# Neural Network Design for Intelligent Mobile Network Optimisation

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

of Doctor of Philosophy

by

**Raid Sakat**

Department of Electronic and Computer Engineering

College of Engineering, Design and Physical Sciences

Brunel University London

United Kingdom

August 2019

*Dedicated to the memory of my Father*

# Abstract

The mobile networks users' demands for data services are increasing exponentially, this is due to two main factors: the first is the evolution of smart phones and their application, and the second is the emerging new technologies for internet of things, smart cities…etc, which keeps pumping more data into the network; 'though most of the data routed in the current mobile network is non-live data'. This increasing of demands arise the necessity for the mobile network operators to keep improving their network to satisfy it, this improvement takes place via adding hardware or increasing the resources or a combination of both. The radio resources are strictly limited due to spectrum licensing and availability, therefore efficient spectrum utilization is a major goal to be achieved for both network operators and developers. Simultaneous and multiple channel access,and adding more cells to the network are ways used to increase the data exchanged between the network nodes. The current 4G mobile system is based on the Orthogonal Frequency Division Multiple Access (OFDMA) for accessing the medium and the intercell interference degrades the link quality at the cell edge, with the introduction of heterogeneity concept to the LTE in Release 10 of the 3GPP the handover process became even more complex. To mitigate the intercell interference at the cell edge, coordinated multipoint and carrier aggregation techniques are utilized for dual connectivity.

This work is focused on designing and proposing enhancing features to improve network performance and sustainability, these features comprises of distributing small cells for data only transmission, handover schemes performance evaluation at cell edge with dual connectivity, and Artificial Intelligence technology for balancing and prediction.

In the proposed model design the data and controls of the Small eNodeB (SeNodeB) are processed at the network edge using a Mobile Edge Computing (MEC) server and the SeNodeBs are used to boost services provided to the users, also the concept of caching data has been investigated, the caching units where implemented in different network levels. The proposed system and resource management are simulated using the OPNET modeller and evaluated through multiple scenarios with and without full load, the UE is reconfigured to accommodate dual connectivity and have two separate connections for uplink and downlink, while maintaining connection to the Macro cell via uplink, the downlink is dedicated for small cells when content is requested from the cache. The results clearly show that the proposed system can decrease the latency while the total throughput delivered by the network has highly improved when SeNodeBs are deployed in the system, rising throughput will incur the rise of overall capacity which leads to better services being provided to the users or more users to join and benefit from the network.

Handover improvement is also considered in this work, with the help of two Artificial Intelligence (AI) entities better handover performance are achieved. Balanced load over the SeNodeBs results in less frequent handover, the proposed load balancer is based on artificial neural network clustering model with self-organizing map as a hidden layer, it’s trained to forecast the network condition and learn to reduce the number of handovers especially for the UEs at the cell edge by performing only necessary ones, and avoid handovers to the Macro cell for the downlink direction. The examined handovers concern the downlinks when routing non live video stored at the small cell’s cache, and a reduction in the frequent handovers was achieved when running the balancer.

Keep revolving in the handover orbit, another way to preserve and utilize network resources is by predicting the handovers before they occur, and allocate the required data in the target SeNodeB, the predictor entity in the proposed system architecture combines the features of Radial Basis Function Neural Network and neural network time series tool to create and update prediction list from the system’s collected data and learn to predict the next SeNodeB to associate with. The prediction entity is simulated using MATLAB, and the results shows that the system was able to deliver up to 92% correct predictions for handovers which led to overall throughput improvement of 75%.

# Publications Based on this Research

1) 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, doi.10.1007/978-3-030-01177-2-43

2) Sakat R., Saadoon R., Abbod M. (2020) “Load Balancing Using Neural Networks Approach for Assisted Content Delivery in Heterogeneous Network”. In: Bi Y., Bhatia R., Kapoor S. (eds) Intelligent Systems and Applications. IntelliSys 2019. Advances in Intelligent Systems and Computing, vol 1038. Springer, Cham

3) 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. doi: 10.1109/FGCT.2017.8103399

4) 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.

5) 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”. BMSB 2019, IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. Jeju, South Korea.

# 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.

*Raid SAKAT*

*August 2019*

*London, UK*

# Acknowledgements

First of all, I thank Almighty God whose wisdom enlightened my mind and whose inspiration alone helped me navigate the complex pathways of applied science and engineering in developing this thesis. All glory to His Name.

During the research and writing, I have had the opportunity of working with extraordinary colleagues to whom I am greatly indebted. I am especially grateful to Dr Maysam Abbod, my thesis supervisor for his thought-provoking discussions, stimulating insights, suggestions for improvements and continuous guidance during this journey.

Finally, I am extremely grateful to my family, especially my late father, my mother, my wife, my children, my brother and sisters, all of whom have patiently supported and encouraged me, gladly surrendering quality family time to my academic work. I dedicate this thesis to all these wonderful people whose unwavering love and support has sustained me over these years.

# Contents

**List of Figures** .... x

**CHAPTER 1** .... 1

**INTRODUCTION** .... 1

**1.1 Introduction** .... 1

**1.2 Motivation** .... 1

**1.3 Aim and Objectives** .... 2

**1.4 Contributions to Knowledge** .... 2

**1.5 Research Methodology** .... 3

**1.6 Thesis Outline** .... 4

**CHAPTER 2** .... 6

**MOBILE NETWORKS THE BEGINNING, EVOLUTION, AND TRENDS** .... 6

**2.1 Background** .... 6

2.1.1 History and Evolution of Mobile Networks .... 7

**2.2 Long Term Evolution - LTE** .... 9

2.2.1 System Architecture Evolution .... 11

2.2.2 Evolved Packet Core .... 11

2.2.3 Interfaces in LTE .... 13

2.2.4 Channels .... 14

2.2.5 Scheduling .... 16

2.2.6 LTE Bearers .... 17

**2.3 LTE Advanced** .... 18

2.3.1 Carrier Aggregation .... 19

2.3.2 Coordinated Multipoint .... 20

2.3.3 Heterogeneous LTE Network .... 20

2.3.4 Cell Types in HetNet .... 21

2.3.4 Small Cells and Dual Connectivity .... 23

2.3.6 Mobile Edge Computing .... 24

**2.4 State-of-the-Arts Work** .... 25

**2.5 Summary** .... 28

**CHAPTER 3** .... 29

**ARTIFICIAL INTELLIGENCE TECHNOLOGIES AND IMPLEMENTATIONS** .... 29

**3.1 Introduction** .... 29

**3.2 Artificial Neural Networks** .... 33

3.2.1 Transfer functions ........ 34
3.2.2 Network Architecture ........ 36
3.2.3 Training the Artificial Neural Networks ........ 39
**3.3 Deep Learning for Mobile Networks** ........ 43
3.3.1 Related Work ........ 44
**3.4 Summary** ........ 48
**CHAPTER 4** ........ 49
**SMALL CELL DEPLOYMENT FOR CONTENT DELIVERY ARCHITECTURE AND CHALLENGES** ........ 49
**4.1 Introduction** ........ 49
**4.2 Architecture** ........ 50
4.2.1 Control Plane Architecture with SeNB ........ 50
4.2.2 User Plane Architecture ........ 51
4.2.3 SeNodeB Commissioning ........ 52
**4.3 Model Design** ........ 53
4.3.1 Cache Modelling ........ 58
4.3.2 Gain Analysis with Dual Connectivity ........ 60
**4.4 Model Simulation** ........ 61
4.4.1 LTE Implementation in OPNET ........ 62
4.4.2 Dual Connectivity for UE Terminal ........ 63
**4.5 Performance Evaluation** ........ 67
**4.6 Results Discussion and Analysis** ........ 69
**4.7 Summary** ........ 74
**CHAPTER 5** ........ 75
**LOAD BALANCING USING NEURAL NETWORKS APPROACH FOR ASSISTED CONTENT DELIVERY IN HETROGENOUS NETWORK** ........ 75
**5.1 Introduction** ........ 75
**5.2 Load Balancing** ........ 77
5.2.1 Reporting Handover Parameters ........ 78
5.2.2. Handover Stages ........ 79
**5.3 Load Balancing Algorithms** ........ 80
5.3.1 Related Work ........ 81
**5.4 Architecture** ........ 84
**5.4.1 Physical Layer Measurements** ........ 85
**5.4.2 Load Balancing Procedure** ........ 86

**5.4.3 Load Balancer Design** ........ 89

**5.5 Simulation Setup** ........ 90

**5.6 Simulation Results and Analysis** ........ 91

**5.7 Summary** ........ 98

**CHAPTER 6 NEURAL NETWORK FOR MOBILITY PREDICTIONS AND RESOURCE ALLOCATION IN LTE** ........ 100

**6.1 Introduction** ........ 100

**6.2 Related Work** ........ 102

**6.3 Handover Prediction** ........ 104

6.3.1 Prediction Parameters ........ 105

6.3.2 Prediction Steps ........ 106

**6.4 Architecture** ........ 108

6.4.1 Predictor Architecture ........ 108

**6.5 Simulation Setup** ........ 109

**6.6 Summary** ........ 115

**CHAPTER 7** ........ 117

**CONCLUSION AND FUTURE WORK** ........ 117

**7.1 Conclusions** ........ 117

**7.2 Future Work** ........ 119

**REFERENCES** ........ 120

# List of Figures

Figure 2.1: Mobile phones everywhere ........................................................................…..6

Figure 2.3: System Architecture evolution ……….…………………………………….....11

Figure 2.4: EPC block diagram……………………………………………………………12

Figure 2.5: Channel Bandwidth and transmission bandwidth…………………………....14

Figure 2.6: LTE Channels…………………………………………………………………15

Figure 2.7: Scheduler in LTE……………………………………………………………...16

Figure 2.8: LTE Bearers…………………………………………………………………...17

Figure 2.9: LTE technologies……………………………………………..………..19

Figure 2.10: Aggregated channel bandwidth…………………….……………………..19

Figure 2.11: CoMP concept………………………………………………………………20

Figure 2.12: Heterogeneous Network (HetNet)…………………………………….……..21

Figure 2.13: HetNet base stations…………………………………………………..……...22

Figure 2.14: Dual Connectivity scenarios…………………………………..….………24

Figure 2.15: MEC deployment in mobile network……………………………….…………25

Figure 3.1: AI classification…………………………………………………………..……31

Figure 3.2: Learning in AI………………………………………………………….……..32

Figure 3.3: Biological neuron………………………………………………….……..…33

Figure 3.4: Artificial neuron………………………...………………………………..……34

Figure 3.5 Hard limit transfer function…………………………………….…..……….35

Figure 3.6 Linear Transfer function……………………………………………………..……35

Figure 3.7: Log Sigmoid Transfer function…………………………………………..……36

Figure 3.8 S Neurons layer………………………………………………………….……..37

Figure 3.9 Three layers feedforward network…………………………………….……….38

Figure 3.10 Delayer & integrator………………………………………………….……...39

Figure 3.11: DNN architecture…………………………………………………………….……41

Figure 3.12: Reinforcement and imitation learning block diagram………………………....44

Figure 4.1: Control plane architecture…………………………………………………….50

Figure 4.2: UP 1A Bearer level split at SGW……………………………………………..51

Figure 4.3: UP 3C Packet level split at MeNodeB………………………………………..52

Figure 4.4: SeNodeB addition procedure…………………………………………………53

Figure 4.5: Scenario #2 of Release 12…………………………………………………….53

Figure 4.6: Proposed system diagram……………………………………………………..55

Figure 4.7: Content delivery procedure flowchart………………………………………..57

Figure 4.8: Caching deployment in LTE network………………………………………..60

Figure 4.9: LTE network using OPNET…………………………………………………62

Figure 4.10: UE node level and equivalent protocol stack……………………………….63

Figure 4.11: UE layers supporting DC (UP 3C)…………………………………………..63

Figure 4.12: Modified UE for DC…………………………………………………………64

Figure 4.13: Attach procedure in process domain………………………………………..66

Figure 4.14: eNodeB node model…………………………………………………………67

Figure 4.15: Basic System model…………………………………………………………68

Figure 4.16: End to End delay…………………………………………………………….70

Figure 4.17: Network throughput………………………………………………………....70

Figure 4.18: End to End delay……………………………………………………………..71

Figure 4.19: Network throughput…………………………………………………………72

Figure 4.20: E2E delay (multiple scenarios)………………………………………………73

Figure 5.1: Global Mobile Data Traffic, 2016 to 2021……………………………………...77

Figure 5.2: Unbalanced network ……………………………………………………….....78

Figure 5.3: HO triggering values selection………………………………………………..79

Figure 5.4: Handover Stages……………………………………………………………....81

Figure 5.5: Proposed EPS architecture…………………………………………………….85

Figure 5.6: Load evaluation and handover decision…………………………………...........88

Figure 5.7: Clustering neural networks……………………………………………………89

Figure 5.8: Network throughput (scenario 1)……………………………………………...92

Figure 5.9: end2end delay (scenario 1)……………………………………………………92

Figure 5.10: Associated eNodeB (scenario 1)……………………………………………..93

Figure 5.11: network throughput (scenario 2)……………………………………………..94

Figure 5.12: end2end delay (scenario 2)…………………………………………………..94

Figure 5.13: associated eNodeB (scenario 2)……………………………………………...95

Figure 5.14: end2end delay (scenario 3)…………………………………………………..96

Figure 5.15: Associated eNodeB (scenario 3)……………………………………………..96

Figure 5.16: end2end delay (Balanced network)…………………………………………..97

Figure 5.17: Associated eNodeB (Balanced Network)…………………………………….98

Figure 6.1: UE mobility path……………………………………………………………..101

Figure 6.2: Prediction parameters collection……………………………………………..107

Figure 6.3: Neural time series (NARX)…………………………………………………..108

Figure 6.4: Mobility path for UE………………………………………………………...110

Figure 6.5: Associated eNodeB sequence…………………………………………..…111

Figure 6.6: Neural time series performance……………………………………….…112

Figure 6.7: Neural Network training regression…………………………………………113

Figure 6.8: Obtained associated eNodeB………………………………………………...114

Figure 6.9: Predicted associated eNodeB………………………………………………...114

Figure 6.10: Throughput evaluation……………………………………………………..115

# List of Tables

Table 2.1: Channel Bandwidths for LTE……………………………………………………14

Table 4.1: Simulation parameters…………………………………………….………...67

Table 4.2: Corresponding Network parameters…………………………….…….……….…68

Table 5.1: Simulation setup……………………………………….…..……………….…..91

Table 5.2: Simulation setup………………………………………………..…………....91

Table 6.1: Simulation parameters……………………………………………….….……109

Table 6.2: Training results……………………………………………………...………..112

# List of Abbreviations

| | |
|---|---|
| **3GPP** | 3rd Generation Partnership Project |
| **AI** | Artificial Intelligence |
| **AP** | Access Point |
| **ANN** | Artificial Neural Network |
| **AODV** | Ad Hoc On-Demand Distance Vector |
| **AR** | Augmented Reality |
| **AS** | Access Stratum |
| **BTS** | Base Transceiver Station |
| **BPTT** | Back Propagation Through Time |
| **CA** | Carrier Aggregation |
| **CRAN** | Cloud Radio Access Network |
| **CHNN** | Continues Hopfield Neural Network |
| **CoMP** | Coordinated Multipoint |
| **DC** | Dual Connectivity |
| **DNN** | Deep Neural Network |
| **D2D** | Device To Device |
| **EPC** | Evolved Packet Core |
| **EPS** | Evolved Packet System |
| **E-RAB** | Evolved Radio Access Bearer |
| **E2E** | End To End |
| **EUTRAN** | Evolved Universal Terrestrial Radio Access Network |
| **GBR** | Guaranteed Bit Rate |
| **GPRS** | General Packet Radio Service |
| **GSM** | Global System for Mobile communication |
| **HetNet** | Heterogeneous Network |
| **HSDPA** | High-Speed Downlink Packet Access |
| **HSS** | Home Subscriber Server |
| **HSUPA** | High Speed Uplink Packet Access |
| **HO** | Hand Over |
| **HOM** | Hand Over Margin |
| **HPC** | Handover Parameter Control |
| **HTTP** | Hypertext Transfer Protocol |
| **ICMP** | Internet Control Message Protocol |
| **IMS** | IP-based Multimedia Services |
| **IP** | Internet Protocol |
| **IoT** | Internet Of Things |
| **IT** | Information Technology |
| **Kpi** | Key Performance Indicator |
| **LAN** | Local Area Network |
| **LRT** | Least Residual Time |
| **LSTM** | Long Short-Term Memory |
| **LTE** | Long Term Evolution |
| **LTE-A** | Long Term Evolution Advanced |
| **LTE-LAA** | Long Term Evolution License Assisted Access |
| **MAC** | Medium Access Control |
| **MANET** | Mobile Ad-hoc Network |
| **MEC** | Mobile Edge Computing |

| | |
|---|---|
| **MeNodeB** | Macro Evolved Node B |
| **MIMO** | Multiple Input Multiple Output |
| **ML** | Machine Learning |
| **MME** | Mobility Management Entity |
| **MOS** | Mean Opinion Score |
| **NARX** | Nonlinear Autoregressive Exogenous model |
| **NAS** | Non-Access Stratum |
| **NB-IoT** | Narrow Band Internet Of Things |
| **PCC** | Policy and Charging Control |
| **PCEF** | Policy and Charging Enforcement Function |
| **PCRF** | Policy and Charging Rule Function |
| **PDCP** | Packet Data Convergence Protocol |
| **PDN** | Packet Data Network |
| **PEGW** | Packet Data Network Edge Gateway |
| **PGW** | Packet Data Network Gateway |
| **PSTN** | Public Switched Telephone Network |
| **QoS** | Quality of Service |
| **RAN** | Radio Access Network |
| **RAp** | Radio Access Point |
| **RB** | Resource Block |
| **RBFNN** | Radial Basis Function Neural Network |
| **ResNet** | Residual Network |
| **RL** | Reinforcement learning |
| **RLC** | Radio Link Control |
| **RNN** | Recurrent Neural Network |
| **RRC** | Radio Resource Control |
| **RRE** | Remote Radio Equipment |
| **RRM** | Radio Resource Management |
| **RSRP** | Reference Signal Received Power |
| **RSRQ** | Reference Signal Received Quality |
| **RSSI** | Received Signal Strength Indicator |
| **RTRL** | Real-time recurrent learning |
| **SeNodeB** | Small Evolved Node B |
| **SC** | Small Cell |
| **SCG** | Small Cell Group |
| **SDR** | Software Defined Router |
| **SGD** | Stochastic Gradient Descent |
| **SGW** | Serving Gateway |
| **SINR** | Signal to Noise and Interference Ratio |
| **SOM** | Self-Organized Map |
| **SNR** | Signal To Noise Ratio |
| **TCP** | Transmission Control Protocol |
| **TTI** | Transmission Time Interval |
| **UDP** | User Datagram Protocol |
| **UE** | User Equipment |
| **UML** | Unified Modeling Language |
| **UMTS** | Universal Mobile Telecommunications System |
| **VoIP** | Voice over IP |
| **VoLTE** | Voice over LTE |
| **VR** | Virtual Reality |
| **WLAN** | wireless local area network |

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

Mobile devices and service provided by cellular networks has been part of every day’s life. The mobile phones are no longer exclusively used to make voice calls; with the emerge of the smart phones and devices, digital revolution, Nano technology, internet popularity and artificial intelligence algorithms, and many other factors boosted the capabilities of mobile devices, made them more powerful, and increased the ways they can be used for. The number of mobile phones and services is growing; the time spent by people using their mobile phones or cellular devices is increasing rapidly. More time on phone means more data is exchanged through the network.  Thus, the network operators keep developing their networks and services to quench the thirst for the bandwidth.  due to the limited availability of spectrum as well as its high cost, the network operators tends to fully exploit what’s available by multiple access techniques, and by using better technologies.

This research’s target is to study and improve the performance of the existing fourth generation mobile network (4G) or the Long Term Evolution (LTE).

## 1.2 Motivation

Since it was first declared, Artificial Intelligence (AI) has been a part of so many industries such as finance, image recognition, healthcare, self-driving cars and telecommunications. AI is a powerful technology that can transform the existing industry by learning the ways to increase the system efficiency and helping the users by advising faster decision making for business development and financial accounting.

In telecommunications, wireless connectivity is a challenge, and maintaining this connectivity for users exchanging data at high rates makes it more complex, especially if those users are moving. Embedding AI technologies in wireless networks can help make intelligent decisions to perfectly manage the network traffic and resources which will eventually result in better performance and increase efficiency and reduce the cost.

Moreover; AI plays an essential role in providing the network operators with data driven solutions based on data collected from the network users, these solutions help providing better services to cover customers' demands. On the other hand; AI can provide the operators with network driven solutions which help improving the network performance efficiently and costly effective in terms of speed and capacity.

## 1.3 Aim and Objectives

The aim of this work is to provide solutions that improve the performance of the current LTE networks and beyond:

i) Extending the network features and capabilities

ii) Enhancing the design and architecture circuitry and software

The objectives of this research can be described as follows:

i) Building and understanding an LTE model using OPNET modeller with small cells deployed for non-live data only transmission.

ii) Running the LTE model with dual connectivity activated for both uplink and downlink, then examining the network behaviour in terms of handover and its effect on the network throughput and delay.

iii) Observing the influence of adding cache units in different points of the network then study the outcomes and compare the benefits of this process.

iv) Adding AI entities to the evolved packet core (EPC), to balance the load distributed over the small cells. Reducing the number of redundant handovers saves network resources, avoids adding unnecessary delays and extends the battery life of the User Equipment (UE).

v) Adding another AI block based on neural networks to the EPC architecture to predict the next effective evolved base station eNodeB for the UE to associate with. The prediction procedure helps the network anticipate the handover prior triggering allowing resources transfer from the source to the target eNodeB.

## 1.4 Contributions to Knowledge

This research discusses the current LTE network features and studies the ways to consolidating these features as well as proposing the addition of new technologies to the LTE architecture, which leads to better Quality of Service (QoS), the contributions of this work, are as follows:

- Building an LTE model with modified architecture that extends the network coverage by adding small cells for non-live content delivery via the downlink, however the principal of small cells and different radio access technologies are proposed, none of the previous researches in this field reinforced the small cells with caching capabilities for storing the contents, non-live content may contain any frequent data (like saved videos or information ) , or temporarily in high demand data (such as operating system updates).
- Handover management for load balancing using neural networks. Maintaining good connectivity in both uplink for signalling and downlink for data transfer especially at the edge of the cell can be improved with suitable handover policy. The neural network is trained using network and user data to learn the best time to trigger handover (HO) and avoid the consequences of any unnecessary HO. The AI entity is embedded in the EPC architecture, but it can be also distributed within the network framework.
- Better resource scheduling can be achieved with prediction of network's behaviour, the UEs mobility can be predicted by using neural networks time series tool with learning algorithms so that all the resources and bandwidth allocated for this user can be transferred to the target SeNodeB before triggering the handover if the target SeNodeB is on the handover list, the implementation of this design had increased the throughput by 50% and was able to effectively decrease the end to end delay.

## 1.5 Research Methodology

The aim of this research is to enhance the performance of the LTE network, thus a model built as a basic framework using Riverbed 18.7 modeller is proposed (formerly known as OPNET). The first step towards achieving the goal was introducing the dual connectivity concept in the system; the Riverbed 18.7 supports only Release 8 of the 3GPP standards. The node models include the full protocol stack from the physical layer up to the application layer represented by modules for the Access Stratum (AS) and Non-Access Stratum (NAS) protocols while the layers representing the U-plane protocol stack are embedded as process modules inside them. The modified model has the same layers of the original node model, except for the Long Term Evolution Dual Connectivity (LTE's DC) which has limited functionality compared to the original one, as it has only the PDCP and RLC layers.

The mobile edge computing (MEC) server was also added at the cell edge to share part of the processing, manage caching and for better traffic handling; this will reduce the signalling overhead between the nodes and the EPC.

Two solutions for better handover initiation were proposed, the solutions includes EPC architecture modifications , AI block that consists of two neural network based objects is embedded in the EPC, this block is used for load balancing , handover decision making and mobility prediction, the neural networks are modelled using MATLAB, the neural networks had two different training schemes ,supervised and non-supervised. The data is collected from the network for training, and the outcomes where fed back to the network for updating the parameters.

## 1.6 Thesis Outline

This thesis comprises totally of seven chapters, starting with the introduction chapter verifying the reasons behind and the motivations for doing this work, brief explanation of the methodology and the contributions, the rest of the thesis is organised as follows:

- Chapter Two, is a survey on the mobile network with literature review about the LTE and its evolutions. Sophisticated definitions of the parts concerned in building up the basic model to provide good comprehension about the network architecture and its components.
- Chapter Three, reviews the beginning of artificial intelligence, and its role in changing the face of industry. It listed some of the basic building units of the artificial neuron and how it is simulated to learn and react as the human brain.
- Chapter Four, this chapter explains the deployment of small cells in an LTE framework simulated using OPNET modeller and how dual connectivity is implemented to mitigate challenges in the coordinated multipoint CoMP and carrier aggregation CA, The second part explains the benefits of adding caching units to enhance content delivery process for the network users.
- Chapter Five, the modified LTE architecture built and tested as explained in chapter four, with added AI blocks is proposed to overcome the problems occurring due to unbalanced load routing and boosting the delivered throughput. The load balancing technique utilizes clustering artificial neural network application for content delivery mechanism in Heterogeneous LTE mobile network. The proposed network design

demonstrated efficient impact on the network performance in terms of power saving and handling data size increase.

- Chapter Six presents a neural network prediction system as an add-on in the LTE EPC, the predictor is able to learn some of the patterns demonstrated by users moving within the network and can then predict the future behaviour of these users and the next associated eNodeB. These predictions are utilized for better radio resource allocation and scheduling.
- Chapter Seven verifies the conclusions extracted from this work, and the future plans for enhancing and improving what has been proposed to help saving the environment and reduce the power consumption and hence the pollution.

# CHAPTER 2

# MOBILE NETWORKS THE BEGINNING, EVOLUTION, AND TRENDS

### Briefing

This chapter comprises three main sections the first illustrates the generations of mobile networks, their beginning, evolution, trends and technologies migrating from predecessor to successor generation, the second explains the architecture and features of the Long Term Evolution (LTE) which is the base of the model built and simulated in this work and the benefits of distributing small cells in LTE mobile networks. The third part views some of the related work and researches in this field.

## 2.1 Background

In the last 100 years the science of telecommunication witnessed huge development alongside with electronic and digital devices revolution. Wireless communication and mobile phones were dreams, now it has not only become true but roaming, video calls, virtual reality VR, Augmented reality AR are now possible because of the powerful and efficient mobile networks. Communication development is governed by the market needs in both business and personal aspects, and since communication became an important part of everyday life; therefore, simplifying the way accessing the communication is the main target of this development.

Figure 2.1: Mobile phones everywhere

### 2.1.1 History and Evolution of Mobile Networks

A cellular network is defined as a wireless network deployed over a geographic territory divided into smaller areas known as cells, provided that each cell is served by one or more transceiver(s). To avoid the interference caused by neighbouring cells while providing the service, all cells must use its own set of frequencies and it must be different from the sets used in all cells that are adjacent to it [1]. Cellular networks are usually deployed over a large geographical area to provide users with wireless connectivity based services; the users' equipment varies from mobile phones used for making phone calls to IoT sensors /actuators proposed for smart cities. The philosophy of cellular networks had grown steadily, starting from first generation, evolving to the second and third generations, and the currently in use the 4G, trending towards the next generation 5G, as shown in figure 2.2

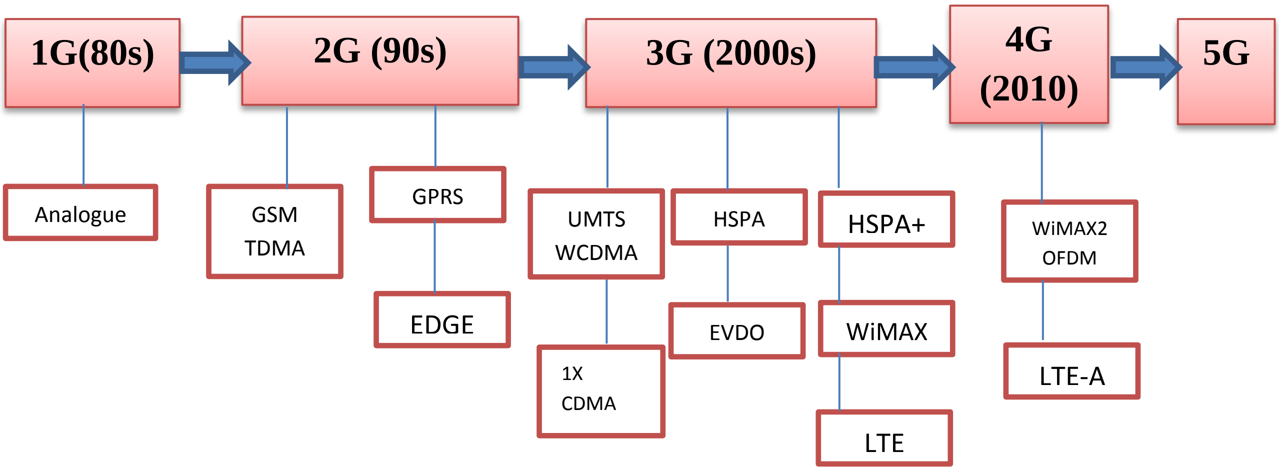


Figure 2.2: Cellular networks evolution history.

The first Generation Mobile Communication System 1G was the first ever wireless commercial phone launched with equivalent to Data rate or speed of 2 kbps ,the utilized technology was analogue with FDMA Multiplexing, the only service this system was able to provide was Voice call which was the main purpose behind launching the 1G, the main backbone Network which the system relied on was the traditional and almost everywhere connected Public Switch Telephone Network (PSTN) , and the radio Frequency used was 800 to 900 MHz, however and due to not having common standards or protocols this system had many drawbacks  such limited coverage, low capacity, equipment cost, and lack of handover [1][2].

The digital revolution and its impact on humans life had its effect on the communication systems which became more available and popular, the introduction of new standards, communication protocols, powerful and supercomputers supported the conversion of the existing analogue systems to more spectrum efficient and less expensive digital system, this led to the birth of second generation mobile communication system 2G in early 1990's with Data capacity of 10 kbps, it was the first digital mobile technology utilizing TDMA Modulation, it provided both voice and data services , and it also relied on the existing PSTN as its backbone network, frequency used 850 MHz to 1900 MHz, due to commissioning of global standards for mobile systems it expanded quickly and dominated the mobile communication market for over a decade and deserved the name Global System for Mobile Communication (GSM) [3][4].

Though the GSM spread widely and many of the services that were introduced during its development had migrated to and still in use with 3G and 4G systems, the need for better coverage and higher data rate was growing, which made the system unable to cope with these demands ,the 2.5G General Packet Radio System (GPRS) was the evolution resultant of 2G due to the increasing demand for data communication so with its modified architecture the Data capacity was increased up to 144 kbps which made it possible to browse the internet and send/receive emails as well as sending Multimedia Messaging Service (MMS). After the millennium and the launch of Wireless Access Protocol (WAP) as a connectivity protocol for the mobile network the Enhanced Data Rate for GSM Evolution (EDGE) considered as 2.75G was the improved version of GPRS it boosted the Data capacity up to 237 kbps, kept the same Multiple Access (TDMA, CDMA) and Frequency (850 MHz to 1900 MHz) [5][6].
As the computers capabilities and their ways of connectivity evolved and the mobile applications that enabled the users to continuously access the world wide web from their handsets, the need for higher capacity and motivated the developers to launch the Universal Terrestrial Mobile System (UMTS) in Europe and the CDMA2000 (1x and EV-DO) in the USA known as 3G- Third Generation Mobile Communication System came with increased Data capacity of up to 384 Kbps, main difference from 2G was the use of Broadband/IP to provide High speed voice, data and video service, Frequency bands changed to 1.6 to 2.5 GHz. 3.5G with higher Data capacity up to 2 Mbps, and new standards High Speed Downlink Packet Access (HSDPA) and High Speed Uplink Packet Access (HSUPA) are introduced to provide High Speed Voice/Data/Video services, it relies the existing GSM as a main backbone network, Frequency: 1.6 to 2.5 GHz. The 3.9G was introduced with higher Data

capacity: 30 Mbps and new standard (EVDO), same access technique (CDMA) is used to provide high speed internet/ Multi-media service, Frequency band: 1.6 to 2.5 GHz [7].

To a certain point the data rate provided by the 3G network was sufficient to fulfil the requirements for many applications, but when the smartphones where launched and 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 growth in the size of cellular networks affects the performance of the 3G due to the rapid increase in the data amount handled by system, and major enhancements in the Quality of Service (QoS) for the network users became a priority need. This in turn led to the need to introduce a new architecture for the new generation that could satisfy the new challenges. In late 2008 the first seed for the Long-Term Evolution (LTE) access network saw light. The fourth generation mobile communication system can be defined as an all IP based wireless network designed to support high speed data transfer, and due to its heterogeneity it can deliver good services indoor and outdoor, the LTE provides service with better quality at high speed and lower cost . The predecessor 3G networks contained a radio network controller (RNC) in its radio network architecture, this controller is responsible of controlling a group or all of the 3G base stations, , while in LTE the network is flat, control functionalities are embedded in the evolved NodeB (eNodeB), therefore the need for the RNC is no longer necessary, and the 4G flat and simple network allows quicker data exchange and eliminate delay caused by the signalling overhead [3][4].

When LTE was launched, it was looked at as the new technology overriding the legacy 3G network from all point of view, similarly ; the LTE-A was launched to expand the capabilities of LTE and to fully cover the requirements for the 4G which was announced by the ITU. The improvements in LTE-A led to commissioning of the LTE-A Pro which further extend the functionalities of the LTE-A and prepares for launching the next generation 5G. Fundamentally the LTE can be considered as 3.5 G, LTE-A is the 4G, and the LTE-A Pro as 4.5G

## 2.2 Long Term Evolution - LTE

Prior to the launch of 4G, mobile communication networks were used mainly to make phone calls or send text messages and transfer Data therefore being online 24/7 were not a major concern. Post- Millennium these networks gradually became dominated by emailing, browsing, and applications usage, all this created huge amounts of data to transfer across mobile communication networks along with the increased development of applications and technology in the form of smart phones this data grew exponentially. In response to this growth, network operators had to increase the capacity of the mobile networks.

One of the ways of achieving this is by improving spectrum utilization using better technologies, so that Data can be transmitted and received faster between mobile phones and the base stations. The main driver behind the introduction of LTE is a change in mobile communication technology, so that it runs faster to achieve higher data transfer rates. Another motivation for LTE is the reduction in delays to figures of around 20 to 30 ms [8], which are suitable for the more recent kind of applications such as interactive remote games, Virtual reality (VR) and Augmented Reality (AR) applications. 3GPP (3rd Generation Partnership Project) is the global organisation that defines specifications for mobile technology standards and can be considered as a centre point for mobile networks beyond 3G.

The ‘Releases’ are the indication units of the specifications issuing and revision, when a Release is completed all the new added features will be ‘frozen’ and ready to be commissioned. The 3GPP might develop more than one Release simultaneously and aims to make the new technology backwards compatible. The LTE was proposed for the first time in R8 and in the following Releases a lot of enhancing features were added to boost the performance starting with interference management and not ending by carrier aggregation in R9, R10, R11 and R12. The R13 and R14 were improvements on the LTE-A technology path and paved the way towards the 5G. 170 features were defined in R13 during its active period (September 2012 – March 2016) [9]. One of the standards based and concluded on R13 and considered as the major contribution was the new narrowband radio technology (NB-IoT) which was developed for IoT in (June 2016) [9].

R14 was built depending on many contributions, researches and projects that were worked out in R13, and it was active for the period (September 2014 – June 2017). As a development of the LTE network, the LTE-A aimed to increase the available bandwidth and deliver higher data rates, furthermore; it extended the range of the connected devices. As a result, the LTE-A successfully improved and optimised the network capacity and performance, functionality and efficiency compared to LTE, to provide better services to the end user.

### 2.2.1 System Architecture Evolution

In 2G and 3G communication systems data is handled separately from the phone calls this is because the packet core was added later to the architecture forming a 2-core network that might be connected to 2 different backbones. Simply in 3G network, the UE is connected to the base station in the radio access network and the radio access network is connected to the core network, the core network has 2 domains the circuit switch (CS) and the packet switch (PS), the CS is connected to the PSTN and the PS is connected to the internet, This architecture is evolved in the LTE in one evolved packet core (EPC) with PS domain embedded and is connected to the internet. Hence the phone calls were processed in the CS domain which no longer exist in the LTE, alternatively the phone calls are placed using 2 techniques the first one is called (CSFB) Circuit Switched FallBack "(CSFB) is a technology where voice and (SMS) services are delivered to LTE devices through the use of GSM or another circuit-switched network", and this causes a significant delay in most cases, the second technique is via the IP Multimedia Subsystem entity (IMS) where all the packets are forwarded to the IMS [11] and the IMS is connected to the outside CS  networks through the PSTN , all signalling and call set up functions are initiated by the IMS. Figure 2.3 illustrates the described system architecture evolution in the form of block diagram.

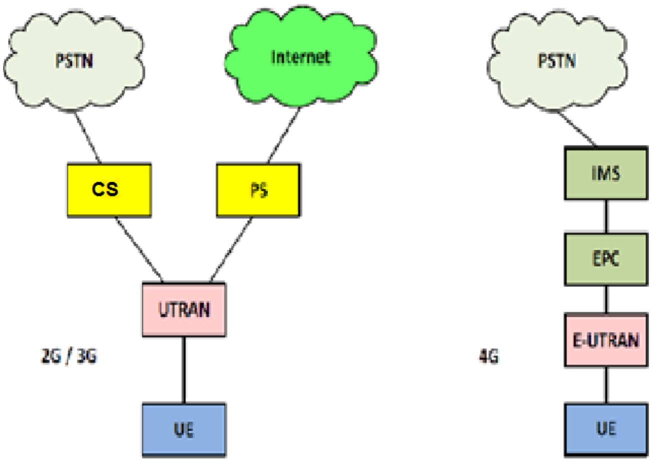


Figure 2.3: System Architecture evolution.

### 2.2.2 Evolved Packet Core

The LTE system evolved packet core consists of the following entities:

- The PDN gateway (PGW) is the part responsible for interfacing the core with the next external system, such as internet or IMS, it's the first entity for packets being transferred to/from the LTE core, it also assigns the IP address for the UE in order to be recognized and connected to the external system.

- The Serving Gateway (SGW) is the entity responsible of forwarding the packets between the PGW and the eNodeB in the E-UTRAN, the reason for forwarding the packets in two stages (PGW & SGW) is to handle the roaming process [7][8].
- The Mobility and Management Entity (MME) is the main control part in the EPC its responsible for mobility management and identify and authenticate the UE and security parameters when the UE request access to the network and is the Non Access Startum (NAS) termination point.
- The Policy Control and Charging Rule Function is responsible for policy control decision making, allowance check and controlling all the flow-based charging functionalities in the Policy Control Enforcement Function (PCEF), which is part of the P-GW and is responsible for (QoS) authorization providing [12].
- Home Subscriber Server (HSS) is the equivalent of the Home Location Register HLR in the GSM with added and evolved functions, it stores all user-related and subscriber related information. It also provides support some functions in mobility management, call and session setup, user authentication and access authorization [12].
- Figure 2.4 shows the block diagram of the LTE system EPC

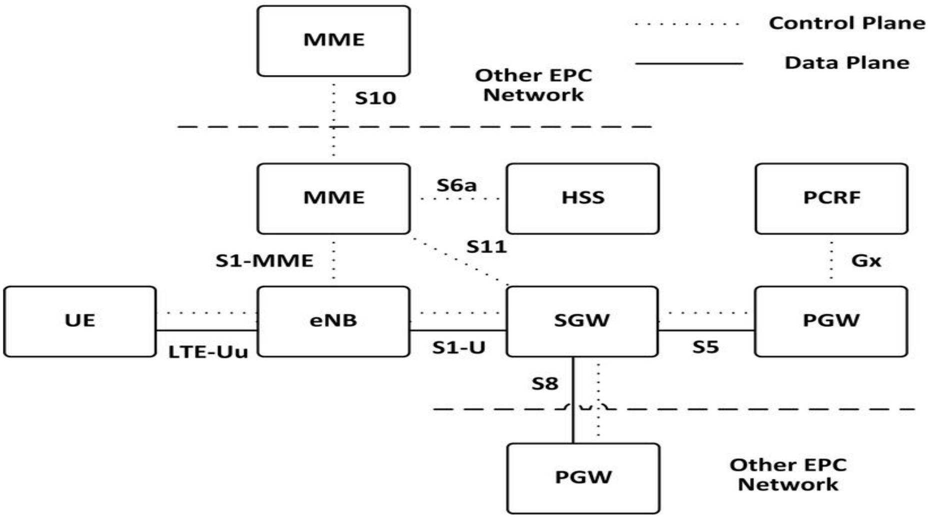


Figure 2.4: EPC block diagram.

Due to increasing number and kinds of the devices in the network and the huge growth of the data routed in the network with all the emerging technologies, the amount of signalling information and traffic data had increased exponentially, and separating the signalling and control functionalities from the traffic data will create easier ways and better procedural steps towards fulfilling the network evolution requirements. Hence the network is split into two main planes; data plane and control plane as shown in Figure 2.4, as a consequence of this separation the S1 interface is split into 2 sub interfaces with same functionality, S1-U is the interface between the E-UTRAN and the SGW for data, and the S1-MME is the interface between the UTRAN and the MME for control [13].

### 2.2.3 Interfaces in LTE

Interface is the point where two different systems or subsystems meet and interact, in LTE network there are several interfaces used to transfer the information between different levels of protocols and network layers, each interface is built in a standard way described by 3GPP [14].

- The Uu interface is the air interface connection between the UE and eNodeB, these two elements use this interface whenever information is trancieved between them.
- The X2 interface connects one eNodeB with another eNodeB. This allows signalling and traffic to be transferred between neighbouring eNodeBs, it is also split into X2-CP interface for signalling exchange between eNodeBs, and X2-UP interface for traffic exchange between eNodeBs directly.
- The S1 interface as explained in the previous section is split into 2 sub interfaces with same functionality S1-U is the interface between the E-UTRAN and the SGW for data, and the S1-MME is the interface between the UTRAN and the MME for control.
- Both X2 and S1 interfaces are IP based and will likely use a physical connection either copper or fibre though they could be connected via spectrum in terrestrial areas or where physical connection is not available, the signalling procedures and the types of exchanged messages sent across these interfaces are defined by the application protocols.
- The S11 interface connects the SGW to the MME, signalling related to bearer generating and mobility management exchanged between these 2 nodes via this interface, and no traffic uses S11 interface.
- The S5 interface represents the main connection for application data across the EPC because it connects the SGW and the PGW which is the provider of the IP services. S5 interface is used by both data plane and control plane.

- The S8 interface is similar to S5 but it connects the SGW in a certain EPC to a PGW in another core network, this interface is used when the UEs are roaming away from their home network.
- The S6a connects the MME to the HSS and is for control plane signalling.
- The GX interface connects the PCRF and the PGW, the PCRF provides the charging information to the PGW, its sometimes denoted as S7 interface [3][9][16].

### 2.2.4 Channels

Channels in LTE are signals or information exchanged between the layers of the protocol stack and between the UE and the eNodeB using a certain bands of frequencies. 3GPP specified 6 channels bandwidths for LTE [17], with characteristics shown in Table 2.1.

| | 1.4MHz | 3 MHz | 5 MHz | 10 MHz | 15 MHz | 20 MHz |
|---|---|---|---|---|---|---|
| Number of resource blocks | 6 | 15 | 25 | 50 | 75 | 100 |
| Number of subcarrier | 72 | 180 | 300 | 600 | 900 | 1200 |
| UL subcarrier bandwidth (MHz) | 1.08 | 2.7 | 4.5 | 9.0 | 13.5 | 18.0 |
| DL subcarrier bandwidth (MHz) | 1.095 | 2.715 | 4.515 | 9.015 | 13.515 | 18.015 |

Table 2.1: Channel Bandwidths for LTE [17].

A resource block (RB) represents the basic unit of resource for LTE air-interface, the eNodeB scheduler allocates resource blocks to the UE when initiating a data transfer, each RB consists of 12 subcarriers and each subcarrier occupies 15 kHz of spectrum, and the total subcarrier bandwidth is smaller than the channel bandwidth to spare some guard band between subcarriers at the band edge, the larger channel bandwidths can support achieving higher throughputs whilst the smaller bandwidths that leads to achieving lower throughput values but are much easier to accommodate specially when low spectrum availability. Figure 2.5 illustrates channel bandwidth and subcarriers allocation for E-UTRA carrier.

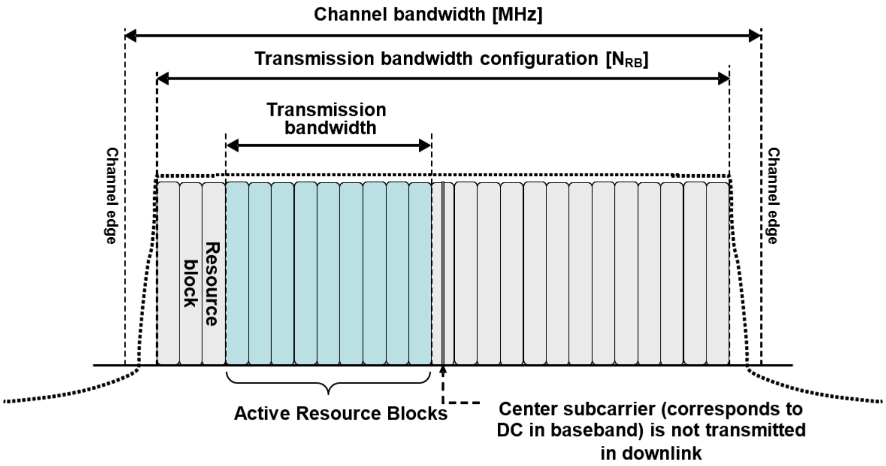

Channel bandwidth [MHz]
Transmission bandwidth configuration [N_RB]
Transmission bandwidth
Channel edge
Channel edge
Resource block
Active Resource Blocks
Center subcarrier (corresponds to DC in baseband) is not transmitted in downlink

Figure 2.5: Channel Bandwidth and transmission bandwidth [17].

Channels can be sorted into physical channels, transport channels and logical channels and are linked as shown in Figure 2.6.

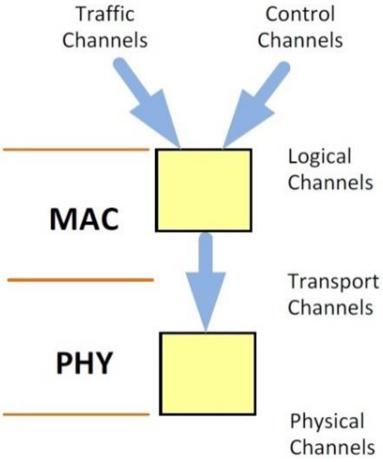


Figure 2.6: LTE Channels.

The physical channels are set between the eNodeB and the UE as streams of information exchanged via the uplink and downlink. The subcarriers are organised in such a way that, at any one moment in time, a particular subcarrier is transmitting on a particular channel [18]. For example the physical uplink control channel (PUCCH) carries the channel quality indicator (CQI) as acknowledgment from the UE to the eNodeB, and the physical broadcast channel PBCH carries information about the bandwidth which can be recognized by the UE and so on. The logical channels are streams of information flowing between the radio link control RLC and the Media Access Control MAC layers in the protocol stack and they are also characterised by the type of information they carry.

One of the significant factors related to the work presented in this thesis is the received radio signal, reference signals are used for 2 purposes: the first is measuring the strength of the signal received at the UE in order to filter the subcarriers, and decide which ones are suitable. The UE scans for the reference signals which are distributed across the time and frequency domains so that the mobile will be able to find and measure them. The second is that they enhance the mobile capabilities to process the received information, by briefly acting as a signal transmitted with known amplitude and phase angle, and this will allow the UE to detect and supress the effect of amplitude and phase changes. On the downlink, the demodulation and sounding signals perform the equivalent tasks for the uplink [18].

### 2.2.5 Scheduling

The scheduler is an entity in the eNodeB it decides which of the connected UEs has the priority for resource allocation and how much resource blocks to allocate depending on the received CQI from the UEs and on the resource availability (Figure 2.7). Scheduling is done at per sub-frame basis i.e. every 1 millisecond.

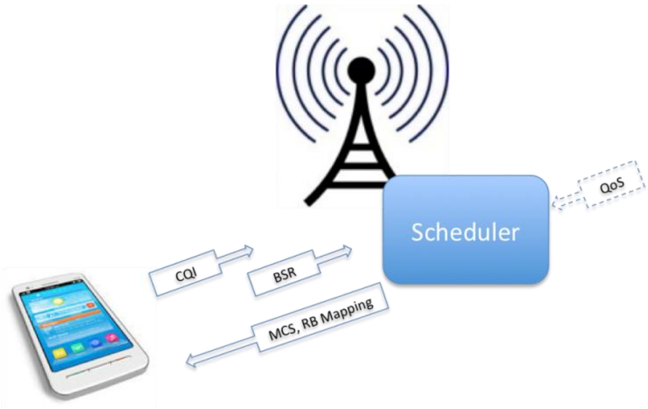


Figure 2.7: Scheduler in LTE.

Scheduling in LTE can be classified into non-persistent (Dynamic) and semi-persistent, as persistent scheduling is not used in the LTE, in dynamic scheduling the network is granted full flexibility to assign the resources to the UE as compared to persistent scheduling where it gives resource allocation is every sub-frame, depending on the channel conditions dynamic scheduling changes the resource allocation to the UEs [20]. In LTE semi-persistent scheduling SPS the process is designed to reduce the control channel overhead for VoIP applications and services. Since voice over LTE (VoLTE) require persistent radio resource allocation at regular intervals (one packet in 20ms from AMR speech codec), SPS allocates radio resources for a long period of time [21].

### 2.2.6 LTE Bearers

The LTE system provides end to end service to the end user using the hierarchy of bearers shown in Figure 2.8 [3].

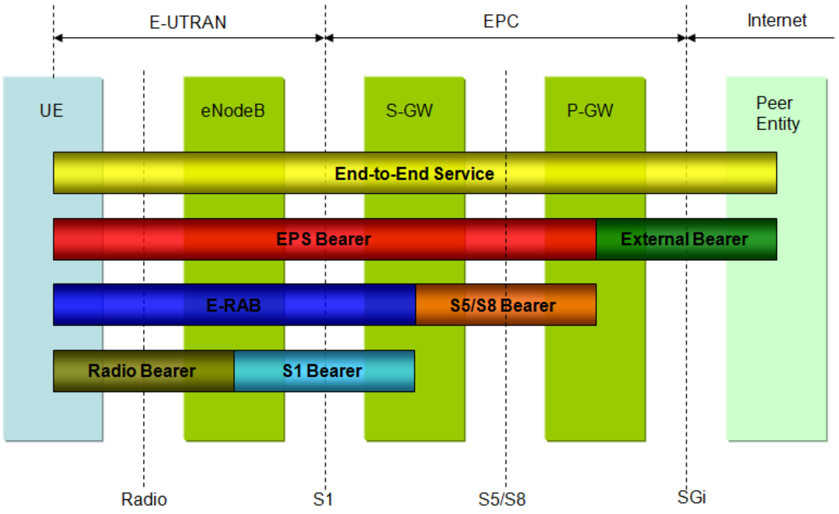


Figure 2.8: LTE Bearers [3].

The radio bearer supports the connection across the air interface and the S1 bearer provides the connection across the transport network, combining these two bearers result in generating the E-UTRAN Radio Access Bearer (E-RAB) bearer. As mentioned in Section 2.2.3 the S5 interface provides connectivity between the SGW and PGW of the same core, while the S8 interface provides the connectivity between the SGW and the PGW belonging to two different cores, combining the S5/S8 bearer and the E-RAB bearer leads to the generation the EPS bearer, and all data plane information transferred by the same EPS bearer will have the same quality of service QoS, the EPS bearer provides data plane connectivity from the user to the PGW, there are many classes of EPS bearers such as; default EPS bearer which is set immediately when the UE request anchoring to the network, other types of bearers ‘known as dedicated EPS bearers' are established either to connect to different PGWs or to provide different QoS in the same PGW. Signalling radio bearers (SRB) are used to transport the signalling messages for Non-Access Stratum (NAS) between the UE and the MME, and the RRC signalling messages between the UE and the eNodeB [7].

## 2.3 LTE Advanced

Increasing capabilities and improving the performance of the LTE lead to the birth of the LTE Advanced (LTE-A) which was introduce in Release 10 of 3GPP specifications and its requirements were specified in [22]. The requirements for LTE-A have been set to comply the requirements of IMT–Advanced specified by the Radio communications division of the International Telecommunication Union (ITU-R). These most important requirements could be summarized by the following; Peak throughput requirements for LTE-A are 10 folds of those in LTE, this huge boost was achieved by increasing the bandwidth and using multiple antenna transmission ,the maximum bandwidth in LTE was 20 MHz, and was increased to a maximum bandwidth of up to 100 MHz in LTE-A, in the downlink direction, the 4×4 MIMO evolved to 8×8 MIMO, and the single antenna transmission was evolved to 4x4 MIMO in the uplink direction. Also and as consequences of these enhancements the peak spectrum efficiency which is a measure of throughput per unit of bandwidth (bps/Hz) was increased by 6 times, and the control plane latency which is the delay in changing the state of the UE to be ready for data exchange via user plane connection was also reduced in LTE-A, if the UE was in the 'connected' mode the latency is reduced from 50ms in LTE to 10 ms in LTE-A, and if the UE was in the 'idle' mode then the latency is reduced from 100 ms in LTE to 50 ms in LTE-A [22].

The main technologies proposed by LTEA are:-

- Carrier Aggregation (CA)
- Coordinated MultiPoint transmission (CoMP)
- 4X4 MIMO in uplink
- 8X8 MIMO in downlink
- Heterogeneous networks
- Introducing Relay nodes

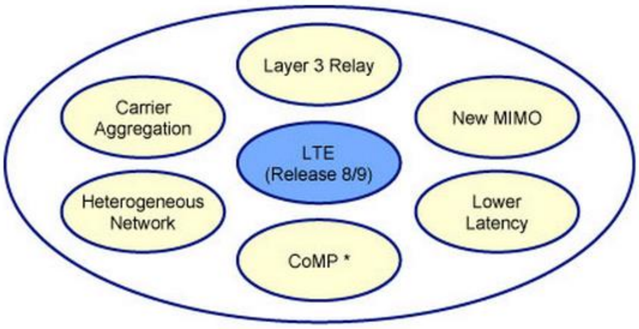


Figure 2.9: LTE-A technologies.

## 2.3.1 Carrier Aggregation

In intra band adjacent carrier aggregation the channel bandwidth is increased by combining multiple carriers, according to Release 10 of 3GPP specification it was defined to support signalling for the combination of up to 5 single carriers resulting in maximum channel bandwidth of 100MHz, these component carriers are not necessarily adjacent and they can be from the same or different operating bands [3][19]. Figure 2.10 shows the aggregated channel bandwidth.

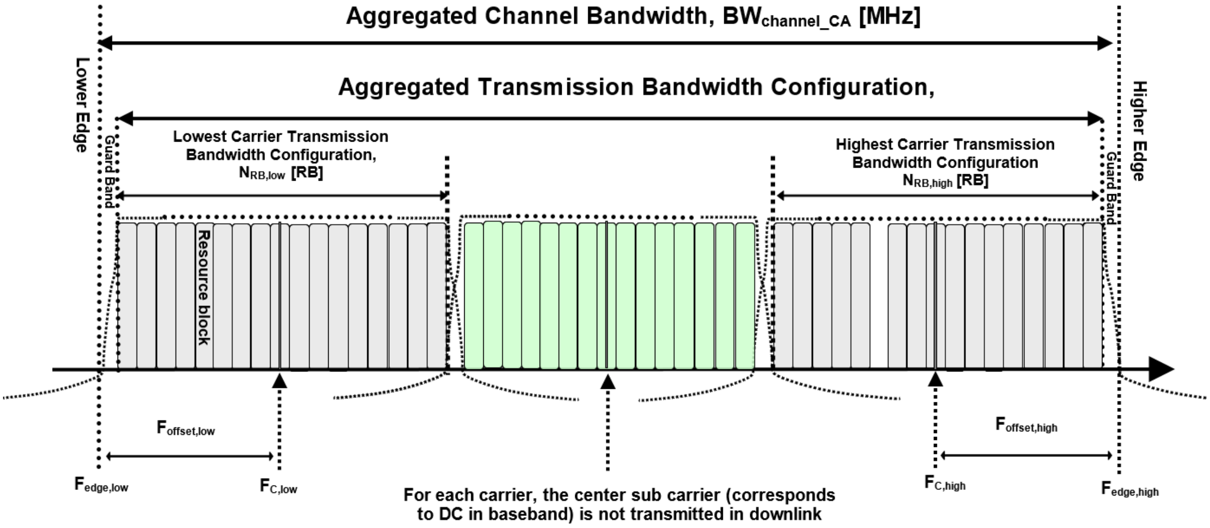

Aggregated Channel Bandwidth, $BW_{channel\_CA}$ [MHz]
Aggregated Transmission Bandwidth Configuration,
Lower Edge
Higher Edge
Guard Band
Lowest Carrier Transmission Bandwidth Configuration, $N_{RB,low}$ [RB]
Highest Carrier Transmission Bandwidth Configuration $N_{RB,high}$ [RB]
Guard Band
Resource block
$F_{offset,low}$
$F_{offset,high}$
$F_{edge,low}$
$F_{C,low}$
For each carrier, the center sub carrier (corresponds to DC in baseband) is not transmitted in downlink
$F_{C,high}$
$F_{edge,high}$

Figure 2.10: Aggregated channel bandwidth [19].

### 2.3.2 Coordinated Multipoint

Coordinated Multi-Point transmission/reception (CoMP) is a DL/UL technique for increasing the capacity of the network and enhance the throughput for users at the cell edge. In a network containing a centralized controlling eNodeB and a Remote Radio Equipment (RRE) as a second transceiver Figure 2.11 shows the eNodeB can centrally control all radio resources by transmitting content data directly from the eNodeB to the  RRE on optical fibre connections, this technique is very robust in areas covered by optical fibre services. There is little signalling delay or other overheads in this technique, but the intra-cell radio resource control is relatively easy, another drawback is that the central eNodeB must be capable to cope with increasing load due to adding and connecting new RREs and this might need more powerful servers or circuitry.

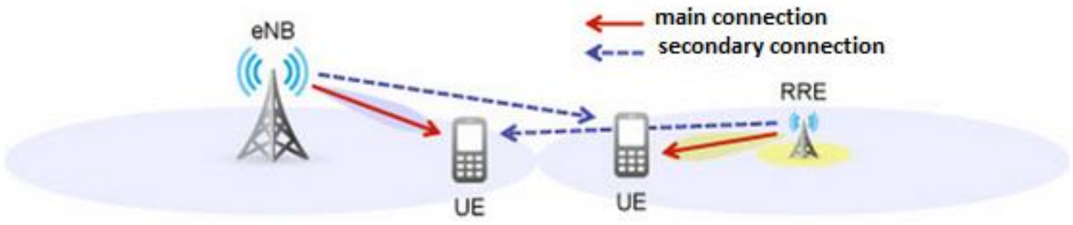


Figure 2.11: CoMP concept.

### 2.3.3 Heterogeneous LTE Network

Pre LTE-A launch, all traditional cellular networks were deployed in homogeneous way by using one fixed type of base transceiver subsystem (BTS), the BTS is commissioned to provide services to the UEs within its coverage. All the BTSs have the same transmission power, however the directions and angles of the sectors can be modified to change the antenna transmission pattern, and when possible the same backhaul connection to the data network [4]. In Heterogeneous (HetNet) network architecture which was defined in Release 10 of 3GPP macrocell as main node of the network and some low-power base stations, i.e. micro, pico, femto and relay nodes, these small cells are robust and effective in compensating the dead zones that might occurred in the traditional systems (with macrocell only), boost the capacity and support other new services that require radio access, Figure 2.12 illustrates a typical HetNet [23].

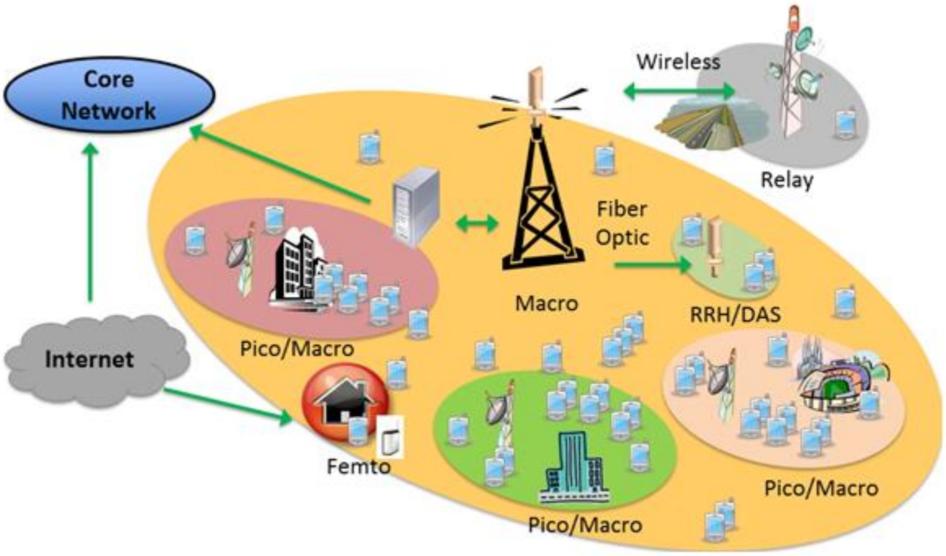


Figure 2.12: Heterogeneous Network (HetNet) [23].

Deploying a HetNet allows wide flexibility and less complexity in service providing to the different types of network users at a cost effective margin. A major issue in this kind of networks is managing and reducing the effect of interference to achieve a worthy influence when implementing such architecture [25]. Higher QoS plus higher signal to noise ratio (SNR) in the network could be scored when packet drops goes as minimum as possible resulting in higher data throughput, and this could be improved when using the correct channel estimation and allocation schemes. To maintain good indoor coverage, the macro eNodeB need to increase the transmission power to serve the unreachable indoor users, but this will certainly cause a serious inter-cell interference which will lead to severe drop in network performance. Introducing femtocells for indoor applications is a good alternative to perfectly cover the indoor areas; they can provide all the network services resulting in network capacity increase. Better traffic handling could be achieved when a large amount of data is delivered to users residing outside the macrocell's coverage, which will be able to serve more users within its coverage while keeping its transmission power unaltered [18][22].

### 2.3.4 Cell Types in HetNet

Small-cell network term is denoted for networks that have, short range and lower transmission power, the purpose of implementing them is to keep network users connected and provide the same QoS to users moving within all the network areas. A heterogeneous network solution should adopt efficient spectrum management schemes by using a range of small-cell nodes depending on the network demands and applications. Figure 2.13 shows different kinds of nodes in HetNet architecture, sorted according to the coverage area.

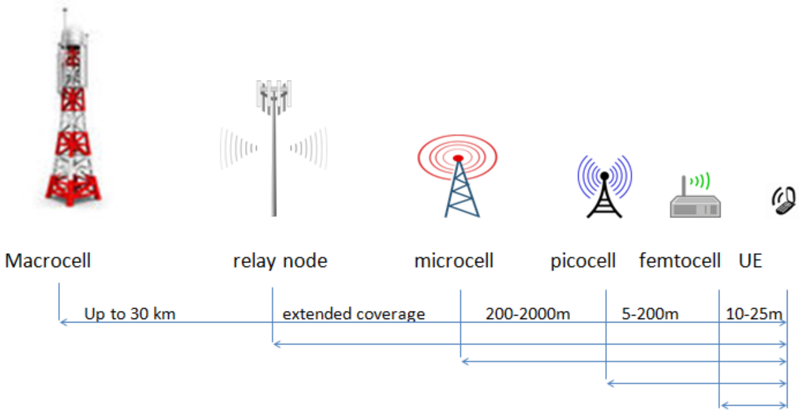


Figure 2.13: HetNet base stations.

- **Macrocell**

Macrocell nodes can cover large outdoor areas of a radius up to 30 km by transmitting high power levels, the power ranges from 40 to 100 Watt. They can provide services to a fixed number of users, the number of users are governed by cell deployment, brand, and available channels; it varies from 200 to 1000 users.

- **Microcell**

Microcell nodes are used in outdoor areas to relief macrocell nodes from users and handle them independently the cell radius is around 2 Km and the transmission power of 2 to 10 Watts depending on the equipment. Microcells have also been used in GSM and 3G cellular networks, due to the ability to cooperate in outdoor areas.

- **Picocell**

Picocells are mainly used to extend the coverage in dead zones in order to support more users, they have lower transmission power compared to macrocells and microcells. Deploying picocells has several drawbacks, the networks using picocell nodes will probably have lower signal to interference ratio, due to the unplanned distributing on the network. Due to the difference in transmitted power levels when deploying macro and pico cells combination in the same network, therefore in the downlink the pico cell will have very limited coverage. However, this doesn’t apply in the uplink direction because the transmission power from UE to any kind of base stationsis the same, Picocell BTS could be used either in indoor or outdoor, the cell radius roughly 200 meters.

- **Relay Node**

Relay Stations (RSs) are introduced to attract user information from the surrounding user equipment (UE) to a local (eNodeB). The RSs are used to increase the total system throughput by boosting the signal they receive from an eNodeB. Relay nodes van be classified into 2 types:

Type 1 (non-transparency relay), which enables a remote UE located outside the coverage of eNB to access to the eNodeB.

Type 2 (transparency relay) enables a UE residing inside the coverage of eNodeB and receiving the service through an active connection to enhance the link capacity and QoS. So it is not used to transmit the signalling and the control information, but they are used to increase the overall system capacity. Hence, the main advantage of combining relay nodes and macrocell in the same network is to boost transmission and reach UEs located outside the coverage area and improving the existing connections if any.

- **Femtocell**

Femtocells are intelligent access point deployed indoor to maintain network availability for 3G and 4G UE. Femto nodes can be deployed within the existing macro networks, therefore the involved users can switch between macro and femto easily and less complexity when the femtocell is used within closed mode HetNet, any user associated with the femto cell will be able to access the network. Femtocells might cause high interference when deployed; therefore it’s very important to follow good switching and handovers schemes to guarantee maximum benefit from the network.

### 2.3.4 Small Cells and Dual Connectivity

According to [29] the definition of small cells is; “any node in the network that transmit lower power than the BS classes of macro node can be considered as small cell”. This will put Pico and femto cells that have transmission power of 0.25 W and 0.1 W respectively under small cells group [3].The 3GPP Release 12 defined Dual connectivity as the technology which extends carrier aggregation (CA) and coordinated multi-point (CoMP) to inter-eNodeB with non-ideal backhaul [26], the focus on inter-eNodeB CA case led to improving throughput per UE and mobility robustness, DC enables UE to have two separate connections to an MeNodeB of macro cells and an SeNodeB of small cells, simultaneously, However, those serving cells for the CA should belong to a single eNodeB that can have multiple serving cells by sectors and operational frequencies. DC in LTE can be deployed according to any of the following three scenarios as represented in Fig2-14, the small cells can adopt intra-frequency or inter-frequency depending on the network resources, the scenarios are classified into;

- Scenario1: Intra-frequency scenario
- Scenario 2: Inter-frequency scenario
- Scenario 3: Out of coverage scenario

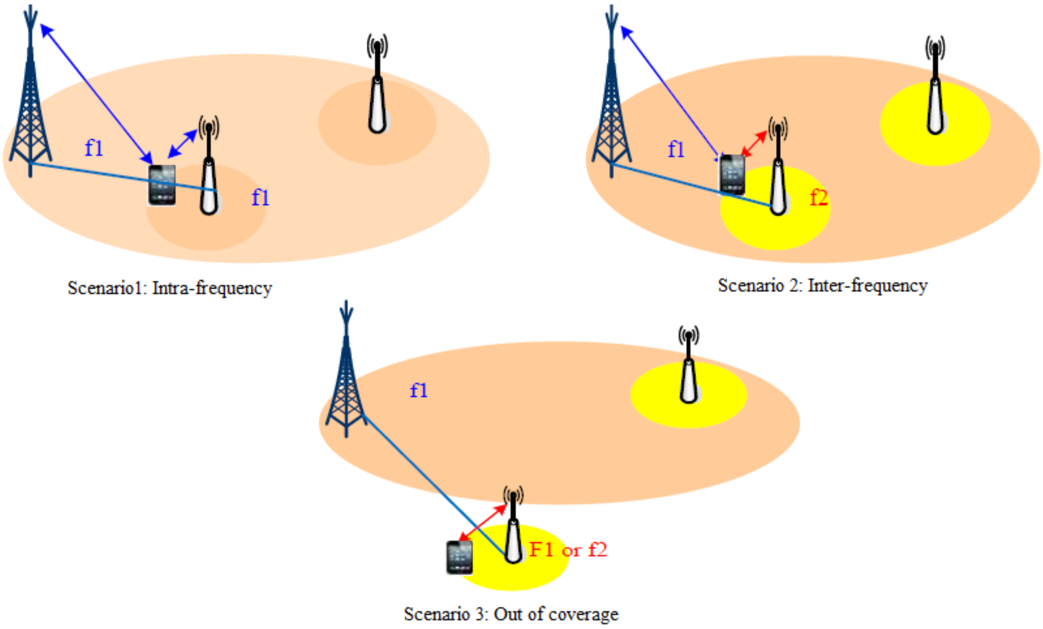

f1
f1
Scenario1: Intra-frequency
f1
f2
Scenario 2: Inter-frequency
f1
F1 or f2
Scenario 3: Out of coverage

Figure 2.14: Dual Connectivity scenarios.

### 2.3.6 Mobile Edge Computing

Edge computing is a technology that enables cloud computing features and IT capabilities to be available at the edge of the network [27]. Since LTE is mainly a mobile network and has flat architecture, which will consequently create a perfect environment for edge computing deployment at the eNodeBs therefore it is denoted as Mobile Edge Computing (MEC), MEC can perform tasks that could not be achieved with traditionally centralized network architectures, this will lead to  improving 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). Deploying MEC within the radio network can decrease latency and efficiently enhance the bandwidth it will also offer processing to some real time performance indicators such as UE position, cell load, and power consumption, processing these factors in the MEC server can be used to efficiently design and deploy applications that can differentiate the mobile broad band quality of experience, reduce cost and increase the performance [28].

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 mobile operators focused option is the ETSI MEC (Multi-access Edge Computing) [29]. Other MEC examples include new emerging technologies such as Internet of things (IoT), Augmented Reality (AR), data caching...etc.

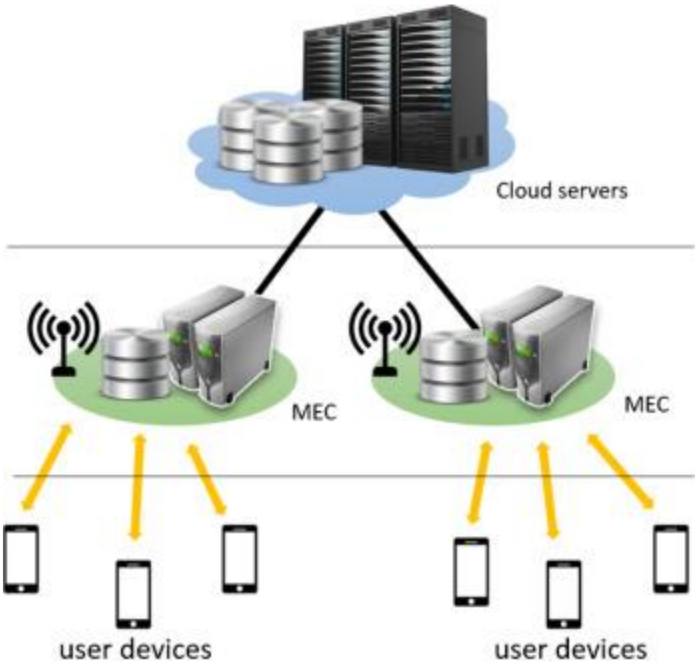


Figure 2.15: MEC deployment in mobile network.

## 2.4 State-of-the-Arts Work

The recent works demonstrated the benefits of small cells, the effect of their range expansion and network performance evaluation in HetNets using different approaches, and how they would be deployed to comply with the next generation(s) requirements and standards.

In [30] a mmWave-based backhaul mesh Ultradense network (UDN) is considered as a model, then an optimization model using IBM CPLEX solver is proposed to find optimal solution in terms of throughput and fairness. Also a heuristic algorithm is proposed to reduce complexity and compare it with the optimal results in evaluation, upper and lower bounds of aggregated UE throughput using different fairness weights in random HetNet topologies were examined, According to simulation results, dense SeNBs and large mmWave bandwidth improve network throughput by increasing access and backhaul link capacity.

In [31] the authors considered the same network model as in [30] but using OPNET modeller, same challenges were investigated, they analysed the potential gains when splitting the data bearer in the uplink in terms of load balancing and per user throughput and compared the drawback of the UE performance when connected to different cells, the results showed an increase in per-user throughput and better load balancing between the MeNodeB and the SeNodeB when implementing data bearer splitting in the uplink, but at the cost of more complexity in the UE functionalities.

The work presented in [32] is a part of project ARTIST4G, it propose a mobile broadband technological framework in which to design interference avoidance solutions using numerical simulations and test bed measurements. The results demonstrate a significant improvement in the cell edge performance in terms of delay and per user throughput when using CoMP and HetNet and considering planned and unplanned deployment.

A detailed survey on the advanced handover techniques, requirements and features for LTE-A system is presented in [33], focusing on advanced handover techniques Fractional Soft Handover (FSHO), Semi Soft Handover (SSHO) and multi-carrier handover (MCHO) that incorporate backward compatibility to the existing systems, and highlighting the limitations when performing HO in terms of high latency, handover procedure unreliability, high outage probability and data loss, and proposed a hybrid HO scheme based on combining FSHA and SSHO with multi-carrier handover techniques the results show that the combination enhanced the system performance in term of latency, outage probability, interruption time and reliability during handover especially at cell edge as well as reducing the transmission overhead on the serving cell, which will consequently balance the load among the network cells.

The authors in [34] proposed a wireless HetNet with split the control-plane and user-plane where the macro cell and SCs are equipped with some caching abilities. Comparisons of throughput and energy efficiency are made between the model and a traditional LTE network,

and numerical results are presented to validate the analysis and the impacts of the key parameters on the performance, the performances are derived from the characteristics of coverage, throughput and energy efficiency, they considered coverage threshold, transmission power, density of cells, and cache size, as key parameters for the numerical analysis.
As an outcome of the comparison the numerical results show that the proposed cache-enabled network has much higher throughput and improved EE than current LTE networks. Similar to the work in [34] but with MEC base caching and virtual RAN, the authors in [35] built a testbed to evaluate the proposed architecture in a proof of concept and test it with typical caching scenarios and different sorts of data. Another work that demonstrated the influence of caching on an LTE-A network is presented in [36] and concentrated on power consumption, in order to formulate the problem of minimizing energy consumption at eNodeB caches a power model is built and optimized it using Lagrangian relaxation technique to find a near-optimal solution to the proposed model, Lagrangian relaxation is a relaxation method in mathematical optimization which approximates a difficult problem of constrained optimization by a simpler problem. A solution to the relaxed problem is an approximate solution to the original problem and provides useful information [37].

According to the presented results, the optimization method reduced the power consumption by 28% while keeping same QoS. The results also indicated how crucial are the content popularity and caching servers are to eNodeB energy efficiency. The authors of [38] also examined the effect of adding caches in the radio network to save the transmission energy, in order to prove that; a system level simulation of LTE-Advanced Release 11has been designed to illustrate the effects of forwarding scheme and adding cache in the network using a simple probabilistic model , platform is co-build with Qualcomm and sponsored with Qualcomm Innovation Fellowship, according to the simulation result, the method was able to save up to 50% (depending on the buffer size) of the total energy consumption compared the consumption in LTE-A network without caches.

In [39] the authors considered a hierarchical collaborative caching framework with the cache deployment at the MEC servers; based on this framework they designed a caching strategy. They formulated the hierarchical collaborative content placement problem as an optimisation problem to reduce latency when running with limited cache capacity. Since finding the optimal solution is an NP-hard problem, they propose a genetic placement algorithm to find the near-optimal solution to reduce the computation complexity. And according to the numerical results, the proposed algorithm had improved the performance compared with the reference algorithms.

The authors of [40] used OMNeT simulator to build the models for centralized cell coordination at both small and large scales, and presented the work as layered solution considering small-scale coordination as the first building block of a large-scale scheme and so on, the results showed improvement in terms of fairness, cell throughput and energy efficiency. In the same domain of HetNet the authors of [41] proposed a rate adaptation and MAC algorithms for video transfer based on the user QoS requirements to improve the throughput, the improvements are done while user satisfaction for video is at an acceptable level, the user satisfaction for video is measured by the Mean Opinion Score (MOS) and is mathematically derived, the simulations were performed using OPNET modeller with certain assumptions that comply with the research motivations. Simulation results show that an increase in system capacity while keeping adequate MOS levels can be achieved for video applications users. Though all the selected researches listed above were orbiting around the LTE-A HetNets and many other researches are on the go in the same criteria the combination of intelligence and content delivery is very rare, many vendors and major players in the telecommunications field are competing and funding projects related to 5G, some are concentrating on specific features like radio network or the core, while others are considering the case from all sides, 3GPP, Qualcomm, Nokia, Eriksson and Huawei and many others working groups are working to bring network infrastructure and UE for 5G roll out by 2020.

## 2.5 Summary

LTE-A system has been able to provide significant improvements in performance for the mobile network. The core network as well as the radio access network seen upgrades and improvements, the result of all the upgrades is that users see significant performance improvements in the service provided, and the operators also gain greater returns compared with the same spectrum resources in previous systems, i.e. the cost per bit is reduced, and with the faster speeds, users tend to consume more data, thereby raising revenues. Accordingly LTE-A has provided improvements to both of the users and operators, as well as those providing additional services such as ISPs, mobile applications developers, and operating systems updates.

# CHAPTER 3

# ARTIFICIAL INTELLIGENCE TECHNOLOGIES AND IMPLEMENTATIONS

## Briefing

This chapter comprises of three main sections the first reviews the start of Artificial Intelligence (AI) , history, technologies, and applications, the second explains the architecture and features of Neural Networks which is the optimization tool used in this work and the benefits of using them in various optimisation approaches in mobile networks. The third part views some of the related work and researches in this field.

## 3.1 Introduction

Starting back in the 1950s, AI is not a newly founded science, it started with simple models and algorithms but with the digital revolution impact and the increase of interaction between humans and machines over the last three decades, AI became more attractive field for research and industry. Since then and due to the rapid and dynamic pace of development of AI it has been difficult to predict its future path and the ways which is going to alter the world's life.

There are many definitions, proposals and statements that customize AI. The authors of [42] proposed grouping AI systems into four categories depending on the used approach:

| | |
|---|---|
| • Systems that think like humans | approach centred around humans |
| • Systems that act like humans | |
| • Systems that think rationally | approach centred around rationality |
| • Systems that act rationally | |

It is an easy process to build an intelligent system as defined above and can be considered a long-term target.

Alternatively, AI can be classified into two categories depending on how the machine responds to data sets:

- Weak AI or Narrow AI: Weak artificial intelligence is a form of AI designed specifically to be focused on limited task and to look very intelligent at it, it can do one specific task, i.e. the machines which are not too intelligent but capable of doing their own work are built in such a way that they seem intelligent. A good example would be a poker game where a machine beats the player; because all the rules and moves are fed into the machine, the level of difficulty is pre-set before the start and each and every possible scenario or tributary is also entered. Each and every narrow AI will contribute to the building of strong AI [43].
- Strong AI: Strong AI is a machine that can actually think and simulate tasks on its own just like a human being. There are no commercially existing examples for this, but some leading industrial corporations are very keen on building a strong AI. [43].

As a resultant of that, AI researchers have tended to focus their efforts on copying the human thinking process or activity and implement them in machines. These efforts were centred in two main areas:

- Knowledge representation- which is the study of finding the language in which knowledge can be encoded and used by a machine. Knowledge representation in AI is intended to reduce problems of intelligent action to a search problem [43]. The integration of search and knowledge representation is considered to be the core of AI.
- Search- is a method of solving a problem by testing a large number of possibilities and finding one (or very few) solution(s) during the search which are best solutions. In real applications, there is a very large number of possibilities (or the search space), which makes the search intractable and computationally complicated. Methods such as heuristics search (wherein one uses existing knowledge) provide practical solutions to this problem.

However, there are many ways in which AI could be implemented or achieved, as illustrated in Figure 3.1

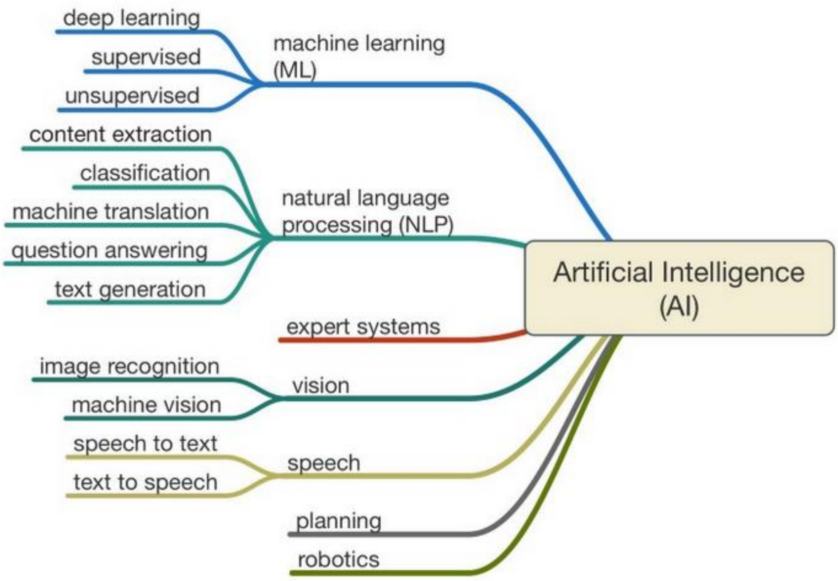


Figure 3.1: AI classification.

Recently, and due to the vast growth in the generated data from communication networks and critical infrastructure, along with the need to analyse this data in an intelligent way, the use of machine learning (ML) algorithms has become an essential in many everyday sectors such as banking, governmental services, surveillance, crime prevention, online gaming…etc. [44].

Building effective and robust models that analyse and predict dynamic system or human behaviour can be achieved using Machine learning algorithms; in such a model the system controller can make intelligent decisions just as human being brain without involvement of human operator. For example, and this is one of the contributions of this work; in a wireless mobile network such as LTE , machine learning tools can be used to analyse big data for mobile edge computing enhancement. Machine learning tasks often depend on the nature of their training data. In machine learning, training is the process that teaches the intelligent machine to achieve the required target. In other words, training enables the machine learning framework to determine the schematic relationships between the input data and output data of the machine learning framework.

Generally; learning schemes can be classified into three key classes as shown in Figure 3.2 [45]:

- Supervised learning, which is driven by the task, e.g. classification and regression.
- Unsupervised learning, which is driven by the type of data, e.g. clustering.

- Reinforcement learning. This is closer to the human learning i.e. the machine learns a policy or rule to guide it acting in certain circumstances to maximize the gained reward.

Supervised learning algorithms are trained using labelled data. When dealing with labelled data, both the input data and its desired output data are known to the system. Unlike supervised learning, the unsupervised learning tasks are done without knowing the output data, instead; unsupervised learning aims to investigate the input data and conclude the output structure directly from the unlabelled data, of course the amount of data needed to train the network is much bigger. A hybrid learning scheme that uses both labelled and unlabelled data for training is called semi-supervised learning, it is used in same applications as supervised learning. Semi-supervised learning is useful when the cost of obtaining fully labelled training data to complete the training process is not possible or relatively high. In contrast to the explained learning methods that need to be trained with historical data, reinforcement learning (RL) is trained by the data from system itself. RL aims to learn about system's environment and find the best strategies for a given agent, in different environments, that makes RL algorithms effective and useful for use in robotics, gaming, and navigation [46].

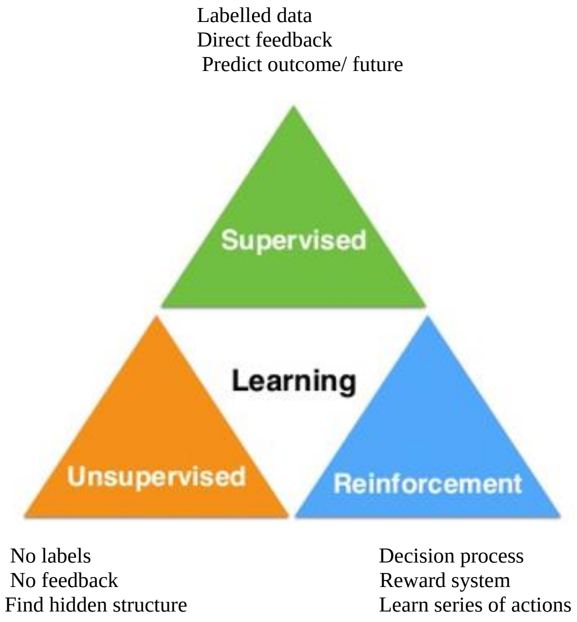


Figure 3.2: Learning in AI.

Several models have been developed to accomplish these learning tasks. The most important one is artificial neural networks (ANN) because they are able to mimic human intelligence in masking complex relationships between inputs and outputs [46]. ANNs can also be used in a self-adaptive manner to learn how to respond and complete tasks depending on the given data for training or previously gained experience.

## 3.2 Artificial Neural Networks

The human brain consists of a large number of biological elements called neurons (approximately $10^{11}$), these neurons are highly interconnected (approximately $10^{4}$ connections per element). These neurons consist of three major parts: the dendrites, the cell body and the axon. The dendrites are tree-like receptive networks of nerve fibres that carry electrical signals into the cell body. Then cell body effectively sums and thresholds these incoming signals. The axon is a single long fibre that carries the signal from the cell body out to other neurons. The point of contact between an axon of one cell and a dendrite of another cell is called a synapse. It is the arrangement of neurons and the strengths of the individual synapses, determined by a complex chemical process that establishes the function of the neural network. Figure 3.3 is a simplified schematic diagram of two biological neurons [47].

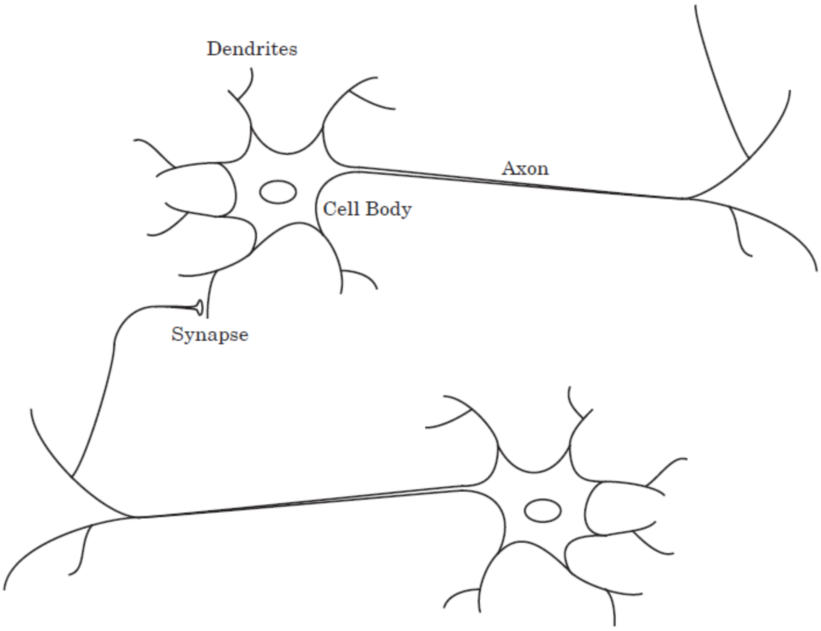


Figure 3.3: Biological neuron.

The artificial equivalent to the biological neurons are entities that act similar to the cell body in mathematical way, the Synapses are modelled by a single variable called weights so that each input is multiplied by a weight before being sent to the artificial neuron [47]. Then, the weighted signals are added together by simple adder to give node activation.

Figure 3.4 shows an artificial neuron model, the unit produces an active high output (+1) if the activation exceeds the threshold, otherwise it delivers a zero.

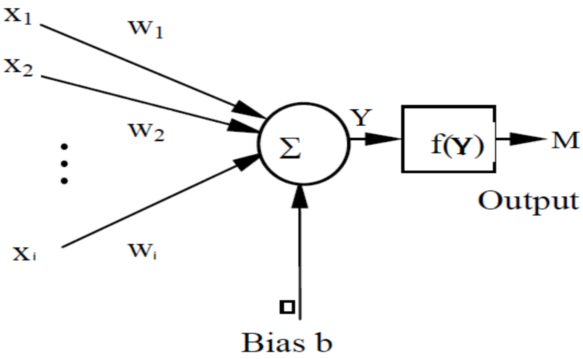


Figure 3.4: The artificial neuron.

The neuron output is simply calculated as

$$M = f\ (\sum_{n=1}^{n=i} X_n\ \ W_n\ \ +b) \qquad (3.1)$$

The actual output depends on the particular transfer function that is chosen. The bias is much like a weight, except that its input value is constant and has the value of 1. It might be omitted as well if it’s not needed. $W$ and b are both adjustable scalar parameters of the neuron. Typically, the transfer function is set according to the design and then the network parameters will be adjusted by the chosen learning rule so that the neuron input/output relationship meets the specific goal [48].

### 3.2.1 Transfer functions

There are different types of transfer function; they can be linear or non-linear, a certain transfer function is chosen to cover some specification of the problem that the neuron is attempting to solve. In general, the three most commonly used transfer functions are; Hard Limit, Linear and Sigmoid.

- Hard Limit transfer function: The hard limit transfer function returns the output of the corresponding neuron to 0 if the function argument is less than 0, or 1 if its argument is greater than or equal to 0 as shown in Figure 3.5, the threshold point could be adjusted with the bias parameter [48]. This function is used to create neurons that can classify inputs into two distinct categories.

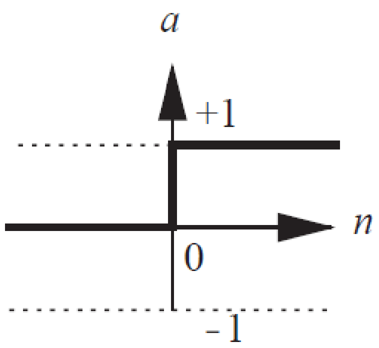


Figure 3.5 Hard limit transfer function

- Linear Transfer function: The linear function is defined as the function that's output is equal to its input, $a = n$ , Figure 3.6 shows a linear function , the output could also be shifted by adjusting the bias and the weights [48].

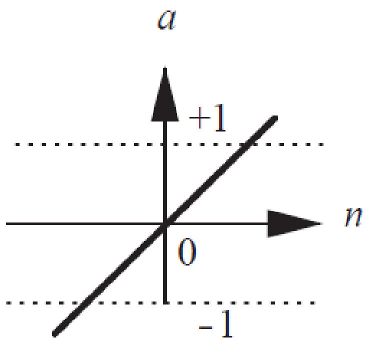


Figure 3.6 Linear Transfer function

- Log Sigmoid Transfer function: This transfer function is widely used in multilayer networks that are trained using the backpropagation algorithm because its differentiable [48], The input (which may have any value between plus and minus infinity) is squashed to generate the output in the range 0 to 1, according to the expression:

$$a = \frac{1}{1 + e^{-n}} \qquad (3.2)$$

The output will be as shown in Figure 3.7 and can also be shifted when implemented in the neuron by using the bias and the weights.

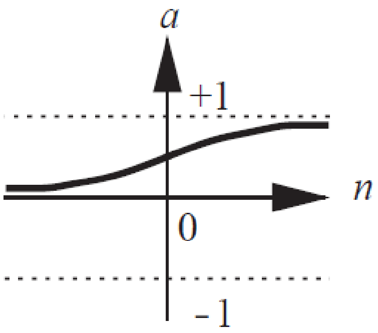


Figure 3.7: Log Sigmoid Transfer function.

In addition to the described transfer functions, there are many other functions that can be used to activate the neuron and they are selected in relation to the tasks assigned to the neuron, such as: Symmetrical Hard Limit, Saturating Linear, Symmetric Saturating, Linear Hyperbolic, Tangent Sigmoid, Positive Linear, Heaviside step function and Competitive.

### 3.2.2 Network Architecture

One neuron (Biological or Artificial) no matter how big or how many inputs are used, won't be an effective entity, hence emerged the need for many interconnected neurons working at the same time, i.e. in parallel, these neurons are lined in the form of layers.

A single-layer network of **(S)** neurons is shown in Figure 3.8, illustrating that each of the inputs is connected to all the neurons and that the weight matrix will also have S rows. The layer comprises of the weight matrix, the adders, the bias vector, the transfer function boxes and the output vector [49].

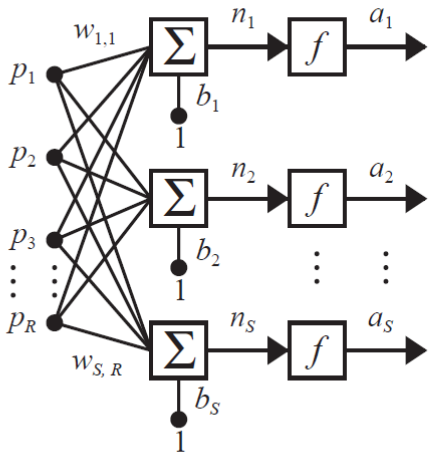


Figure 3.8 S Neurons layer

Each element of the input vector $P_R$ is connected to all the neurons and is multiplied by the weight matrix $W_{S,R}$. Each neuron has a bias input, an adder, a transfer function and an output of $a_S$ vector.

The weight matrix is represented as follow:

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,R} \\ \vdots & \vdots & & \vdots \\ w_{S,1} & w_{S,2} & \cdots & w_{S,R} \end{bmatrix}$$

## a. Multi-Layer Neural Network

In order to maximize the benefits of parallelism for solving complicated problems and bring the response to be more like the human brain, cascading neuron layers and interconnecting the outputs of each layer as inputs for the following layer and using different kinds of activation functions will result in more powerful network.

The feedforward network shown in Figure 3.9 has three layers which are separated: One input layer, one output layer and one hidden processing layer, it is called hidden because of its invisible from the outside. In most networks the connections are linked to neurons of the following layer [50], but some of them allow shortcut connections that skip one or more levels and go directly towards the output layer.

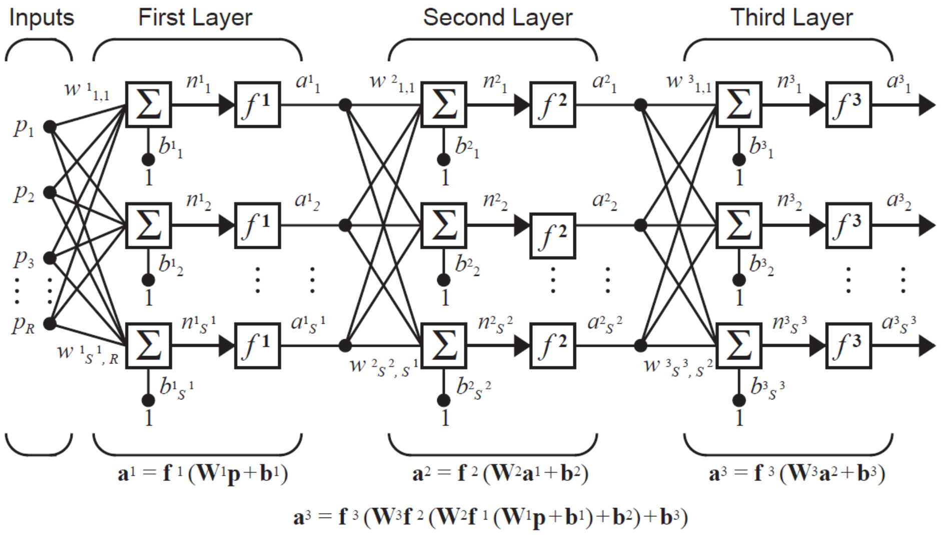


Figure 3.9 Three layers feedforward network [46]

Most practical neural networks have just two or three layers. However; four or more layers do exist, but they are rarely used in applications. In Figure 3.9 all the neurons have biases. The bias gives the network an extra variable, therefore the neurons with biases are expected to be more powerful than the one without biases, it is a technical trick to consider threshold values as connection weights [48][50].

Another important and powerful type of neural networks is the Recurrent Neural Network (RNN), recurrence is the process of a neuron influencing its status by a feedback or by any connection, recurrent networks do not always have strait defined input or output neurons, this is because some of its outputs are connected to its inputs. Depending on the feedback processing unit which is mostly a delayer or an integrator as shown in Figure 3.10, the recurrent network can be classified into direct, indirect or lateral, in direct recurrent network the neuron's output is fed back to itself with the weight assigned to this connection as $W_{nn}$ with a noticeable feature in this matrix ( all values below diagonal are zeros) [50], While with indirect recurrent network the feedback is connected to the first layer in the network and the weights matrix has the feature of symmetry, finally the lateral recurrent network is the network that allows feedback exclusively within the same layer.

Recurrent networks are potentially more powerful than feedforward networks and can exhibit temporal behaviour.

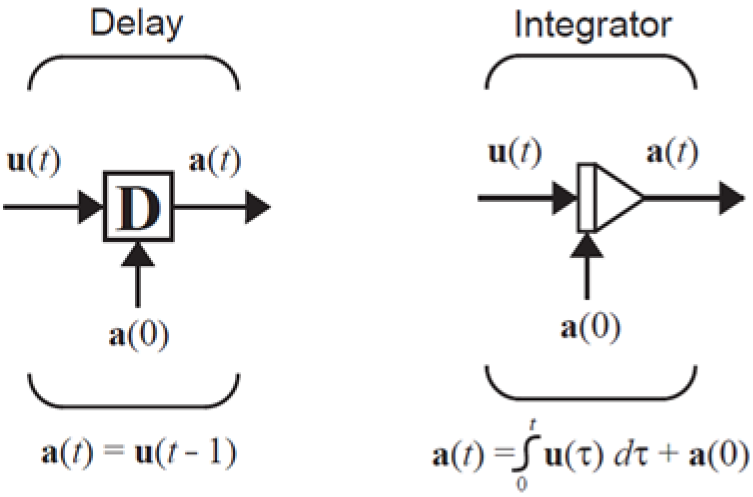


Figure 3.10 Delayer & integrator [46]

### 3.2.3 Training the Artificial Neural Networks

Training process of the ANN is defined as the process of adjusting and updating the weights of the connections between the neurons, training enables the ANN to extract information from the input data. Different learning schemes require different learning algorithms these training algorithms can be classified according to the assigned tasks into supervised and unsupervised learning. In supervised learning tasks, the objective of the training algorithm is to decrease the errors between the desired output data sets and actual output data sets [51]. This error can be expressed in the following formula:

$$E(W,b) = 0.5 \sum_t (\| Y(W,b,p) - Y \|_2) \qquad (3.3)$$

where ($t$) is the training data sets, $W$ is the weights matrices of the connections in the input and hidden layers, $b$ is the vector representing the bias, $p$ is the input matrix, $Y$ is the optimum output, and $Y(W,b,p)$ is the actual output which needs to be as close as possible to $Y$. and scaling the error by 2 is to simplify the differentiation [51].

The best learning algorithms that can fulfil the above requirements for supervised learning are the ones including gradient descent and particularly the backpropagation. However, there are many learning algorithms for supervised learning that uses the gradient descent in different ways as a basic building formula.

Minimizing the error E is achieved by updating the weights of each neuron by implementing the learning algorithm: for a neuron N, the error between the actual output and the desired output can be calculated using Equation (3.3).

$$E_j(W_j,b_j) = 0.5 \parallel Y_j\ (\ W_j, b_j\ , p_j) - Y_j\ ) \parallel_2 \qquad (3.4)$$

In order to minimize E every element of vector $W_j$ and vector $b_j$ is updated and can be calculated using gradient descent based algorithms :

$$W_{j,\,new} = \ W_{j,\,old}\ - \lambda \frac{\partial \mathrm{Ej}(Wj,bj)}{\partial Wj} \qquad (3.5)$$

$$b_{j\,new} = \ b_{j,\,old}\ - \lambda \frac{\partial \mathrm{Ej}(Wj,bj)}{\partial bj} \qquad (3.6)$$

where λ is the learning rate and is in the range of (0, 1.0). The iteration continues and both weight and bias values continue to be updated until the value of E for all the neurons is at minimum value [52], However this minimum value might not reach its best, this is because backpropagation algorithm converges to produce solution as long as its fed with training data sets which may not lead perfect results, on the other side if the size of the training set is too large then using backpropagation will be time consuming [52][53], in order to mitigate such a drawback many gradient descent base algorithms were have been proposed from which: stochastic gradient descent (SGD) algorithm [54] , mini batch gradient descent[55], momentum SGD [56], nesterov accelerated gradient[57], and Adagard [55]. Many more methods were proposed for different cases, in addition to that the size of the network can be shrunk using pruning techniques [58], the concept of pruning is excluding all neurons that do not affect the output of the ANN because they are not involved in the learning task, as a result of the pruning process ;the ANN will become faster, smaller, and more efficient.

Training RNN quite different and this is due to the architecture and the feedback connections between the neurons, therefore gradient descent algorithms such as backpropagation cannot be directly used, because the error backpropagation calculated using backpropagation algorithm cannot have feedback cycles in the connections between the neurons. Hence the backpropagation through time (BPTT) algorithm is one of the proposed and one of the most commonly used training algorithms for the RNNs [59], this approach simplifies the RNN and converts into a kind of feedforward ANN by simply unfold the network and process a group of links step by step, all feedbacks are fed forward to a copy of the original RNN and the process continue forth to the next copy. However, because of the feedback connections in RNNs, the BPTT algorithm may generate quite large number of sub-optimal results compared to the gradient descent algorithms used for training feedforward ANNs. Moreover, the gradient in BPTT is computed based on the complete training set, and if the size of the training data is large it will end up with relatively long training time just like backpropagation algorithm.

Real-Time Recurrent Learning (RTRL) is one of the learning schemes proposed to mitigate the drawbacks occurring when using the BPTT for training an RNN, RTRL computes the error gradient and can deliver it at any time, Unlike the BPTT that unfolds RNNs in time, the RTRL propagates error forward in time [60].

In RTRL, the weights (*W*) are update depending on the gradient value at time t and on the gradient value at the previous time, i.e. $W_{t+1} = W_t - \lambda \sum \frac{\partial \mathrm{E}(t)}{\partial Wt}$ , and because of that the RNN trained using RTRL doesn't need to be unfolded like in BPTT as the error vector is related to time and propagates forward only, (time never come back).

### a. Deep Learning

Any artificial neural network that includes multiple hidden layers in its architecture is considered a deep neural network DNN [61], Figure 3.11 shows the architecture of a DNN.

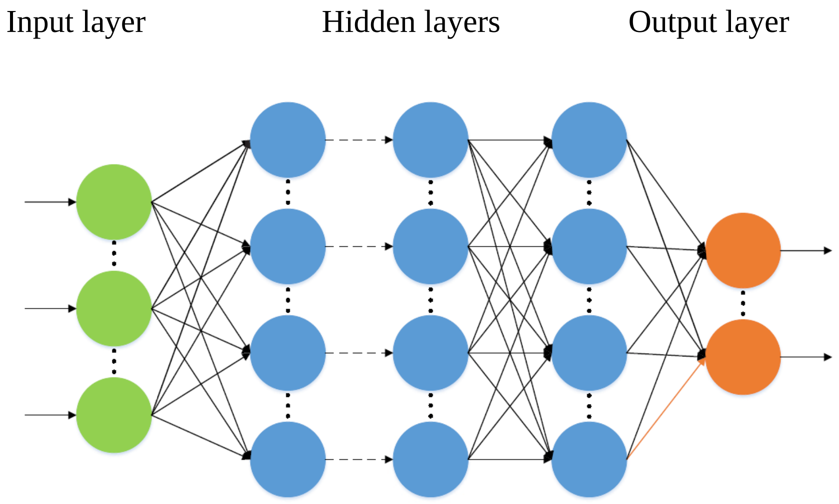


Figure 3.11: DNN architecture. [61]

Because of its architecture DNN models can deliver high level refined data, the data travels through multiple layers and is subject to several linear and non-linear transformations, and consequently better data relations are abstracted. There are several types of DNNs such as: deep RNNs, deep feedforward networks, deep convolutional networks, long-short term memory (LSTM), and deep Q-learning…etc [62][63].

The new emerging technologies with high computing capabilities, the availability of huge amount of data generated by the digital systems and the effective programming languages for writing and implementing learning algorithms, all opened the way towards utilizing DNN and deep learning in problems solving.

Unlike the traditional ANNs that have only one hidden layer, a DNN with multiple layers is more constructive due to the following reasons:

- For the same performance level; the number of neurons in a DNN required to solve a certain problem is far less than the neurons required to solve the same problem using shallow ANN, because the number of neurons in a shallow ANN is exponentially proportional with problem complexity.
- Complex task learning; Shallow ANN are effective in solving small scale problems, however they might turn impractical when used to solve complex problems, this is because these networks can learn quickly and memorize but are not good in generalizing the learned rules, thus the DNNs are more practical for many everyday life tasks that contain complex problems that require partitioning the target function into a chain or hierarchy of smaller functions to simplify and speed up the learning process.

As every working system, DNNs suffer from some drawbacks and have to mitigate some challenges; because of their high capacity and capability to process large number of parameters, the possibility of overfitting will increase. To overcome this glitch, several advanced regularization techniques have been designed, such as dataset augmentation, early stopping, and weight decay [53]. These methods alter the learning algorithm in a way so that the test error can be reduced but this will cause increasing the training error.

## b. Training Deep Neural Network

Training a DNN with a gradient based method will end with high instability in the error gradient; as explained earlier the connections weights' are updated according to the computed gradient of the error depending on their current values and the process is repeated until reaching the minimal, in each iteration, the gradient will pass via the weighted connections and hence the magnitude will be affected, and since the gradient is computed using chain rule. Therefore, multiplying the gradients at each layer will make the gradients exponentially decrease (for small gradient values) or increase (for large gradient values), these values depends on the number of the gradients and the layers of the DNN and are within range (-1, 1), respectively. This problem is not critical in the shallow ANN models because they contain only one hidden layer, these two problems are known as the exploding gradient problem and vanishing gradient and their effect on the DNN can cause the layers learning at different speed or levels i.e. last layers learn well while beginning layers learn very little. Several DNN learning algorithms were proposed to overcome this instability such as LSTM [64], adaptive learning rate algorithms [65][66][67], multi-level hierarchy [68], and residual networks ResNets [69].

## 3.3 Deep Learning for Mobile Networks

Data growth and user preference of using wireless connectivity drives the Internet Service Providers (ISPs) to involve intelligent systems and tools that can be used in the currently 4G LTE system(s) and migrate to the next generation of mobile systems (5G) to help manage the rise in data volumes and algorithm-driven applications and satisfy the end-user demands. Therefore, embedding machine intelligence into future mobile networks is being a point of interest for research and industry [70][71]. Most of the works are focusing on problem optimization using machine learning (ML), the criteria where these solutions are needed range from radio access network technology selection, developing the architecture of the core network and introduce machine learning technologies to cope with the existing technology.

In wireless system ML can be used to extract valuable data from the traffic and automatically discover the mutual relationship or connection between system components leading to better network optimization and faster response to the users requests, Due to the high data volumes in the network data mining and information abstraction is a hard job for human, even for the ones who designed and implement the intelligent system.

### 3.3.1 Related Work

Deep Learning is a very powerful tool in function approximation, for this reason it has been widely used in improving reinforcement learning and imitation learning, both approaches have a high impact in solving problems relating mobile network control that were considered hard to control or deal with, and complex: the difference between these two approaches is while the admin in the reinforcement learning is in direct contact with the environment to learn the best action, with continuous probing and analysis, the agent will learn how to maximize its gain, while in the imitation learning , the agent is not in direct contact with the environment and the learning is achieved via a demonstrating entity that teaches the agent how to respond with a suitable action to a specific case, after sufficient teaching , the agent will learn how to imitate the demonstrator, copy its behaviour and can work without supervision [72], Figure 3.12 shows the block diagram and data flow of both approaches.

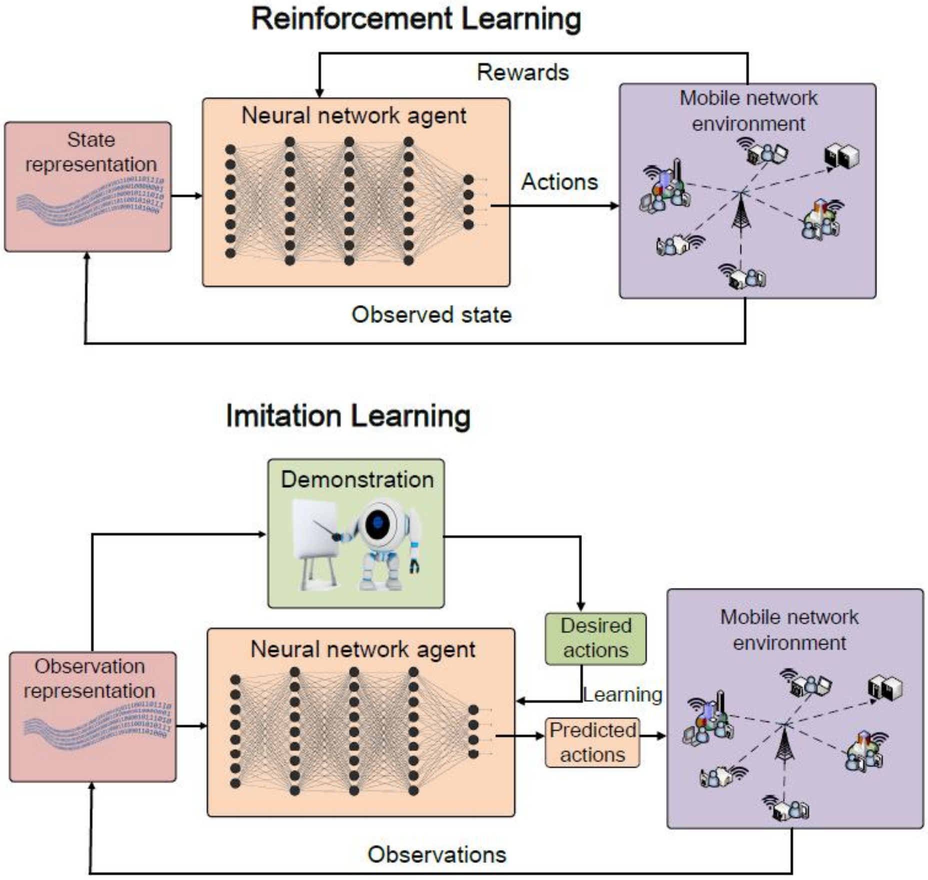


Figure 3.12: Reinforcement and imitation learning block diagram.

Most of the research considered in this field: Radio Control, Resource allocation, Scheduling, Routing and Network optimization. The authors in [73] considered the problem of dynamic spectrum access for network utility maximization in multichannel wireless networks, and proposed an algorithm called Deep Q-learning for Spectrum Access (DQSA), enabling each user to learn good policies in an online and distributed manner, the experimental results showed strong performance of the algorithm in complex multi-user scenarios. The authors in [74] discussed the case when small cells are deployed, they proposed a new inter-cell interference management (IIM) scheme for small cell networks with power control using NN. According to the results the NN system delivered almost the same performance as that of the ideal scheme and superior to that of the belief propagation, especially in MIMO environments. Same authors presented a similar work in [75].

In [76] the authors used RF measurements to study the suitability of deep recurrent neural networks for band selection in land mobile radio (LMR) bands. The results showed that RNN was able to improve the performance of spectrum sharing in dynamic wireless environments. Nguyen Cong Luong *et.al.* presented in [77] some IoT service improvement using cognitive radio technique, they proposed using blockchain with mining pool to achieve that, they used deep reinforcement learning algorithm to derive an optimal transaction transmission policy for the secondary user, according to their simulation results; it is shown that the proposed deep reinforcement learning algorithm perform better than the conventional Q-learning scheme in terms of gain and learning speed.

Working in resource allocation criteria, in [78] H. Sun *et al.* proposed a learning-based approach for wireless resource management, the algorithm considers its input and output as unknown non-linear mapping parameters and to use a deep neural network (DNN) to approximate it, they demonstrated the DNN performance using extensive numerical simulations for approximating two complex algorithms designed for power allocation in wireless transmit signal design, while giving orders of magnitude speed increase in computational performance.

In [79], the authors proposed a resource allocation framework for collaboration between LTE-LAA and Wi-Fi in the unlicensed spectrum, and developed a deep learning algorithm based on LSTM, in this algorithm taught each Small Base Station (SBS) to select its spectrum allocation scheme independently, according to their simulation results, the proposed scheme demonstrated better performance compared to the conventional methods that consider network fairness Simulation results have shown that the proposed approach yields significant performance gains in terms of rate compared to conventional approaches that considers only instantaneous network parameters such as instantaneous equal weighted fairness, proportional fairness and total network throughput maximization. Results have also shown that the proposed scheme is more stable regarding Wi-Fi connectivity in the case large number of LTE-LAA is deployed in the unlicensed spectrum.

In [80] the authors discussed  solutions for the under development next generation mobile network they considered the limitations of dynamic TDD-based resource assignment in a 5G Ultra dense network UDN when massive MIMO with beamforming capabilities is fitted in the eNodeB, the research addressed the vulnerability to congestions when conventional traffic control strategy is implemented. The deep LSTM algorithm aims to avoid the congestion events by predicting them.

In [81] Xu *et al.* designed a framework for power-efficient resource allocation in cloud RANs that applies Deep Neural Network (DNN) to approximate the action-value function. According to the simulation results which show that the framework can achieve significant power savings while meeting user demands, and it can well handle highly dynamic cases. The authors of [82] considered the build of intelligent transport system, addressing both safety and Quality-of-Service (QoS) as concerns in a green Vehicle-to-Infrastructure communication scenario, they presented a deep reinforcement learning model, that learns an energy-efficient scheduling policy from inputs corresponding to the specifications and requirements of vehicles running within a RoadSide Unit's (RSU) coverage .the results listed a comparison of the proposed algorithm and three other algorithms: RVS: Random Vehicle Selection algorithm, LRT: Least Residual Time algorithm and GPC: Greedy Power Conservation algorithm, the proposed algorithm showed better performance.

In [83] the authors worked in MEC criteria and cared of user scheduling, they adopted a model with small cells deployment and proposed a deep RL algorithm to optimize the probabilistic policy and minimize the average transmission delay, with Boltzmann distribution rule used as the parameterized policy to generate probabilistic actions, and the gradient ascent method to update the parameters. According to their simulation results show the advantage of the proposed solution.

The collaboration between different wireless networks sharing the same spectrum and handling large number of devices is and interest of the authors in [84], they presented two ANN algorithms to predict free slots in a Multiple Frequencies Time Division Multiple Access (MF-TDMA) network using function approximation, these algorithms use a low dimensional NN to predict the probabilities of using a slot in the next frame based on spectrum observation. The simulation results showed that using the proposed approach reduced the collisions between the networks by 50% compared to the case when using the traditional non-collaborative scheduler.

Continuous Hopfield Neural Network (CHNN) is used in [85] to seek an optimal route, which can improve the utilization and survivability of MANET, compared to the same network; but using of Ad hoc On-demand Distance Vector (*AODV*) routing protocol, the simulation results show that CHNN AODV can perform better compared to AODV in terms of average delay and successful packet, however; adding CHNN will increase the power consumption which is a critical factor in Ad-hoc networks.

The authors of [86] explained the router architectures. They reviewed the current Software Defined Router (SDR) architectures and suggested using a supervised deep learning to compute the routing paths instead of the conventional routing protocol in order to enhance the traffic control in backbone network. The simulation result show that the proposed deep learning based routing strategy exceeded the conventional OSPF in terms of the overall throughput and end-end delay per hop. Moreover, the proposed routing strategy was analysed to prove that the GPU-accelerated SDR better to run the proposed algorithms than the CPU-based SDR. Same authors presented [87] in same field, a smart packet routing scheme using Tensor-based Deep Belief Architectures (TDBAs) that learn from the network traffic and Kpi , they used the tensors to perform weights , biases and the units in all the layers, the proposed TDBAs was trained to predict the best routes for every edge router. According to the simulation results the proposed TDBA algorithm performs better than the conventional Open Shortest Path First (OSPF) protocol in terms of packet loss rate and average delay per hop the network experiences a high traffic load.

The authors in [88] worked on caching and interference alignment (IA), unlike the ideal models assumed in most researches, they considered realistic time-varying channels, especially the channel that is formulated as a finite-state Markov channel (FSMC), which forms a highly complex system. And propose a big data deep reinforcement learning approach to obtain the optimal policy for user selection in cache-enabled wireless networks, according to the simulation results show that the performance of cache-enabled IA networks has been improved by using the proposed big data reinforcement learning approach. Same authors presented [89] and considered same model but used Google Tensor Flow to implement deep reinforcement learning in order to obtain the best IA user selection policy in the cache-enabled wireless networks.

Finally, in [90] the authors considered handover (HO) as a core for the work, they aimed to decrease the HO rate but at guaranteed system throughput, they developed an asynchronous deep reinforcement learning scheme to control the handover (HO) process across multiple (UEs), supervised learning was used in initializing the DNN, simulation results demonstrate that the proposed framework can achieve better performance than the traditional schemes, in terms of HO rates, and the adopted framework could train faster when the number of UEs is increased, which a positive point supporting the scalability issue and suitability for large networks.

## 3.4 Summary

Neural networks are very good for solving variety of problems by finding trends in large quantities of data; they are better solver to problems which humans are good at than traditional computer, such as image recognition, approximation, prediction…etc.

Most ANN applications are built using computational entities and perform the propagation of continuous variables from one processing unit to the next. Compared biological neural networks which communicate through electrical pulses, both use the timing of the pulses to send information and perform computation. This realization has created research fields on neural networks, including theoretical analyses and model development.

Neurons are outlined in the form of layers; input, hidden and output layers and are trained using different learning schemes and algorithms. Well-trained networks are able to classify correctly patterns unseen during training. If this does not occur, then the net is denoted as over-fitted the decision plane and does not generalize well.

# CHAPTER 4

# SMALL CELL DEPLOYMENT FOR CONTENT DELIVERY ARCHITECTURE AND CHALLENGES

### Briefing

The first part of this Chapter explains the deployment of small cells in an LTE framework simulated using OPNET modeller and how dual connectivity is implemented to mitigate challenges in the coordinated multipoint CoMP and carrier aggregation CA. The second part explains the benefits of adding caching units to enhance content delivery process for the network users.

## 4.1 Introduction

All the improvements and evolutions on the cellular mobile networks since the beginning aimed on providing better services within shorter time, i.e. high throughput and low latency. The current 4G mobile system so far had successfully support the users and covered all their requirements and demands , but with the new technologies and increasing capabilities of the mobile devices, launched applications, IT integration in telecommunication, all these factors generated and injected big amounts of data to the network, on the other hand the way the people are using their smart phones and the time spent using them for whatever reason, chatting , watching videos, online gaming …etc. had increased the burden and made the network operators develop new strategies to guarantee sustainability and survivability.

In this aspect small cells were one of the solutions applied by developers to meet the increasing demand for higher data rates. Small cells can be considered as low power hotspots and the macrocell as a parasol shading them, the UE can simultaneously connect to the Sc and MeNodeB using dual connectivity technique specified in 3GPP TR 36.842 [91]. This technology had mended several challenges when deployed such as: lack of radio resource, Mobility management, Signalling overhead.

Dual connectivity adds capacity for the UEs at the cell edge because they can dynamically adapt and choose the best radio resources among several cells i.e. make a handover. In heterogeneous networks frequent handovers might not be completed successfully which will result in service interruption, the LTE DC can eliminate the handover failures because the UE maintains the connection with MeNodeB as the coverage layer.

## 4.2 Architecture

In LTE DC, the UE can receive/transmit data from/to multiple eNodeBs. There is a Main eNodeB (MeNB) and one or more Small cells (SeNodeB), only the case of one MeNodeB and one SeNodeB is considered in 3GPP Release 12 specifications [91]. In order to simplify the architecture and its comprehension, it's separated into control plane and user plane.

### 4.2.1 Control Plane Architecture with SeNB

In the case of handover between SeNodeBs and in order to reduce the control signalling, it was agreed to assign only one S1-MME connection for each UE and is established between the MeNodeB and the core network, and the RRC of the UE is connected to the RRC of the MeNodeB since there is no RRC block in the SeNodeB this procedure will help by not adding extra signalling or increase complexity. The control plane architecture is shown in Figure 4.1

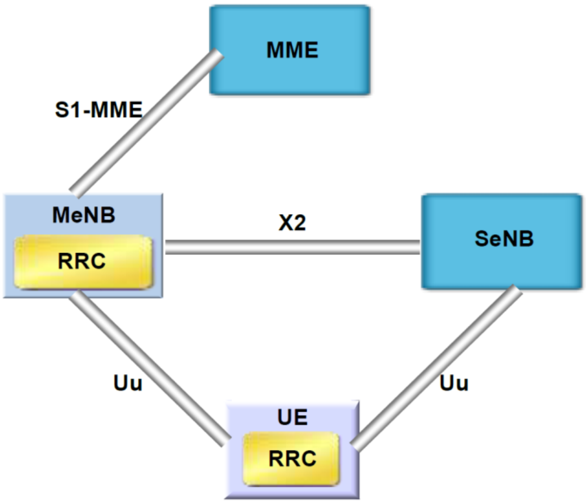


Figure 4.1: Control plane architecture [91].

### 4.2.2 User Plane Architecture

3GPP has defined two cases for user plane architecture; the traffic can either split at the MeNodeB or at the SGW. If the bearer level split takes place in the SGW the packets are sent via two S1 bearers to both MeNodeB and SeNodeB, and the control signalling is exchanged via X2 interface. This architecture is denoted as UP 1A, as shown in Figure 4.2.

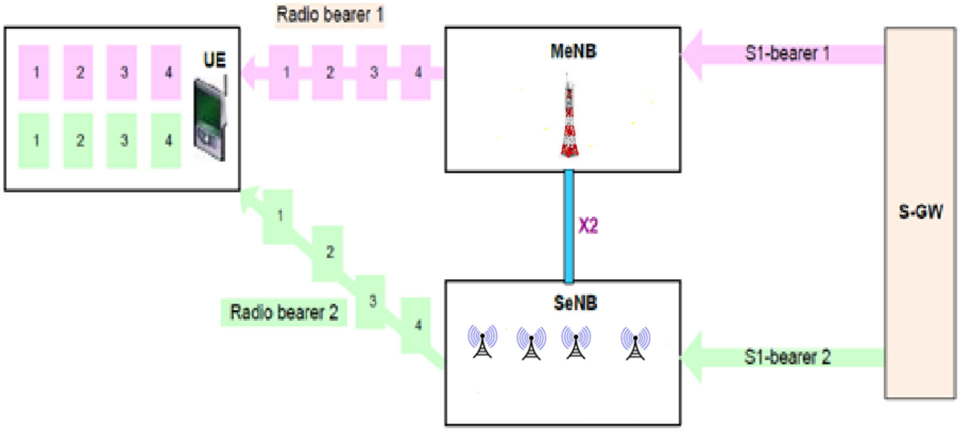


Figure 4.2: UP 1A Bearer level split at SGW [91].

The advantages of such architecture is the buffer independency so the MeNodeB is handling the traffic for Radio bearer 2, another advantage is only basic specifications are required for the link between SeNodeB and MeNodeB as it will carry no traffic. The drawbacks of this technique are that SeNodeB is visible to the core network, and additional overhead is required where coding and ciphering is duplicated to enable security in both MeNodeB and SeNodeB.

Alternatively, if the split takes place at the MeNodeB, it can be either bearer level or packet level split. For the packet level split, user data may be routed between MeNodeB or SeNodeB and the UE and can be split as IP packets. While for the bearer level split, all user data of a radio bearer are routed between the SeNodeB and the UE, the architecture is denoted as UP 3C and is illustrated in Figure 4.3

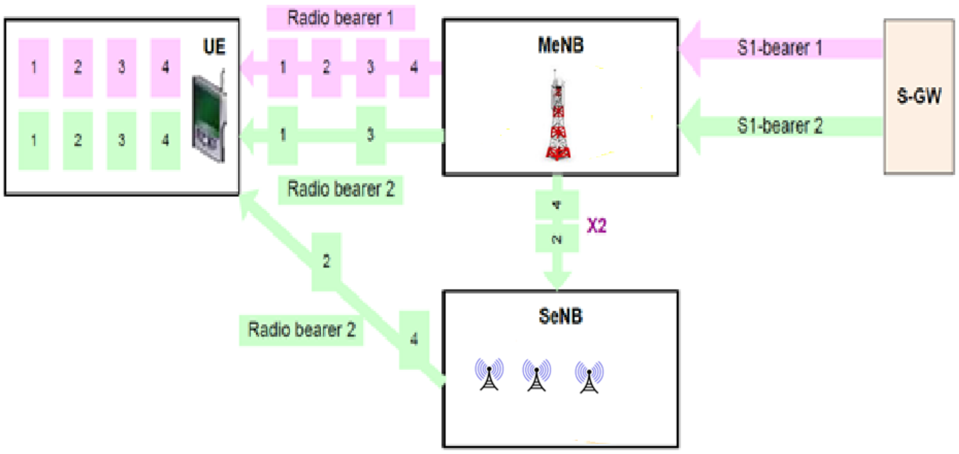


Figure 4.3: UP 3C Packet level split at MeNodeB [91].

In contrast to alternative 1A the 3C the following advantage points: the SeNodeB is hidden from the core network, all the security steps and ciphering are carried out in the MeNodeB only, but on the other hand, it has the following considerable disadvantages: where all data needs to be buffered at the MeNodeB and then routed to the SeNodeB, the second is the need for flow controller between the MeNodeB and SeNodeB to handle that traffic.

### 4.2.3 SeNodeB Commissioning

Adding SeNodeB(s) under the coverage of a MeNodeB is done following these steps:

- MeNodeB broadcasts the Small Cells Group (SCG) addition indication message containing the configuration
- SeNodeB responds with SCG addition/ modification request message containing radio resource configuration.
- MeNodeB sends RRC connection reconfiguration message to the UE, with configuration of both MeNodeB and SeNodeB
- UE sets the new configuration and sends acknowledgment RRC connection reconfiguration complete to the MeNodeB
- MeNodeB forwards RRC connection reconfiguration complete message to SeNodeB

The steps are illustrated in Figure 4.4

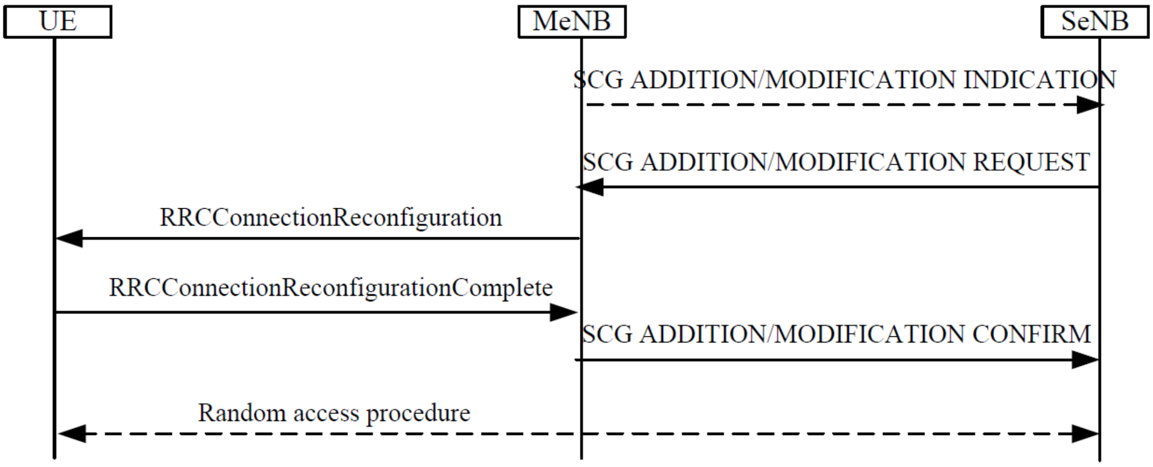

UE
MeNB
SeNB
SCG ADDITION/MODIFICATION INDICATION
SCG ADDITION/MODIFICATION REQUEST
RRCConnectionReconfiguration
RRCConnectionReconfigurationComplete
SCG ADDITION/MODIFICATION CONFIRM
Random access procedure

Figure 4.4: SeNodeB addition procedure [91].

## 4.3 Model Design

The architecture is based on the 3GPP LTE 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 [91] which was previously explained in section 2.3.4 and Figure 2.14. This architecture was also the main model in publications [92] and [93]. In this scenario the deployment of the macro and small cells are on different carrier frequencies (inter-frequency) where the Small Cells (SC) will be distributed as hot spots covering specific areas under the coverage of the macro-cell layer. The small cell 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 small cells layer. The macro and small cells layers are assumed to be connected via an ideal backhaul. Scenario #2 of Release 12 is mapped in Figure 4.5.

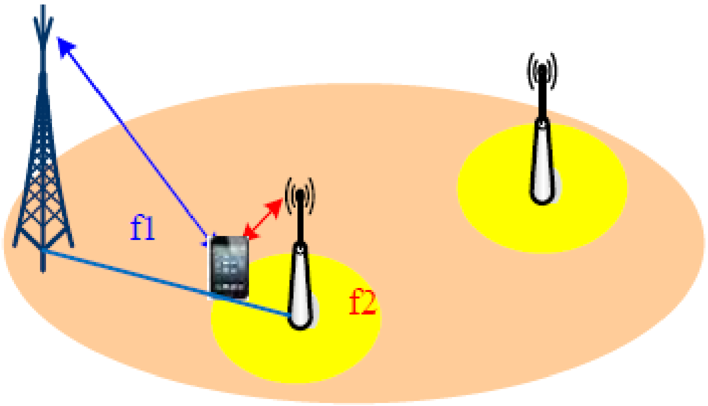


Figure 4.5: Scenario #2 of Release 12 3GPP [91].

In such a scenario, bringing the computation and storage capacity units from the core to the edge of the network with dense deployment of low-power small-cell nodes in which the distance between the radio access points (RAP) and terminals is reduced, virtualization and centralized processing would improve the throughput and maintain the low latency without adding any additional overhead [92][93].

In this work, the concept of adding computing and storing capacity to the main eNodeB is considered; 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 fibre cables connection, as in the C-RAN architecture. In this way the resources of several cells can be pooled in one centralized entity [92]. In LTE network, resource allocation takes place at the level of cells, and scheduling of the resource units called Resource Block (RB) takes place every Transmission Time Intervals (TTIs). A UE is associated to a cell, and transmission of neighbouring cells on the same RBs count as interference, interference-prone transmission imply lower Signal to Noise and Interference Ratio (SINR), leading to more RBs being used to transmit the same payload, this obviously reduces the capacity of the network, allowing fewer UEs to be served simultaneously, and affecting the quality of service being introduced to the end user, on the other hand, it will negatively affects the energy efficiency, which also depends on the number of bits per RB. [93][94].

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. Figure 4.6 shows the system diagram.

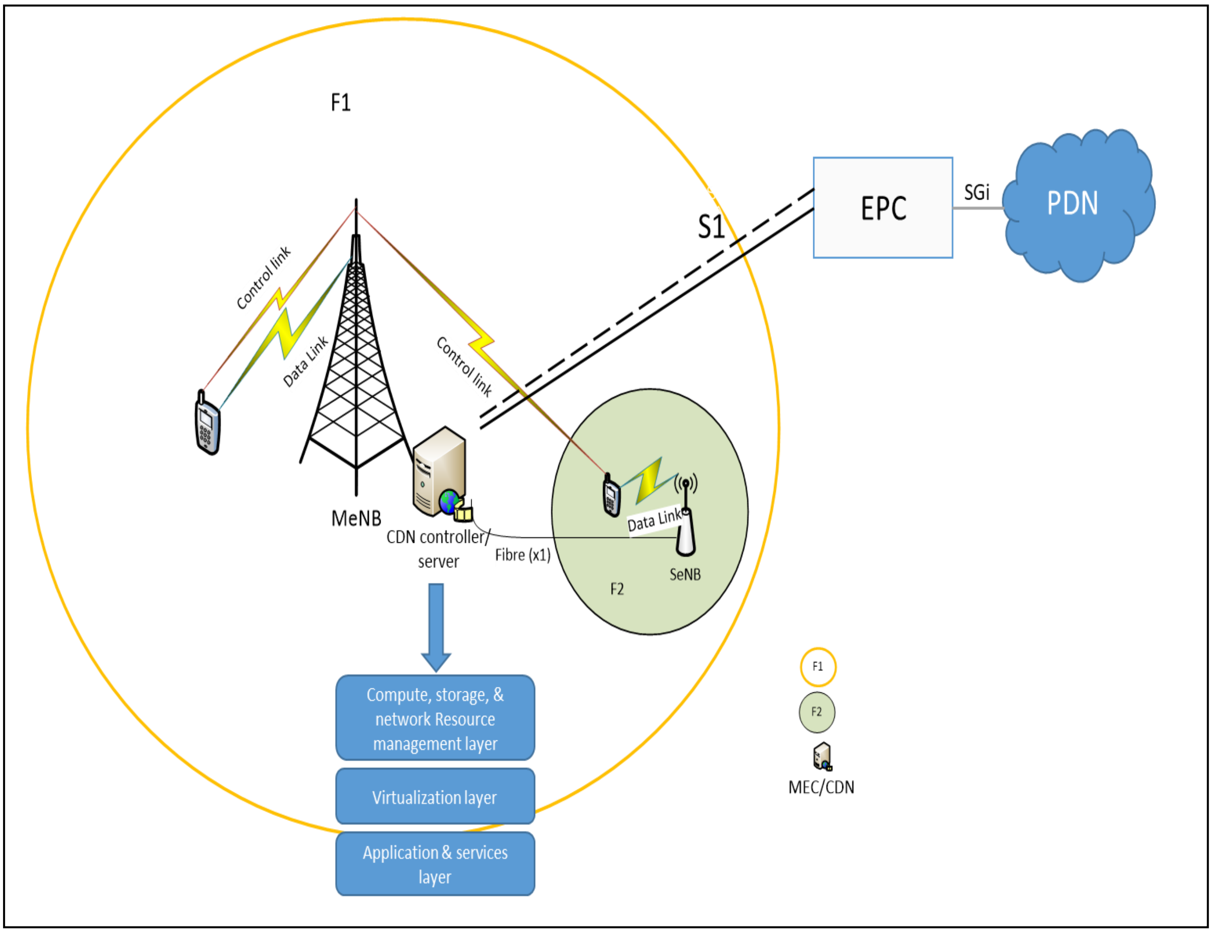


Figure 4.6 proposed system diagram [92]

In this architecture, the UE in the proposed form of dual connectivity maintain a normal connection with MeNodeB and will establish a U-Plane connection with a SeNodeB 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 (Mobile Edge Computing) server could be used for this purpose), which can add computing capabilities for the radio access networks (RATs) or could be used as an aggregation point in the IP transport layer.

As in LTE Release 12 specifications [91], LTE small cell enhancement by dual connectivity is defined as a technology which extends carrier aggregation (CA) and coordinated multi-point (CoMP), in which the small cells are typically deployed as hotspots within macro cell coverage, where any UE has the ability to receive/send data from/to two or more eNodeBs simultaneously. Some of the expected benefits from such enhancement are:

- Rising UE throughput for cell edge UEs in particular.

- Reduce the overhead occurred from signalling towards the core requesting handovers.

Information exchanged between the MeNodeB server and UE may took place on different layers, such as MAC, PDCP and RRC layers. A UE in RRC – connected mode first obtain access to the MeNodeB and keep C-plane connection with this node, which is the only RAT 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 big size 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. The flowing steps, could explain such a procedure:

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

Step 2: content is requested by a UE

Step 3: MeNodeB decides which node of the SCs will the UE 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 node of small cell decided by the controller at the MeNodeB.

Figure 4.7 shows the procedure in the form of flowchart

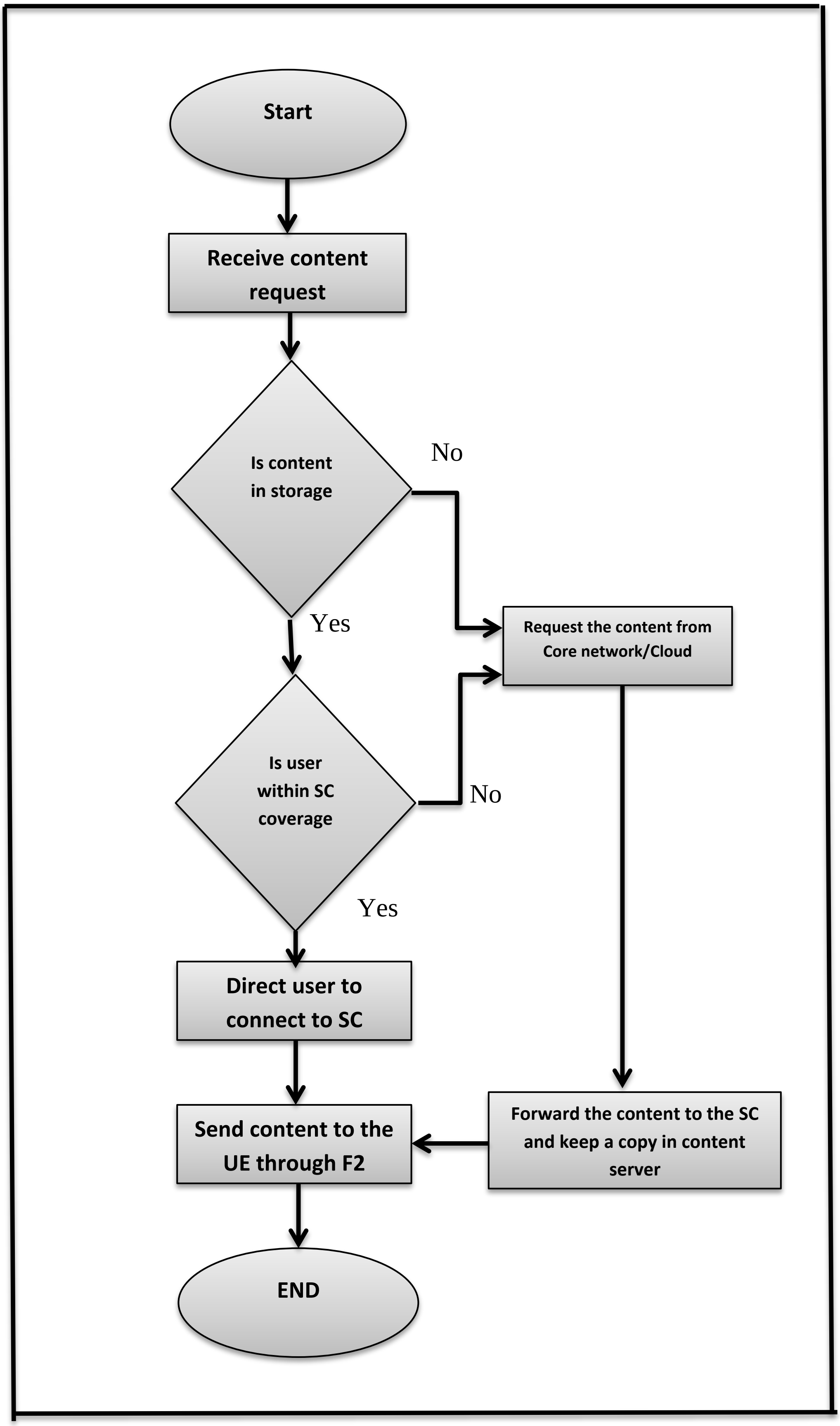

Start
Receive content request
Is content in storage
No
Yes
Request the content from Core network/Cloud
Is user within SC coverage
No
Yes
Direct user to connect to SC
Send content to the UE through F2
Forward the content to the SC and keep a copy in content server
END

Figure 4.7: Content delivery procedure flowchart.

### 4.3.1 Cache Modelling

The idea of caching frequent information in a nearby storage has been introduced since the beginning of computer operating systems [95], the term cache refers to a memory with fast access but limited storage, caching was utilised for the internet ,when internet became more popular and easy to access. Popular webpages were saved in small servers (caches) instead of retrieving them from a central server and this significantly reduced access time as the distance between the user and the requested data had decrease and also reduced central server congestion, and saved bandwidth to use it to respond for different demands [96].

In wireless networks; the challenges are seen from two standpoints: The Delay and the Bandwidth (hence the throughput), the second scenario shows that caching at eNodeB can lead to many benefits for both mobile operators and end users:

- Delivery cost (Scenario1) caching in the eNodeB

  Selecting N to be the number of eNodeBs in the network, $n_j$ be the number of data requests received by the $j_{th}$ eNodeB and P be the mean size (bytes) of a requested object [97]. We also denote the cost components to be as follows:

  - U is the cost per byte from UE to eNodeB.
  - C is the backhaul link cost per byte from eNodeB to PGW.
  - W is the transit cost per byte between core network and the content provider.

If no caching is provided in the cellular network, each request will incur the costs U+C+W, and the total costs for complying all the requests in the network can be calculated as follows:

$$M_{nocache}=\sum_{j=1} (n_j) \times (U+C+W) \times P \qquad (4.1)$$

In the case of adding cache to the eNodeB the following parameters are consider for the equation:

(i) $n_{jc}$ be the number of requests for objects that are cached at the $j_{th}$ eNodeB

(ii) K is the additional cost per byte of caching objects at the eNodeBs (Server, storage cost, etc.). With the existence of eNodeB caching, the cost for the mobile operator to serve the requests will be the sum of the following parameters:

$$M1=\sum_{j=1} (n_j - n_{jc}) \times (U+C+W) \times P \qquad (4.2)$$

$$M2=\sum_{j=1} n_{jc}\times U \times P \qquad (4.3)$$

$$M3=\sum_{j=1} n_{jc}\times K \times P \qquad (4.4)$$

where M1 is the cost when requested objects need to be fetched from the origin server, M2 is the cost incurred on the network path from the UEs to the caching eNodeB, and M3 is the cost when adding cache to the eNodeB. So $M_{cache}$ can be calculated by adding the three parameters M1, M2 and M3 as follows:

$$M_{cache}= M_1+ M_{2+} M_3$$

$$M_{cache}= \sum_{j=1} n_{jc}\times U \times P + \sum_{j=1} (n_j - n_{jc})* (U+C+W) \times P +\sum_{j=1} n_{jc}\times K \times P \qquad (4.5)$$

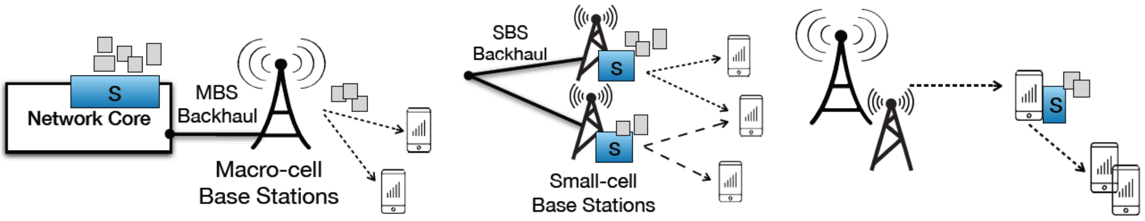
By subtracting (5) from (1) this will result in the benefit we get from adding cache to the network and as follows:

$$M_{Benefit}=\sum_{j=1} n_{jc}\times (C+W-K) \times P \qquad (4.6)$$

- In-network cache (Scenario 2)
  The new trend of in-network caching can achieve additional delay reduction for end users [95]. However, the advantage of in-network cache might be a disadvantage if the in-network cache cannot be efficiently utilized by the main eNodeB. An example of this is the situation when the cached content is duplicated in both the eNodeB and the in-network caches, which most likely occur in the case of full load.

- Device to Device (D2D) caching (Scenario 3)
  Caching content at the edge of a wireless networks using the (UE) is different from caching techniques in CDN and had incurred many challenges, such as caching decisions are coupled, security, and power management [96] [98], However; this scenario is out of the scope of this work due to the above mentioned drawbacks and the difficulties associated with simulating cache units inside the UE using OPNET.

Figure 4.8 illustrates all possible caching scenarios.



In Network Caching eNodeB Caching D2D caching

Figure 4.8: Caching deployment in LTE network [95].

### 4.3.2 Gain Analysis with Dual Connectivity

Gain is defined as a proportional value that shows the relationship between the magnitudes of the input to the magnitude of the output signal at steady state [93]. Gain can be enhanced by using changing parameters, adding circuitry, or adopting new schemes, increasing the gain obtained from any system or media is the goal of users and operators, introducing DC to the network and provisioning the layers to Macro and Small, first it is calculated theoretically according to Shannon-Hartley equation as follow:

$$C_i = B_i \log_2 (1+ SNR_i ) \qquad (4.6)$$

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 Small and Macro, i.e. $im = \arg\max (Ci)$ for $i\epsilon\ M$ and $is = \arg\max (Ci)$ for $i\epsilon\ S$.

When there is no DC, the user is considered to be served by one cell all the time, when introducing DC, the users are assumed to receive data from the small cells and the controls from the macro cell. The serving cells are selected according to their performance. A cell either Macro or Small is picked from a list of candidate cells if it has the best throughput estimation. The Shannon capacity equation for a user with DC is:

$$C_{DC} = C_{im} + C_{is}$$

The user throughput gain with DC will be

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

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

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

If the same bandwidth is set in both two layers (i.e. Bq/Bp =1) it will explicitly show that introducing DC delivers its most benefit when the users are exposed to the same link 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.

## 4.4 Model Simulation

The proposed cache enabled dual connectivity architecture that integrate the SCs to the LTE MeNodeB at the PDCP layer, is implemented using Riverbed 18.5 Modeller based on the 3GPP technical requirement for small cell enhancement [24]. The system model is then investigated in terms of throughput and delay.

The Riverbed (formerly known as OPNET) is a powerful simulation software, offers libraries that contains more than 400 ‘out of the box’ protocols and vendor device models including TCP/UDP, IPv6, VoIP/Video/FTP/HTTP/Email, WiMAX, WLAN (a/b/g/n), and LTE to support accurate event driven simulation scenarios. Nevertheless, The LTE model features supported by this modeller are based on 3GPP Release 8 & 9, that don’t support dual connectivity. 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. Furthermore, visualization functionality is also not supported but there is big number of devices belonging to the computer networking category and they can be used instead to simulate the network using the program capabilities to assign tasks and choose the statistics of each node. A measurement entity is also 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.

### 4.4.1 LTE Implementation in OPNET

Communication networks and distributed systems typically encompass a wide range of technologies ranging from low-level communications hardware to high-level decision-making software. A successful system modelling must represent each of these subsystems and their interactions at a level of detail that is sufficient to obtain valid predictions of performance and behaviour. Because the nature of these subsystems varies significantly from level to level, the traditional single level frame work does not meet these expectations, hence the need for a multi-tier system becomes a requirement. Any wireless system modelled using OPNET contains the following three domains:

- The Network Domain: concerned with the specification of a system in terms of high-level devices called nodes, and communication links between them
- The Node Domain: concerned with the specification of node capability in terms of applications, processing, queueing, and communications interfaces.
- The Process Domain: concerned with the specification of behaviour for the processes that operate within the nodes of the system. Fully general decision making processes and algorithms can be created.

The basic modules for building an LTE framework using OPNET are: UE, eNodeB, and an EPC, these modules can be duplicated, altered or combined to perform the required system functionalities, Figure 4.9 shows the basic units simulated using OPNET at module level.

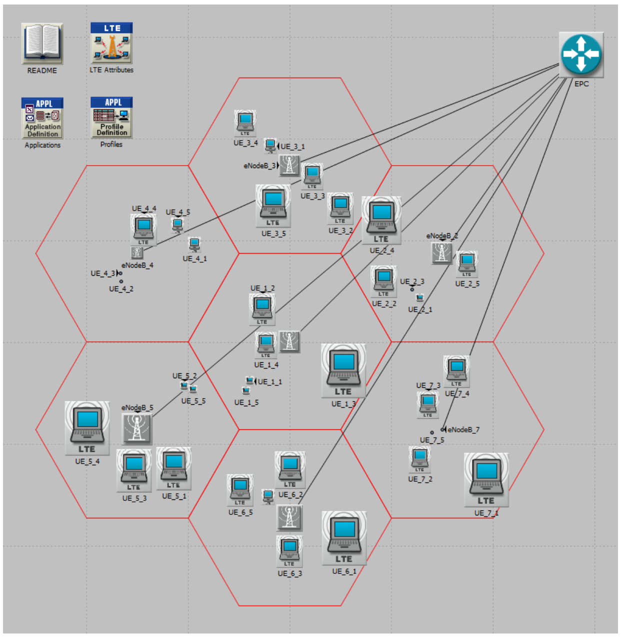


Figure 4.9 LTE network using OPNET

### 4.4.2 Dual Connectivity for UE Terminal

OPNET 18.5 modeller contains many ready built modules to use for creating a network, 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. Figure 4.10 shows the protocol stack layers of the UE without DC and its equivalent in OPNET node domain.

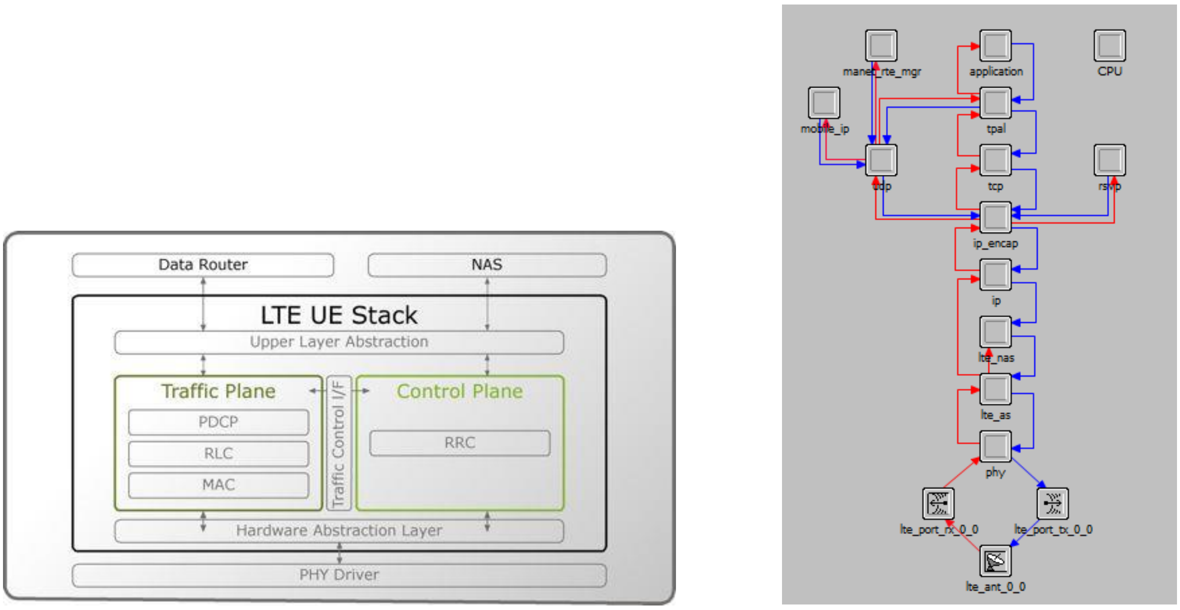


Figure 4.10: UE node level and equivalent protocol stack.

Because the dual connectivity allows a UE to have two simultaneous connections to a main eNodeB (MeNodeB) of macro cells and a secondary eNodeB (SeNodeB) of small cells while exchange of 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 shown in Figure 4.11.

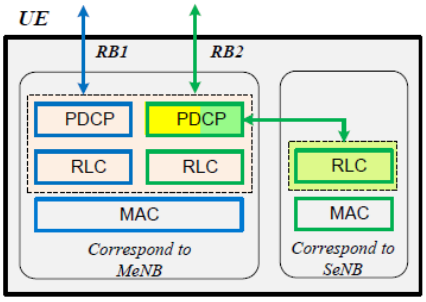


Figure 4.11: UE layers supporting DC (UP 3C) [26].

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 [92] [93].

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's DC) which has limited functionality compared to the original one, as it has only the PDCP and RLC layers.

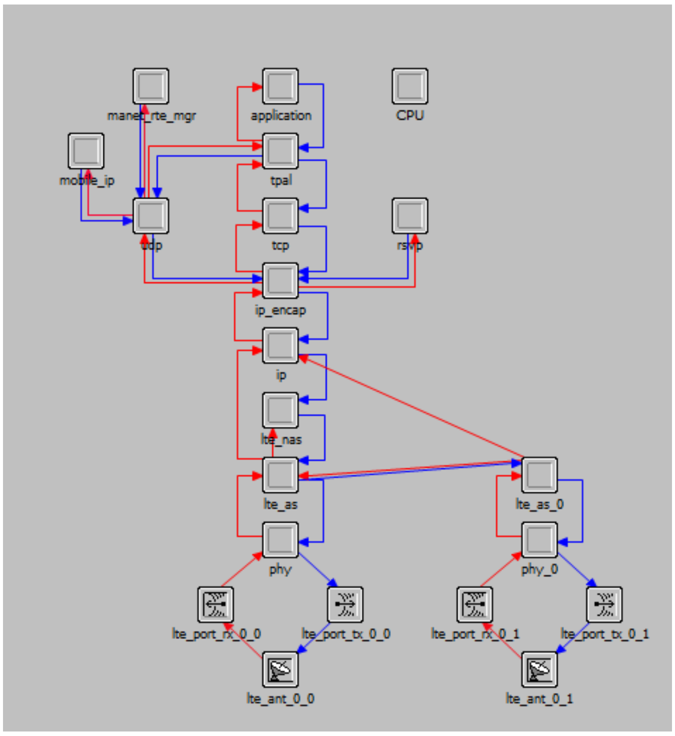


Figure 4.12: Modified UE for DC.

The process of AS protocol will be done only by the original protocol through the attached procedure as follows:

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 shows the procedure in process domain.

After this procedure the UE is 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, In the Idle mode the UE is just paged for the downlink data while in the connected mode, and 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.

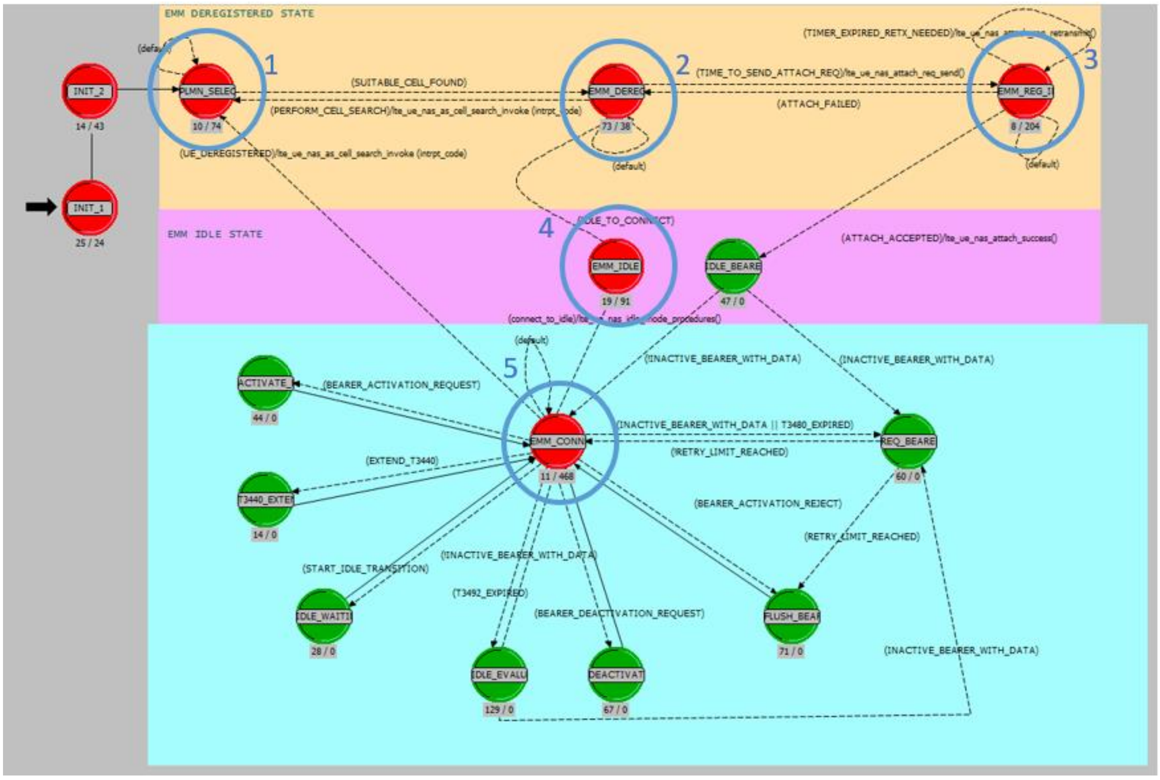

EMM DEREGISTERED STATE
(default)
INIT_2
14 / 43
INIT_1
25 / 24
1
PLMN_SELEC
10 / 74
(SUITABLE_CELL_FOUND)
(PERFORM_CELL_SEARCH)/lte_ue_nas_as_cell_search_invoke (intrpt_code)
(UE_DEREGISTERED)/lte_ue_nas_as_cell_search_invoke (intrpt_code)
2
EMM_DEREG
73 / 38
(default)
(TIME_TO_SEND_ATTACH_REQ)/lte_ue_nas_attach_req_send()
(ATTACH_FAILED)
(TIMER_EXPIRED_RETX_NEEDED)/lte_ue_nas_attach_req_retransmit()
3
EMM_REG_I
8 / 204
(default)
EMM IDLE STATE
4
(IDLE_TO_CONNECT)
EMM_IDLE
19 / 91
IDLE_BEARE
47 / 0
(ATTACH_ACCEPTED)/lte_ue_nas_attach_success()
(connect_to_idle)/lte_ue_nas_idle_mode_procedures()
(default)
5
(!INACTIVE_BEARER_WITH_DATA)
(INACTIVE_BEARER_WITH_DATA)
ACTIVATE_
44 / 0
(BEARER_ACTIVATION_REQUEST)
EMM_CONN
11 / 468
(INACTIVE_BEARER_WITH_DATA || T3480_EXPIRED)
(!RETRY_LIMIT_REACHED)
REQ_BEARE
60 / 0
(EXTEND_T3440)
T3440_EXTEN
14 / 0
(BEARER_ACTIVATION_REJECT)
(RETRY_LIMIT_REACHED)
(!INACTIVE_BEARER_WITH_DATA)
(START_IDLE_TRANSITION)
(T3492_EXPIRED)
(BEARER_DEACTIVATION_REQUEST)
IDLE_WAITI
28 / 0
IDLE_EVALU
129 / 0
DEACTIVAT
67 / 0
FLUSH_BEAR
71 / 0
(INACTIVE_BEARER_WITH_DATA)

Figure 4.13: Attach procedure in process domain.

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 as an MME to the SCs. In similar functionality to the HeNodeB [100] with some modification made to support Dual Connectivity.

Figure 4.14 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.

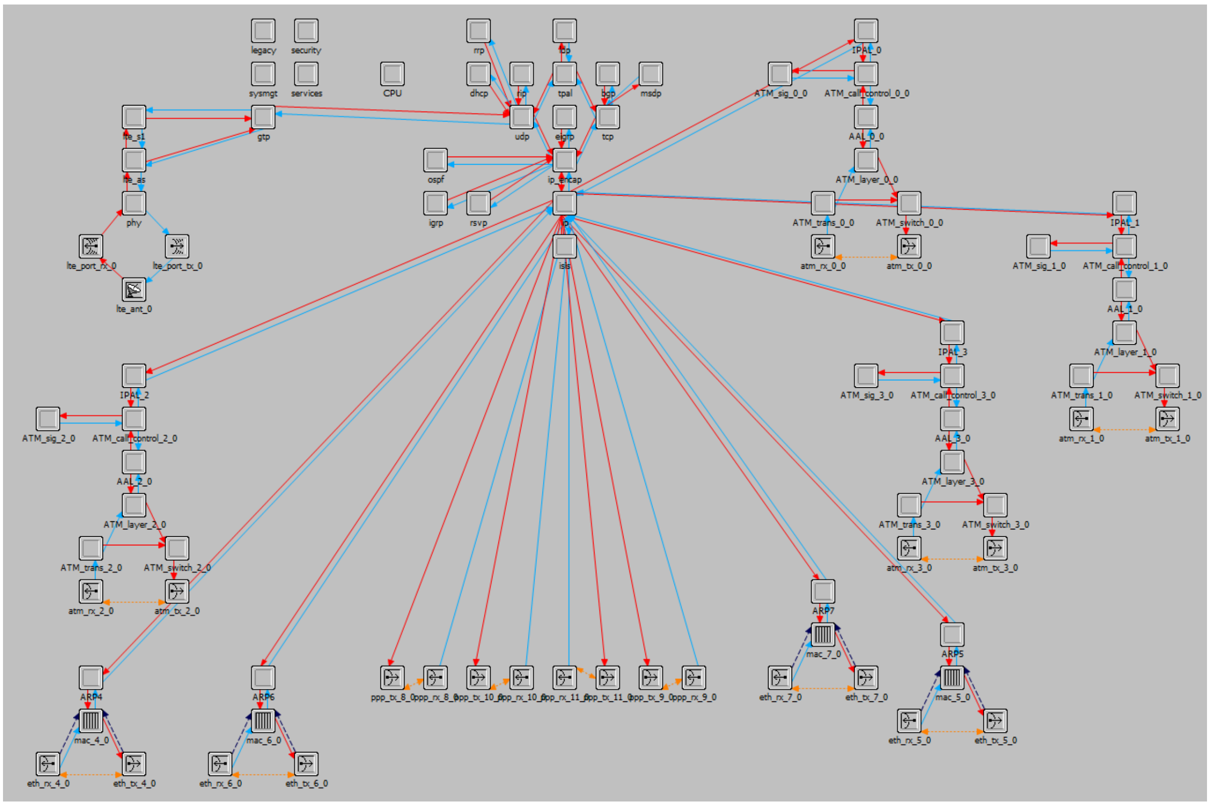


Figure 4.14: eNodeB node model.

## 4.5 Performance Evaluation

The system performance is evaluated over 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 [91] and summarized in Table 4.1.

| Parameter | Value |
|---|---|
| Type of Service | HTTP with Video |
| Video Type | On demand – Non live |
| File Size | 200 Mbytes |
| File inter-arrival distribution | Exponential |
| Average File inter-arrival Time | 16s, 20s, 24s |
| 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 |
| MeNodeB Carrier Frequency (F1) | 2 GHz |
| SC Carrier Frequency (F2) | 3.5 GHz |
| LTE Bandwidth/Duplexing | 20 MHz/FDD |
| Sub-carrier spacing | 15 kHz |
| Sub-frame length (TTI) | 1 ms |
| Symbols per TTI | 14 |
| Data/control symbols per TTI | 11/3 |

Table 4.1: Simulation parameters.

LTE system contains 1 MeNodeB, with variable number of hotspots (Small Cells) and UEs as shown in Table 4.2. The SCs and UEs are randomly distributed under the MeNodeB coverage. Adaptive modulation and coding were enabled in order to enable the UE to communicate with the eNodeB in variable channel conditions.  The interference and multi-path are modelled. IP traffic is established between the UEs and HTTP server is connected to the LTE network through internet backbone as shown in Figure 4.15.

| SC | UEs |
|---|---|
| 0 | 5 |
| 1 | 5 |
| 2 | 5 |
| 2 | 10 |
| 3 | 10 |
| 3 | 20 |

Table 4.2: Corresponding Network parameters.

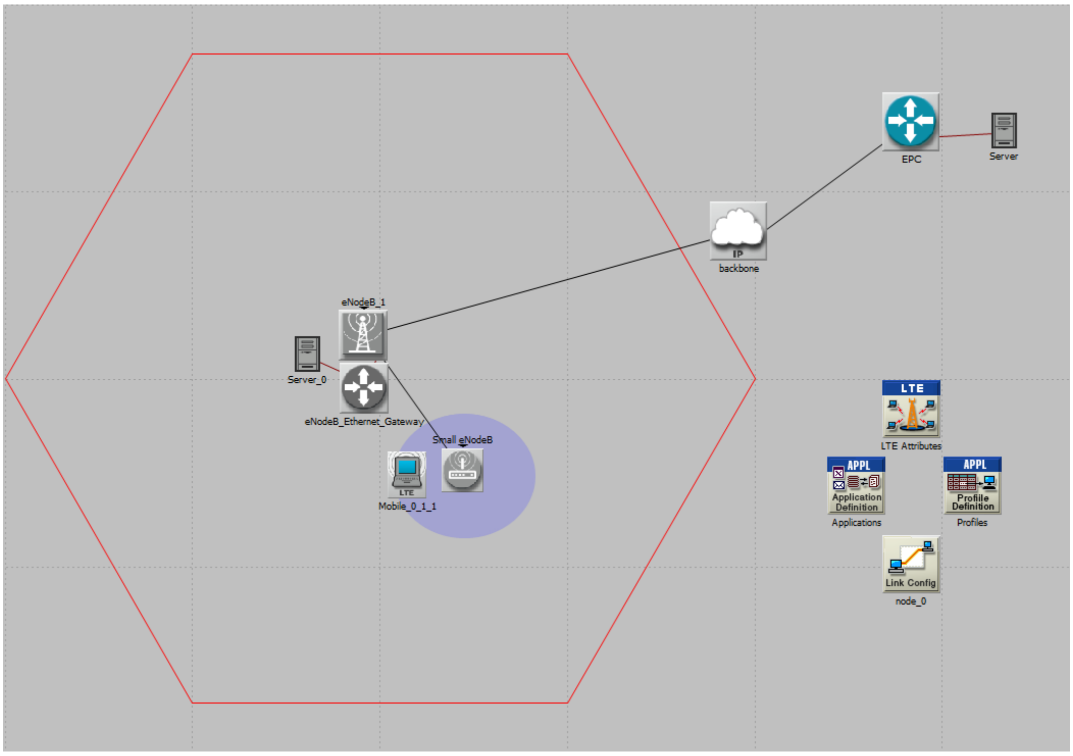


Figure 4.15: Basic System model.

For the first scenario set to start with 1 MeNodeB and 5 UEs randomly distributed within the MeNodeB coverage area, then for the succeeding scenarios, the number of SCs and UEs will be increased to be as 0, 1, 2, 3 for the SCs, and 5, 10, 20 for the users as shown in Table II. 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 proposed scheme is analysed based on the previously specified settings and scenarios. The IP data packets (both sent and received) are also examined over the LTE network. The key performance factors chosen for investigation are the throughput and packet end-to-end (E2E) delay. In the first scenario, the network is configured with low load traffic to decrease the probability of packet loss due to either the buffer overflow or repeated re-transmissions due to the traffic congestion.

Three main cases are considered in evaluating the network performance

- Content is delivered from the cloud (No content is cached).
- Content cached in M-eNodeB
- The UE with active dual connectivity is connected to the M-eNodeB in the UL and to the S-eNodeB in the DL, with Content cached in the small cell.

## 4.6 Results Discussion and Analysis

Figure 4.16 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 M-eNodeB provided that no data is cached in the network while it is acceptable when the contents are cached in the M-eNodeB and has dropped significantly when the UE is connected to the SC-eNodeB and using the proposed scheme.

The explanation of drop in the delay is the fact that the distance to the M-eNodeB is quite larger than the distance to the SCeNodeB, provided that the DL frequency differs from the UL frequency which reduces the interference in the network in order to help decreasing the losses as proposed

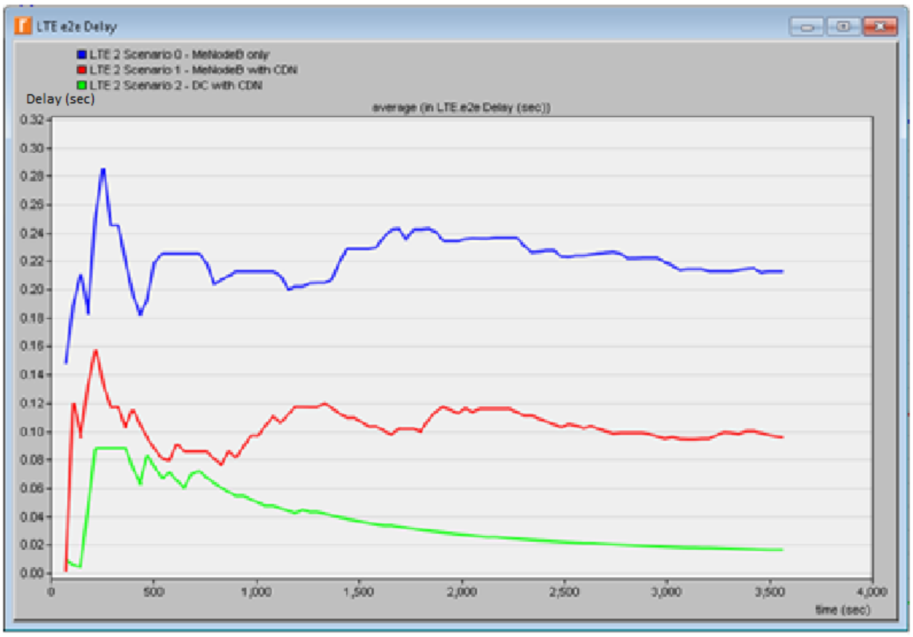


Figure 4.16: End to End delay.

Considering the same network simulation and examining the results in terms of the throughput. Figure 4.17 illustrates the throughput 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 UE downlink is connected to the SC-eNodeB and using the proposed scheme.

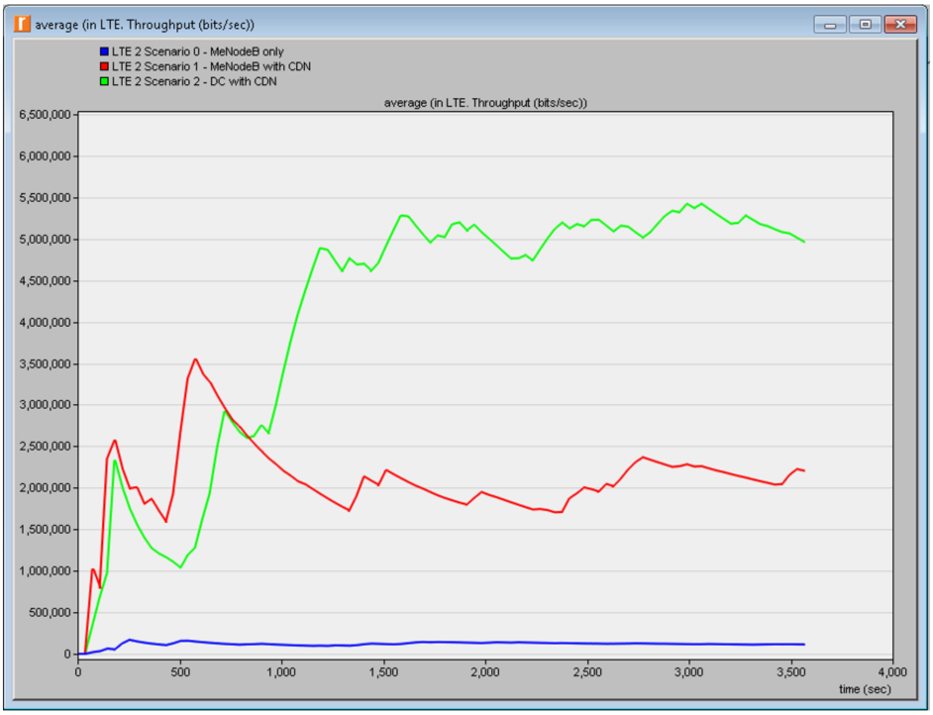


Figure 4.17: Network throughput.

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

i) If Data traffic is increasing within the same period

ii) If the elapsed time to transfer the same amount of data decreases.

iii) Or both of them though it's very rare to happen

Hence in the proposed model, when the content is cached in the M-eNodeB the network delivers and performs best at the beginning of the simulation because the data has been fetched and cached closer to the UE. However, it starts to perform even better in the third scenario when the DL is connected to the S-eNodeB after 20% of the simulation time, this is due to S-eNodeB initialization and time spent fetching the content from the main sever to the S-eNodeB.

In the second run of the simulation the same network is considered with the same parameters but will examine the case when the network is configured and routes full load in its data plane.  Figures 4.18 and 4.19 show the response of the network in terms of End to End delay and throughput for the three scenarios.

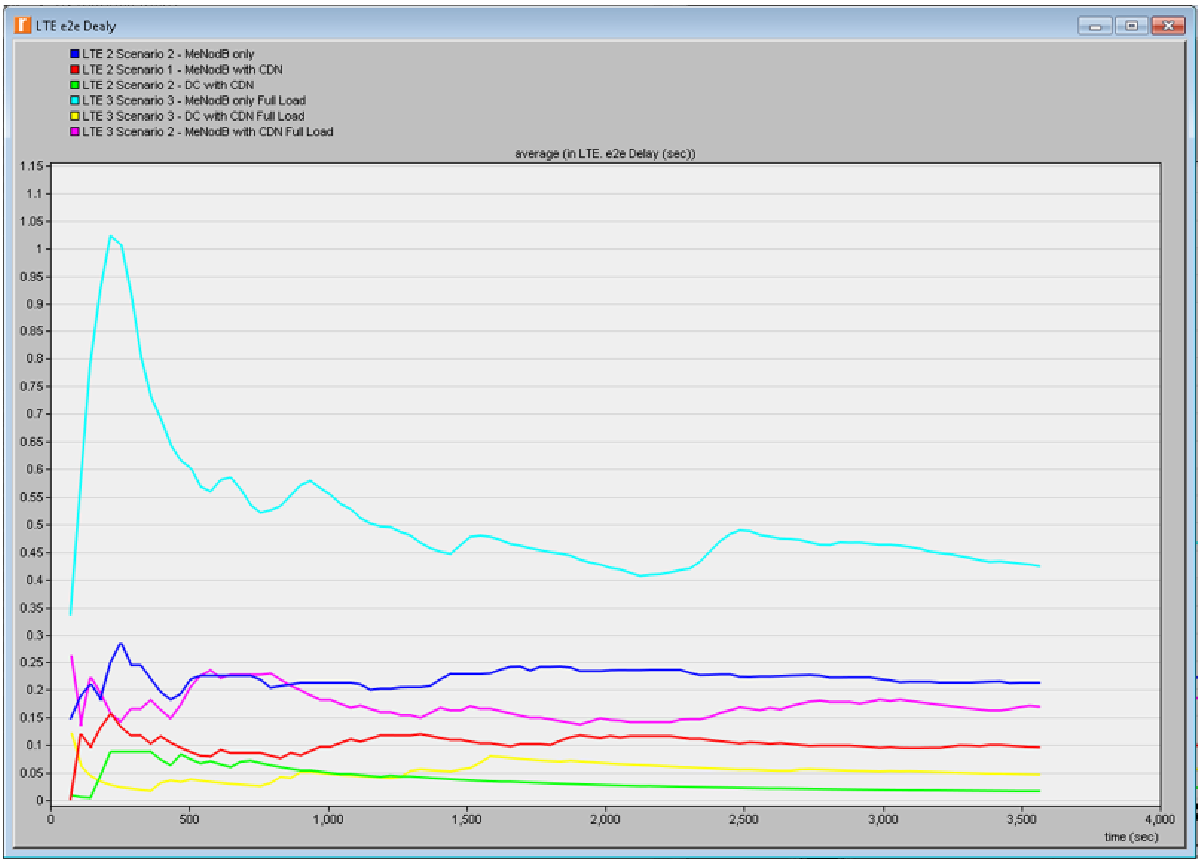


Figure 4.18: End to End delay

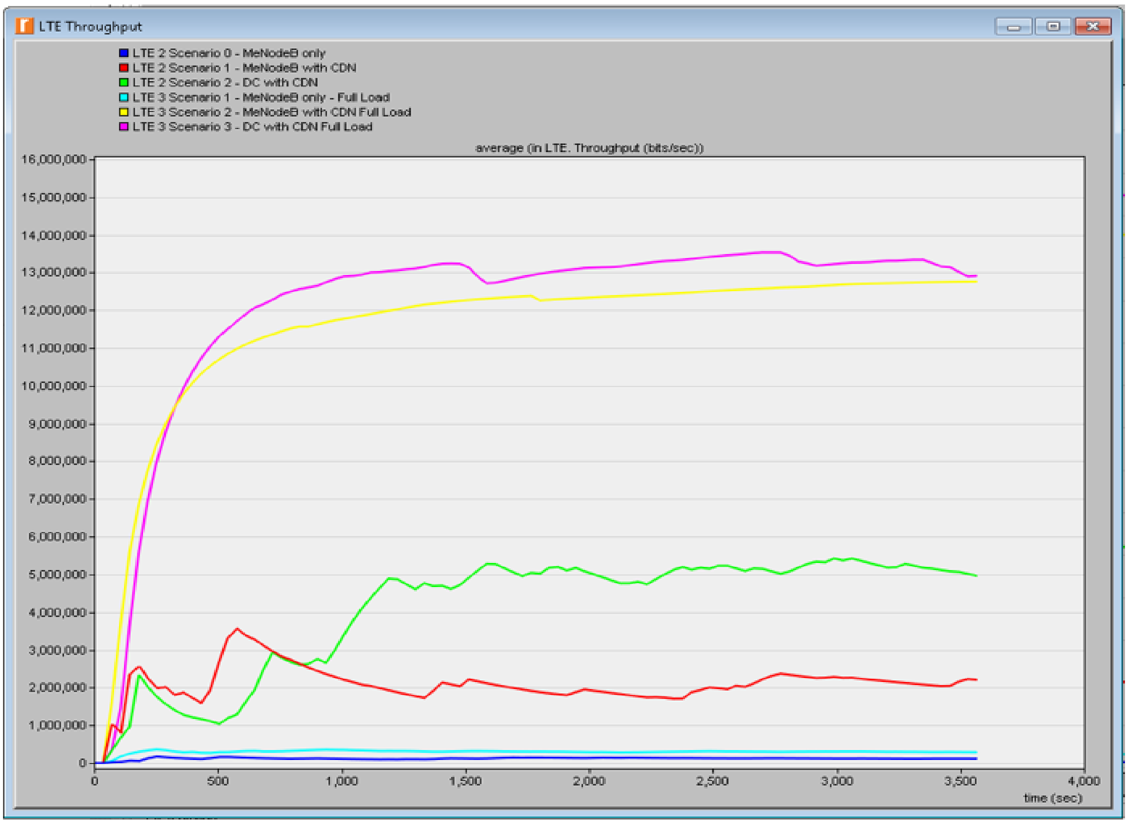


Figure 4.19: Network throughput.

It is observed that the delay is at its highest value when the UE is connected to the MeNodeB with no data server available in the RAN, which is considered as a normal result. On the other hand it is more acceptable compared to the first scenario when the content server is attached to the MeNodeB and it has dropped significantly when the UEs are connected to the SeNodeB and using the proposed Dual connectivity scheme.

The same thing applies regarding the throughput delivered in bits/seconds. When the content server is getting closer to the UE, it can be observed that throughput starts to increase achieving its best when the UE downlink is connected to the SeNodeB. Noting that when the content server is placed in the M-eNodeB the throughput drops remarkably 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 SeNodeBs and UEs as set in table 4.2, Figure 4.20 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 is more than one SeNodeB to serve the same number of

UEs equally or when the number of UEs under the coverage of the SeNodeBs is close. This is the expected response as the burden increases on the MeNodeB since the same content is routed in the network in every scenario. Whilst the incremental increase in SeNodeB numbers in the entire network efficiently decreases the delay as the time elapsed to fetch data from the cloud is narrowed or sub-zeroed, once the data is cached. The difference between the distance to the SeNodeB and to the MeNodeB is major factor to the rise and drop in the delay. In other words, the drop in the delay is because the distance from the UE to the SeNodeB is quite smaller than distance to the MeNodeB, provided that the access time is increasing when increasing the number of UEs to be attached to the network and requesting same contents to be delivered from the server.

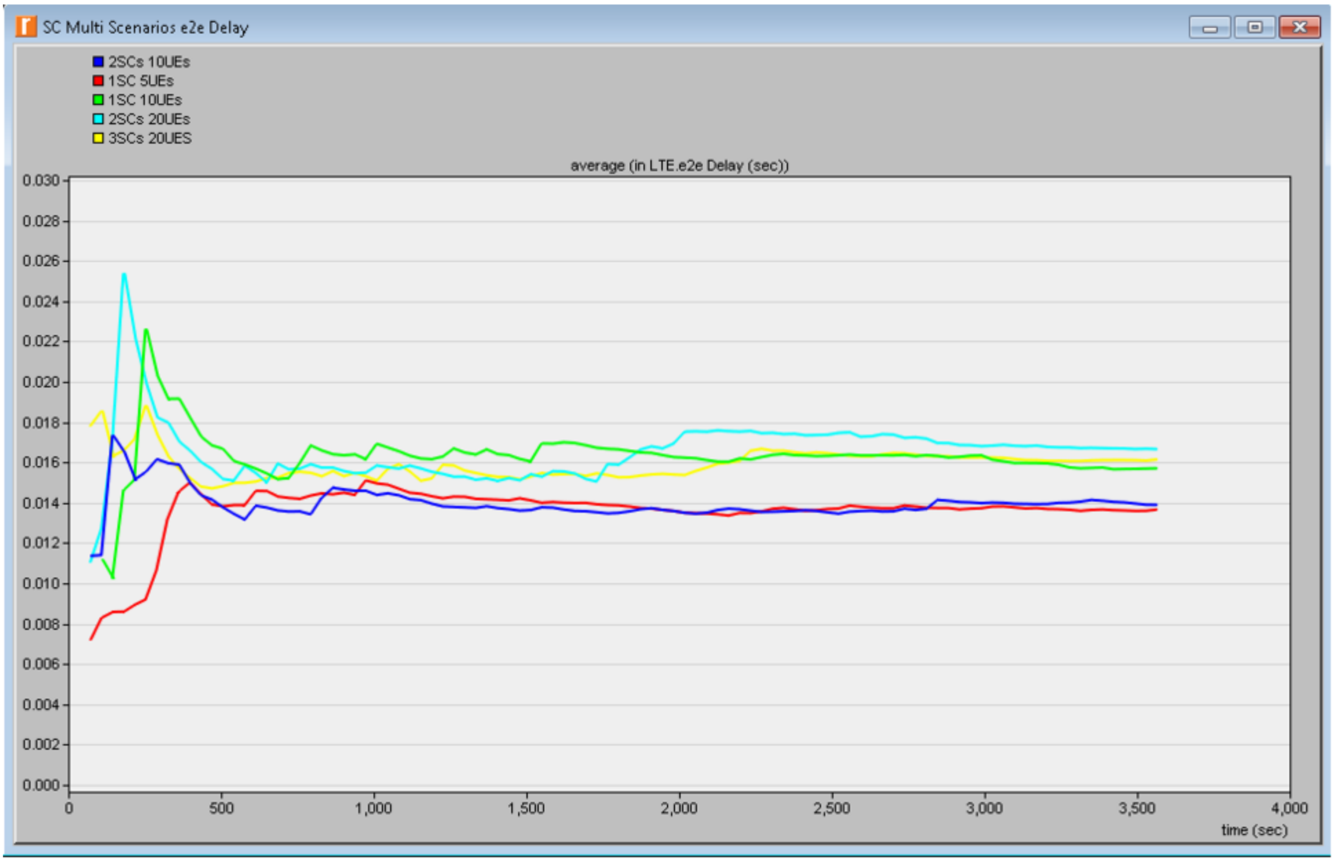


Figure 4.20: E2E delay (multiple scenarios).

## 4.7 Summary

The increasing demand for data connectivity especially indoor drives both operators and developers towards improving the network in terms of capacity, integration of new technologies, spectrum and architecture options. And one of the promising solutions is small cells deployment with in-network caching capabilities; caching techniques have an essential role in communication systems and networks.

In order to fully utilize the facilities provided by small cells without adding burden to the network, dual connectivity was introduced for the UE, DC is a technology to extend CA and CoMP to simultaneous double connection,

This chapter described a heterogeneous network, design and implementation based on the LTE system that supports dual connectivity and data delivery at the RAN. In the proposed design the data and controls of the SeNodeB is processed at the network edge using a MEC server, and the SeNodeBs are used to boost services provided to the users. The proposed system and resource management are simulated using the OPNET modeller and evaluated through multiple scenarios with and without full load. The results clearly show that the proposed system can decrease the latency while the total throughput delivered by the network has highly improved when SeNodeBs are deployed in the system. Rising throughput will incur the rise of overall capacity which leads to better services being provided to the users or more users to join and benefit from the network.

# CHAPTER 5

# LOAD BALANCING USING NEURAL NETWORKS APPROACH FOR ASSISTED CONTENT DELIVERY IN HETROGENOUS NETWORK

## Briefing

In this chapter, a modified LTE architecture with added AI blocks is proposed to overcome the problems occurring due to unbalanced load routing and boosting the delivered throughput. The load balancing technique utilizes Hopfield artificial neural network and Radial Basis Function Neural Network for content delivery mechanism in Heterogeneous LTE mobile network. The proposed network design demonstrated efficient impact on the network performance in terms of power saving and handling data size increase [102].

## 5.1 Introduction

Surfing through the different mobile generations from 1G to 4G, the mobile networks evolution examined various fundamental changes and challenges. Early systems migrated from analogue networks providing voice only service to, nowadays, full IP packet core networks providing multimedia services. During this evolution journey both parts of the mobile network, the core and the radio, evolved through essential changes and enhancements to their structure as well as the way that the user equipment accesses the network. The mobile equipment or user equipment is associated with the end user of the network and is the first link in the mobile network chain; consequently, satisfying the end user by keeping a good quality of experience. The latter is a fundamental requirement to gain substantial revenue levels.

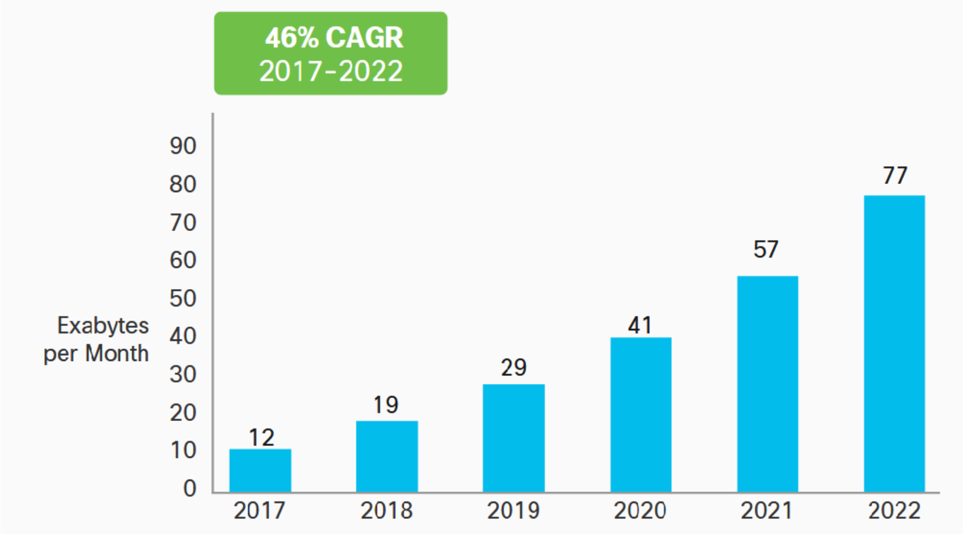
Revenue generation is the target for the mobile network operators; hence they provide their maximum to the customer to increase income. As the mobile networks evolved, the customers’ demands grew and their satisfaction level demands became harder to maintain.

Mobility is one of the major challenges in the mobile network! The fact that any mobile user can move within the network, remain connected and use the service at the same time is one of the fundamental reasons that keeps the user preference of the mobile networks over fixed or landline networks. In order to sustain customers' desired level of satisfaction, the mobile network operators must provide the mobile users with seamless connectivity and continuous service. This is particularly important since mobile devices are no longer a complement or luxury items but have become an essential part of people's everyday life and business [92]. People use their mobile devices while on the move for different purposes and needs, browse the internet, check their e-mails, connect with each other through social media, and streaming audio and video files.

The integration between mobile and computer networks with the advanced capabilities of the devices have led to large amounts of data to be generated. The majority of this data is due to mobile networks and the attached devices, such as mobile smart phones, tablets, wearables and many other devices, routing Big Data which is another major challenge.

According to Cisco Visual Networking Index (VNI) 2019 [102], by 2021, one year after the 3GPP to submit the final specifications of the 5G at the ITU-R WP5D meeting in February 2020, there will be an estimated 58% of the world population using the internet, 4 networked devices per person, global IP traffic will reach 3.3 Zettabyte as shown in figure 5.1. The traffic from wireless and mobile devices will represent 63% of total IP traffic, Smartphones will exceed 86 % that is four-fifths of mobile data traffic, and over 78% of that is three-fourths of the world's mobile data traffic will be video. The requirements for the 5G mobile network include higher connection speed of up to 10 Gbps, latency of about 1 ms, increase in the bandwidth per unit area, 100% coverage and almost the same for availability. Taking into consideration the estimates above and the potential requirements for 5G mobile network; the current architecture of the 4G, represented by the structure of LTE , will not be able to cope with such needs in its present form, but could be the base of the future mobile network or 5G.

Before the full Release of 5G, it is expected that 4G will continue to develop to support many new uses and applications, by making some modification and enhancement on its structure. 4G can handle many new features that could be seen as 5G specific. Such enhancements include the use of small cells in Heterogeneous networks, cloud computing with intelligent load balancing, software defined networks (SDN), and network slicing.



Overall mobile data traffic is expected to grow to 77 exabytes per month by 2022, a sevenfold increase over 2017. Mobile data traffic will grow at a CAGR of 46 percent from 2017 to 2022

Figure 5.1: Global Mobile Data Traffic, 2017 to 2022 [102].

## 5.2 Load Balancing

Mobility while routing Big Data in wireless networks is one of the hardest challenges [103]. Call drops or transmission gaps, which may appear at the users' end, must be prevented as much as possible. This becomes even more critical with LTE since this technology was proposed and designed to support mobile terminals moving at high speeds. While soft and softer handovers mechanisms were implemented in the GSM and 3G mobile networks, they are not applicable in LTE; all handovers performed in LTE are known as "hard handovers". Hard handover means that the reception is interrupted, i.e. connection with the network is lost for a short period [104]. The occurrence of these interruptions has to be reduced as well as their effective periods keeping them as low as possible in order to satisfy the quality of service (QoS) requirements for Voice-over Internet Protocol (VoIP). The users at the edges of cells with heavily loaded links can be transferred to less loaded cells within the neighbouring eNodeBs by making inter-eNodeB and intra-eNodeB handovers similar to cell breathing, to efficiently host the imbalance load over the links, load balancing is needed, figure 5.2 shows balancing diagram. At a certain time, the offered network load, through the bottleneck link in the network link interface, can be reallocated to other links that are not congested. Moreover, from the point of view of the radio network, diverting traffic to the less congested cells will reduce the cell overloading [105]. The radio network can be improved by applying knowledge-based adaptive handovers; thus, enabling a guaranteed QoS for end users. There are several methods for load balancing in the LTE mobile networks such as cell coverage control (CCC) and handover parameter control (HPC) [106], both mechanisms has advantages and disadvantages.

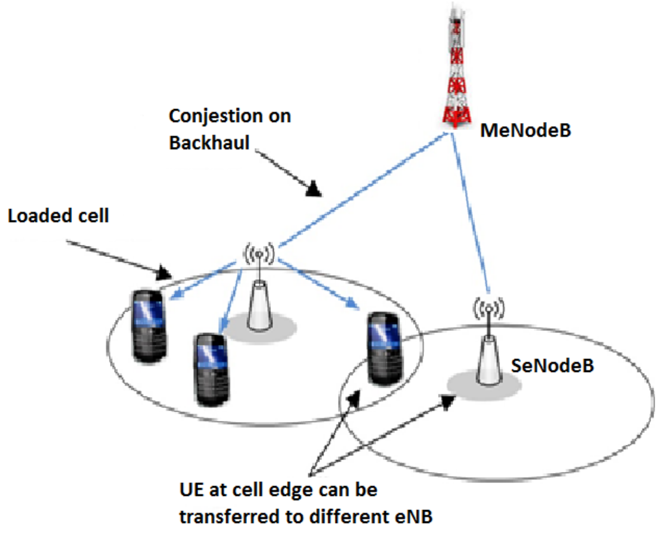


Figure 5.2: Unbalanced network [105].

### 5.2.1 Reporting Handover Parameters

When the UE is in RRC connected mode, it keeps measuring the cell signals of the SeNodeB and the neighbouring cell, according to 3GPP TS 36.331 [107] the UE can be configured to comply with wide individual and separate measurements. These network measurements are related to the reference signals which are generated and transmitted within the control frame. By keeping measures of the reference signal received power (RSRP) and the reference signal received quality (RSRQ) signals the UE generates a measurement table and send it to the SeNodeB with the triggering response or Time To Trigger (TTT) that is determined according to the measurements values as specified in [107]. The TTT is the timespan required for the entering condition to be fulfilled without triggering the leaving condition, which in turn would trigger the handover, while Handover margin (HOM), is a critical value that when is reached by the measured signal to trigger the entering condition Figure 5.3 gives general description of the process within time [108].

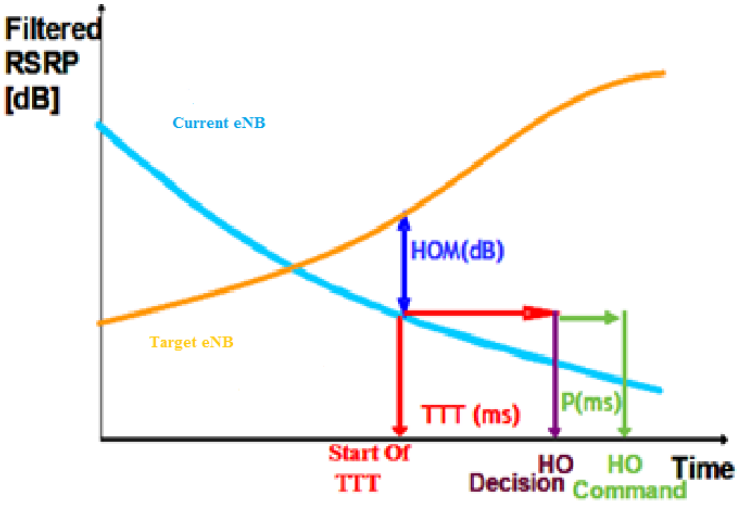


Figure 5.3: HO triggering values selection [108].

In general, and in order to have as minimum as but necessary number of handovers; the mobile service providers must set the parameter values to make sure neither unnecessary nor repeated handovers occur. This is because each handover consumes valuable network resources as well as UE resources (such as battery, processing…etc.) that can be used to deliver better services to network users. If these settings are not carefully selected, then unnecessary handovers may be triggered. For example, if the received RSRP from the neighbouring cell goes high for a very short period of time and if the selected TTT value is too small, a handover will be triggered. These handovers might happen in a frequent way at cell edges, because the received signal strength of the two adjacent cells changes many times, this is denoted as the ping-pong effect. Therefore, the parameters have to be selected carefully by the network operator in such a way that optimal network performance is delivered. Hence, the operator must set these parameters considering the requirements and conditions of the network.

### 5.2.2. Handover Stages

The handover process comprises of three stages [109], handover preparation, handover execution, and handover completion.

- Handover preparation

  The Handover preparation procedure is initiated by the source eNodeB if it determines the necessity to initiate the handover via the S1 interface, then the target eNodeB will perform the admission control to determine whether it has enough resources to provide the EPS bearers for the new added user while maintaining acceptable services for the UEs within its coverage, After finishing the admission control the target eNodeB transmits a handover-request containing result of the admission control to the source eNodeB within acknowledge message., the  source eNodeB approves the handover.

- Handover execution

  In the handover execution, the UE disconnects from the source eNodeB and sequence number status transfer message is generated, all handover necessary data are included in this message, the sequence number status transfer is sent to the target eNodeB to establish a forwarding tunnel, Once this has been done, the SeNodeB will forward all incoming downlink data to the target eNodeB over the X2 interface without the EPC involvement, the target eNodeB buffers all the forwarded until the UE reconnects again.

- Handover completion

  After the UE is reconnected to the target eNodeB, the eNodeB will inform the MME and the SGW to switch the downlink to it and send end marker to the source eNodeB to terminate the old path and no further data related to the specific UE will be sent, and will consider such data as a duplicate and discard it. If any packets are transmitted during the handover procedure, they might go through either old or new path via dual connectivity and the PDCP layer set them in order deliver in the correct sequence.

Figure 5.4 shows the phases of handover according to [108].

## 5.3 Load Balancing Algorithms

There are several types of traffic distribution algorithms which can be considered for load balancing, from which: random, round robin, least load, and least hops…etc. all these algorithms had not achieve full utilization neither for the links nor for used equipment. Hence the need for better algorithms had increased due to the limited resources availability and the growth of user data exchanged in the network, many proposals for new algorithms has been presented considering different sides of the problem.

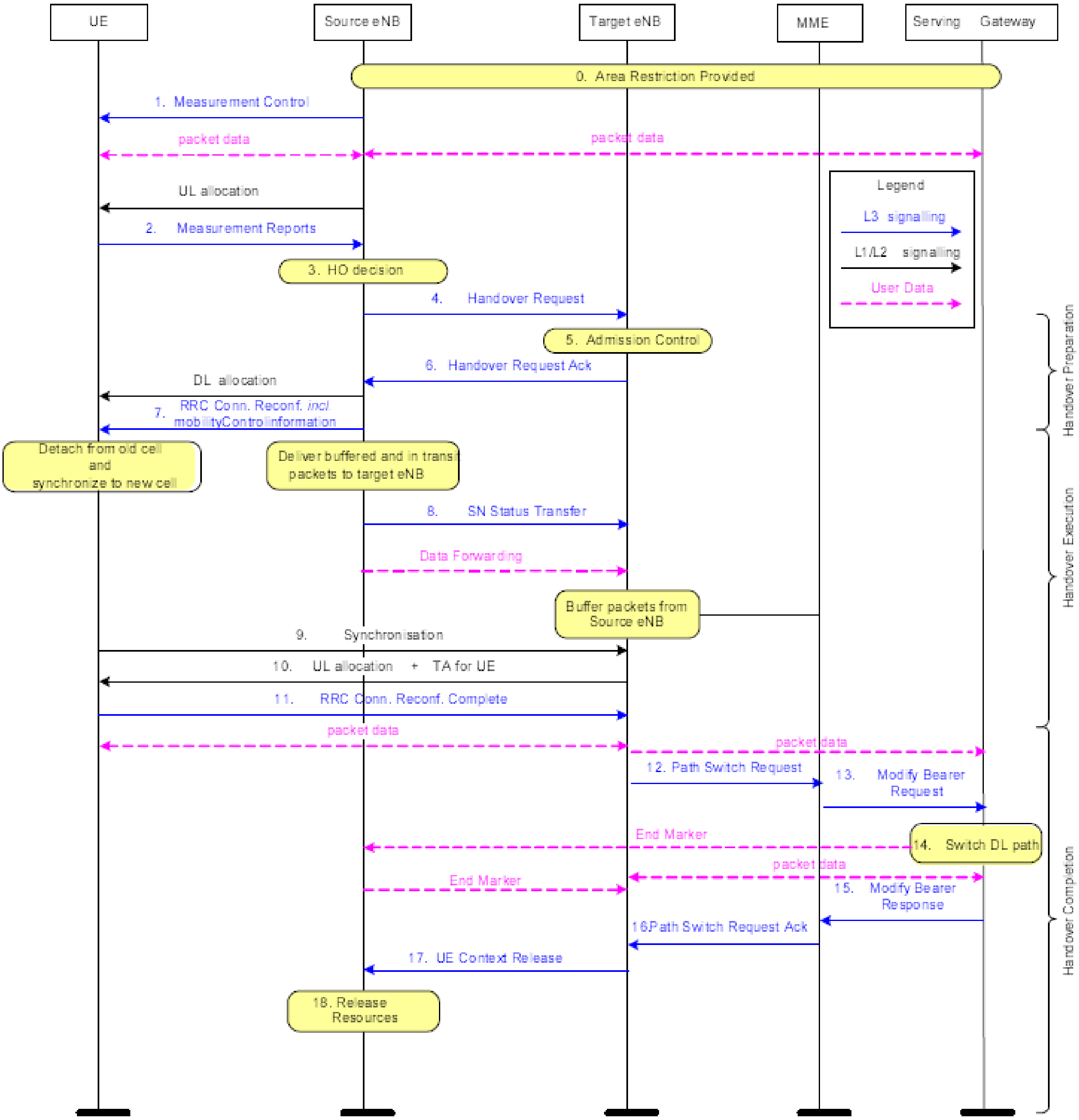


Figure 5.4: Handover Stages [109].

### 5.3.1 Related Work

Some research studies have recently been focusing on achieving load balancing among heterogeneous networks. In [110] the authors presented policy framework for resource allocation in combined cellular/WLAN network, admission control, and selection schemes for access network/access points (APs) where they are re-designed to achieve load balancing. According to the presented results high utilization was achieved when the offered traffic loads are dynamically balanced over the two networks via admission control and vertical handoff, also significant performance improvement is observed in comparison with other reference schemes.

Similar network to [110] was considered by the authors of [111], they proposed performing dynamic load balancing through joint session admission control based on user mobility cognition and service awareness in a tightly coupled 3G/WLAN network, the numerical validation showed enhancement in terms of delivered throughput and call blocking probability of service.

The authors in [105] worked on congestion in the transport network to implement handover toward less loaded cells to help redistribute the load of the bottleneck link; they designed and implemented a handover and load-balancing mechanism for an LTE system model. They considered simulating various handover schemes and different load scenarios with various traffic classes, the results show that the load balancing algorithm can help in balancing the load among the network components. The simulation was carried out using OPNET.

The authors in [112] worked on radio resource allocation in a heterogeneous wireless access medium with constant and variable bit rate services, they proposed a distributed algorithm for resources allocation in each part of the network, Numerical results demonstrated the validity of the proposed algorithm and showed better resources allocation when the number of eNodeBs is reduced or the number of UEs is increased.

The authors in [113] mathematically proved NP-completeness of the problem and develop two algorithms to approximate the optimal solution for big instance sizes, the first algorithm allocates the most demanding service requirements first, considering the average cost of interfaces' resources. The second one calculates the demanding resource shares and allocates the most demanding of them first by choosing randomly among equally demanding shares. The numerical results show the role of the activation cost in the services' splits and distribution among the interfaces, moreover the results demonstrate relation between the number of rounds and the total cost.

While in [114] Song *et al*. propose a load balancing algorithm based on Radial Basis function neural networks, they implemented the algorithm to conduct prediction through network load rate and achieve the network admission of new service. This is by combining an admission control optimization algorithm, and by analysing network performance, some services of heavy load network are transferred to overlay light load network, according to the simulation results the proposed algorithm was able to well realize the load balancing of heterogeneous wireless network and provide high resource utilization.

The authors in [115] introduced an inter-access system anchor based load balancing mechanism to performs load monitoring and evaluation for access gateways and networks, they also proposed a load balancing algorithm for heterogeneous integrated networks with introduction of the utility function concept which supports both single type service. Numerical results demonstrate that load balancing between access networks can be achieved, and the optimal number of handoff users corresponding to the maximal joint network utility can be obtained. Same authors extended their work [116] and improved the network to support multimedia services.

In [117] the authors combined fuzzy logic control (FLC) and multiple preparation (MP) for self-optimization of HO parameters, MP approach is adopted to overcome the hard HO (HHO) drawbacks, such as the large delay and unreliable procedures caused by the break-before-make process. According to the results of the work; the proposed method was capable of reducing HOF, HOPP, and packet loss ratio (PLR) at various UE speeds compared to the HHO and the enhanced weighted performance HO parameter optimization (EWPHPO) algorithms.

The authors of [118] proposed load balance technique based on artificial neural network to equally distribute the workload across all the nodes by using back propagation learning algorithm to train feed forward Artificial Neural Network (ANN). The ANN is used to predict the demand and thus allocates resources according to that demand, the work and results were compared to another 17 load balancing techniques.

[119] demonstrated small cell traffic balancing by jointly optimizing the use of the licensed and unlicensed bands in LTE License-Assisted Access (LTE-LAA), the authors derived a closed form solution for this optimization problem and proposed a transmission mechanism for the operation of the LTE-LAA small cell on both bands. According to the numerical and simulation results, the proposed traffic balancing scheme, successfully provided better LTE-WiFi coexistence and more efficient utilization of the radio resources relative to the existing

traffic balancing scheme, it also provided a better trade-off between maximizing the total network throughput and achieving fairness among all network flows compared to alternative approaches.

## 5.4 Architecture

The proposed network model is based on the 3GPP LTE  Evolved Packet System (EPS) built and validated in chapter 4 of this work, main components for radio and core networks, (the core network simulated using OPNET is available in the form of only one node called the EPC) remained the same, but two additional entities are introduced into the EPC, which will be denoted as Predictor and Balancer, users in various heterogeneous access networks transmit data to these entities via several access gateways, i.e. ePDGs for small cells and MMEs for Macro eNBs MeNB, and the traffic load of each access network can be evaluated through summing up the load of the corresponding access gateways. After receiving the RSSI, the gateways in each access network forward it with monitoring response message containing their load information to the Predictor/Balancer block. In the radio network, and in a similar way to scenario#2 in the 3GPP TR 36.842 V12.0.0 (2013-12) (Release 12) [91].

In this scenario, a UE with dual connectivity will maintain an RRC Connection with the MeNodeB at all times, while receiving user plane data from the MeNodeB and SeNBs. Hence there will be only one S1-MME connection per UE. In such way, the mobility of the UE will be controlled through the MeNodeB, as the MeNodeB will be responsible of the RRC signalling with the MME and there is no need to move the RRC context of the UE between the SeNodeB's. In this technique, handover between SeNodeB's will be as adding and removing new cells as all the information and radio resource management (RRM) are in the MeNodeB server. The main service required by the UE will be the responsibility of the MeNodeB as it keeps all time connection with the UE, while SeNodeBs could be used for specific services such as content delivery. Figure 5.5 shows the suggested EPS with Predictor/Balancer block.

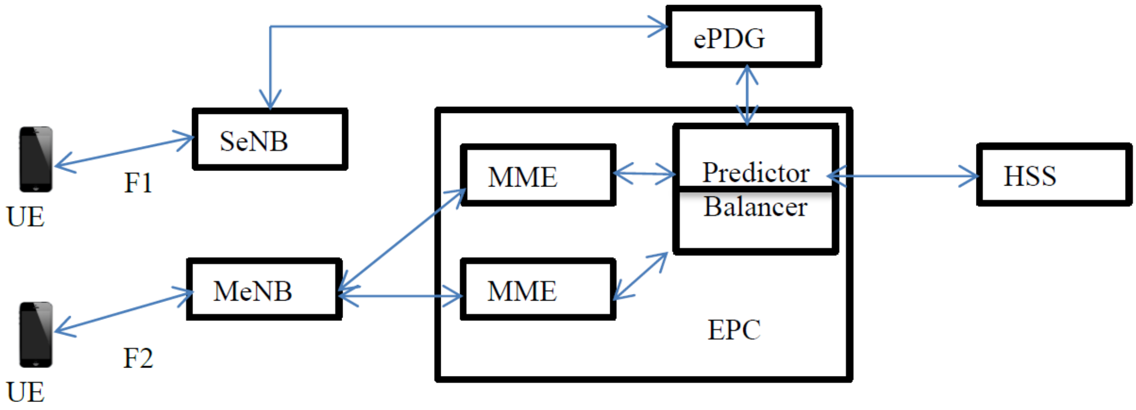


Figure 5.5: Proposed EPS architecture [101].

The Balancer needs to be trained before involvement in decision making; the training scheme is non-supervised which justifies the need for large amount of data that can also be used for the test. The proposed network prototype here consists of 4 clusters based on the data extracted from the LTE core, these groups of data concerns CPU utilization, Link utilization, end to end delay, and throughput. The training data were collected from several OPNET simulations ran 24 hours and fed to the Balancer slices.

The Predictor part is trained with fewer amounts of data when based on RBFNN (supervised training), and the training samples are divided into number of classes equal to the number of hidden neurons, also the same data used for training could be further used for testing the model.

### 5.4.1 Physical Layer Measurements

Physical layer measurements of SINR are collected for both UL-SCH and DL-SCH by a measurement entity in each UE; the measurement entity computes the average load and the size of the measurement window. The link adaptation parameters values can be set in the LTE attributes. RSRP and RSRQ are also collected by the measurement entity in the UE for every eNodeB in range.

- **Transmission power**

  Maximum transmission power is the transmission power of the device over the entire channel bandwidth. It's measured in Watts (W) and is configured via the node attributes.

- **Uplink Power Control**

  LTE uses uplink power control to manage the transmit power of UEs. Hence, each UE adjusts its transmit power to an optimal level for achieving the desired received signal quality at its eNodeB while keeping the power consumption and co-channel interference to neighboring cells as minimum as possible. The Physical Uplink Shared Channel (PUSCH) and the Physical Random Access Channel (PRACH) parameters are also measured from the eNodeB by the measurement entity.

- **Power Headroom**

  Power headroom is the difference between computed and maximum transmits power of a UE. It's computed by the UE and reported to the associated eNodeB.

Both parts of Power Consumption Measurement for eNodeB (operating power and battery capacity) are also reported as physical layer measurements.

Reporting of the measurement required to perform the handover 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 specifications [120]. The RSRP is the average power received by the UE from a single cell and can be measured as:

$$RSRP_i\,(UE) = P_i - L_{UE} - L_f \qquad (5.1)$$

where P is the transmission power of $eNodeB_i$, $L_{UE}$ is the loss gain from the UE to the source eNodeB, and Lf is the fast fading channel gain loss.

The reference signal received quality RSRQ is calculated as follows:

$$RSRQ = B \quad X \ \frac{RSRP}{RSSI} \qquad (5.2)$$

where B is the number of resource blocks and RSSI is the reference signal received indicator of the total power received by the UE.

### 5.4.2 Load Balancing Procedure

According to the proposed architecture shown if figure 5.5 regarding load balancing among the small cells, it is assumed to have one MeNodeB and four SeNodeBs for this simulation, three cases were considered (all cells overloaded, some cells overloaded, no cell is overloaded).

- All cells overloaded: in the situation when all the SeNodeB are congested, the balancer will provide information to the chosen UEs to perform a handover to the MeNodeB at a certain triggering time and the procedure explained in 5.2.2 will start.
- Some cells are overloaded: in the situation when all the SeNodeB are congested, the balancer will provide information to the chosen UEs to perform a handover to another less loaded SeNodeB, the balancer decides which UEs and which SeBodeBs are involved in the handover and also when to perform the handover , depending on the learned scheme. Resulting to the optimal load balancing
- No cell is overloaded: in the situation when the load is below the critical value for all the SeNodeB, no handover is needed, and the balancer will keep monitoring the load status.

The load status in a certain SeNodeB is governed by many factors such as: Number of the connected UEs, allocated bandwidth for each UE and the actually available bandwidth, and this will lead to:

$$L_i = \frac{\mathrm{Bi}}{\mathrm{Bmaxi}} \qquad (5.3)$$

where L is the load, B is the bandwidth and $B_{max}$ is the maximum bandwidth for the $i_{th}$ eNodeB.

The $i_{th}$ eNodeB is considered congested if $L_i \geq K$
where K is a threshold value defined for this purpose
Figure 5.6 shows the flow chart for the algorithm.

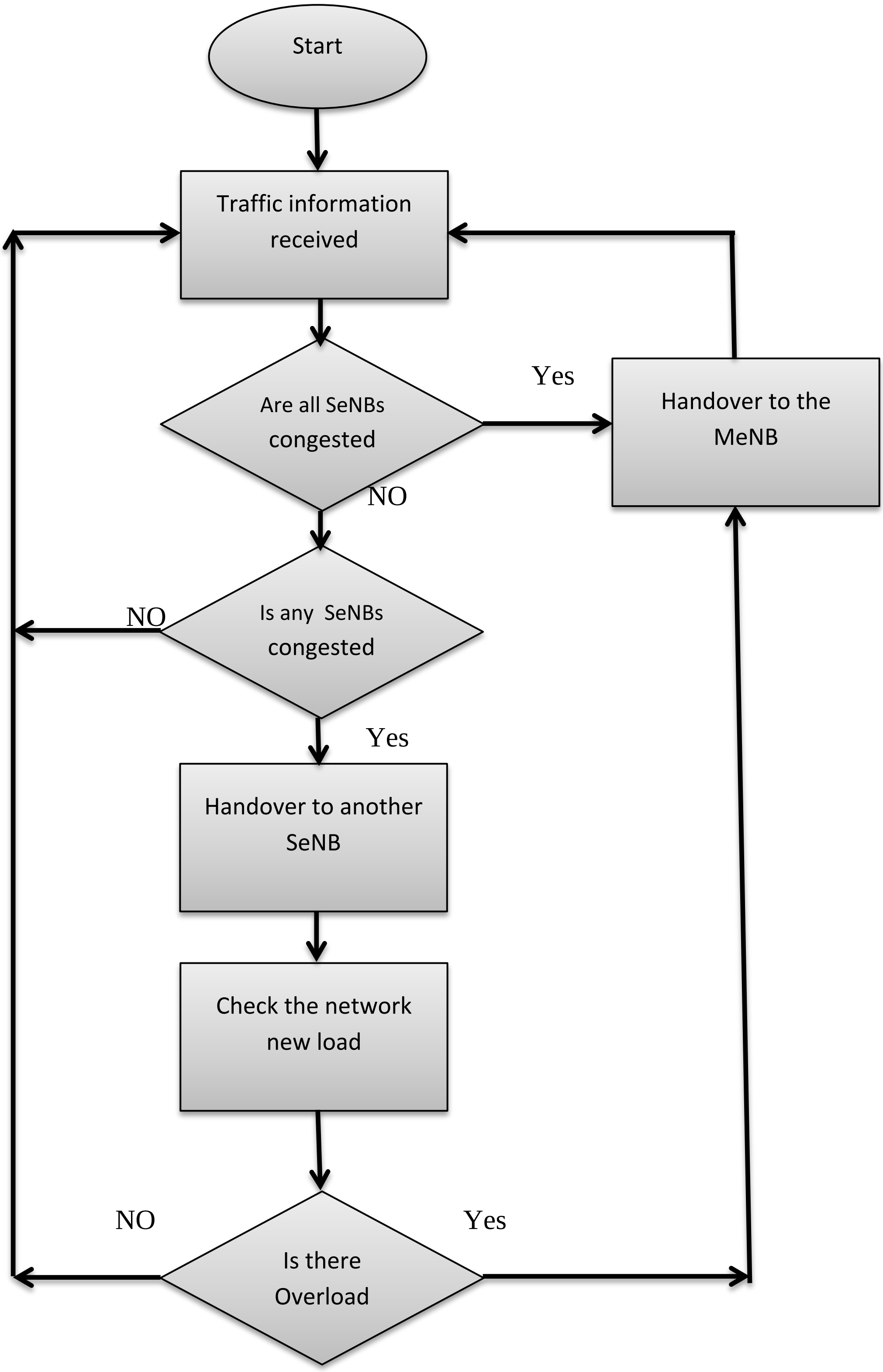


Figure 5.6: Load evaluation and handover decision.

### 5.4.3 Load Balancer Design

The proposed load balancer utilises neural network clustering mechanism to sort the UEs in clusters within the SeNodeBs and gives results to control the handover triggering for any congested cell in the network. MATLAB software is used to implement the predictor, and as mentioned earlier, simulation run takes place over two platforms and the data exchange is made manually for the prototype, because of that the balancer can be distributed in one or many pints of the network e.g. the MEC server, or any other computing capable machine, however; in the proposed network the balancer is located in the EPC and the connections to and between the cells are assumed to be optical fibre.

The balancer consists of three layers, Input, Self-Organizing map SOM, and output layer. MATLAB software supports two approaches to simulate clustering with neural networks; the first is numerical modelling with which all the parameters are defined numerically, with system coding to solve the problem and obtain the output tables, the second approach is the neural clustering app which can be set using graphical user interface (GUI), the input is called and fed to the input layer, the SOM learn to cluster data based on similarity, topology, with a preference of assigning the same number of instances to each class, they are also used to reduce the dimensionality of data [121]. Figure 5.7 shows the layers of clustering neural network.

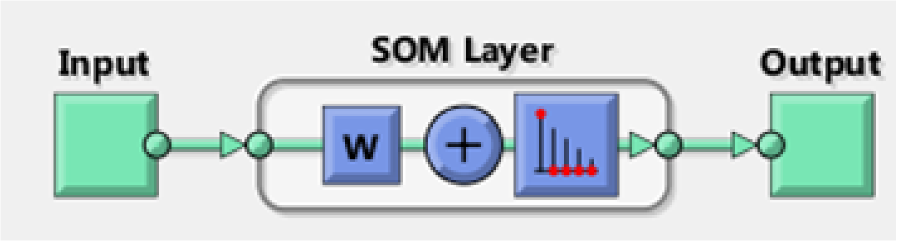


Figure 5.7: Clustering neural networks [48].

For SOM training, the weight vector associated with each neuron moves to become the centre of the cluster of input vectors [48].

The number of hidden neurons determines total number of clusters in the first layer. The larger the hidden layer the more clusters the first layer can learn, and the more complex mapping of input to target classes can be made. The relative number of first layer clusters assigned to each target class is determined according to the distribution of target classes at the time of network initialization [48].

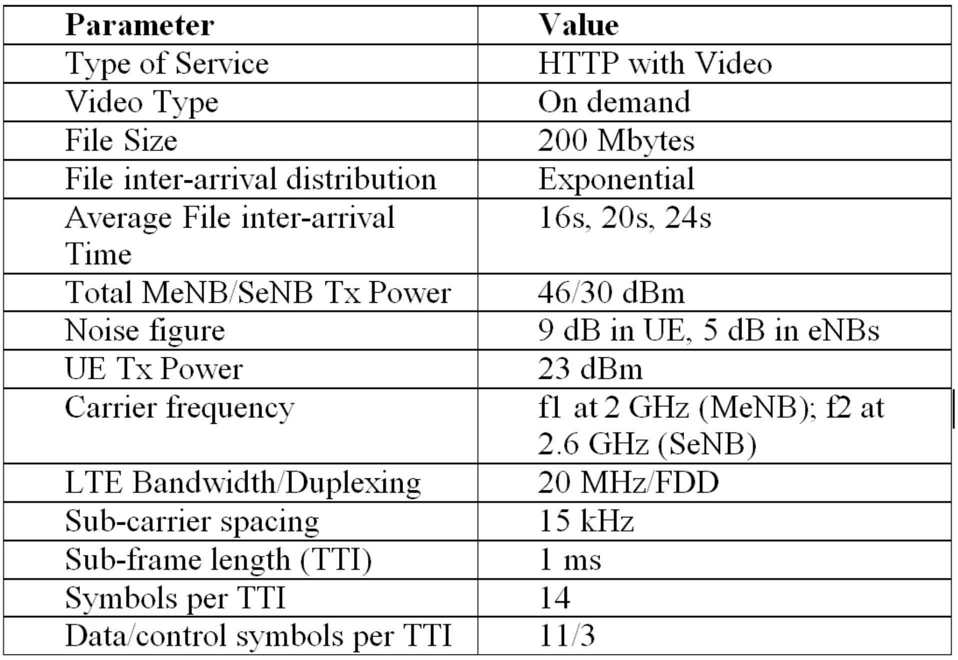
## 5.5 Simulation Setup

As mentioned in section 5.3 the proposed network model is based on the 3GPP LTE Evolved Packet System (EPS) built and validated in chapter 4 of this work, main components for radio and core networks, (the core network simulated using OPNET is available in the form of only one node called the EPC) remained the same, The AI part, is designed, simulated and run using MATLAB. The behaviour of the proposed traffic balancing scheme in various scenarios using a combination OPNET/MATLAB results.

In order to compare results of the same class and distinguish the impact of the enhancing circuitry added to the network, the simulation considered two projects;

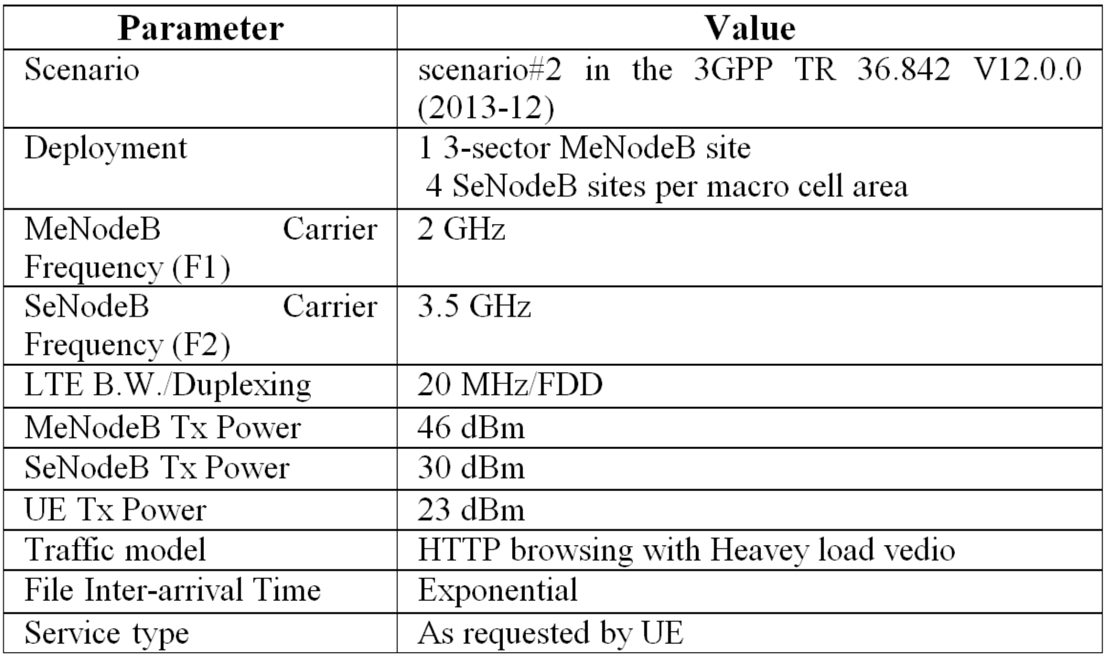
a) Using the available balancing technique in three scenarios; where the first one presents the case when no congested cells, the second incur the same condition but with more users (causing congestion on some cells), and the last one with all SeNodeBs congested. The simulation parameters are listed in Table 5.1.

b) The other project in which the proposed balancing with handover mechanism is activated, it comprises of one scenario examining and comparing the behaviour of the network. The simulation parameters are set as in Table 5.2.



| Parameter | Value |
|---|---|
| Type of Service | HTTP with Video |
| Video Type | On demand |
| File Size | 200 Mbytes |
| File inter-arrival distribution | Exponential |
| Average File inter-arrival Time | 16s, 20s, 24s |
| Total MeNB/SeNB Tx Power | 46/30 dBm |
| Noise figure | 9 dB in UE, 5 dB in eNBs |
| UE Tx Power | 23 dBm |
| Carrier frequency | f1 at 2 GHz (MeNB); f2 at 2.6 GHz (SeNB) |
| LTE Bandwidth/Duplexing | 20 MHz/FDD |
| Sub-carrier spacing | 15 kHz |
| Sub-frame length (TTI) | 1 ms |
| Symbols per TTI | 14 |
| Data/control symbols per TTI | 11/3 |

Table 5.1: Simulation setup (Balancing).



| Parameter | Value |
|---|---|
| Scenario | scenario#2 in the 3GPP TR 36.842 V12.0.0 (2013-12) |
| Deployment | 1 3-sector MeNodeB site<br>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 | HTTP browsing with Heavey load vedio |
| File Inter-arrival Time | Exponential |
| Service type | As requested by UE |

Table 5.2: Simulation setup ( proposed).

## 5.6 Simulation Results and Analysis

**Project 1.** In this section, the network is run and analysed based on the scenarios and settings discussed in the previous section. The sent and received IP data packets all over the LTE network were examined. The throughput, packet end-to-end (E2E) delay, CPU utilization and the associated eNodeBs for both uplink and downlink were chosen as the key performance factors. In the first scenario, the network was running normally with different load conditions and the handovers performed were due normal mobility, the network behaviour with increasing probability of packet loss due to either the buffer overflow or repeated retransmissions caused by link congestion, however; since the number of users was limited then the cached content are neither large nor variant and the small cells handles the network efficiently. Figure 5.8 shows the network throughput

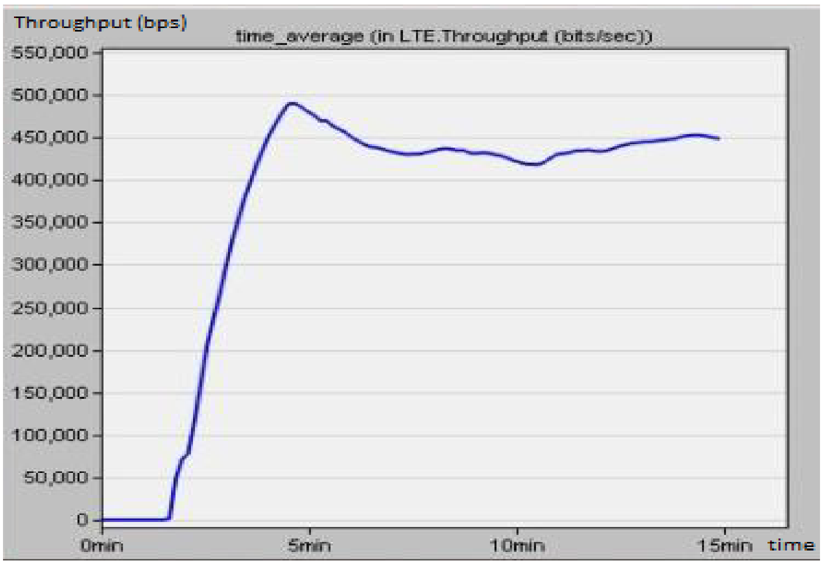


Figure 5.8: Network throughput (scenario 1).

Figure 5.9 shows the packets end to end delay for each small cell in the network, it can be observed that they have close values to each other, the differences corresponds to several reasons, such as distance between UE and the eNodeB and content availability.

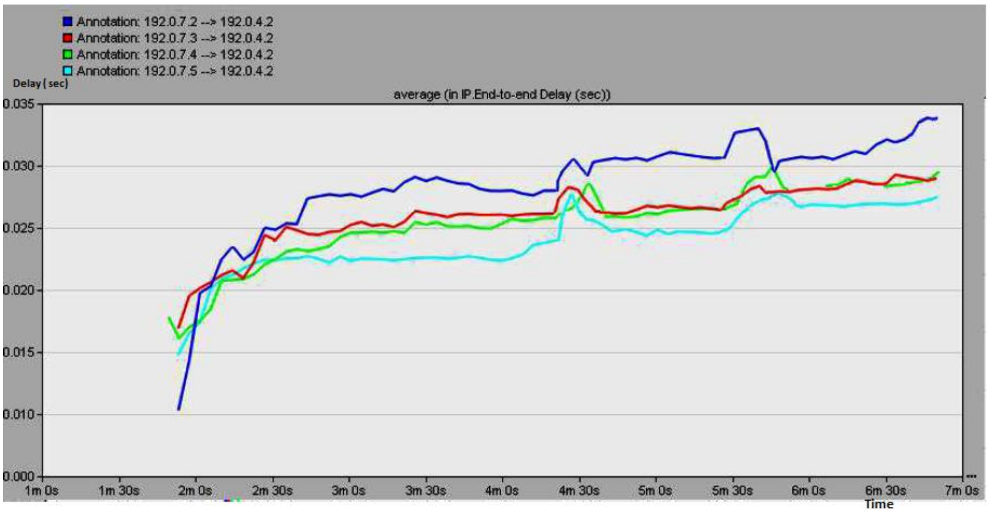


Figure 5.9: end2end delay (scenario 1).

Figure 5.10 shows the associated eNodeB for the UEs at the cell edge. It can be noted that the UEs are connected and transmitting/ receiving for the whole period of the simulation.

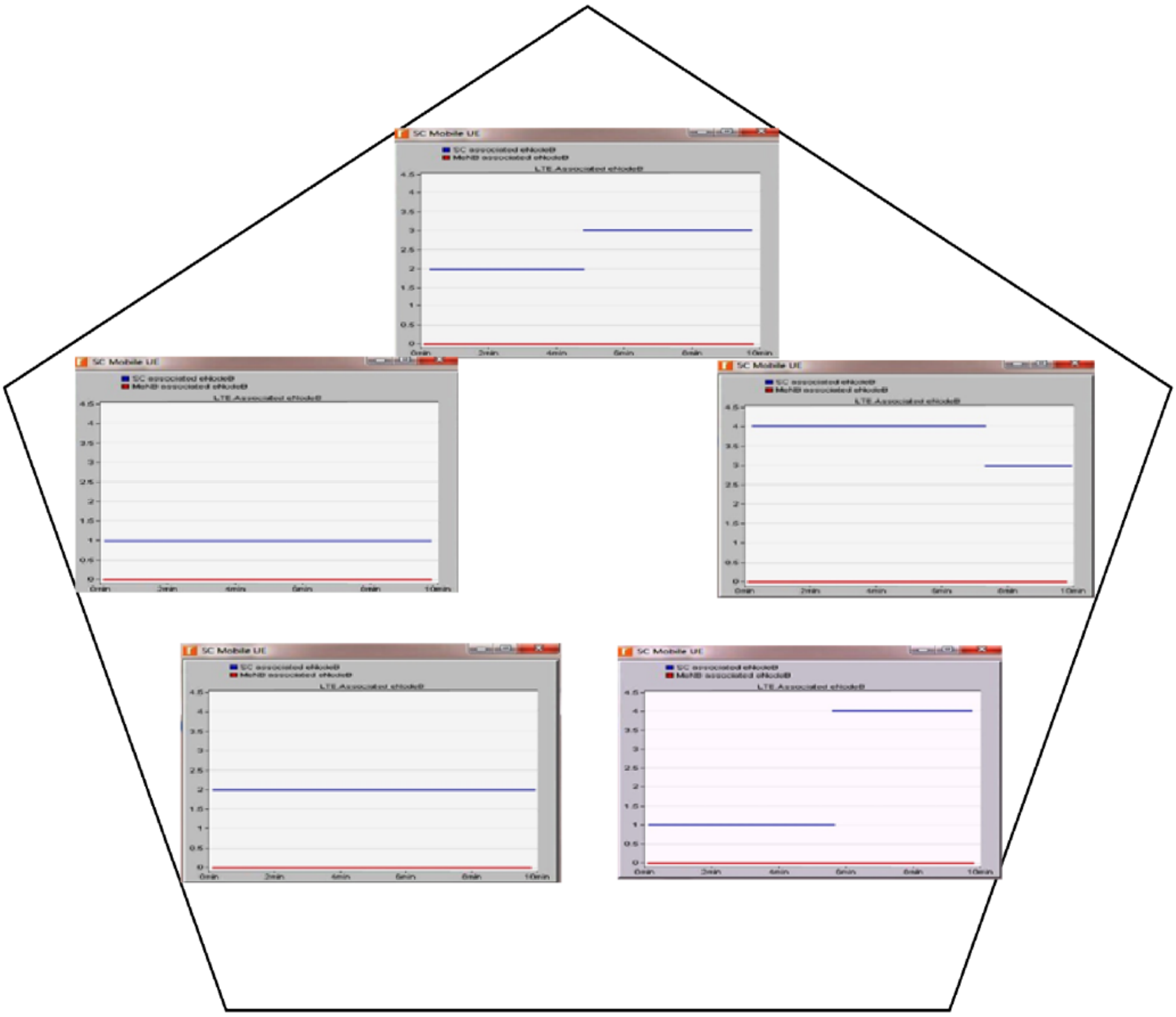


Figure 5.10: Associated eNodeB (scenario 1).

In the second scenario, the load was increased by increasing number of UEs; however the new added UEs were chosen to be close to the eNodeB with limited or zero mobility in order to maintain connectivity and to affect the resources and channel conditions for the UEs at the cell edge. The throughput and the delay were not significantly altered compared to scenario 1, and this is due to the caching facility in the SeNodeB, which saves the time elapsed for fetching the content .Figure 5.11 shows the throughput and Figure 5.12 shows the end to end delay of SeNodeBs in the network.

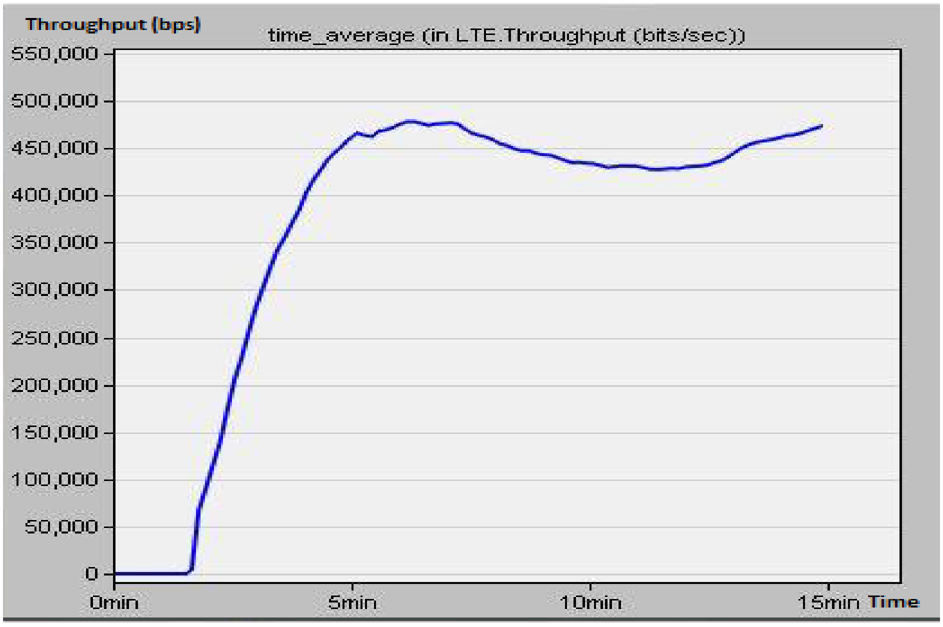


Figure 5.11: network throughput (scenario 2).

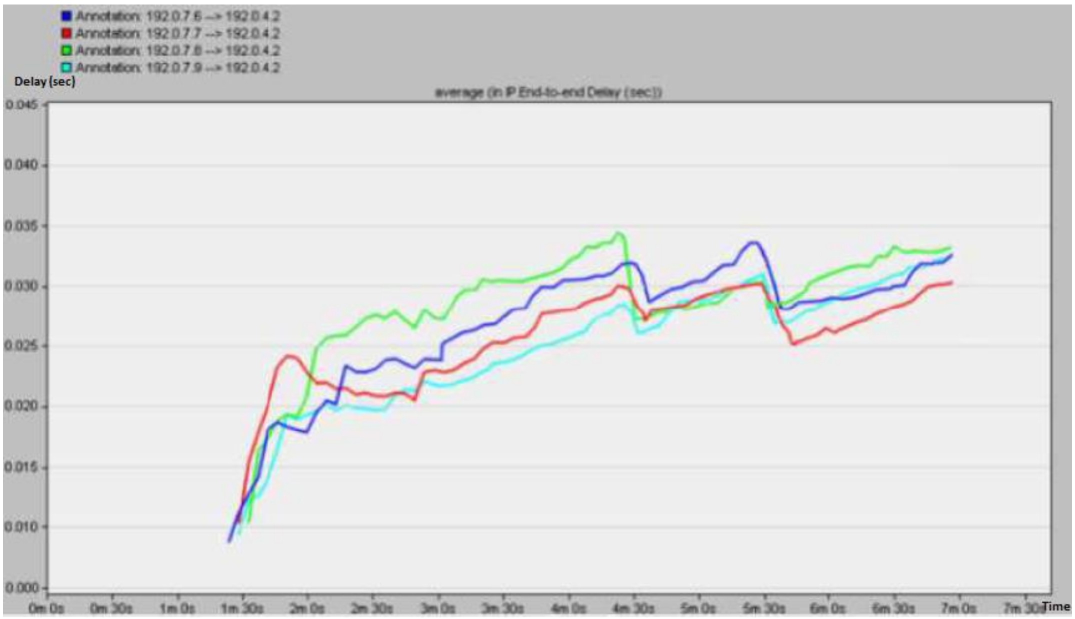


Figure 5.12: end2end delay (scenario 2).

Figure 5.13 shows the associated eNodeB for the UEs at the cell edge. It can be noted that the UEs maintaining connectivity and keeps transmitting/receiving for the whole period of the simulation.

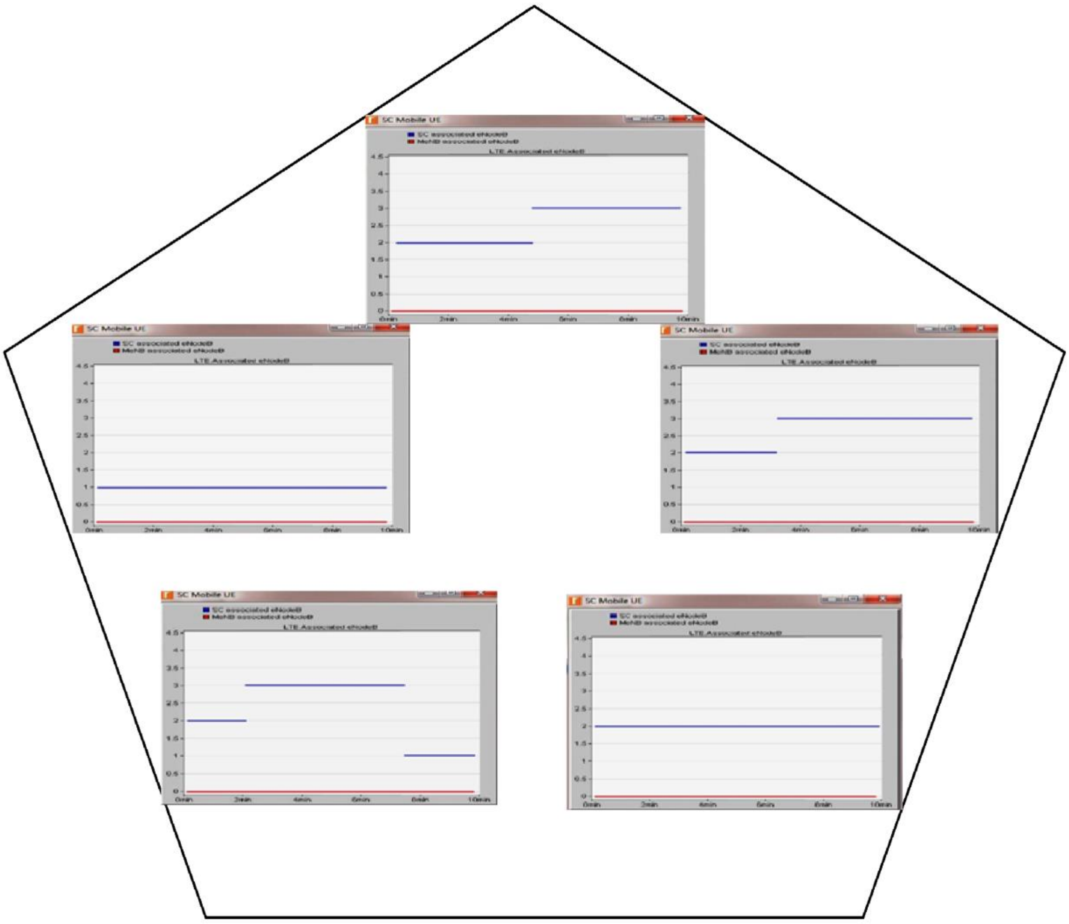


Figure 5.13: associated eNodeB (scenario 2).

In the third scenario, the network load was further increased by increasing number of UEs; similar to scenario 2, the new added UEs were chosen to be close to the eNodeB with limited or zero mobility in order to maintain connectivity and to absorb as much resources and channel conditions resulting in shortage for the UEs at the cell edge. The delay started to increase rapidly compared to scenarios 1 and 2. This increase is caused by frequent handovers performed by the UEs due to the link congestion; Figure 5.14 shows the end to end delay for the 4 participating SeNodeBs.

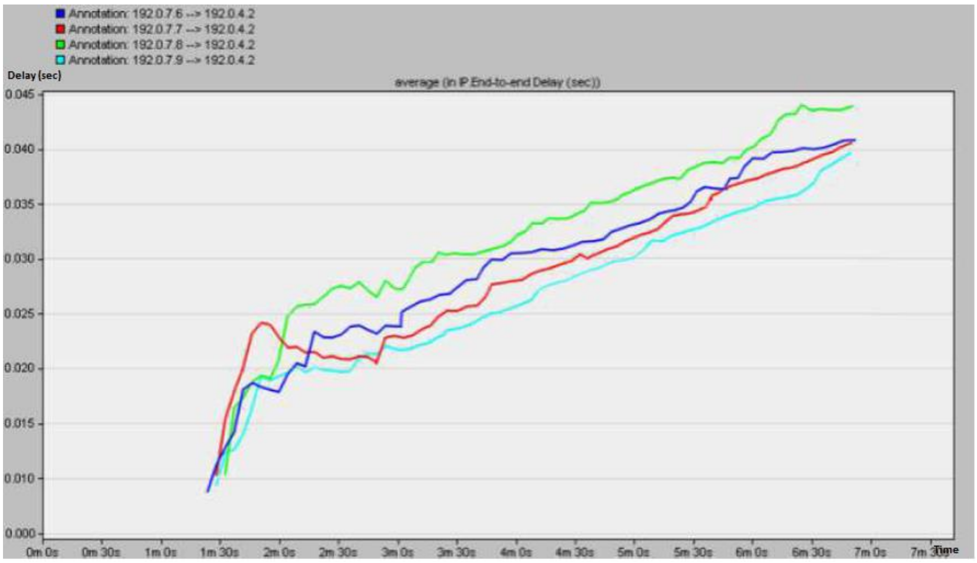


Figure 5.14: end2end delay (scenario 3).

In Figure 5.15 where the associated eNodeB for the UEs at the cell edge is viewed, it can be seen that the UEs in dual connectivity mode are performing handovers continuously in the downlink direction, either from SeNodeB to another or from SeNodeB to the MeNodeB; this will result in huge amount of resources to be allocated and hence consumed for this purpose as well as more burden on the MeNodeB due to the uneven resources sharing, it can be observed that SeNodeB 3 is providing less service than the other SeNodeBs.

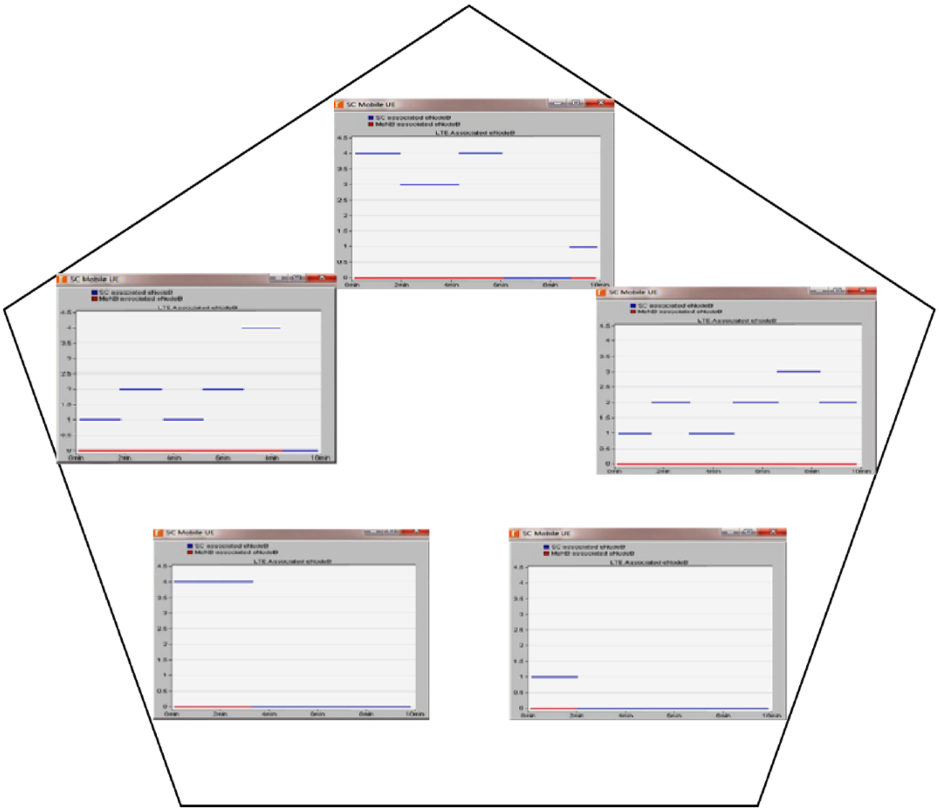


Figure 5.15: Associated eNodeB (scenario 3).

**Project 2.** In the second part of the simulation, in order to investigate and compare the benefits of the proposed architecture, same conditions of scenario 3 were considered but with the load balancing entity ON. The balancer is trained to gain knowledge from the network load, delay and handover parameters, the outputs can be used to apply load balancing, the propose load balancing add-on allows defining new threshold values to recognize congested cells, upon these values the decision of performing handover is made, depending on the learned pattern load balancing is activated in the PDCP layer of the MeNodeB and SeNodeBs to decide the which UEs to remain associated with and which ones to issue a handover.

The results in Figure 5.16, shows the end to end delay for each SeNodeB in the network, it can be observed that the average delay doesn’t exceed 30 msec in its worst case and remain consistent when compared to the results showing the delay of unbalanced network (Figure 5.14). The delay can be reduced either by adding more resources or reducing the number of connected users, however better resources utilization can also reduce the delay.

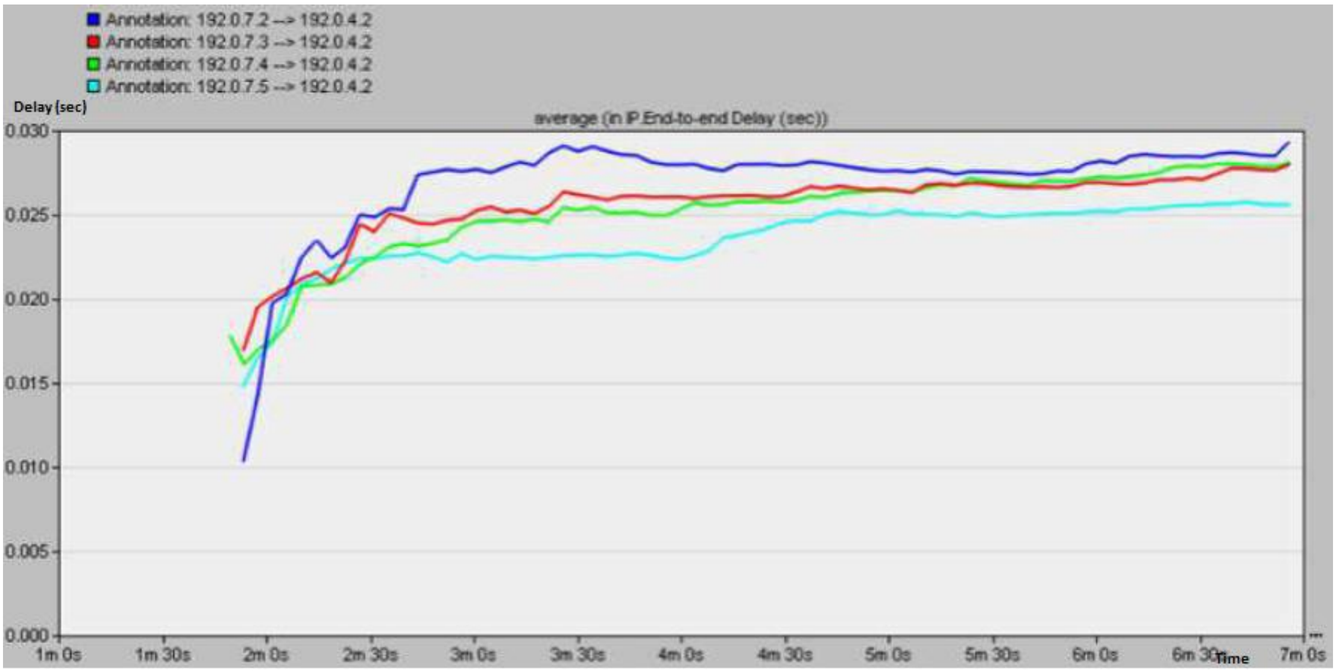


Figure 5.16: end2end delay (Balanced network).

Another benefit of the adding the balancer is noted when comparing Figures 5.15 and 5.17 in which the number of handovers occurring in the network are shown for UEs at the edge of the cell, the balanced network demonstrated less handovers compared to the unbalanced and this is due to the better exploitation of the network resources.

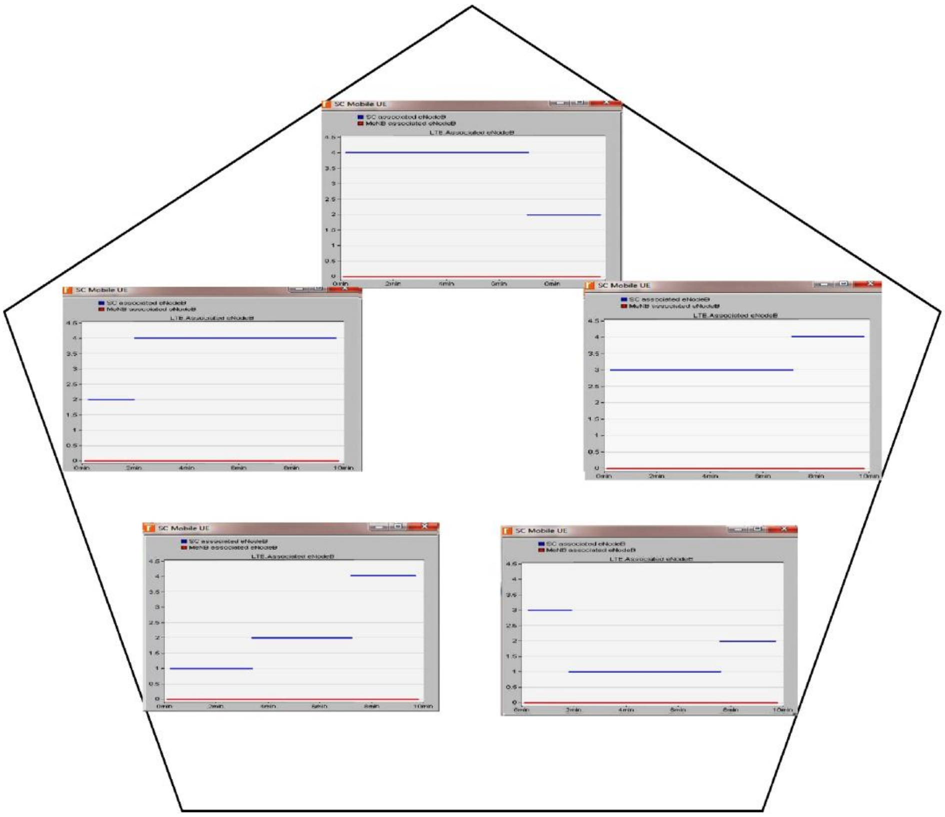


Figure 5.17: Associated eNodeB (Balanced Network).

However; and because of its nature, the balancer in some cases may not give the optimum solution, and will output data for triggering the handover not in the direction of the best cell, this might happen when very large sets of data are used which increases dimensionality, or deficiency of the used algorithm, or unexpected reaction or speed mobility of the UE, which is the reason for proposing another AI entity for prediction in the next chapter of this work, the AI entity is merely a software package installed on a computing platform to perform specific tasks.

## 5.7 Summary

Maintaining good mobile communication service at the cell edge is a critical point and requires improvement, especially with limited resources availability and increasing demand for service, in this chapter an AI based load-balancing block is introduced within the architecture to help congested cells handle traffic dynamically by changing the handover triggering parameters depending on the values tables obtained from the AI entity. The balancing algorithm is based on clustering approach and was evaluated in various conditions with comparison to the traditional built in load balancing.

The results show a better distribution of the load in the previously congested cell, less handovers were initiated and less end to end delay in all the small cells; this will consequently result in significant reduction in the signalling allocated for handovers and will save the resources just like in a lightly-loaded cells.

# CHAPTER 6

# NEURAL NETWORK FOR MOBILITY PREDICTIONS AND RESOURCE ALLOCATION IN LTE

## Briefing

This chapter presents a neural network prediction system as an add-on in the LTE EPC, the predictor is able to learn some of the patterns demonstrated by users moving within the network and can then predict the future behaviour of these users and the next associated eNodeB. These predictions are utilized for better radio resource allocation and scheduling.

## 6.1 Introduction

Prediction is defined as knowing what will might happen in the future under certain conditions and circumstances before its occurrence, prediction is an outstanding approach in many fields such as disease health care, telecommunications, weather and climate forecasting, and natural disasters precautions. Prediction studies address the development of a prediction model, the validation of a prediction model or both. In mobile communication networks, all the UEs must have free mobility while maintaining connection and keep acceptable level of QoS. As a consequence of mobility, any moving UE in the wireless environment will have to have different points of contact with the network, i.e. it has to perform handovers to all the eNodeBs lying in its path.

The time elapsed for handovers is a critical problem in wireless mobile networks, and it has to be minimized along with decreasing the occurrence of redundant handovers or unfinished handovers especially when the UE is moving fast at the edge of the cell, this kind of handovers is known as the ping pong effect, this kind of handovers reduces the throughput and degrades the QoS because of the management signalling overhead and the channel interference. Predicting the next eNodeB on the path and directing the resources towards it, could be helpful in solving this problem.

Prediction process needs to address the following points: location, the type of knowledge, type of next action, time of next action, and accuracy level of the prediction,

- The location of the user can be considered as the location of the fixed eNodeB that is associated with, from the view of the prediction unit, though it can easily be either calculated or accurately estimated, the exact geographical location is not very important for the process, therefore the UE location and its motion through the network can be considered as a sequential list of associated eNodeBs and their links, as shown in Figure 6.1.

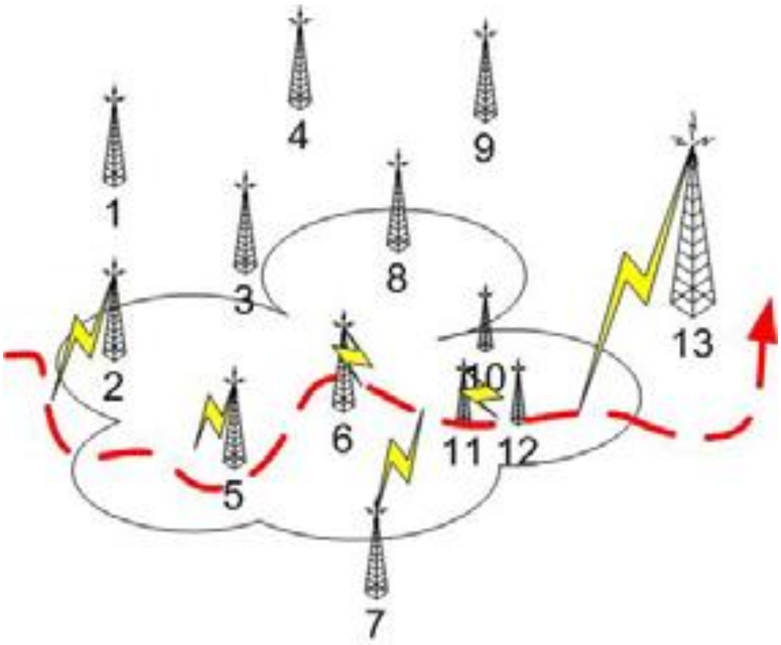


Figure 6.1: UE mobility path [122].

- Type of previous knowledge used for prediction: The exact specifications of the data used, and the knowledge acquired for prediction is very necessary for evaluating and determine how appropriate will be the prediction result. If the predictor needs data that is not available at a certain time interval or it does exist but not accessible due to privacy setting requirements, then the predictor will not output a correct result.
- The type and time of the next event: Prediction can be classified into: a predictor that knows the event and predicts the time of the next occurrence, *e.g.* the predictor knows a handover to another eNodeB is needed but needs to predict the exact moment to trigger, or a predictor that must predict what event or action needs to be adopted in a specific time, *e.g.* the predictor knows it's the right time to trigger the handover but needs to predict the target eNodeB.

- Accuracy level of the prediction: is an essential factor when implementing prediction, it specify how granulated is the prediction, i.e. if prediction is about time then what is the accuracy level (sec, msec, nsec…etc.), or if it's about location then it must match the definition of location in the design (coordinates, eNodeB ID, whole path…etc.)

## 6.2 Related Work

Many handover prediction approaches were proposed in academic and industry, these approaches predict the next station by utilizing combination of many network's Kpi such as RSSI, battery power, latency, and throughput. The RSSI is the most used parameter in the prediction process because of simplicity, easy to be measured.

The authors of [123] presented a prediction method based on RNN with LSTM to predict the next eNodeB for the UE to associate with. The RNN is trained using sequences of RSSI values, for using to predict next eNodeB association, according to the simulation results; the proposed machine learning method achieved an accuracy of over 98% to predict the optimal virtual cell topology in the time required based on the mobility of users.

In the field of UE future location prediction in mobile network to benefit from both intra and inter cell based techniques for network and services enhancement the authors of [124] proposed Intra Cell Movement Prediction (ICMP), this method depends on map based intra-cell prediction and utilizes the network database and handovers history to extract the user trajectories and movement styles. The performance of the proposed algorithm is evaluated and compared with two similar works. The simulation results show that the proposed method is more efficient, and it can also be used to enhance location based services with satisfactory accuracy.

Davaslioglu *et.al.* in [125] considered the deployment of small cells in LTE environment and studied critical issues affecting the performance focusing on cell selection part, they proposed interference-based cell selection algorithm as a solution to provide better load balancing among the base stations in the system to improve the uplink user rates compared to traditional cell selection schemes such as RSRP, CRE and PL-based strategies that don't consider the network traffic load when applied, they presented the implementation steps in a typical LTE network and examined its performance through simulations. According to the simulation results the proposed cell selection algorithm based on interference levels was able to double the cell-edge user SINR and to raise the SINR of the median user by 50% compared to the RSRP-based cell selection.

In [126] the authors proposed call admission control using neural networks to predict the future location of the UE based on its mobility history. The neural network model used error back propagation technique for prediction. The proposed scheme predicts the time and location of the next handover, a call admission control scheme is also used to decide whether a call will be accepted or rejected in the current traffic which is based on the available resources after a handover. According to the results the proposed model increased the ratio of mobility prediction and decreased the call dropping probability and the probability of ping pong effect compared to the conventional schemes.

Kaaniche and Kamoun presented [127] for prediction model in wireless Ad-Hoc network, because mobility prediction is a related to time series prediction the predictor utilized recurrent neural network for long-term time series prediction, this neural network is composite of three layers with feedback and is trained using BPTT algorithm, the nodes mobility style follows Random Waypoint Model (RWM) with variable but limited speed. The depicted results show high similarity between predicted and actual nodes paths.

The work presented in [128] refers to predicting next position of the network user in a GSM network, the applied pattern recognition algorithm used to generate the model is based on the Support Vector Machine (SVM), which provides a powerful solution to nonlinear classification. The SVM is used to calculate the optimal separating hyper plane maximizing the distance between the plane and the training examples, the authors used RapiMiner Data mining framework to implement the model and OneR as a reference algorithm to compare with, and the results demonstrated higher efficiency and accurate next cell prediction.

The work in [129] considers handover prediction between Wi-Fi access points in a portable and completely decentralized way, by exploiting RSSI monitoring and with no need of external global positioning systems GPS. The authors focused on proposing and comparing different filtering techniques for mitigating sudden variations in the Received Signal Strength Indication. Grey Model, Fourier Transform, Discrete Kalman, and Particle, techniques were used for filtering the RSSI.

The authors in [130] proposed a prediction technique dedicated for reducing the handover failure rate and ping-pong handover rate, in order to avoid the delay increasing and burdening eNodeB with too much data. Two network models were simulated to compare the results, the proposed method showed better performance in terms of handover failure rate and ping-pong handover rate.

Similar to [127], the authors of [131] also utilized RSSI, delay and handover cost as parameters for prediction, they used nonlinear autoregressive exogenous model (NARX), and used integer linear programming (ILP) mathematical model to solve the handover decision problem; the solution was then used to train the NARX. After the training the NARX was able to decide handover triggering depending on RSSI and the delay data. The simulation results showed that the proposed handover mechanism can avoid redundant handovers (ping pong effect) within a short time, which leads to better QoS.

In [132] the authors considered the handover procedure between different radio access technology, and the cellular network and the Wi-Fi access points. A comparison between the existing association policies and a policy they proposed based on fairly maximizing the minimal utility of the user and how they achieve sustainability for the network; the minimal utilities were obtained from quality of experience essential values for users with different requirements. According to the results, the proposed algorithm for selection sustainably enhanced the total satisfaction by 62.2%. The problem was considered in more specific cases regarding fairness, worst satisfaction and load balancing.

## 6.3 Handover Prediction

Predicting the target eNodeB before handovers depends on network information such as , radio channels quality , trajectory history and handovers history, these information are updated and stored to the target and source eNodeBs whenever a handover is completed, in addition to that the eNodeBs gather more information from the UEs such as the UE location , RSSI strength , and possible target eNodeBs , because handover is only possible between adjacent or neighbour eNodeBs. Prediction can be also made by using the information and the history of the previous associated eNodeBs.

In order to successfully predict the next UE position, the following information requires to be collected [133]:

- Accurate network domain deployment with the current and previous positions of the UE, *e.g.* similar to GPS coordinates [134].
- Knowledge about UE’s frequently visited locations and time spent, *e.g.* going from/to home to/from work Monday to Friday 9:00 – 17:00, and the usual route

### 6.3.1 Prediction Parameters

In most prediction algorithms the RSSI is used to evaluate the radio link quality, this indicator helps the UE decide the best next eNodeB to associate with. The proposed prediction scheme is based on continuous reporting of the RSSI from the eNodeBs in range. The UE performs the scanning process on frequencies of the MeNodeB and SeNodeB. Continuous scanning will drain the UE battery power faster, especially if there are more than one eNodeB covering the area, additional reasons of power dissipation are Ideal listing, collisions detection, frame errors, protocol overhead, and overhearing.

Two threshold parameters are defined for the predicted handover algorithm, The first threshold ($S\_threshold_{AB}$) is determined by the RSSI value of the source eNodeB when the UE performs the handover, and the second threshold ($P\_threshold_{AB}$) is determined by the RSSI value of the predicted eNodeB at the same time , these two parameters at the specific and due to the fact they are measured and considered at the same time , they would have close but not the same values to each other [135].

The handover is triggered if the target eNodeB offers better radio resources quality than the source eNodeB. at this specific moment the value of the threshold, ($P\_threshold_{AB}$) is considerably higher than the corresponding value of the threshold, ($S\_threshold_{AB}$) and will continue to increase until the handover if finished:

```
If
RSSI of eNodeB$_A$ $<$ $S\_threshold_{AB}$
And
RSSI of eNodeB$_B$ $>$ $P\_threshold_{AB}$
Then
UE is handover from eNodeB$_A$ to eNodeB$_B$
$A$ and $B$ are base station ID
```

The target eNodeB should also accommodate a maximum value of SNR and maintain enough bandwidth at the cell edge in order to be included in the candidate eNodeBs list, otherwise they are listed as blocked or currently not suitable for handover and this applies for the neighbouring eNodeBs without enough bandwidth. The UE keeps scanning the RSSI values of its associated and neighbouring eNodeBs. All determined SNR, RSSI and data rates values are reported to the predictor in the EPC in a continuous manner.

### 6.3.2 Prediction Steps

The predictor compares the available bandwidth at the neighbouring eNodeB(s) and the data rate needed by the UE to switch to perform a successful handover, to calculate a new parameter for preparing the prediction list:

$$D_{rate} = \min(\text{eNodeB}_{bandwidth}, \text{UE}_{Datarate})$$

and the prediction follows the steps below:

1. UE scan RSSI values for source eNodeB and all neighbouring stations and report them to the predictor.
2. If RSSI of the source eNodeB is less than $S\_threshold_{AB}$, then start prediction
3. Sort eNodeBs according to the RSSI values
4. Determine eNodeB $_{bandwidth}$ for the list
5. Obtain D $_{rate}$
6. Sort eNodeBs according to $D_{rate}$ Values

The prediction list is updated continuously to keep optimum number of handovers initiation. Figure 6.2 shows the parameters collection and calculation by the network.

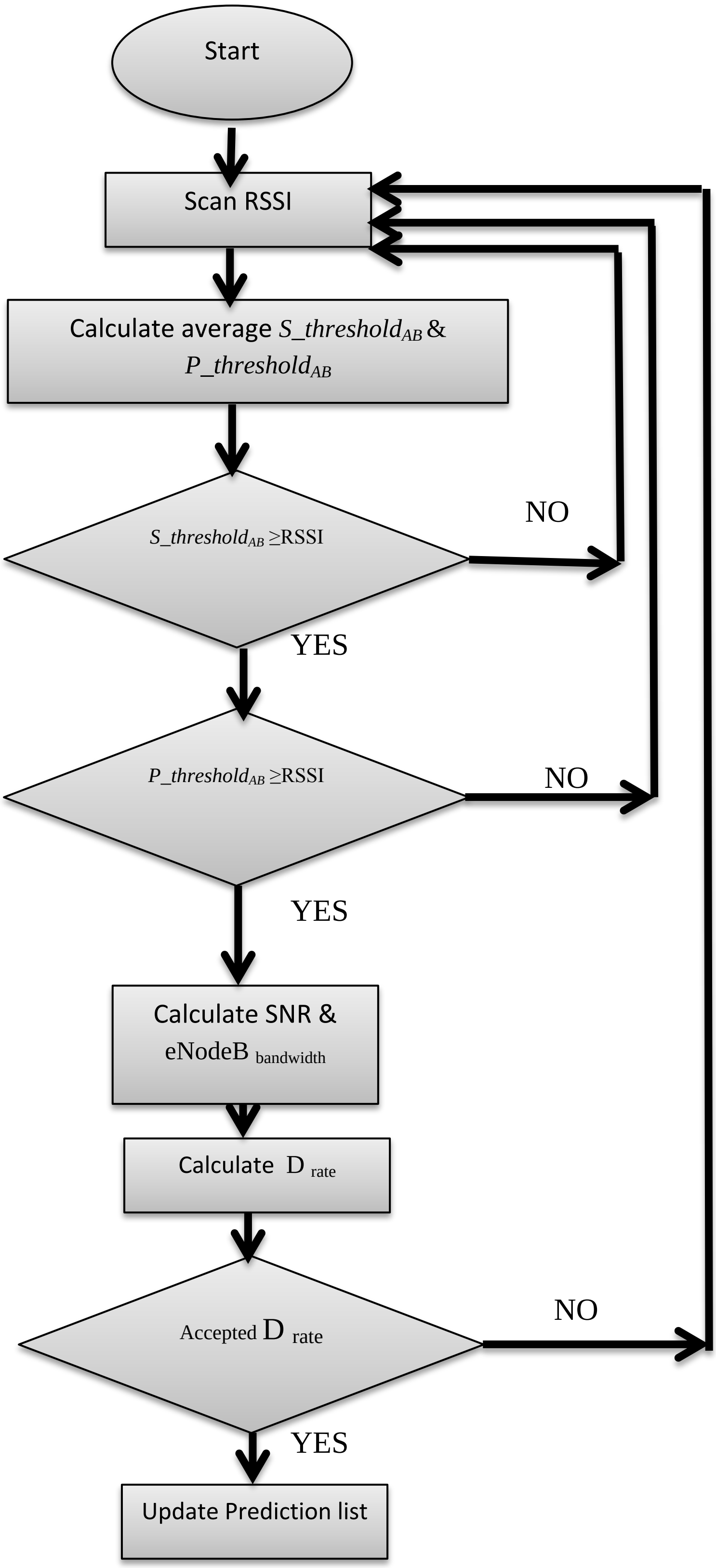

Start
Scan RSSI
Calculate average $S\_threshold_{AB}$ & $P\_threshold_{AB}$
$S_threshold_{AB} \geq$RSSI
NO
YES
$P_threshold_{AB} \geq$RSSI
NO
YES
Calculate SNR & eNodeB bandwidth
Calculate D rate
Accepted D rate
NO
YES
Update Prediction list

Figure 6.2: Prediction parameters collection.

## 6.4 Architecture

The proposed network model is based on the 3GPP LTE Evolved Packet System (EPS) built and tested in Chapter 4 of this work, main components for radio and core networks, (the core network simulated using OPNET is available in the form of only one node called the EPC) remained the same, but two additional entities are connected to the EPC, which will be denoted as Predictor and Balancer as previously explained in Chapter 5 and shown in Figure 5.5. In the model the data is exchanged between the predictor and other network components manually as these is no interface or suitable software that combines the OPNET and MATLAB capabilities.

### 6.4.1 Predictor Architecture

From the definition of prediction; it can be considered as dynamic filter, which uses the previous results to predict the future results. Dynamic neural networks can be used in nonlinear filtering and hence prediction.

MATLAB software provide a readymade tool packages to solve the nonlinear time series and selecting one of these tools depends on the nature of the problem and the characteristics of the tool itself.

- Non-Linear Autoregressive Network with Exogenous inputs (NARX)
- Non-Linear Autoregressive (NAR)
- Non-Linear input-output

The NARX prediction is based on the time series previous values and another time series, while the NAR doesn't have an input time series (only the feedback), and the Non Linear input-output keeps the values from the previous session and use it for the prediction because it has no feedback. Figure 6.3 shows the layer architecture of the NARX tool.

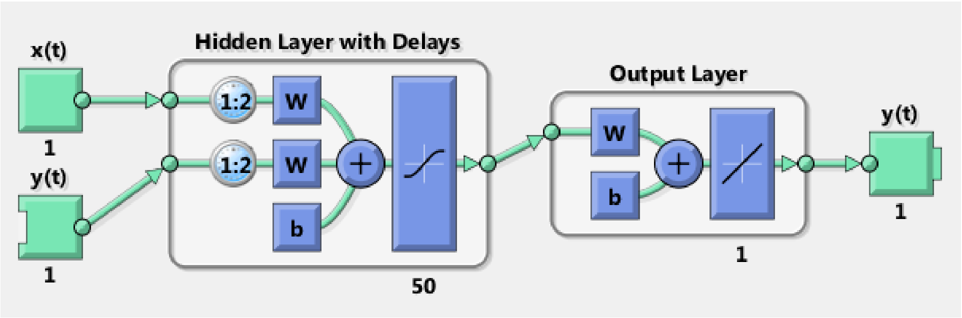


Figure 6.3 Neural time series (NARX) [48].

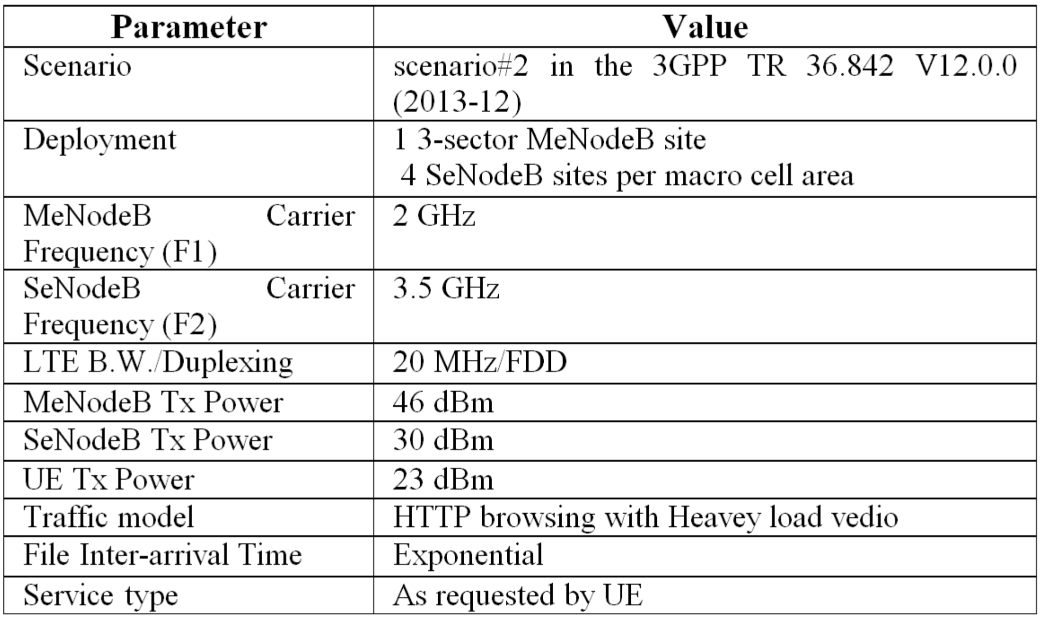
The network must be trained to learn the prediction rules, and since the output sets are also known which means the need for supervised learning and no large amounts of data is required. MATLAB offers 3 learning algorithms for training the neural time series:

- The Levenberg-Marquardt
- Bayesian Regularization
- Scaled conjugate gradient

Each one of these algorithms has different features and must be selected on that basis. The Bayesian algorithm needs more training time, but it returns good generalization solutions especially when difficult or distorted datasets are used, stopping is governed by minimizing the network weights. On the other hand, the Levenberg-Marquardt is faster and very robust for simple problems. While the Scaled conjugate gradient needs less memory and is very effective when small datasets are used, both algorithms stop the training once the generalization stops to improve or when the mean square error of the validation data begins to increase. However; in the handover problem that needs prediction based solution the exact location of the UE is not an essential necessity, the predictor deals with eNodeBs list and, therefore the Levenberg-Marquardt method is used to train the network.

## 6.5 Simulation Setup

In order to fully discuss and compresence the benefits of prediction each of the simulation scenarios has 3 phases, the first is running the network before prediction, training the predictor, and running based on prediction results. The same model that was proposed in chapter 4 is used for performance evaluation; one user is selected to have a predefined path which will be considered for network behaviour investigation in terms of handovers and throughput. All the other nodes were set to have random trajectory. The simulation parameters for dual connectivity setting are customized according to table 6.1.



| Parameter | Value |
|---|---|
| Scenario | scenario#2 in the 3GPP TR 36.842 V12.0.0 (2013-12) |
| Deployment | 1 3-sector MeNodeB site<br>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 | HTTP browsing with Heavey load vedio |
| File Inter-arrival Time | Exponential |
| Service type | As requested by UE |

Table 6.1: Simulation parameters (LTE model with predictor).

The trajectory of the chosen UE is selected as shown in Figure 6.4, and the cruising speed is assumed constant and selected to have an average of 5 miles/hour.

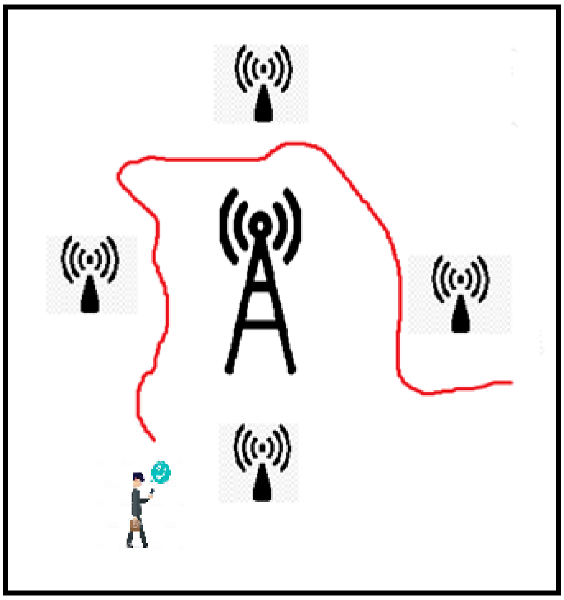


Figure 6.4: Mobility path for UE.

Figure 6.5 shows the associated eNodeB for the entire period of the simulation.

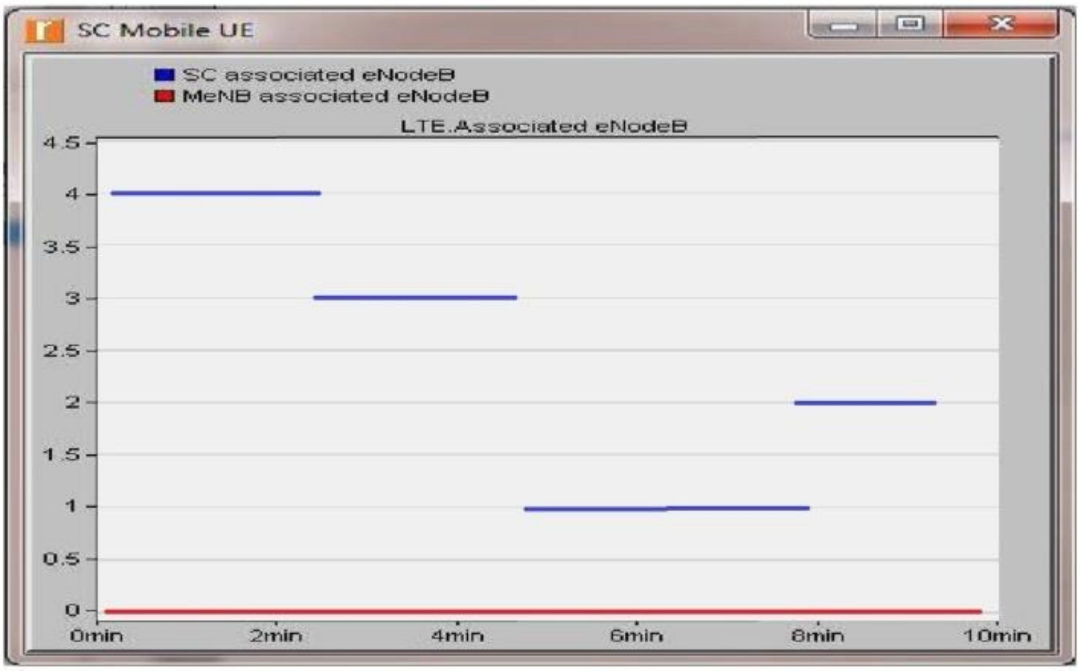


Figure 6.5: Associated eNodeB sequence.

In the second phase the predictor is being trained for handover triggering to the target eNodeB. The network validation and testing data are divided into time steps, these time steps are then grouped for using them in training, validation and testing, the percentiles are set as :70% for training, 15% for validation, and 15% for testing. According to the simulation results, the network did 17 epochs in total and the mean square error was at its minimum magnitude in epoch 11.

The epoch is “the number of times all of the training vectors are used once to update the weights. For batch training all of the training samples pass through the learning algorithm simultaneously in one epoch before weights are updated” [48]. The Mean Squared Error (MSE) is the average squared difference between outputs and targets. Lower values are better. Zero means no error, and Regression R Values measure the correlation between outputs and targets. An R value of 1 means a close relationship, 0 a random relationship. Figure 6.6 shows the performance of the time series [48].

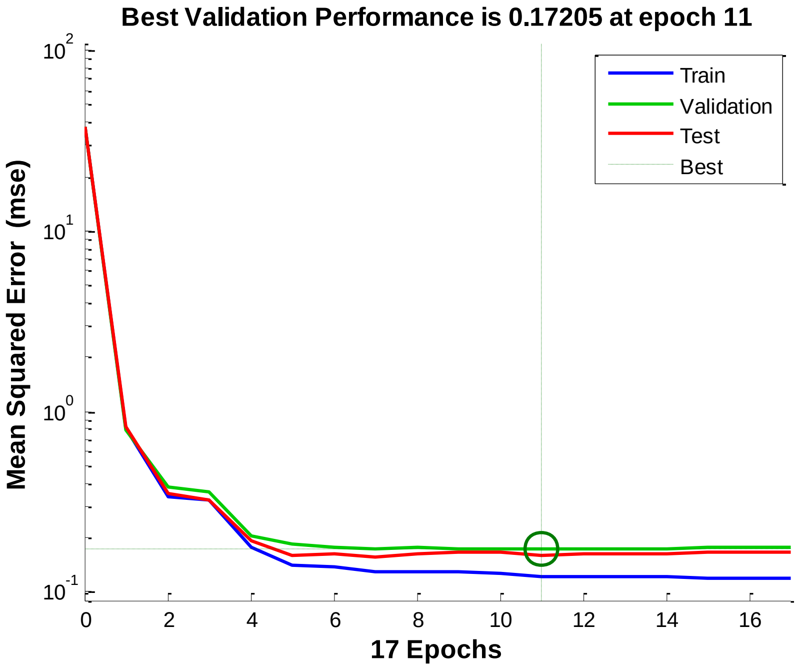


Figure 6.6: Neural time series performance.

Table 6.2 demonstrates the results obtained from the network

| | Target values | Mean Squared Error | Regression |
|---|---|---|---|
| Training | 1400 | 1,2289 $e^{-1}$ | 9,77886 $e^{-1}$ |
| Validating | 300 | 1,72054 $e^{-1}$ | 9,69093 $e^{-1}$ |
| Testing | 300 | 1,60436 $e^{-1}$ | 9,70192 $e^{-1}$ |

Table 6.2: Training results.

The regression is illustrated in Figure 6.7 shows good correlation between the outputs and the target with very small number of errors.

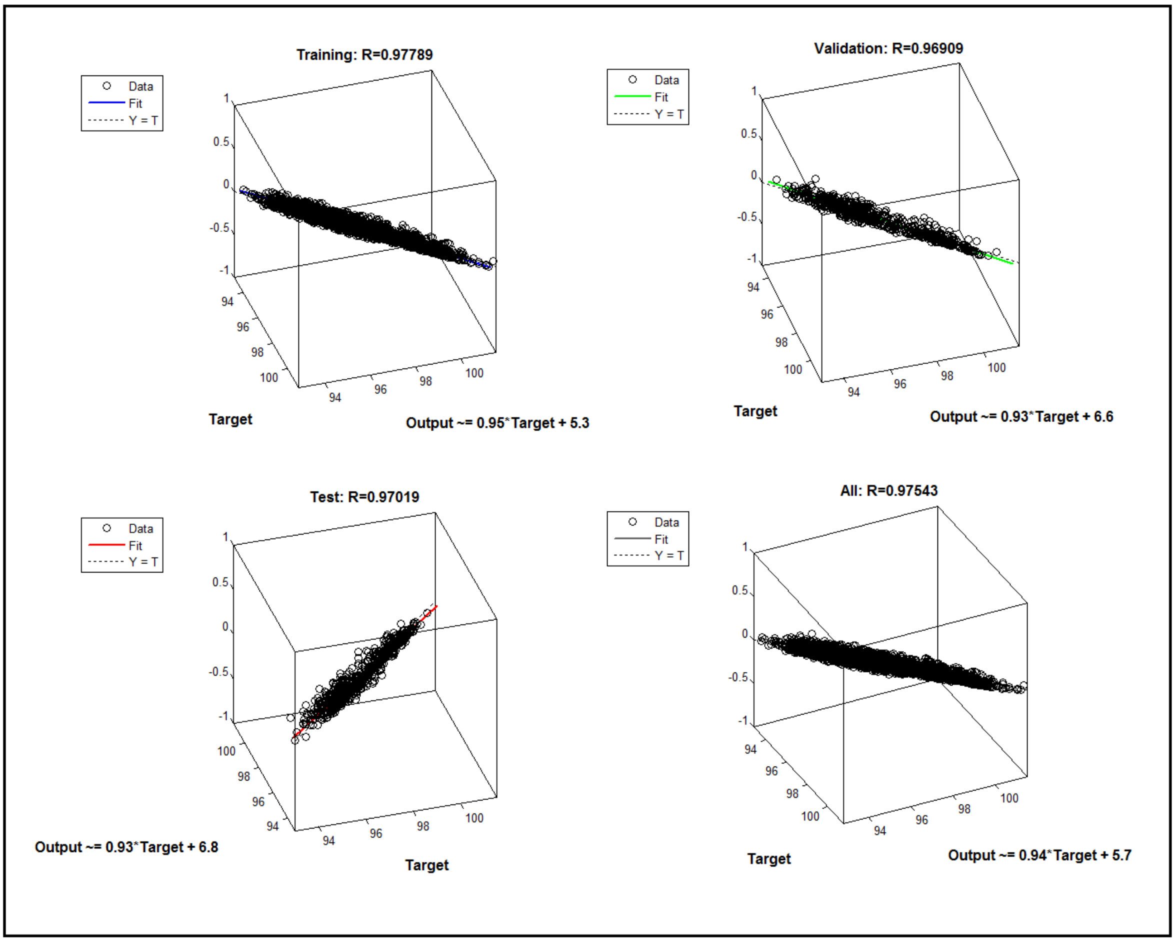


Figure 6.7: Neural Network training regression.

In the third phase, the first scenario is repeated with same trajectory and load conditions but with new information for the associated eNodeB and time, Figure 6.8 shows the actual handovers results and Figure 6.9 shows the predicted ones. From the graphs it's obvious that the neural network was able to accurately predict the sequence of the handovers and the triggering time is also predicted.

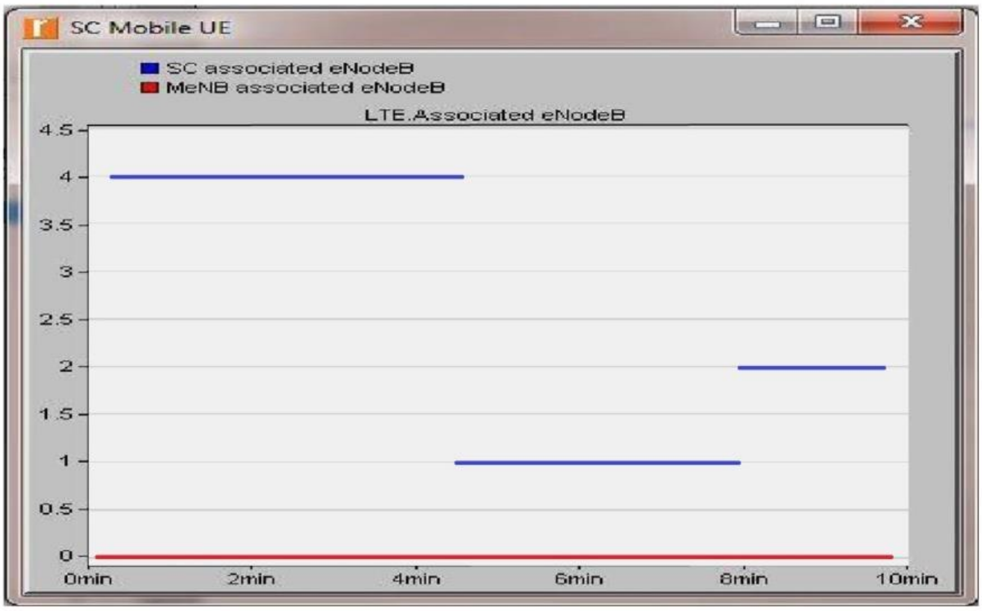


Figure 6.8: Obtained associated eNodeB.

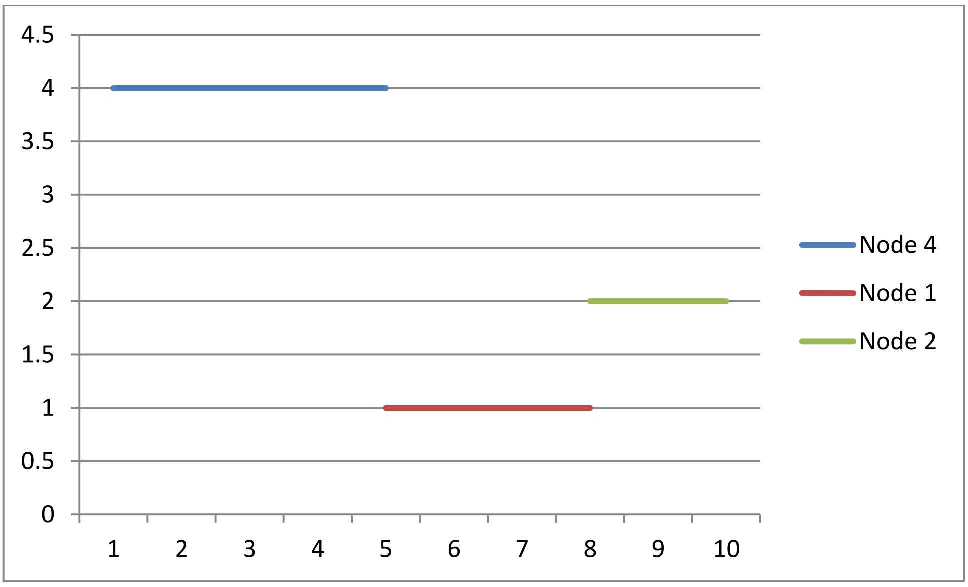


Figure 6.9: Predicted associated eNodeB.

Comparing the results in Figure 6.5 and Figure 6.8, it can be noted that the UE never initiated a handover towards eNodeB 3; this is because the predictor decided for the UE not to perform a handover and associate with eNodeB 3. The prediction is made for the downlink only and the MeNodeB has not been involved in the prediction procedure.

The throughput of the network is evaluated for three scenarios:

- No prediction is made

- Prediction performed with normal load (automatically set by OPNET)
- Prediction performed with full load ( file size is set to 3 Giga Byte)

Figure 6.10 shows the impact of prediction procedure upon the throughput and how it is affected by the reducing the number of handovers.

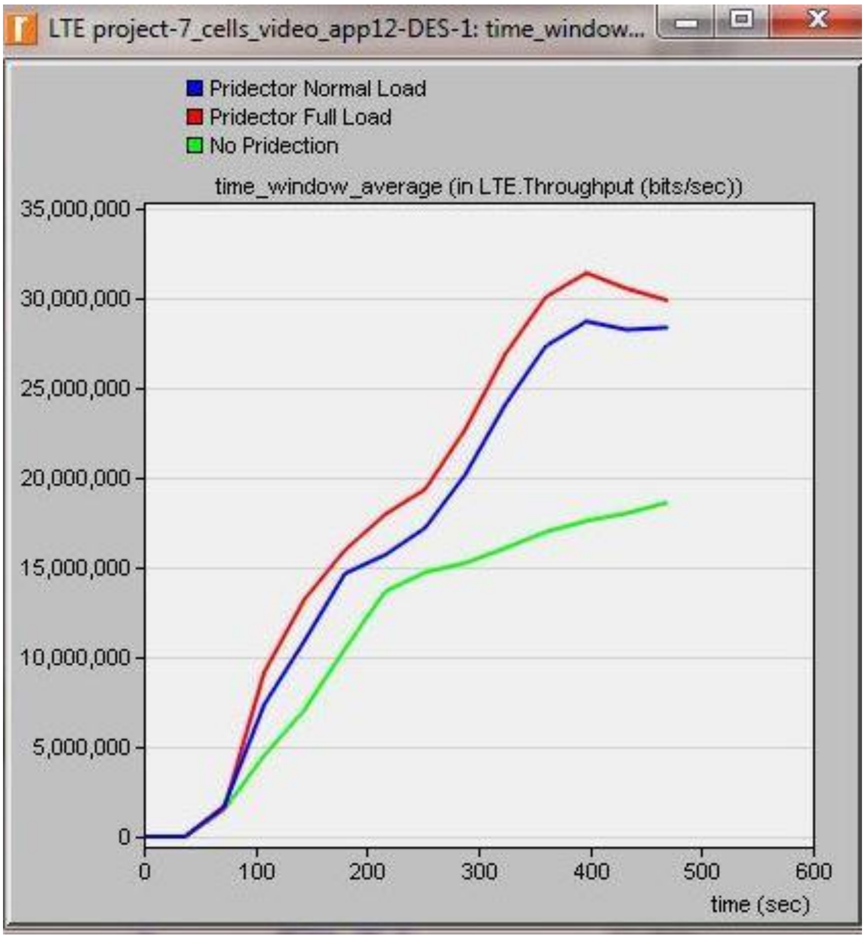


Figure 6.10: Throughput evaluation.

## 6.6 Summary

Predicting the handovers before they occur, and allocate the required data in the target SeNodeB will help the network to utilise its resources in a better way and save time, the predictor entity in the proposed system architecture combines the features of Radial Basis Function Neural Network and neural network time series tool to create and update prediction list from the system's collected data and learn to predict the next SeNodeB to associate with. The prediction entity is simulated using MATLAB, and the simulation results show that the system was able to deliver up to 92% correct predictions for handovers, these predictions assisted increasing the throughput of the network to its optimal value, and it can be observed they led to overall throughput improvement of 50% when the network saturates.

# CHAPTER 7

# CONCLUSION AND FUTURE WORK

## 7.1 Conclusions

While looking forward for the 5G Release the current 4G network will continue to provid services, many of the features of the 4G will also be used in the 5G. This thesis considers maximizing the outcomes and the efficiency of the LTE heterogeneous cellular system by fully utilizing the available resources. Small cells, caching, dual connectivity for separate data and control transmission, artificial intelligence, all these tools have been exploited either separately or jointly to boost the performance of the model framework.

This thesis initially presents the how the heterogeneous deployment of macrocell and smallcell is beneficial and listed the restrictions and solutions when they are in cooperation, followed by building the model and using resource management techniques in the network to present and test the simulations. The general information about the OPNET modeller network simulator, standard deviation and confidence interval, and also network statistical configurations are presented, which are widely used in all the delivered contributions. The behaviour of the system was evaluated in terms of throughput, end to end delay and handover performance.

According to the repeated system simulations, the obtained results specify the following:

- Separating the control and data and allocating small cells to transmit non-live content is proposed as a solution to relief the burden on the network that is caused by multiple requests for the same content, the system performance is evaluated over multiple scenarios using Riverbed simulator to investigate the network from all points and diagnose the optimal solution. According to the results, the proposed design had effectively reduced the delay and increased the throughput, while it had a higher impact when caching units were distributed and brought closer to the UE, the concept of caching is similar to the one used in microprocessors for short term content storing, but the policy is different because the requests in the mobile network are governed by the users.

- For the same network model and to furtherly increase the efficiency and optimise the network, Artificial neural network clustering tool with self-organizing map had been introduce to balance the load over the small cells, The balancer consists of three layers, Input, Self-Organizing map SOM, and output layer. The proposed balancer designed using neural network had successfully balanced the load among the clusters; the influence was sensed through consistency and the drop in the number of handovers, and reduction in the end to end delay in each cluster. The handover mechanism requires resources and signalling, it sometimes backfires by keeping the UEs at the cell edge switching between the cells (ping pong effect); therefore any reduction helps saving network resources, reducing signalling overhead and saving UE battery power.
- The third approach for enhancing the performance of the network is proposed using another form of artificial neural network techniques, neural networks are like a black box, they can learn and derive the relations between inputs and return outputs, similar to the human way of thinking. For the proposed model and when the users are in high mobility and frequent handovers, in order to maintain uninterrupted services optimising the available resources is very necessary. The proposed solution is by using Artificial Neural network time series for mobility prediction, predicting the handovers before they occur, and allocate the required data in the target SeNodeB will help the network to utilise its resources in a better way and save time. The designed prediction entity was able to successfully predict the next target cell to associate with when there is a need for a handover. Though the exact position was not desperately required, the new position could be any spot within the coverage area; the ANN did the prediction very accurate with very small error (8% of total attempts). The prediction resulted in rising the throughput by 50%, because no additional data was injected into the network the throughput had risen due to the time saving which is normally wasted for handover preparation when there is no prediction.

The overall performance evaluation showed robustness when the proposed add-ons are added, and the network survivability scored 100% as shown in Figure 7.1 . Survivability is defined as the ability of a system to minimize the impact of a finite disturbance on value delivery, achieved through either (1) the satisfaction of a minimally acceptable level of value delivery during and after a finite disturbance or (2) the reduction of the likelihood or magnitude of a disturbance

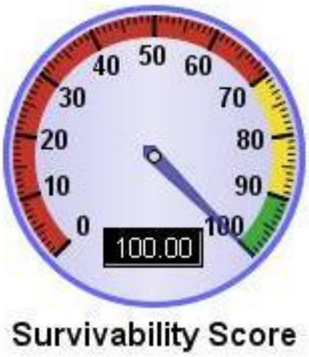


Figure 7.1 : Survivability Score

## 7.2 Future Work

Integrating AI with wireless network was considered from two aspect, balancing and prediction. The AI capabilities nowadays exceeded the expectations and the learning algorithms are in continuous evolution. But for the proposed system and due to platforms incompatibility between OPNET and MATLAB, the dialogues and data transfer were exchanged  manually, as part of the future work concerning this thesis, an interface that is capable of linking OPNET and MATLAB is needed , this will convert the whole system and make it operate  in a more convenient way. The presence of the interface will also make it possible to operate both the balancer and predictor at the same time. Testing more learning algorithms and different ANN designs to provide solutions for various optimization problems in the wireless network , for example ; the predictor can be designed to predict the type of the content to be cached based on the first time it was broadcasted live , and predict how long time to keep it in the cache, are to be considered in the future researches or publications as well as possible interaction from different AI technologies usage, such as: genetic algorithms, particle swarm optimization, and ant colony.

# REFERENCES

1. Rappaport, T. S. Wireless communications: Principles and practice. Upper Saddle River, N.J: Prentice Hall PTR, 2002.
2. C. X. Mavromoustakis, G. Mastorakis, and C. Dobre, “Advances in Mobile Cloud Computing and Big Data in the 5G Era, vol. 22, 2017.
3. Chris Johnson, “Long Term Evolution in Bullets”, 2nd ed., ver. 1. Northampton: England, (2012).
4. Sesia S, Toufik I, Books24x7 I. LTE – The UMTS Long Term Evolution From Theory to Practice, Second Edition. 2nd ed. John Wiley & Sons; 2011.
5. Timo Halonen, Javier Romero, Juan Melero GSM, GPRS, and edge performance: evolution towards 3G/UMTS / 2nd edition.
6. Friedhelm Hillebrand, GSM and UMTS: The Creation of Global Mobile Communication Copyright q 2001 John Wiley & Sons Ltd.
7. ETSI TS 136 300, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description," 3GPP, vol. 11.5.0, 2013.
8. S. Singh; N. Saxena; A. Roy; P. De “Proximity-based video delivery architecture for LTE networks” Electronics Letters Year: 2016 Volume: 52, Issue: 11 Pages: 984 - 986 IET Journals & Magazines.
9. 3GPP TS 36.141 version 13.6.0 Release 13 (Evolved Universal Terrestrial Radio Access (E-UTRA); Base stations conformance testing )
10. 3GPP TS 36.302 version 14.4.0 Release 14 (Evolved Universal Terrestrial Radio Access (E-UTRA); Services provided by the physical layer)
11. H. Holma and A. Toskala, “LTE for UMTS: Evolution to LTE-Advanced”. John Wiley & Sons, 2011.
12. The LTE Network Architecture, A comprehensive tutorial, Alcatel-Lucent
13. A. Perez, “LTE and LTE advanced: 4G network radio interface”; Wiley-ISTE, 2012
14. ETSI TS 123 401 V12.6.0 (2014-09), General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access.
15. C. Cox, “An Introduction to LTE: LTE, LTE-Advanced”, SAE and 4G Mobile Communications, Wiley, 2012
16. Alcatel-Lucent, The LTE Network Architecture, A comprehensive tutorial.

17. 3GPP TS 36.106 version 10.0.0 Release 10 (Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN);Overall description; Stage 2 ).
18. ETSI TS 136 213, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 3GPP, 2013.
19. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); FDD repeater radio transmission and reception (3GPP TS 36.106 version 10.0.0 Release 10).
20. Mehlfuhrer, et al., The Vienna LTE Simulators – Enabling Reproducibility in Wireless Communications Research; EURASIP Journal on Advances in Signal Processing, Vol. 29, July 2011.
21. H. Wang, C. Rosa, and K. Pedersen, “Uplink Component Carrier Selection for LTE-Advanced Systems with Carrier Aggregation,” in IEEE Int. Conf. Comm. (ICC), pp. 1 –5.
22. ETSI TR 136 913 V9.0.0 (2010-02)
23. Guillaume de la Roche, Alan Taylor,” A new wave in wireless: Small cells for a heterogeneous network”, Mindspeed Technologies, Inc. November 2011
24. 3GPP TR 36.932 V14.0.0 (2017-03)
25. S. Brueck,” Heterogeneous networks in LTE advanced “, Invited paper, IEEE 8th International Symposium on Wireless Communication Systems, Aachen
26. J. Zang, Q. Zeng, T. Mahmoodi, A. Georgakopoulos, P. Demestichas. “ LTE Small Cell Enhancement by Dual Connectivity” ,white paper November 2014
27. D. Sabella, et al. “Mobile-Edge Computing Architecture: The role of MEC in the Internet of Things”. IEEE Consumer Electronics Magazine. 5. 84-91. 10.1109/MCE.2016.2590118.
28. Mobile Edge computing use cases and & deployment options, Juniper white paper, July 2016.
29. ETSI White Paper No. 24 MEC Deployments in 4G and Evolution Towards 5G, First edition – February 2018.
30. W. Kim, “Dual Connectivity in Heterogeneous Small Cell Networks with mmWave Backhauls” Mobile Information Systems Volume 2016, Article ID 3983467, dx.doi.org/10.1155/2016/3983467
31. S. C. Jha, K.i Sivanesan, R. Vannithamby and A. Koc.” Dual Connectivity in LTE Small Cell Networks” IEEE GLOBCOM 2014

32. V. D’Amico, et al., “Advanced interference management in ARTIST4G: Interference Avoidance” 3rd European Wireless Technology Conference,2010

33. I. Shayea, M. Ismail, R. Nordin, “Advanced Handover Techniques in LTE- Advanced system” IEEE International Conference on Computer and Communication Engineering (ICCCE 2012).

34. J. Zhang, X. Zhang, W. Wang “Cache-Enabled Software Defined Heterogeneous Networks for Green and Flexible 5G Networks” IEEE Access 10.1109/ACCESS.2016.2588883.

35. L. Lei, X. Xiong, L. Hou, K. Zheng,” Collaborative Edge Caching through Service Function Chaining: Architecture and Challenges” IEEE Wireless Communications , June 2018

36. Y. Xu, et al: “Coordinated Caching Model for Minimizing Energy Consumption in Radio Access Network” IEEE ICC 2014 - Mobile and Wireless Networking Symposium

37. C. Kwon, “Julia Programming for Operations Research: A Primer on Computing”, chapter 9.

38. Y. Gao, “Energy Efficient Content Aware Cache and Forward Operation in 3GPP LTE-Advanced Base Stations”, IEEE 2013 3rd International Conference on Computer Science and Network Technology.

39. Q.Tang, R. Xie, Ta. Huang, Y. Liu, “Hierarchical collaborative caching in 5G Networks”, IET Commun., 2018, Vol. 12 Iss. 18, pp. 2357-2365

40. G. Nardini, G. Stea, A. Virdis, M. Caretti, D Sabell, “Improving Network Performance via Optimization-Based Centralized Coordination of LTE-A Cells”, IEEE WCNC 2014.

41. C. He, G. Arrobo, R. Gitlin, “Improving System Capacity Based upon User-Specific QoS for Heterogeneous Networks” 2015 IEEE Wireless Communications and Networking Conference (WCNC).

42. S. Russell, P. Norvig. 1995.” Artificial Intelligence—A Modern Approach”. Upper Saddle River, N.J.: Prentice Hall.

43. M. Ginsberg, 1993. “Essentials of Artificial Intelligence”. San Francisco, Calif.: Morgan Kaufmann Publishers.

44. E. Alpaydin, “Introduction to machine learning”. MIT press, 2014.

45. A. Ferdowsi, U. Challita, and W. Saad, “Deep learning for reliable mobile edge analytics in intelligent transportation systems" arXiv:1712.04135, Dec. 2017.

46. H. Demuth, M. Beale, O. De Jess, M. Hagan, “Neural network Design “second edition. 2014.

47. K. Gurney, “An introduction to Neural Networks” 2004 ISBN 0-203-45151-1 Master e-book

48. M. Beale, M. Hagan, H. Demuth “Neural Network Toolbox User’s Guide” ,Version 7 , MathWorks Inc, 2010.

49. P. Winston, “Artificial Intelligence”, Third Edition, Addison Wesley publishing, 1992.

50. D. Kriesel “A Brief Introduction to Neural Networks”, 2005, http://www.dkriesel.com/en/science/neural_networks.

51. B. Muller, J. Reinhardt, M. Strickland, “Neural networks: An introduction”, Springer Science & Business Media, 2012.

52. M.Nielsen, “Neural Networks and Deep Learning”, online book, June 2019 ,http://neuralnetworksanddeeplearning.com.

53. I. Goodfellow, Y. Bengio, A. Courville, ”Deep Learning” An MIT Press book, ISBN-13: 978-0262035613

54. S. Song, K. Chaudhuri, A. Sarwati, “Stochastic gradient descent with differentially private updates”, 2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP), DOI: 10.1109/GlobalSIP.2013.6736861.

55. S. Ruder, “An overview of gradient descent optimization algorithms”, Insight Centre for Data Analytics, NUI Galway Aylien Ltd., Dublin, June. 2017.

56. I. Sutskever, J. Martens, G. Dahl, G. Hinton, “On the importance of initialization and momentum in deep learning”, 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

57. A. Botev, G. Lever, D. Barber, “Nesterov's accelerated gradient and momentum as approximations to regularised update descent”, International Joint Conference on Neural Networks (IJCNN) Volume: 2017,10.1109/IJCNN.2017.7966082.

58. T. Kavzoglu, P. Mather, “Assessing artificial neural network pruning algorithms”, 24th Annual Conference and Exhibition of the Remote Sensing Society, Greenwich, UK, pp. 603-609

59. J. Mazumdar, R. Harley, "Recurrent Neural Networks Trained With Backpropagation Through Time Algorithm to Estimate Nonlinear Load Harmonic Currents", IEEE Transactions on Industrial Electronics, vol. 55, no. 9, pp. 3484-3491, Sept. 2008. doi: 10.1109/TIE.2008.925315

60. S. Haykin, “Neural networks and learning machines”,3rd edition, Prentice Hall 2009, ISBN-13: 978-0-13-147139-9

61. X. Glorot,Y. Bengio,”Understanding the difficulty of training deep feedforward neural networks”, 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Italy. Volume 9 of JMLR: W&CP 9

62. R. Miikkulainen, J. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzyan, N. Duffy, B. Hodjat,”Artificial Intelligence in the Age of Neural Networks and Brain Computing”, Sentient Technologies, Inc., San Francisco, CA, United States, 2018 https://doi.org/10.1016/B978-0-12-815480-9.00015-3

63. A. Baliyan,K. Gaurav, S. Mishra, “A Review of Short Term Load Forecasting using Artificial Neural Network Models”, International Conference on Intelligent Computing, Communication & Convergence (ICCC-2015)

64. H. Sak, A. Senior, F. Beaufays,” Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling”, 15th Annual Conference of the International Speech Communication Association, 2014

65. M. Riedmiller,” Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms”, Elsevier, Computer Standards & Interfaces Volume 16, Issue 3, 1994, Pages 265-278

66. M. Zeiler, “ADADELTA: AN ADAPTIVE LEARNING RATE METHOD”, Google Inc., USA, arXiv:1212.5701v1 [cs.LG] , 2012.

67. D. P. Mandic, "A generalized normalized gradient descent algorithm," in *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 115-118, 2004, doi: 10.1109/LSP.2003.821649

68. Y. Xu, Q.Huang, W. Wang, M.. Plumbley,” HIERARCHICAL LEARNING FOR DNN-BASED ACOUSTIC SCENE CLASSIFICATION”, Detection and Classification of Acoustic Scenes and Events 2016, experiarXiv:1607.03682v3 [cs.SD] 13 Aug 2016

69. C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi,” Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”, Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

70. K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, W. Xiang, “Big data-driven optimization for mobile networks toward 5G”. IEEE network, 30(1):44–51, 2016.

71. C. Jiang, H. Zhang, Y. Ren, Z. Han, K.Chen, L. Hanzo, “Machine learning paradigms for next generation wireless networks”. IEEE Wireless Communications, 24(2):98–105, 2017

72. A. Hussein, M. Gaber, E. Elyan, C Jayne, “Imitation learning: A survey of learning methods”, ACM Computing Surveys (CSUR), 50(2):21:1–21:35, 2017.

73. O. Naparstek, K. Cohen, “Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks”, IEEE GLOBCOM, 2017.

74. M. A. Wijaya, K. Fukawa and H. Suzuki, "Intercell-Interference Cancellation and Neural Network Transmit Power Optimization for MIMO Channels," *2015 IEEE 82nd Vehicular Technology Conference (VTC2015),* 2015, pp. 1-5. doi: 10.1109/VTCFall.2015.7390988

75. M. A. Wijaya, K. Fukawa and H. Suzuki,” Neural Network Based Transmit Power Control and Interference Cancellation for MIMO Small Cell Networks”, IEICE Transactions on Communications, 2016, Volume E99.B, Issue 5, Pages 1157-1169, Released May 01, 2016, Online ISSN 1745-1345, Print ISSN 0916-8516

76. H. Rutagemwa, A. Ghasemi and S. Liu, "Dynamic Spectrum Assignment for Land Mobile Radio with Deep Recurrent Neural Networks," *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, Kansas City, MO, 2018, pp. 1-6.doi: 10.1109/ICCW.2018.8403659

77. N. Luong, T. Anh, H. Binh, D. Niyato, D. Kim, Y. Liang,” Joint Transaction Transmission and Channel Selection in Cognitive Radio Based Blockchain Networks: A Deep Reinforcement Learning Approach”, arXiv preprint arXiv:1810.10139, 2018

78. H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, N. Sidiropoulos,” Learning to optimize: Training deep neural networks for wireless resource management”, 18th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pages 1–6, 2017.

79. U. Challita, L. Dong and W. Saad, "Proactive Resource Management for LTE in Unlicensed Spectrum: A Deep Learning Perspective", in *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4674-4689, July 2018. doi: 10.1109/TWC.2018.2829773

80. Y. Zhou, Z. M. Fadlullah, *B. Mao and N. Kato, "A Deep-Learning-Based Radio Resource Assignment Technique for 5G Ultra Dense Networks," in IEEE Network, vol. 32, no. 6, pp. 28-34,2018. doi: 10.1109/MNET.2018.1800085*

81. *Z. Xu, Y. Wang, J. Tang, J. Wang, M. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs" ,2017 IEEE International Conference on Communications (ICC), pp. 1-6. doi: 10.1109/ICC.2017.7997286*

82. R. Atallah, C. Assi and M. Khabbaz, "Deep reinforcement learning-based scheduling for roadside communication networks" *,15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), Paris, 2017, pp. 1-8.doi: 10.23919/WIOPT.2017.7959912*

83. *Y. Wei, Z. Zhang, F. R. Yu and Z. Han, "Joint User* Scheduling and Content Caching Strategy for Mobile Edge Networks Using Deep Reinforcement Learning," *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1-6. doi: 10.1109/ICCW.2018.8403711

84. R. Mennes, M. Camelo, M. Claeys and S. Latré, "A neural-network-based MF-TDMA MAC scheduler for collaborative wireless networks," *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, 2018, pp. 1-6. doi: 10.1109/WCNC.2018.8377044

85. H. Yang, Z. Li, Z.g Liu, “Neural networks for MANET AODV: an optimization approach”, Cluster Computing, pp 1–9, 2017, 20:3369–3377, doi: 10.1007/s10586-017-1086-y

86. B. Mao, Z. Fadlullah, F. Tang, N. Kato, O. Akashi, T. Inoue, K. Mizutani, "Routing or Computing The Paradigm Shift Towards Intelligent Computer Network Packet Transmission Based on Deep Learning," *IEEE Transactions on Computers*, vol. 66, no. 11, pp.1946-1960, 2017. doi: 10.1109/TC.2017.2709742

87. B. Mao, Z. Fadlullah, F. Tang, N. Kato, O. Akashi, T. Inoue, K. Mizutani,” "A Tensor Based Deep Learning Technique for Intelligent Packet Routing," *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1-6. doi: 10.1109/GLOCOM.2017.8254036

88. Y. He, C. Liang, F. R. Yu, N. Zhao and H. Yin, "Optimization of cache-enabled opportunistic interference alignment wireless networks: A big data deep reinforcement learning approach," *2017 IEEE International Conference on Communications (ICC)*, Paris, 2017, pp. 1-6. doi: 10.1109/ICC.2017.7996332

89. Y. He, C. Liang, F. R. Yu, N. Zhao and H. Yin, V. Leung, and Y. Zhang, “Deep-Reinforcement-Learning-Based Optimization for Cache-Enabled Opportunistic Interference Alignment Wireless Networks", *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp.10433-10445, doi: 10.1109/TVT.2017.2751641

90. Z. Wang, L. Li, Y. Xu, H. Tian and S. Cui, "Handover Optimization via Asynchronous Multi-User Deep Reinforcement Learning," *2018 IEEE International Conference on*

*Communications (ICC)*, Kansas City, MO, 2018, pp. 1-6. doi: 10.1109/ICC.2018.8422824

91. 3GPP TR 36.842 V12.0.0 (2013-12).

92. 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. doi: 10.1109/FGCT.2017.8103399

93. R. Sakat, R. Saadoon, M. Abbod (2019) Small Cells Solution for Enhanced Traffic Handling in LTE-A Networks. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Computing. SAI 2018. Advances in Intelligent Systems and Computing, vol 857. Springer, Cham, doi.10.1007/978-3-030-01177-2-43

94. G. Nardini, G. Stea, A. Virdis, M. Caretti, and D. Sabell, "Practical large-scale coordinated scheduling in LTE-Advanced networks", Wireless Networks, Springer US, January 2016, Volume 22, Issue 1, pp 11–31, https://doi.org/10.1007/s11276-015-0948-6

95. G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley and G. Caire, "The Role of Caching in Future Communication Systems and Networks," in *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1111-1125, June 2018. doi: 10.1109/JSAC.2018.2844939

96. E. Baştuğ, M. Bennis and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, pp. 649-653. doi: 10.1109/ISWCS.2014.6933434

97. D. Henriet, and H. Moulin. "Traffic-Based Cost Allocation in a Network." The RAND Journal of Economics, vol. 27, no. 2, 1996, pp. 332–345. JSTOR, www.jstor.org/stable/2555930.

98. M. Ji, G. Caire and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," *2013 IEEE International Symposium on Information Theory*, Istanbul, 2013, pp. 1461-1465. doi: 10.1109/ISIT.2013.6620469

99. A. Pyattaev, K. Johnsson, S. Andreev and Y. Koucheryavy, "3GPP LTE traffic offloading onto WiFi Direct," *2013 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Shanghai, 2013, pp. 135-140. doi: 10.1109/WCNCW.2013.6533328

100. ETSI TS 132 593, V9.0.0, "Home enhanced Node B (HeNB) Operations, Administration, Procedure flows for Type 1 interface HeNB to HeNB Management System (HeMS)", (2010-02)

101. Sakat R., Saadoon R., Abbod M. (2020) Load Balancing Using Neural Networks Approach for Assisted Content Delivery in Heterogeneous Network. In: Bi Y., Bhatia R., Kapoor S. (eds) Intelligent Systems and Applications. IntelliSys 2019. Advances in Intelligent Systems and Computing, vol 1038. Springer, Cham

102. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022

103. Sadoon, R.S.: Explosion of data (BIGDATA), Chapter "3" in the Internet of Things and Big Data Analysis: Recent Trends and Challenges", November 2016. ISBN-10:0692809929

104. A. A. Neghabi, N. Jafari Navimipour, M. Hosseinzadeh and A. Rezaee, "Load Balancing Mechanisms in the Software Defined Networks: A Systematic and Comprehensive Review of the Literature," in *IEEE Access*, vol. 6, pp. 14159-14178, 2018. doi: 10.1109/ACCESS.2018.2805842

105. S. Marwat, S. Meyer, T. Weerawardane and C. Görg, "Congestion-Aware Handover in LTE Systems for Load Balancing in Transport Network", ETRI Journal. 36. 761-771. (2014). 10.4218/etrij.14.0113.1034

106. T. Warabino, S. Kaneko, S. Nanba and Y. Kishi, "Advanced load balancing in LTE/LTE-A cellular network," *2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications - (PIMRC)*, Sydney, NSW, 2012, pp. 530-535. doi: 10.1109/PIMRC.2012.6362842

107. ETSI TS 136 331 V11.3.0 (2013-04), LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification

108. C. Lin, K. Sandrasegaran, H. Ramli, R. Basukala, "Optimized performance evaluation of LTE hard handover algorithm with average RSRP constraint", International Journal of Wireless & Mobile Networks (IJWMN) Vol. 3, No. 2, DOI: 10.5121/ijwmn.2011.3201.

109. 3GPP Technical Specification 36.300 V 11.5.0," Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description"; Stage 2

110. W. Song, W. Zhuang and Y. Cheng, "Load balancing for cellular/WLAN integrated networks," in *IEEE Network,* vol. 21, no. 1, pp. 27-33, Jan.-Feb. 2007. doi: 10.1109/MNET.2007.314535

111. Q. Liu, J. Yuan, X. Shan, Y. Wang and W. Su, "Dynamic load balance scheme based on mobility and service awareness in integrated 3G/WLAN networks," *2010 Global Mobile Congress,* Shanghai, 2010, pp. 1-6. doi: 10.1109/GMC.2010.5634553

112. M. Ismail and W. Zhuang, "A Distributed Multi-Service Resource Allocation Algorithm in Heterogeneous Wireless Access Medium," in *IEEE Journal on Selected Areas in Communications,* vol. 30, no. 2, pp. 425-432, February 2012. doi: 10.1109/JSAC.2012.120222

113. V. Angelakis, I. Avgouleas, N. Pappas, E. Fitzgerald and D. Yuan, "Allocation of Heterogeneous Resources of an IoT Device to Flexible Services," in *IEEE Internet of Things Journal,* vol. 3, no. 5, pp. 691-700, Oct. 2016. doi: 10.1109/JIOT.2016.2535163

114. X. Song, L. Wu, X. Ren, and J. Gao. “Load Balancing Algorithm Based on Neural Network in Heterogeneous Wireless Networks”. In: Hu X., Xia Y., Zhang Y., Zhao D. (eds) Advances in Neural Networks – ISNN 2015. ISNN 2015. Lecture Notes in Computer Science, vol. 9377. Springer, Cham, 2015.

115. R. Chai, X. Y. Dong, J. Ma, & Q. B. Chen. An optimal IASA load balancing scheme in heterogeneous wireless networks, In Proceedings of 6th International ICST Conference on Communications and Networking in China (CHINACOM), 2011.

116. R. Chai, H., Zhang, X., Dong, et al. “Optimal joint utility based load balancing algorithm for heterogeneous wireless networks”, 2014. 20: 1557. https://doi.org/10.1007/s11276-014-06950.

117. Y. Hussein, B. Ali, M. Rasid, A.Sali, “Handover in LTE Networks with Proactive Multiple Preparation Approach and Adaptive Parameters Using Fuzzy Logic Control”, KSII Transactions on Internet and Information Systems. vol. 9, No.7, July 31, 2015 ,10.3837/tiis.2015.07.004

118. N. M. Al-Sallami, A. Al-Daoud and S. A. Al-Alousi, “Load Balancing with Neural Network” International Journal of Advanced Computer Science and Applications(IJACSA), 4(10), 2013.http://dx.doi.org/10.14569/IJACSA.2013.041021

119. U. Challita, M. Marina,” Holistic Small Cell Traffic Balancing across Licensed and Unlicensed Bands”, MSWiM ’16, November 13-17, 2016, Malta, 2016 ACM. ISBN 978-1-4503-4502, http://dx.doi.org/10.1145/2988287.2989143

120. 3GPP TS 36.214 V8.6.0 (Release 8).

121. O. Altrad, S. Muhaidat, "Load balancing based on clustering methods for LTE networks", *Cyber Journals: J. of Selected Areas in Telecommun.*, vol. 3, no. 2, pp. 1-6, 2013.

122. Capka J., Boutaba R. (2004) Mobility Prediction in Wireless Networks Using Neural Networks. In: Vicente J., Hutchison D. (eds) Management of Multimedia Networks and Services. MMNS 2004. Lecture Notes in Computer Science, vol 3271. Springer, Berlin, Heidelberg

123. D. S. Wickramasuriya, C. A. Perumalla, K. Davaslioglu and R. D. Gitlin, "Base station prediction and proactive mobility management in virtual cells using recurrent neural networks," *2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON),* Cocoa Beach, FL, 2017, pp. 1-6. doi: 10.1109/WAMICON.2017.7930254.

124. M. Abo-Zahhad, S. M. Ahmed and M. Mourad, "Future location prediction of mobile subscriber over mobile network using Intra Cell Movement pattern algorithm," *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA),* Sharjah, 2013, pp. 1-6. doi: 10.1109/ICCSPA.2013.6487272

125. K. Davaslioglu and E. Ayanoglu, "Interference-based cell selection in heterogeneous networks," *2013 Information Theory and Applications Workshop (ITA),* San Diego, CA, 2013, pp. 1-6. doi: 10.1109/ITA.2013.6502931

126. S. Kumar, K. Kumar and P. Kumar, "Mobility based call admission control and resource estimation in mobile multimedia networks using artificial neural networks," *2015 1st International Conference on Next Generation Computing Technologies (NGCT),* Dehradun, 2015, pp. 852-857. doi: 10.1109/NGCT.2015.7375240

127. H. Kaaniche, F. Kamoun, “Mobility Prediction in Wireless Ad Hoc Networks using Neural Networks”, Journal of Telecommunication, vol 2, iss. 1, April 2010.

128. Stefan Michaelis, Nico Piatkowski, Katharina Morik,” Predicting next network cell IDs for moving users with Discriminative and Generative Models”, Mobile Data Challenge 2012 (by Nokia) Workshop; June 18-19, 2012; Newcastle, UK

129. P. Bellavista, A. Corradi and C. Giannelli, "Evaluating Filtering Strategies for Decentralized Handover Prediction in the Wireless Internet," *11th IEEE Symposium on Computers and Communications (ISCC'06),* Cagliari, Italy, 2006, pp. 167-174. doi: 10.1109/ISCC.2006.70

130. Y. Wang, J. Chang and G. Huang, "A Handover Prediction Mechanism Based on LTE-A UE History Information," *2015 18th International Conference on Network-Based Information Systems*, Taipei, 2015, pp. 167-172. doi: 10.1109/NBiS.2015.29

131. Y. Luo, P. N. Tran, C. An, J. Eymann, L. Kreft and A. Timm-Giel, "A Novel Handover Prediction Scheme in Content Centric Networking Using Nonlinear Autoregressive Exogenous Model," *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, Dresden, 2013, pp. 1-5. doi: 10.1109/VTCSpring.2013.6691837

132. Kobayashi H, Kameda E., Terashima Y., and Shinomiya N. Towards sustainable heterogeneous wireless networks: A decision strategy for AP selection with dynamic graphs. Computer Networks. 2018; 132:99–107.

133. A. Nadembega, A. Hafid and T. Taleb, "A Destination and Mobility Path Prediction Scheme for Mobile Networks," in *IEEE Transactions on Vehicular Technology*, vol. 64, no. 6, pp. 2577-2590, June 2015. doi: 10.1109/TVT.2014.2345263

134. M. Grewal, L. Weill, A. Andrews. “Global positioning systems, inertial navigation, and integration”, John Wiley & Sons; 2007

135. B. Hui, J. Kim, H. Chung and I. Kim, "Creation and control of handover zone using antenna radiation pattern for high-speed train communications in unidirectional networks," *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, 2016, pp. 737-740.