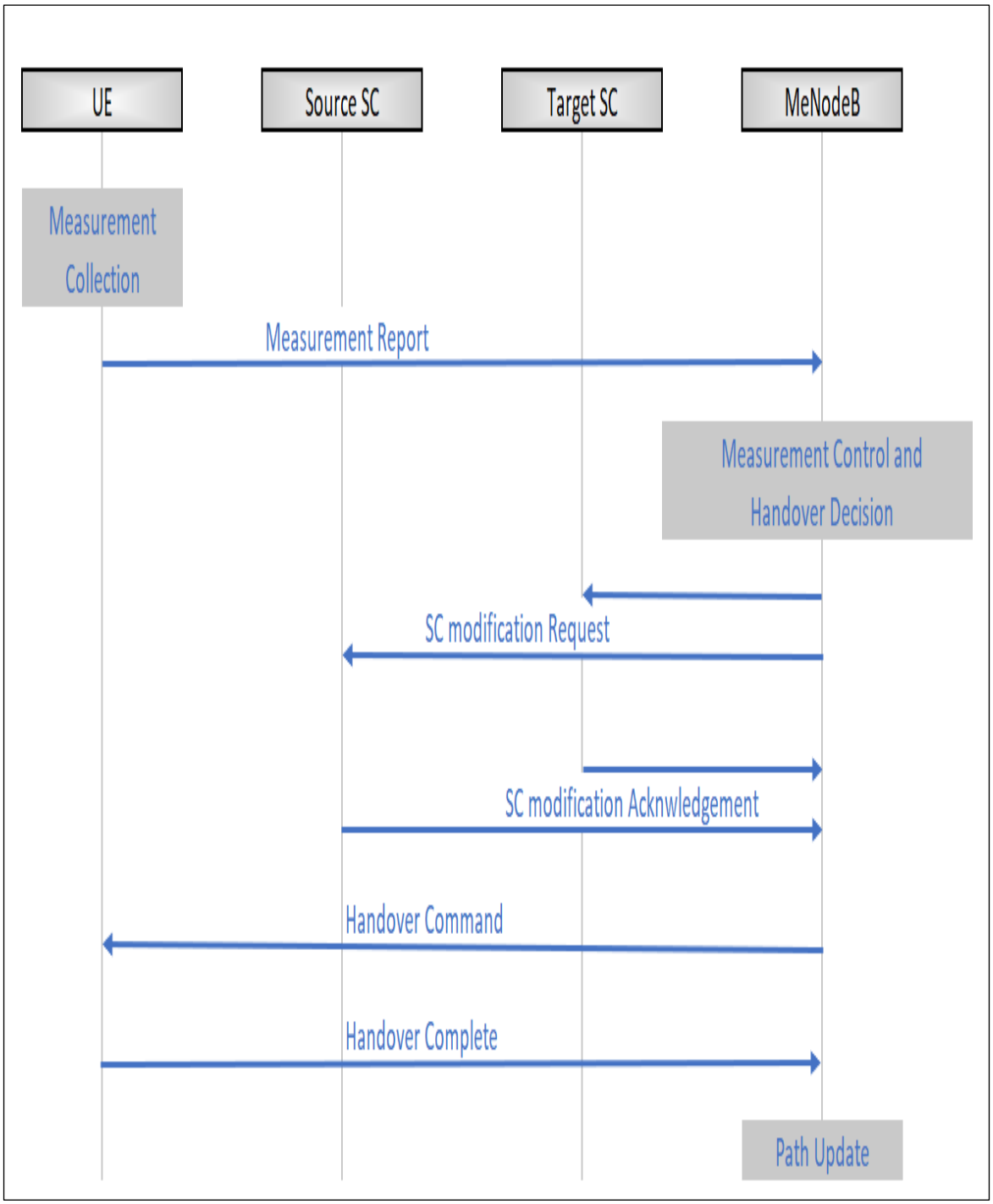


Figure 5.7: SC modification procedure.

The system performance of a UE under the coverage of MeNodeB and receiving data from both MeNodeB and one of the SeNodeBs while moving in a specified defined trajectory that insure the UE will be under the coverage of one SC within the MeNodeB coverage area and will be examined based on the parameters in Table 5.2.

Parameters	Values
Scenario	scenario#2 in the 3GPP TR 36.842 V12.0.0 (2013-12)
Deployment	1 3-sector MeNodeB site 4 SeNodeB sites per macro cell area
MeNodeB Carrier Frequency (F1)	2 GHz
SeNodeB Carrier Frequency (F2)	3.5 GHz
LTE B.W./Duplexing	20 MHz/FDD
MeNodeB Tx Power	46 dBm
SeNodeB Tx Power	30 dBm
UE Tx Power	23 dBm
Traffic model	FTTP
File Inter-arrival Time	Exponential
Service type	As requested by UE

Table 5.2: Simulation environment parameters.

5.5.2 Results Discussion and Analysis

In this work, HetNet scenario is considered with SCs deployed under the coverage of MeNodeB. The scheme proposed is analysed based on the discussed settings and scenarios explained in the proceeded sections of this chapter.

The trajectory setup took in consideration the speed of the UE, a pedestrian mobility profile was created for the UE in order to avoid passing the SCs in high speed, this was done for two reasons; the first one is instability. When HO is performed to a SC, there will be a short time of stay within its coverage (when the UE moves fast compared to the cell size then the UE has to find a new cell and starts new measurements). The second reason is to avoid instantaneous changes in the load of the SC because there can only be a limited number of users connected to a SC.

In the preparation phase of this deployment we investigate the relative distance between UE and SC based on the randomly generated trajectories, the UE's movement and the coverage area of the SC and the MeNodeB are depicted in figure 5.8.

The observation is that a separation distance >10 meter from the SC results in drops in the services which means loss of the connection between the UE and the SC as shown in Figure 5.9. and represented by the effective coverage area in dashed green line in figure 5.8.

In figure 5.9, the red line represents the effective coverage of the SC. After this distance the UE will lose connection with the SC and no data will be received from the SC, approximately at time (30 s) and this is represented in figure 5.8 by point (a) as this is the point when the

UE is out of the coverage area of the SC. The UE will reconnect to the SC at time about (44s) in figure 5.9 and this is represented in figure 5.8 by point (b).

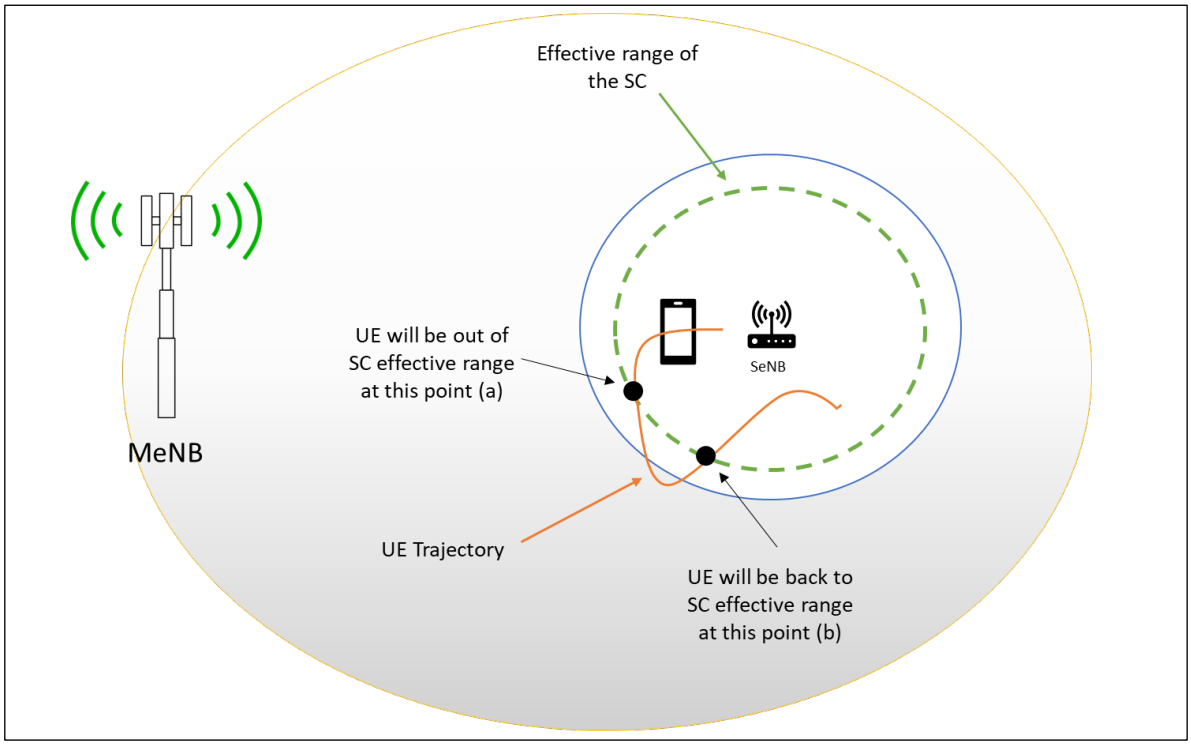


Figure 5.8: UE trajectory and effective range of the SC

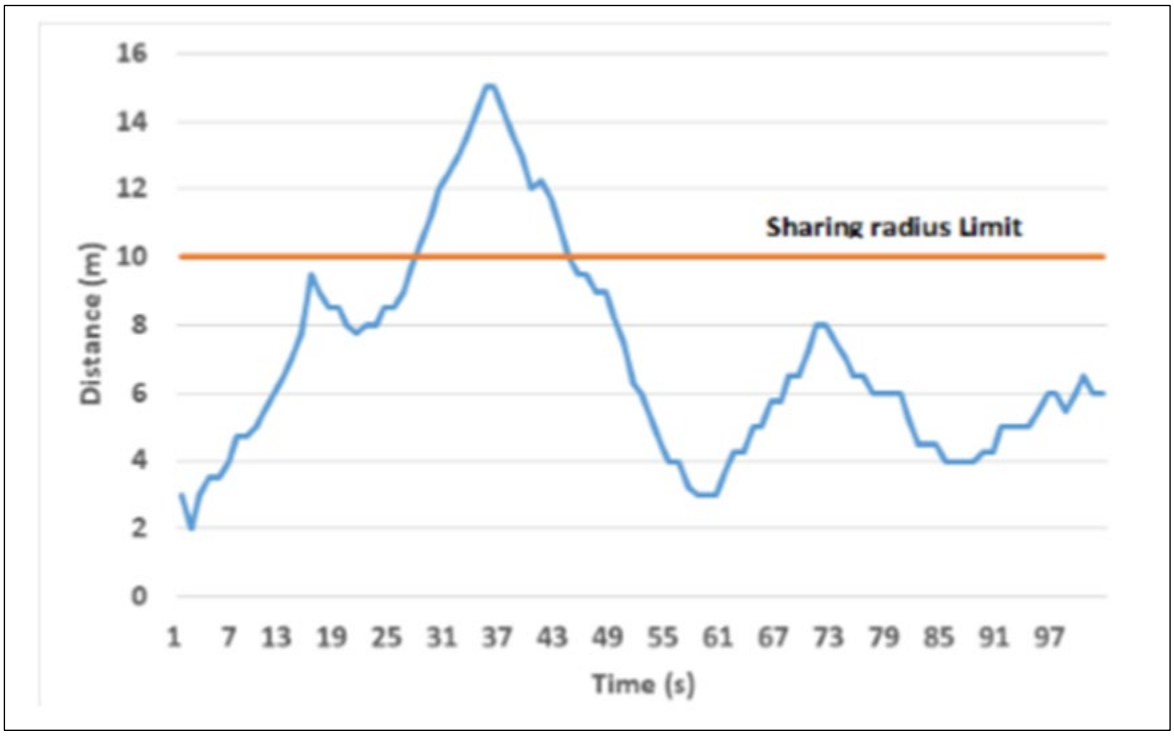


Figure 5.9: Separation distance [107].

In the same regard, we investigated the performance of the UE when approaching the SC with pedestrian mobility speed. The UE is connected to the SC and following a trajectory towards the SC of the Hotspot. The UE has 1200 kbps of downlink IP flows from a server connected to the CN. Around t = 160 s, the UE is within the coverage area of the SC. Therefore, we see that UE's throughput increases as it is getting closer to the SC, until t = 250 s and then decreases as it moves away from the SC. At t = 270 s, we notice that the received traffic is 0 kbps, as the UE is getting out of the coverage of the SC as depicted in Figure 5.10 and Figure 5.11. Figure 5.10 depicts the trajectory of the UE within the coverage area of the MeNodeB and the SC while Figure 5.11 shows the resulted traffic flow between the UE and the SC.

The dashed green line in figure

The dashed green line in figure 5.10 represents the effective coverage area of the SC and the two red dotes represent the entry and exit point of the UE to the effective coverage area of the SC at times 160 s and 270 s respectively.

In Figure 5.11, the red curve represents the traffic sent by the server to the UE throughout the SC and the blue curve represents the traffic received by the UE during its movement.

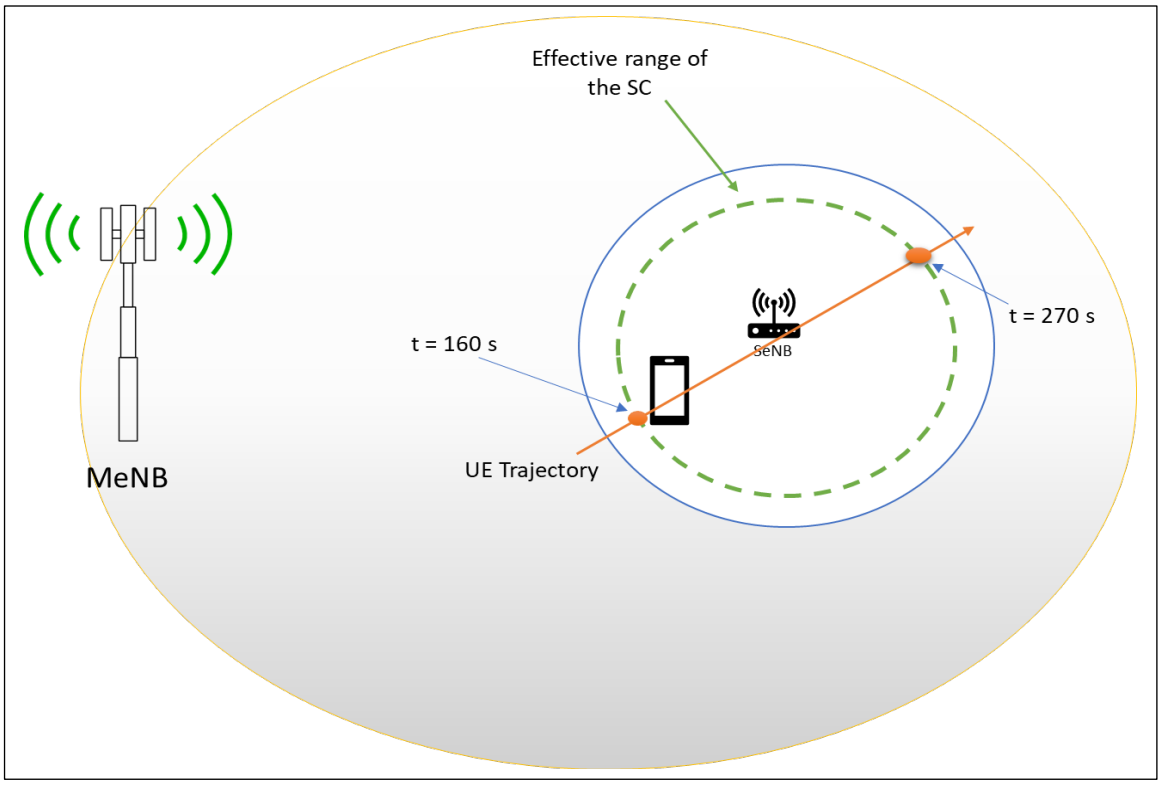


Figure 5.10: UE Trajectory within the effective SC coverage

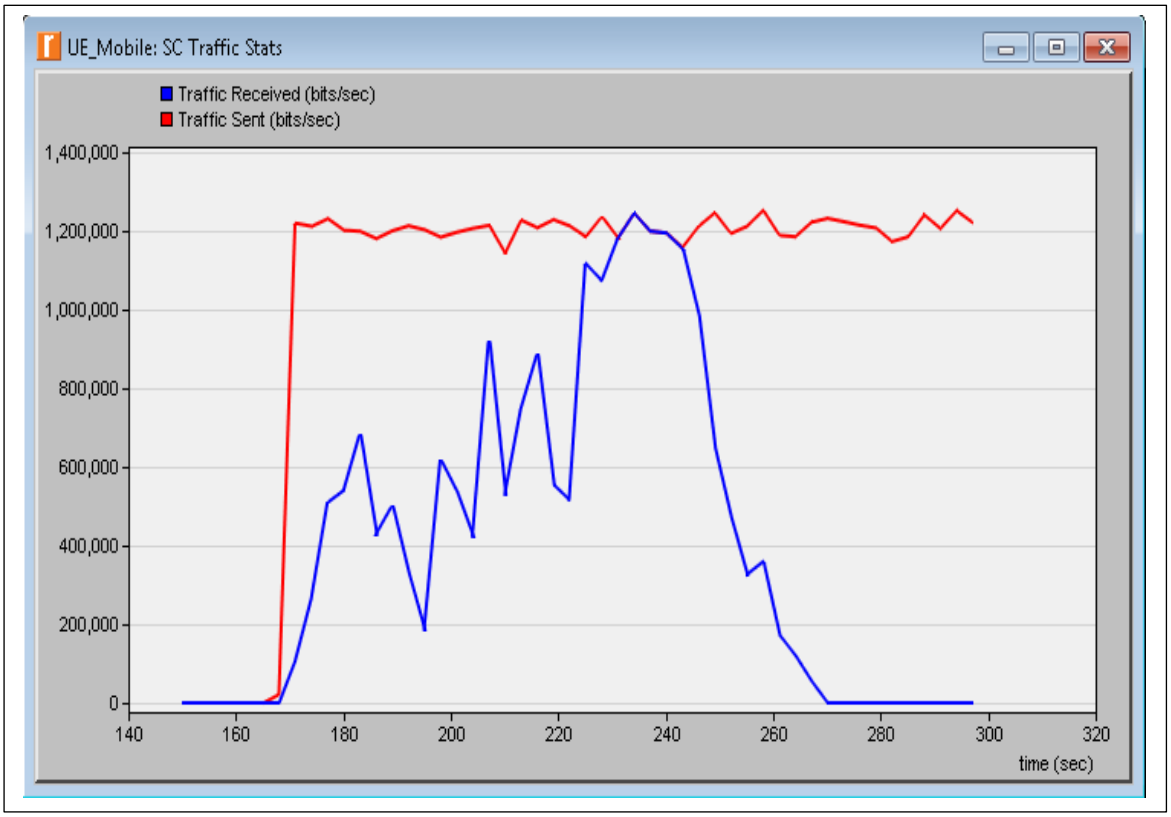


Figure 5.11: UE mobility traffic states.

Based on the above observations HO commands will be configured for a pedestrian (low) speed UE, enhancement on the trajectory of the UE will be focused on locating the UE within the coverage areas of the SCs.

In the simulated scenario, trajectory is set for the UE for mobility within the MeNodeB coverage area and UE configured to perform cell search and select the best suitable SC, the UE model supports the measurement of RSRP for each cell to aid in the cell selection process. RSRP is measured for the MIB packets. Cell selection threshold that's configured on the SCs is compared with the Scanned SC RSRP. Figure 5.13 shows the RSRP distribution during mobility.

Recalling from figure 5.6, Mobile_UE initially is associated to eNodeB_0, serves as MeNodeB, and to eNodeB_3, serves as SC. This is shown as well in 5.12. As eNodeB_0 is represented by the red line that shows the connection with MeNodeB for the whole simulation interval, keeping all time connection for the UE with the system, while the blue

lines represent the SCs. MeNodeB is configured to receive periodic measurements from the Mobile_UE.

When Mobile_UE moves and approaching another SC coverage area, the MeNodeB will trigger measurement reporting to the UE and the UE starts to generate reports to perform HO. This appears in Figure 5.12, represented by the dashed purple lines at times 4, 5, & 6 mins, at these times UE will perform HO trigger between SC, while keeping attached to the MenodeB, as shown in Figure 5.13. The red line represents the MeNodeB to which the UE is attached for the simulation period, while the blue lines represents the SCs. As seen in the results in figure 5.13 show that HO is performed when the serving cell measurements reported by Mobile_UE violates the HO triggers. As explained earlier in the simulation setup.

In figure 5.13, the red curve represents the RSRP of the MeNodeB (eNodeB_0) which is steady during the simulation intervals as the movement of the UE is within its coverage with low speed. The blue curves represent the RSRP of the SCs and the HO event is triggered when the signal of one SC is going blow the threshold.

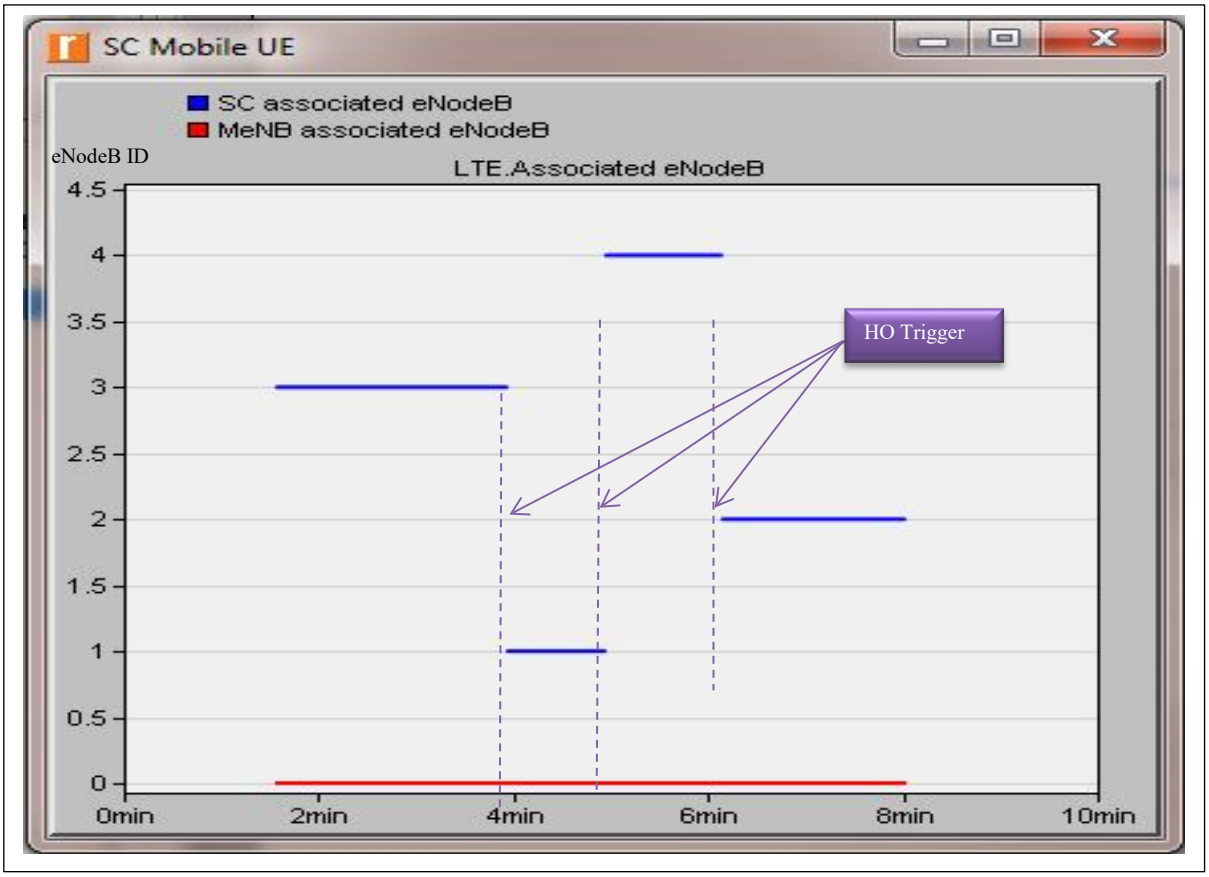


Figure 5.12: LTE associated eNodeB.

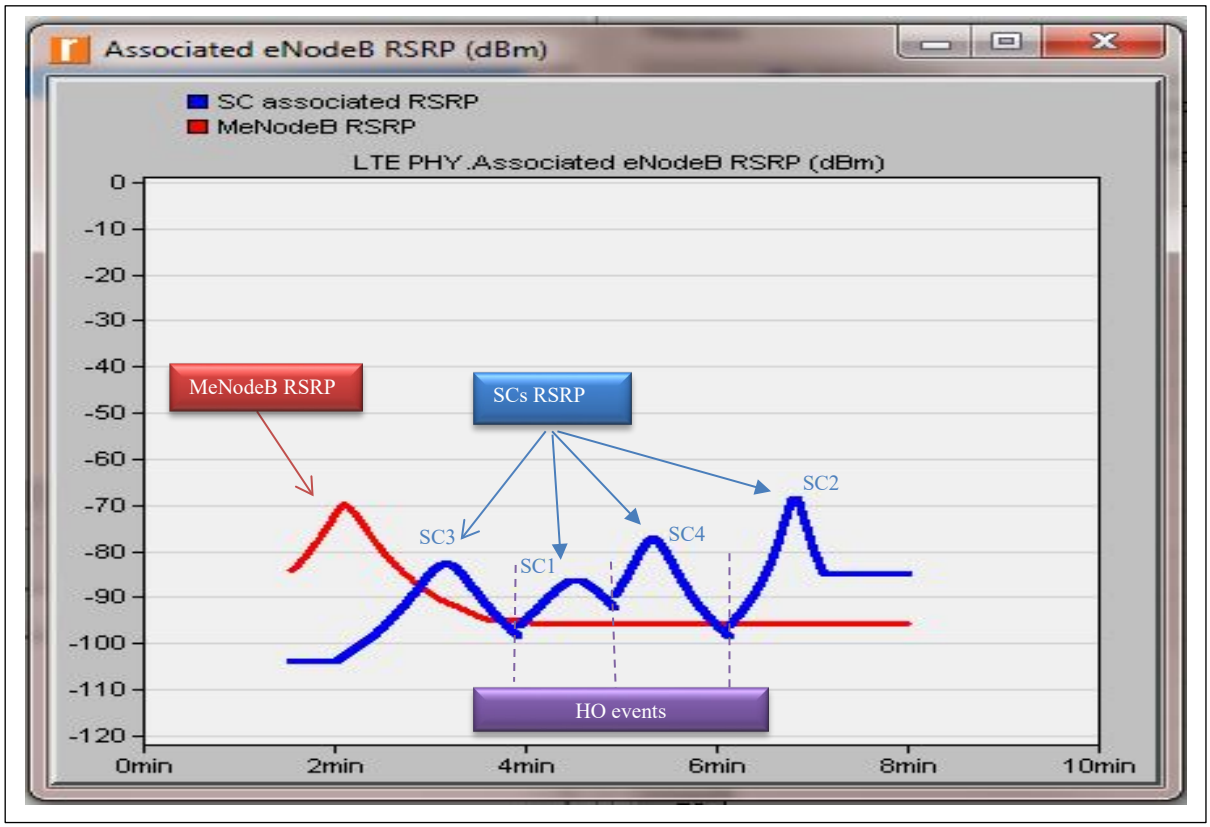


Figure 5.13: HO Region based on the RSRP.

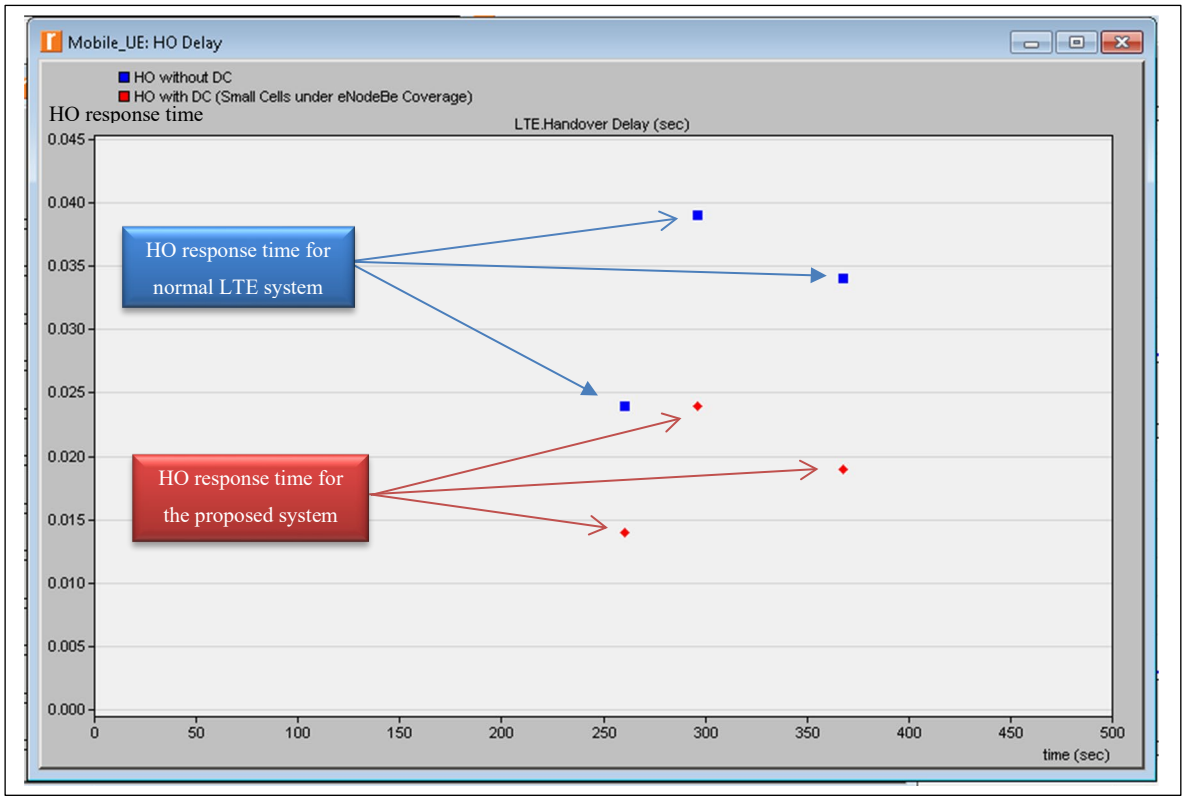


Figure 5.14: HO response time

Figure 5.14 shows the HO response time. In this scenario a comparison of the HO in terms of delay between the HO in a normal LTE system and the HO in the proposed DC system. The results show a higher response time for the HO in the case of the normal LTE system; HO response time are shown in Blue, In the panel "Mobile_UE: HO Delay". In this case the HO is taking place between normal eNodeBs. The red dotes represent the HO response taking place between SCs in the proposed system and showing less response time as the HO procedure is controlled by the main sever connected to the MeNodeB and controlling all the signalling required for HO trigger to take place.

The interesting observation is that proposed DC HO takes lesser time than the normal LTE HO, since the SCs communicate directly through the MeNodeB gateway.

In term of traffic, there are two FTP applications configured for Mobile_UE:

FTP 1: Between 100 s and 225 s.

FTP 2: Between 260 s and 410 s.

The result demonstrates that the total traffic received is the same in both cases; the case of the normal Lte system and the case of the proposed DC system. Due to TCP, no traffic is lost; what is different is the delays encountered during HO in the two cases.

The panel "UE_Mobile: SC Traffic States" as shown in Figure 5.15 demonstrates the traffic sent and the traffic received in the two cases. The upper graph represents the traffic received for the case of normal Lte system, i.e. the HO is taking place between normal eNodeBs while the lower graph is for the case of the proposed Lte DC system, i.e. the HO is taking place between the SCs.

The long-dropped lines out of both curves represent the time for the HO to take place and as it can be seen the HO is taking longer time to trigger in the case of the basic Lte system, which represents the case of the normal system while it took less time in the case of the proposed system. Other than that, there is not any other difference between the two cases as the used traffic is configured to be FTP application, therefore, no traffic is to be lost.

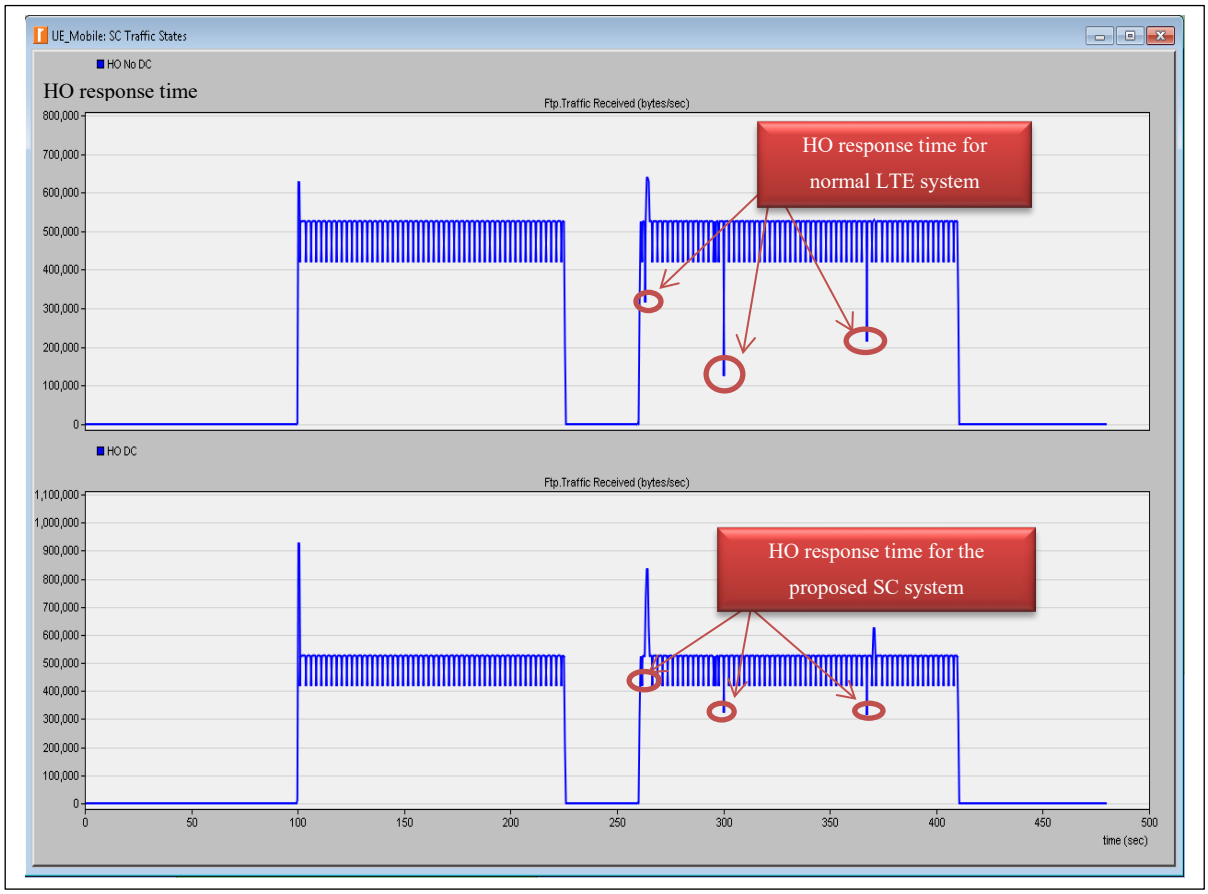


Figure 5.15: Traffic states.

5.6 Summary

The increased demand of wireless data services driven by smart devices capabilities and mobile internet led to an exponential growth on the traffic generated by wireless systems. In order to cope with this explosive growth of generated data, traditional macro-cell only networks have evolved to HetNets in order to improve the performance of the system. In such system, the MeNodeB will act as the main point of connection for the UEs' while SCs are deployed to provide specific services for the users throughout DC. This architecture of user centric network is considered as one of the most promising techniques for the next generation system. One of the main issues with such architecture is the frequent HO due to the large number of the SCs deployed under a relatively small area. In this chapter, a centralized mobility management system was proposed to overcome such challenges. Under the proposed architecture, the Macro-cell (MeNodeB) is the anchor for UEs movement, while the other SCs will act as the SeNodeBs, which only provide service for the users under their coverage.

Chapter 6

Differentiated Service and Network Slicing

6.1 Introduction

The vision of the next generation wireless network demands the various network operators to provide differentiated services targeting numerous applications to users with wide variety of network performance characteristics. Hence, the QoE for the users of the network should be pretty much the same whatever the service is. Traditional networks are service specific, having a one-size-fits-all approach. One of the problems with current implementation is that the networks are providing the same service characteristics no matter what the service is, versus a monolithic network architecture that carry services for multitude of applications with a broad range of requirements.

The future wireless networks should be designed to provide the necessary QoS to applications in an optimized environment capable of delivering a reliable user QoE across a wide variety of use cases. Network slicing is considered to be key technique for meeting diverse requirements with high degree of flexibility in enabling differentiated services to users which will be discussed in this chapter.

6.2 Differentiate Services (DiffServ)

Differentiated Services, in general, is a method of categorizing services into different classes by addressing the clear need for each class of service. We all examine differentiated services in our daily lives; for example, VIPs are afford special privileges due to their position or importance; they are granted a better class of service based on their ranks, such as the

different classes of tickets in airlines or events like big matches, where first class passengers or ticket holders provided with better services and seats and get immediate or faster access while common people have to wait in line. Another example is the parking areas reserved for the different classes of cars or people with special needs or parking areas reserved for families. It's worth mention that such differentiation in services is provided among different classes, and not among the same classes. As all people with the same class would receive the same treatment for the same service. For example, first class ticket holders all handled in the same way with no special treatment or differentiation between them.

Differentiated services arises to fulfil the need for different requirement and it is not necessarily a privilege. The concept of providing multiple classes of service (i.e. providing different levels of service to different classes of traffic) was clearly considered in the early steps of network's design (i.e. about 5 decades ago), and a Type of Service (ToS) byte was included in the header of IPv4; (equivalent "Traffic class" byte was added later to the header of IPv6); to distinguish the different types of IP datagrams from each other [80], as shown in Figure 6.1.

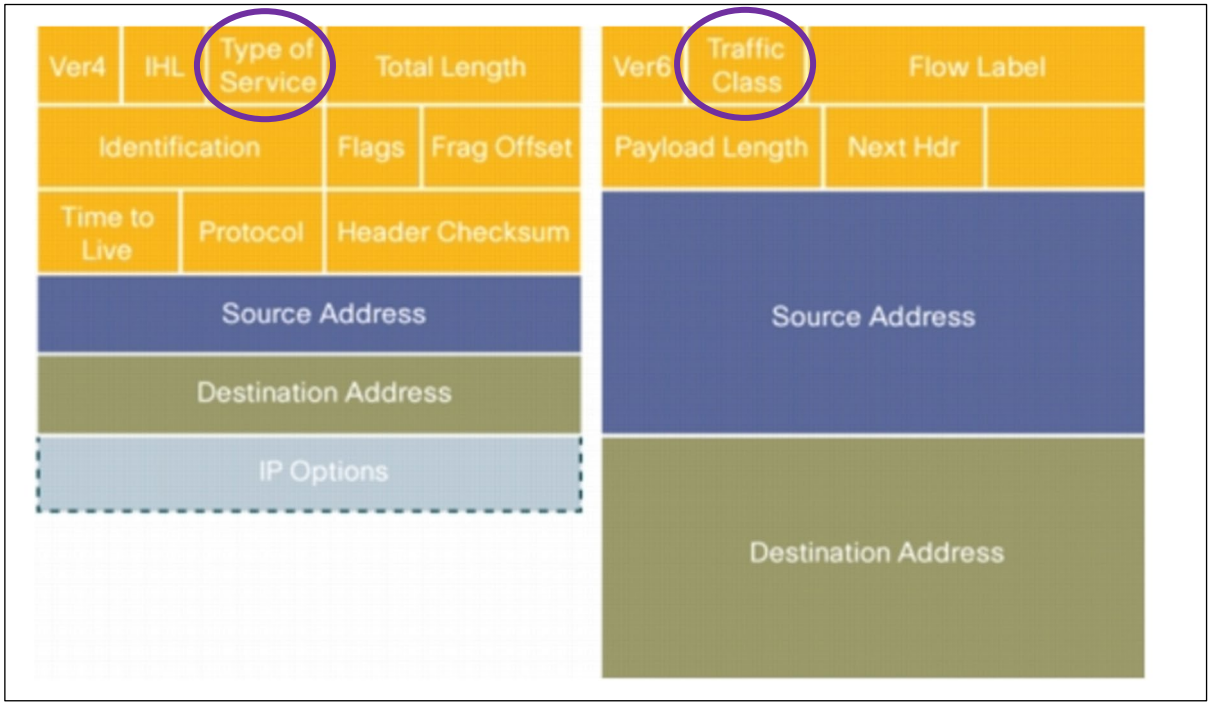


Figure 6.1: IPv4 and IPv6 Headers showing the ToS & Traffic Class Bytes [109].

Although it was designed to provide best-effort service for delivery of data packets, the Internet Protocol (IP) has allowed the development of a global internet, and the increasing

popularity of IP as an acceptable internet standard; developed by the Internet Engineering Task Force (IETF) [105]; has shifted the paradigm from "IP over everything," to "everything over IP." Dragging the internet to become an infrastructure that provides services to applications and spanning from computer networks to support the development of applications in mobile networks.

such applications are referred to as distributed applications , including traditional applications such as Web surfing and e-mail services, in addition to mobile smartphone and tablet applications such as internet messaging, online social networks, video conferencing, streaming different contents from the cloud, multi-person online games, and location-based recommendation systems.

These multitude of applications are running in multiple end systems and involve them to interact with each other and exchange data of different precedence, consequently, networks are required to provide Quality of Service (QoS) for applications in addition to best-effort service, as different applications have different needs for the parameters that form the basis of QoS (i.e. bandwidth, delay, jitter, packet loss, and availability). For this reason, networks should be designed to provide the necessary QoS to applications.

Having this motivation for providing multiple classes of service in mind, the IETF has defined the model of Differentiated Services (DiffServ). Where DiffServ works to facilitate true end-to-end QoS on an IP-network by categorizing traffic into multiple classes, also termed Class of Service (CoS), and applies varying QoS parameters to those classes [106].

6.3 Evolution and Slicing Theory Technology

Although the vision of providing differentiated services for different classes of traffic was in mind of the designers in the early stages of the implementation of the networks, however, it seems to take long period of time to be realized. The reason behind that could be that the early networks were designed to support a specific type of service for all users. For example, the Digital Subscriber Line (DSL) that was used in the 1980s and 1990s to bring an internet connection and information into homes and businesses, was providing price plans based on the speed the users want or are willing to get. Yet, the service provided was the same for all users. Similarly, the early generations of mobile networks were providing services based on

payment models, where users who are not qualified for telephone subscription plans, can have prepaid models to the same service [107].

The rise and widespread of mobile telephony (mainly, 2G and 3G systems), has brought with it a huge number of implications, influenced by the fertile environment surrounding the mobile systems, in terms of the competitions and challenges from other technologies started during the same time as 2G and 3G mobile systems, especially the rapid growth in the use of the internet. But probably the most crucial impact is the development and creation of new internet-based services for mobile devices, making them the foremost point of access to both basic telecommunication services, and to information and communication technologies (ICT). Leading as a natural next step to the creation of what is known today as mobile broadband [79].

Followed by that remarkable evolution in mobile system (i.e. mobile broadband), came possibilities for new services; supported by the higher data rate along with the emerging of smart phones, and the IP protocol being supported and later being adopted as the primary design target of the mobile network; this evolution enables network operators to offer wide range of more advanced services than ever before, leading to the appearance of applications not previously available which represent a big shift on the trend of mobile phone usage. Such a shift brings new variant requirements that require new management capabilities for the network and the services.

Being able to provide variety of services with different requirement represents a major challenge in the current and future mobile networks [79], therefore, the concept of bearer was used in the LTE system in order to support the divergence of use cases that require different quality of services (as will be explained in the later section) [24].

As a result, and as new use cases are emerging, the traditional structure of the network need to be broken down into layers (Slices) of network functions to classify and manage different type of services [68].

6.3.1 Network Slicing

The possibility to design logically divergent clusters of end-to-end networks tailored to allow differentiated services targeting different types of users with different QoS is where

the concept of ‘network slicing’ emerges from [30]. Network Slicing is a technology where multiple logical networks each with different performance characteristics can be created on the top of a common shared physical infrastructure [108].

As many use cases and service propositions are emerging quickly and even before the introduction of 5G, and as these use cases have divergent needs in terms of technical capabilities, network slicing can be used as key feature for the next generation network allowing the implementation of the differentiated services that deliver different guarantees such as availability, reliability and low latency.

These use cases that are broadly categorized as eMBB, URLLC and mIoT [30], and characterised by their different performance characteristics can be hosted on a common infrastructure allowing differentiated services targeting different types of users. Figure 6.2 shows different network slicing implemented on the same infrastructure and competing for the same network resources to allow differentiated services. One network can support one or more network slices, likewise a single network slice can provide the functionality of a complete network [109]. As the connectivity becomes differentiated, this will provide high degree of flexibility enabling innovative business models to be easily configured and several use cases to be active concurrently demonstrating additional value [110] [111].

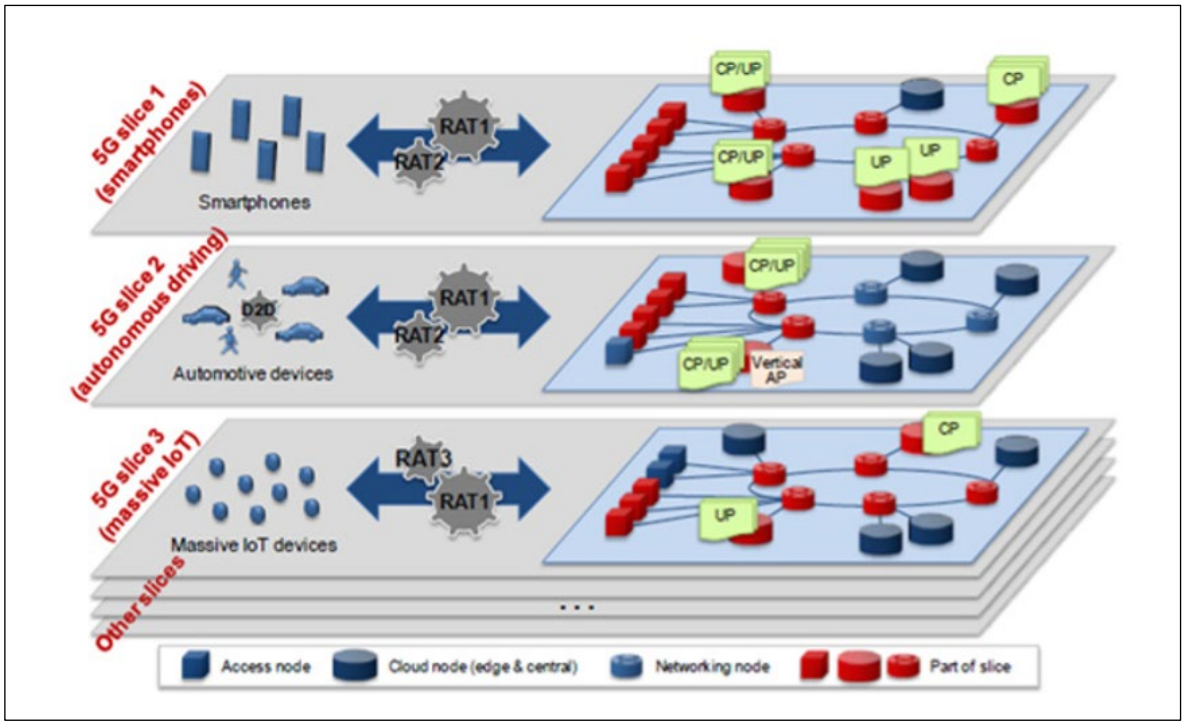


Figure 6.2: 5G network slices implemented on the same infrastructure [30].

The network slicing architecture contains access slices (RAN access), core network (CN) slices and the selection function. The selection function connects both the RAN and the CN slices into a complete network slice and routes communications to a suitable CN slice that is tailored to provide specific services. The need to meet different requirements for service/applications and for communication regulate the criteria of defining the RAN and CN slices.

A network slice would provide the connected devices a full network function support and would last throughout the intended service lifetime. A set of network functions (NFs) are used to setup CN slice. Some NFs are tailored to a specific slice or can be used across multiple slices. The mapping between devices and slices can be 1:1:1 or 1:M: N. For example, a device could use multiple RAN slices, and a RAN slice could connect to multiple CN slices as shown in Figure 6.3.

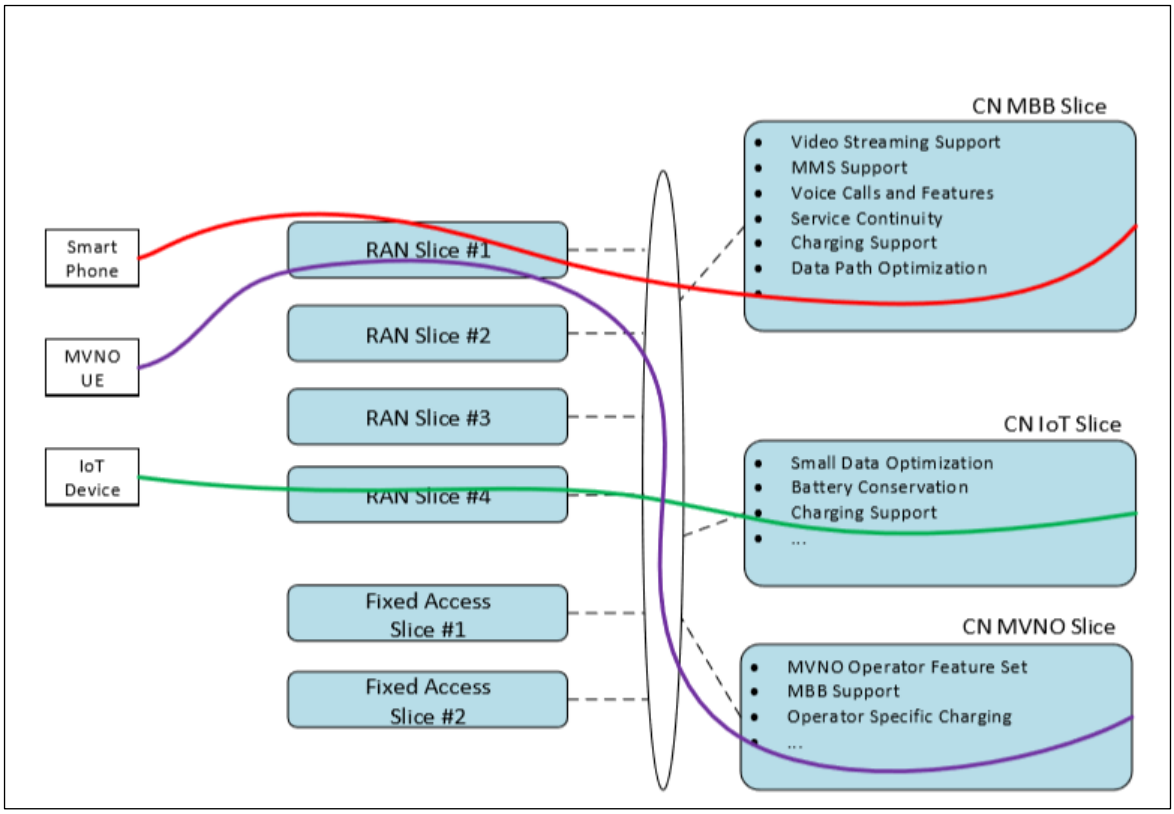


Figure 6.3: Network slice examples [110].

6.4 Multiple Classes of Service in LTE

Differentiated services has had history with telecom networks, the efforts for creating technical solutions to provide different level of services to different users were considered in the design of the networks, such as IntServ, DiffServ, and QCI [112].

The number of applications that require different demand of QoS is rising so fast, and because of that, the LTE system in its design use the concept of bearers as a technique to handle the IP traffic of such different application [24].

"*A bearer is an IP packet flow (i.e. transmission path) of a defined quality of service (QoS) between the gateway and the UE*." [24].

Typically, multiple applications having different requirements for the quality of service may be running in a UE at any time. For example, a UE might be engaged in a Voice over IP (VoIP) call while at the same time performing an FTP file download or browsing a web page. As these services have different QoS requirement, so it becomes important to differentiate between the different traffic in order to provide the required QoS for each service. Therefore, different bearers can be established for a UE, each being associated with the necessary QoS for each service. As VoIP has more stringent requirements for QoS in terms of delay and delay jitter, thus a VoIP bearer would provide the necessary QoS for the VoIP call, while FTP download and web browsing services can be assigned to a best-effort bearer as they require a much lower packet loss rate.

Bearers represent the flow of the IP traffic of the UE in the LTE system, when a UE is first attached to the network, it is assigned an IP address and at least one bearer called the default bearer is established. The default bearer provides the UE with always-on IP connectivity and remains established as long as the UE is in connection with the network. When required, the UE may request or be assigned one or more dedicated bearers that connect it to the same network for different type of services. No new IP address to be assigned to the UE for the creation of the dedicated bearers, instead the dedicated bearers share the same IP address and runs as a child to the default bearer.

The bearer used to provide an end-to-end service to the users in the LTE system (i.e. the EPS Bearer) is generated from a combination of sub-bearers. The hierarchy of the bearers is shown in Figure 6.4 [32].

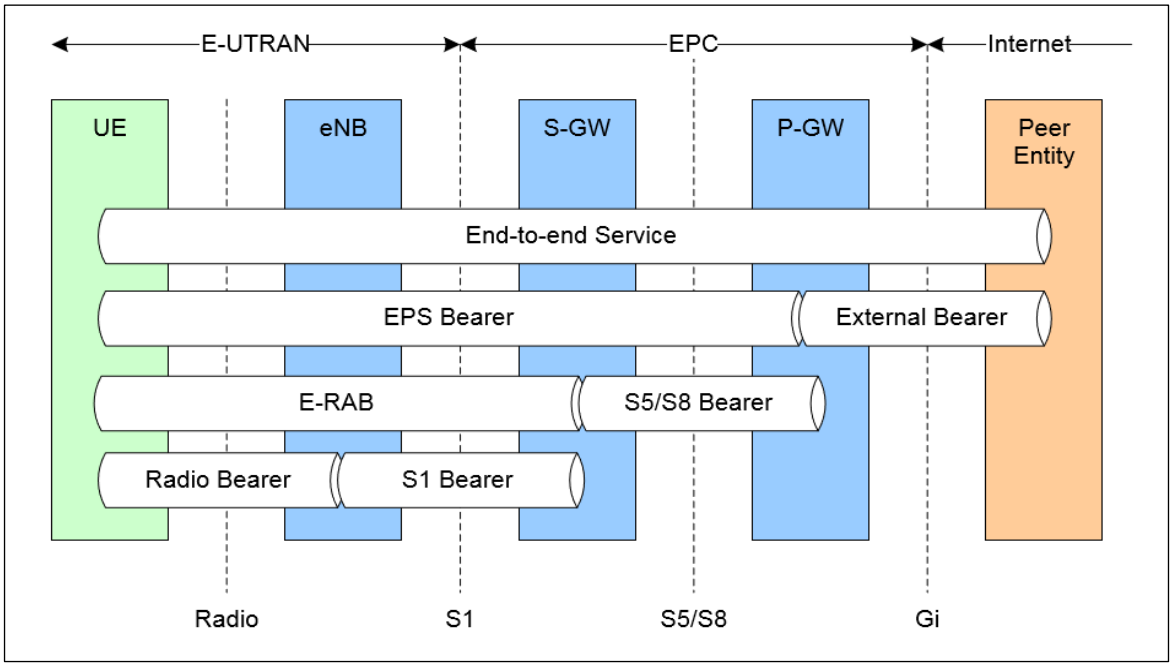


Figure 6.4: Bearers for LTE [32].

An EPS bearer is typically associated with a QoS in order to allow differentiation of both services and subscribers. If multiple bearers are established for a given UE, then each bearer is associated with a QoS that best suit the required service. Worth mentioning that all UE traffic handled by the same EPS Bearer has same the QoS. EPS bearers are classified into two classes: Guaranteed Bit Rate (GBR) and non-GBR bearer, based on the provided QoS [32].

GBR bearers have an associated minimum guaranteed bit rate value for which dedicated resources are permanently allocated at bearer establishment or modification time and can be used for real-time applications such as VoIP. Conversely, non-GBR bearers have no guarantees to any bit rate and no resources are to be allocated permanently to such bearers. These bearers can be used for non-delay sensitive applications such as web browsing or FTP transfer.

On the other hand, and as mentioned earlier, EPS bearers are best referred to as default bearer and dedicated bearer. While a dedicated bearer can be either a GBR or a non-GBR

bearer, the default bearer; since it is permanently established, has always to be a non-GBR bearer [24].

Moreover, bearers are associated with a pointer called the QoS Class Identifier (QCI). The QCI has been adopted by 3GPP to determines which bearers are considered as GBR and which are considered as non-GBR base on a set of standardised QoS characteristics, which are represented by priority, packet delay budget and acceptable packet loss rate [32].

The QCI label for a bearer determines how it is handled based on the underlying service characteristics and thus provide corresponding treatment for the service by the network.

6.5 Network Slicing in 4G

In today's 4G architecture, network slicing was experimented using multiple existing methodologies. Some of them include [110]:

- Service Specific APNs
- Service Specific PLMN
- DÉCOR/EDCOR

6.5.1 Service Specific APNs:

A methodology to isolate multiple services and consider them as slices by having multiple APNs on the PGW or having PGWs dedicated for certain APNs. Example: IMS APNs and MVNO partner APNs.

6.5.2 Service Specific PLMNs:

While Service Specific APNs used to isolate services, dedicated PLMNS are used to serve certain type of devices using an overlay EPC network. Example: overlay EPC network implementation for IoT (Internet of things) or Critical MTC as a separate PLMN.

6.5.3 DÉCOR/EDCOR:

DÉCOR (Dedicated Core) is a 3GPP feature standardized in release 13 [113] and is implemented using the 3GPP architecture for LTE and can be deployed for GERAN and

UTRAN as well. DÉCOR consists of one or multiple core network entities that allows an operator to deploy multiple Dedicated Core Networks (DCNs) within a single PLMN, therefore aligns quite well with the concept of "Network Slicing".

The DÉCOR feature does not require any modification or configuration of the UE and slice selection is based on an operator configured subscriber parameter (i.e. UE Usage Type). The process for selecting the right CN will have a negative impact on access times due to new evaluation steps and there will be also increased signalling due to possible re-direction messages between the CN and the RAN as well as additional request to the HSS may be required. Thus, enhancements beyond DÉCOR are required and therefore EDCOR is being standardized in 3GPP Release 14 as an evolution of DÉCOR with the objective to "*Improve DCN selection mechanism by providing assistance information from the UE* [114]. It also addresses some of the key challenges with DÉCOR related to signalling and increased delay. but since EDECOR requires UE support then it cannot be used with legacy devices. In addition, it also increases the dependency on the core network.

6.6 System Design and Implementation

The EPS bearer is a transition of path of defined parameters such as quality, priority and capacity. These parameters can be configured using OPNET Modeller in the attribute of the LTE configuration node and can be assigned to a specific application within a UE. Based on that, the same network architecture that are explained in the previous chapters can be used in a way that split traffic flows among available paths of the different eNodeB's (i.e. the MeNodeB and the SCs).

Complete configuration of the EPS bearer will be well known by the MeNodeB GW. Therefore, each eNodeB will always be able to identify the EPS bearer of arriving data traffic even if that bearer is inactive at the moment of transmission. In this case, each UE has one Non-GBR type EPS bearer (i.e. Default Bearer) which is established at the moment the UE gets attached to the CN. And when additional data traffic with high priority is requested, new GBR bearer can be used to deliver the required data without interrupting other running services.

In the normal situation, when the cell is congested, the resources needed by an active GBR bearer in order to guarantee the QoS required may no longer be available in the cell under

high load conditions. In that case, the EPS bearer may be released if there are not enough resources to sustain the bearer. In addition to that, GBR radio bearers with low priority may be pre-empted and resources to be granted to the EPS bearer with higher priority if necessary.

An architecture scheme of a HetNet network with some management possibilities can use a RAT that best suits the required services and attaches the devices to the core network to get the services they need. So when a UE is running multi services using different EPS Bearers that are each tailored to a specific purpose, the RAN can be tailored to provide specific service and have to be able to differentiate between services, such that when a device wants to carry on services like voice or video calls, while at the same time to continue downloading functions, no service will be impacted by the other service as different RAN nodes will serve different services.

6.6.1 Configuring Application Models

But the compatibility of an application model will likely be dependent on the objective of the study. Dependent on the underlying network, application architecture may differ.

The workflow for building the application is depicted in Figure 6.5 [115]:

1) Configuring individual applications for application modelling on the application configuration node.

2) Configuring the profile in the profile configuration node.

3) Deploying the applications and profiles to network objects.

are that once the profiles and applications are defined, they can be reused across the entire topology.

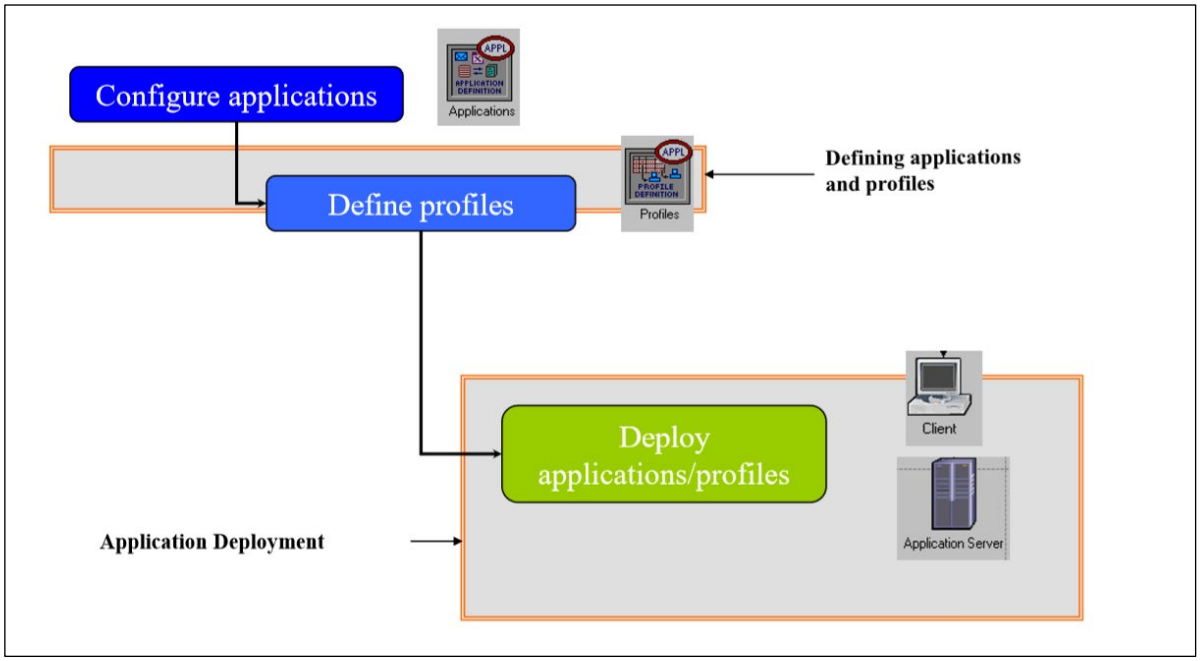


Figure 6.5: Configuring applications workflow in OPNET.

A custom type of application can be created that can be used to model a broad class of applications. This type of application provides attributes that allow to configure various aspects of the application in details including the source and destination in addition to other parameters like the ToS used to assign QoS parameter and priority to the application. A profile contains a list of applications. The profile Definition Object defines all profiles that can be used within a scenario. Application configured within a profile can be executed in different manner, i.e. at the same time, one after another – randomly or in a specific order. In most cases, applications are performed in serial action, however using the applications and profile configuration capabilities, more than one application can be setup to run for the same user at the same time.

After configuring the applications and profiles, these functions need to be deployed in the nodes (e.g. client and server) since they are the source of the traffic. In addition to deploying the application and profiles to the nodes, the supported services must be deployed to the different servers. For example, one server may support one application like E-mail, while another server may support different application like video. Server might support only one application or might support multi applications.

6.6.2 LTE QoS Model

In general, QoS is the concept of providing particular quality guarantee for a specific service type, LTE introduces the concept of EPS bearers for the QoS support. LTE Configuration node in OPNET defines the EPS bearers and their properties, each by a unique name and QoS Class Identifier (QCI) that determines scheduling priority by associate the bearer with QCI. QCI ranges from 1 to 9. In which [3GPP TS 23.203]: {5} > {1-4} > {6-9}.

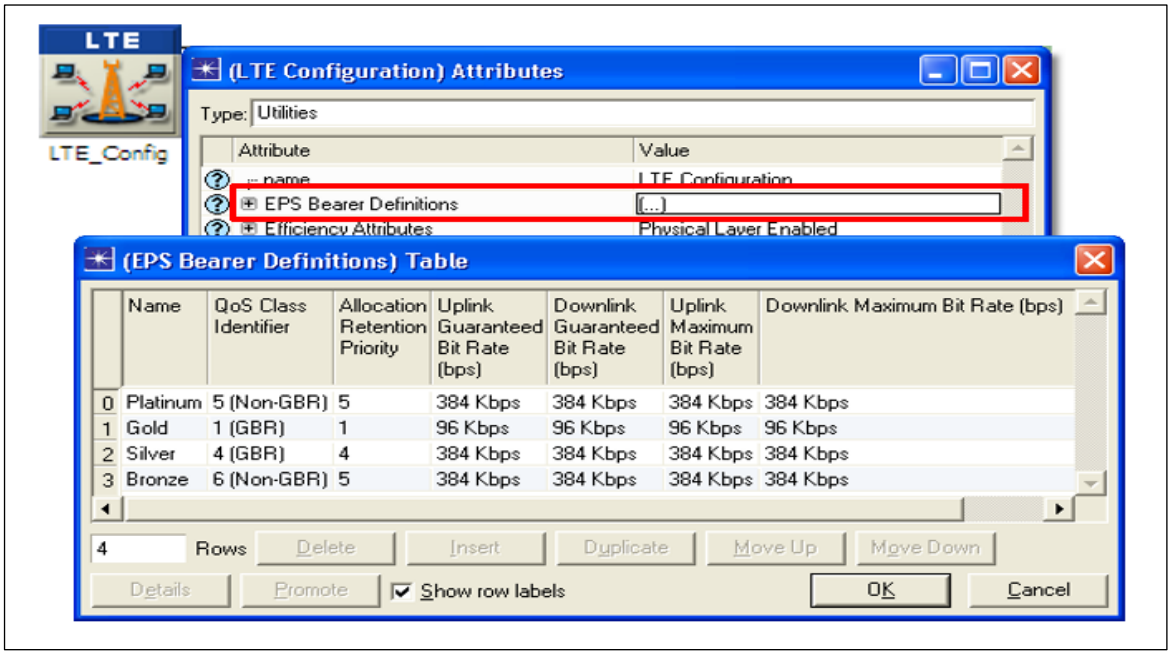


Figure 6.6: Example of EPS bearer definition.

In the UE side, the EPS bearers are deployed by their names with definition of which application are mapped to each bearer. Application's packets that are not mapped to any bearer (or all packets when no bearer defined) are mapped to "Default" bearer with QCI = 9.

Figures 6.6 and 6.7 above shows an example of the QoS configuration for the EPS in the OPNET LTE node models.

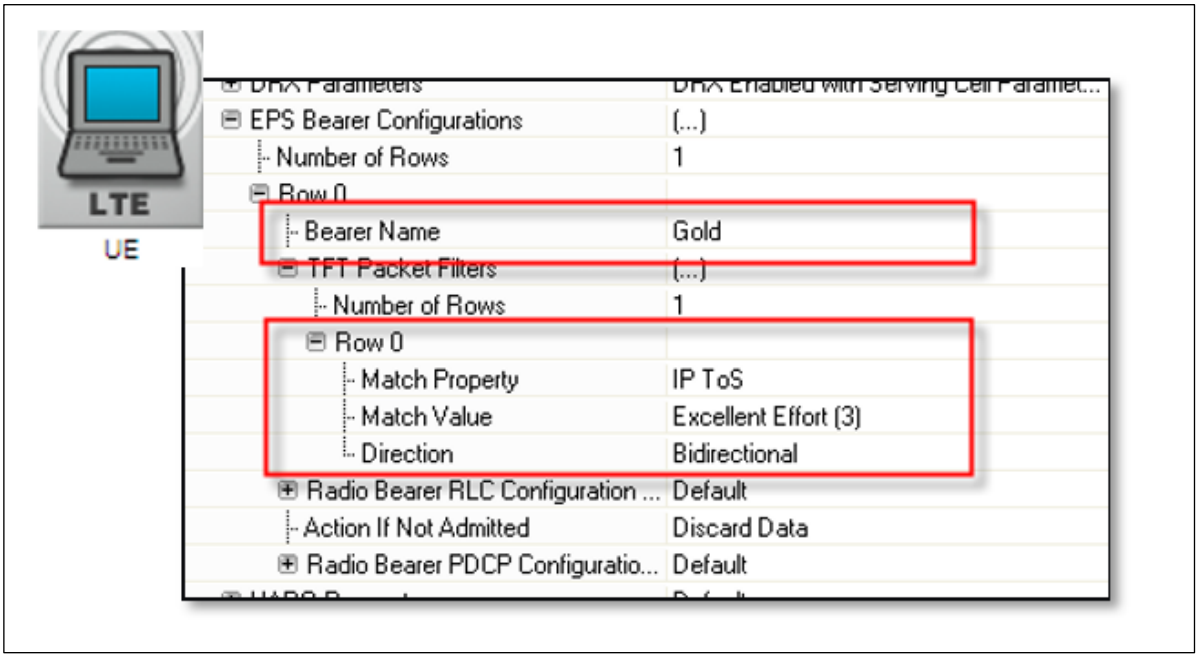


Figure 6.7: Example of EPS Bearer Definition in the UE.

As the current LTE model features supported by this OPNET modeller (Riverbed Modeler/Release 18.7) are based on 3GPP release 8 & 9, that doesn't support features of later releases such as DC and NS. An imitation of the NS in the RAN and CN will be implemented using the differentiated service provided by the capabilities of the applications and profile configuration in addition to the QoS provided by the definition of the EPS bearer. As will be shown in the discussion of the results.

6.7 Performance Evaluation

The evaluation of the performance of the system will be based on the network architecture shown in Figure 6.8 as a base model.

The network is implemented in OPNET using 1 cell eNodeB, 1 server for the traffic, and 3 UEs referred to as (UE_GBR, UE_nonGBR_1 and UE_nonGBR_2), the traffic configuration is as follow:

- UE_GBR is running 2 services:
  - 550 kbps of IP flow carried by an Excellent Effort (GBR) bearer with guaranteed bit rate of 64 kbps (starts around 140 sec until 500 sec).

- 32 kbps of IP flow carried by the Multimedia (GBR) bearer with guaranteed bit rate of 0.5 Mbps (starts around 260 sec until 360 sec).

- UE_nonGBR_1 and UE_nonGBR_2 are running only one service for each of them:
  - 550 kbps of IP flow carried by the Best Effort (non-GBR) bearer (starts around 90 sec until 440 sec).

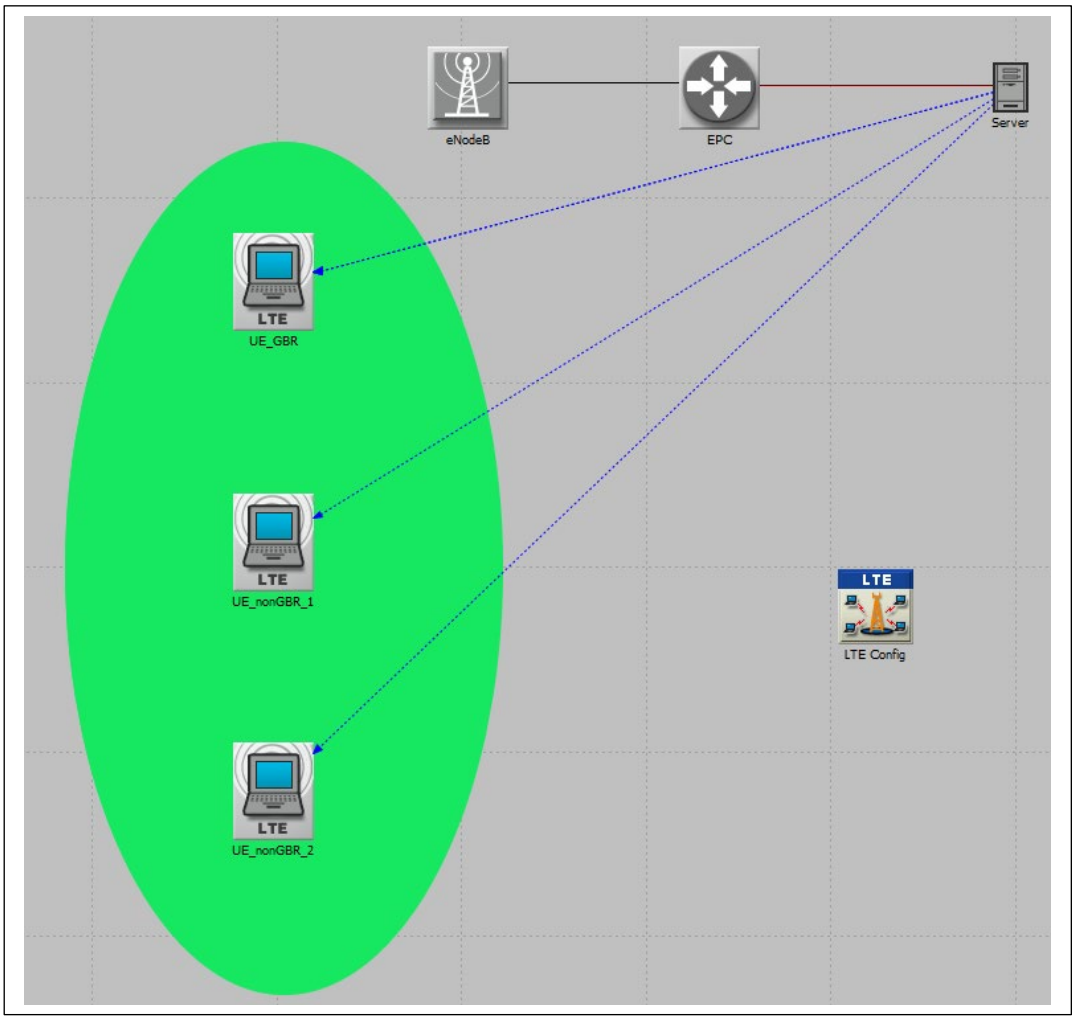


Figure 6.8: Network Topology.

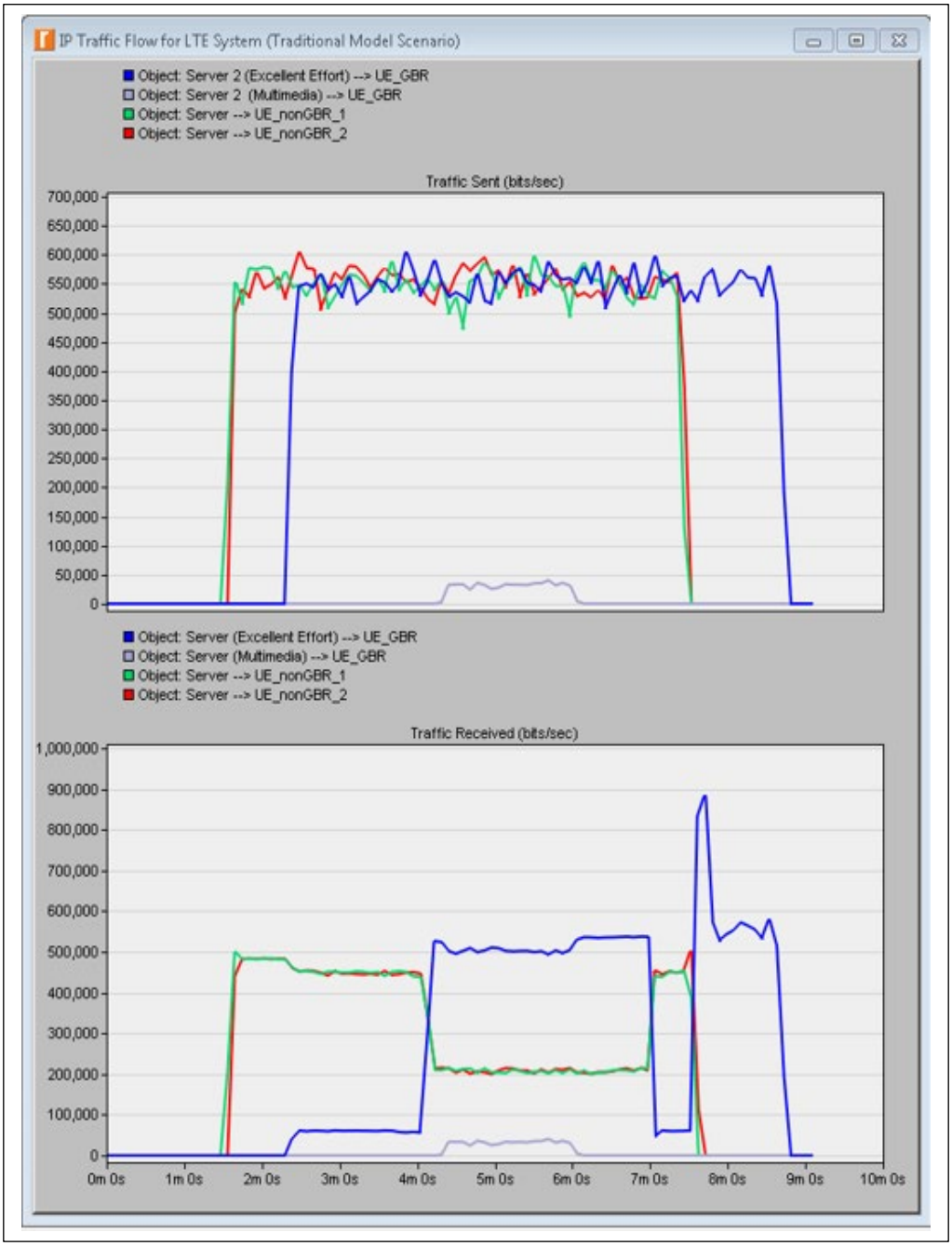


Figure 6.9: IP traffic Flow for LTE System, Traditional Model Scenario.

Figure 6.9 shows the "IP traffic Flow Statistics" panel, first UE_nonGBR_1 and UE_nonGBR_2 start at around ~90 and their demands reach a throughput of ~510 kbps. At around ~140 seconds UE_GBR starts receiving 550 kbps traffic via its 64kbps bearer (Excellent Effort). This activity of UE_GBR reduces the throughput of both non_GBR UE's demands. The GBR bearer reaches its guaranteed bit rate (approximately 60 kbps for the IP flow). At around ~260 seconds, UE_GBR now starts receiving 32 kbps of traffic via its 0.5

Mbps bearer (Multimedia). It can be noticed that this new bearer is merged at the eNodeB with the existing one (i.e. Excellent Effort) as they are part of the same Logical Channel Group of the same eNodeB. For both GBR bearers, there is now a reservation of 564 kbps. With that, the Excellent Effort bearer is able increase the throughput of the IP flow up to 520 kbps. This reduces the throughput of both non-GBR IP flows further down to about ~ 250 kbps each.

When the load for Multimedia GBR demand stops at around ~360 seconds, the bearer is still active (inactivity timer is 60 seconds), so the Excellent Effort bearer uses the entire 564 kbps reserved rate (it slightly increases its throughput). And when Multimedia GBR is torn down due to inactivity at ~420 seconds, the Excellent Effort bearer goes back to a congestion state as it loses the combined guaranteed rate and goes back to a 64kbps guaranteed rate. As a result, the non-GBR bearers increase their throughput back to ~ 480 kbps. In addition, when the traffic goes to zero for all non_GBR bearers, the Excellent Effort bearer is able to use the full bandwidth even if it's above the guaranteed bit rate. Since there is a lot of traffic accumulated in the LTE MAC buffers of this bearer, therefore, a jump in throughput around ~ 455 seconds can be noticed.

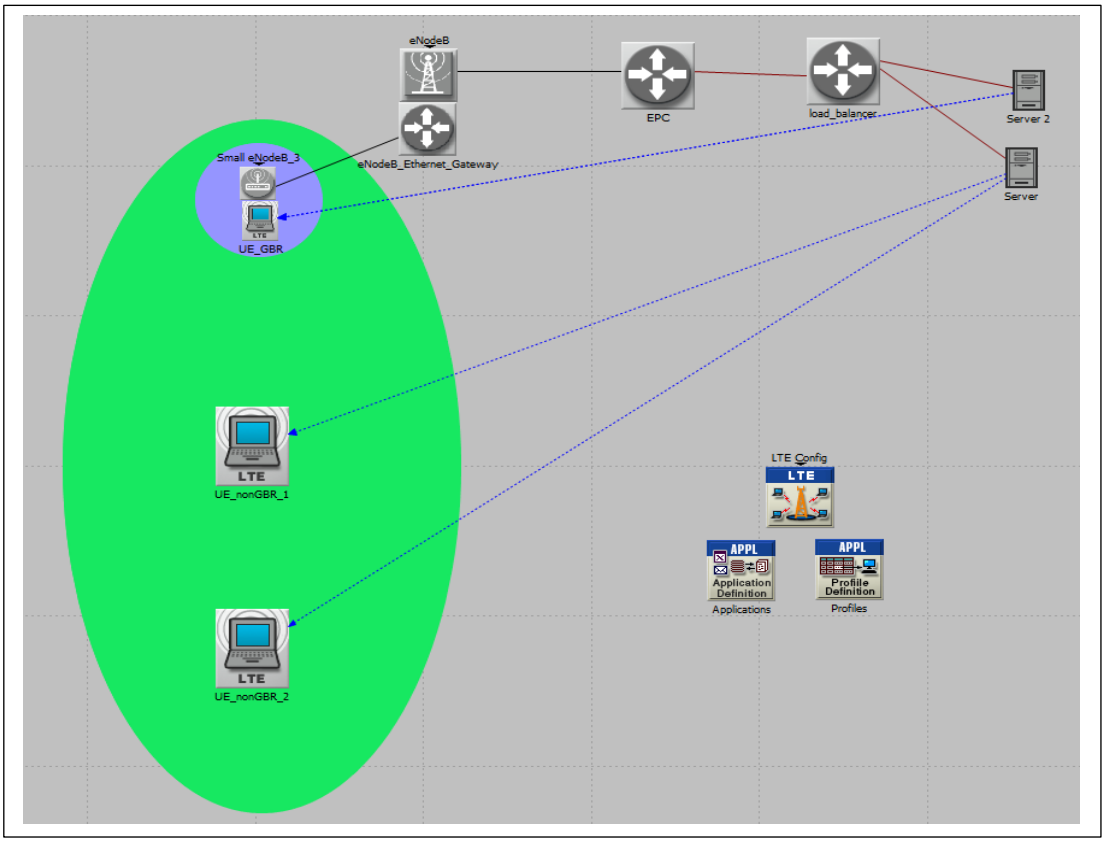


Figure 6.10: Modified Network Topology.

In the second scenario, the same architecture is used with additional modification to the network. In this scenario the network is implemented using 1 cell eNodeB acting as MeNodeB, 1 SC, 2 servers for the load and the same 3 UEs with the same traffic configuration as shown in Figure 6.10.

In this network, the SC and the additional server will be used to handle the traffic of the UE_GBR. The SC will act as RAN slice while the server will act as CN slice. The performance of the network is shown in Figure 6.11.

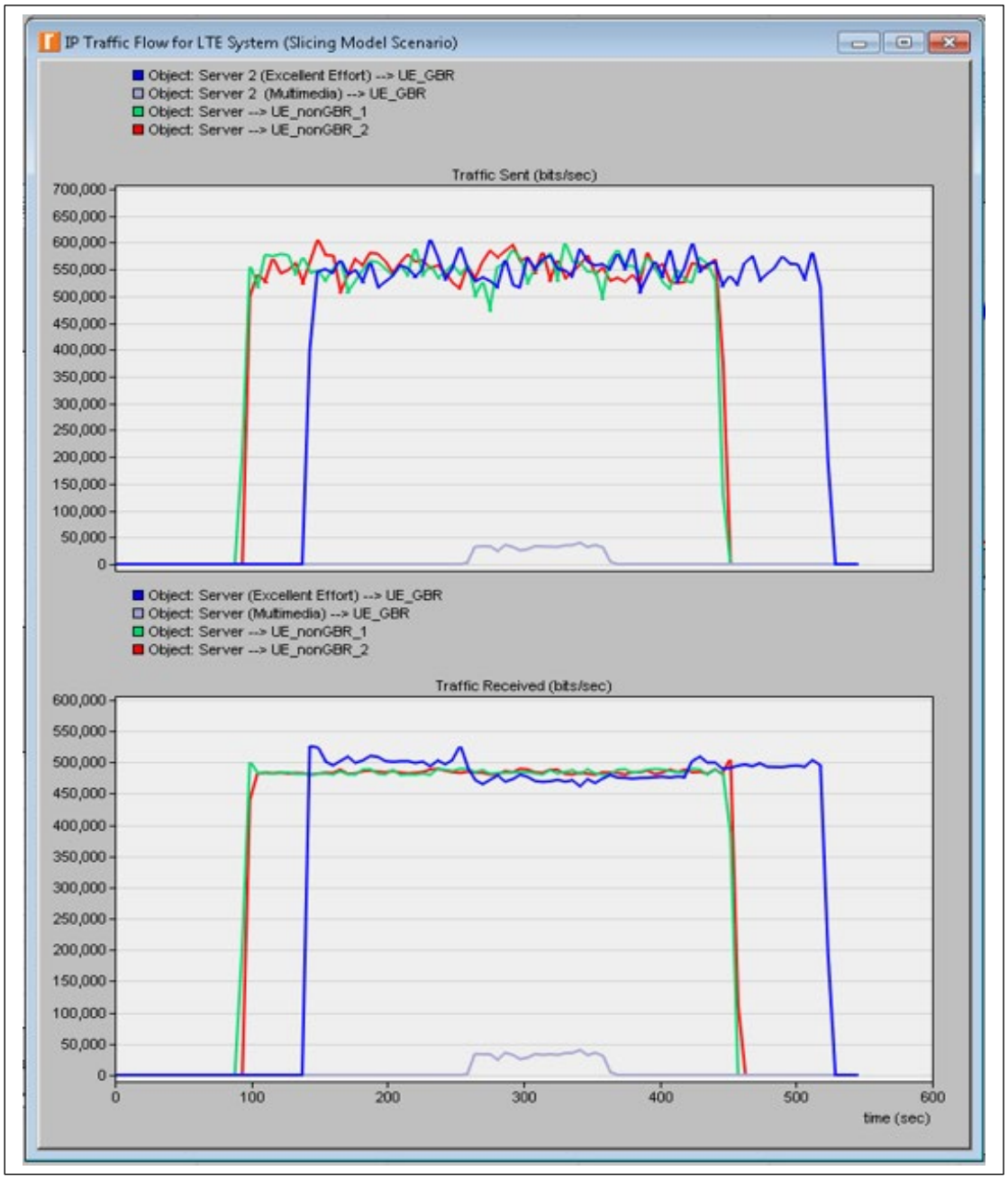


Figure 6.11: IP traffic flow for LTE System, Slicing Model Scenario.

First UE_nonGBR_1 and UE_nonGBR_2 start at around ~90 (same time as the base model network in the first scenario) and their demands reach a throughput of ~510 kbps. At around ~140 seconds UE_GBR starts receiving 550 kbps traffic via its 64kbps bearer (Excellent Effort). In contrast to the first implementation, this activity of UE_GBR has no effect on the throughput of the two non_GBR UE's as it is served from different server using the SC eNodeB. Also, it can be noticed that the GBR bearer uses the maximum bit rate although its guaranteed bit rate is 64 kbps. This is because it is the only bearer in use in the SC and all the resources are available for its traffic. At around ~260 seconds, UE_GBR now starts receiving 32 kbps of traffic via its 0.5 Mbps bearer (Multimedia). Again, and in contrast to the first scenario, the resources for this new bearer is now deducted from the SC resources and it slightly reduces its throughput. This activity still with no effect on the throughput of both non-GBR IP flows as it is running in different channel group.

When the load for Multimedia GBR demand stops at around ~360 seconds, the bearer is still active (same as first scenario), and the resources still available for this bearer till the Multimedia GBR is torn down due to inactivity at ~420 seconds, then the Excellent Effort bearer goes back to benefit from all the resources available in the SC. It can be noticed that in this scenario there is no traffic accumulated in the LTE MAC buffers of this bearer, therefore, there is no jump in throughput of the Excellent Effort bearer as it is able to use the full bandwidth even if it's above the guaranteed bit rate.

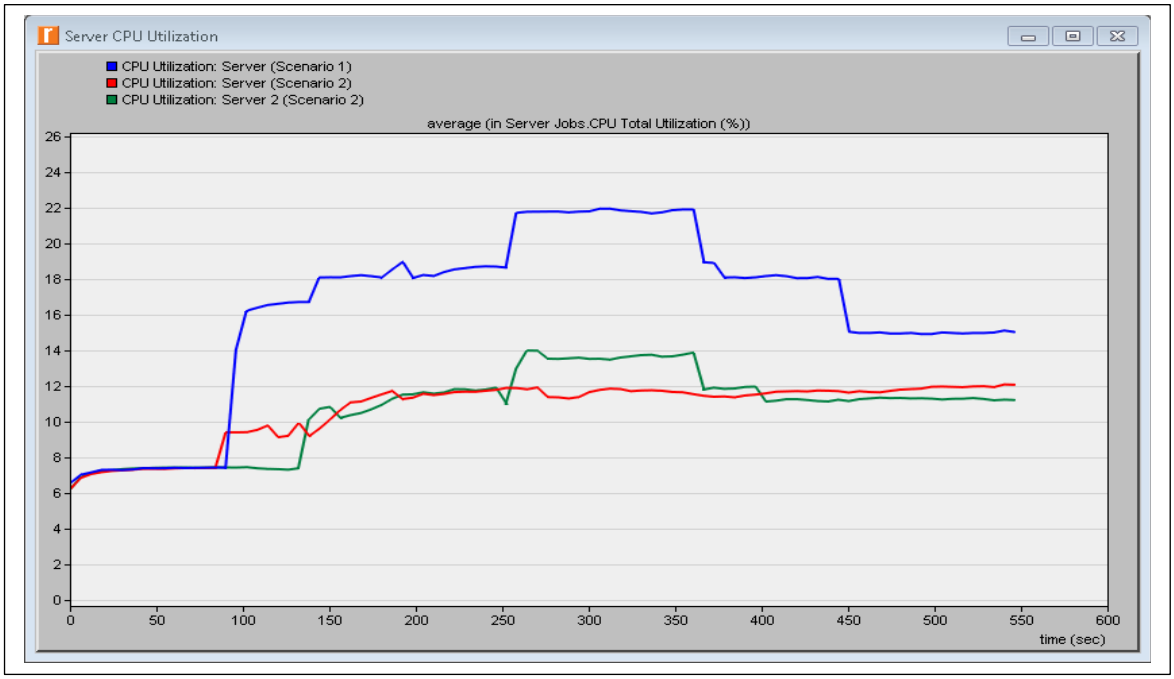


Figure 6.12: Servers CPU utilization.

In regard to the servers used in the CN as an imitation to the CNS, the improvement in the performance can be noticed in term of the % CPU utilization as shown in Figure 6.12.

Same as the effect noticed for the bearers, the CPU Utilization has better performance when the traffic is split between more than one server. The server in the first scenario was handle all the traffic for all the traffic flow of the applications, therefore, the CPU utilization is relatively high for few user (about 20%) then increase to about (22%) when more applications are running. While in the second scenario, the applications are split between two server and the load relatively balanced between them. In this scenario, the first server (i.e. server) is handling the traffic of the two non_GBR UEs, while the second server (i.e. server 2) is handling the traffic for the GBR UE. In this case, the CPU utilization is improved as the load on the server is reduced and both servers have a utilization of about (10%). In addition, it can be noticed that there is a slight increase in the utilization of server 2 at around ~260 seconds, this is because UE_GBR now starts receiving traffic via its Multimedia bearer.

6.8 Summary

The next generation system must be capable of supporting a diverse set of devices, from the very smallest to the most powerful, in an efficient manner. This means that it must be possible to configure the next generation system to match the unique needs of both each device and the applications running on that device. Based on that vision, the ability of the network to provide users with differentiated services based on their needs without any services being interrupted or impacted by other service require drastic change in the architecture of the mobile network.

Network slicing, in its simplest description, is the ability to tailor a set of functions to optimize use of the network for each mobile device. It is true that all of the functionality is needed, but at the same time only the functionality that is needed have to be assembled in a way that optimizes that the ability of the network to deliver the required service in an efficient manner. Furthermore, network slicing providing the ability to the network operators to deploy only the specific functions needed to satisfy their customers' needs. This ability to focus on 'what is needed' reduces investment in unnecessary features, saving operator's money and making them more competitive.

In this respect, an imitation of the NS was introduced in this chapter based on more flexible network infrastructure that can deliver the variety of network performance characteristics targeting various applications of different QoS requirements than the traditional one size-fits-all approach.

Chapter 7

Conclusions and Future Work

7.1 Introduction

The vision of the next generation mobile network will transcend the capabilities of previous systems. The stringent requirement of the next generation mobile system simply imply that the design of the network must be able to fulfil the needs of the most challenging use cases that most of them require high demands for data in addition to higher network availability and very low latency. In the previous generations, communication; in most cases; was between people using their mobile devices in an interactive way. In addition to the classic communication type that include voice calls and broadband connections, communication in the next generation system will take place between automated devices that will come from the Internet of Things (IoT) which will in turn to enable the Machine Type Communication (MTC). And the next system will be required to satisfy the demands of the use cases and applications of such connections which are in most cases unpredicted or unknown.

Therefore, the next generation system, will be built to integrate networking, computing and storage resources into one programmable and unified infrastructure. So, it could be seen as a network of networks, that integrate most of the existing technologies and the diverse types of devices connected to these technologies in a unified system that will allow for an optimized usage of all distributed resources, with powerful programmable capabilities represents the convergence between the IT and telecom systems. In addition, the next generation system will enable prioritization and differentiation of services in a multi tenancy model that will enable the operators and other players to collaborate in new ways.

Therefore, it has become necessary for the next generation mobile system to be implemented in a flexible architecture that is capable to support diverse solutions and to accelerate the service delivery for the new and various services and the diverse of the use cases. As was the aim of this research which will be summarized in the next section followed by some vulnerabilities in the system which are considered as future work.

7.2 Conclusion

The main idea in this research is to make use of the principles of: Heterogeneous Network, Dual Connectivity and Content Delivery.

The proposed design attempts to improve the performance of the system by serving the user from close proximity by bringing the data source to a location nearer to the consumer and provide the services by the means of dedicated services small cell in order to be able to reduce the time (latency) and also reduce the consumed bandwidth. The main structure of the system was introduced and examined for multiple scenarios in Chapter 4. The implementation and simulation of the network was done in OPNET (Riverbed) Modeller.

In Chapter 5, we examine the network for handover performance between the small cells for a UE in dual connectivity. Inspection of service delivery during handover was checked.

Providing differentiated services which is a main concept of the next generation mobile network; (which use the term network slicing), was also imitated in Chapter 6, using different QoS to differentiate between different traffic.

Different results show that using such architecture of centralized management of heterogenous network could be promising solution to deal with the ever increase in new services and their stringent requirement as the services can be provided from close proximity using dedicated channel for this purpose while the user is attached to the network using another channel, in this case the process time used to transfer the data and bandwidth available can be enhanced as the user can be served by different channels used for different services without losing connection to the network.

7.3 Future Work

Although the proposed system architecture provides better user experiences and efficient resource utilization, but when considering the full picture (if the system to be implemented as a whole network), then many concerns must be considered.

The two main concerns to be considered are the power consumption and security of the system. As those are the most important issues arising with development of the next generation system.

Regarding to power consumption, there is a trade-off between the service quality and the consumed power, in general, the better quality of service encourages the longer time of use of the devices leading to fast degradation in battery charge. In addition, the new solution for better quality of services mostly exhaust the battery charge in one way or another. Dual connectivity as an example requires the use of multi channels by the devices using the same battery capacity.

Unfortunately, there are not many solutions to deal with device power consumption in telecom and researches need to be made in this regard. Most of the solutions are traditional solutions which include:

- Expanding the battery cells and/or capacity.
- Using sleep mode and power management solutions.

Methodology that take in consideration renewable energy for mobile devices must be taken in consideration, example of that is the use of solar system, friction or body heat when the device is in the pocket or in the hand.

The other factor which is of high importance and may impact the overall performance of the system is security. Ranging from violating the privacy of the individual users to the whole shutdown of the services and may be the system residing the security risks. Therefore, security is a critical aspect to deal with in any system including the wireless communication system, and especially in the case of small cells, since their form factor and deployment scenarios make them potential targets of various security attacks. Such kind of attacks that are very likely to occur includes, unauthorised changing of small cell location, Booting small

cell with fraudulent software, Man-in-the-middle attacks on small cell first network access and many more [116]. Deployment scenarios of small cells make security a critical aspect to be considered in the development and deployment of small cell networks.

Security must be a central consideration in the next generation mobile system, taking in consideration the various number of devices and application and the huge amount of data to be exchanged in such architecture. In such system, everything is expected to be connected, from surveillance systems, utility meters, and applications including self-driving cars, telemedicine, online gaming which are all in direct contact with people's lives, security and privacy. The demand for high security in the network is increasing and concerns of the personal security and privacy could be a major reason for less adoption of the new system.

References

1. Mobile Network Design and Deployment: How Incumbent Operators Plan for Technology Upgrades and Related Spectrum Needs. Rysavy Research, June 2012.
2. Beyond LTE: Enabling the Mobile Broadband Explosion, LTE and 5G Innovation: Igniting Mobile Broadband. Rysavy Research/4G Americas, August 2015.
3. Amit Kumar, Dr. Yunfei Liu ,Dr. Jyotsna Sengupta, Divya, Evolution of Mobile Wireless Communication Networks 1G to 4G, International Journal of Electronics & Communication Technology, IJECT Vol. 1, Issue 1, December 2010.
4. J. Agrqwal, R. Patel, P. Mor, P. Dubey, and J.M. Keller, "Evolution of mobile communication network: From 1G to 4G", International Journal of Multidisciplinary and Current Research, Vol. 3, No. Nov. /Dec., pp. 1100-1103, 2015.
5. https://www.etsi.org/technologies/mobile/2g
6. Sesia S, Toufik I, Books24x7 I. LTE – The UMTS Long Term Evolution From Theory to Practice, Second Edition. 2nd ed. Chichester, West Sussex, U.K: John Wiley & Sons; 2011.
7. https://www.gsma.com/aboutus/history
8. T. A. Yahiya, "Understanding LTE and its Performance," Springer New York Dordrecht Heidelberg London, Jan. 2011.
9. https://www.etsi.org/technologies/mobile/4g
10. S. Ahmadi, LTE-Advanced A Practical Systems Approach to Understanding the 3GPP LTE Releases 10 and 11 Radio Access Technologies, New York, NY, USA: Academic, 2014.
11. https://www.etsi.org
12. http://www.3gpp.org
13. Mishra, Ajay K. "Fundamentals of Cellular Network Planning and Optimization, 2G/2.5G/3G…Evolution of 4G", John Wiley and Sons, 2004.
14. Recommendation ITU-R M.1457 "Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2000 (IMT-2000)".
15. Introduction to ITU-R WP5D Regional Workshop Kyu-Jin WEE PhD Vice-Chairman, ITU-R WP5D
16. Recommendation ITU-R M.1645 "Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000" (06/2003).

17. ITU global standard for international mobile telecommunications ´IMT-Advanced´: http://www.itu.int/en/ITU-R/study-groups
18. Qualcomm halts UMB project, Reuters, 13 November 2008.
19. ITU-R. *ITU paves way for next-generation 4G mobile technologies; ITU-R IMT-advanced 4G standards to usher new era of mobile broadband communications*. ITU Press Release 21 October 2010.
20. https://www.cnet.com/news/sk-telecom-launches-worlds-first-lte-advanced-network/
21. http://4g5gworld.com/blog/hard-realities-wimax-lte-conversion
22. ATT Best Practices: LTE Performance & Optimization, LTE Call Flows, Ericsson
23. The LTE Network Architecture, A comprehensive tutorial, Alcatel-Lucent
24. http://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Pages/default.aspx
25. 5G: WHAT IS IT? OCTOBER 2014, www.ericsson.com
26. Network architecture for the 5G era, Nokia Solutions and Networks 2015, networks.nokia.com
27. IMT-2020 Tech performance requirements recommendation M.2083
28. SCF197, "URLLC and network slicing in 5G enterprise small cell network", Small Cell Forum, Feb. 2018.
29. 3GPP TR 36.932 V14.0.0 (2017-03)
30. Chris Johnson, "Long Term Evolution in Bullets", 2nd ed., ver. 1. Northampton: England, (2012).
31. Harri Holma, et. Al., "LTE Small Cell Optimization: 3GPP Evolution to Release 13," John Wiley & Sons, (2015).
32. 3GPP TR 36.842 V 11.0.0, June 2012; Technical Specification Group Radio Access Network; Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects, Release 12
33. 3GPP TR 36.872 V 12.1.0, December 2013; Technical Specification Group Radio Access Network; Small Cell enhancements for E-UTRA and E-UTRAN – Physical layer aspects, Release 12
34. 3GPP TR 36.932 V 12.1.0, March 2013; Technical Specification Group Radio Access Network; Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN, Release 12

35. Rohde & Schwarz LTE- Advanced (3GPP Rel.12) Technology Introduction, A. Roessler, J. Schlienz, S. Merkel, M. Kottkamp 6.2014 – 1MA252_2E
36. A. Goldsmith, Wireless communications. Cambridge University Press, 2005.
37. ETSI GS MEC-IEG 004 V1.1.1 (2015-11)
38. SCF014_Edge-Computing-made-simple, Report title: Edge computing made simple Issue date: 04 December 2017 Version: 014-10-01
39. Mobile Edge computing use cases and & deployment options, Juniper white paper, July 2016.
40. 3GPP TR 36.933 V14.0.0 (2017-03)
41. ETSI White Paper No. 24 MEC Deployments in 4G and Evolution Towards 5G, First edition – February 2018.
42. Network Functions Virtualisation - Introductory White Paper, October, 2012: http://portal.etsi.org/NFV/NFV_White_Paper.pdf
43. Network Functions Virtualisation – White Paper on NFV priorities for 5G, February 21st, 2017: https://portal.etsi.org/nfv/nfv_white_paper_5g.pdf
44. Understanding 5G: Perspectives on future technological advancements in mobile, GSMA Intelligence, December 2014
45. Hakan Ohlsen Introduction to IMT-2020, ITU-R Working Party 5D, October 24, 2018.
46. Anritsu Understanding 5G, Rev1 02/16.
47. Parvez et al.: Survey on Low Latency 5G: RAN, Core Network and Caching Solutions, IEEE Communication Surveys and Tutorials, vol. 20, no. 4, Fourth Quarter 2018.
48. A. Gupta and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," IEEE Access, vol. 3, pp. 1206–1232, 2015.
49. C. Amali1, B. Ramachandran, "Enabling Key Technologies and Emerging Research Challenges Ahead of 5G Networks: An Extensive Survey", International Journal in Informatics Visualisation, vol 2, no 3, 2018.
50. M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," IEEE Commun. Surveys Tuts, vol. 18, no. 3, pp. 1617–1655, 3rd Quart, 2016.
51. V.-G. Nguyen, A. Brunstrom, K.-J. Grinnemo, and J. Taheri, "SDN/NFV-based mobile packet core network architectures: A survey," IEEE Commun. Surveys Tuts, vol. 19, no. 3, pp. 1567–1602, 3rd Quart, 2017.

52. T. Taleb et al., “On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration,” IEEE Commun. Surveys Tuts, vol. 19, no. 3, pp. 1657–1681, 3rd Quart, 2017.
53. W. Xia, P. Zhao, Y. Wen, and H. Xie, “A survey on data center networking (DCN): Infrastructure and operations,” IEEE Commun. Surveys Tuts, vol. 19, no. 1, pp. 640–656, 1st Quart, 2017.
54. M. F. Bari et al., “Data center network virtualization: A survey,” IEEE Commun. Surveys Tuts, vol. 15, no. 2, pp. 909–928, 2nd Quart, 2013.
55. A. Ioannou and S. Weber, “A survey of caching policies and forwarding mechanisms in information-centric networking,” IEEE Commun. Surveys Tuts, vol. 18, no. 4, pp. 2847–2886, 4th Quart, 2016.
56. M. Zhang, H. Luo, and H. Zhang, “A survey of caching mechanisms in information-centric networking,” IEEE Commun. Surveys Tuts, vol. 17, no. 3, pp. 1473–1499, 3rd Quart, 2015.
57. M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, “5G backhaul challenges and emerging research directions: A survey,” IEEE Access, vol. 4, pp. 1743–1766, 2016.
58. T. O. Olwal, K. Djouani, and A. M. Kurien, “A survey of resource management toward 5G radio access networks,” IEEE Commun. Surveys Tuts, vol. 18, no. 3, pp. 1656–1686, 3rd Quart, 2016.
59. F. Haider et al., ‘‘Spectral efficiency analysis of mobile Femtocell based cellular systems,’’ in Proc. IEEE ICCT, Jinan, China, Sep. 2011, pp. 347–351.
60. Z. Wang, H. Li, H. Wang, and S. Ci, “Probability weighted based spectral resources allocation algorithm in Hetnet under Cloud-RAN architecture,” in Proc. Int. Conf. Commun. China Workshops, 2013, pp. 88–92.
61. E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, “Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective,” IEEE Wireless Commun., vol. 21, no. 3, pp. 118– 127, Jun. 2014.
62. J. Xu et al., “Cooperative distributed optimization for the hyper-dense small cell deployment,” IEEE Commun. Mag., vol. 52, no. 5, pp. 61–67, May 2014.
63. S. Talwar, D. Choudhury, K. Dimou, E. Aryafar, B. Bangerter, and K. Stewart, “Enabling technologies and architectures for 5G wireless,” in Proc. MTT-S Int. Microw. Symp. (IMS), 2014, pp. 1–4.

64. X. Costa-Perez et al., “5G-crosshaul: An SDN/NFV integrated fronthaul/backhaul transport network architecture,” IEEE Wireless Commun., vol. 24, no. 1, pp. 38–45, Feb. 2017.
65. G. Wang, G. Feng, S. Qin, and R. Wen, “Efficient traffic engineering for 5G core and backhaul networks,” J. Commun. Netw., vol. 19, no. 1, pp. 80–92, Feb. 2017.
66. Shuo Wang, et al. (2016). A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications.
67. A. Ahmed and E. Ahmed, "A survey on mobile edge computing," 2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2016, pp. 1-8.
68. J. Zhang, X. Zhang, and W. Wang, “Cache-enabled software defined heterogeneous networks for green and flexible 5G networks,” IEEE Access, vol. 4, pp. 3591–3604, 2016.
69. Huawei, “5G a technology vision,” White paper, 2013.
70. Huawei, 5G Network Architecture A High-Level Perspective, 2016.
71. K. Chen and R. Duan, “C-RAN: The road towards green RAN,” China Mobile Research Institute, Beijing, White paper, 2011.
72. Content Delivery Network Optimization, Ericsson. www.ericsson.com
73. Ericsson, “5G radio access,” White paper, 2015.
74. Nokia Networks, “Looking ahead to 5G: Building a virtual zero latency gigabit experience,” White paper, 2014.
75. NTT Docomo, “5G radio access: Requirements concepts technologies,” White paper, 2015.
76. Qualcomm Technologies, Inc., “Qualcomm’s 5G vision,” White paper, 2014.
77. https://5g-ppp.eu/
78. https://www.surrey.ac.uk/5gic/
79. http://www.5gforum.org/
80. http://www.ieee-5g.org/
81. E. Dahlman, S. Parkvall and J. Sköld, 4G: LTE/LTE-Advanced for Mobile Broadband. (Second ed.) 2014.
82. Kurose, James, and Keith Ross. Computer Networking: A Top-Down Approach, Global Edition, Pearson Education Limited, 2017. ProQuest ebook Central,

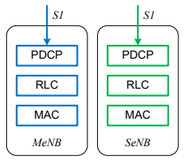
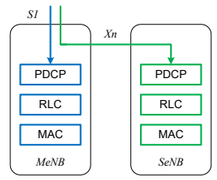
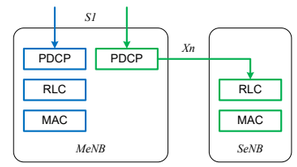
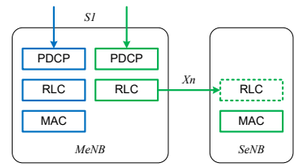
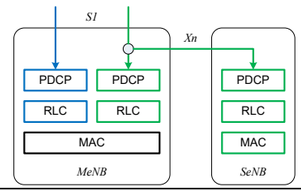
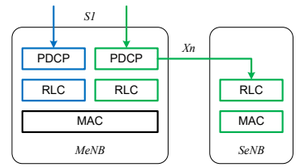
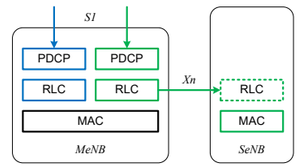
http://ebookcentral.proquest.com/lib/brunelu/detail.action?docID=5187270.
Created from brunelu on 2019-04-26 02:47:43.

83. 3GPP TS 22.261: “Service requirements for the 5G system” V16.0.0 (2017-06)
84. M. Bennis, M. Debbah and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," in Proceedings of the IEEE, vol. 106, no. 10, pp. 1834-1853, Oct. 2018. doi: 10.1109/JPROC.2018.2867029
85. C. A. García-Pérez and P. Merino, “Experimental evaluation of fog computing techniques to reduce latency in LTE networks,” Transactions on Emerging Telecommunications Technologies, vol. 29, no. 4, Article ID e3201, 2018.
86. 3GPP TR 21.905: “Vocabulary for 3GPP Specifications“, V10.3.0 (2011-03).
87. The Cisco Learning Network, Bandwidth vs Speed, Posted by Chandan Singh Takuli in VIP Perspectives on 16-Oct-2015 07:45:39.
88. NGMN Alliance Rachid El Hattachi, Javan Erfanian, “NGMN 5G Initiative,” White paper, 2015.
89. Nokia Solutions and Networks, “5G Use Cases and Requirements,” White paper, 2014.
90. 5G PPP Architecture Working Group: “View on 5G Architecture”, July 2016.
91. Qualcomm, Inc., “LTE Advanced An evolution built for the long-haul”, October 2013.
92. ETSI White Paper No. 23, “Cloud RAN and MEC: A Perfect Pairing”, February 2018.
93. ETSI White Paper No.11, “Mobile Edge Computing A key technology towards 5G”, September 2015.
94. ETSI White Paper No. 28, “MEC in 5G networks”, June 2018.
95. 5G PPP White Paper – Software Networks WG, “Vision on Software Networks and 5G”, January 2017.
96. Cisco 5G cloud-to-client network, “Multi-Access Edge Computing”, October 2018.
97. https://www.akamai.com/us/en/cdn/
98. https://cloud.google.com/cdn/docs/
99. Riverbed Modeler/Release 18.6, https://www.riverbed.com
100. Riverbed Modeler/Release 18.6, chapter 23 “LTE Model Description”, https://www.riverbed.com.
101. A. Dutta, H. Schulzrinne and I. Books24x7, Mobility Protocols and Handover Optimization: Design, Evaluation and Application. (1st ed.) 2014.
102. 3GPP TS 36.323 V8.6.0 Rel. 8.
103. 3GPP TS 36.331 V13.0.0 Re. 13.

104. A. Karandikar et al., “Mobility Management in LTE Heterogeneous Networks”, Springer 2017.
105. A. Prasad, O. Tirkkonen, P. Lunden, O. N. Yilmaz, L. Dalsgaard and C. Wijting, "Energyefficient inter-frequency small cell discovery techniques for LTE-advanced heterogeneous network deployments," Communications Magazine, IEEE, vol. 51, pp. 72-81, 2013.
106. J. Lorca and A. Sierra, "A simple speed estimation algorithm for mobility-aware SON RRM strategies in LTE," in Wireless Days (WD), 2013 IFIP, 2013, pp. 1-3.
107. DIFFSERV—THE SCALABLE END-TO-END QUALITY OF SERVICE MODEL. Cisco Systems, Inc. 2005. www.cisco.com
108. https://ietf.org/standards/
109. V. Feldmann, “Mobile Overtakes Fixed: Implication for Policy and Regulation”, International Telecommunication Union (ITU), 2003.
110. Network Slicing can be a piece of cake, Ericsson, May 2018.
111. 5G Americas, Network Slicing for 5G and Beyond, Nov. 2016.
112. Study on Implications of 5G Deployment on Future Business Models, DotEcon Ltd and Axon Partners Group, March 2018.
113. Network Slicing Use Case Requirements, GSM Association, April 2018.
114. 3GPP TR 23.707 V13.0.0. "Architecture Enhancements for Dedicated Core Networks" (Release 13), Dec 2014.
115. 3GPP TS 22.891 v14.0.0, “Feasibility Study on New Services and Markets Technology Enablers; Stage 1”, Release 14, Sep. 2016.
116. Riverbed Modeller/Release 18.6, Chapter 2 “Application”
117. SCF 099, “Security for small cells”, Small Cell Forum, June. 2014.

Appendix A

Table A: Comparison table of user plane architecture alternatives

Alternative	Alternative 1A	Alternative 2A	Alternative 2C	Alternative2D	Alternative 3A	Alternative 3C	Alternative 3D
Overview	S1-U terminates in SeNB + independent PDCPs (no bearer split). 	S1-U terminates in MeNB + no bearer split in MeNB + independent PDCP at SeNB. 	S1-U terminates in MeNB + no bearer split in MeNB + independent RLC at SeNB. 	S1-U terminates in MeNB + no bearer split in MeNB + master-slave RLC for SeNB bearers. 	S1-U terminates in MeNB + bearer split in MeNB + independent PDCPs for split bearers. 	S1-U terminates in MeNB + bearer split in MeNB + independent RLCs for split bearers. 	S1-U terminates in MeNB + bearer split in MeNB + master-slave RLCs for split bearers. 
Description	This option terminates the currently defined air-interface U-plane protocol stack completely per bearer at a given eNB, and is tailored to realize transmission of one EPS bearer by one node. The transmission of different bearers may still happen simultaneously from the MeNB and a SeNB	This option terminates the currently defined air-interface U-plane protocol stack completely per bearer at a given eNB. With the help of an additional function at MeNB, it supports routing of a bearer towards SeNB. The transmission of different bearers may still happen simultaneously from the MeNB and a SeNB.	This option assumes that S1-U terminates in MeNB with the PDCP layer residing in the MeNB always. The transmission of different bearers may still happen simultaneously from the MeNB and a SeNB.	This option assumes that S1-U terminates in MeNB with the PDCP layer and part of the RLC layer residing in the MeNB always. While requiring only one RLC entity in the UE for the EPS bearer, on the network side the RLC functionality is distributed between the nodes involved, with a "slave RLC" operating in the SeNB. In downlink, the slave RLC takes care of the delay-critical RLC operation needed at the SeNB: it receives from the master RLC at the MeNB readily built RLC PDUs (with Sequence Number already assigned by the master) that the master has assigned for transmission by the slave, and transmits them to the UE. The custom-fitting of these PDUs into the grants from the MAC scheduler is achieved by re-using the currently defined re-segmentation mechanism.	This option terminates the currently defined air-interface U-plane protocol stack completely per bearer at a given eNB. With the help of an additional layer (or functional entity), it supports splitting of a single EPS bearer for transmission by MeNB and SeNB.	This option assumes that S1-U terminates in MeNB with the PDCP layer residing in the MeNB always. There is a separate and independent RLC bearer (SAP above RLC), also at UE side, per eNB configured to deliver PDCP PDUs of the PDCP bearer (SAP above PDCP), terminated at the MeNB.	This option assumes that S1-U terminates in MeNB with the PDCP layer and part of the RLC layer residing in the MeNB. While requiring only one RLC entity in the UE for the EPS bearer, on the network side the RLC functionality is distributed between the nodes involved, with a "slave RLC" operating in the SeNB. In downlink, the slave RLC takes care of the delay-critical RLC operation needed at the SeNB: it receives from the master RLC at the MeNB readily built RLC PDUs (with Sequence Number already assigned by the master) that the master has assigned for transmission by the slave, and transmits them to the UE. The custom-fitting of these PDUs into the grants from the MAC scheduler is achieved by re-using the currently defined re-segmentation mechanism.
Implementation & Specs Impacts	**Alternative 1A**	**Alternative 2A**	**Alternative 2C**	**Alternative2D**	**Alternative 3A**	**Alternative 3C**	**Alternative 3D**
Xn interface	Must introduce signalling to support the interaction between MeNB and SeNB on RRM, power control… ☹	Must introduce signalling to support the interaction between MeNB and SeNB on RRM, power control… ☹ Transfer of PDCP SDUs 😐	Must introduce signalling to support the interaction between MeNB and SeNB on RRM, power control… ☹ Transfer of PDCP PDUs 😐 Opens interaction between RLC and PDCP ☹ Flow control required ☹	Must introduce signalling to support the interaction between MeNB and SeNB on RRM, power control… ☹ Transfer of RLC PDUs ☹ Transfer of RLC Status Reports ☹ Flow control required ☹	Must introduce signalling to support the interaction between MeNB and SeNB on RRM, power control… ☹ Transfer of PDCP SDUs 😐 Flow control required ☹	Must introduce signalling to support the interaction between MeNB and SeNB on RRM, power control… ☹ Transfer of PDCP PDUs 😐 Opens interaction between RLC and PDCP ☹ Flow control required ☹	Must introduce signalling to support the interaction between MeNB and SeNB on RRM, power control… ☹ Transfer of RLC PDUs ☹ Transfer of RLC Status Reports (unless MAC is made aware to transmit them over MeNB always) 😐 Flow control required ☹
Above PDCP	Nothing required in the MeNB to handle packets of SeNB ☺	Routing function needed in MeNB 😐	Nothing required ☺	Nothing required ☺	New layer required to split the bearers, route the packets towards the appropriate eNB, and reorder packets ☹	Nothing required ☺	Nothing required ☺
PDCP Security	Security impacts due to ciphering being required in both MeNB and SeNB ☹	Security impacts due to ciphering being required in both MeNB and SeNB ☹	No security impacts with ciphering taking place in MeNB only ☺	No security impacts with ciphering taking place in MeNB only ☺	Security impacts due to ciphering being required in both MeNB and SeNB ☹	No security impacts with ciphering taking place in MeNB only ☺	No security impacts with ciphering taking place in MeNB only ☺
PDCP TX eNB	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ Two PDCP entities required for split bearer ☹	RLC bearer selection required ☹ One PDCP entity always, even for split bearer ☺	No impact ☺ One PDCP entity always, even for split bearer ☺
PDCP RX UE	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	One PDCP entity per bearer always ☺	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ Two PDCP entities required for split bearer ☹	One PDCP entity always, even for split bearer ☺ PDCP to become responsible for reordering data from two parallel RLC bearers ☹ PDCP buffer for reordering need to be dimensioned to cope with Xn latencies ☹	No impact ☺ One PDCP entity always, even for split bearer ☺
PDCP TX UE	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ Two PDCP entities required for split bearer ☹	RLC bearer selection required ☹ One PDCP entity always, even for split bearer ☺	No impact ☺ One PDCP entity always, even for split bearer ☺

Appendix A

Table A: Comparison table of user plane architecture alternatives

PDCP RX eNB	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	One PDCP entity per bearer ☺	No impact ☺ One PDCP entity per bearer ☺	No impact ☺ Two PDCP entities required for split bearer ☹	One PDCP entity per bearer ☺ PDCP to become responsible for reordering data from two parallel RLC bearers ☹ PDCP buffer for reordering need to be dimensioned to cope with Xn latencies ☹	No impact ☺ One PDCP entity per bearer ☺
RLC	No impact ☺ One RLC entity per bearer ☺	No impact ☺ One RLC entity per bearer ☺	No impact ☺ One RLC entity per bearer ☺	One Master + one Slave RLC entity per bearer on eNB side 😐 Backhaul delay becomes part of RLC RTT and extension of RLC SN space may be required, which in turn may increase buffering requirements ☹ application with RLC UM requires adoption of UMD PDU Segment ☹ Re-segmentation header (SO - 2bytes) always added to SeNB RLC PDUs during segmentation, therefore increasing overhead (also in status PDUs) ☹ care needs to be taken at SeNB that RLC Status PDU cannot be segmented ☹	No impact ☺ Two RLC entities required for split bearer ☹	No impact ☺ Two RLC entities required for split bearer ☹	One Master + one Slave RLC entity per bearer on eNB side 😐 Backhaul delay becomes part of RLC RTT and extension of RLC SN space may be required, which in turn may increase buffering requirements ☹ eNB selection needed at master RLC transmitter ☹ application with RLC UM requires adoption of UMD PDU Segment ☹ Re-segmentation header (SO - 2bytes) always added to SeNB RLC PDUs during segmentation, therefore increasing overhead (also in status PDUs) ☹ care needs to be taken at SeNB that RLC Status PDU cannot be segmented ☹
MAC TX eNB	Prioritisation of traffic between MeNB and SeNB must be done through signalling over S1/X2 ☹ One MAC entity per eNB 😐	Prioritisation of traffic between MeNB and SeNB must be done through signalling over Xn ☹ One MAC entity per eNB 😐	Prioritisation of traffic between MeNB and SeNB must be done through signalling over Xn ☹ One MAC entity per eNB 😐	Prioritisation of traffic between MeNB and SeNB must be done through signalling over Xn ☹ One MAC entity per eNB 😐	Prioritisation of traffic between MeNB and SeNB must be done through signalling over Xn ☹ One MAC entity per eNB and two MAC entities per split bearer ☹	Prioritisation of traffic between MeNB and SeNB must be done through signalling over Xn ☹ One MAC entity per eNB and two MAC entities per split bearer ☹	Prioritisation of traffic between MeNB and SeNB must be done through signalling over Xn ☹ One MAC entity per eNB and two MAC entities per split bearer ☹
MAC RX UE	No impact ☺ One MAC entity per eNB ☹	No impact ☺ One MAC entity per eNB ☹	No impact ☺ One MAC entity per eNB ☹	No impact ☺ One MAC entity per eNB ☹	No impact ☺ One MAC entity per eNB and two MAC entities per split bearer ☹	No impact ☺ One MAC entity per eNB and two MAC entities per split bearer ☹	No impact ☺ One MAC entity per eNB and two MAC entities per split bearer ☹
MAC TX UE RLC PDUs	Radio resource allocation is restricted to the eNB where the Radio Bearer terminates and the UE needs to be aware of the correspondence → need mapping rules between grants and corresponding RLC entities related to the eNB issuing the grants ☹	Radio resource allocation is restricted to the eNB where the Radio Bearer terminates and the UE needs to be aware of the correspondence → need mapping rules between grants and corresponding RLC entities related to the eNB issuing the grants ☹	Radio resource allocation is restricted to the eNB where the Radio Bearer terminates and the UE needs to be aware of the correspondence → need mapping rules between grants and corresponding RLC entities related to the eNB issuing the grants ☹	Radio resource allocation is restricted to the eNB where the Radio Bearer terminates and the UE needs to be aware of the correspondence → need mapping rules between grants and corresponding RLC entities related to the eNB issuing the grants ☹	Radio resource allocation is restricted to the eNB where the Radio Bearer terminates and the UE needs to be aware of the correspondence → need mapping rules between grants and corresponding RLC entities related to the eNB issuing the grants ☹ For split bearers, although the UE is free to select an eNB where to send data first, RLC retransmissions must always target the same eNB → one MAC entity per eNB + one to one mapping between MAC entity and RLC entities ☹ Tocken bucket algorithm in LCP needs to take the transmission over different eNBs into account ☹	For bearers contained in MeNB, radio resource allocation is restricted to the MeNB and the UE needs to be aware of the correspondence → need mapping rules between MeNB grants and corresponding RLC entities For split bearers, although the UE is free to select an eNB where to send data first, RLC retransmissions must always target the same eNB → one MAC entity per eNB + one to one mapping between MAC entity and RLC entities ☹ Tocken bucket algorithm in LCP needs to take the transmission over different eNBs into account ☹	For bearers contained in MeNB, radio resource allocation is restricted to the MeNB and the UE needs to be aware of the correspondence → need mapping rules between MeNB grants and corresponding RLC entities ☹ For split bearers, the UE is always free to select an eNB where to send data (first transmissions and re-transmissions) ☺ Tocken bucket algorithm in LCP needs to take the transmission over different eNBs into account ☹
MAC TX UE BSR, PHR, SR	BSR, PHR and SR either have to be sent separately or we assume some interaction between MeNB and SeNB via S1/X2 ☹	BSR, PHR and SR either have to be sent separately or we assume some interaction between MeNB and SeNB via Xn ☹	BSR, PHR and SR either have to be sent separately or we assume some interaction between MeNB and SeNB via Xn ☹	BSR, PHR and SR either have to be sent separately or we assume some interaction between MeNB and SeNB via Xn ☹	BSR, PHR and SR either have to be sent separately or we assume some interaction between MeNB and SeNB via Xn ☹	BSR, PHR and SR either have to be sent separately or we assume some interaction between MeNB and SeNB via Xn ☹	BSR, PHR and SR either have to be sent separately or we assume some interaction between MeNB and SeNB via Xn ☹
MAC RX eNB	One MAC entity per eNB 😐 If BSR, PHR and SR are not sent separately, the eNB needs to forward the information to the other node ☹	One MAC entity per eNB 😐 If BSR, PHR and SR are not sent separately, the eNB needs to forward the information to the other node ☹	One MAC entity per eNB 😐 If BSR, PHR and SR are not sent separately, the eNB needs to forward the information to the other node ☹	One MAC entity per eNB 😐 If BSR, PHR and SR are not sent separately, the eNB needs to forward the information to the other node ☹	One MAC entity per eNB 😐 If BSR, PHR and SR are not sent separately, the eNB needs to forward the information to the other node ☹	One MAC entity per eNB 😐 If BSR, PHR and SR are not sent separately, the eNB needs to forward the information to the other node ☹	One MAC entity per eNB 😐 If BSR, PHR and SR are not sent separately, the eNB needs to forward the information to the other node ☹
PHY Aspects	Separate PUCCH, CQI and power control loops required ☹	Separate PUCCH, CQI and power control loops required ☹	Separate PUCCH, CQI and power control loops required ☹	Separate PUCCH, CQI and power control loops required ☹	Separate PUCCH, CQI and power control loops required ☹	Separate PUCCH, CQI and power control loops required ☹	Separate PUCCH, CQI and power control loops required ☹

Appendix A

Table A: Comparison table of user plane architecture alternatives

Performance	Alternative 1A	Alternative 2A	Atlernative 2C	Alternative2D	Alternative 3A	Alternative 3C	Atlernative 3D
SeNB Mobility	Not hidden to CN ☹ Forwarding between SeNBs ☹ Interruption visible due to MeNB unable to support SeNB bearer ☹	Not hidden to CN (unless security can solely be handled by MeNB) ☹ Forwarding between SeNBs (unless MeNB buffers the data) ☹ Interruption visible due to MeNB unable to support SeNB bearer ☹	Hidden to CN ☺ No forwarding between SeNBs ☺ Interruption visible due to MeNB unable to support SeNB bearer ☹	Hidden to CN ☺ No forwarding between SeNBs ☺ Interruption visible due to MeNB unable to support SeNB bearer ☹	Not hidden to CN (unless security can solely be handled by MeNB) ☹ Forwarding between SeNBs (unless MeNB buffers the data) ☹ Interruption limited thanks to the ability of the MeNB to transmit data for the split bearers ☺	Hidden to CN ☺ No forwarding between SeNBs ☺ Interruption limited thanks to the ability of the MeNB to transmit data for the split bearers ☺	Hidden to CN ☺ No forwarding between SeNBs ☺ Interruption limited thanks to the ability of the MeNB to transmit data for the split bearers ☺
Utilisation of radio resources across MeNB and SeNB	Not possible for the same bearer, requires at least two DRBs for having user plane traffics in MeNB and SeNB ☹	Not possible for the same bearer, requires at least two DRBs for having user plane traffics in MeNB and SeNB ☹	Not possible for the same bearer, requires at least two DRBs for having user plane traffics in MeNB and SeNB ☹	Not possible for the same bearer, requires at least two DRBs for having user plane traffics in MeNB and SeNB ☹	Possible for the same bearer ☺	Possible for the same bearer ☺	Possible for the same bearer ☺
Dynamic Offload	Need to involve MME, very static ☹	Controlled by MeNB, static as it can only use one eNB and reconfig required each time 😐	Controlled by MeNB, static as it can only use one eNB and reconfig required each time 😐	Controlled by MeNB, static as it can only use one eNB and reconfig required each time 😐	Controlled by MeNB, can be dynamic as long SCells are setup ☺	Controlled by MeNB, can be dynamic as long SCells are setup ☺	Controlled by MeNB, can be dynamic as long SCells are setup ☺
MeNB processing of SeNB traffic	None ☺	Limited to routing, with or without buffering above PDCP 😐	Down to PDCP level followed by routing ☹	Down to RLC level followed by routing ☹	Limited to routing, with buffering and reordering above PDCP ☹	Down to PDCP level followed by routing ☹	Down to RLC level followed by routing ☹
Backhaul Traffic	Low requirements ☺	Need to carry traffic offloaded to SeNB ☹ If the router is above MeNB, the MeNB-router link wil see the SeNB traffic twice ☹	Need to carry traffic offloaded to SeNB ☹ If the router is above MeNB, the MeNB-router link wil see the SeNB traffic twice (once as PDCP SDUs, and once as PDCP PDUs) ☹	Need to carry traffic offloaded to SeNB ☹ If the router is above MeNB, the MeNB-router link wil see the SeNB traffic twice (once as PDCP SDUs and once as RLC PDUs) ☹	Need to carry traffic offloaded to SeNB ☹ If the router is above MeNB, the MeNB-router link wil see part of the traffic twice (the offloaded part) ☹	Need to carry traffic offloaded to SeNB ☹ If the router is above MeNB, the MeNB-router link wil see part of the traffic twice (the offloaded part, once as PDCP SDUs, and once as PDCP PDUs) ☹	Need to carry traffic offloaded to SeNB ☹ If the router is above MeNB, the MeNB-router link wil see part of the traffic twice (the offloaded part, once as PDCP SDUs and once as RLC PDUs) ☹
Buffering Requirements	Full termination of EPS bearer at SeNB offloads PDCP buffering from MeNB ☺	Full termination of EPS bearer at SeNB offloads PDCP buffering from MeNB ☺	Partially overlapping buffering needed at both MeNB and SeNB 😐	Partially overlapping buffering needed at both MeNB and SeNB 😐 RLC SN space may require extension, because Xn delay becomes part of ARQ RTT 😐	Bearer splitting implies increased reordering-buffering requirement, either to UE or MeNB ☹	Bearer splitting implies increased reordering-buffering requirement, either to UE or MeNB ☹	Bearer splitting implies increased reordering-buffering requirement, either to UE or MeNB ☹ RLC SN space may require extension, because Xn delay becomes part of ARQ RTT 😐