Load Balancing using Cell Range Expansion in LTE Advanced Heterogeneous Networks

Wiseman Nkosingiphile Nyembe

This dissertation is submitted in partial fulfillment of the academic requirements

for the degree of

Master of Science in Electrical Engineering

in the Faculty of Engineering and The Built Environment

University of Cape Town

2015

Declaration

I, Karen Schwabe, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

Signed by candidate

(Signature)

22/11/2006

(Date)

Declaration

I declare that this dissertation is my own work. Where collaboration with other people has taken place, or material generated by other researchers is included, the parties and/or materials are indicated in the acknowledgements or are explicitly stated with references as appropriate.

This work is being submitted for the Master of Science in Electrical Engineering at the University of Cape Town. It has not been submitted to any other university for any other degree or examination.

Wiseman Nkosingiphile Nyembe

Signed by candidate

27/01/2016

Name

Date

EBE Faculty: Assessment of Ethics in Research Projects

Any person planning to undertake research in the Faculty of Engineering and the Built Environment at the University of Cape Town is required to complete this form before collecting or analysing data. When completed it should be submitted to the supervisor (where applicable) and from there to the Head of Department. If any of the questions below have been answered YES, and the applicant is NOT a fourth year student, the Head should forward this form for approval by the Faculty EIR committee: submit to Ms Zakiya Chikte (Zakiya.chikte@uct.ac.za); New EBE Building, Ph 021 650 5739).

Please note – It is important to keep a signed copy of this form as students must include a copy of the completed form with the dissertation/thesis when it is submitted for examination.

Name of Principal Researcher/Student:
Wiseman Nkosingiphile Nyembe

Department:
Electrical Engineering

If a Student: **Degree:**
Master of Science

Supervisor:
A/Prof Mqhele Dlodlo

If a Research Contract indicate source of funding/sponsorship:

Research Project Title:
Load Balancing using Cell Range Expansion in LTE Advanced Heterogeneous Networks

Overview of ethics issues in your research project:

Question 1: Is there a possibility that your research could cause harm to a third party (i.e. a person not involved in your project)?		NO
Question 2: Is your research making use of human subjects as sources of data? If your answer is YES, please complete Addendum 2.		NO
Question 3: Does your research involve the participation of or provision of services to communities? If your answer is YES, please complete Addendum 3.		NO
Question 4: If your research is sponsored, is there any potential for conflicts of interest? If your answer is YES, please complete Addendum 4.		NO

If you have answered YES to any of the above questions, please append a copy of your research proposal, as well as any interview schedules or questionnaires (Addendum 1) and please complete further addenda as appropriate.

I hereby undertake to carry out my research in such a way that

- there is no apparent legal objection to the nature or the method of research; and
- the research will not compromise staff or students or the other responsibilities of the University;
- the stated objective will be achieved, and the findings will have a high degree of validity;
- limitations and alternative interpretations will be considered;
- the findings could be subject to peer review and publicly available; and
- I will comply with the conventions of copyright and avoid any practice that would constitute plagiarism.

Signed by:

	Full name and signature	Date
Principal Researcher/Student:	**Wiseman Nyembe**	11/02/2016

This application is approved by:

Supervisor (if applicable): **Mqhele Dlodlo**		12/02/20116
HOD (or delegated nominee): Final authority for all assessments with NO to all questions and for all undergraduate research.		17/2/16
Chair : Faculty EIR Committee For applicants other than undergraduate students who have answered YES to any of the above questions.		

ADDENDUM 1:
Please append a copy of the research proposal here, as well as any interview schedules or questionnaires:

ADDENDUM 2: To be completed if you answered YES to Question 2:

It is assumed that you have read the UCT Code for Research involving Human Subjects (available at http://web.uct.ac.za/depts/educate/download/uctcodeforresearchinvolvinghumansubjects.pdf) in order to be able to answer the questions in this addendum.

2.1 Does the research discriminate against participation by individuals, or differentiate between participants, on the grounds of gender, race or ethnic group, age range, religion, income, handicap, illness or any similar classification?	YES	NO
2.2 Does the research require the participation of socially or physically vulnerable people (children, aged, disabled, etc) or legally restricted groups?	YES	NO
2.3 Will you not be able to secure the informed consent of all participants in the research? (In the case of children, will you not be able to obtain the consent of their guardians or parents?)	YES	NO
2.4 Will any confidential data be collected or will identifiable records of individuals be kept?	YES	NO
2.5 In reporting on this research is there any possibility that you will not be able to keep the identities of the individuals involved anonymous?	YES	NO
2.6 Are there any foreseeable risks of physical, psychological or social harm to participants that might occur in the course of the research?	YES	NO
2.7 Does the research include making payments or giving gifts to any participants?	YES	NO

If you have answered YES to any of these questions, please describe how you plan to address these issues (append to form):

ADDENDUM 3: To be completed if you answered YES to Question 3:

3.1 Is the community expected to make decisions for, during or based on the research?	YES	NO
3.2 At the end of the research will any economic or social process be terminated or left unsupported, or equipment or facilities used in the research be recovered from the participants or community?	YES	NO
3.3 Will any service be provided at a level below the generally accepted standards?	YES	NO

If you have answered YES to any of these questions, please describe how you plan to address these issues (append to form)

ADDENDUM 4: To be completed if you answered YES to Question 4

4.1 Is there any existing or potential conflict of interest between a research sponsor, academic supervisor, other researchers or participants?	YES	NO
4.2 Will information that reveals the identity of participants be supplied to a research sponsor, other than with the permission of the individuals?	YES	NO
4.3 Does the proposed research potentially conflict with the research of any other individual or group within the University?	YES	NO

If you have answered YES to any of these questions, please describe how you plan to address these issues(append to form)

Dedication

To my lovely wife, Dudu and the boys, Dumo and Seke.

Abstract

The use of heterogeneous networks is on the increase, fueled by consumer demand for more data. The main objective of heterogeneous networks is to increase capacity. They offer solutions for efficient use of spectrum, load balancing and improvement of cell edge coverage amongst others. However, these solutions have inherent challenges such as inter-cell interference and poor mobility management. In heterogeneous networks there is transmit power disparity between macro cell and pico cell tiers, which causes load imbalance between the tiers. Due to the conventional user-cell association strategy, whereby users associate to a base station with the strongest received signal strength, few users associate to small cells compared to macro cells. To counter the effects of transmit power disparity, cell range expansion is used instead of the conventional strategy.

The focus of our work is on load balancing using cell range expansion (CRE) and network utility optimization techniques to ensure fair sharing of load in a macro and pico cell LTE Advanced heterogeneous network. The aim is to investigate how to use an adaptive cell range expansion bias to optimize Pico cell coverage for load balancing. Reviewed literature points out several approaches to solve the load balancing problem in heterogeneous networks, which include, cell range expansion and utility function optimization. Then, we use cell range expansion, and logarithmic utility functions to design a load balancing algorithm. In the algorithm, user and base station associations are optimized by adapting CRE bias to pico base station load status. A price update mechanism based on a suboptimal solution of a network utility optimization problem is used to adapt the CRE bias. The price is derived from the load status of each pico base station.

The performance of the algorithm was evaluated by means of an LTE MATLAB toolbox. Simulations were conducted according to 3GPP and ITU guidelines for modelling heterogeneous networks and propagation environment respectively. Compared to a static CRE configuration, the algorithm achieved more fairness in load distribution. Further, it achieved a better trade-off between cell edge and cell centre user throughputs.

Acknowledgements

I would like to sincerely thank my supervisor, Associate Professor Mqhele Dlodlo for his unwavering support and advices during the course of the study. It is through your constructive criticism and insights that I have been successful in this endeavor.

My gratitude also goes to my colleagues in the MED Lab for creating a lively environment conducive for exchange of ideas. I have truly learnt a lot through you guys, keep up the good spirit. Specifically, I would like to thank Fred Kumi and Henry Ohize for reviewing my dissertation.

Finally, a special thanks goes to my family, Dudu, Seke and Dumo for their support and encouragement when it was needed the most. Your love has truly carried me throughout this period.

Table of Contents

List of Figures

List of Abbreviations

3GPP	Third Generation Partnership Project
BS	Base Station
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CCH	Control Channel
CRE	Cell Range Expansion
CoMP	Coordinated Multi-Point Transmission Reception
eICIC	Enhanced Inter-Cell Interference Coordination
EPC	Evolved Packet Core
EPS	Evolved Packet System
eNodeB	Evolved Node Base station
ETSI	European Telecommunication Standards Institute
E-UTRAN	Evolved Universal Terrestrial Access Network
GSM	Global System for Mobile Communications
HSDPA	High Speed Downlink Packet Access
HSUPA	High Speed Uplink Packe Access
IMT	International Mobile Telecommunications
ITU	International Telecommunications Union
LTE	Long Term Evolution
MAC	Medium Access Control
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
NAS	Non Access Stratum

OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiplexing Access
PAPR	Peak to Average Power Ratio
PBCH	Physical Broadcast Channel
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDN	Public Data Network
PDN-GW	Public Data Network Gateway
PDSCH	Physical Downlink Shared Channel
PRB	Physical Resource Block
PUCCH	Physical Uplink Control Channel
QoS	Quality of Service
RLC	Radio Link Control
RSS	Received Signal Strength
RSRP	Reference Signal Received Power
SAE	System Architecture Evolution
SC-FDMA	Single Carrier Frequency Division Multiple Access
SINR	Signal to Interference and Noise Ration
UE	User Equipment
UMTS	Universal Terrestrial Radio Access Network
VNI	Visual Networking Index
WCDMA	Wide Band Code Division Multiple Access
WiMAX	Worldwide Interoperability for Microwave Access

Chapter 1
Introduction

1.1 Background and Motivation

The demand for data in wireless networks is in a steep incline, and will continue in the same trajectory at least for a while. The primary drivers of this demand are the proliferation of bandwidth-hungry mobile applications and the insatiable desire by consumers to be "always connected". This projection is backed by data presented by the Cisco visual network index (VNI) white paper in 2015, which states that "global mobile data traffic grew nearly 70 percent in 2014 and it is expected to grow by about 20 percent by 2019" [1]. In view of these projections, there are challenges and opportunities for the research community and operators to design and optimise network solutions.

The increase in data traffic thereby calls for innovative system designs geared towards rapid wireless network traffic growth. As a consequence of this demand, the International Telecommunications Union – Radio sector (ITU-R), the radio organ of the ITU, laid down a framework for building cohesive solutions for 4^{th} generation technologies in the form of the IMT-Advanced requirements specifications[2]. In response to these requirements, 3GPP came up with the Long Term Evolution (LTE) Advanced standard, which is built on old LTE standards[3]. To meet the requirements of the ITU-R the LTE standard went through major technology upgrades in the radio interface. These upgrades enhanced the performance of LTE even beyond the requirements set by the ITU-R. As one of the enhancements, the third generation partnership project (3GPP) included heterogeneous networks (HetNets) in the radio access network of LTE-Advanced.

The design of wireless networks is centred on a concept of cells, where the cell comprises of users who are associated to a base-station. This design is predominantly macro-cell based; small base stations are not an integral part of the initial network planning. However, in HetNets, small cells are key to enhancing network capacity hence the need for relevant network planning strategies. As highlighted, in LTE Advanced, 3GPP introduced HetNets to improve network capacity. In HetNets, macro and small base stations are deployed in the same location, where typically macro cells provide an umbrella coverage for a large area, overlaid with smaller cells, to

enhance network capacity or extend network coverage. Small base stations relieve overloaded macro base stations by providing resources to offloaded macro base station users. They are also a good option for handling spatially diverse data traffic and varying user densities [4]. For a rapidly growing traffic demand, HetNets present a compelling business case to operators for fast and scalable deployments. Figure 1.1 illustrates a typical HetNet deployment. It consists of a macro-cell overlaid with pico-cells, femto-cells and relay nodes [5].

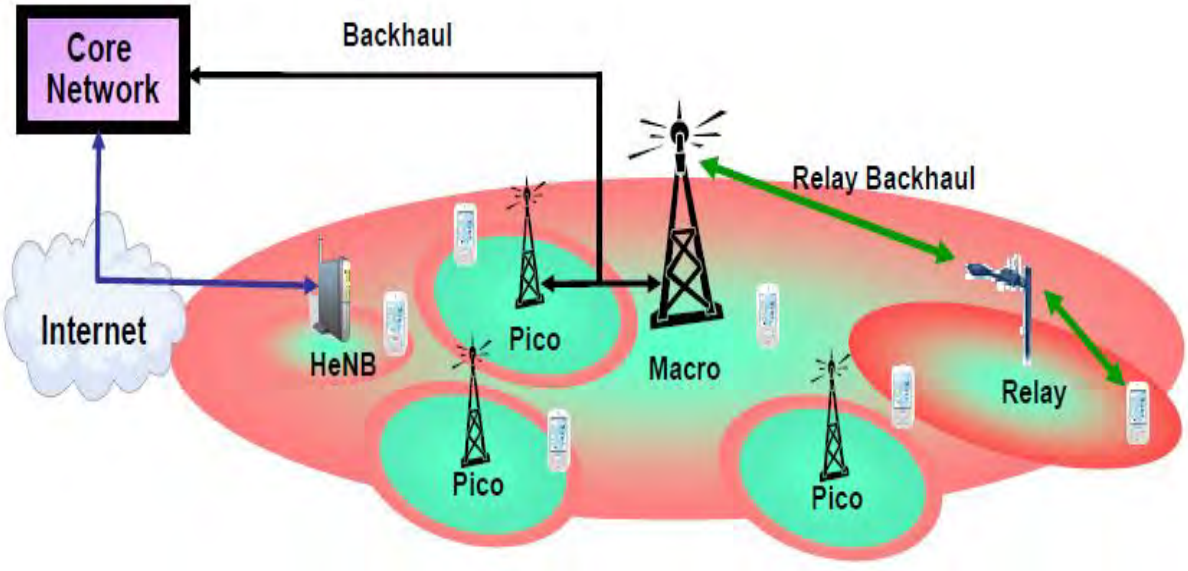

Figure 1.1. Illustration of a Heterogeneous Network [5]

HetNet solutions are also beneficial for their cost effectiveness, flexibility, and providing capacity in high-density areas. The capacity is achieved through the splitting of macro cells into smaller cells with spectrum reuse. This approach yields more capacity, coverage and average throughput gains for a limited amount of spectrum resources [5].

Small cell deployments, however, present their own challenges in network management and planning. This is mainly due to the nature of deployment, which is somehow "ad hoc". A less coordinated small cell deployment is certainly likely to result in a poor performance and inferior quality of service (QoS). This is because, unlike macro-cell based networks, HetNets are a combination of macro and small base stations, where the macro transmit power is higher than the

small base station transmit power. The transmit power disparity between the network tiers does not allow the macro base stations to sufficiently offload users when overloaded. This is because users will normally associate to a base station with a higher received signal strength. Hence, many users will associate with macro base stations even when small base stations have sufficient capacity to accommodate more users. This user-cell association approach leads to load imbalances, therefore is not suitable for HetNets.

To this end, several HetNet suitable user-cell association strategies have been proposed for enabling a balanced utilisation of network resources [4], [6]–[9]. However, some of the proposed strategies are complex to implement, due to the computational rigour and signalling required for optimising resource utilisation in a dynamic network environment. Cell Range Expansion (CRE), utilised in our work, is one simpler strategy for offloading users from macro-cells to small cells. CRE is a practical and effective method that has its roots in the power control techniques of legacy mobile technologies such as GSM and WCDMA. Traditionally, in mobile networks, including LTE, a User Equipment (UE) connects to a node that provides the strongest downlink received signal strength. Cell range expansion increases the coverage of small cells by adding a bias value to the measured received signal strength during cell association [10]. This user-cell association strategy relies on the received signal strength and CRE bias. The technique allows for offloading of more users to small base stations. However if the bias is large, small cell users in the cell CRE region may experience severe interference from the macro base stations, especially in a co-channel deployment [11].

In general, load balancing has long been a study item in telecommunications networks, as a mechanism to optimise resource utilisation and network congestion control. In HetNets, however, balancing load amongst network tiers is challenging due to varying transmit powers, unpredictable user distributions, and ad hoc deployment of small cells.

1.2 Related Work

The focus of this work is to investigate optimisation of cell range expansion for load balancing in HetNets by tuning the cell range bias in accordance with the load on each small cell. Cell range expansion, as a cell association strategy for macro-cell traffic offloading, is covered in several studies as discussed in [10]–[12]. However, its potential for performing load balancing in HetNets needs more consideration. Generally, the reviewed literature does not give sufficient

theoretical guidance on how an optimal CRE bias is determined. The strategy for achieving a load-balanced network is by some manual trial and error method. It is important, therefore, to have a dynamic CRE bias mechanism that is responsive to cell load of each base stations. In most of works, the assignment of CRE bias is uniform for an entire network tier [10], [12]–[14]. Hence the need for more non-uniform and adaptive range bias strategies that factor in the spatial and traffic diversity of each base station.

Considerable amount of work on load balancing is on macro-cell only networks [15]–[17]. In Universal Mobile Telecommunications System (UMTS) for instance, the adjustment of transmit power levels were studied in [15]. Some research works approach the load balancing problem by using utility functions such as max-min fairness [18], and logarithmic functions [19]. Cell Range Expansion, as a load balancing strategy, is explored in [4], [20]–[23]. These studies use uniform network wide bias values, except for [4] who proposed the use of cell specific bias values, computed from a centralised entity. A comprehensive review of the cited and other relevant works is provided in chapter three.

The load balancing problem has been explored in the past in our research group, for generic next generation network systems (NGNS) [24]. The focus was on enabling the call admission control (CAC) function of the core network to perform load balancing on heterogeneous networks. This work builds on that knowledge base for an LTE Advanced load balancing solution. However, the mathematical formulation follows the approach in [19] and adopts the use of logarithmic utility functions to optimise cell range bias of pico cells in a distributed manner. The load-balancing problem is formulated as an optimisation problem, taking into account the distributed nature of an LTE system. We then design a load-balancing algorithm based on a sub-optimal solution to the load-balancing problem and CRE. The objective is to exploit the theoretical aspects of optimisation techniques, whilst avoiding their computational rigour and signalling overheads. To determine the fairness of the algorithm, we use the Jain fairness index [25].

1.3 Problem Statement

The issue of load balancing in heterogeneous networks is considered as one of the challenges in LTE Advanced [26]. HetNets provide increased capacity for the network. However, if the network's load balancing parameters are not appropriately tuned that capacity may not be

fully realised. A well-balanced network improves the utilisation of small base stations, enabling more users to exploit the capacity of the network and thereby resulting in higher data rates.

The core of the load-balancing problem is in network planning and optimisation. Presently, network planning and optimisation is static yet the network is dynamic in terms of load patterns driven by user behaviour and patterns [26]. Users tend to form groupings in places of activity; as an example, users tend to converge in various spots for a given period depending on their activities. In a typical working day, offices and other business areas generate a lot of network traffic, which is later on observed in residential areas in the evening. Similarly, on weekends the traffic surge is observed in areas of recreation. These user movements clearly results in fluctuations in network load and renders fixed network planning ineffective.

This therefore calls for a provision of agile and responsive planning and optimisation mechanisms, which can dynamically allocate network resources according to the load. Cell range expansion is a simple yet practical solution proposed for LTE Advanced HetNets to perform load balancing. CRE is a technique that adds a virtual bias to the received power from small base stations for handover to expand small cell coverage area. This enables more users to access the small base station even when there is a power disparity between the macro cell tier and small cell tier. CRE however, increases interference on range expansion users particularly when macro-cells and small cells share the same channel. This can be mitigated by enhanced inter cell interference coordination (eICIC). This is a frequency-time domain spectrum scheduling technique, which alternately schedules macro cell and small cell users on provided sub-frames to ensure that there is no interference between the macro-cell and small cell network tiers.

Even though CRE provides offloading gains, the bias value is fixed and there is no tacit guidance in literature on how it is determined. The proposed topic seeks to investigate how an adaptive CRE bias can be used in cell range expansion to adjust small cell coverage for load balancing in a macro and pico cell LTE Advanced HetNet for network optimisation. Since the wireless network environment is dynamic, base stations are subjected to varying loads and interferences, we believe the CRE bias should be non-uniform and adaptive to reflect the prevailing traffic load experienced by each base station. There is a need therefore, for an optimisation framework, which ensures a balance between throughput and fairness towards solving this problem.

1.4 Hypothesis

The use of non-uniform and adaptive cell-range expansion biases in small base stations for load balancing in a dynamic and disparate traffic density heterogeneous network will enhance user throughput, fairness and macro cell user offload ratios.

1.5 Research questions

- How can an adaptive cell range expansion bias that is optimised for load balancing in a two tier heterogeneous network, be determined?

- How can a cell range expansion based load balancing algorithm that utilises optimised adaptive cell range expansion biases be designed?

- How can the performance of the algorithm be evaluated to determine its effectiveness in terms of fairness, throughput and user offloading?

1.6 Research Objectives

- To investigate the use of individual cell range expansion bias in an optimisation strategy for network load balancing in a two tier heterogeneous network.

- Design a cell range expansion region optimising algorithm that adjusts the bias according to individual cell loads.

- Evaluate the performance of the algorithm in terms of fairness, throughput and offload ratio.

1.7 Scope of Research

This work is focused on designing a load balancing algorithm that is compliant with LTE Advanced standard and related guidelines. Therefore, it may not be suitable for other wireless standards in its current form. According to 3GPP standards, as stated in [27], cell range expansion is applied to users in connected mode only. Hence, we limit our algorithm to perform load balancing only on active users. For the sake of simplicity, we assumed that all users in the network

are in connected mode. We recognize the negative impact of using very large cell range bias values, especially on cell edge users which exposes them to severe downlink macro eNodeB interference. To counter the effects of interference in co-channel HetNet deployments, eICIC techniques are usually combined with cell range expansion. However, we limited our work to using cell range expansion without eICIC techniques. The rationale behind the exclusion of eICIC techniques is to focus our efforts on designing a cell specific range expansion based load balancing algorithm. To evaluate the performance of the load balancing algorithm we use MATLAB version 8.3.0.532 (R2014a) and HetNet tool box, which contains channel modelling and macro and small cell network layout functions [28]. The toolbox is built on 3GPP recommended guidelines for simulating heterogeneous networks. As a baseline configuration for performance comparison we set scenarios where the load balancing algorithm uses static cell range expansion biasing i.e. CRE bias = {0dB, 6dB, and 12dB}.

1.8 Thesis outline

The thesis consists of six chapters. Chapter 1 provides a background and motivation to the study. It describes the problem to be solved and states key research questions and objectives. The rest of the thesis is outlined as follows:

- Chapter 2 presents a technical background on LTE-Advanced, touching on the general architecture of LTE radio interface and core network. It further highlights the evolution of 3GPP standards through to LTE-Advanced.
- Chapter 3 presents a review of literature on load balancing in wireless networks. It consists of earlier work and recent work on the subject which includes cell range expansion and utility function based optimization techniques.
- Chapter 4 proposes a cell range expansion based load balancing algorithm. It further includes a system model, corresponding problem formulation, and performance metrics.
- Chapter 5 presents implementation details and performance evaluation of the load balancing algorithm.
- Chapter 6 presents concluding remarks and pointers to possible future research work.

Chapter 2
Technical background of LTE-Advanced

2.1 Introduction

This chapter presents the technical background of LTE-Advanced and a brief discourse on the evolution of 3GPP standards. The aim is to provide the reader with understanding of key features and network architecture concepts of the standard. First, it discusses the evolutionary path of 3GPP standards. Then it describes the network architecture, which consists of the evolved packet core and radio access network. Furthermore, it includes some aspects of the radio access network such as radio resource management and physical layer technologies. Finally, it presents a discussion on heterogeneous networks.

2.2 Evolution of 3GPP Network Standards

The surge of data traffic in mobile networks has been widely discussed and acknowledged [1]. This traffic growth is mainly driven by user demand and enabled by technological innovations in mobile networks [29]. These innovations are notably evolutionary by nature, gradually tracking the traffic demand trajectory. The emergence of LTE as a prominent mobile network standard is a result of incremental changes from older technologies. These changes in 3GPP technology standards, often referred to as an evolution, are evident when comparing GSM voice only networks and LTE Advanced. In the past, mobile networks were designed for voice communication only, and end-to-end communication was established by circuit switching. When the need for data communication initially arose, the necessary capabilities were merely a patchwork to voice systems. It took some time for data communication capabilities to actually mature. Post voice only communication, the evolutionary process, spans three radio access technologies, increasing richness in features and capacity with each release. Figure 2.1 illustrates an evolutionary path of 3GPP standards over the years [30]. Since its inception in 1998, the 3GPP started work on 3G technology standard, which was the UMTS release 99. The standard was later improved with an addition of enhanced dedicated downlink and uplink channels, called high speed downlink packet access (HSDPA) and high speed uplink packet access (HSUPA). The combination of these two developments led to what is now called high speed packet access (HSPA) [31]. These standards

use wideband code division multiple access (WCDMA) as their access technology. The need for mobile broadband then prompted 3GPP to work on the LTE standard to pave way for a 4G standard. Orthogonal frequency division multiple access (OFDMA) was introduced as the choice access technology for LTE.

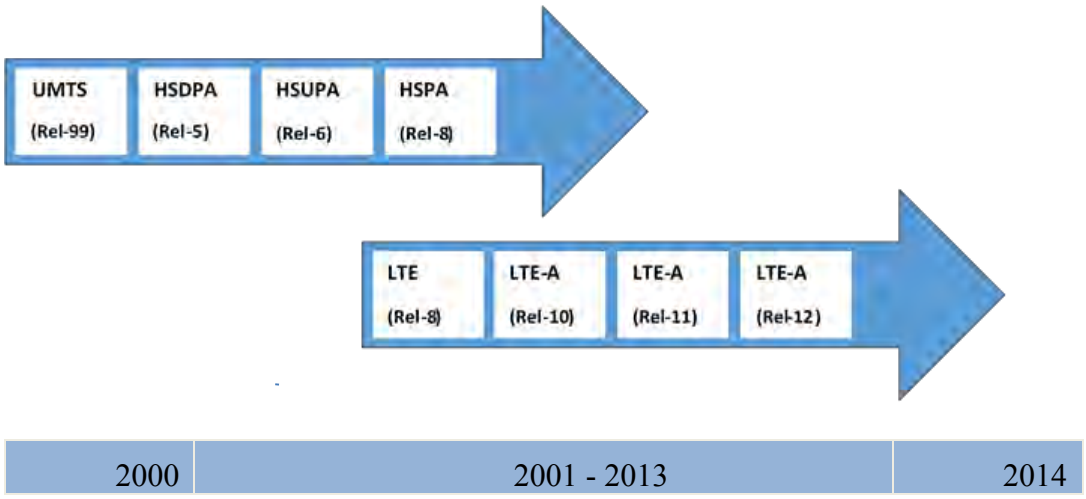

Figure 2.1. Evolution of 3GPP standards

LTE debuted with release 8 and has evolved through a series of incremental enhancements to what is now release 12. The establishment of LTE Advanced was motivated by the ITU-R IMT-Advanced requirements for 4G technology standards [3]. LTE Advanced was qualified as one of the technologies to meet the requirements of the ITU-R.

Even though it would seem that LTE-A is emerging as the de-facto mobile network standard, the reality is that it will have to co-exist with other legacy technologies for a while. A 2014 report on mobility by Ericson [29] suggests that mobile networks, at least in the near term, will essentially be an ecosystem of mobile technologies serving their niche areas. LTE will be gradually eased in, as the need for mobile broadband increases. This is observed in figure 2.2, which shows user subscription trends per mobile network technology. Being cognisant of this reality, when designing the LTE standard, 3GPP ensured that the standard was open and interoperable as much as possible. The introduction of an open architecture core network, known as the evolved packet core, is an indicator to that effect.

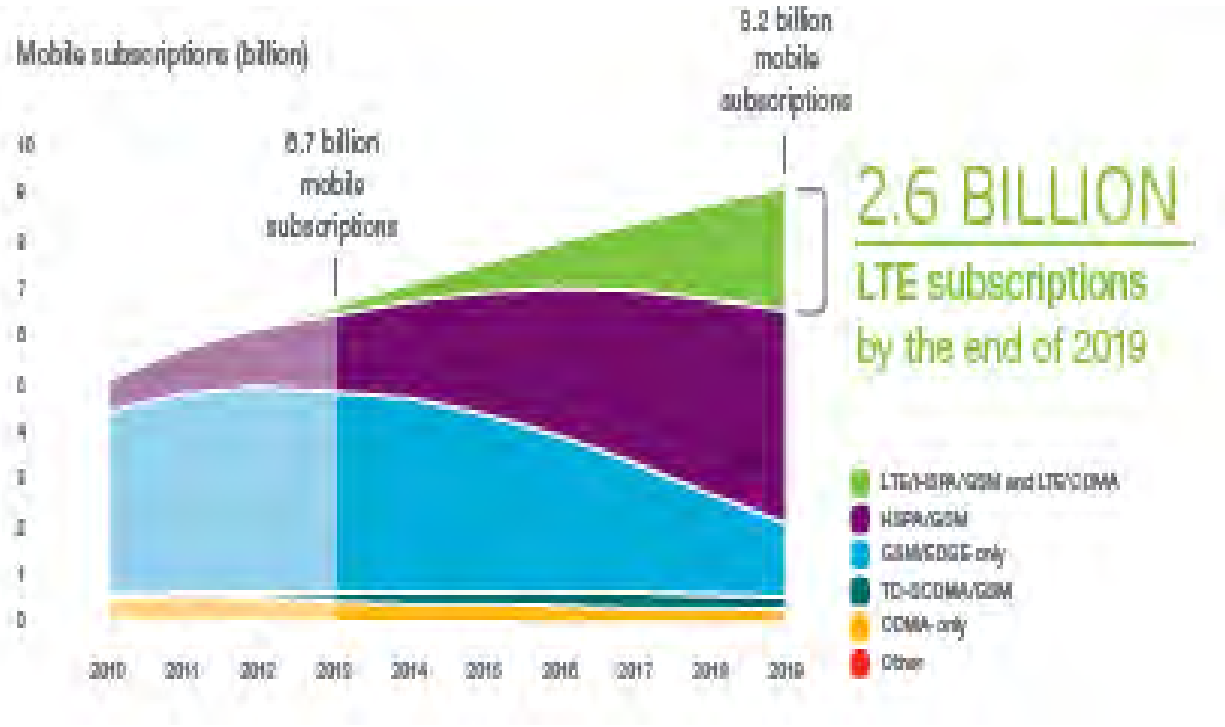

Figure 2.2. Mobile subscriptions trends per mobile network technology [29]**.**

2.3 Key Features of LTE-Advanced

The objective for further development of LTE to LTE-Advanced was to improve data rates, reduce cost per bit, and improve user experience, whilst meeting the requirements of the ITU for IMT-Advanced. The development was further driven by the need to flexibly exploit new and existing spectrum [32]. For the standard to meet the ITU requirements, it went through major enhancements in the radio access network. Enabling technologies were included, which provided the necessary functionalities to meet the requirements. The 3GPP also developed their own requirements specifications, and they exceeded the requirements set by the ITU-R. Some of the features of LTE Advanced include downlink peak data rates of 3 Gbits/sec for low mobility UEs and uplink peak data rates of 1 Gbit/sec, high spectral efficiency, interoperability with a wide variety of services and applications, and improved cell edge performance. Listed below are some of the key functionalities and capabilities of LTE-Advanced as highlighted in [32] .

- **Carrier aggregation:** To increase data rates, wider bandwidth is necessary; release 8 of LTE specifies the use of up to 20MHz, which is not sufficient to meet the data rate requirements of the IMT-Advanced. Therefore, the spectrum in LTE Advanced was increased to supports up to 100MHz bandwidth. However, it is currently not feasible to have a contiguous spectrum this wide. Hence, the idea of carrier aggregation. Carrier aggregation is the logical joining of spectrum from adjacent bands (contiguous) or different bands (non-contiguous) so that they may be used by a UE. If the UE supports multiple transceivers it can utilise non-contiguous spectrum as if it was one wide band. Otherwise, it can only utilise contiguous spectrum only.

- **Enhanced multiple antenna techniques:** LTE Advanced supports up to 8 antennas in the downlink, and 4 antennas in the uplink. This is a huge improvement in comparison to release 8, which supports 4 antennas in the downlink and 1 antenna in the uplink.

- **Relay nodes:** To improve or add coverage in poorly covered areas or dead zones, relay nodes are introduced in LTE Advanced as a backhaul mechanism. These relay nodes have both traffic forwarding and routing capabilities, as opposed to legacy radio frequency repeaters. They may be used to improve throughput in high dense or indoor environments and extend coverage to sparsely populated environments.

- **Coordinated multipoint (CoMP) operation:** Coordinated transmission of data from different eNodeBs to UEs, known as CoMP, is effective in managing interference in heterogeneous networks [5].

- **Support for heterogeneous networks:** The heterogeneous network deployment concept uses cells of different sizes and access technologies, working together in co-ordinated manner to extend network coverage. In LTE Advanced, they are a part of the network planning strategy.

An over-arching theme on all the features is that of increasing spectrum efficiency per unit area in a flexible and scalable manner, from the manipulation of the spectrum to physical infrastructure.

2.4 Network architecture

The LTE network is composed of two major entities, which are the core network known as the Evolved Packet Core (EPC) and the radio access network known as the Evolved Universal Terrestrial Radio Access Network (E-UTRAN). UEs connect to the EPC via the E-UTRAN [31]. The combined system from the two entities form the Evolved Packet System (EPS). The architectural diagram of the EPS as shown in figure 2.3, is redrawn from [31].

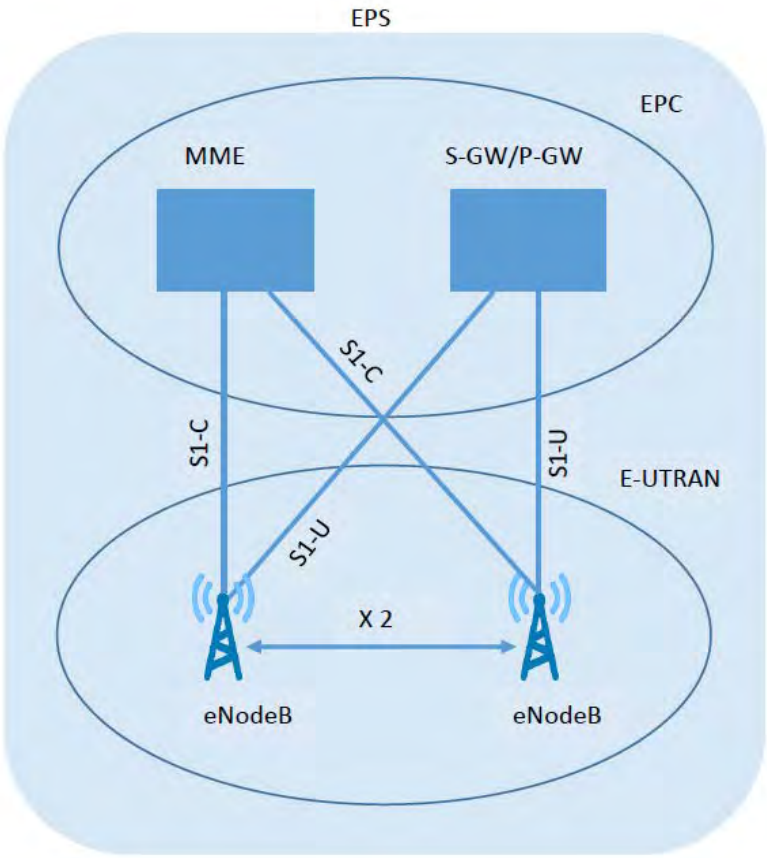

Figure 2.3. LTE network architecture.

Generally, the function of the EPC is to facilitate end-to-end connection of UEs by providing functions such as UE authentication, session establishment, and accounting. The EPC is composed of several logical nodes, which are the mobility management entity (MME), packet data Network gateway (P-GW), and serving gateway(S-GW). On the other hand, the function of the E-UTRAN is to provide physical access to the network by means of radio links. The E-UTRAN is

composed of one node type, known as the enhanced NodeB (eNodeB). The eNodeB combines functions, which were previously provided by the base transceiver station (BTS) and radio network controller (RNC) in UMTS, thereby flattening the structure of the radio access network. The LTE architecture is an all IP implementation. The combination of an all-IP packet switching core and flat E-UTRAN increases the performance of LTE by far, as compared to its hierarchically structured predecessors, such as GSM and UMTS. The architecture lends itself to easy scalability and interworking with other communication technologies. Further discussion on the EPC and the E-UTRAN follows in subsequent sections. The following sections present the description of the EPC and the E-UTRAN functionalities provided by the prescribed protocols in each layer. A comprehensive discussion on LTE architecture is covered in detail in [27].

2.5 The Evolved Packet Core (EPC)

The focus on the design of the EPC was in providing a flat architecture that could efficiently handle data traffic whilst ensuring that data traverses only a few network nodes, and conversion between protocol layers is kept to a minimum. As referred to earlier, the function of the EPC is to establish end-to-end connections between entities that need to communicate; it is only after the establishment of an end-to-end logical connection that a data session can begin. The logical connection is called an EPS bearer. The EPS bearer is between the UE and P-GW going through the S-GW, figure 2.4 illustrates [31].

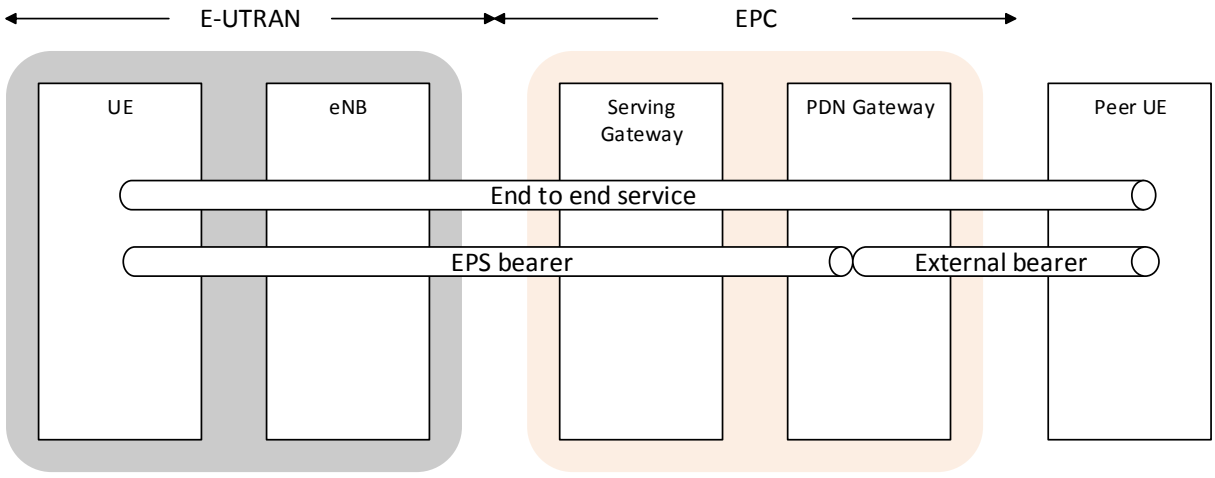

Figure 2.4. Illustration of EPS Bearer

The EPS bearer may be viewed as an end-to-end packet flow with specified quality of service (QoS) guarantees. It is the responsibility of the EPC to provide different bearers, as requested for various IP-packet flows. The architecture of the EPC is such that the user plane and control planes are separated, an approach which allows for flexible scaling in the core network.

The control plane functionality lies with the MME. The MME performs non access stratum (NAS) signalling between the UE and EPC. NAS refers to those functions that do not deal with access, but rather the transfer of data. To understand the functions of the MME better, an illustration suffices. When a UE is switched on and associates itself to a network, it is the responsibility of the MME to establish and manage EPS bearers related to the UE. The MME also has to ensure seamless UE mobility between LTE and other non-LTE networks.

The user-plane functions are performed by the P-GW and S-GW. The P-GW is placed at the edge of the network as an interface with external IP packet based networks, including the internet. As a gateway, it enables UEs to connect to external entities. It achieves the connection by allocating IP addresses to UEs so that data packets can be routed accordingly. It also controls data flow and enforces QoS requirements to and from the UE. The S-GW performs routing and forwarding of packets coming and leaving the UE. In addition, it is responsible for mobility of UEs between LTE networks, acting as a mobility anchor. When a UE moves from one network to another, packets destined for the UE are firstly buffered in the S-GW whilst the MME attempts to locate the UE.

It is worth noting that the implementation of the EPC nodes, that is the MME, S-GW and P-GW could be logical or they can be standalone physical entities, and this to an extent depends on the operator's requirements. The MME, S-GW, and P-GW may be implemented as separate entities. They may also be put together as a unit.

2.6 LTE Radio Access Network

The E-UTRAN is composed of a single node, which is the eNodeB. The eNodeB performs functions, which are mainly required for proper operation of the radio access network. The wireless channel is noisy and unreliable compared to fixed line channel conditions. As a result, transmitted signals are subject to sporadic amplitude and phase attenuation. Further, transmission in the wireless channel is generally broadcast oriented by nature, which means when signals are

transmitted in the channel from co-located nodes (UE or eNodeB), they interfere with one another. Therefore, for a successful reception of a transmitted signal, it is the responsibility of the E-UTRAN to perform signal processing to recover that signal. Radio resource management (RRM) and medium access control mechanisms are necessary to mitigate errors introduced by noise and interference, and allocate channel resources in an efficient and fair manner. Sometimes errors in received signal cannot be corrected, thus requiring a retransmission. The radio link control layer handles the retransmission process [27]. To ensure protocol cohesion, just like the EPC, the E-UTRAN also has control plane and user plane functionalities. A brief discussion related to the properties and operation of radio interface protocols follows.

2.7 Radio interface protocols

The radio interface protocol stack redrawn from [31] is shown in figure 2.5. The protocols are terminated between the UE and eNodeB. The stack consists of the radio resource control (RRC), packet decompression and compression protocol (PDCP), radio link control (RLC), medium access control (MAC), and physical (PHY) layers. The RRC layer ensures that broadcast information related to Access Stratum (AS) control signalling is performed to allow UEs to access the network. The RRC layer also supports procedures necessary to perform tasks such as NAS control signalling, handover control, paging, radio bearer management and measurement, and configuration management [33].

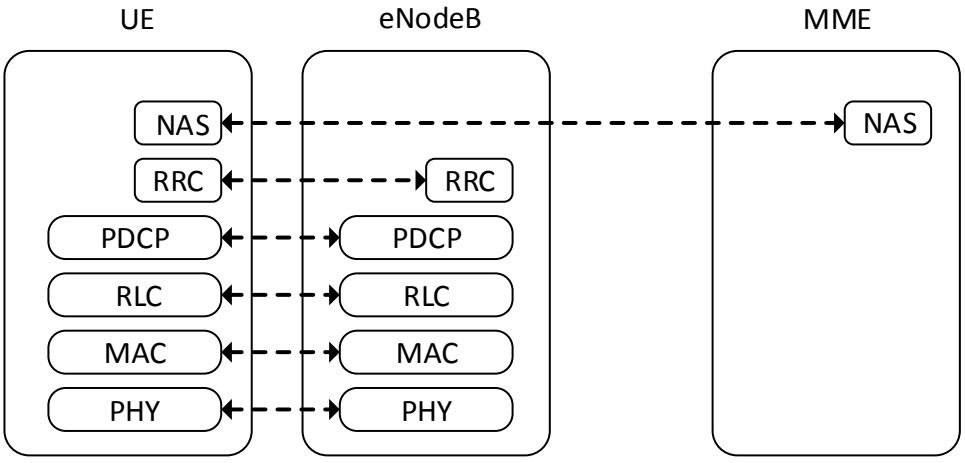

Figure 2.5. Radio interface protocol stack.

The PDCP layer provides security measures, which include ciphering to ensure integrity of data during transmission. It also performs IP header compression to reduce packet sizes. To assist receiving terminals, the appended PDCP header includes information for packet ordering and removal of duplicates. The RLC layer is responsible for segmentation and concatenation of PDCP packet data unit to ensure they are the appropriate size [31]. An RLC header included in this layer to support automatic repeat request (ARQ) procedure, if errors are detected in lower layers, are not correctible. The MAC layer is responsible for scheduling channel resources in time and frequency between UEs. In the downlink, it provides logical channels to RLC PDUs and multiplexes them to be transported by the physical layer. The multiplexed channels form transport block, whose size depends on the link adaptation mechanism. In the uplink, it schedules UEs to allocated channel resources. It also performs hybrid ARQ error correction and synchronisation of uplink transmission timing. The PHY layer is responsible for transmission and reception of signals. It prepares PDUs from the MAC layer and transmits them as signals. In addition, it carries out cyclic redundancy check (CRC) procedure and includes the check sum for correcting errors in the receiving terminal. It uses Adaptive Modulation and Coding techniques (AMC) to transmit according to the conditions of the channel. Finally, it also performs transmission power control [34].

2.8 Communication channels

In the first three layers of the protocol stack, services between the layers are provided via communication channels. The MAC layer uses logical channels to serve the RLC layer above and transport channels to serve the physical layer below. The type of information carried differentiates logical channels, which either carry control or traffic oriented information. LTE specified logical channels include the broadcast control channel (BCCH), paging control channel (PCCH), common control channel (CCCH), dedicated control channel (DCCH), multicast control channel (MCCH), dedicated traffic channel (DTCH), and multicast traffic channel (MTCH). Transport channels are differentiated by the characteristics of the information they transmit; for instance, there is a channel for paging UEs. When transporting information across the channel, information is packaged into transport blocks of varying sizes and formats. The transport format carries information on how the information is to be transported in relation to the channel quality and available resources. Hence, it affects the data rates. LTE specified transport channels include the downlink shared channel

(DL-SCH), broadcast channel (BCH), paging channel (PCH), multicast channel (MCH), and the random access channel (RACH). The DL-SCH is the main transport channel, as it transmits downlink data with the support of data adaptation and scheduling mechanisms. To form a conduit of communication, logical channels, transport channels, and physical channels to be discussed below, are mapped together, as shown in figure 2.6 [31].

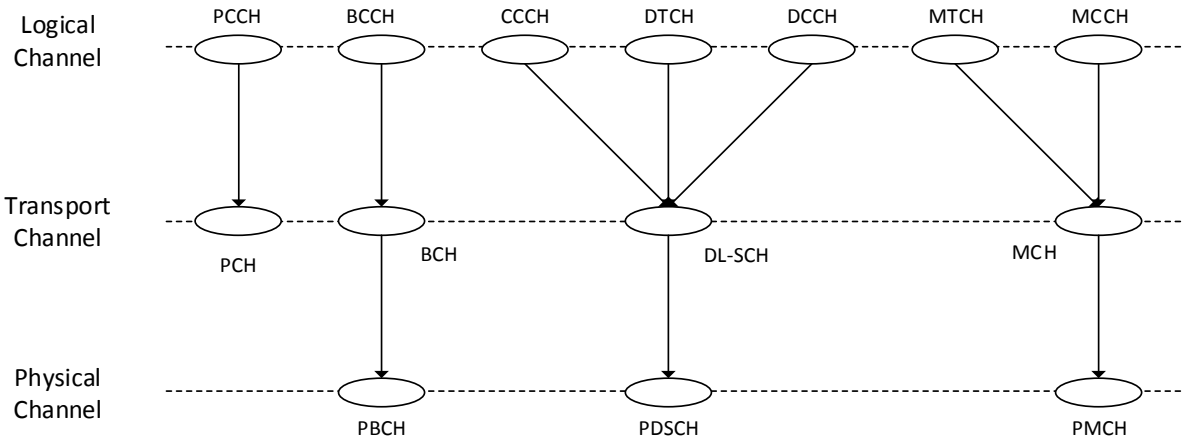

Figure 2.6. Illustration of channel mapping

In the physical layer, services to UEs are provided via physical channels. Each physical channel is related to a corresponding mapped transport channel, which is allocated enough resources [31]. LTE specified resources include physical downlink shared channel (PDCH), physical broadcast channel (PBCH), physical multicast channel (PMCH), physical downlink control channel (PDCCH), physical hybrid-ARQ indicator channel (PHICH), physical control format indicator channel (PHICH), physical uplink shared channel (PUSCH), and the physical uplink control channel (PUCCH).

2.9 LTE physical layer

The technologies used in the physical layer are the multiple access based OFDM and MIMO. In the downlink, the radio access network uses the conventional OFDMA, and in the uplink, it uses SC-FDMA. OFDMA was selected in LTE radio access network for its robustness against flat fading caused by frequency selective channels, flexibility, and capacity improvement.

A typical implementation of OFDMA, however, is not suitable in the uplink due to the high power consumption of OFDMA signal processing, yet UEs have limited battery power. The power consumption is a result of high peak to average power ratios (PAPR) between OFDMA subcarriers. For low PAPR and longer battery life, SC-FDMA is used in the uplink. MIMO antenna systems enhance the performance of LTE by taking advantage of the OFDMA parallel subcarrier property, which enables spatial diversity. For the sake of the scope of this work, details of MIMO systems, including the theoretical background are not covered here, but can be obtained from [31]. The following section presents an overview of OFDMA, resource grid and frame structures.

2.9.1 Orthogonal Frequency Division Multiple Access (OFDMA)

OFDMA combines modulation, multiplexing, and multiple user access. The OFDMA scheme is built on OFDM. Therefore at this point we start by presenting a brief overview of OFDM.

In the OFDM scheme, first, the available bandwidth is divided into multiple independent subcarriers, which are then modulated and transmitted as parallel streams. Compared to a single carrier transmission such as WCDMA, the scheme is robust to unpredictable mobile propagation conditions. The robustness is a result of parallel transmission of sub carriers at low symbol rates. The key feature of OFDM is orthogonality of the sub-carriers used during modulation. Orthogonality allows for parallel transmission of multiple sub-carriers in a tight frequency band without interference amongst the sub-carriers. The OFDM scheme is used in many telecommunication and broadcasting standards such as WiFi 802.11X, LTE, digital video broadcasting (DVB) [35].

The OFDMA scheme works by allocating users to sub carrier in the time and frequency domains, not just time domain only as is the case with OFDM. The flexibility of the scheme was one of the reasons it was adopted for LTE [31]. Introducing flexibility in the frequency domain for instance, allows for frequency dependent scheduling, as described in the ensuing sub-section.

2.9.2 Resource Grid Structure

The LTE scheduler, maps control and user data streams onto OFDMA signals and sub carriers. The information is organised in a time and frequency resource grid. An illustration of the resource grid is shown in figure 2.7 [36].

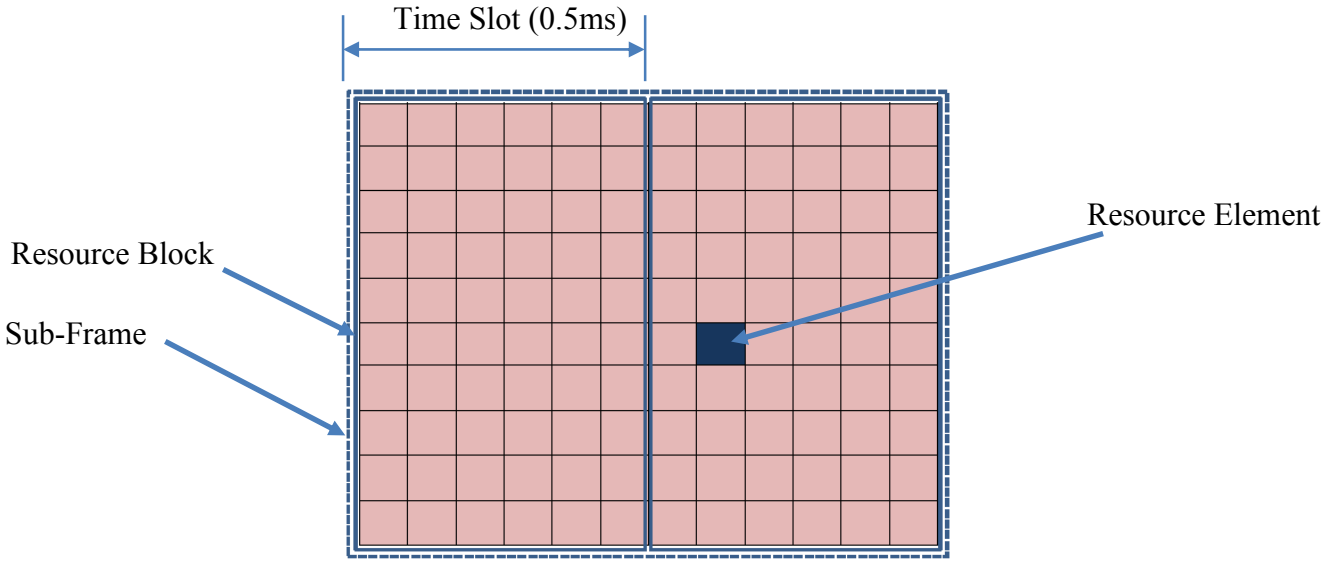

Figure 2.7. An illustration of a resource grid

The smallest physical unit in LTE is a resource element. It consists of one sub-carrier across a symbol. The resource elements are grouped into resource blocks, which are structured into 12 sub-carriers and 0.5 ms long time slot. Each sub-carrier spacing is 15 KHz, so the resource block is 180 KHz (12*15KHz). Each time slot consists of seven OFDM symbols, which makes a resource block to have 84 resource elements. For dynamic scheduling, sub-frames are used. Sub-frames are time domain units, which consist of two time slots. The scheduling of two consecutive resource blocks within a sub-frame is called a resource block pair. The number of resource blocks in a carrier depend on the transmission bandwidth. This bandwidth ranges from 1.4 MHz to 20 MHz and from a minimum of six resource blocks to a maximum of 100 resource blocks. For LTE release 10, the bandwidth can be increased to 100 MHz using carrier aggregation [31].

2.10 The X2 Interface

The X2 interface enables connectivity between eNodeBs for the exchange of information related to load balancing, interference co-ordination, and mobility management. The interface is a point-to-point logical connection. The physical implementation depends on the available physical medium. For direct physical connectivity, fiber, microwave, or Ethernet maybe used. However, in situations where a direct physical link is not possible, an X2 interface implementation might be realised by connecting eNodeBs via the MME in the EPC. Certainly, the choice of implementation has an effect on the latency of the X2 interface link; for instance, some procedures have QoS requirements, which may not be satisfied if there is a long delay in the link [27].

The X2 interface plays a key role in balancing traffic load between eNodeBs. The load balancing procedure uses exchanged load status information, which enables eNodeBs to adjust handover and cell reselection parameters. The adjustment of the parameters offloads UEs from heavily loaded to lightly loaded eNodeBs in a uniform distribution. The load status information exchanged can be of various types, such as physical resource block usage, classified into real time and non-real time traffic, hardware usage, and processing time. The X2 interface is also responsible for UE handover between eNodeBs. For a comprehensive discussion on the X2 interface the reader is referred to [37] and [27].

2.11 Radio Resource Management

The basic understanding of radio resource management is essential to our work. In this section, therefore, we present a brief discussion on radio resource management. This includes packet scheduling, handover and load balancing.

In LTE, radio resource management involves the co-operation of the first three layers of the protocol stack. Layer 3 performs QoS management and admission control for new data flows. Then in layer 1and 2 link adaptation, hybrid ARQ, and packet scheduling functions are performed. For scheduling decisions, the scheduler uses channel quality indicator (CQI) reports for the downlink channels and services reference signal (SRS) reports for the uplink channels. The CQI and SRS reports assist eNodeBs to make appropriate link adaptation decisions. Due to the rapid fluctuations of radio channel conditions and UE, spatial diversity link adaption is necessary. The

scheduler also schedules UEs in time and frequency dimensions, thereby providing frequency and time diversity [31].

2.12 Packet Scheduling

Scheduling is concerned with the dynamic allocation of time-frequency resources to users according to their requests, QoS requirements, and the prevailing channel conditions. A scheduler is a key part of the MAC layer, which controls the allocation of downlink and uplink resources. The scheduler makes scheduling decisions every 1ms, which is called a sub-frame. The basic unit of resources allocated to UEs is a resource block. The resource block is 1ms long and 180 KHz wide in dimensions, as highlighted in an earlier discussion on resource blocks. Scheduling is the responsibility of the eNodeB in both the downlink and uplink. However, the downlink and uplink procedures are independent. In the downlink scheduling, the scheduler decides which terminals to transmit to and the channel to use. The scheduling process starts by the UE measuring the CQI from downlink channel and sends CQI and buffer size reports to the eNodeB using uplink control channel. Then the eNodeB evaluates the buffer size and CQI reports, and includes the QoS requirements corresponding to radio bearers of the data to be transmitted. The evaluation determines the modulation and coding scheme to be used and an appropriate physical resource block mapping pattern, which are then sent to the UE [31].

2.13 Handover

In order for users to associate with suitable eNodeBs, cell selection or handover occurs. In LTE, handover is only performed if a user is currently connected, that is it is an RRC connected state, otherwise cell selection is performed [38]. A handover procedure between cells that share the same MME and have an X2 interface is normally implemented via the X2 interface. However, if the X2 interface is not available the S1-C interface is used. A proper execution of handover procedure is important for good user experience, especially for delay sensitive traffic, but most importantly for conservation of backhaul resources.

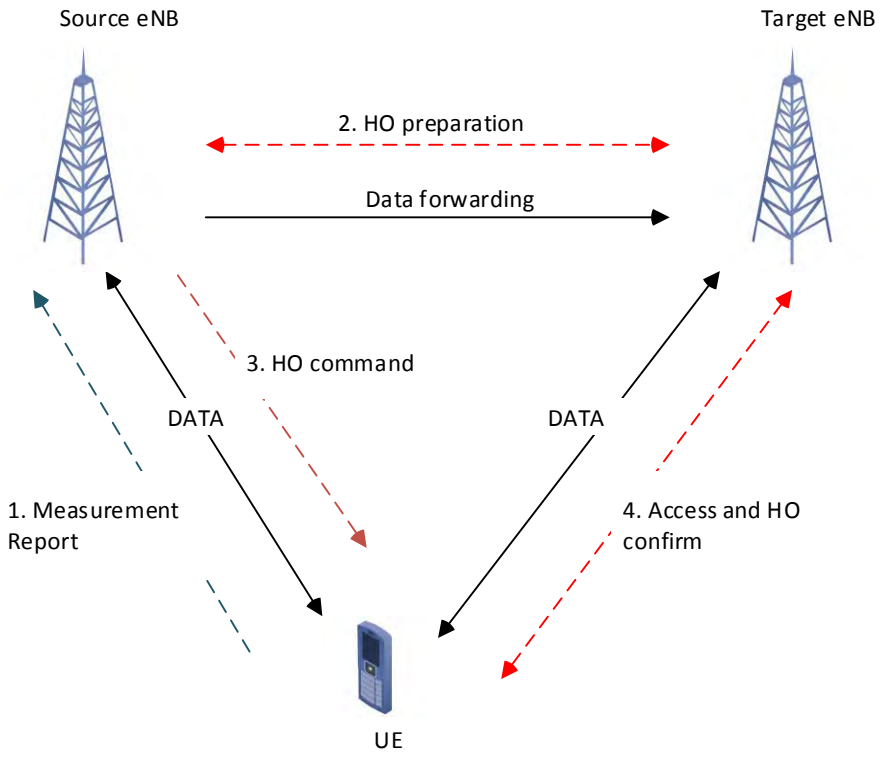

Figure 2.8. Handover procedure [26]

During a handover, user plane and control plane information is transferred to the target eNodeB. This ensures that packet losses are kept to a minimum, delivered in an ordered sequence, and the backhaul is efficiently used. For TCP based traffic, for instance, packet losses are undesirable since they trigger TCP retransmissions, and this could have adverse effects on the backhaul. A basic illustration of an LTE X2 handover procedure is shown in figure 2.8 [26], the procedure is divided into four stages. In the first stage, UE measures the RSRP from the downlink channel and reports to the eNodeB. Then in the second stage, the source eNodeB prepares for a handover to a target eNodeB and forwards user and control plane contexts. In the third stage, the source eNodeB sends a handover command to the UE. During the period when the UE processes the handover command and the time when the target eNodeB confirms the handover, there is a brief break of UE connectivity. Then finally, it connects to the target cell and the session is restored [26].

2.14 Heterogeneous Networks

Conventionally, planning in mobile networks is centred on homogeneous networks where the transmit power, antenna patterns and backhaul connectivity are generally similar. However, as the traffic demand increases this approach becomes cumbersome and expensive [5]. This calls for a more cost effective and scalable approach. The rationale behind the concept of heterogeneous networks is to cater for increasing data traffic whilst controlling operating expenses. Heterogeneous networks increase capacity by increasing network density. The network density is increased by adding a layer of small cells on a layer of macro cell network. A distinct feature that is common to small cells is the low transmit power and small coverage area. By definition a heterogeneous network or HetNet is a combination of small cells and macro cells [27]. The term heterogeneous network has been used before to denote co-located networks that are based on different access technologies, however, in the context of this work, it also means layered networks based on transmit power disparity. This follows how other similar works have contextualised the concept [5], [14], [27], [39].

Small cells allow for flexible deployments, which could depend on traffic demands or coverage gaps without serious network planning. The low transmit power improves energy efficiency and reduces operational expenditure. There are three broad types of small cells: which are pico cells, femto cells, and relay node based cells [5].

Pico cells are an operator-managed solution. They are normally open for public use, subject to operator access controls, but may give preference to users of a certain establishment. They are deployed in high-density areas to provide increased capacity or extended network coverage. In LTE, pico eNodeBs are connected to the core network via the S1 interface and can communicate with other eNodeBs via the X2 interface. Femto cells have similar characteristics as pico cells, except that they are user-owned and self-organising. They are normally only accessible to a closed subscriber group (CSG). Femto cells do not have the X2 and S1 interfaces but use a backhaul such as a fixed line, ADSL to connect to the core network [37]. Then finally, relay nodes, provide UEs with indirect access to donor pico eNodeBs. Relay nodes are wireless backhaul enhancement in LTE heterogeneous infrastructure. The purpose of relay nodes is to provide backhaul facilities in poor coverage areas such as cell edges.

2.15 Cell range expansion

As referred to earlier, macro base stations transmit at higher power than small base stations (pico-eNodeB or home-eNodeB). A lack of a proper strategy in deploying small base stations, in most cases results in smaller coverage areas by small base stations. The coverage area is particularly impacted by the high transmit power disparity in the macro and LPN network tiers [26]. Conventionally, in mobile networks, including LTE, UEs connect to a base station, which provides the strongest downlink signal strength. This strategy provides each UE with the best available channel conditions but may lead to load imbalance in HetNet deployments, and the resources provided by LPN will be underutilised. The macro base stations will be overloaded, as most UEs will be attracted by the high downlink signal strength but without sufficient resources to provide to the UEs. There will be an unfair distribution of load and user experiences [26].

To address the problem, CRE was adopted by 3GPP, starting from release 9 [40]. An illustration of cell range expansion is shown in figure 2.9. In the figure it is shown that, range expansion allows more UEs in the vicinity of the LPN to associate with the LPN. This technique results in more users being offloaded from the macro base station. CRE therefore, increases the average size of small cells by adding a bias value to the measured signal strength during cell association. The user-cell association strategy is then based on the signal strength and bias. An illustration is shown in figure 2.10.

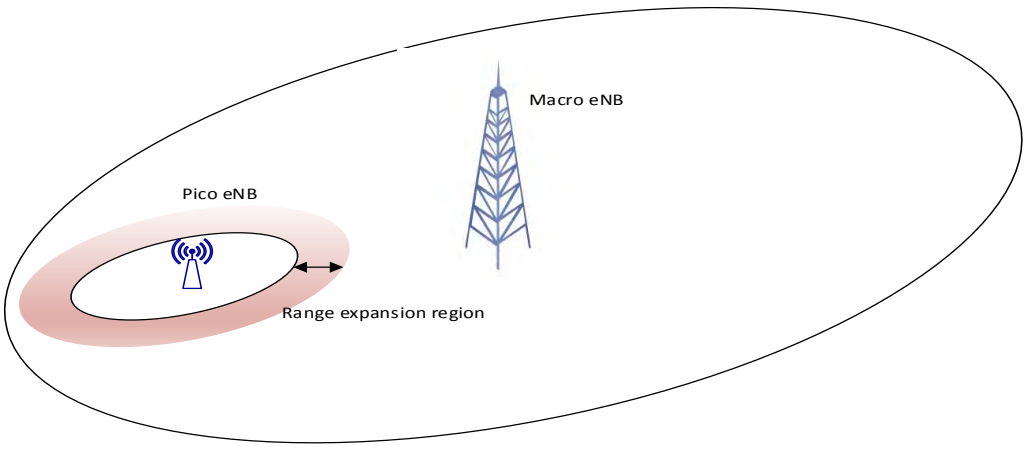

Figure 2.9. Illustration of cell range expansion

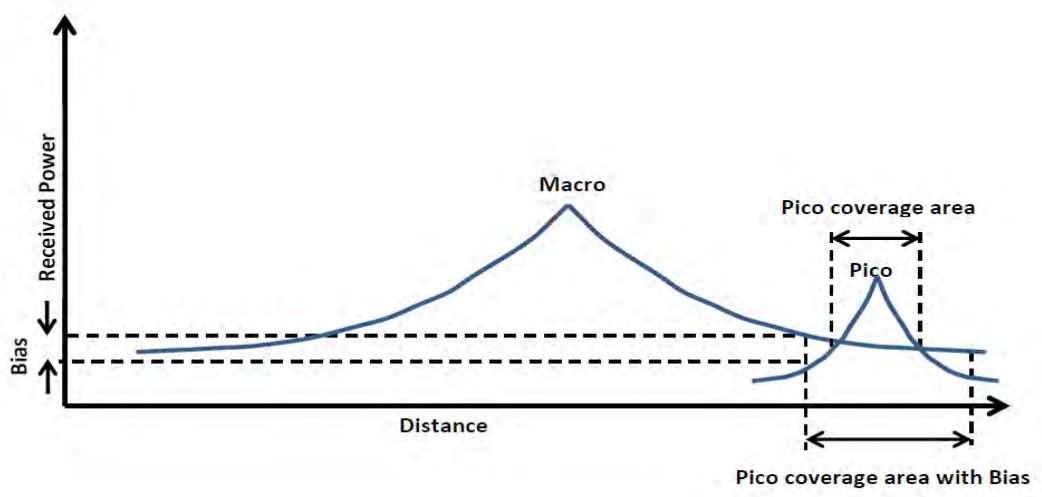

Figure 2.10. Signal strength curve [14]

2.16 Chapter Summary

This chapter presented a basic summary of LTE architectural and related aspects of the technology to aid in understanding the load balancing solutions being proposed in this work.

First, highlight of the evolution of 3GPP technologies from GSM to LTE was provided. Then, the network architecture was discussed, putting an emphasis on the radio access network. The radio access network architecture was described as flat since it uses one node unlike in previous technologies. Further, description of radio access network protocol stack, communication channels, and the resource grid was provided. Since the study is in the radio resource management space, an overview on packet scheduling and handover was also provided. Finally, a discussion on HetNets and cell range expansion as a technique for traffic offloading was provided.

Chapter 3
Load balancing in LTE Advanced

3.1 Introduction

The random distribution and variation in traffic densities are inherent challenges in mobile networks due to uneven load distributions. Load balancing is key to equitable sharing of load amongst deployed nodes of a radio access network. Load balancing is the process of equitably distributing load in a network to ensure that resources are efficiently utilised. Load balancing can be static or dynamic. In legacy systems, load balancing was performed statically without considering the current load and resource status of the network. This was achieved by provisioning for peak traffic densities when configuring the system during the planning stages of the network and then later on manually optimise it during operation [41]. The load balancing process was performed in long time intervals and the network resources were not effectively utilised. In LTE, load balancing is dynamic, network nodes are self-organised based on load status and network resource measurements and parameter tuning. This is part of a broader set of functionalities known as self-organising network (SON) functions [32]. Automation of load balancing enables the network to be responsive to traffic density changes and allocate resources where they are needed timeously.

This chapter presents the state of the art on load balancing with particular focus on LTE/LTE-Advanced load balancing approaches. A general classification of load balancing approaches in wireless networks is presented, then load balancing schemes on each of the classified approaches are discussed. The last section discusses LTE load balancing schemes.

3.2 Load balancing approaches

Load balancing in mobile networks can be generally classified into resource allocation scheme and traffic steering schemes. Resource allocation schemes use channel borrowing and resource scheduling strategies to balance load. Traffic steering schemes actively shift traffic to lightly loaded cells. In traffic steering schemes load balancing is achieved by tuning mobility related parameters and configurations such as handover parameters, cell selection parameters, and offsets. Figure 3.1 shows a general classification of load balancing schemes [19][42].

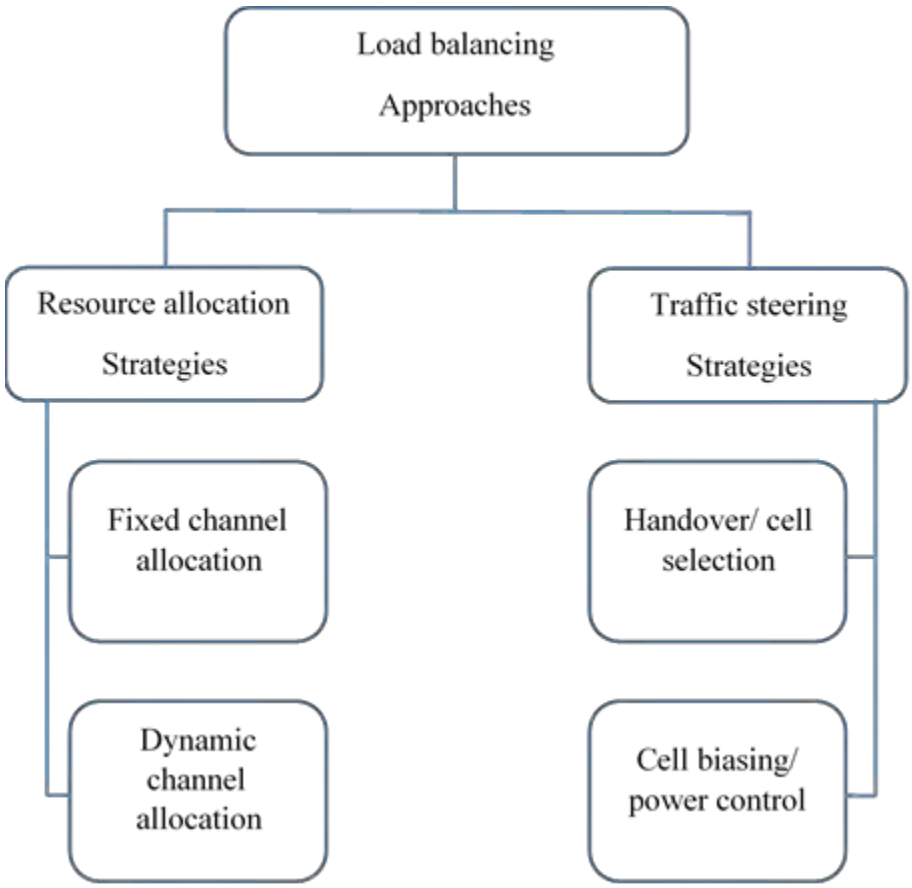

Figure 3.1. A general classification of load balancing schemes.

3.3 Resource allocation based load balancing schemes

Resource allocation schemes are concerned with moving resources to where they are needed, that is assigning network resources to heavily loaded cells in the case of load balancing in mobile networks. Unused network resources may be assigned through a centralised controller or a distributed mechanism [42]. Centralised resource allocation approaches normally have a hierarchical command structure, which is simpler to manage but is architecturally rigid and is susceptible to the "single point failure problem". Distributed resource allocation approaches are characterised by a simpler architecture, where nodes exchange resource status information. The information exchange, however, increases complexity of resource allocation schemes [42].

The strategies employed in resource allocation schemes are predominantly channel borrowing oriented. A highlight of channel borrowing strategies for legacy cellular networks such as hybrid channel assignment, channel borrowing without locking and load balancing with selective borrowing is presented in [19]. Since the interest of this work is on traffic steering based approaches the discussion on resource allocation schemes for load balancing is brief.

3.4 Traffic steering based load balancing schemes

In contrast to resource allocation schemes, in traffic steering schemes, traffic is directed to where resources are available. Therefore, traffic steering is generally defined as the directing and controlling of traffic from an overloaded cell to a cell that has underutilised resources [41]. Traffic steering is key in LTE heterogeneous networks, since they are characterised by multi-tiered networks, composed of macro, pico and femto cells. The network tiers provide the end user with variety of access networks. By using traffic steering, network resources could be exploited to the advantage of the end user by optimising network capacity and user experience. To achieve load balancing by traffic steering, network parameters and configurations such as handover offsets and cell selection thresholds are optimised for a given geographical area, such as a hotspot, according to traffic density.

3.5 Earlier work

Earlier works on traffic steering oriented load balancing were predominantly based on the concept of cell breathing in macro cell only networks. In this section, efforts will be devoted to discussing works related to cell breathing, particularly because they provide groundwork for understanding the ideas proposed for our work. A brief review of other techniques will then be given towards the end.

Cell breathing techniques were employed in legacy mobile networks such as GSM, WCDMA, and later on, other wireless networks such as wifi and WiMaX [16], [17], [42]. Cell breathing is defined as a traffic congestion relieving mechanism that adjusts the coverage area of a cell according to the load [16]. An illustration of cell breathing is shown in figure 3.2. An overloaded cell's coverage area shrinks whilst a lightly loaded cell's coverage area expands to compensate. The shrinking of the overloaded cell offloads users from overloaded cell, forcing them to associate with lightly loaded neighbouring cells to balance the overall network load.

Usually, cell breathing is implemented by controlling base station or access point transmit power to adjust the coverage area. There are different variants of cell breathing techniques. In CDMA systems for instance, the coverage area of a cell is adjusted by controlling the pilot signal power, whilst in GSM it could be done by adjusting power of other control channels [16].

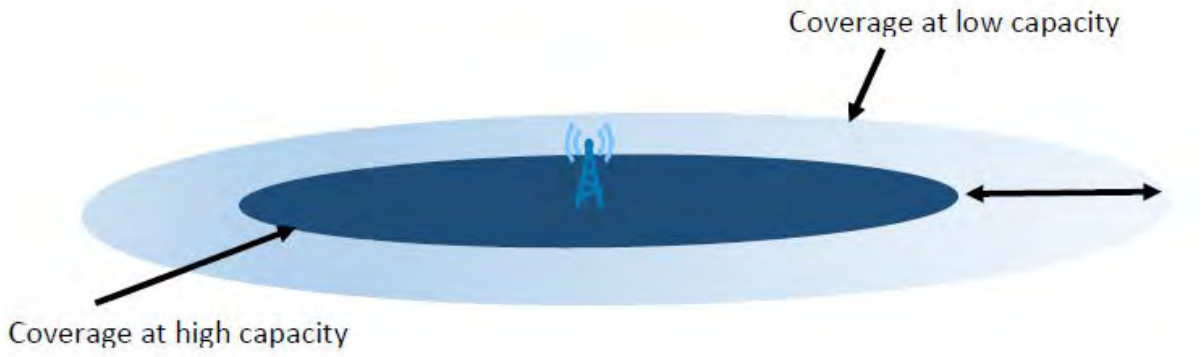

Figure 3.2. Cell breathing

Even though the idea of adjusting transmit power to cover specific geographical areas has been in existence ever since the conception of cellular wireless networks it was not initially used for traffic steering. Transmit power adjustment was mainly used for network planning purposes and interference mitigation amongst neighbouring cells. As a load balancing mechanism, earlier work on cell breathing/power control was promulgated by [16], [17], [43]. In [43] an observation, based on mathematical analysis, was made that cells could expand and contract as a result of power adaptation according to traffic density, which by extension meant relief for congested cells. In [17], an algorithm that combines the adaptation of user transmit power and cell selection to achieve cell breathing was proposed. This work was motivated by findings in [43]. It was assumed that the network layout is such that cells could overlap by expanding and contracting. The algorithm uses a decentralised approach where the users and base stations independently measure interference and determine load status from other users and base stations and compute their transmit powers accordingly. In as much as this algorithm produces an optimal solution for load sharing it adds computational complexity and signalling overhead, rendering it impractical for real systems. In

[16] users in overloaded cells were distributed by cell breathing technique for a CDMA system. A simplified approach was employed, where users only measure the base station SINR to ascertain whether a cell is crowded, hence, a need to associate with another cell.

This technique was also adopted in some load balancing schemes for WCDMA networks, such as schemes quoted in [4] and [5]. Cell breathing based techniques were further explored for other technologies such as WiMAX and wireless LAN. Cell breathing for load balancing in WiMAX was applied in [46] where cooperative signalling was used to coordinate transmit power levels of base stations from a centralised command entity. To ensure quick response an optical network was used to connect the centralised command entity to the wireless domain. The concept of cell breathing has also been used in [47]–[49] for wireless LANs. The common theme in the quoted work is that the purpose for this technique was mainly to regulate traffic density in hotspots to enhance QoS.

Other techniques for traffic steering based load balancing have been proposed, such as the use of smart antennas and related power control methods [50], [51]. A multi agent mechanism for controlling cell coverage is proposed in [50]. Cell coverage is controlled by shaping antenna patterns. A framework for base station negotiation is devised, which gives base stations autonomy to quickly adapt to impromptu traffic changes without significant signalling overheads. In [51], a technique that emulates bubble oscillations for a realistic irregular shape of cell coverage is proposed. A bubble represents a cell, and the air inside the bubble, traffic density. The shape of the coverage area is achieved by manipulating smart antennas radiation patterns. The antenna patterns change dynamically according to changes in traffic density. The technique is formulated as a multidimensional optimisation problem where the network capacity is maximised whilst transmit power is kept to a minimum.

Cell breathing is a concept that is usually associated with CDMA systems, where it is used as a mechanism to contain interference and congestion. As traffic increases, the effects of this technique may not be desirable for users who could experience call dropping and blocking. However, from the work reviewed here it is observed that the power behind the concept has been harnessed to achieve load balancing in varying contexts. To ensure that there are no coverage holes during cell breathing most of the techniques apply some cooperative mechanisms, or the base stations are managed from a centralised controller. Another common theme among the techniques

is the use of objective functions for optimisation. The use of objective functions provides optimal solutions, but the computational rigour required may render some of the techniques impractical. The ideas proposed in our work apply cell range expansion for load balancing in LTE. This is a concept that has distinctly similar traits as cell breathing and by extension there are transferable ideas from some of the works. The difference between cell breathing and cell range expansion is that cell breathing adapts base station transmit power to adjust coverage area of a cell, but cell range expansion, as discussed in section 2.4.5, applies a virtual bias value to handover margin of users in the vicinity of a lightly loaded cell.

3.6 Recent work

This section presents a literature review of recent works related to load balancing, especially in HetNets. Usually load balancing in HetNets is studied with interference coordination by resource partitioning between network tiers. LTE advanced uses enhanced interference coordination (eICIC) as interference coordination scheme. Base stations take turns to mute certain sub-frames to allow each tier to transmit without interference. Actually network performance is dramatically enhanced by the addition of eICIC techniques when used in conjunction with cell range expansion in small cells [52], [53]. Our study, however, focuses on the user-cell association aspect of load balancing, hence, the focus is on load balancing approaches only. The review is in two categories; cell range expansion based approaches are presented first, then followed by utility function based approaches. In some of the works the joint usage of cell range expansion and utility functions proved to yield better load balancing performance, which is a concept that is also adopted in our work.

3.6.1 Cell range expansion based approaches

Cell range expansion was conceived as a strategy for macro cell traffic offloading in HetNets and adopted by 3GPP as part of the features for LTE release 8 [54]. The offloading is achieved by extending the coverage area of small base stations overlaid on macro cells. A number of studies have investigated and confirmed the traffic offloading gains attained by the strategy [12], [13], [55], [56]. Work presented in [13] demonstrated the benefits of CRE in HetNets, where path loss and predetermined sum data rates were used as criteria for CRE. In [55], a scheme that uses a binary cell range expansion bias mechanism is proposed. All UEs, which report RSS below

a particular threshold value, are assigned a high-predetermined bias value, otherwise, they are assigned a low bias value. It is a simple and intuitive approach, but it does not provide mathematical guidance, especially the mathematical framework to determine bias values. Furthermore, the proposed approach is rigid; the use of a two level bias mechanism might not translate to the desired offload gains.

Another scheme that adapts range expansion is proposed in [56], however the bias is determined by traffic demands of user equipment. In this work, each UE is assigned a unique bias value according to its traffic needs. UEs with higher uplink traffic demand are assigned a proportionately larger bias to increase their chances of associating with pico base stations. UEs with larger downlink traffic demands are assigned smaller bias values to force them to remain connected to macro base stations. By assigning most uplink traffic to pico base station this reduces the impact of interference on UEs in the CRE region, but might be at a cost of underutilising pico base station resources. This assertion is based on the understanding that, uplink traffic usually accounts for only one third of total traffic in the network. The computational and signalling requirement could be significant in a real system, rendering this approach impractical.

Work in [12] presents an insightful two pronged range expansion scheme, which considers downlink and uplink interference coordination. Pico cell coverage area is expanded from either an equal received signal strength boundary (i.e. the point where the macro and pico base station signal strengths are equal) to an equal path loss or hot spot boundary. Mathematical formulae are derived based on pathloss and downlink received signal strength to determine range expansion values for desired pico cell-coverage areas. The reasoning behind this work was to provide a framework to determine CRE bias upper bounds beyond which the received signal of CRE region UEs becomes very poor to meet SINR requirements of PDDCH decoding in the downlink. Another almost similar technique to estimate upper bound CRE bias values is proposed in [57].

Cell range expansion, as a cell association strategy for macro cell traffic offloading, is sufficiently explored in literature, [21], [52], [55], [56], but its potential for load balancing in HetNets needs more consideration. Traffic offloading and load balancing are related concepts, but they differ in the sense that load balancing requires an optimisation mechanism to ensure fair sharing of load amongst network entities, whilst traffic offloading is geared towards congestion relief on overloaded entities. Considerable work on load balancing in macro cell only networks

has been done as referred to earlier, where transmit power control (cell breathing) based and smart antenna techniques were used. From the reviewed literature, the use of cell range expansion for load balancing is explored in [20]–[22], [58].

The work in [20] proposes two adaptive schemes, which dynamically adapts sub-frame muting ratio and CRE bias according to network load and UE traffic demands. In the first scheme, which is centralised, signalling is coordinated from a macro cell base station. The load and traffic demand are periodically estimated from information reported by UEs. CRE bias values and muting ratios for all the base station are then computed from the estimates. The process of computing the values of the two parameters is iterative; they have to be continuously updated until the network is balanced. The centralised system is accurate in load balancing, but is heavy on signalling overheads. In the decentralised scheme, each base station performs its own estimations, and CRE bias value updates. Another scheme in [21] adapts the CRE bias according to cell load and uses an adaptive power allocation mechanism to mitigate inter-tier interference. The idea behind this scheme is to balance throughput of macro cell centre and edge users by using logarithmic utility functions. The power allocation scheme ensures that a macro base station's maximum transmission power on muted sub-frames does not interfere with pico UEs when they are transmitting on those sub-frames.

Reinforcement learning techniques are used in [22] where a joint mobility management and UE handover approach was proposed for load balancing. Macro and pico base stations continually learn to optimise their CRE bias and UE associations from interactions with UEs. Each base station optimises its load and associations with limited coordination to achieve load balancing. Another learning based technique for load balancing is proposed in [58], where two distributed learning algorithms are devised, the log-linear learning algorithm(LLLA) and binary log-linear learning algorithm (BLLLA). Base stations determine their optimal bias values by using a game theory oriented model for interaction amongst themselves. The LLLA provides sub-optimal solutions, but is fast enough for practical applications, as compared to the BLLLA.

Most of the reviewed works related to cell range expansion propose the use of fixed bias values. However, the use of fixed bias values has limitations; small base stations are deployed in spatially diverse and constantly changing traffic densities. Therefore, the coverage area has to respond quickly to these traffic density changes. It is also apparent that setting the bias value by

trial and error may not provide the desired offloading. If a small bias value is used, the offloading may be insufficient. However, when the bias is too large, the CRE region UEs may suffer severe macro cell interference and not be able to satisfy the SINR requirements for PDCCH decoding. In this work, the use of a dynamic CRE bias strategy is at the centre of our load-balancing proposition, hence, the need to review relevant works. Some works in literature, such as some we have discussed here, consider the adjustment of CRE bias to reflect the dynamic nature of a wireless network. The evaluation of the power allocation strategy in [21], which is discussed earlier in the section, demonstrates the benefits of using dynamic CRE bias values, which props up overall network throughput. Whilst in [9] a Q-learning technique was applied to determine user equipment specific CRE bias values. UEs are agents, which learn from their environment by measuring received signal strength from target base stations, and update new information in a Q-table that resides in the UE. Based on information learned UEs independently determine appropriate bias values. In this approach, communication with target base station is minimal; however, mobile UEs have to constantly rebuild their Q-tables as they are handed over from one cell to the other. This challenge could impede the learning algorithm's ability to decide on optimal CRE bias values, thereby degrading system performance.

In [4], an approach similar to our work is proposed, where cell specific bias values are used for cell range expansion in low powered nodes. The algorithm uses a CRE bias optimisation mechanism that is built on a bounding scheme to enhance response time. Additionally, a range optimisation framework that recognises a cell load coupling relationship between cells is derived. This framework is characterised by a system of nonlinear equations each representing load of each cell. To allow for a quick convergence of the solutions these equations are consolidated into a bounding scheme that estimates upper and lower bound cell loads. The estimated loads are then fed into a load-balancing algorithm, which uses a statistical method called the design of experiments (DOE). A predefined vector of CRE bias values is used to select the most fitting bias values for estimated loads. In the load-balancing algorithm, the bounding scheme is repeatedly called until all the values, which match the load estimates, have been found. Even though it is not explained how the algorithm could be effected in a practical application, it is clear that it would have to be centralised. Although, such an approach could result in a single point of failure and may not be suitable for a flat architecture such as that of LTE. In our work, we take a decentralised

approach, and CRE bias values are not selected from a vector, but are uniquely defined for each load balancing situation.

3.6.2 Utility function based approaches

Performing load balancing by associating users to corresponding cells in HetNets is a complex optimisation problem, partly due to the power disparities between network tiers and interference in the network. The complexity increases exponentially as the network increases, which calls for different approaches to solving load-balancing problems. This section presents a review of various utility function based load optimisation techniques. When using utility functions for load balancing the load-balancing problem is characterised by coupled multiple objective functions, maximisation of the prescribed utilities in the functions is difficult even for reasonably sized networks. The use of utility functions has been explored for load balancing in optimisation problems in [59].

The work in [59] proposes a user association policy-based framework for distributed user cell associations in wireless networks. Several user association policies, which are grouped and denoted as alpha-optimal user associations, are derived. An iterative distributed user-association policy algorithm is then developed which adapts these policies according to changing loads without shared information amongst the cells. The user association problem is formulated as a convex optimisation problem, where, for simplicity, a fully loaded model is assumed, and users are allowed to associate to more than one base station. In [60], logarithmic utility functions are employed to a network utility maximisation problem. The utility function is based on achievable data rates, and the constraints are unique user-base station associations and base station power control. The work proposes an asynchronous and distributed price update strategy, where users are assigned to a base station according to the value of a utility function. Due to the complexity of the user cell association optimisation problem, the work resorts to a heuristic approach, where user-base station association and power levels are optimised iteratively to obtain a sub-optimal solution.

In [19], logarithmic utility functions are also used for a joint user-cell association and resource allocation optimisation problem. The work investigates optimal and sub-optimal solutions to user cell association that results in load balancing. The approach taken emphasises on distributed coordination of information. Since user-cell association and resources in that cell are

coupled, the optimisation problem is simplified by assuming that each user can associate with at least two base stations. This assumption decouples the problem and provides an upper bound to the network utility that can be achieved. Although, this assumption reduces complexity to the problem it is hard to solve. A more feasible method is then formulated by using logarithmic utility functions. Based on this method a sub-optimal distributed algorithm is developed. In the algorithm a price-based method characterising the user-base station associations and resources such as the one in [60] is used. The load balancing proposed in our work is closely related to [19], [60]. We adopt the use of logarithmic utility functions in a joint optimisation of resource allocation and user-cell association, and the idea of using a price strategy. However, in our case the price moves the CRE bias for load balancing.

Other works that follow similar approach include [39], [61]. In [61] load balancing, by client-access point associations, is investigated on 60GHz millimetre wave short range communication links. The utility function in this case is based on access point(AP) utilisation, which represents the demanded data rates and link quality. A distributed load-balancing algorithm based on the langrangian duality theory and subgradient methods is proposed. The langrangian is particularly used here to simplify the problem by relaxing the constraints of the utility function. Work in [39] proposes a distributed load balancing algorithm, which adapts the bias to distribute load fairly amongst base stations. This work differs from the other works discussed in that the optimisation problem is formulated as a local utility maximisation problem, where each base station determines its own local optimal user association. A heuristic approach is taken for adapting the bias based on the premise that, intuitively if a cell is over loaded it must shrink and expand otherwise. It is assumed there are operator-determined upper bounds for the bias. The bias is adapted in a step-size fashion in such a way that if a cell is overloaded it reduces its bias by one-step at a time, thereby offloading users in the range expansion region to lightly loaded cells. Likewise, if the cell is under loaded it increases its bias in steps. Unlike this approach, in our work, we use a subgradient technique to determine bias values.

3.7 Other approaches

Other works employ markov decision processes as a foundation for investigating and modelling sequence oriented optimisation schemes [6]. Generally, the whole idea in using markov decision processes is to influence the current state of system for future desired outcomes. In

HetNets, markov processes have usually been used in modelling handover, and by extension they can be used in load balancing problems. Work in [6], for instance, proposes an energy aware sleep-wake up scheme for load balancing in a macro and femto cell HetNet, whereby the system of switching is modelled by markov decision processes. In this scheme, femto cells are switched off when the femto cell is not heavily loaded and provided the macro cell can handle all the traffic.

For decentralised load balancing approaches game theory has been used in some studies such as in [7]. Game theory, as strategic decision-making tool, allows for analysing negotiation based decision-making mechanisms, and this trait has been exploited in HetNets for solving decentralised optimisation problems with low signalling overhead. As an example, in [7], a hybrid and cooperative load balancing scheme is proposed. In this scheme users make decisions based on broadcasted load information from base stations, however, if there are network-wide issues such as traffic congestion a centralised entity assists the users. In other studies graph theory has been used to solve the optimisation problem in HetNets, such as in [8], which proposes a graph theory-based network flow oriented optimisation approach. An algorithm that performs load balancing and reduces unnecessary handovers is designed. The algorithm sequentially adjusts handover parameters of each cell according to the cell's load status, starting from the least loaded to the most loaded cell. The manner in which this algorithm is designed limits its efficiency since the execution is from a central point, and it is performed in sequence. Therefore, it might not respond quickly enough for UE handoffs. Additionally, the algorithm assumes, there will always be under loaded cells surrounding an overloaded cell, ready to receive distressed UEs.

3.8 Performance metrics

A performance metric is a measure of a system's performance based on defined system specific criteria. On top of providing a measure of improvement on a system, a performance metric may elucidate on the behaviour of a system when subjected to various relevant stressors. We subject our load balancing algorithm to three performance metrics. These are fairness, throughput, and UE offloading to determine the effectiveness of the algorithm.

3.8.1 Fairness index

Fairness is a desirable metric in wireless networks. It assists in determining if there is fair sharing of network resources amongst users. Application areas, where this metric may be used, include investigating TCP Fairness for congestion control and fair sharing of spectrum [62]. Various methods of evaluating fairness are proposed in literature, which include min-max fairness, TCP fairness, and Jain's fairness index. The evaluation of a system for fairness generally depends on the resources shared amongst entities of a system. The resources can either be time or frequency based resources or a combination of both resource dimensions, and are measured as time slots, sub-channels and resource blocks respectively. To evaluate fairness in our algorithm we use the Jain's Fairness Index. The fairness index is a dimensionless metric that provides a perspective on the fairness of a system in sharing resources amongst its users [25]. It is defined as:

$$Fairness\ Index(I) = \frac{(\Sigma_j L_d^j)^2}{N_{ue} \Sigma_j (L_d^j)^2} \tag{3.1}$$

The value of the Jain's Fairness Index varies from $1/N_{ue}$ to 1. The range of the fairness index indicates the extent to which a system is fair. In our case, $1/N_{ue}$ implies that the load is heavily skewed towards one eNodeB, which takes the whole load. Obviously, this is not possible since eNBs are spatially distributed. It is an extreme case of an unbalanced network. On the other hand a value of 1 means the load is equally shared amongst the eNodeBs. The performance of the algorithm in terms of fairness will, therefore, be determined by the load balancing index; a larger load balancing index will imply the system performs better, otherwise the performance will be deemed poor.

3.8.2 System throughput

Throughput is a good measurement to determine the efficiency of network resources utilisation. High throughput is always desirable in a system, since this implies network resources are efficiently utilised. Throughput can be measured in different ways depending on the circumstance. The measurement units include, bits per second (bps), packets per second and packet arrival to packet departure ratios. From a fundamental understanding of transmission of information, the over-arching tenet for all throughput measuring approaches is the Shannon

Hartley theorem, which defines the upper bounds on information that can be transmitted on a given channel [46].

Generally, there is a trade-off between throughput and fairness in a system. To increase the throughput of a user in a given system, for instance, requires that other users receive a lesser share of network resources [56]. In the case of HetNets, normally overall network throughput is throttled by the overloading of macrocells because of a poor association strategy. The idea in this work, therefore, is to improve overall network throughput by efficiently utilising low power node resources, as users are actively pushed to low powered nodes.

3.8.3 UE offloading

UE offloading effect could be a useful performance measurement to determine the effectiveness of a load balancing algorithm as illustrated in [22], [39],[54]. The expectation, when a load-balancing algorithm is used, is that UEs are actively pushed to low powered nodes. The offloading effect can be measured as a percentage of UEs associated with LPNs as suggested in [54]. We use this measurement to ascertain the responsiveness of our load balancing algorithm in macro user offloading to pico nodes. We expect that for over loaded macro nodes the algorithm will offload users aggressively in keeping with the load disparity between macro node and pico nodes.

3.9 Chapter summary

This chapter presented a review of load balancing and CRE approaches related to our study. We first considered the general classification of load balancing approaches in wireless networks, where we identified that they are broadly categorised into resource allocation and traffic steering schemes. On review of relevant CRE works, the lack of a CRE bias optimisation framework was noted. Such a framework could act as a guide for CRE bias assignment. The second main aspect of the review considered utility function based optimisation schemes for load balancing. Utility optimisation schemes provide optimal solutions and insightful theoretical fundamentals to optimisation problems, however, the computation rigour involved renders them impractical for wireless network application. Finally, we presented a brief review of the performance metrics used in evaluating the effectiveness of the algorithm in terms of fairness, throughput and user offload.

The discussion in the following chapter is built on the backdrop of the reviewed works, where we present the formulation of the load-balancing problem and the design of the algorithm.

Chapter 4
System design

4.1 Introduction

This chapter presents a detailed description of the general system model, the optimisation model, and the proposed cell range expansion load-balancing scheme. HetNets are a flexible and cost effective way to scale up wireless networks. However, the challenge, is power disparity between tiers, which unfortunately reduces coverage of small base stations, and the expected capacity boost is not realised. Furthermore, this results in load imbalance and poor utilisation of deployed network resources. This work uses cell range expansion to improve load balance between macro and pico cells. The proposed load-balancing scheme is built on the capabilities of the cell range expansion technique, which was merely meant for traffic offloading. To achieve load balancing, this technique is jointly used with utility function based optimisation techniques. Due to the geographical disparate traffic densities in wireless networks, the scheme uses adaptive cell range biases to reflect local traffic densities.

4.2 System model

We model our network as a two-tier LTE Advanced cellular network consisting of macro-eNodeBs and uniformly distributed pico-eNodeBs. Both macro-eNodeBs and pico-eNodeBs are deployed on the same channel. The pico-eNodeBs are deployed to provide hotspot coverage for places such as malls, airports, and schools. The power disparity that is observed in HetNets causes downlink and uplink coverage mismatch, it is, however, severe in the downlink, as it consequently causes load imbalance. For that reason, we focus on the downlink cell coverage. Interference co-ordination between network tiers is not considered for this work, as the emphasis is to investigate load balancing by user association. In any case, time domain interference coordination techniques could be included in future work. The pico eNodeB's coverage area is divided into CRE and cell centre regions; an illustration is shown in Figure 4.1. The network coverage area is defined as the combination of the coverages of all the eNodeBs. We assume the eNodeBs are connected to the core network via a backhaul, which may be a physical or wireless link and can communicate amongst themselves via the X2 interface. The purpose of the X2 interface is to communicate Load

status information and transmit data during UE handover. In the physical layer, an OFDMA system is considered, where all the eNodeBs have the same bandwidth. The bandwidth is split into smaller sub-carriers which are grouped in units of 12 and scheduled as resource blocks. Physical resource blocks (PRB) are the basic time-frequency resource units that can be assigned to users.

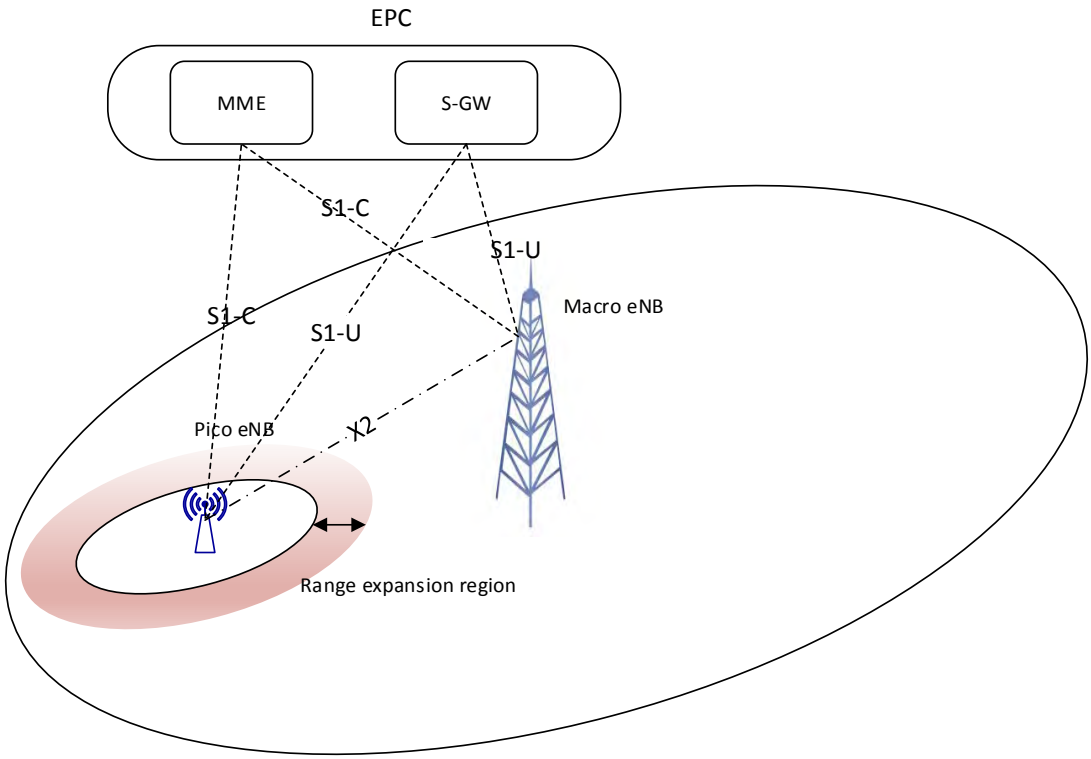

Figure 4.1. An illustration of macro-eNodeB and pico-eNodeB HetNet that uses CRE

We implement CRE bias optimisation on pico eNodeBs only, and the bias values dynamically change in proportion to the load status of each cell. Conventionally UEs associate with an eNodeB that has the highest measured Reference Signal Received Power (RSRP) value. To rebalance loads, CRE is employed, whereby the low RSRP of a pico eNodeB is virtually biased

to actively push cell edge macro-eNodeB UEs to pico cells. Therefore, the association rule of UE i to neighbouring pico eNodeB j is defined as,

$$j^* = argmax_j\{RSRP_{ij} + Bias_j\}, \tag{4.1}$$

where j^* denotes the most suitable pico eNodeB to associate with UE i. $Bias_j$ is the cell range bias of eNodeB j, a varying positive value for pico-eNodeBs and fixed to zero for macro eNodeBs. The variation of the bias is determined by a price-based utility optimisation technique, yet to be discussed in section 4.3.

It is further assumed that users are continuously transmitting at a constant rate and transmit power during their association period. Since a hotspot deployment is considered, users are assumed to have limited mobility. Each user can associate with one eNodeB at a time when it accesses the network via the wireless channel. We use J and I to denote the set of eNodeBs and UEs respectively. The notations i and j are used as identifiers of each UE and eNodeB respectively.

To denote user-cell association between UE i and eNodeB j, we use x_{ij} as an association index. This index is binary; it is one during the connection period of UE i to eNodeB j, otherwise, it stays at zero, as shown in (4.2a). Furthermore, a user can only associate to one eNodeB as shown in (4.2b).

$$x_{ij} = \begin{cases} 1, & when\ i\ is\ associated\ to\ eNB\ j \\ 0, & otherwise, \end{cases} \tag{4.2a}$$

$$\sum_{j \in J} x_{ij} = 1. \tag{4.2b}$$

The data rates of UEs are determined from the SINR measurements taken from resource blocks allocated to each UE. First the received power at UE i from eNodeB j is determined by $P_{ij}.g_{ij}$, where P_{ij} is the transmit power and g_{ij} is the channel gain. The channel gain is composed of path loss, shadow fading effects and antenna feeder losses. The SINR is defined as the ratio of the received signal power at UE i from eNodeB j, and the sum of the total interference power received from neighbouring eNodeBs and noise, as shown in (4.3):

$$SINR_{ij} = \frac{P_{ij} \cdot g_{ij}}{\sigma^2 + \sum_{k \neq j} P_{ik} \cdot g_{ik}} . \tag{4.3}$$

For modelling the channel of an established UE-eNodeB association, the channel model for hotspot heterogeneous deployments, configuration #4b, is used, as recommended by 3GPP. The channel model includes path-loss, antenna gain, and shadowing effects, appropriate for clustered indoor users in an urban environment. A detailed description of the model is found in [63]. Based on the Shannon Hartley channel capacity theorem, the maximum achievable data rate of UE i provided by eNodeB j on the allocated resource blocks is defined as [46]:

$$Rij = x_{ij} \cdot bw \cdot \log_2(1 + SINR_{ij}). \tag{4.4}$$

When a UE connects to an eNodeB it requires a portion of the available bandwidth. The overall load exerted by UEs to an eNodeB can be defined as the ratio of the sum of the data rates from each UE and the capacity. The capacity of an eNodeB is the total bandwidth that an eNodeB can allocate to UEs for communication. That portion of the bandwidth demanded by a user, we call, service demand. For the sake of simplicity, the bandwidth is allocated equally for all the users per eNodeB. Then the load in each eNodeB is denoted by L_d^j and is defined as:

$$L_d^j = \frac{\sum_i^{Nue} \omega_d^i}{W_c^j}. \tag{4.5}$$

The notations ω_d^i and W_c^j denote the service demand from each UE to an eNodeB and the capacity in each eNodeB respectively.

4.3 Problem formulation

Following [19], where the problem formulation was initially proposed, we use logarithmic utility functions. In our problem formulation, a user demands ω_d^i resources and receives a portion of W_c^j. This is essentially, a joint user association and load-balancing problem. By getting access to eNodeB resources users are adding load to an associated eNodeB. The optimisation problem takes into consideration the service demand and user associations. Using the logarithmic function approach makes such a problem manageable, as seen in [19] and [61]. The objective of the utility function is to maximise the utilisation of eNodeB resources by associating as many users as possible to each eNodeB. The user-cell association binary variable x_{ij} is included in the function to indicate when users are attached to an eNodeB. The idea is that, for every user-cell association

x_{ij} an eNodeB j must provide the required service in the form of ω_d^i. Therefore, the objective function that maximises the overall network utility of eNodeB resources and user associations is defined as:

$$\max_x \sum_{i\in I}\sum_{j\in J} x_{ij} \log\left(\frac{\omega_d^i}{W_c^j}\right) \tag{4.5a}$$

Subject to:

$$\sum_{j\in J} x_{ij} = 1, \qquad \forall\, i \in I \tag{4.5b}$$

$$\sum_{i\in I} x_{ij} \cdot \omega_d^i \leq W_c^j, \qquad \forall i \in I, j \in J \tag{4.5c}$$

$$x \in \{0,1\}, \qquad \forall\, i \in I, j \in J. \tag{4.5d}$$

The constraint in equation 4.5c is put in place to ensure that the demand from users does not exceed the capacity offered by the eNodeB. Additionally, the constraint in 4.*5d* ensures that a user can only be in one of two states, either attached or disconnected.

Based on the assumption that users receive an equal share of an eNodeB's bandwidth W_c^j, the bandwidth demanded ω_d^i by each user is equal to the bandwidth offered and shared among the users, that is;

$$\omega_d^i = \frac{W_c^j}{N_{ue}} \ , \tag{4.6}$$

where N_{ue} is the total number of users, which is equivalent to the number of associations in each eNodeB j :

$$N_{ue} = \sum_{i\in I} x_{ij}. \tag{4.7}$$

For simplification the expressions in 4.5 can be rewritten in an equivalent format, where a load variable C_j is introduced to signify the load on eNodeB j in relation to x_{ij}, as used in [61]. The introduction of this variable is based on the assumption that users equitably share network resources, and the load is directly proportional to user-cell associations. Furthermore, a parameter u_{ij}, which represents the utility of UE i attached to eNodeB, is introduced to make the objective function compact. The variables C_j and parameter u_{ij} are defined as;

$$C_j = \sum_{i \in I} x_{ij}, \tag{4.8}$$

$$u_{ij} = \log(\omega_d^i). \tag{4.9}$$

The introduction of the variable C_j to the expression emphasises the coupling relationship that exists between the eNodeB capacity and user associations, which could otherwise not be noticeable, and is useful for analysing the problem further. The expression then becomes:

$$\max_x \sum_{i \in I} \sum_{j \in J} u_{ij}\, x_{ij} - \sum_{j \in J} C_j \log(C_j) \tag{4.10a}$$

Subject to:

$$\sum_{j \in J} x_{ij} = 1, \qquad \forall\, i \in I \tag{4.10b}$$

$$\sum_{i \in I} x_{ij} = C_j, \qquad \forall\, j \in J \tag{4.10c}$$

$$x \in \{0,1\}, C_j \geq 0, \qquad \forall\, i \in I, j \in J. \tag{4.10d}$$

Since the optimisation problem has coupling constraints, to solve it further, we need to decompose it into smaller manageable sub-problems. Following the work in [61] we use the

Lagrangian dual decomposition technique, which is useful in separating coupling constraints. To decompose the problem we introduce a Lagrangian multiplier, λ_j to the coupling constraint, and then transfer it to the objective function, such that after grouping like terms a Lagrangian function is formed, as shown in (4.11) ;

$$\mathrm{L}(\mathrm{x}, \mathrm{C}, \lambda) = \sum_{i \in I} \sum_{j \in J} x_{ij} \left(uij - \lambda_j\right) + \sum_{j \in J} C_j \left(\lambda_j - \log(C_j)\right). \quad (4.11)$$

Then the optimisation problem is separated into two optimisation sub-problems, $f(\lambda)$ and $g(\lambda)$. Each sub-problem has a Lagrangian multiplier that is related to each eNodeB as shown in (4.11). In compact form the Lagrange function is written as:

$$\mathrm{L}(\lambda) = f(\lambda) + g(\lambda). \quad (4.12)$$

Then the optimisation of the sub-problems $f(\lambda)$ and $g(\lambda)$ can be written as:

1. For $f(\lambda)$;

$$\max_{x} \sum_{i \in I} \sum_{j \in J} x_{ij} \left(uij - \lambda_j\right) \quad (4.13a)$$

Subject to

$$\sum_{j \in J} x_{ij} = 1, \qquad \forall\, i \in I \quad (4.13b)$$

$$x \in \{0,1\}, C_j \geq 0, \qquad \forall\, i \in I, j \in J. \quad (4.13c)$$

2. For $g(\lambda)$:

$$\max_{C} \sum_{j \in J} C_j \left(\lambda_j - \log(C_j)\right) \quad (4.14a)$$

Subject to:

$$C_j \leq N_{ue}, \qquad \forall\, j \in J. \quad (4.14b)$$

To solve the dual sub-problem for optimal solutions, first the multiplier λ_j must be obtained from function $g(\lambda)$. However, $g(\lambda)$ is not differentiable, in [19] a sub-gradient technique is used to solve for lambda, so we follow the same approach. The solution is achieved iteratively, where λ_j is updated at every iteration according to the following expression:

$$\lambda_j^+ = \lambda_j - \delta \cdot \left(C_j - \sum_{i \in I} x_{ij} \right), \qquad \forall j \in J. \tag{4.15}$$

The iteration is performed until an optimal solution is found. The parameter δ is a step size function, which for our solution is kept constant and sufficiently small.

The multiplier obtained from the function $g(\lambda)$ can be used as a price of each utility, an interpretation that is suitable for the optimisation problem at hand. An eNodedB could have a unique price to regulate the number of user-cell associations it can handle. The expression in (4.13*a*) portrays a good interplay of the users and eNodeB load balancing. In the expression each UE i that is associated to eNodeB j maximises the UE utility u_{ij} reduced by the price. On the other hand, an eNodeB determines the price based on its load. The idea is to encourage users to associate with lightly loaded cells by advertising low prices, and higher prices for heavily loaded cells. Therefore, the price λ_j can be broadcasted to users to enable them in making association decisions that result to a balanced network.

4.4 Cell range expansion algorithm

We use the formulation in section 4.3 to optimise the cell range of pico eNodeBs for balancing the load of the network. We then build on the fundamental understanding of the price concept to vary cell range for load balancing. The two sub-problems are appropriate for a distributed system and can be implementable in an LTE system. From sub-problem $f(\lambda)$ we can deduce that a user associates with an eNodeB that maximises its utility. The association rule of user *i* to eNodeB *j* can be defined as,

$$j^* = arg \max_j \{u_{ij} - \lambda_j\}. \tag{4.16}$$

Our earlier discussion on the user utility and price relationship can be reduced to the rule in (4.16). It is clear at this point that the price influences the association decisions of the users.

According to the rule, each user chooses an eNodeB that offers the highest utility u_{ij}, less the price for all the possible user-cell associations. A low price tends to increase the utility, thereby incentivising users to associate to an eNodeB, and a high price does the opposite. The second sub-problem is essential for price updates using expression (4.15). Without losing our fundamental understanding of C_j and $\sum_{i \in I} x_{ij}$, we substitute C_j and $\sum_{i \in I} x_{ij}$ in (4.15) with the eNodeB capacity W_c^j and the eNodeB load L_d^j respectively, and can be written as follows:

$$\lambda_j^+ = \lambda_j - \delta \cdot \left(W_c^j - L_d^j \right), \qquad \forall j \in J. \tag{4.17}$$

Now, the expression exhibits a supply and demand behaviour in the optimisation problem. From observing the expression, it shows that the price is driven by eNodeB load. Similar expressions have been used in Least Mean Square (LMS) based algorithms, which have a variety of applications such as machine learning, scheduling in computer networks and signal processing. In [64] for instance, they use an LMS approach to design an adaptive scheduler for 802.11 systems. The basic idea behind the LMS approach is to determine the optimal price. This is achieved by updating the price in such a manner that the system converges to an optimal solution. We use this expression to obtain the price in our algorithm, where we assume the algorithm starts at a price of 1 and at each iteration the price is updated.

On expansion of the utility u_{ij} and exponentiation of the terms, the expression in (4.16) can alternatively be written as;

$$j^* = arg \max_j \{\omega_d^i \cdot e^{-\lambda_j}\}. \tag{4.18}$$

Our interest in (4.18) is the second term of the expression, which could be interpreted as a bias, $B = e^{-\lambda_j}$. This term is a multiplicative bias to the service demand ω_d^i of user j in eNodeB j. The bias is a decaying exponential expression, whose value follows a corresponding trajectory as the price moves. This is a desirable property for our solution, as it makes the load balancing algorithm naturally responsive to load changes. In a way the bias mechanism behaves like some exponentially decaying systems. A typical example is a damped system, such as a hydraulic door damper. A door damper allows a door to return to its original closed position at the quickest time possible whilst ensuring a rapid decrease of its speed. In the same manner we expect the bias mechanism of a pico eNodeB to emulate this behaviour. Now when the price is low a pico eNodeB

will advertise a larger bias, which means that pico can receive many users who can be provided with better utility. However, as the price increases, the bias rapidly decreases, thereby discouraging user-cell association. The rapid change of the bias in relation to the price lends itself to an agile load balancing mechanism.

Normally in LTE HetNets, user-cell association by CRE follows the rule of association shown in (4.1). In our formulation so far the cell association strategy is essentially based on UE data rate maximising optimisation approach, where the bias is a multiplicative factor to the service demand. However this biasing approach is not compliant with the LTE HetNets 3GPP standards. Therefore for compliance and practicality we resort to a sub-optimal solution whose underpinnings are derived from the formulation. The aim is to achieve realistic user-cell associations which provide sufficient load balancing at a cost of minimal association errors.

Considering the traffic offloading impact that CRE has on multi-tiered networks, we set all macro eNodeB biases to zero and the algorithm is predominantly executed in pico-eNodeBs. The role of the macro-eNB is minimal; it is to ensure synchronisation of load balancing related processes in pico-eNodeBs. The bias term from expression (4.18) is introduced in expression (4.1), then the association rule is rewritten as;

$$j^* = argmax_j\{RSRP_{ij} + \beta e^{-\lambda_j}\}. \quad (4.19)$$

A coefficient β is included for controlling the extent to which the bias could be aggressive. A high value of β, for instance, will result in an aggressive user-cell association, a very low value of β will result in a sluggish up-take of users during the user-cell association process. Aggressive user-cell association is not always desirable especially in the cell edge of range expanded regions of small base stations [54].

Since eNodeBs experience varying user traffic densities spatially and in time, the prices from each eNodeB will also vary. This means the value of the price will vary causing the bias to change accordingly. The biases are uniquely determined, that is they are not selected from a bias vector as seen in [4] but identified from the price- bias relationship. The price-bias relationship is illustrated in figure 4.2.

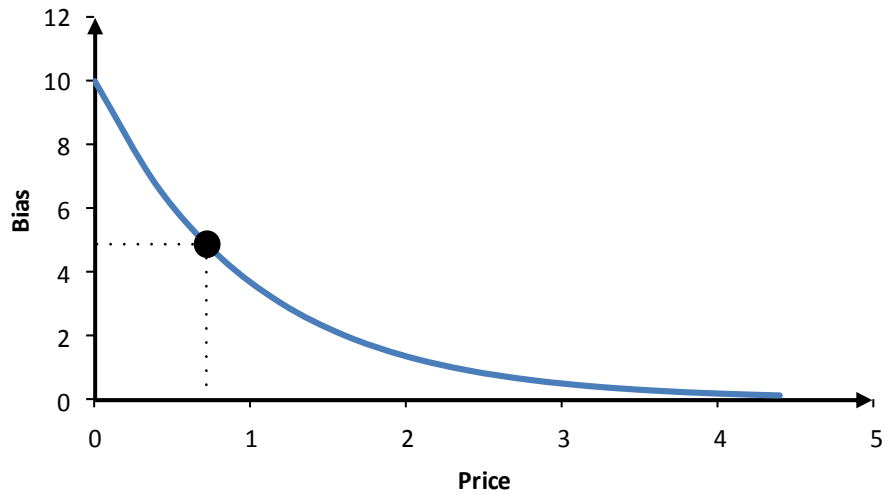

Figure 4.2. Illustration of price-bias relationship

As the price increases the price-bias decays exponentially as shown in figure 4.2. A point of interest is point P, which lies on the curve, this could be described as an optimality point signifying an optimal bias value in relation to a given price. It is interesting to observe the interplay between service demand, price and the bias; the relationship between service demand and the price are defined by expression 4.17, then the price and the bias are defined by expression 4.19. From the two expressions we understand that if the service demand is low, the price will be low. A low price will increase the bias and subsequently increase the service demand. In the same manner if the load is high the price will be high, causing the price to increase and then reducing the service demand.

The algorithm follows a two-pronged approach based on the formulation discussed; there are procedures on the user side and eNodeB side. This is a distributed approach, hence, the inherent benefits characteristic of distributed networks such as low traffic overheads and fast execution. On the eNodeB side, the eNodeB computes prices based on load. It has already been mentioned that the prices are updated by expression 4.17. Each eNodeB has a unique price that is used to calculate the bias. All eNodeBs perform the same procedure in a synchronised manner. On the user side, for the user to associate with an eNodeB, it measures pilot signals from eNodeBs that are possible targets and reports accordingly. In LTE the users measure the RSRP from the PDDCH channel and report to eNodeBs through the UPCCH channel. The signal reports are then added to

corresponding price variables to determine the eNodeB that is most suitable for association. We assume the prices are shared amongst neighbouring eNodeBs. The most suitable eNodeB for association, according to expression 4.19, is one which has the highest sum of RSRP and price. Most of the processes in the approach described are performed by the eNodeB, the role of the user is to assist by performing measurements.

Depending on the requirements of an operator, minimum and maximum bias constraints may need to be set according to the purpose for which a pico-eNodeB has been deployed. Some pico-eNodeBs, for instance, are deployed to enhance coverage gaps and others for increasing capacity in a high traffic density area. In our algorithm, we include a maximum bias, B_{max} , to ensure that bias values are not issued beyond this point. The load balancing algorithm is executed in iterations to search for a bias that maximises fairness. For each iteration, price and control information are exchanged amongst pico-eNodeBs and macro-eNodeBs. We use the Jain Fairness Index as our metric for fairness in the algorithm. We refer to it as the load balancing index *(I)*, defined as shown in expression 3.1. This index is computed from all the eNodeB loads of the network, as shown in expression 3.1. The iteration of the algorithm, therefore, is terminated only after the highest load balancing index has been found.

A lot of information exchange could generate unnecessary overheads in a practical system. Hence, to mitigate it we include a predetermined macro eNodeB load threshold, Th. The threshold ensures that range expansion is only effected when the macro eNodeB needs user offloading. In this approach the Macro-eNodeBs are responsible for controlling when to offload users to pico-eNodeBs within their coverage area and synchronise the load balancing process. A macro eNodeB will command pico eNodeBs to perform load balancing only when the load threshold is exceeded.

An outline of the steps for a load balancing cycle, when the algorithm is executed, is shown below.

1. Compute the load L_d^j, for each macro-eNodeB j .
2. If the load for macro-eNodeB j is above Th, send a "balance load" command to pico-eNodeBs within the coverage area to start load balancing. Else go to last step.
3. Compute the loads for pico-eNodeBs and their biases.
4. Re-associate UEs according to the obtained biases.

5. Compute the load balancing index*(I)*
6. While the previous *I* is less than the current *I*, go to step 3, else go to step 7.
7. Stop.

A complementary flow chart that captures the load balancing steps of the algorithm through a load balancing cycle is shown in figure 4.3(a). The shaded area of the flow chart is expanded in figure 4.3(b) to show more details.

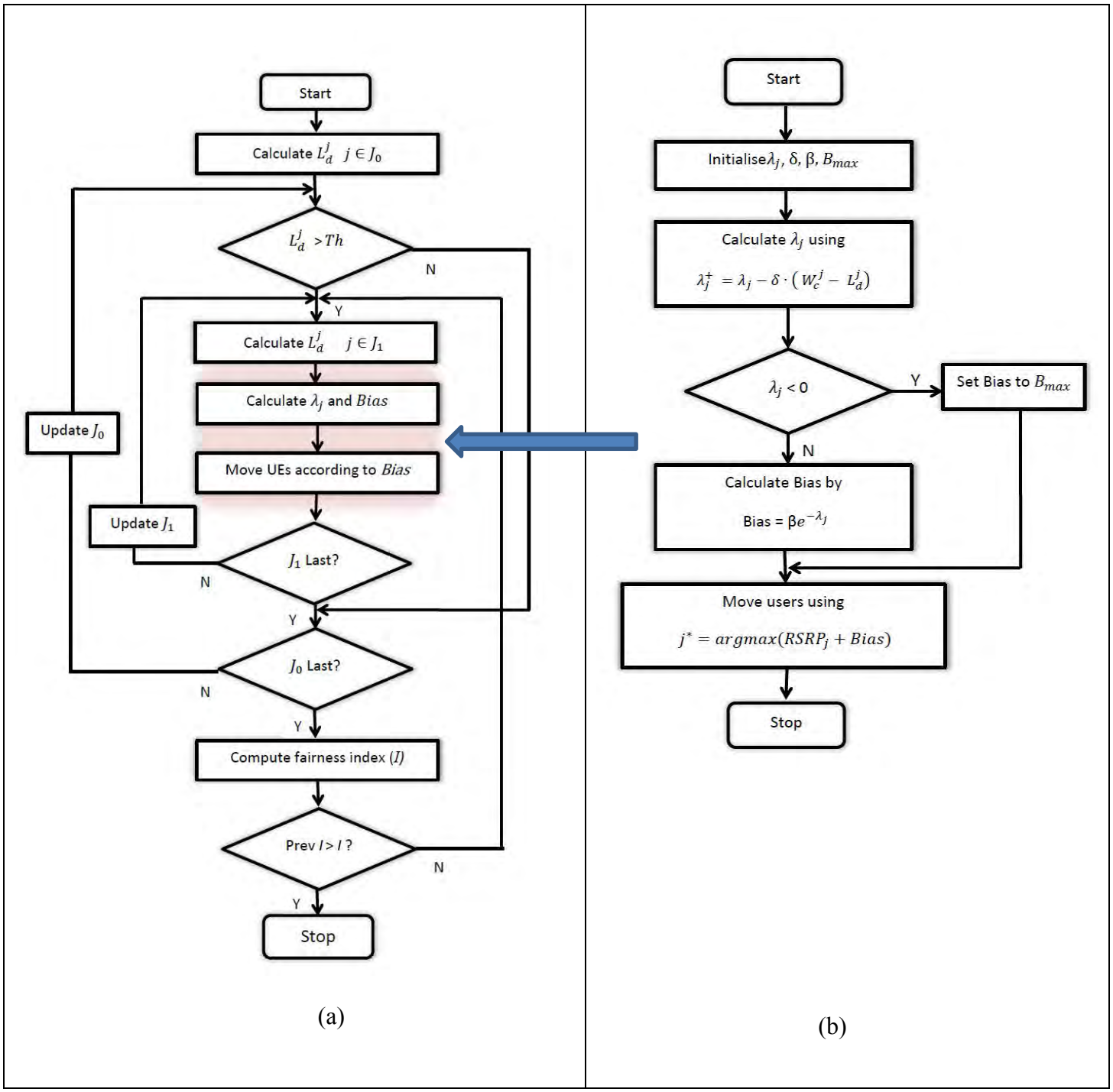

Figure 4.3. Flow charts of load balancing algorithm

4.5 Chapter Summary

In this chapter, we presented the design aspects of the proposed load balancing algorithm. This algorithm achieves load balancing among eNodeBs in a macro-pico HetNet by using CRE and logarithmic utility function. The system model on which the proposed algorithm is based considers full buffer downlink transmission only and assumes the users are static during transmission. First, the load balancing problem was formulated as a logarithmic utility optimisation problem, where users associate with eNodeBs that maximise their utility. Then based on the formulation and CRE, a sub-optimal load balancing solution was proposed.

Chapter 5
Implementation and Performance Evaluation

5.1 Introduction

This chapter presents a discussion on the implementation and performance evaluation of the proposed load-balancing algorithm. We begin by providing an overview of the tools used for the implementation, followed by the description of the network layout and system parameters. Finally, we discuss the performance evaluation of the algorithm and provide a summary of the results.

5.2 Implementation

To simulate and evaluate the effectiveness of the load-balancing algorithm, we use MATLAB version 8.3.0.532 (R2014a) and an LTE MATLAB toolbox developed by Hitachi Wireless Systems Research Lab. The toolbox generates macro and small cell network topologies, which includes channel modelling, network layout and user distributions [28]. The design of the toolbox follows 3GPP guidelines for modelling heterogeneous networks [63] and ITU guidelines for evaluating radio interface technologies [2]. It provides various propagation modelling options such as macrocell, microcell, ordinary HetNet configuration, referred to as HetNet conf1a, and hotspot configuration, referred to as HetNet conf4b. For purposes of our work, we built and incorporated the following functions to implement the algorithm:

- Bias function (Bias-fun) - it calculates a bias value for each eNodeB, using the load, which acts as a price.
- User association function (reattach-UE) - it performs handover of UEs according to equation 4.12. First, it calculates the price for each eNodeB and then obtains a bias by calling the bias function.
- Fairness index function (FI-index) – it calculates fairness index of the overall network after a load balancing cycle.

5.3 Network layout and configuration

We consider a network consisting of 7 hexagonal macrocell sites which have three sectors, where each sector has a macro eNodeB. Each of the sectors have four uniformly distributed pico eNodeBs, which are deployed as hotspots. All macro eNodeBs and pico eNodeBs have the same maximum transmission power, 46dB and 30dB, respectively. Furthermore, 60 users are randomly placed within each sector where 2/3 of the users are positioned around hotspots and the remaining are spread around macroeNodeB coverage areas. The main idea behind the described placement of users is to emulate a realistic distribution. An illustration of the network topology is shown in figure 5.1. In the figure, 3 macro eNodeBs are placed at the centre of hexagonal cells, pico eNodeBs are represented by circles (o) and users by crosses (x).

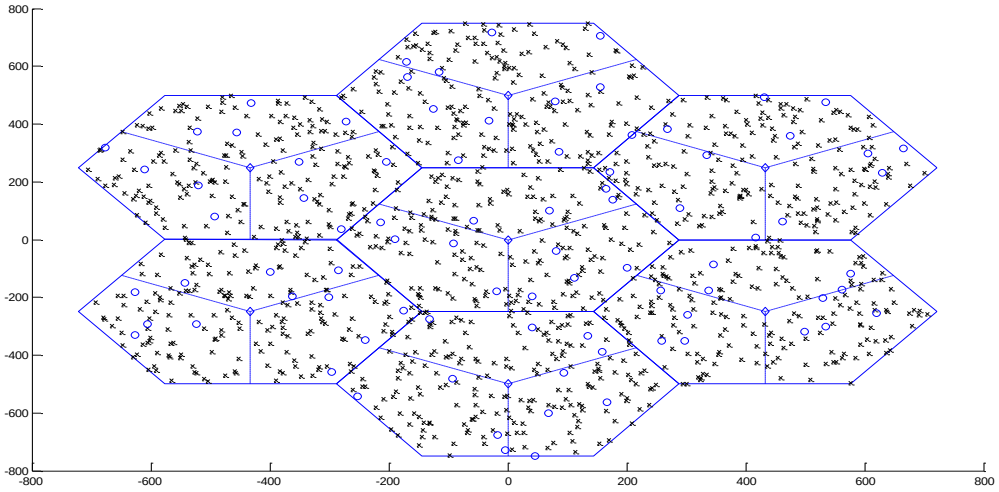

Figure 5.1. Illustration of network layout

The propagation environment, which includes shadowing effects and multipath fading between a user and an eNodeB, is handled by the toolbox. To cater for an urban macro-cell environment, where users are located outdoors, the ITU-R urban macro (UMa) channel model is selected. Then for pico-cell environment, where signal propagation is from outdoors to indoors, and there is an expectation of concentrated user density, the ITU-R urban micro (UMi) channel model is selected [2]. We assume that users are generating full buffer traffic when they are connected and they achieve Shannon capacity on the allocated resources. Table 5.1 presents a summary of system parameters for modelling the load balancing algorithm.

Table 5.1. System parameters

Attribute	Macrocell (value)	Picocell (value)
Centre frequency	2GHz	2GHz
Network layout	7 macrocell sites, 3 sectors	4 pico cells per sector
Path-loss model	ITU-R urban macro	ITU-R urban micro
Shadowing standard deviation	4 dB (LoS), 6 dB (nLoS)	3 dB(LoS), 4 dB(nLoS)
Bandwidth	20MHz	20MHz
Maximum transmit power	46 dBm	30 dBm
Number of UEs per sector	60	
Traffic model	Full buffer	Full buffer
Intersite distance	500 m	

5.4 Performance Evaluation

Performance of the algorithm is evaluated using metrics described in the previous chapter namely, fairness, throughput, and offloading percentage. As a baseline configuration, when evaluating the performance of the algorithm, we use the equation in (4.1) and fix the CRE bias to Bias = {0dB, 6dB, 12dB}. Then we compare the results obtained against the load balancing algorithm's results.

5.4.1 Distribution of users and load

To provide an information backdrop for analysing the load balancing algorithm's performance, we start by looking at the effect of CRE on traffic offloading; analysing the

distribution of users on the network for different CRE biases and the load balancing algorithm. Figure 5.2 presents the comparison of users in pico and macrocells for CRE Bias = {0dB, 6dB, 12dB} and the load balancing algorithm. When the CRE bias = 0dB, the user-cell associations are highly unbalanced with most users preferring to associate with macro eNodeBs, whilst fewer associate to pico eNodeBs. When the CRE = 6dB, the users are shifted to pico eNodeBs, which suggests that CRE reduces the macro eNodeB overloading problem in HetNets. When the CRE-Bias = 12dB, more users are shifted to pico eNodeBs. This occurs because a larger bias extends the CRE coverage area, thereby encouraging more users to prefer associating with pico eNodeBs. The load balancing obtains very close results to the 6dB bias settings.

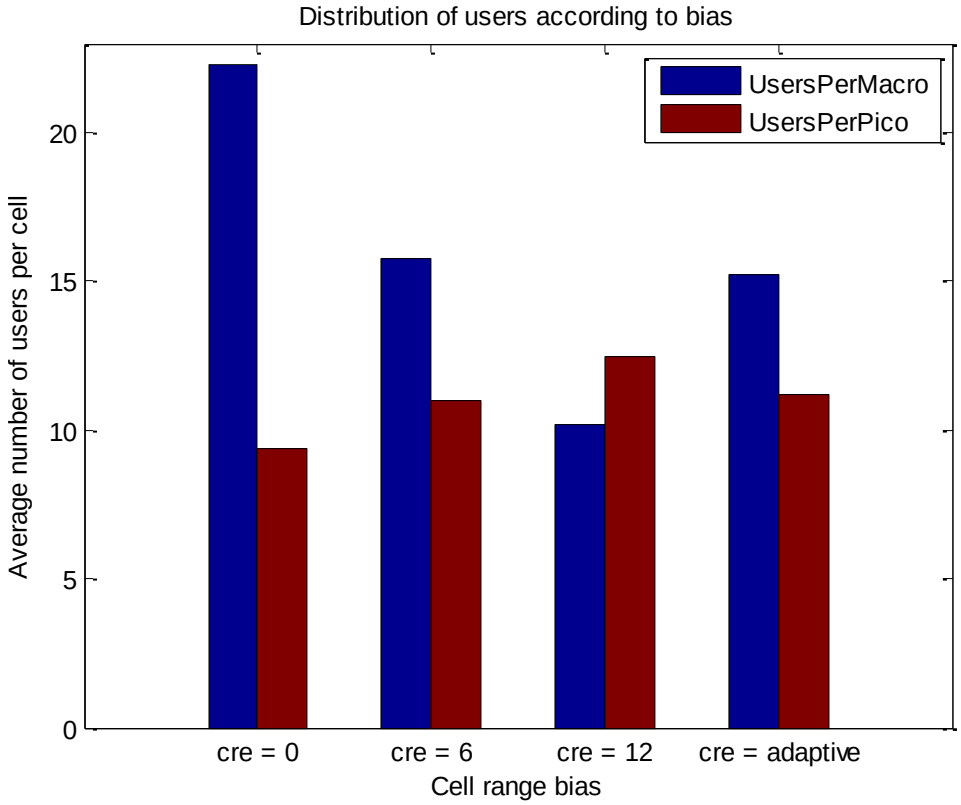

Figure 5.2. Comparison of User Distribution according to static CRE Bias and Load Balancing algorithm

Figure 5.3 compares distribution of users according to their association, either to overloaded cells or under loaded cells for different cell bias configurations. The figures demonstrate the shift of macro users to pico eNodeBs due to CRE. The pie chart on the top right

hand corner shows results for the baseline configuration, where 36% of the users are associated to overloaded macro eNodeBs and 61 % of users associated to underloaded pico eNodeBs. Despite the fact that each macro cell is capacitated with 4 pico eNodeBs most macro cells experience overloading when CRE biasing is not used. From the user distribution trend depicted by the figures, it is observed that, with increasing CRE bias, the number of users in heavily loaded pico eNodeBs increases. However, the load-balancing algorithm adapts to the change in load by assigning smaller CRE biases to pico eNodeBs that are heavily loaded, and larger CRE biases to lightly loaded pico eNodeBs, thereby moderating the shift of users to pico eNodeBs.

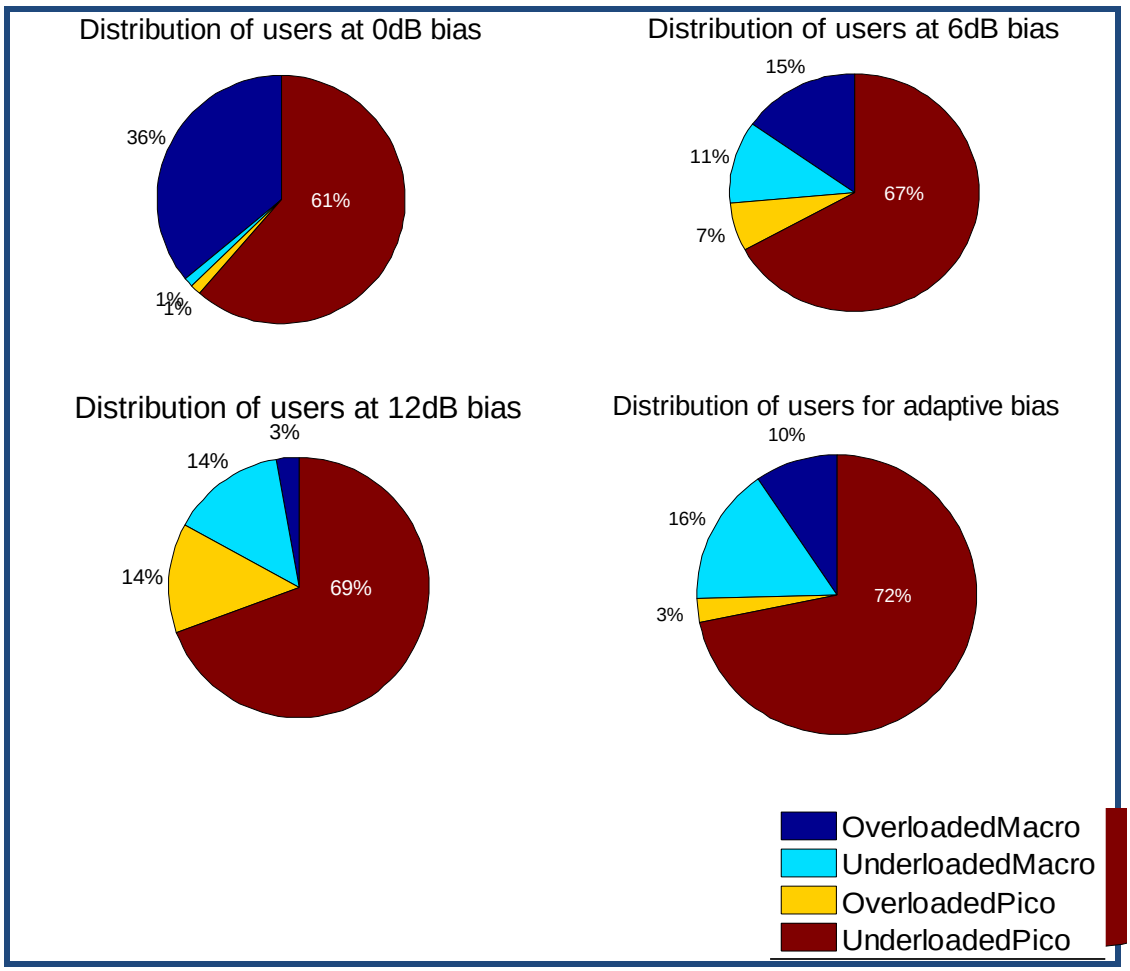

Figure 5.3. Comparison of user associations according to load status for different bias configurations.

A static CRE bias strategy, as opposed to an adaptive bias strategy, does not react to varying loads in the network. Consequently, there is a notable performance gap between the strategies in terms of fair user distribution.

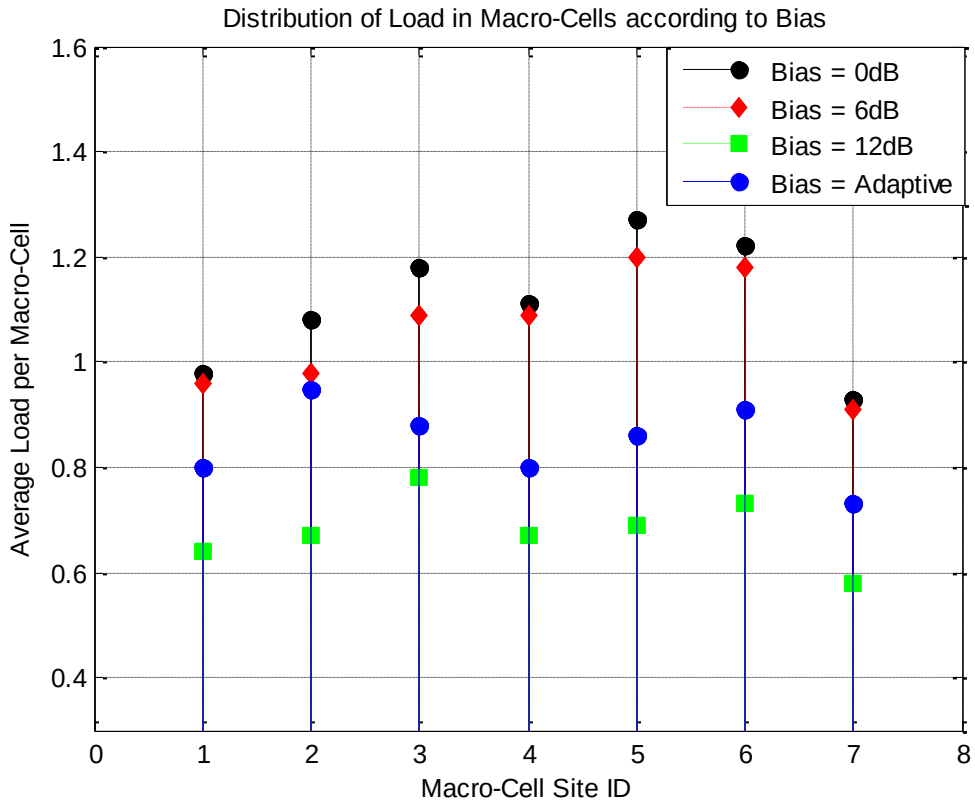

Figure 5.4 Distribution of load in macro-cells for static and adaptive CRE biases

Figures 5.4 and 5.5 show average loads per Macro-cell site of Macro eNodeBs and Pico eNodeBs for the 4 bias configurations. Each of the sites consists of 3 Macro eNodeBs and 12 Pico eNodeBs. The figures compare the results obtained by the load-balancing algorithm to those of the static CRE bias configurations. It is observed that, the load-balancing algorithm produces relatively uniform load among Pico cells when compared to the three static CRE bias configurations. This is caused by the algorithm's ability to adapt the bias according to the load. When 0dB bias configuration is used, some Macro eNodeBs are overloaded whilst most pico eNodeBs are moderately loaded. On the other extreme, when 12dB bias configuration is used, Macro eNodeB loads are reduced whilst loads in some Pico eNodeBs are significantly increased.

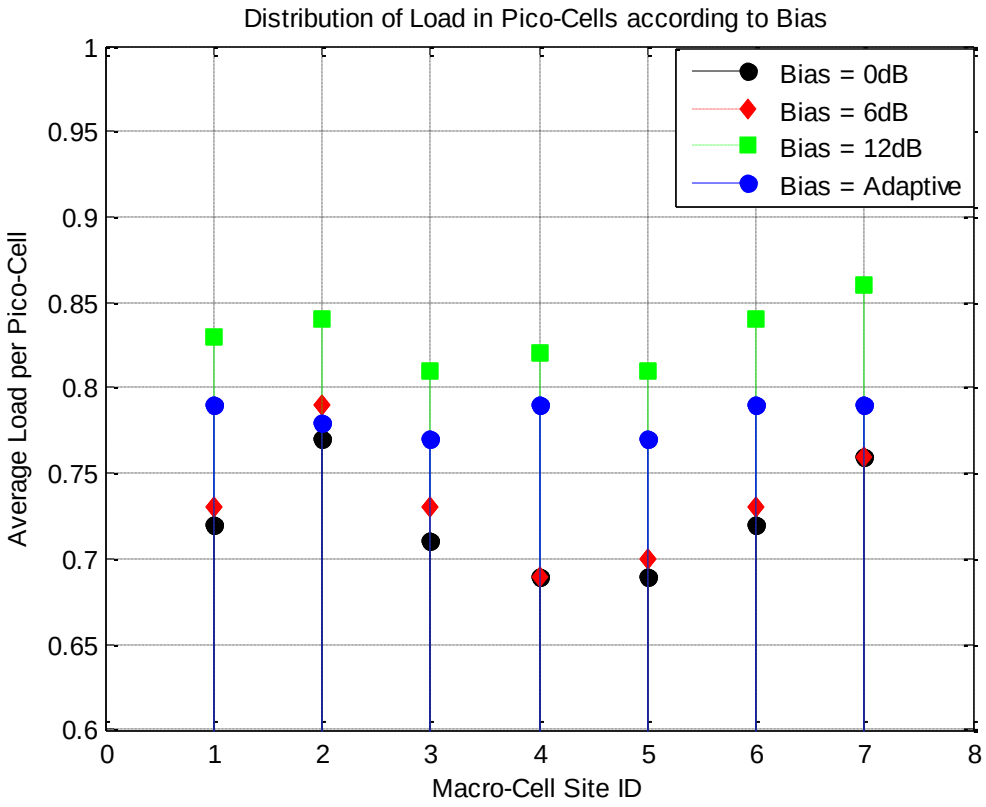

Figure 5.5 Distribution of load in pico-cells for static and adaptive CRE biases

5.4.2 Fairness

Figure 5.6 shows the fairness index versus iterations of the sub-gradient expression 4.12 compared to the fairness indexes of the three static CRE configurations. The static CRE configurations do not depend on the iteration of the sub-gradient, but are included for comparison. As expected, the load-balancing algorithm provides a better performance when compared to the static configuration solutions. The results show that within eight iterations the load-balancing index achieves a fairness index of 0.98, whilst the others achieve 0.79, 0.88 and 0.91 for 0dB, 6dB and 12dB bias configurations, respectively. However, the results also show that static CRE configuration achieves a lower degree of load balancing as compared to the algorithm, since it only seeks to maximise the offloading of macro cell users to pico cells. Further, we observe that the adjustment of the CRE bias has an impact on the fairness index; when the bias is increased, the

fairness index also increases. Therefore, the load balancing algorithm outperforms the static CRE configurations because of its ability to optimise user offloading by adapting the CRE bias of each Pico eNodeB according to its load.

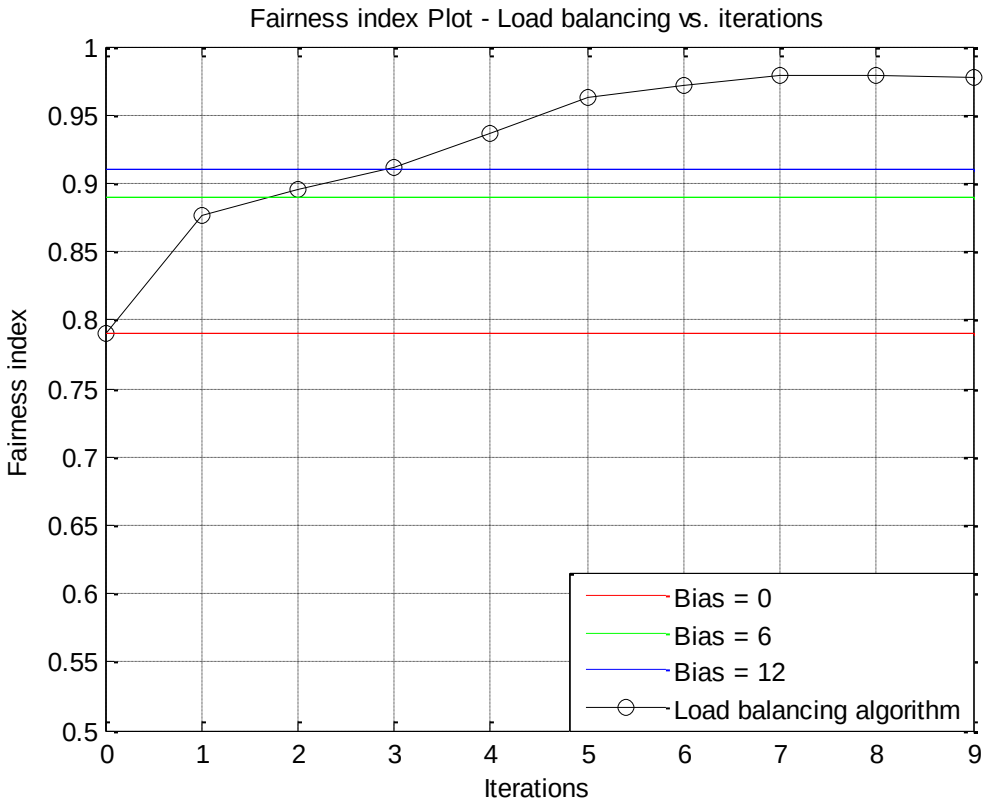

Figure 5.6. Fairness index plot of load balancing algorithm versus static CRE biasing

5.4.3 User Throughput

Figure 5.7 shows the cumulative distribution function (CDF) curve of user throughputs for the static CRE configurations and the load balancing algorithm. It is observed that the CDF curves of 6dB, 12dB CRE biasing, and the load-balancing algorithm improve significantly for low user throughputs. For instance, 5th percentile user throughput, which normally represents cell edge users or coverage throughput, improved by up to 4.5x, 8.9x and 5.5x, for 6dB,12dB biasing, and the load balancing algorithm, respectively. This improvement is a result of offloading users from macro eNodeBs to the pico eNodeBs as the CRE region increases. Further, since pico eNodeBs are nearer

to users, path-loss is low and their resources may be underutilised in comparison to distant macro eNodeBs. We note that, the 12dB bias configuration leads with high throughput gains due to aggressive offloading of cell edge users. The load-balancing algorithm on the other hand achieves moderate throughput gains; this is consistent with the objective of fair distribution of load amongst eNodeBs.

In the 95th percentile, which represents cell-centre region users, the user throughput decreases by 23%, 11%, and 19 % for 12dB, 6dB CRE bias configurations and the load-balancing algorithm, respectively. This is caused by the rescheduling of pico eNodeB resources used by cell centre users to CRE users. The 12dB configuration causes more resource scheduling due to its aggressive user offloading, as explained in section 4.4. The load-balancing algorithm's user throughput gain closely follows that of the 12dB biasing. However, it achieves it by a CRE adapting strategy.

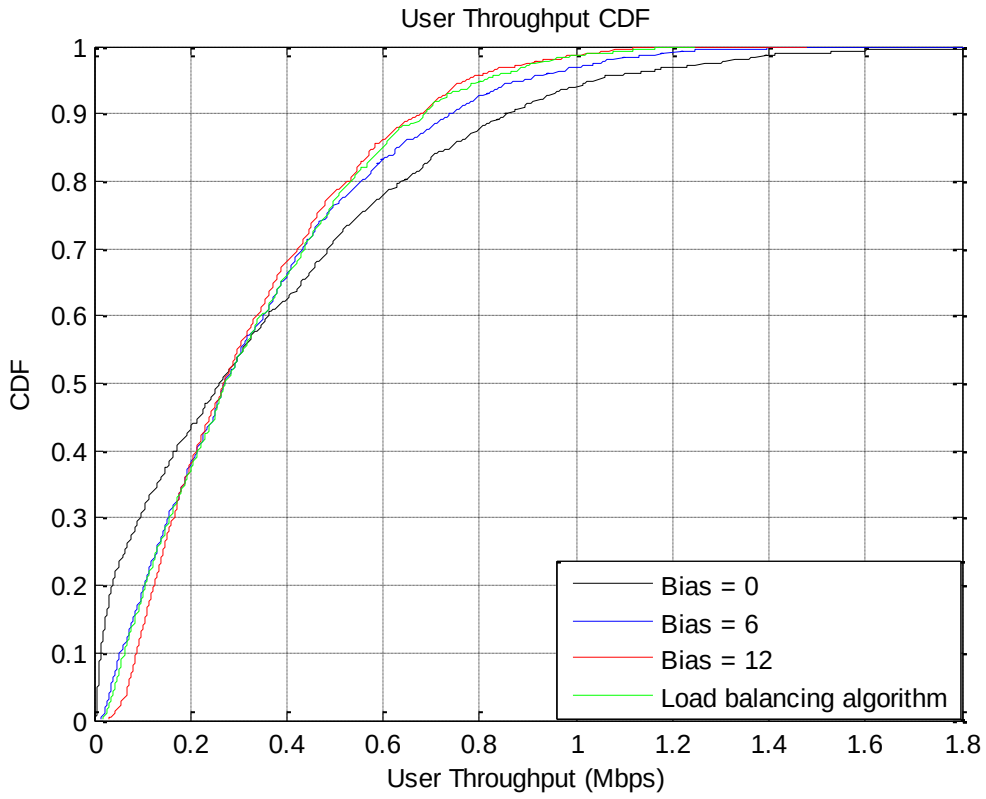

Figure 5.7. Cumulative distribution function plot of load balancing algorithm versus static CRE biasing

5.4.4 Impact of Bias Adaptation to Load

Figures 5.8 and 5.9 show the adaptation of the CRE bias and load versus the iterations of the load-balancing algorithm for pico eNodeB ID number 65, 78, and 69. These three pico eNodeBs are sampled from the 84 deployed in the network to demonstrate the performance of the algorithm in different load scenarios. Pico eNodeB ID number 65 represents an under loaded scenario, whilst 78 and 69 represent moderately loaded and overloaded scenarios, respectively.

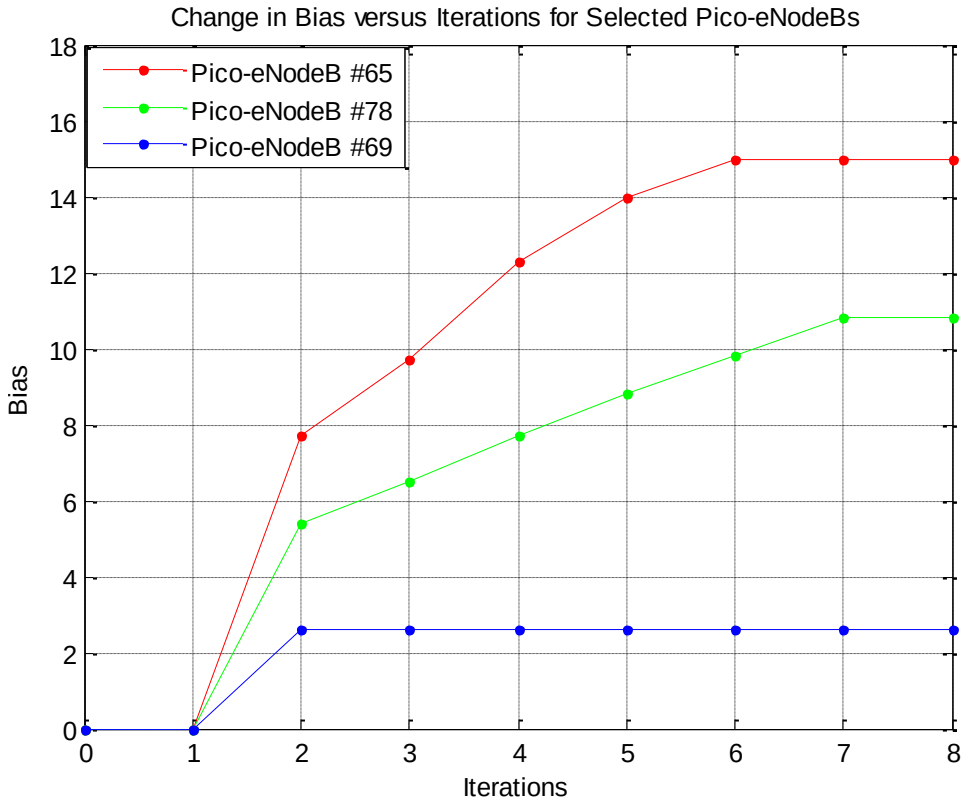

Figure 5.8. Comparison of change in bias against iterations for selected pico eNodeBs

We observe that, when a pico eNodeB is under loaded, the bias increases sharply to a high value until it reaches a steady state. On the other extreme, when a pico eNodeB is overloaded, the bias increases only slightly and stabilises at a much lower value as shown in figure 5.8. Since an underloaded pico eNodeB has underutilised resources, the algorithm increases the CRE bias relative to the resources utilised. For an over loaded pico eNodeB, the resources are strained and the pico eNodeB can hardly accommodate more users, hence the small increase of CRE bias.

However, we note that in the overloaded pico eNodeB the load slightly goes down despite the added bias as shown in figure 5.9. This is because of the redistribution of the load as other adjacent pico eNodeBs extend their CRE regions.

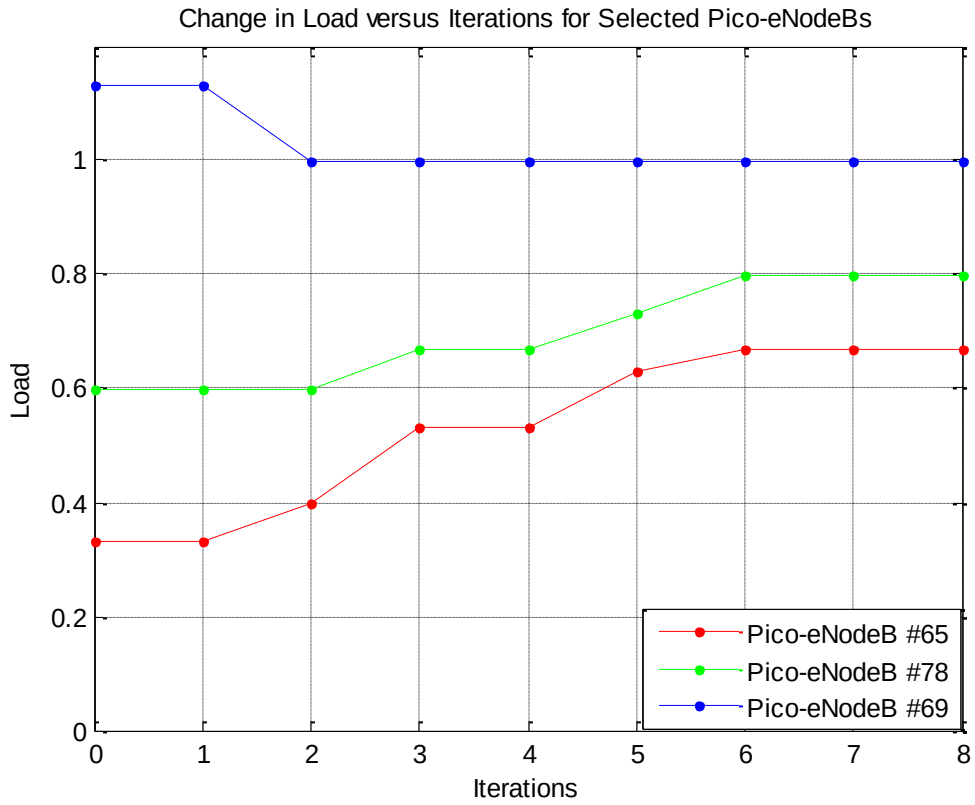

Figure 5.9. Comparison of change in load against iterations for selected pico eNodeBs

To provide a statistical perspective to the variation of the CRE bias among pico eNodeBs in the network, we use a CDF plot of CRE biases compared to static CRE bias configurations, as shown in figure 5.10. We observe that, the CRE biases across the network range between 2dB and 15dB. Further, 75% of pico eNodeBs use biases below 12dB, and 20%dB pico eNodeBs use biases below 6dB. This means the load balancing algorithm, in an attempt to avoid unnecessary overloading and underutilisation of network resources, assigns CRE biases that correspond to the load of each pico eNodeB. By so doing, it ensures that there is a fair distribution of load in the network.

Besides regulating load, by varying the CRE bias the load balancing algorithm also limits the penalty of low SINR values to CRE users in large CRE regions. A large CRE region is caused by using a large CRE bias. Some of the CRE users are very close to the path-loss boundary of the macro and pico cell, they consequently experience high macro eNodeB interference that leads to low SINR values. Users with low SINR values have challenges correctly decoding information from downlink control channels as demonstrated in [54].

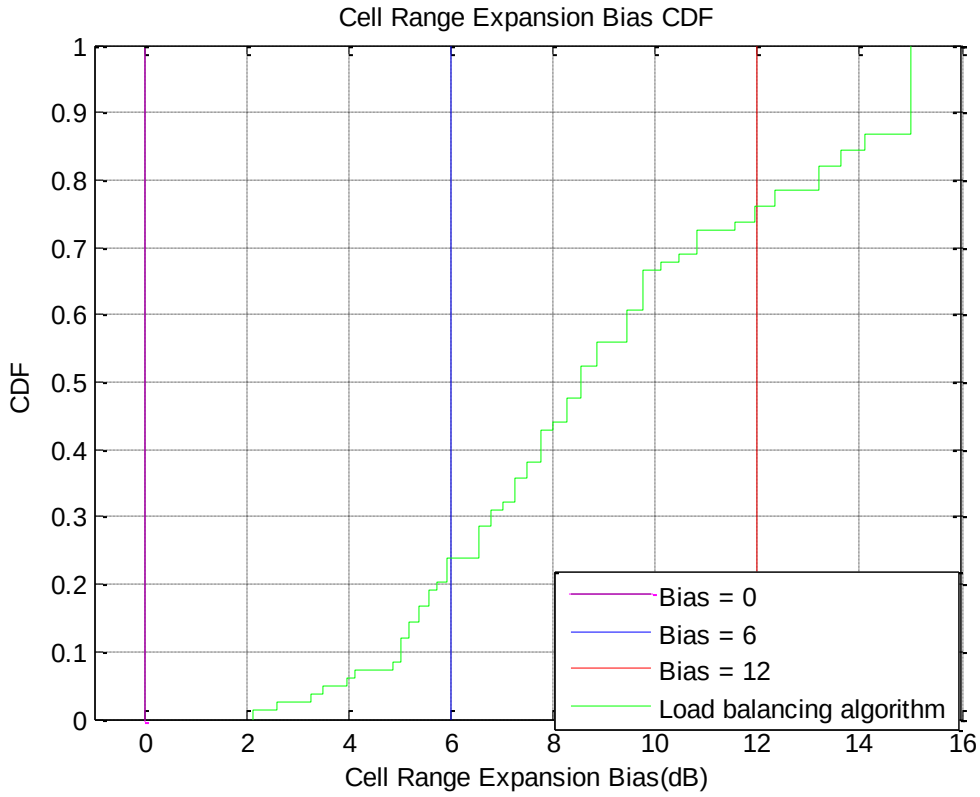

Figure 5.10. Cumulative distribution function plot of adaptive CRE bias compared to static CRE Bias

5.5 Chapter Summary

This chapter presented the performance evaluation of the adaptive CRE bias based load-balancing algorithm. We compared the performance of the load-balancing algorithm to three static CRE bias configurations used as baseline cases. The value of the adaptation of the CRE bias according to the load status of pico eNodeBs was demonstrated. The results showed that the load balancing algorithm achieves a fair distribution of load when compare with static CRE

configuration. It further achieves a better trade-off between cell edge and cell centre user throughputs. The variation of the bias limits the impact of low SINR values associated with large CRE biasing, which could cause problems in downlink channel control decoding, thereby leading to radio link failures. Therefore, we consider the proposed algorithm as a viable solution for a traffic steering oriented load balancing solution for LTE Advanced.

Chapter 6
Conclusion and Future Work

6.1 Introduction

This chapter presents a summary of the research work described in this dissertation. It highlights the achievements of the stated objectives. Then it suggests areas which can be explored further in the future.

6.2 Conclusion

The goal of this research was to investigate the use of adaptive CRE biasing as a load balancing strategy in a macro and pico LTE Advanced HetNets. The idea of using adaptive CRE biasing was prompted by the fact that pico eNodeBs experience varying user densities, which are often related to movement of users. The use of static CRE biasing is effective in traffic offloading, but could result in unfair sharing of the load, due to these varying user densities. However, by using an adaptive CRE biasing strategy, assigning cell specific CRE bias values to pico eNodeBs according to the load status, results in a fair distribution of load.

The strategy requires that CRE bias values be optimized to ensure that each pico eNodeB has a fair share of network load. To achieve this, the load balancing problem was first formulated as a logarithmic utility optimisation problem, where users associate to eNodeBs that maximize their utility. Since the problem is combinatorial, considering both user-cell association and pico eNodeB load, a Langragian multiplier was introduced for purposes of decomposing the problem into two sub-problems. The expression in the second sub-problem is not differentiable. Hence a sub-gradient technique was used, where the multiplier is updated by iterating the sub-gradient expression until an optimal solution is found.

Based on the sub-problems, we proposed an adaptive CRE bias load balancing algorithm that achieves sub-optimal solutions. We used the expression from the first sub-problem to determine the CRE bias and the expression from the second sub-problem to determine the Lagrangian multiplier. The Lagrangian multiplier was interpreted as a price of pico eNodeB load. It is on the basis of the price that the bias is calculated in the expression of the first sub-problem.

From the results obtained, we demonstrated that an adaptive CRE bias strategy achieves better load distribution in an unbalanced HetNet compared to a fixed CRE bias strategy. The load balancing algorithm achieved a high Jain's Fairness Index value compared to fixed CRE bias configurations. Further, the results demonstrate the efficiency of the optimization method used, as a balance between cell edge and cell centre region user throughputs was achieved. The 5^{th} percentile user throughput increased 5.5 times and the 95^{th} percentile user throughputs decreased by 19%. A fixed CRE bias configuration proved to be effective only in offloading users. However, the bias needs to be carefully selected, as it was observed that if it is too high it offloads users aggressively, thereby drastically reducing cell centre region user throughputs. On the other hand when it is too low it results in insufficient offloading resulting in lower cell edge user throughputs. Therefore to fully exploit the advantages of HetNets, the use of an adaptive CRE bias is essential.

6.3 Future work

Concerning directions for extending the work in the future, there are many areas worth exploring, but we will just highlight three, namely integrating CRE and eICIC techniques, Coordinated Multipoint transmission and reception (CoMP), mobility load balancing and base station power control.

The incorporation of eICIC techniques to load balancing strategy could reduce the severity of interference experienced by cell edge users when large CRE biases are assigned. The eICIC technique logically partitions the use of allocated spectrum by scheduling pico users and macro users to transmit and receive in different sub-frames in a co-channel deployment. When a macro user is using a particular sub-frame pico eNodeBs blank out that sub-frame so that pico users may not use it, thereby avoiding interference. To combat the problem of interference at the cell edges, another approach would be to use CoMP, particularly joint transmission and reception. CoMP is a technique that allows for co-operative transmission and reception between multiple base stations to and from users such that received signals are not adversely impacted by interference but rather boosted. The idea of employing CoMP techniques for interference mitigation would even be more viable, considering the imminent densification of HetNets.

In the model used for our work users were stationary; it would be interesting to adapt the model to investigate the effectiveness of the algorithm when users are in motion. This could

possibly lead to some aspects of mobility load balancing, where the tuning of handover parameters would have to be considered.

Finally the formulation of the load balancing algorithm could be used in base station power control for an energy efficiency load balancing strategy.

References

[1] "Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update, 2014-2019," *Cisco Public Information*. Cisco, San Jose, USA, pp. 1–42, 2015.

[2] "Guidelines for evaluation of radio interface technologies for IMTadvanced," *Report ITU-R M.215-1*, vol. 93, no. 3. ITU-R, Geneva, 2010. [Online]. Available: http://www.ncbi .nlm.nih.gov/pubmed/19923880. [Accessed: 06-Aug-2015].

[3] "Radiocommunication Study Groups -Background on IMT-Advanced," *Document IMT-ADV/1-E*. ITU-R, pp. 10–12, 2008. [Online]. Available: itu.int/ITU-R/go/rsg5-imt-advanced. [Accessed: 08-Aug-2015].

[4] I. Siomina and D. Yuan, "Load balancing in heterogeneous LTE: Range optimization via cell offset and load-coupling characterization," in *2012 IEEE International Conference on Communications (ICC)*, 2012, pp. 1357–1361.

[5] Qualcomm, "LTE Advanced : Heterogeneous Networks," *White Paper*. Qualcomm, Jan-2011. [Online]. Available: https://www.qualcomm.com/documents/lte-heterogeneous-networks. [Accessed: 12-Aug-2014].

[6] L. Saker, S. E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 664–672, Apr. 2012.

[7] S. E. Elayoubi and S. Antipolis, "A hybrid decision approach for the association problem in heterogeneous networks," in *INFOCOM, 2010 Proceedings IEEE*, 2010, pp. 1 – 5.

[8] S. Jia, W. Li, X. Zhang, Y. Liu, and X. Gu, "Advanced Load Balancing Based on Network Flow Approach in LTE-A Heterogeneous Network," *International Journal of Antennas and Propagation*, 2014. [Online]. Available: http://www.hindawi.com/journals/ijap/ 2014/93 4101/abs/. [Accessed: 19-Aug-2015].

[9] T. Kudo and T. Ohtsuki, "Cell range expansion using distributed Q -learning in heterogeneous networks," *EURASIP J. Wirel. Commun. Netw.*, vol. 1, pp. 1–10, Jun. 2013.

[10] "Aspects of Pico Node Range Extension," *Technical Document R1-103824*. Nokia Siemens Networks, Dresden, 2010. [Online]. Available: http://www.3gpp.org/ftp/tsg_ran/WG1_RL 1/TSGR1_61b/Docs/R1-103824.zip. [Accessed: 13-Aug-2015].

[11] G. Parkvall, Stefan, Dahlman, Erik, Jongren and L. eorge, Landstrom, Sara, Lindbom, "Heterogeneous network deployments in LTE," *Ericsson Review (English Edition)*, vol. 88, no. 2. Ericsson, pp. 34–38, 2011.

[12] D. López-Pérez, X. Chu, and I. Güvenç, "On the expanded region of picocells in heterogeneous networks," *IEEE J. Sel. Top. Signal Process.*, vol. 6, no. 3, pp. 281–294, 2012.

[13] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell Association and Interference Coordination in Heterogeneous LTE-A Cellular Networks," *IEEE J. Sel.*

Areas Commun., vol. 28, no. 9, pp. 1479–1489, 2010.

[14] S. Barbera and P. Michaelsen, "Improved mobility performance in LTE co-channel hetnets through speed differentiated enhancements," *IEEE GlobeCom Workshops*. IEEE, Anaheim, USA, pp. 426–430, Sep-2012.

[15] T. Buot, R. Nagaike, and S. Harmen, "Load balancing in WCDMA systems by adjusting pilot power," *5th Int. Symp. Wirel. Pers. Multimed. Commun.*, vol. 3, pp. 936–940, 2002.

[16] J. X. Qiu and J. W. Mark, "A dynamic load sharing algorithm through power control in cellularCDMA," *Ninth IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. (Cat. No.98TH8361)*, vol. 3, pp. 1280–1284, 1998.

[17] S. V Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity.pdf," *IEEE J. Sel. Areas Commun.*, vol. 13, no. I, pp. 1332–1340, 1995.

[18] Y. Bejerano, S. Han, L. E. Li, B. Laboratories, L. Technologies, M. Avenue, and M. Hill, "Fairness and Load Balancing in Wireless LANs Using Association Control," *Proceedings of the ACM MobiCom*. IEEE, Philadelphia, USA, pp. 315–329, Sep-2004.

[19] Q. Ye, B. Rong, Y. Chen, M. Al-shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *ArXiv*, vol. 2, pp. 1–24, Nov. 2012.

[20] H. Zhang, Y. Li, and Y. Li, "Traffic-based adaptive resource management for eICIC and CRE in heterogeneous networks," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*. IEEE, London, UK, pp. 117–121, 2013.

[21] Z. Ning, Q. Song, L. Guo, M. Dai, and M. Yue, "Dynamic Cell Range Expansion-based interference coordination scheme in next generation wireless networks," *Commun. China*, vol. 11, no. 5, pp. 98–104, May 2014.

[22] M. Simsek, M. Bennis, and I. Guvenc, "Mobility management in HetNets: a learning-based perspective," *EURASIP J. Wirel. Commun. Netw.*, vol. 2015, no. 1, p. 26, 2015.

[23] I. Guvenc, "Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination," *IEEE Commun. Lett.*, vol. 15, no. 10, pp. 1084–1087, Oct. 2011.

[24] K. H. Suleiman, H. A. Chan, and M. E. Dlodlo, "Load Balancing in the Call Admission Control of Heterogeneous Wireless Networks," in *International Conference on Wireless Communications and Mobile Computing*, 2006, pp. 245–250.

[25] R. Jain, D.-M. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," *DEC technical report TR301*, vol. cs.NI/9809, no. DEC-TR-301. pp. 1–38, 1984.

[26] Qualcomm, "LTE Mobility Enhancements," *White Paper*. Qualcomm, Feb-2010. [Online]. Available: https://www.qualcomm.com/media/.../lte-mobility-enhancements.pdf. [Accessed: 23-Aug-2015].

[27] J. Acharya, L. Gao, and S. Gaur, *Heterogeneous Networks in LTE-advanced*. John Wiley & Sons, 2014.

[28] J. Acharya, “Cellular Network Topology Toolbox.” Santa Clara, CA:Hitachi Wireless Systems Research Laboratory, 2014.

[29] R. Qureshi, “Ericsson Mobility Report: On the Pulse of the Networked Society,” *Technical Report*. Ericsson, Stockholm, Sweden, Jun-2015.

[30] ETSI, “Long Term Evolution,” *ETSI*, Nov-2013. [Online]. Available: http://www.etsi.org/technologies-clusters/technologies/mobile/long-term-evolution. [Accessed: 02-Oct-2015].

[31] J. S. Erik Dahlman, Stefan Parkvall, *4G: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.

[32] T. Nakamura, “Proposal for Candidate Radio Interface Technologies for IMT-Advanced Based on LTE Release 10 and Beyond,” *ITU-R WP 5D 3rd Workshop on IMT-Advanced*. 3GPP, Oct-2009.

[33] 3GPP, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) Radio Resource Control (RRC); Protocol Specification (Release 8),” *3GPP TS 36.331 V8.0.0*. 3GPP, Dec-2007.

[34] ETSI, “Evolved Universal Terrestrial Radio Access (E-UTRA); Long Term Evolution (LTE) physical layer,” *ETSI TS 136.201 V8.3.0*. ETSI, Apr-2009.

[35] B. Jang, “Orthogonal Frequency Division Modulation (OFDM).” National Chengchi University, Taiwan, 2008. [Online]. Available: http://www.cs.nccu.edu.tw/~jang/teaching/NextMobCom_files/Orthogonal Frequency Division Multiplexing.pdf. [Accessed: 02-Aug-2015].

[36] G. Ku, “Resource Allocation in LTE.” Drexel University, Philadelphia, USA, 2011. [Online]. Available: www.ece.drexel.edu/walsh/Gwanmo-Nov11-2.pdf. [Accessed: 10-Jun-2015].

[37] S. Palat and P. Godin, “The LTE Network Architecture,” Alcatel Lucent, 2009. [Online]. Available: www.cse.unt.edu/...2013...NETWORKS/LTE_Alcatel_White_Paper.pdf. [Accessed: 03-Jul-2015].

[38] 3GPP Technical Specification 36.331, “Evolved Universal Terrestrial Radio Access (E-UTRA); radio resource control (RRC); protocol specification.” 3GPP, 2012.

[39] R. Castro-Hernandez, Diego and Paranjape, “A Distributed Load Balancing Algorithm for LTE / LTE-A Heterogeneous Networks,” in *Wireless Communications and Networking Conference Workshops (WCNCW)*, 2015, pp. 380–385.

[40] 3GPP Technical Report and 36.839, “Evolved Universal Terrestrial Radio Access (EUTRA); mobility enhancements in heterogeneous networks.” 3GPP, 2012. [Online]. Available: http://www.etsi.org/deliver/etsi_ts/136300_136399/136331/10.07.00_60/ts_136331v100700p.pdf. [Accessed: 02-Jul-2015].

[41] S. Hämäläinen, H. Sanneck, and C. Sartori, *LTE self-organising networks (SON): network management automation for operational efficiency*. John Wiley & Sons, 2012.

[42] C. Askarian and H. Beigy, “A Survey for Load Balancing in Mobile WiMAX Networks,” *ACIJ*, vol. 3, no. 2, Mar. 2012. [Online]. Available: http://www.airccse.org/journal/acij/papers/0312acij13.pdf. [Accessed: 03-May-2015].

[43] S. V. Hanly, “Information capacity of radio networks,” University of Cambridge, 1993. [Online]. Available: http://people.eng.unimelb.edu.au/hanly/Thesis/title.pdf. [Accessed: 20-Jun-2015].

[44] K. A. Ali, H. S. Hassanein, A.-E. M. Taha, and H. T. Mouftah, “Directional Cell Breathing : A Module for Congestion Control and Load Balancing in WCDMA Networks,” in *Proceedings of the 2006 International Conference on Wireless Communications and Mobile Computing*, 2006, pp. 317–323.

[45] K. A. Ali, H. S. Hassanein, and H. T. Mouftah, “A Novel Dynamic Directional Cell Breathing Mechanism with Rate Adaptation for Congestion Control in WCDMA Networks,” *Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE*. IEEE, Las Vegas, USA, pp. 2927–2932, 2008.

[46] Y. Yan, L. Dittman, S.-W. Wong, and L. G. Kazovsky, “Integrated Control Platform with Load Balancing Algorithm in Hybrid Optical Wireless Networks,” *2009 3rd International Conference on New Technologies, Mobility and Security(NTMS)*. IEEE, Cairo, Egypt, 2009.

[47] A. Balachandran, P. Bahl, and G. M. Voelker, “Hot-Spot Congestion Relief and Service Guarantees in Public-Area Wireless Networks,” in *Mobile Computing Systems and Applications, 2002. Proceedings Fourth IEEE Workshop*, 2002, pp. 70 – 80.

[48] O. Brickley, S. Rea, and D. Pesch, “Load Balancing for QoS Enhancement in IEEE802 . 11e WLANs Using Cell Breathing Techniques,” *7th IFIP International Conference on Mobile and Wireless Communications Networks, Maroc*, 2005. [Online]. Available: http://137.73.11.49/MWCN2005/Paper/C200548.pdf. [Accessed: 25-Jul-2015].

[49] O. Brickley, S. Rea, and D. Pesch, “Load balancing for QoS optimisation in wireless LANs utilising advanced cell breathing techniques,” in *2005 IEEE 61st Vehicular Technology Conference*, 2005, vol. 3, pp. 2105–2109.

[50] J. Bigham and L. Du, “Cooperative negotiation in a multi-agent system for real-time load balancing of a mobile cellular network,” in *Proceedings of the second ACM international joint conference on Autonomous agents and multiagent systems - AAMAS ’03*, 2003, pp. 568–575.

[51] L. D. L. Du, J. Bigham, and L. Cuthbert, “Geographic load balancing for WCDMA mobile networks using a bubble oscillation algorithm,” in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, 2004, vol. 1, pp. 330–338.

[52] M. Vajapeyam, A. Damnjanovic, J. Montojo, T. Ji, Y. Wei, and D. Malladi, “Downlink FTP performance of heterogeneous networks for LTE-advanced,” *Communications Workshops (ICC), 2011 IEEE International Conference*. IEEE, Kyoto, Japan, pp. 1–5, Jun-2011.

[53] Y. Wang and K. I. Pedersen, “Performance Analysis of Enhanced Inter-Cell Interference Coordination in LTE-Advanced Heterogeneous Networks,” *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)*. IEEE, Yokohama, Japan, pp. 1–5, May-2012.

[54] Nokia Siemens, “Aspects of Pico Node Range Extension,” *3GPP TSG RAN WG1 #61bis Meeting: R1-103824*. 3GPP, Dresden, Germany, Jul-2010. [Online]. Available:

http://www.3gpp.org/DynaReport/TDocExMtg--R1-61b--28159.htm. [Accessed: 23-Jun-2015].

[55] K. Kikuchi and H. Otsuka, "Proposal of Adaptive Control CRE in Heterogeneous Networks," *Personal Indoor and Mobile Radio Communications (PIMRC), 2012 IEEE 23rd International Symposium*. IEEE, Sydney, Australia, pp. 910–914, Sep-2012.

[56] S. Sun, W. Liao, and W. Chen, "Traffic Offloading with Rate-Based Cell Range Expansion Offsets in Heterogeneous Networks," *Wireless Communications and Networking Conference (WCNC)*. IEEE, Istanbul, Turkey, pp. 2833–2838, Apr-2014.

[57] Y. Sun, T. Deng, Y. Fang, M. Wang, and Y. Wu, "A method for pico-specific upper bound CRE bias setting in HetNet," *2013 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, Shanghai, China, pp. 80–84, 2013.

[58] M. S. Ali, P. Coucheney, and M. Coupechoux, "Load Balancing in Heterogeneous Networks Based on Distributed Learning in Potential Games," 2015. [Online]. Available: http://perso.telecom-paristech.fr/~coupecho/publis/wiopt15.pdf. [Accessed: 20-Aug-2015].

[59] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed \alpha -Optimal User Association and Cell Load Balancing in Wireless Networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[60] K. Shen and W. Yu, "Distributed Pricing-Based User Association for Downlink Heterogeneous Cellular Networks," *Sel. Areas Commun. IEEE J.*, vol. 32, no. 6, pp. 1100– 1113, Jun. 2014.

[61] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing Client Association for Load Balancing and Fairness in Millimeter-Wave Wireless Networks," *Networking, IEEE/ACM Trans.*, vol. 23, no. 3, pp. 836–850, Jun. 2015.

[62] M. Pokhrel, S. and Panda, M. and Vu, H.L. and Mandjes, "TCP Performance over Wi-Fi: Joint Impact of Buffer and Channel Losses," *Mob. Comput. IEEE Trans.*, vol. PP, no. 99, pp. 1–1, Jul. 2015.

[63] 3GPP, "Further advancements for E-UTRA physical layer aspects (Release 9)," *3GPP TR 36814*, Mar-2010. [Online]. Available: http://www.3gpp.org/dynareport/36814.htm. [Accessed: 08-Sep-2015].

[64] X. He, B. Mugdim, H. Jyri, and J. Riku, "Self-organised and Bio-inspired Radio Resource Management for WiMAX," in *Radio Resource Management in WiMAX: From Theoretical Capacity to System Simulations*, Second., E. Vivier and G. Vivier, Eds. Hobkon, USA: John Wiley & Sons, 2013, pp. 245 – 267.