Energy Efficient Resource and Topology Management for Heterogeneous Cellular Networks

Abimbola Adeola Fisusi

PhD

University of York

Electronics

January 2016

Abstract

This thesis investigates how resource and topology management techniques can be applied to achieve energy efficiency while maintaining acceptable quality of service (QoS) in heterogeneous cellular networks comprising high power macrocells and dense deployment of low power small cells. Partially centralised resource and topology management algorithms involving the sharing of decision making responsibilities regarding resource utilization and activation or deactivation of small cells among macrocells, small cells and a central node are developed. Resource management techniques are proposed to enable mobile users to be served by resources of a few small cells. A topology management scheme is applied to switch off idle small cells and switch on sleeping cells in accordance with traffic load and QoS. Resource management techniques, when combined with the topology management technique, achieve significant energy efficiency.

A choice restriction technique that restricts users to resources from only a subset of suitable small cells is proposed to mitigate interference and improve QoS. A good balance between energy efficiency and QoS is achieved through this approach. Furthermore, energy saving under different generations of small cell base stations is investigated to provide insights to guide the design of energy saving strategies and the enhancement of existing ones. Also, an online, adaptive energy efficient joint resource and topology management technique is developed to correct deteriorating QoS conditions automatically by using a novel confidence level strategy to estimate QoS and regulate decision making epochs at the central node. Finally, a novel linear search scheme is applied together with database records of performance metrics to select appropriate resource and topology management policies for different traffic loads. This approach achieves better balance between QoS and energy efficiency than previous schemes proposed in the literature.

Table of Contents

List of Figures

List of Tables

Dedication

This thesis is dedicated to the memory of my late sister, Omobolanle Sanni (nee Fisusi). You were a rare gem and a blessing to everyone you came in contact with. I miss you. Continue to rest in peace.

Acknowledgement

First and foremost, I want to appreciate the Almighty God, my strength and song, for giving me the strength and grace to carry out this study. I am immensely grateful to Him for standing by me and preserving throughout the course of this study. I am sincerely grateful to the Petroleum Technology Development Fund (PTDF), Nigeria for granting me the full scholarship to carry out this study. I am also grateful to Obafemi Awolowo University, Ile-Ife, Nigeria for granting me study leave from my job to pursue the PhD degree.

I am immensely grateful to my first supervisor, Prof. David Grace, for his consistent support, guidance, encouragement, and motivation throughout the course of this study. I am very grateful to my second supervisor, Dr. Paul Mitchell, as well for his support and guidance throughout this study. This thesis would not have been possible without the excellent mentorship of my supervisors. Also, I want to thank the members of the Communication Research Group for their support and the friendly atmosphere I enjoyed while in the Group. I want to especially thank Abdulkarim Oloyede, Lawal Mohammed Bello and Salahedin Rehan for their support, advice and friendship during our times together on and off campus.

I am immensely grateful to my parents, Chief & Mrs. R.Z. Fisusi, for their unfailing love, support and encouragement during my study and always. I would not be here without your sacrifices over the years. I also want to thank my siblings – Dolapo Agboola, Funmilola, Olabisi, and Opeyemi - for their love, support and encouragement. I would also like to appreciate my brother-in-law, Mr. Tope Agboola for his constant care and support.

Also, I want to appreciate Dr. & Dr. (Mrs.) Sowole for their prayers, support and encouragement during the course of my PhD programme. I want to appreciate Prof. G.E. Erhabor for his encouragement and motivation always. I am very grateful to Eseoghene Umukoro for her support and motivation. I, also, want to appreciate Yewande Ogundeji for her support and advice during my PhD study. Finally, I want to appreciate my good friend, Kingsley Benjamin, for his support, advice, encouragement and consistent care from the moment I applied for the PhD funding until I completed the thesis.

Declaration

Some of the research presented in this thesis have led to publications or have been submitted for publication in conferences or journals. A list of publications is provided below.

This work has not been presented for an award at this or any other university. All the contributions presented in this work as original are as such to the best knowledge of the author. References and acknowledgements to other researchers have been included as appropriate.

Publications

Conference Papers

A. Fisusi, D. Grace, and P. Mitchell, "Energy efficient cluster-based resource allocation and topology management for beyond next generation mobile broadband networks," in ***IEEE International Conference on Communications Workshops (ICC)***, 2013, pp. 576-580.

A. Fisusi, D. Grace, and P. Mitchell, "Interference aware, energy efficient resource allocation for beyond next generation mobile networks," in ***IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC),*** 2013, pp. 2197-2202.

Journal Articles

A. Fisusi, D. Grace, and P. Mitchell, “Energy Saving in a 5G Separation Architecture under Different Power Model Assumptions”, ***submitted to Computer Communications***

Chapter 1. Introduction

1.1 Background and Motivation

The demand for mobile traffic is expected to increase exponentially in future wireless networks [1] with this estimated high traffic level predominantly data traffic rather than voice [2]. According to [3], mobile data traffic has already surpassed voice on a global scale and is expected to continue to increase rapidly. Furthermore, data-centric devices (like smart phones and tablets) with inbuilt cellular access are now common place and are responsible for the demand for increased capacity [4]. In addition, it is expected in the future that subscribers may consume mobile data of the order of several gigabytes (GB) in a month and the mobile industry is therefore preparing to support data rates of the order of tens of megabits per second (Mbps) and gigabytes data volumes [5]. Moreover, it is believed that future fifth generation (5G) networks should be able to support 1000 times the system capacity of the current fourth generation (4G) networks [6]. Hence, future wireless networks are expected to be high data rate and ultra-high capacity networks.

Apart from the requirement of support for high data rate and high capacity, future wireless networks are also expected to be energy efficient. This is due to increasing energy consumption and environmental impact of wireless networks in recent years. Already more than 4 million base stations (BSs) have been deployed to serve cellular traffic and on average each of these BSs consume 25MWh per year [7]. Also, the total energy consumption of cellular networks, wired networks and the internet is estimated to be over 3% of worldwide electricity consumption [8]. Furthermore, the cost of energy and electricity has been projected to increase significantly over the decade leading to 2020 and could result in high cost of operation for the wireless communication industry [9]. It is estimated that Information and Communications Technology (ICT) could lead to 15% reduction in CO_2 emission by 2020 in other sectors through the incorporation of the smart features of ICT [10]. However, ICT already accounts for about 2% of the global CO_2 emission and this trend is not sustainable [11].

In order to reduce energy consumption of cellular systems, close attention has to be paid to the radio access network. This is because the radio access network is estimated to account for over 80% of the energy consumption of cellular networks and the BSs are the dominant contributor [12]. Traditionally, cellular network deployments have been based on acquisition of costly cell sites and utilization of high power macrocell BSs (or simply macrocells). However, exponentially increasing demand for cellular data traffic requires massive deployment of BSs that will not be cost-efficient nor energy-efficient to achieve with macrocells [13].

Unlike macrocells, small cell BSs (or simply small cells) operate at lower power levels and serve smaller geographic areas. As a result, most users served by a small cell BS are closer to the serving BS than those under a macrocell coverage. Hence, most small cell users experience lower degradation (path loss) in transmitted signals with distance and can potentially achieve high data rates at lower power levels relative to the macrocell BS case. Therefore, heterogeneous networks consisting of high power macrocells overlaid with low power small cells have been proposed as a viable solution for the dual need for high capacity and low energy consumption [1, 14]. It is envisaged that the small cells will be deployed in locations not covered properly by macrocells and also in densely populated urban locations to provide high data rates not viable with the macrocell coverage at these locations [13].

WiFi offloading which involves serving cellular traffic by a WiFi access point rather than by the conventional macrocell is also considered as a suitable solution for meeting the increasing cellular traffic demand [15]. This is because WiFi access points can be widely deployed to serve small coverage areas like hotspots and homes at low cost [15] and they also have sufficient unlicensed spectrum available to achieve high throughput [16]. However, this thesis focuses on heterogeneous network of small cells and macrocells utilizing licensed cellular spectrum only.

In such a heterogeneous network, excessive overhead signalling can be experienced if small cells are densely deployed and each small cell transmits its own control and reference signals to support mobile users [17]. As up to 20% of the maximum transmission power may be used by BSs to transmit overhead information, overhead signalling in dense small cell deployments can constitute significant energy

consumption [18]. Hence, a new architecture, described as Hyper-cellular network [19] or Separation Architecture [20], has been proposed to reduce overhead signalling and also optimise resource utilization and energy efficiency in heterogeneous network [19].

The separation architecture concept proposes the separation of the data and control planes, with high power macrocells handling the control while low power small cells serve user data only [21, 22] as shown in Figure 1.1. The macrocells take on fully or partially the responsibility of transmitting essential overhead signals (control, synchronization and reference signals) needed by mobile users to select a cell, receive information about allocated resources and acquire other important information on the network. This allows significant reduction of overhead signal transmission at the small cells under the coverage of macrocells. This approach will enable coverage of the service area to be provided by the high power BSs while the capacity needs are met by the low power BSs. As a result, at low traffic load most of the low power BSs can be switched off without compromising on the coverage requirement of the network. In addition to handling coverage, the macro BSs can be configured to handle low-data rate user requests; while small cells handle high-data rate requests [20, 23]. This thesis investigates radio resource management and topology management strategies that can achieve energy efficiency in such a separation architecture based heterogeneous cellular network.

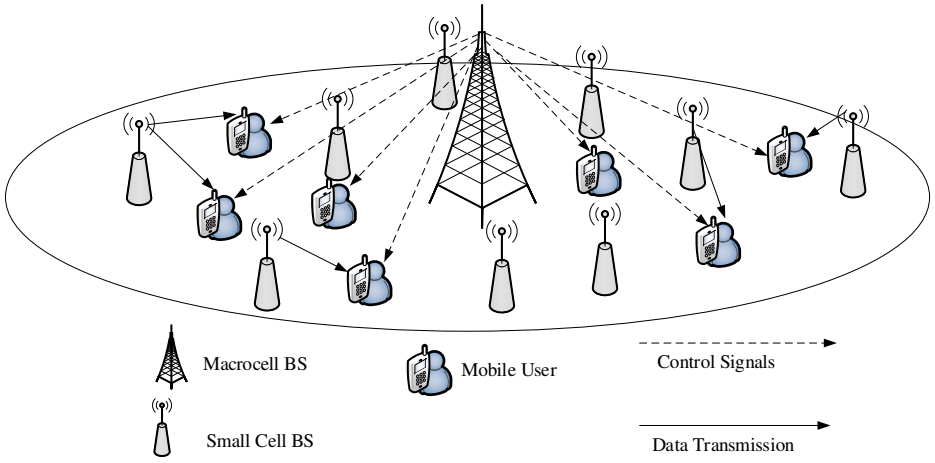


Figure 1.1 Separation Architecture Based Heterogeneous Network

1.2 Hypothesis

The hypothesis of this thesis is that partially centralised radio resource management and topology management can lead to significant energy saving in future cellular networks.

Heterogeneous networks based on the separation architecture concept are considered as viable solution for providing the ultra-high capacity needed in the future. These networks will consist of high power macrocells and low power small cells deployed within the coverage of the macrocells. In dense urban locations, small cells will be densely deployed to meet high capacity demands. In order to save energy, some of the small cells can be switched off while macrocells should remain on to maintain universal coverage.

Partially centralised Radio Resource Management (RRM) algorithms can concentrate, or cluster, mobile users on to few small cells and adapt the number of active small cells to the traffic load supported in a separation architecture based heterogeneous network. Idle small cells can be switched off using a partially centralised Topology Management (TM) algorithm. The partially centralised radio resource management and topology management algorithms involve the sharing of decision making responsibilities among small cells, macrocells and a central node. The radio resource management and topology management algorithms can be combined to provide significant energy saving.

A good balance between energy efficiency and quality of service (QoS) can be achieved by restricting mobile users to utilise resources from a subset of suitable small cells in their vicinity. Furthermore, energy saving can still be possible even when small cells are not switched off in future heterogeneous networks based on advanced small cells with very low power consumption when idle. In addition, an online, adaptive joint radio resource management and topology management scheme can dynamically adapt the policies governing radio resource management and topology management decisions to meet QoS targets and achieve energy efficiency. Also, a better balance between QoS and energy efficiency can be achieved by storing, analyzing offline and utilizing past performances of different policies for making more informed decisions.

1.3 Summary of Novel Contributions

Several novel techniques and methods have been developed in this thesis to achieve energy efficiency under the constraint of satisfactory quality of service (QoS) through the application of partially centralised radio resource and topology management. These novel contributions are presented in the following:

1.3.1 Clustering Capacity Rating for Energy Efficient Resource Management

A radio resource management concept that prioritises more centrally located small cells to serve user requests over those closer to the edge of the service area is proposed in Chapter 4 to reduce the number of active small cells. This concept is termed clustering capability rating (CCR) and different variants of the CCR are evaluated. The first one, Normalized Clustering Capability Rating (NCCR), considers only location and traffic load information of the small cells in deciding which ones to serve users. The second one, Controllable Quality Clustering Capability Rating (CQ-CCR), also considers the relative magnitude of the Signal to Interference plus Noise Ratios (SINRs) of candidate cells in addition to the location and traffic load information using a tunable QoS parameter. It is shown that the QoS parameter can be tuned to enable users to connect to small cells with higher SINR in the second case than the first case, resulting in better QoS across all traffic load levels but lower energy reduction gain at low traffic load.

The NCCR and CQ-CCR schemes when combined with the topology management scheme, applied in this work to switch idle BSs off, are shown to achieve significant energy efficiency. The NCCR and CQ-CCR schemes are the first set of schemes that utilise a clustering based rating to reduce the number of active small cells in a dense small cell network in the literature. Before the start of this thesis, previous energy efficient resource management were based on handing over users to highly loaded cells and switching of lowly loaded cells such as in [24]. The NCCR and CQ-CCR schemes do not require any handing over of users to achieve energy savings. The above contributions on CCR have been published in [25] and presented at the ***IEEE International Conference on Communications, 2013.***

1.3.2 BS Choice Restriction for Interference Mitigation and Energy Efficiency

In chapter 4, a BS choice restriction technique which restricts mobile users to utilise resources from only a high SINR subset of suitable small cell BSs is proposed to mitigate interference while still achieving significant energy efficiency. Joint interference mitigation and energy efficiency have been achieved in the literature by partitioning of available spectrum among different tiers in HetNets [26], interference power constraints to reduce cross-tier interference [27] and matrix of conflicts of interference to avoid allocating similar channels in different cells to address cross-tier and co-tier interference [28]. The choice restriction technique addresses co-tier inter-cell interference among small cells indirectly by placing limits on the choice of BSs that can serve users rather than directly controlling interference power or spectrum allocation. The choice restriction rules are defined by the central node in the network.

It is shown that by allowing users to be served by high order choices rather than lower order choices, higher inter-cell interference among small cells is introduced. This is why inter-cell interference can be mitigated by choice restrictions. An RRM scheme termed Interference Aware Clustering Capability Rating (IA-CCR) that first restricts the small cell choices of user before clustering them to more central choices (with the NCCR scheme) is developed. The IA-CCR scheme in combination with the topology management scheme is shown to achieve significant energy saving across different traffic load depending on the choice restriction applied. This contribution on BS choice restriction has been published in [29] and presented at the ***IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), 2013.***

1.3.3 Confidence Level Based Adaptive Joint Resource and Topology Management

A technique that estimates the QoS value from traffic statistics collected from macrocells and regulates the decision making epochs of the central node according to a predefined confidence level is proposed in Chapter 6. This is utilised in the development of an online, adaptive joint resource and topology management scheme in Chapter 6 to detect and correct QoS deterioration. RRM policies are defined in

terms of choice restrictions while TM policies are defined in terms of sleep state transition permission or prohibition. The central node is configured to make decisions to improve QoS without compromising energy efficiency completely through continuous adjustment of RRM and TM policies until QoS target is achieved.

It is shown that the adaptive scheme can successfully detect QoS deterioration and rectify it at different predefined confidence levels, when macrocells enforce new policy decisions from the central node. This is the first time such confidence level based collection of statistics and adaptation of policies have been utilised for QoS adaptation and energy efficiency studies in heterogeneous cellular networks to the best of my knowledge. In [30] traffic statistics were used to detect cell outages, however no measure of confidence was utilised in statistics collection and energy saving was not considered. The contributions on the confidence level based adaptive joint resource and topology management are being prepared for submission to ***Engineering Applications of Artificial Intelligence***.

1.3.4 Linear Search and Database Aided RRM and TM Policy Selection

A linear search method that searches for the best combination of RRM and TM policies that best balances QoS and energy efficiency using past performance metrics stored in a database is proposed in Chapter 7. Combinations of RRM and TM policies are mapped to different traffic load covering the range supported by the network using the linear search method. The mapping information are stored in the database and used to guide the selection of appropriate combination of RRM and TM policies for new traffic loads not previously mapped in the database.

It is shown that the combination of the linear scheme and database records for policy selection can achieve better QoS and energy efficiency balance over the range of different load supported by the network compared to the adaptive scheme developed in Chapter 6 and also previous schemes proposed in the literature. This is the first study that uses a linear search method and database records to select a combination of RRM and TM policy for balancing QoS and energy efficiency in heterogeneous cellular networks. The contributions on the Linear Search and Database Aided RRM

and TM Policy Selection are being prepared for submission to ***IEEE Transactions on Vehicular Technology.***

1.3.5 Impact of Different Power Model Assumptions on Energy Saving

The impact of different power model assumptions on energy saving is investigated for the separation architecture in Chapter 5 to understand the effect of future improvement in the idle and sleep state power consumption of small cells on energy efficiency. This is the first time such an investigation is carried out for a separation architecture. Previous energy efficiency studies on separation architecture have been based on single power models and most usually consider state-of-the-art BSs. However, since the separation architecture has been proposed for future cellular networks and BS component energy efficiency enhancements is ongoing, it is reasonable to consider more advanced power models aside the state-of-the-art types.

A framework, termed Low Power State Saving (LPSS), is developed and used to estimate energy saving resulting from operating small cells at lower power consumption states rather high power consumption states. Energy saving is evaluated across six different power models. Based on this framework, it is shown that significant energy saving is possible only if BSs are allowed to transition to the sleep state in existing small cells. However, significant energy saving can still be achieved without sleep state transition in future small cell BSs with low idle state consumption, even when energy saving is sought through idle state transition only. The contributions on the impact of power model assumptions on energy saving have been submitted to ***Computer Communications***.

1.3.6 Partially Centralised Topology Management Scheme

The distributed topology management scheme proposed in [31] is enhanced in this work with the introduction of small cell activation when QoS deteriorates and dynamic duration instead of fixed duration for the waiting period before small cell activation/deactivation. Furthermore, the partially centralised paradigm has been applied to the topology management strategy with the macrocells responsible for activating small cells when QoS deteriorates and the central node defining global topology management policies – permission or prohibition of sleep state transitions.

It is shown that the topology management scheme when combined with the developed radio resource management schemes can achieve significant energy efficiency without comprising the QoS in the network. Some aspects of the contribution on the topology management scheme have been published in [25] and presented at the ***IEEE International Conference on Communications, 2013.*** Other aspects of this contribution are being prepared for submission to ***Engineering Applications of Artificial Intelligence*** alongside the contributions on confidence level based adaptive joint resource and topology management (1.3.3 above).

1.3.7 Partially Centralised Paradigm for RRM and TM

A novel partially centralised paradigm for resource and topology management has been utilised in this thesis in order to enjoy the benefit of the global information access of a central node without its disadvantage of single point of failure. This paradigm involves the sharing of RRM and TM responsibilities across three hierarchical node types in the network: the central node, macrocell BSs and small cell BSs. The operation of the network is not fully dependent on the central node, rather the central node only defines and modifies RRM and TM policies which enhance the performance of the network. Even when the central node develops a fault, network operation can still continue based on the latest RRM and TM policies defined. Thus, RRM and TM are based on a partially centralised approach.

This is the first partially centralised RRM and TM paradigm based on policy definition at a central node that has been proposed for the separation architecture HetNet. A partially centralised paradigm was proposed in [32], however the central node was used to provide information for load estimation and cell deactivation at the macrocell rather than policies governing how RRM and TM decisions are made by the macrocell.

1.4 Outline

The rest of the thesis is organised as follows:

Chapter 2 presents a review of previous studies carried out on the energy efficiency of wireless networks. Energy models and energy metrics proposed for evaluating energy efficiency are discussed. Different approaches for achieving energy

efficiency, including base station component enhancements, network topology modification, and also resource and topology management schemes are presented.

Chapter 3 discusses the system modeling and performance evaluation techniques used for investigating the evaluated telecommunication network. Simulation software, traffic model, energy models and propagation models applied in the investigation are discussed. The performance metrics used for evaluating QoS and energy efficiency of the proposed schemes are also explained. In addition, methods utilised in validating the results of the system level evaluation carried out in subsequent chapters are also presented.

Chapter 4 introduces the partially centralised paradigm of radio resource management and topology management which involves sharing of decision making responsibilities among small cells, macrocells and a central node. Radio resource management schemes are proposed that enable mobile users to be clustered or concentrated on few small cells in order to reduce the number of active cells. Furthermore, radio resource management choice restrictions are introduced, usually defined at the central node but enforced by macrocells, to improve QoS through the restriction of mobile users to resources from only a subset of suitable small cells. Also, a topology management scheme is presented which switches small cells on and off in accordance to traffic load and QoS based on decisions of the different network nodes. Radio resource management and topology management schemes are combined to achieve significant energy saving.

Chapter 5 investigates the impact of different power model assumptions on the energy saving in separation architecture based heterogeneous networks. Unlike in Chapter 4, where a single power model is used for estimating power consumption, six power models are considered in this chapter. A framework referred to as Low Power State Saving (LPSS) is developed which is based on operating BSs in low power consumption state rather than high power consumption state in order to save energy. This framework is used to study energy saving when small cells are operated in idle state and sleep state rather than active state in the network for different power models. This investigation is expected to provide insights that can be useful in designing energy saving strategies and enhancing existing ones.

Chapter 6 exploits the insight gained in Chapter 4 regarding how choice restriction affects QoS and in Chapter 5 about the energy saving due to operating small cells in idle state. An online, adaptive joint radio resource management and topology management scheme is developed to enable the network to autonomously detect and correct deteriorating QoS conditions. This scheme requires the modification of choice restriction and the permission or prohibition of small cell transitions to sleep state by the central node to meet QoS targets. A novel confidence level based method is applied to regulate the time interval between adaptation decisions made by the central node. It is expected that when central node decisions are adopted locally by the small cells and macrocells, QoS targets can be satisfied while moderate energy saving can still be achieved at medium and high traffic loads.

Chapter 7 exploits extra information apart from the relationships among choice restriction, QoS, energy saving and sleep state configurations utilised in Chapter 6. A database is created and used for storing QoS and energy efficiency performance metrics of the network under different combinations of choice restrictions and sleep state configurations for different traffic loads. A linear search method is developed to select offline the combination of choice restriction and sleep state configuration that best balances QoS and energy efficiency for each traffic load. The mapping of choice restriction and sleep state configuration combinations to different traffic loads is stored in the database and utilised by the central node to guide future radio resource management and topology management decisions. This approach is expected to achieve improved balance between QoS and energy efficiency.

Chapter 8 provides a summary of the studies carried out in this thesis. Chapter 9 presents future work that can follow on from the research carried out in this thesis.

Chapter 2. Literature Review

2.1 Introduction

This chapter provides a review of the studies carried out in the literature, which are related to energy efficient wireless networks in general, and energy efficient resource and topology management in heterogeneous cellular networks in particular. The review provides the necessary background needed to understand the studies carried out in this thesis. The remaining section of this chapter is organised as follows. Firstly, the opportunities for energy saving are presented in section 2.2. Then, energy models are discussed in section 2.3. This is followed by the description of common energy efficiency metrics in section 2.4, while enhancements of base station components studies are presented in section 2.5. Studies on energy efficient network deployments are provided in section 2.6, while those on resource and topology management in heterogeneous cellular networks are presented in section 2.7. Finally, the chapter is concluded in section 2.8.

2.2 Opportunities for Energy Saving in Cellular Networks

Cellular networks usually consist of the core network and the access network [33]. The core network consists of switches, gateways and databases that connect the access network to the internet and the public switched networks. In the second generation (2G) and third generation (3G) cellular networks, the access network consists of cell sites and cell site controllers [34]. Cell sites are managed by base stations (BSs) while cell site controllers manage several BSs connected to them. A large cellular network can contain several cell site controllers interconnected to serve mobile users in the network. The state-of-the art fourth generation (4G) Long Term Evolution (LTE)/Long Term Evolution Advanced (LTE-A) networks have only eNodeBs (BSs in 4G LTE/LTE-A) in the access network with the cell site controller function integrated in the eNodeBs [35]. Usually, mobile users access the resources in the network through the BSs.

In terms of energy consumption, the access network is responsible for 80% of the energy consumption attributed to cellular networks, with the BSs mainly responsible [12]. Hence, a reduction of the energy consumption of these BSs by enhancing the

energy efficiency of their components will improve the energy efficiency of cellular networks. Furthermore, access network technologies and algorithms of cellular network are traditionally optimised for full load condition; however most times the traffic demand is low or medium [36]. Thus, dynamic adaptation of BSs and their resources to the instantaneous traffic demand, for example switching off BSs during periods of no load, can provide opportunity to save energy.

Also, traditionally BSs with high transmission power, termed macro BSs or macrocells, capable of serving users over a wide area (coverage area) have been utilised for building cellular networks. However, the ultra-high capacity requirement [4] and exponentially increasing data rates forecast [1] for future networks cannot be achieved in a cost-effective nor energy-efficient manner with conventional macro BSs [13]. This is because even higher transmission power will be required to meet the high data rate requirements. Moreover, macrocell sites are expensively acquired and the macro BSs themselves are very costly network systems. However, low cost, low power, small cell BSs with much smaller coverage area can provide high data rates due to much closer proximity of users to BSs and are considered as a viable solution [13] . Small cells can be integrated with the macrocells in future networks to provide the desired user demands while keeping network energy consumption low. It is important to note that small cells can be classified according to the size of their coverage area in an ascending order as femtocells, picocells and microcells [37].

Opportunities abound for energy saving in cellular networks, as described above this includes energy efficient design of BS components, BS and resource adaptation to real-time traffic demand, and small cell consideration in cellular networks. Hence, studies on enhancing the energy efficiencies of the constituent components of BSs have proposed. Also, different network deployments have been proposed to bring access base stations closer to the mobile users in order to reduce transmission power and thereby reduce energy consumption. Finally, different BS activation and deactivation, topology management, and resource management techniques have been proposed to adapt the access network to real time traffic and reduce energy consumption. These different approaches to energy efficiency are discussed in more detail in the following section. However, energy models proposed for quantifying the power consumption of BSs and energy efficiency metrics proposed for

characterizing energy efficiency in cellular network are first discussed. This is because they are the fundamental tools required for comparing different systems, algorithms and scenarios.

2.3 Energy Models

Energy models are required for the estimation of the energy consumption of wireless networks and unless the energy consumption is determined the energy efficiency of the network cannot be evaluated. Hence, energy models have been developed for cellular networks to compute the overall energy consumption over a given period. Although, overall energy consumption should include both embodied energy (the total energy required to produce goods or services [38]) and operational energy [39], the contribution of embodied energy have been largely ignored in most energy models in the literature. Furthermore, most models consider only the contribution of the base station in estimating energy consumption. This is probably due to the large share attributed to the base stations in the energy consumed in cellular networks.

Typically, a base station consists of several transceivers and each transceiver is connected to a transmit antenna element [40]. Each transceiver is made up of a power amplifier (PA), a small signal radio frequency (RF) transceiver section which comprises of a transmitter and a receiver for downlink and uplink communication respectively, a baseband (BB) engine for signal processing (such as filtering and modulation/demodulation), a DC-DC power supply, an active cooling system and an AC-DC unit (mains supply) which connects to the power grid [41]. The block diagram of a typical base station transceiver is shown in Figure 2.1. The total power consumption of a base station is dependent on the summation of these individual components' contributions.

In macrocell base stations, the PA is the major contributor to the overall power consumption of the base station. The PA consumes about 55-60% of the overall power consumption of the base station in macrocells at full load, however this figure reduces as the cell size reduces and falls to less than 30% in small cell base stations like picocells and femtocells [41]. The breakdown of the power consumption of the components of a macrocell base station is shown in Figure 2.2. As can be observed from the figure, the cooling system and signal processing section, which incorporates

the baseband engine, are also main contributors to the power consumption of the base station apart from the PA.

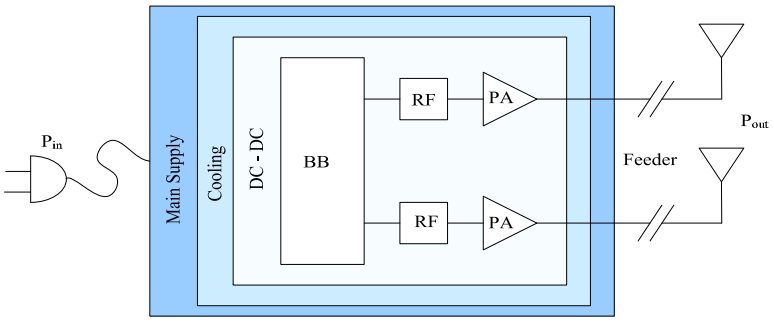


Figure 2.1 Block diagram of a base station transceiver (directly reproduced from [40])

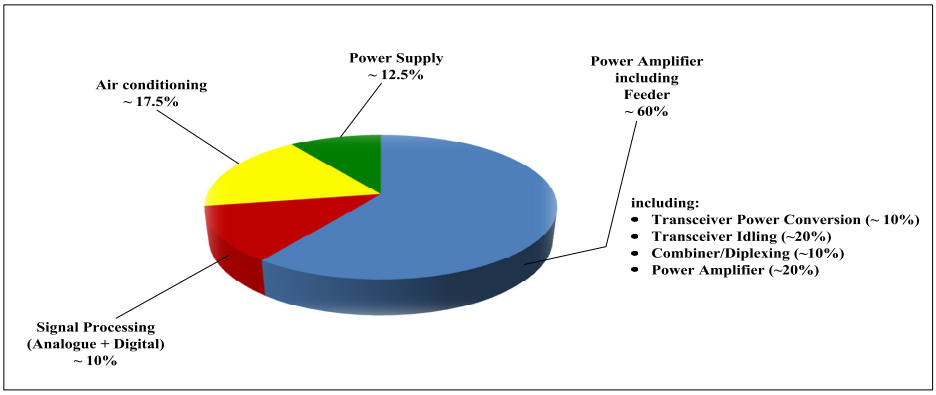


Figure 2.2 Energy Consumption of Macrocell Base Station (directly reproduced from [42])

Generally, a power or energy model comprises of a static part and a dynamic part. The static part is independent of the traffic load and transmission power. It incorporates the losses in the power supply, signal processing, and cooling systems. The dynamic part is dependent on the traffic load level supported by the base station.

A generic power model for all types of base stations (macro, micro, pico, and femto cells) is proposed in [41]. Different equations are provided for calculating power consumption at full load and variable load. Power consumption at full load has only static part while power consumption at variable load has both static and dynamic part. Although the equations are relevant to all types of base stations, the parameter values are different for different base station types. The parameters have been

obtained from measurements for LTE macro, micro, pico and femto base stations as of year 2010. The power consumption at full load is given by [41]:

$$P_{in} = N_{TRX}.\frac{P_{PA}+P_{RF}+P_{BB}}{(1-\sigma_{DC})(1-\sigma_{MS})(1-\sigma_{cool})} \quad (2.1)$$

N_{TRX} is the number of transceiver chain of the base station type; while $P_{PA}, P_{RF},$ and P_{BB} are the power consumed by the power amplifier, small signal RF transceiver and the baseband engine respectively. σ_{DC}, σ_{MS} and σ_{cool} are the losses incurred by the DC-DC power supply, mains supply and cooling system respectively. The power consumed at variable load is approximated by a linear function as follows [41]:

$$P_{in} = N_{TRX}.(P_o + \Delta p\,.P_{out}), \qquad 0 \le P_{out} \le P_{max} \quad (2.2)$$

P_o is the power consumed at the minimum possible output power and this represents the static part. This is also the power consumption of the BS when on no load. Δp is the slope of the load dependent power consumption while P_{out} is the RF output power. P_{max} is the maximum RF output power at maximum load.

The power consumption of the BS can be further reduced under the no load condition by deactivating some components such as the PA. The BS is often said to be in sleep state or sleep mode under this condition. P_{sleep} is the sleep mode power consumption of one base station transceiver chain. $P_{sleep} < P_o$ since it is assumed that some BS components can be deactivated in the sleep state to further reduce power consumption [43]. While in the deepest sleep the BS may consume close to zero watt of power and requiring 10-20 seconds to wake up, in a light sleep state with few components deactivated less power saving is possible, but the BS can wake up in about 30 µs [44]. Such a light sleep state is considered in this study which enables ABSs in sleep state to be activated to serve users almost instantly.

The linear model [41] described above is further enhanced in [40] by the introduction of the power consumption when a base station is in sleep mode given by:

$$P_{in} = N_{TRX}.\ P_{sleep}\,, \quad P_{out} = 0 \quad (2.3)$$

This linear model of [40] incorporating the functions of (2.2) and (2.3) has been widely used in the literature (e.g. in [45-50]).

The linear models of [40, 41] are further extended in [51] to include the energy consumption of the backhaul links (the links which connect the access network to the core network [52]). The authors noted that it is important to consider the backhaul consumption in the bid to determine the optimum network deployment i.e. deciding between the choice of low power BSs and higher power BSs for a given coverage area.

The energy consumption of the mobile network, P_{net}, is then modelled as follows taking into account the backhaul power consumption [51, 53]:

$$P_{net} = \sum_{i}^{N}\left(P_i + P_{bh}^{i}\right) \tag{2.4}$$

where N is the number of BSs, P_i is the power consumption of a BS i in the network which is equivalent to the P_{in} obtained by (2.2). P_{bh}^{i} is the power consumption of the backhaul associated with BS i. It comprises the uplink and downlink power required to transfer information over the backhaul link between the base station and aggregation switch(es) where traffic from different BSs are accumulated before being transferred to the core network. The power consumption of the aggregation switch(es) which is a function of the traffic transferred is also included in the backhaul power consumption [51, 53].

Furthermore, in [31] an energy model is proposed for the Beyond Next Generation (BuNGee) mobile broadband network architecture, which is used as the test network in the evaluation of the schemes proposed in this thesis. The model computes the energy consumption for two different types of base station: a small cell base station referred to as the Access Base Station (ABS) and a macro type base station referred to as the Hub Base Station (HBS). The ABSs are assumed to be low power nodes and do not require cooling systems like the HBSs. The energy consumption of the ABS, E_{ABS} is calculated as follows:

$$E_{ABS} = \sum_{j=1}^{n_{ABS}} \left(t_{ABS,sleep,j}\, P_{ABS,sleep,j} + t_{ABS,Rx,j} \frac{P_{ABS,Rx,j}}{\mu_{RF}} + t_{ABS,Tx,j} \frac{P_{ABS,Tx,j}}{\mu_{RF}} + t_{ABS,idle,j}\, P_{ABS,idle,j} + n_{wakeup,j}\, E_{wakeup} \right) \left(\frac{1}{1-\mu_c} \right) \quad (2.5)$$

n_{ABS} is the number of ABS, $P_{ABS,sleep,j}$ is the power consumed by the *jth* ABS when in sleep state. $P_{ABS,idle,j}$ is the power consumed when the *jth* ABS is on but not receiving or transmitting, instead it is waiting in line to serve users. This power is due to the non-radio-frequency components such as the battery backup and power supply. $P_{ABS,Rx,j}$ and $P_{ABS,Tx,j}$ are the power consumed in the receiving and transmitting states respectively by the *jth* ABS. $t_{ABS,sleep,j}$,$t_{ABS,idle,j}$, $t_{ABS,Rx,j}$ and $t_{ABS,Tx,j}$ are the total time the *jth* ABS spends in sleep, idle, receiving and transmitting states respectively. μ_{RF} is the efficiency of the power amplifier while μ_c represents the losses in the power supply and battery. n_{wakeup} is the number of times the ABS switches from the sleep state to the idle state. Finally, E_{wakeup} is the energy consumed in the process of waking up the ABS.

The energy consumption of the HBS, E_{HBS}, is calculated as follows:

$$E_{HBS} = \sum_{j=1}^{n_{HBS}} \left(t_{HBS,Rx,j} \frac{P_{HBS,Rx,j}}{\mu_{RF}} + t_{HBS,Tx,j} \frac{P_{HBS,Tx,j}}{\mu_{RF}} + t_{HBS,idle,j} P_{HBS,idle,j} \right) \left(\frac{1}{1-\mu_c-\mu_{cool}} \right) \quad (2.6)$$

n_{HBS} is the number of HBS, $P_{HBS,Rx,j}$, $P_{HBS,Tx,j}$ and $P_{HBS,idle,j}$ are the power consumed in the receiving, transmitting and idle state respectively by the *jth* HBS. $t_{HBS,idle,j}$ $t_{HBS,Rx,j}$ and $t_{HBS,Tx,j}$ are the total time the *jth* HBS spends in idle, receiving and transmitting states respectively. μ_{cool} is the losses due to the cooling system.

It is clear from the energy models presented above that energy consumption has been estimated using different assumptions. Energy models have not yet been standardised and as such different models are still being used in the literature. However, it is important to note that the choice of energy model greatly impacts the results of energy efficiency investigation. This is demonstrated in [54] where different power models have been applied and different power consumption (power

supply, P_s, required for different base station transmit power, P_{Tx}) results are obtained for different models in a single macrocell scenario as shown in Figure 2.3.

The SotA 2010 is based on the linear model proposed in [40, 41] while the other models have the same general form as in (2.2) and (2.3) but the BS parameters are different to reflect energy efficiency improvement envisaged for BSs in the future (Market 2014 [55] and Improved DTX [44]) or the ideal power consumption that BSs would be expected to have (Future Model [54]). The impact of these models on energy efficiency is investigated in this thesis (in Chapter 5) for the separation architecture based heterogeneous cellular network in a multiple cell scenario. This will provide a comprehensive understanding of the energy efficiency capabilities of the network and algorithms as BS technology evolves.

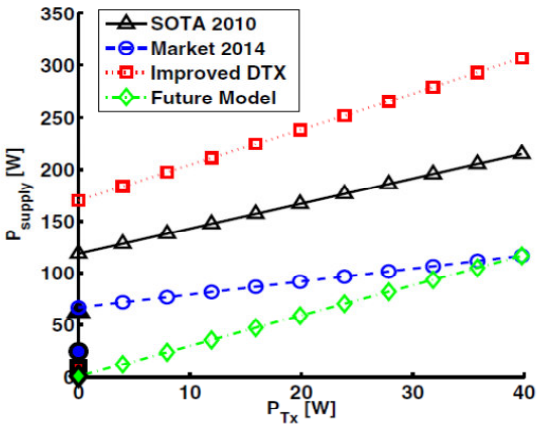


Figure 2.3 Power consumption estimation using different power models (directly reproduced from [54])

2.4 Energy Efficiency Metrics

The energy efficiency performance of a cellular network can be estimated once the energy consumption is determined from a suitable energy or power model. Several metrics have been proposed for estimation of the energy efficiency of cellular networks. The Energy Consumption Rating (ECR) Initiative proposed an energy efficiency metric based on the concept of normalizing network energy consumption to the highest throughput achieved in a test [56]. The metric is referred to as Energy Consumption Rating (ECR) and it is the ratio of the energy consumption to the maximum throughput. However, in the Green Radio Project (a research project

aimed at reducing power consumption of cellular networks) [57], the ECR is interpreted differently by consideration of successfully delivered bits rather than maximum throughput achieved. This is a more practical approach since it focuses on the information bits [57] and does include the overhead bits from the physical and link layers in the throughput like in the ECR initiative [58]. ECR has been defined as the energy consumed per information bit delivered, it is measured in joule per bit in the Green Radio project and expressed as follows [57]:

$$ECR = \frac{Energy\ Consumed\ in\ the\ Network}{Successfully\ transmitted\ information\ bit} \tag{2.7}$$

Another widely used energy efficiency metric, also proposed under the Green Radio project, is the Energy Consumption Gain (ECG). The ECG is a relative energy efficiency metric, in that it is calculated by the comparison of the energy consumption of a test scenario to a baseline scenario [57]. The ECG for the test scenario can be computed as a ratio of the ECR of the baseline scenario to the ECR of the test scenario as follows [57]:

$$ECG = \frac{ECR_{baseline}}{ECR_{test}} \tag{2.8}$$

Hence, a high value of ECG indicates that the test scenario is more energy efficient than the baseline scenario.

Another metric based on relative performance of two schemes used in the literature is the Energy Reduction Gain (ERG) [59-61]. The ERG is often expressed as a percentage and it can be obtained as follows:

$$ERG = \frac{ECR_{baseline} - ECR_{test}}{ECR_{baseline}} \times 100\% \tag{2.9}$$

$$\text{Hence, } ERG = \left(1 - \frac{1}{ECG}\right) \times 100\% \tag{2.10}$$

It can be observed from (2.9) that a positive value of ERG indicates energy efficiency under the test scenario relative to the baseline, while a negative value indicates the test scenario is less energy efficient than the baseline.

The bits/joule, φ, is another energy efficiency metric, which has been commonly used for single links but has been extended to the evaluation of whole networks [53]. It is evaluated as the ratio of the total capacity of the network, C_{net}, in bits/s to the total power consumption of the network, P_{net}, in Watts and thus expressed as follows [53]:

$$\varphi = \frac{C_{net}}{P_{net}} \tag{2.11}$$

The Area Power Consumption is another energy efficiency metric that is being used in the literature. The area power consumption, P_A, proposed in [12] is the ratio of total power consumed in a network, P_{net}, to coverage area of the network, A_{net}; it has a unit of $Watt/km^2$ and given as follows [12, 53]:

$$P_A = \frac{P_{net}}{A_{net}} \tag{2.12}$$

This energy efficiency metric is particularly relevant when comparing the energy efficiency of different network designs, especially different heterogeneous networks, where the densities and coverage areas of different BS types are not the same for the various designs [12, 62].

The ERG has been used extensively in this thesis because it allows the energy saving of the proposed schemes to be easily compared with a baseline scheme since the improvement in performance is simply expressed as a percentage relative to the baseline. This is not as clear and as straight forward with the other metrics discussed here.

2.5 Enhancements of Base Station Components

An understanding of the energy consumption of different components of the base station obtained from the energy models provides the motivation for improving the performance of energy hungry components such as the power amplifiers and the cooling system. Improvement in energy efficiency of these components will directly impact on the energy efficiency of the base stations and also the overall network energy consumption.

The power amplifier is the highest energy consuming component of a macro BS (approximately 60% as shown in Figure 2.2 under section 2.3), thus an energy efficient PA is desirable. Usually, the most energy efficient operating point of the PA is close to its maximum output power which is near the saturation [43]. However, for signals with nonconstant envelope, such as an Orthogonal Frequency Division Multiplex (OFDM) signal, linear amplification is necessary to avoid distortion and associated adjacent channel interference [63]. Therefore, PAs have to be operated in a more linear region below saturation [64] in state-of-the-art OFDM based cellular systems (e.g. 4G LTE/LTE-A). In addition, for a typical PA, high linearity is achieved in the low power region, while high energy efficiency is achieved in the high power region [65]. Hence, there is a fundamental tradeoff between linearity and efficiency [66].

Different approaches to improve energy efficiency without compromising the linearity requirement is investigated in the literature. The Doherty technique which involves a combination of two different types of PA (carrier and peaking PAs) can achieve high efficiency and linearity over a wide range of frequencies and power levels and is commonly used in start-of-the-art cellular systems [65]. In [67] a Digital Doherty PA architecture which digitally controls the adaptive power distribution of input power between the carrier and peaking PAs is proposed. It is shown to achieve improved efficiency over an existing analog Doherty PA.

Furthermore, in [14] a PA switching/selection technique comprising of different PAs with different output power levels and efficiencies is proposed for improved energy efficiency in place of a conventional single PA with different power levels. Improvement in energy efficiency is achieved by the selection of the most efficient PA that can support the target transmission rate with the least power consumption at a particular instance.

Also, envelope elimination and restoration (EER) technique, which involves separate amplification of the phase and envelope of the input signal and eventual combination, can facilitate high PA efficiency and linearity over a wide range of frequencies and power levels [65]. An amplifier with high efficiency and linearity

based on the EER technique, comprising of a nonlinear PA for phase modulation and a wide bandwidth envelope amplifier, is proposed in [68].

Efforts are also made to reduce the energy consumption of other energy hungry components of the BS. Modern base stations (such as in [69]) are designed to be naturally cooled by air, thus eliminating the energy consumption associated with the cooling system in traditional macro base stations [70]. In addition, as the size of transistors shrink it is expected that the power consumption of the digital integrated circuits will reduce and lead to improved power efficiencies of baseband signal processing circuits [9]. Also, advanced antenna technologies such as the multiple inputs multiple outputs (MIMO) techniques require lower transmission energy than conventional single antenna systems for the same bit error rate [71].

2.6 Energy Efficient Network Deployments

At the network level, in order to address the problem of high power transmission associated with macrocell BSs, the relationship between path loss and distance has been exploited. Path loss is an attenuation of the transmitted power [72]. It is directly proportional to the distance between two communicating entities (such as a BS and a mobile user) as follows: $PL \propto d^{\alpha}$; where PL is the path loss, d is the distance and α is the path loss exponent [39]. Hence, when the base station is closer to the mobile users the path loss can be reduced and the received signal strength can be improved leading to higher data rates, or alternatively by keeping the signal strength constant the transmission power required can be reduced.

This fundamental relationship has been applied in recently proposed network deployments which involve communication over shorter distances (on average) than associated with macrocell only deployment strategies. The application of this relationship in the literature includes studies investigating the optimal energy efficient cell size for densifying an area, the energy efficiency potentials of heterogeneous networks incorporating a mixture of macrocell base stations and smaller cell base stations. Also, the energy efficiency evaluations of full scale small cell deployments for urban hotspots have been investigated. These studies are discussed in detail in the following sections.

2.6.1 Optimum Cell Size for Energy Efficiency

The effect of cell size reduction on energy efficiency is investigated in [39] by evaluating the energy consumption ratio (ECR) and the energy consumption gain (ECG) for a fixed user density in a fixed service area. Different cell sizes, representing large, medium-sized and small cells, are evaluated. The total transmission power of the cells in the service area is kept constant regardless of the size of cells deployed. It is shown that the ECR per cell site decreases with reduction in the cell size, while the radio access network (RAN) capacity in bits/second increases with reduction in cell sizes. A sleep mode algorithm is applied to switch off cells without active users and trade off capacity for energy consumption. The ECG is evaluated with the largest cell size as reference. Under this sleep regime, the cell ECR and RAN energy consumption both decreases with reduction in cell size while the ECG increases with reduction in cell size without degrading the RAN capacity. Increasing ECG or decreasing ECR is equivalent to increasing energy efficiency. Thus the energy efficiency improves with decreasing cell size.

Similarly, in [73] the impact of cell size on energy efficiency is evaluated with macrocells, microcells, picocells and femtocells with 1km, 500m, 100m and 10m coverage respectively. The transmit power requirement of cells is defined as the power required to achieve a given signal to noise ratio (SNR) at the cell boundary under a fixed regime or to achieve the same received power everywhere in the cell under a power control regime. Closed form expressions under fixed and power control regimes are derived for the ratio of the transmit power requirement of smaller cells to the macrocell case for the same coverage area of 1km. Under the assumption of full traffic load, energy consumption is shown to decrease with the cell size in both the fixed power and variable power scenarios. Also, system capacity is shown to increase with decreasing cell size. Finally, the system capacity per unit energy consumption is shown to increase with decreasing cell size as expected after the findings of the initial cases.

It is observed that only the transmit power consumption is considered while the static consumption is not included in [39] and [73]. Different conclusions may arise when the static consumption is considered as observed in [74]. In [74], the relationships

between energy efficiency, cell size and area capacity are investigated under full load conditions. However, in this case, the effect of interference, noise and contribution of backhaul and static base station power consumption are also considered. The results show that smaller cell deployments lead to reduced transmit power but increased backhaul and idle power consumption. Furthermore, energy efficiency is optimised for the largest feasible cell size that meets the capacity requirement. However, with increasing capacity requirements larger cells are incapable of meeting the requirements, and the optimum cell size for energy efficiency reduces with increasing capacity requirement. This observation favours the deployment of smaller cells.

In [75] optimum cell sizes and transmit powers for energy efficiency are studied in a heterogeneous environment comprising four layers of BSs. The layers of BSs are 2G macrocells, 3G macrocells, 4G macrocells, and 4G microcells. They are deployed at the same location within the same macrocell coverage area. Closed form expressions are derived for energy efficiency in terms of the area power consumption and bits per joule per unit area for a reference macrocell coverage area. Static power consumption is considered in the evaluation of energy efficiency. The impact of the transmit power, inter-site distances between neighbouring BSs and number of macro and micro BSs are evaluated. It is shown that energy efficiency is optimised in a macro BS only deployment when macro BS transmit powers are between 15 and 20W while inter-site distance is between 1.5 and 2.0 km. Furthermore, addition of 4G micro BSs leads to improved energy efficiency in terms of both energy efficiency metrics evaluated.

In [76] the impact of cell discontinuous transmission (DTX), which is the deactivation of some BS components during periods when the BS is idle, on the optimal cell size for energy efficiency is studied. The energy efficiency problem is formulated analytically as a minimization of the daily average area power consumption with or without cell DTX in the service area. The minimization is subjected to coverage and QoS requirements by optimizing the cell range. This problem is solved by a time-static iterative algorithm, which determines the daily average area power consumption and busy hour data rates for different cell ranges individually. It then searches over all cell ranges for the optimum cell range. It is

shown that incorporation of cell DTX in the network deployment task leads to higher BSs deployment but significantly better energy saving relative to no cell DTX consideration.

It can be concluded from the investigations on appropriate cell sizes for energy efficiency described above that small cells are a necessity in future networks. How small the cell needs to be is a function of the capacity requirement. Nevertheless, macrocells are no longer the choice deployment of the future. Small cells are particularly seen as essential for crowded public places like rail stations and city centres. Studies on full scale deployment of small cells in such scenarios are presented in the next section.

2.6.2 Full Scale Small Cell Network Deployments

Homogeneous small cell networks have been proposed for urban hotspots (such as city centres and train stations) with high population densities. Small cell deployments are considered energy efficient because of the shorter distances between most mobile users and base stations on average compared to traditional macrocell networks. Also, as mentioned earlier, higher data rates can be achieved as a result of proximity of the mobile users to the base stations. These features make small cell networks (SCN) attractive.

Already, test runs of SCN have been carried out. Real Wireless carried out SCN trial in Newcastle and Bristol City Centres, United Kingdom with small cells mounted on existing structures such as lamp posts [77]. However, the goals of this deployment is the quality of service (QoS) performance not energy efficiency. The evaluation focused on the capacity density and throughput performances of the SCN. The findings of this trial reveal that small cells when compared to the macrocell deployment have the potential to triple the indoor throughputs and double outdoor median throughputs and also offer capacity offload for macrocells because of extensive indoor and outdoor coverage [77]. The significant margin of capacity and throughput of the SCN should translate to better energy efficiency relative to macrocells in terms of joules/bits.

The Beyond Next Generation (BuNGee) mobile broadband network is another SCN based proposal. The access network comprises of a dense deployment of only small cells referred to as access base stations (ABSs). BuNGee has been described as a cost-, spectrum- and energy-efficient architecture [78] and its primary goal is to achieve high capacity density of 1Gbps/km^2 [79]. Although, initial performance studies focussed on the capacity density and the QoS [80], more recent studies [31, 81] investigated the energy efficient operation of this network. Particular emphasis has been on how to reduce energy consumption when traffic demand is low or medium since the dense deployment of small cells is optimised for high capacity density during peak traffic condition.

In [31] topology management algorithms are proposed and applied to reduce the energy consumption of the BuNGee network by turning off lightly loaded ABS especially at low traffic load. Substantial energy savings are achieved relative to the scenario with all ABSs always kept on. Furthermore, in [81] energy efficient resource allocation schemes are combined with topology management algorithms to improve energy efficiency over the BuNGee network. Higher energy saving is achieved in this study compared to [31]. The schemes in [31] and [81] are discussed in more detail in section 2.7. The access network of the BuNGee Architecture has been modified in this thesis to conform with the separation architecture paradigm with the introduction of control base stations. All performance evaluation is evaluated on this modified BuNGee Architecture, which is presented in Chapter 3.

In [82] homogeneous microcell deployments are evaluated and compared to homogeneous macrocell deployments and heterogeneous networks consisting of a macrocell network overlaid with a number microcell base stations. The energy consumption of the different deployments are calculated based on an energy model proposed in [83], which assumes macrocell base station power consumption to be independent of the load while microcell power consumption adapts to the load. It is shown that dense deployment of microcells whether in the homogeneous case or the heterogeneous case leads to increase in the power consumption relative to the macrocell case. However, the increase in power is offset by substantial increase in average throughput when microcells are deployed leading to superior energy efficiency. Finally, it was observed that homogeneous microcell deployments are the

most efficient at medium and high load conditions, while at low load a heterogeneous mixture of macrocell and dense microcell is as efficient as a homogeneous microcell deployment.

Studies presented so far in this chapter have shown that small cells are really good for providing high capacity density and high data rates. However, apart from capacity, adequate coverage of the service area is also essential in cellular networks. Since, more small cells will be required to cover an area compared to a macrocell, it will not be practical to create a wide area network with small cells alone. Hence, a mixture of small cells and macrocells will be expected even in the future. Studies on the energy efficient operation of such joint macrocell and small cell deployment are discussed in the next section.

2.6.3 Heterogeneous Networks

Heterogeneous networks (HetNets) are usually made of a network of macrocell base stations overlaid with smaller cells such as femtocells, picocells or microcells. Femtocells are usually deployed indoors by mobile subscribers rather than network operators in areas with poor coverage and connect to the operator's core network via internet broadband connection owned by the subscribers [84]. Femtocells can be classified according to the access control approach utilised as open subscriber group, closed subscriber group, or hybrid [85]. Whereas anyone can connect to an open subscriber group femtocell, only pre-registered users can connect to a closed subscriber group femtocell; an hybrid femtocell may permit fixed number of unregistered users in addition to registered ones [85]. Picocells and microcells are usually deployed by the network operators.

In conventional HetNets, even when small cells are within the coverage area of macrocells, aside from handling data transmission and reception, they are still responsible for handling control plane functionalities. Control plane functionalities include the transmission of overhead (non-data) signals required to ensure mobile users can connect to a cell and enjoy reliable data service from a cell in the network. Another type of HetNet architecture, referred to as a Separation Architecture [20] or Hyper-cellular Network [19], has been proposed to significantly reduce small cell overhead signalling and also for better resource utilization and energy efficiency

[19]. In this type of architecture macrocells mainly handle the control plane functionalities and ensure ubiquitous coverage while the small cells within the macrocell coverage areas are responsible mainly for handling data services. Energy efficiency of the joint deployment of small cells and macrocells in both conventional and separation architecture based HetNets has received attention in the literature.

2.6.3.1 Conventional Heterogeneous Networks

In [86], energy efficiency of joint closed subscriber group (CSG) femtocell and macrocell deployment is compared with the macrocell only deployment. The femtocells are uniformly distributed in a random manner in the coverage areas of the macrocells. The impact of interference on the throughput and energy efficiency, when femtocells and macrocells use non-orthogonal spectrum, is studied for a downlink scenario. It is shown that energy efficiency increases with increasing density of femtocells, albeit at the price of deterioration in the throughput when the number of deployed femtocell per macrocell site is above 5. With reference to the macrocell only deployments, ECG of about 100 times can be obtained with deployments of 70 femtocells per macrocell site; on the other hand, up to 8% deterioration in throughput is observed when 80 femtocells are deployed per macrocell site. Hence, there is a trade-off between energy efficiency and spectral efficiency for joint femtocell and macrocell deployments. Therefore, the right number of femtocells per macrocell that improves energy efficiency without degrading the throughput should be selected.

In [87] the impact of the joint deployment of publicly accessible residential picocells and traditional macrocells on energy efficiency is considered. The evaluation was done with a varying number of randomly distributed picocells. Picocells are given priority to serve users and remaining users that cannot be served by the picocells are associated with the macrocells. The results show that the picocell contribution to the annual network energy consumption increases with number of picocells deployed. Also, reduction in the annual network energy consumption up to 60% can be achieved with the joint deployment of picocells and macrocells when high data rate is demanded by users relative to macrocell only deployment. The effect of future technological improvement of both macrocell and picocell base station power

consumptions on energy efficiency was also evaluated and it was shown that though the energy consumption is reduced for both macrocells and picocells, the improvement in energy efficiency due to picocell deployment is still sustained.

In a similar study [88], the energy efficiency and spectral efficiency of the joint picocell and macrocell deployment is evaluated under full load conditions. The number of picocells per cell is the same for all macrocells. Picocells are assumed to serve hotspots and used to offload traffic from the macrocells such that users being served by the macrocells can use more resources. The area spectral efficiency (ASE), "achievable rates in a network per unit bandwidth per unit area" [88] (bits/s/Hz/km^2), is found to decrease with increasing inter-site distance (ISD) and increase with increasing number of picocells per macrocell. The area power consumption (APC) also has a similar trend of decreasing with increasing ISD and increasing with increasing picocell deployment. The higher power consumption due to picocell deployment is offset by the leap in spectral efficiency. Thus, at a given ISD, the energy efficiency is observed to improve with increasing picocell deployments as compared to the macro-only deployment.

In [1] the energy efficiency of the joint picocell and macrocell deployments at full load and low load with the application of sleep mode for picocells is studied. Power consumption is evaluated for a varying number of picocell deployments per cell. The fibre point to point interfaces (which are part of the backhaul) and idle picocells are switched to the sleep mode to save energy. Power saving of up to 9% with reference to the macro-only scenario is observed at full load (or busy hour) when an optimum number of picocells are deployed. Exceeding the optimum picocell deployment leads to power gain for the macro-only over the joint deployment. On the other hand, at low load (or non-busy hour) negative power saving is observed without application of sleep mode at the picocells. However, the application of the sleep mode is shown to reduce the power loss for the joint deployment to almost half.

2.6.3.2 Separation Architecture Based Heterogeneous Networks

In [89] a novel separation architecture based on control and user (or data) plane split is proposed for capacity improvement in LTE networks but also incorporates energy

saving enabling features. In the proposed architecture, macro BSs provide continuous coverage while small cells BSs are introduced within the macro coverage to support high capacity demand. Macro BSs operate at the usual frequencies of LTE (below 2.5 GHz) and features both control and user plane signalling of conventional LTE. However, small cell BSs operate at higher frequencies of 3.5 GHz and higher and only support user plane signalling. Therefore, small cell BSs do not transmit the conventional control signals (e.g. cell specific reference signal and primary/secondary synchronization signals) needed by a mobile station or user equipment (UE) to associate with the small cell BSs. Hence, these small cells are referred to as Phantom cells. As shown in Figure 2.4, a UE can have a dual connection to a macrocell BS and a small cell BS at the same time, receiving control signals from the macrocell and data from the small cell. Macro BSs help transmit control plane signals on behalf of the Phantom cells and the connection of the UEs to Phantom cells are aided by Macro BSs in a sort of master-slave relationship through a new interface, X3. A new discovery signal for UEs to detect Phantom cells is proposed and it is assumed that the signal is time synchronised with the Macro BS to reduce the UE effort at detecting the Phantom cells. In order to save energy, it is expected that some of the Phantom cells can be switched off in networks with dense Phantom cell deployment.

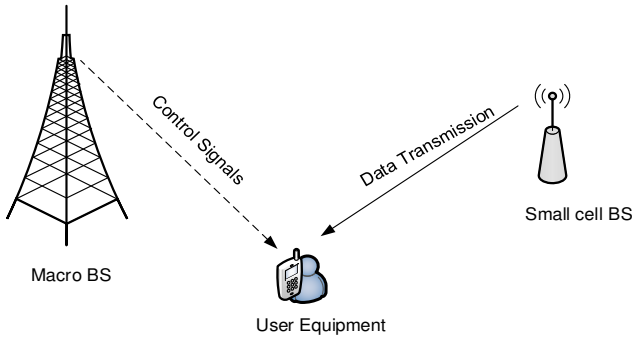


Figure 2.4 Control and Data Plane Separation

A similar architecture is proposed in [90], albeit with particular emphasis on energy saving rather than system capacity. Control signal transmission is separated from data transmission as well and transmitted by different types of BSs. On one hand, signalling (or control) BSs, optimised for long range and low data rate transmission, transmit control signals and maintain anytime, anywhere coverage utilizing a small

frequency bandwidth. On the other hand, data BSs optimised for short range and high data rate transmission handles data transmission only. As control signals require low data rates, signalling BSs are designed to be highly energy efficient with relatively larger coverage area compared to conventional architectures. In addition, the data BSs are activated on demand and switched off when no user is active in their vicinity. The activation of appropriate data BSs to serve a user request is done at the signalling BSs. Location information and past channel measurements during active sessions of BSs prior to deactivation are suggested for use in the selection process.

The sort of simultaneous connection of a user to macrocell and small cell described in [89] above has also recently been standardised by the 3rd Generation Partnership Project (3GPP) [91]. 3GPP is the standard body that provides specification for the operation of 3G technologies and currently does the same for the state-of-the-art 4G LTE/LTE-Advanced networks. The simultaneous connection is referred to as *dual connectivity* by 3GPP. Dual connectivity has been described as a process where a user can concurrently utilise radio resources from at least two access points (Master eNodeB and Secondary eNodeB) connected by non-ideal backhaul [91, 92]. A non-ideal backhaul has latency of several milliseconds to tens of milliseconds [93].

The dual connectivity concept has been proposed mainly to improve per-user throughput, provide robust mobility and reduce handover signalling in the small cell layer of such joint macro and small cells scenario described above [91, 92]. Unlike the previous proposals [89, 90], in dual connectivity both macro and small cells can serve as the Master node [94] and thus, the anchor point for control plane signalling. However, when the control plane is handled by the macrocell, mobility robustness is enhanced [92]. This is because loss of user connection due to incomplete handover process can be avoided when control signals, including handover commands, are transmitted by the macro layer [95]. In addition, when the macro layer handles the control signals and provide coverage, some small cells can be switched off at low load to save energy. Such a concept has been exploited in [32] and is discussed in section 2.7.3.

Some studies did not only present the energy efficiency or QoS benefit of the separation architecture HetNet but also evaluated the energy efficiency performance

as well unlike the previous studies above [89, 90]. In [49] a separation architecture is considered in a single cell scenario and achieved with the replacement of a conventional macro BS with a low power, coverage BS (CBS) and several small cell, traffic BS (TBS). While the CBS handles coverage, the TBS handles data services. Based on linear power models, the power consumption of the separation architecture is estimated using numerical computations and shown to achieve significant energy saving (more than 50%) relative to the conventional macro BS approach. Furthermore, closed form expressions are derived for the adaptation of TBS intensity to changes in user traffic intensity and the optimal TBS intensity for fixed user traffic intensity. The optimal TBS deployment and TBS intensity adaptation are shown to achieve almost 60% energy saving compared to the conventional macro BS.

Similarly, the joint optimization of density of BSs, number of antennas and spectrum allocation for energy efficiency in a separation architecture based HetNet of macro BSs and small cell BSs is investigated in [96]. The optimization problem is formulated and solved in two stages. Firstly, the optimal BS density and number of antennas for each BS tier are determined under the assumption of a known spectrum allocation, which is expressed in terms of the share of bandwidth allocated to each BS tier. The optimal spectrum allocation that results in minimum network energy consumption is then determined. The combination of small cell BSs and multiple antennas is shown to provide significant energy saving relative to a single antenna macro BSs only system. However, the most energy efficient of the two concepts is a function of the design target of the system.

Dense deployment of small cells is proposed to meet the high data traffic demands of the future; albeit macrocells are still required to provide umbrella coverage [13]. Both conventional and separation architecture based HetNets of macrocells and small cells are shown to achieve better energy efficiency in comparison with macrocell only deployments in the studies discussed above. However, apart from the energy efficiency gains of introduction of small cells alongside macrocells, there is benefit in adapting the network state in terms of active cells and resources utilised to instantaneous traffic demands. Dense deployments of small cells are believed to be essential to meeting the goals of the high capacity and high data rate demands envisaged in future wireless networks [97] (i.e. 5G and beyond). However, as the

traffic demand will not be at the peak value for most of the time, techniques to dynamically adapt the number of active cells (or BSs) and the resources (bandwidth, timeslot and power) to the current traffic load are essential. Resource management and topology management techniques that have been proposed in the literature for such dynamic adaptation in both conventional and separation architecture based HetNets are discussed in the next section. Although, there have been studies on homogeneous macrocell deployments, only those on HetNets, which is the focus of this thesis, are presented.

2.7 Existing Energy Efficient Resource and Topology Management Techniques for Heterogeneous Networks

A brief overview of resource management and topology management as they relate to heterogeneous networks and energy efficiency are first provided as background to studies reviewed subsequently and schemes proposed in this thesis. Then, RRM and TM algorithms proposed in the literature for conventional HetNets are presented. This is followed by those proposed for separation architecture based HetNets, which is the type studied in this thesis.

2.7.1 Overview of Resource Management and Topology Management

In cellular networks, radio resource management (RRM) algorithms are applied to share resources among active users in order to meet a certain goal (e.g. fairness among users or high system capacity). Traditionally, radio resources include power, frequency band, and time slots and the goal of a RRM algorithm is to assign these resources to achieve a predefined goal. In [17], a new paradigm of resource management is considered for a densely deployed LTE network with mainly small cells. Instead of the conventional LTE approach in which BSs or eNodeBs individually allocate their resources, BSs in a geographical area are abstracted as a virtual big-base station with radio elements (the BSs) at different locations in the area under the management of a centralised controller. Furthermore, a BS index is considered as part of the radio resource alongside the conventional time slot and frequency band considered in LTE. Hence, the individual BSs can be seen as part of the radio resource of the central controller. The controller helps to determine the best

BS and the resource on the BS to serve a user in the area. This type of paradigm is adopted in the RRM algorithms proposed in this thesis.

Furthermore, RRM algorithms can be proposed for a single cell scenario where the interference from other cells due to reusing similar resources is not taking into account. Hence, the resource utilization can be optimised without consideration of other cell effects. However, in a multi-cell scenario, inter-cell interference which may arise due to the broadcast nature of radio transmission and the reuse of similar radio frequencies in neighbouring cells cannot be neglected [17]. Hence, multi-cell RRM algorithms usually take inter-cell interference into account when assigning resources [98, 99]. Different approaches of reusing the shared frequency bandwidth have been studied in the literature to mitigate inter-cell interference in multi-cell scenarios. However, only approaches which mitigate inter-cell interference with the aim of improving energy efficiency are presented. In this thesis, inter-cell interference and it is impact on energy efficiency is studied in a separation architecture based HetNet and a novel interference aware energy efficient resource management scheme is proposed. In addition to RRM algorithms, topology management algorithms have also been proposed to ensure energy efficient operation of HetNets.

Topology management in the wireless sensor network domain involves understanding the interconnections and relationships between nodes as well as the subset of nodes that needs to communicate at a certain time to conserve energy [100]. While the interconnection and relationship between nodes may not be as important in a cellular system without the multiple hops of the sensor network, the selection of communicating set and energy conservation is relevant. Topology management (TM) is defined herein as the dynamic control of the network topology in order to balance the state of the network in terms of the active BSs with the traffic being served and conserve energy. This includes algorithms that select which BSs to put to sleep (or switch off) and those to keep on in accordance with traffic load. Hence, the cell activation/deactivation, sleep mode and cell switch off algorithms have been classified under the broad heading of topology management. Some studies considered topology management schemes in isolation while others considered a joint resource and topology management algorithm.

2.7.2 RRM and TM Techniques for Conventional HetNets

In [101], an optimal sleep and wake up scheme is proposed for the downlink of a HetNet of macrocells and femtocells deployed under the coverage of macrocells. The authors use a Markov Decision Process based approach to derive optimal policies for selection of femtocells to switch off or wake up at a macrocell based on the traffic load and user location information. The cases of complete, partial, delayed or no location information are investigated. It is shown that reduction in energy consumption in the network is possible in all cases even with partial or no user location information when femtocells are switched off. However, the user perceived QoS, measured in terms of the user throughput, is poorer without complete location information.

The authors in [26] proposed a partial spectrum reuse (PSR) scheme which requires micro BSs to reuse only a portion of the system spectrum in order to mitigate interference caused towards the macro-tier. Specifically, a PSR factor, which is the portion of system spectrum the micro BSs can reuse, is determined. A closed form expression for the optimal PSR factor is derived and based on this the threshold for energy cost of micro BS is derived. It is shown that if the energy cost is lower than the threshold, more micro BS should be deployed for capacity extension, or more macro BSs should be switched off for energy saving. The reverse is the case when the energy cost is higher than the threshold. A significant reduction in the energy consumption is achieved in the network with the PSR approach relative to a non-PSR scheme.

Energy efficient resource allocation is studied in a HetNet where small cell BSs are permitted to share the spectrum of a macro BS under an interference power constraint and incomplete Channel State Information (CSI) in [27]. This is formulated as a Stackelberg game where the macro BS acts as the leader by setting an interference price as its revenue, while the small cell BSs are followers. The small cell BSs determine the transmit power for a subchannel based on the interference price and the channel gains of their users. A closed form expression for the Stackelberg equilibrium is obtained that jointly maximises the interference revenue of the macro BS and the energy efficiency utility of the small cell BSs. The proposed

approach is shown to outperform an existing scheme with complete CSI in terms of the total interference revenue of macro BS and the total utility of the small cell BSs for different interference power constraints.

In [102] a cell association and switch off algorithm is proposed for a HetNet comprising of smaller cell BSs which belong to lowers layers and are located within the coverage of larger cell BSs of higher layers. Energy saving is achieved by exploring the possibility of serving locations previously served by smaller cell BSs with larger cell BSs of another layer on the condition that capacity demanded from these locations can be supported. Subsequently, if the smaller cell BSs are idle they are switched off. This concept is mapped to a 0-1 Knapsack-like problem with profit of an action taken being the achievable energy reduction, while the weight of the action is the capacity demanded at the location. Mean energy saving of 35% is achieved in a three layer system across several scenarios relative to an always on approach for BSs in all layers.

The increasing complexities of mobile networks due to deployment of different cell types (HetNets), coexistence of different access technologies (i.e. 2G, 3G and 4G) and different user QoS requirements mandate automation of mobile network processes [103]. Such an automated network is term a Self-Organising network (SON) [103]. The resource and topology management schemes discussed above inherently require autonomous co-ordination between BSs that are not explicitly described in terms of SON. However, in some studies this SON autonomous functionality is explicitly stated. A more detailed explanation of SON is given in Chapter 6 where an adaptive scheme which utilises the SON paradigm is proposed.

In [28] a SON based RRM scheme is proposed for energy saving in a HetNet of macro BSs and small cell BSs through interference management. A matrix of potential conflicts of interference, referred to as matrix of conflict, between BSs on each subcarrier is created. When there is no potential conflict between BSs, resources are allocated by the BSs individually in a distributed manner. However, when there is interference conflict, a central SON node is used to coordinate the resource allocation and avoid the potential interference. The proposed scheme achieves better QoS and energy efficiency with respect to existing schemes.

Also, in [104] energy efficient load balancing is studied in a LTE-A network where infrastructures are shared among different operators according to service level agreements. A load balancing algorithm is proposed which utilise a centralised SON module to make BS selection decisions based on channel quality measurements reported by mobile users and traffic load counters of BSs. The BS traffic load of an operator involved in the sharing agreement is offloaded to neighbouring BS belonging to another operator if a specified load threshold is exceeded. However, if the load threshold is not exceeded the user association to BSs is done to minimise energy consumption. The proposed algorithm is evaluated in a network comprising two operators with an overlapping coverage area. The coverage area of each operator is served by a macro BS (eNodeBs in LTE) aided by a relay node. It is shown that the proposed algorithm leads to significant improvement in the average cell-edge user throughput and a reduction in energy consumption under the sharing regime relative to the non-sharing case.

In the context of the BuNGee Architecture evaluated in this thesis, there have been some investigations on the original architecture which features a full scale small cell deployment. Also, a modified version where macrocells are introduced into the access network, creating a HetNet access network has also been investigated. However, this is a conventional HetNet without separation of the control and data plane. The work presented in this thesis is the first proposal of separation architecture for BuNGee and its energy efficiency evaluation. The studies on the baseline and the modified architecture are presented as follows to provide a good background for the energy efficiency aspect of the BuNGee Architecture.

Different topology management strategies are proposed for the original BuNGee Architecture in [31] to put base stations to sleep in the full scale small cell access network. The strategies are defined to switch a BS on or off based on instantaneous traffic load and average load thresholds of the BS and its neighbours. The threshold utilised varies from strategy to strategy. In some cases the working state of neighbour BSs are also considered when switching BSs off. The strategies are compared with regard to their blocking probabilities and energy efficiencies. Energy savings between 35% and 70% are achieved at low traffic load with respect to the conventional approach without BS deactivation.

In [81] joint resource and topology management schemes are proposed for the original BuNGee Architecture. The topology management algorithm in this study also considers the traffic load threshold of BSs and neighbours in deciding which BSs to switch on/off and when to perform this action. However, BSs considered as neighbours are defined differently from the case in [31]. Two resource management schemes are proposed and combined with the topology management scheme. The first one, capacity based channel assignment (CBCA) scheme, assigns users to the highest loaded base station as long as the signal to interference plus noise ratio (SINR) threshold is satisfied. The second scheme, the priority based channel assignment (PBCA) scheme, assigns random priority values to BSs. BSs with higher number of neighbours are usually assigned higher priority values than their counterparts with lower number of neighbours. The results show that both CBCA and PBCA schemes when combined with the topology management algorithm achieves higher energy savings (15% more on average) than the conventional highest SINR based scheme with sleep mode. Furthermore, even though the CBCA achieves slightly higher energy savings than the PBCA, at low traffic loads the blocking probability of the CBCA is significantly higher than the PBCA scheme.

In [105], the original BuNGee architecture is modified to include macro BSs for provision of continuous coverage throughout the service area. Energy saving based on switching off the small cell BSs and macro BSs are studied. A topology management scheme that switches off small cell BSs and macro BSs, depending on traffic load and average delay per cell, is proposed. It is shown that the proposed scheme achieves higher energy saving and better QoS relative to the previous work in [81] on the BuNGee architecture where macrocell layer is not considered.

2.7.3 RRM and TM Techniques for Separation Architecture Based HetNets

In [23] the separation architecture is achieved through data and signalling BSs respectively. The signalling BS carries out prediction of the best data BS to serve users based on the user location unlike conventional pilot signal based estimations, which utilise received signal strength. The data BSs are activated on demand (i.e. cell on demand) and are switched off when no user is active within their coverage area. Two statistical models are developed to analytically determine the probability

of activation of data BSs: a Poisson model and an integral geometry model. The Poisson model assumes that activation occurs if at least one user is in the coverage area of the data BS, while the integral geometry model activates data BSs if a user is not already covered by adjacent data BSs. The integral geometry model matches simulation results of activation probability for a typical daily traffic profile better than the Poisson model. It is shown that the energy consumption of the cell on demand approach is 50 times more energy efficient than conventional macro BS only systems using the integral geometry model.

Similarly, in [106] the energy efficiency of a separation architecture comprising of data handling small cells, which carry no cell-specific reference signals and are called phantom cells, and control handling macrocells is evaluated. The phantom cells are operated on different frequency bands from the macro layer and thus there is no cross layer interference. A closed form expression for energy efficiency in terms of the ratio of the spectral efficiency to the power consumption is derived for phantom cells. Energy efficiency as well as the spectrum efficiency of the small cell layer is shown to be better in phantom based approach relative to conventional HetNet with shared frequency bands and no plane separation.

Whereas [23] and [106] do not seek consistency with existing standards, in [20] a signalling approach using existing wireless standards (such as UMTS and LTE) is considered in developing the separation architecture. As a result in [20], the small cell data BSs still carry a reduced set of overhead signals including pilot (or reference) signals. However, coverage and low-rate data services are handled by macro control BSs, while data BSs handle high-rate data services. A scenario comprising of a macro BS and several small cell BSs is simulated. It is assumed that in a non-separation architecture if the number of users served by a small cell BS falls below a given threshold for a specified period of time it can go into sleep mode. Small cell BSs are assumed to enter sleep mode quicker in separation architecture since user access is guaranteed by the macro BS. The separation architecture is shown to achieve significant energy savings of more than one-third of the energy consumption of the non-separation approach.

A novel database aided cell activation scheme is proposed for a separation architecture based on the Phantom Cell Concept (PCC) in [107]. Small cell BSs are put into sleep mode and RF receiving and transmitting chains are deactivated. This prevents reception of UE wake up signal and detection of pilot signals by UEs. At each macrocell BS, the proposed scheme builds a database of Signal to Noise Ratio (SNR) of each small cell to different geographical location within the coverage area from UE measurement reports. The SNR values are then used to estimate the SINR of small cell BSs in sleep. Hence, an active user can be connected to the small cell BS with the best SINR either in sleep mode or already active. Simulation results show that the proposed scheme can achieve energy saving of up to 40% relative to a system where small cell BSs are not switched off in dense user deployments.

Energy saving in a heterogeneous network of Macro eNodeBs (MeNBs) and Small cell eNodeBs (SeNBs) based on the LTE-A dual connectivity concept, is studied in [32]. An energy efficient scheme is proposed to activate sleeping SeNBs and offload the MeNB traffic load to SeNBs only when network energy consumption will be reduced whilst taking into account the backhaul link power consumption. SeNBs are deactivated by MeNBs when the energy efficient condition is no longer satisfied. Energy saving of 20% or lower is achieved by the proposed scheme relative to conventional cell activation schemes based on small cell proximity and traffic load.

Possible energy saving resulting from switching off small cell BSs in a Phantom Cell Concept (PCC) based HetNet comprising control macro BSs and phantom small cells is studied in [108]. The authors propose a heuristic algorithm that requires macro BSs to activate the small cell with the best SINR and sufficient resources to guarantee a user data rate requirement. The algorithm is applied in three energy saving schemes (uplink signalling, downlink signalling and database aided schemes), each with a different method of signalling sleeping small cells. The power consumption model is derived for the small cell BSs under the PCC architecture and the sleep mode consumption portion for the three energy saving scheme is determined. The delay involved in connection of a user to an activated small cell is also determined for the three schemes. It is shown relative to a baseline scheme with all small cells always on that energy saving of up to 45% and throughput

improvement of up to 25% is possible with the macro BS based activation schemes in spite of connection delays.

In [109] the tradeoff between total energy consumption and overall delay for a Data BS under the coverage of a control BS in a separation architecture is studied. The data BS is modelled as an M/G/1 vacation queue and closed form expressions are derived for BS switch off policies based on close-down time before sleep, total packet arrivals before waking up and setup time after BS wakes up. The relationship between energy consumption and mean delay is shown to be linear under varying close-down time. However, this relationship is non-linear with varying total packet arrival before BS wake up.

In [110] the ratio of small cell BSs that can be put to sleep is investigated in a separation architecture. Closed form expressions are derived for the outage probabilities of users connected to macro BSs and small cell BSs under a random sleep scheme, which puts small cell BSs to sleep with an equal probability, and a repulsive sleep scheme, which puts only small cell BSs which are a certain threshold distance away from macro BSs to sleep. The ratio of sleeping small cell BSs are determined for the random and repulsive sleep schemes based on derived outage probabilities. With random sleep, the ratio of sleeping small cell BSs is inversely proportional to the density of small cell BS users but linearly decreases with the density of macro BS users. The repulsive scheme is less sensitive to traffic variation but more beneficial to high traffic load situations.

In [111] a separation architecture is proposed for future 5G cellular systems. The architecture is design with macro BSs responsible for handling the bulk of control signal transmission and co-ordinating the allocation of the resources on the small cell BSs, which are used primarily for data service and thus only transmit data-related control signals. In addition, small cell BSs are operated at higher frequencies like 3.5 GHz while the Macro BSs utilise conventional lower frequencies of 2 GHz similar to LTE. Simulation results show that the proposed architecture with centralised small cell BSs on/off approach achieves 17% more throughput than an LTE distributed on/off approach due to reduction in overhead and absence of cross-tier interference. In addition, nearly 75% percent energy efficiency gain is achieved since unlike in

LTE most small cell BSs can go into deep sleep while macro BSs guarantee coverage.

2.7.4 Important Observations on RRM and TM Techniques

Based on the RRM and TM studies discussed above under the conventional HetNet and Separation Architecture based HetNet, some important observations are noted. Firstly, RRM and TM strategies enable HetNet to adapt the number of active cells to traffic load and assign resources in an energy efficient manner, resulting in better energy efficiency than the macrocell deployment only. Also, while a conventional HetNet has to deal with interference between macro and small cell layers, this is avoided in separation architecture based HetNets since different frequencies are utilised, and this results in better resource utilization. Furthermore, when conventional HetNets are compared directly with separation architecture based HetNets, energy efficiency performance are shown to be better under the separation architecture paradigm. This is due to the reduced signalling overheads and better opportunity for small cells to be put to sleep since they are not deployed for coverage reasons. Since the separation architecture is a relatively new architecture compared to the conventional architecture, there are still many aspects that have not been looked into.

The energy efficient operation of the small cell BSs when uplink traffic is being served has not been investigated. It is the downlink that has been usually considered. However, the power consumption of small cell BSs in state-of-the-art systems in the uplink is comparable to the downlink and should not be overlooked. In addition, the impact of power model assumption, which can be interpreted as the impact of different future enhancements of BSs, has not been considered in previous studies on separation architecture. These two issues are considered and investigated in this thesis.

Furthermore, in most studies on energy efficient RRM and TM including both conventional and separation architectures, the interference is assumed constant throughout the duration of user transmission. Hence, the SINR and data rates are constant during the duration of transmission. Hence, when RRM algorithms assign resources to users under the condition of meeting a QoS requirement, it is assumed

that once the condition is satisfied at the beginning of the transmission it will not be violated throughout the transmission duration. This is acceptable when the system is modelled at packet level and RRM decisions are made at every time slot of few milliseconds. However, users' perception of network performance is reflected by flow and session behaviour rather than packets, and performance evaluation is better done by flow level modelling [112].

A flow can be referred to as a series of packets associated with a particular type of document such as a file or a music track [113] and requires longer transmission duration than a packet. Flow level modelling considers random arrival of flows and once flows are admitted into the system they remain in the system until they are successfully delivered. Hence, interference experienced by a particular flow may vary during the transmission duration and therefore, constant interference cannot be assumed as a result. Furthermore, QoS requirements may be violated during the course of the transmission. In this thesis, flow level modelling is used and dynamic inter-cell interference is considered. Also, detection and rectification of QoS deterioration is a key feature of the schemes proposed in this study.

2.8 Conclusion

In this chapter different studies focussed on achieving energy efficiency in cellular networks are discussed, with particular emphasis on heterogeneous networks. This literature review provides a background for the studies eventually carried out in this thesis. Specifically, energy models and energy efficiency metrics which are fundamental tools for measuring energy efficiency are discussed. Also, opportunities for saving energy through enhancement of base stations and network deployments are presented. Finally, resource and topology management techniques proposed for energy efficient operation of both conventional and separation architecture based heterogeneous cellular networks are discussed.

Chapter 3. System Modelling and Performance Evaluation Techniques

3.1 Introduction

This chapter describes the techniques that are used to model the system scenario and evaluate the performance of the proposed schemes in this thesis. The performance metrics used in the evaluation are also explained. Furthermore, the approaches used to verify that the system model and proposed schemes deliver reliable results are explained.

Telecommunication systems can be modelled by analytical, practical or simulation tools. Practical models utilise real devices in evaluating the system at smaller scales (referred to as test beds) before full deployment since large scale implementation will be too costly. Analytical models are used to derive closed form mathematical expressions that describe the behaviour of the system and can be used to predict the performance of the system. On its part, simulation models are virtual representations of a real system created using software programmes and also capable of providing system performance results. Analytical and simulation models are not limited by the problem of cost like practical models.

The increasing complexity of telecommunication systems makes it often impractical to model them with analytical models, as a result simulation tools are the main candidate for the complex cases [114]. The telecommunication network considered in this study comprises a dense deployment of small cell base stations (BSs) directly serving mobile stations (MSs) under the coverage of macro base stations providing control signalling. In addition, energy efficiency through the use of joint Radio Resource Management (RRM) and Topology Management (TM) schemes is evaluated in this network. The mutual impact of allocated resources of different small cell BSs on one another through interference and the switching off and on of small cell BSs at different time instances in this network constitute a complex process.

Therefore, the complexity that will be involved in effectively evaluating energy efficiency performance in the entire network with analytical models make this

approach a less suitable candidate for thorough evaluations. Practical approaches are limited by the cost of acquiring sufficient devices to effectively study the energy efficiency for the extensive network considered. Hence, simulation techniques are the major tools used for the evaluation of the schemes in this study. However, a performance bound is derived analytically for the number of active small cell BSs required to achieve good QoS and energy efficiency. Furthermore, mathematical expressions leading to the algorithms developed and the energy models used are clearly derived and explained. This is the case with the schemes and models presented in Chapters 4 to 7. In addition, confidence intervals are used to verify that the models and algorithms deliver reliable results for the different simulation based evaluations.

The rest of the chapter is organised as follows. The simulation techniques are discussed in section 3.2, while system modeling approaches are presented in section 3.3. The performance metrics used in this work are explained in section 3.4, followed by a brief overview of the energy efficient RRM and TM processes in 3.5. The methods utilised for verification of results are discussed in section 3.6 and the chapter is concluded in section 3.7.

3.2 Simulation Techniques

3.2.1 Link Level and System Level Simulation

The evaluation of wireless networks can be done at different levels including the link level and the system level. While the link level models a radio link of the network at the bit level, the system level models the full network [115]. Link level simulations usually evaluate the performance of the link between a base station and a mobile station [116]. Physical layer parameters such as antenna gains and bit error rates can be obtained from this type of simulation. On the other hand, system level simulations model network of multiple mobile stations and multiple base stations. They are concerned with network related issues such as interference management, user mobility and RRM [117].

In order to fully understand the performance of a wireless network, both link level and system level simulations have to be carried out. Although, a single simulator that

incorporates everything is desirable, this kind of simulator will be far too complex because of different simulation resolutions and durations usually required for link level and system level simulations [118]. Hence, the two simulation types are carried out separately and the result of the link level simulation is abstracted and utilised in the system level simulation. The interfacing of the system level and link level simulations is usually accomplished by a link level abstraction model. The link abstraction model feeds system level parameters such as the signal to interference plus noise ratio (SINR), a measure of radio link quality at the receiver, into the link level and produces as output link quality parameters such as the Block Error Ratio (BLER) and the data rates. In practice, look-up tables or performance graphs are used in system level simulations for mapping SINR to data rates. The system level simulation is implemented in this thesis with a programming language called MATLAB. The reason for the choice of MATLAB is explained next.

3.2.2 Simulation Software

Several software applications are available for modelling communication systems such as C, OPNET, and MATLAB. C is a popular programming language that has been around for decades; it is powerful and flexible and has been utilised for diverse projects including operating systems and word processors [119]. It is highly efficient, nevertheless more difficult than other programming languages due to large number of operators but few keywords [120].

OPNET is a simulation tool developed specifically for modelling communication networks rather than being a general purpose programming language like C. It has the capability to model and evaluate complete heterogeneous networks with graphical user interface for selecting different network components and choosing parameters of these network components [121]. Furthermore, it boasts a large database of existing network components, supports network protocols, produces graphical outcomes and statistics and also free for academic institutions [122]. These features make OPNET user-friendly and suitable for evaluating performances of real networks.

MATLAB is a programming language that combines the capability for users to develop models from scratch and the availability of ready-made models in the form

of toolboxes for diverse application including communication, control, and financial analysis among others. MATLAB features a rich set of built-in mathematical functions and graphical tools and user-friendly visual interfaces [123]. These features make it easy to solve mathematical expressions, view plots of analysis, and debug written programmes. In addition, programming syntax is a lot easier in MATLAB than C. For example, it is not necessary to declare variables and array sizes before being utilised [123]. This implies that system models can be built quicker with MATLAB than C. MATLAB has therefore been used in modelling and evaluating the telecommunication system in this work, as a result of its flexibility, ease of coding, and user-friendly, built-in mathematical and graphical tools.

3.3 System Modelling

The telecommunication network considered in this study is the modified version of the Beyond Next Generation mobile broadband network (BuNGee). BuNGee is a mobile network based on a two-tier deployment of backhaul network and access network[124]. It has been modified in this study to incorporate high power control BSs, referred to as the Zone Base Stations (ZBSs), in the access network. As explained earlier in Chapter 2, this allows the separation of the control and data planes in order to provide ubiquitous coverage for Mobile Stations (MSs) and reduce signalling overhead. It also provides the opportunity to switch off all or almost all small cell BSs in the access network. The layout of the different network nodes in the modified BuNGee Architecture is explained later in this section. The unique features of each node are described and are implemented in the simulator as described.

Several other models are implemented on top of the network model to simulate the different related processes going on over the network. The traffic model implements the arrival of user requests into the system and the departure from the system after being served. The radio propagation model estimates the signal level at the receiver given the signal level at the transmitter. The energy model outputs the energy consumption of the network nodes based on the usage over the monitored period.

These models are fundamental to the accurate implementation of the operations in the network and they are interrelated. Clearly, a user arriving into the network,

modelled by the traffic model, can be served by the base station only when the quality of the communication link, determined by the radio propagation model, is good enough. Furthermore, the energy consumed in serving this user, estimated with an energy model, depends on the time the user spends in the network which is a function of the link quality and thus, the radio propagation model as well. These models are discussed in details later in this section. In addition, the performance metrics used in characterizing the performance of the system are also discussed.

3.3.1 Network Layout

The BuNGee Architecture considered in this work is a high capacity, cost-, spectrum- and energy-efficient architecture [78], with the goal of 1 Gbps/km^2 [79], proposed within the framework of the FP7 BuNGee Research project. As mentioned earlier, in this work ZBSs are introduced in the access network to facilitate the separation of the control and data plane. This results in a separation architecture based access network as discussed in Chapter 2. The location of all the other network components remains the same as in the initial architecture. A brief overview of the architecture is provided here and specific detail relevant to each evaluation are emphasised in the subsequent chapters.

The BuNGee Architecture is based on a two-tier deployment of access and backhaul network [124] as mentioned earlier. The access network consists of the BSs that provide resources (time, frequency, power) used to serve MS requests while the backhaul network connects the access network to the core network of the cellular system. The BuNGee access network consists of a dense deployment of small cell BSs, referred to as Access Base Stations (ABSs), in a regular pattern as shown in Figure 3.1. The ABSs are spaced 90 meters from each other along the streets and are deployed outdoor below rooftop on existing structures like street lamp. The building blocks have a square dimension of 75m by 75m and the spacing between buildings forms the streets in the service area. Each street has a width of 15m.

In this work, the architecture is subdivided into square zones and each zone is served by a ZBS. Each ZBS provides coverage and always-on connection throughout the zone it is serving to the MSs residing in the zone. This is achieved through provision of essential control, pilot (or reference), and synchronization signals by the ZBSs.

The data services required by MSs are provided by the ABSs under the control of the ZBSs and ABSs can be switched off to save energy while the ZBSs are always on to ensure ubiquitous coverage. The co-ordination between the ZBSs and MSs are explained in detail in Chapter 4 while the control and data plane separation are explained in Chapter 5.

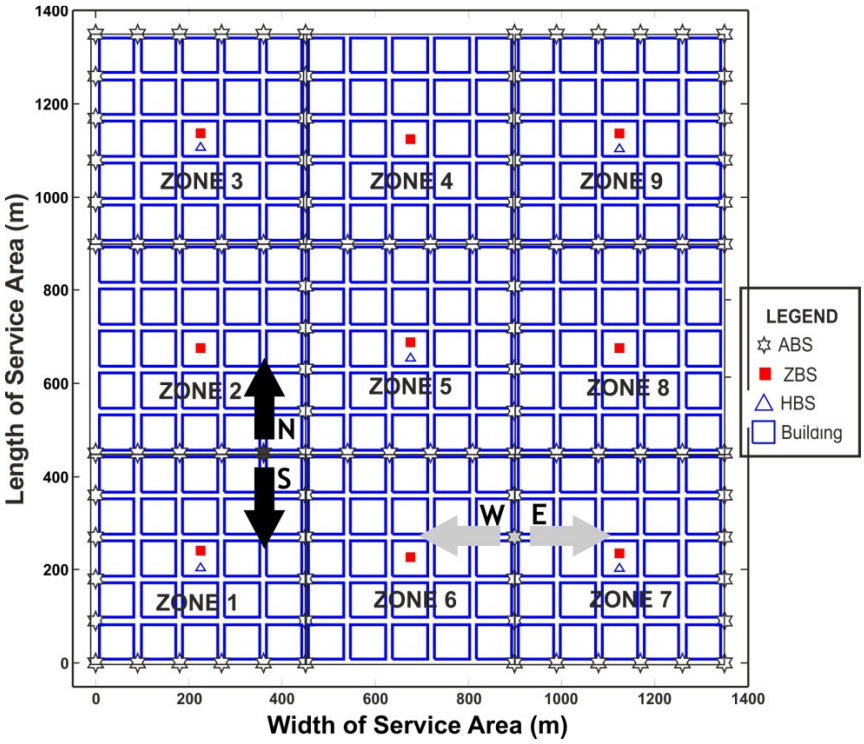


Figure 3.1 BuNGee Topology

Each ABS is equipped with two directional antennas pointing in opposite directions along the street. The ABS antennas are deployed at a height of 5m above the ground. ABSs have been classified as East-West ABSs or North-South ABSs depending on the directions of their antennas. East-West ABSs are located along the north-south streets and have antenna beams in the east and west directions like the ABS shaded in grey in Figure 3.1. Similarly, North-South ABSs are located along the east-west streets and have antenna beams in the north and south directions like the ABS shaded in black. The gain experienced by an MS signal at an ABS antenna beam is obtained from an antenna gain profile that has been derived for the ABSs in the BuNGee project for different combination of the elevation and azimuth angles of an MS to an ABS beam [125]. Such elevation and azimuth angles about an ABS beam are shown in the 3D antenna pattern in Figure 3.2. The ABS antenna is a directional antenna

with a beam width of approximately 25 x 25 degree and a gain of 17 dBi [125]. The MS antenna is assumed to be omnidirectional with a gain of 0 dBi [80]. It is important to note that the ABS and MS antennas are deployed below rooftops.

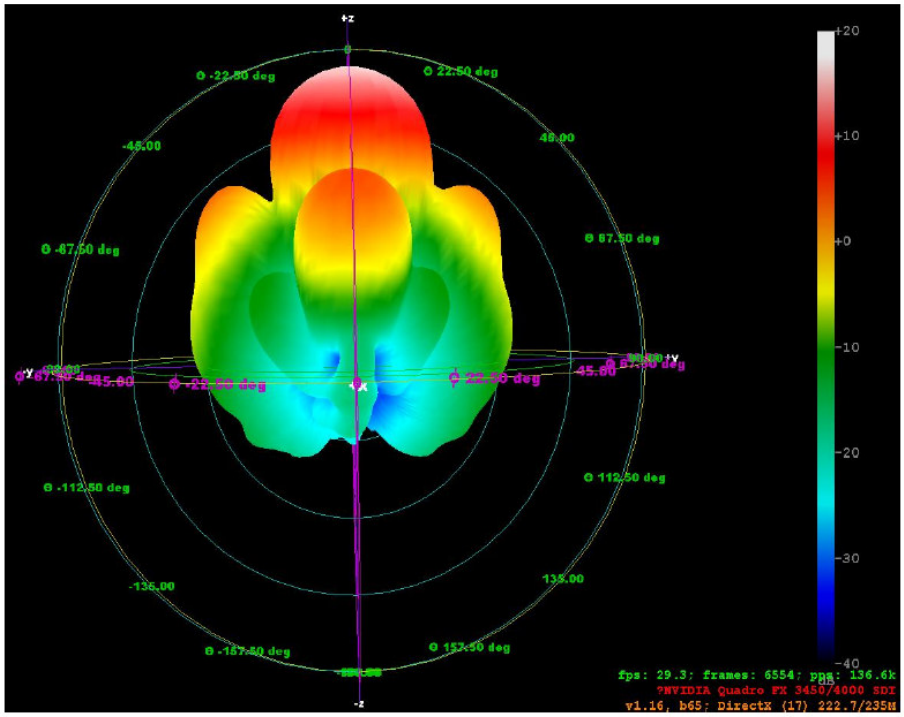


Figure 3.2 ABS 3D antenna pattern (directly reproduced from [80])

The backhaul network consists of Hub Base Stations (HBSs) and Backhaul Subscriber Stations (BHSSs).The BHSSs are also referred to as Hub Subscriber Stations (HSSs) [80, 125]. The BHSSs are co-located with the ABSs, so at each ABS location there is a complementary BHSS for backhauling MS information to a HBS. Backhauling is achieved through in-band and millimeter Wave (mmWave) backhauling. The in-band backhauling involves the link between a BHSS and an HBS while the mmWave backhauling involves shorter link distances between BHSSs, due to the shorter range of millimeter waves. It is assumed that separate frequency bands are used for the access and backhaul network and interference between the two tiers are completely avoided.

The HBSs are high power, wide coverage base stations deployed solely for backhauling. They serve as high capacity hubs for the network through 24-beam, dual-polarised antennas at 3.5GHz [125]. HBS antennas are deployed above rooftops

at heights of 25m above the ground. HBSs are located in a manner to ensure each BHSS can connect to at least one HBS. Each BHSS supports both in-band backhauling (at 3.5GHz) to an HBS and mmWave backhauling to another BHSS (at 60 GHz). The in-band backhauling is based on a directional antenna with a beam width of approximately 40 x 40 degree and a gain of 13dBi [125]. The backhaul network is assumed to be on always in this study and the focus is on the energy saving in the access network through the switching off of some small cell BSs while receiving uplink data from MSs.

3.3.2 Radio Propagation Models

Signals propagated over wireless channels usually experience deterioration in signal strength due to adverse effects caused by natural phenomenon and man-made structures. These channels are usually assumed to be characterised by three main effects: path loss, shadowing and multipath fading [116]. Path loss is an attenuation of the transmitted power [71] and it is directly proportional to the distance between communicating nodes [39]. Shadowing is the attenuation due to such phenomena like reflections and diffraction resulting from obstruction of the radio path by large objects [114]. Multipath fading is the fluctuation in received signal as a result of transmitted signals arriving at the receiver through different paths with different attenuation and delay [114].

Path loss and shadowing are large scale fading effects [126] that are dependent on the positions of the communicating nodes and they remain the same as time passes [117]. On the other hand, multipath fading changes quickly with time and/or frequency and it is also referred to as fast fading. In this study, it is assumed either that the duration of file transmissions over the wireless channel are much longer than the rate at which multipath fading changes or that frequency diversity can be applied. Therefore, multipath fading is observed by its average and assumed to be incorporated in the location dependent path loss like in [112]. Path loss and shadowing are modelled using the WINNER II B1 urban micro-cell model [127]. This is specified for the scenario where the BS and MS are outdoors and below rooftop, which is studied in this work.

In [127], the path loss model for line of sight (LOS) communication between a MS and an ABS is different from the non-line of sight (NLOS) model. The LOS path loss, $PL_{LOS,}$ is obtained as follows:

$$PL_{LOS} = \begin{cases} 22.7\log_{10}(d_1) + 41 + 20\log_{10}(f_c/5);\ 10m < d_1 < d'_{BP} \\ 40\log_{10}(d_1) + 9.45 - 17.3\log_{10}(h'_{BS}) - 17.3\log_{10}(h'_{MS}) + 2.7\log_{10}(f_c/5);\quad d'_{BP} < d_1 < 5km \end{cases} \tag{3.1}$$

d_1 is the distance between the MS and the ABS, f_c is the carrier frequency, d'_{BP} is the breakpoint distance, h'_{BS} and h'_{MS} are the effective ABS and MS antenna heights for ABS and MS antenna heights of h_{BS} and h_{MS} dimensions respectively.

$$d'_{BP} = 4\,.\,h'_{BS}\,.\,h'_{MS}\,.\frac{f_c}{c};\ c = 3 \times 10^8\ m/s \tag{3.2}$$

$$h'_{BS} = h_{BS} - 1.0\ m \tag{3.3}$$

$$h'_{MS} = h_{MS} - 1.0\ m \tag{3.4}$$

The NLOS path loss, PL_{NLOS}, is given by:

$$PL_{NLOS} = \min\left(PL(d_1, d_2), PL(d_2, d_1)\right);\ 10m < d_1 < 5km,\ w/2 < d_2 < 2km \tag{3.5}$$

$$PL(d_k, d_l) = PL_{LOS}(d_k) + 20 - 12.5n_j + 10n_j\log_{10}(d_l) + 3\log_{10}(f_c/5);$$
$$k, l \in \{1,2\} \tag{3.6}$$

$$n_j = \max\left(2.8 - 0.0024d_k, 1.84\right) \tag{3.7}$$

w is the street width and when $0 < d_2 < w/2$ LOS Path loss is applied. d_1 is the straight line distance of the ABS to the centre of the perpendicular street on which the NLOS MS lies while d_2 is the straight line distance of the MS to the centre of the street which the ABS is located as shown in Figure 3.3.

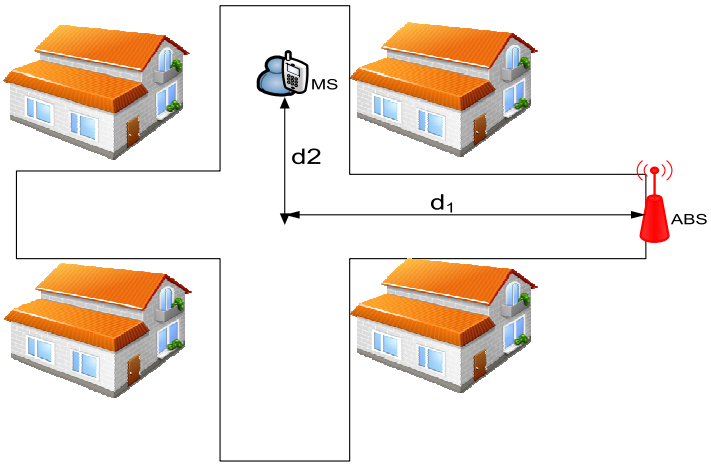


Figure 3.3 NLOS Path loss

The shadowing can be modelled as a log-normal distribution with a mean of zero and a specific standard deviation as follows [114]:

$$S = 10^{\varepsilon/10} \tag{3.8}$$

where ε is a Gaussian random variable with mean 0 and standard deviation of 3-12 dB. In [127], the standard deviation is 3dB for the LOS paths but has a value of 4dB for the NLOS paths. The effective signal at the receiver is obtained by accounting for the gains of the MS and ABS antennas, shadowing, path loss, noise and interference from other users using similar frequency channels.

3.3.3 Traffic Model

Traffic models are used to describe and predict the different services available on a wireless network [128]. They are usually characterised by the inter-arrival times between different user entities (calls, packets, connections, files etc.) in the system and the length of time these entities utilise the resources in the system. The inter-arrival time between user entities determine how frequently the system resources are requested; while, the length of time the entities stay in the system is determined explicitly by the call duration for voice calls, but for data traffic this depends on the file size.

Data traffic can be modelled at the packet, flow or session level. According to [113], a flow refers to a series of packets associated with a particular type of document such as a file or a music track, while a session refers to a series of flow associated with a particular individual. As stated earlier in Chapter 2, users' perception of network performance is reflected by flow and session behaviour rather than packets, and performance evaluation is better done by flow level modelling [112]. Hence, flow level modelling is adopted in this work.

The user entity considered is the file generated by an MS and transmitted to an ABS. In addition, a fixed file size is assumed for all MSs. MSs are uniformly distributed over the entire coverage area and each MS transmits one file at a time. User arrival into the system is modelled by a Poisson process, where the inter-arrival times between users are exponentially distributed.

For the Poisson process, if τ is a random variable that represents the first file arrival in the system after an arbitrary time *x* and the arrival rate of files is represented by λ, the inter-arrival time distribution is characterised by the exponential cumulative distribution function as follows [129]:

$$P(\tau \leq x) = 1 - e^{-\lambda x} \tag{3.9}$$

In this work, a user arriving into the system is served if the SINR threshold is satisfied.

3.3.4 Power Models

The power consumption of the base stations are calculated using power models. Six power models have been considered in this study in order to understand the impact of power model assumption on energy saving. The outcome of this evaluation is presented in Chapter 5. Five out of the six models are based on the generic linear relationship proposed in [40, 41] and described earlier in Chapter 2 (section 2.3). The sixth model proposed in [31], and described earlier in Chapter 2 also, utilises a larger parameter set in estimating power consumption. This model is utilised in the initial performance evaluation in Chapter 4 but it is adapted to fit with the generic linear format used for the remaining models in Chapter 5 for comparison purpose. These models are discussed in more detail in Chapters 4 and 5.

3.4 Performance Metrics

The performance of the system is evaluated in terms of both QoS and energy efficiency. There is a trade-off between QoS and energy efficiency, such that although high energy efficiency is achieved, it may be achieved at the expense of satisfactory QoS. Emphasis is placed on energy efficiency being achieved with QoS targets being satisfied. The QoS metrics considered are the SINR, blocking probability, delay, and throughput. The energy reduction gain (ERG) described earlier in Chapter 2 (section 2.4) and the effective energy saving are the energy efficiency metrics considered.

3.4.1 Signal to Interference Plus Noise Ratio

The signal to interference plus ratio (SINR) is a fundamental performance metrics in wireless communication systems and it is described as the ratio of the desired received signal power to the sum of the powers of interfering signals and noise power at the receiver [130]. The SINR accounts for the propagation effects (path loss, shadowing and multipath fading) and the gains of the transmitting and receiving antenna. Given that signal power at the transmitter is P_{TX} , the SINR at the receiver is given by [131]:

$$SINR = \frac{G.P_{TX}}{I+N} \tag{3.10}$$

where G is the effective gain accounting for the propagation effects and antenna gains, while I is interference from other users and N is the receiver noise power.

The SINR determines the data rate at which a user (or MS) can transmit to or receiver from the BS. MS requests are served only if their SINR satisfies the SINR threshold condition. The SINR threshold is set to a value higher than the minimum SINR acceptable over the BuNGee network in order to guarantee the quality of on-going traffic when new users are admitted.

The data rates of MS are dependent on the SINR achieved at the ABS. In this work, the data rate, R, is estimated using the Truncated Shannon bound [132] as follows:

$$R = \begin{cases} 0\,; & for\ SINR < SINR_{min} \\ \alpha \log_2(1 + SINR)\,; & for\ SINR_{min} < SINR < SINR_{max} \\ R_{\max}\,; & for\ SINR > SINR_{max} \end{cases} \tag{3.11}$$

α is the attenuation factor, $SINR_{min}$ is the minimum SINR required for reception, and $SINR_{max}$ is the SINR at which the maximum data rate, R_{max}, can be achieved.

3.4.2 Blocking Probability

In conventional telephone system, the blocking probability (or grade of service) is the probability that a call arriving at a switch will be blocked [133]. In a system where calls that cannot be served immediately are put in a queue, this is interpreted as the probability of a call being delayed [134]. In this work, admission control is

used to determine whether a user data request would be served or not. If the SINR achieved by the user exceeds the admission threshold, the user is admitted into the network but it the SINR achieved is lower than the threshold, the user is denied access or blocked. The blocking probability is measured in terms of the file requests that are blocked by the network. Hence, the blocking probability, P_b, is given by:

$$P_b = \frac{N_B}{N_T} \tag{3.12}$$

where N_B is the total number of blocked file transmission requests and N_T is the total number of file requests. Blocking can occur as a result of unavailability of free channels for file transmissions or due to poor channel quality as a result of low SINR on free channels.

3.4.3 Average File Transfer Delay

The delay is another important measure of the QoS of a telecommunication system. Delay generally measures the waiting time before a service is provided. File transfer delay is considered in this work. This is measured as the time between the instance an initial file transfer request is made and the instance the file is received in entirety at the receiver. Queuing delay is not considered, once free resource is available to serve a file request it is processed; otherwise it is blocked and retransmitted at a later time. The retransmission time is assumed to be exponentially distributed with a mean equivalent to the current mean inter-arrival rate. The file transfer delay of all successfully transmitted files is averaged to obtain the **average file transfer delay**. Thus the average file transfer delay, $\overline{D}$, is given as:

$$\overline{D} = \frac{D_T}{N} \tag{3.13}$$

where D_T is the sum of the file transfer delay of all successfully transmitted files and N is total number of successfully transmitted files.

3.4.4 Throughput

The throughput is another measure of QoS in telecommunication systems. It is particularly relevant to data transmission as it measures the rate at which the system delivers data offered to it. The throughput, Thr, is measured herein in terms of the

ratio of the total files successfully delivered, F_T, (measured in bits) to the duration of observation, T, (measured in seconds). Thus the throughput, is given by:

$$Thr = \frac{F_T}{T} \tag{3.14}$$

3.4.5 Energy Reduction Gain

The energy reduction gain (ERG) described earlier in Chapter 2 is used to measure energy efficiency in this work and evaluated according to (2.9). The energy efficiency of the schemes proposed in this work are evaluated relative to a baseline scheme with the objective of serving user requests at the highest SINR available from the small cell BSs in the vicinity. In addition, all BSs are always on whether ZBSs or ABSs. The baseline scheme is, therefore, a high data rate centric scheme rather an energy efficiency centric type.

3.4.6 Effective Energy Saving

The effective energy saving is an energy efficiency metric proposed in this work to measure how well a scheme balances energy efficiency with QoS. It is estimated from the difference between the ERG and the percentage increase in delay of a test scheme relative to the baseline scheme.

If delay degradation (DD) is defined as follows:

$$DD = \frac{\overline{D_t} - \overline{D_b}}{\overline{D_b}} \times 100\% \tag{3.15}$$

where $\overline{D_b}$ and $\overline{D_t}$ are the average delays achieved by the baseline and test schemes respectively under the same system conditions.

Then the effective energy saving (EES) is given by:

$$EES = ERG - DD \tag{3.16}$$

The EES metric effectively offset the energy savings calculated in terms of ERG by the loss of QoS in terms of delay degradation. Hence, with this metric the balance between energy efficiency and QoS can be easily observed with only the plot of the EES. This is unlike the previous energy efficiency metrics which require comparison

of delay or blocking probability graphs with the plot of these energy efficiency metrics to determine the balance between QoS and energy efficiency.

3.5 Overview of Energy Efficient Radio Resource Management and Topology Management Schemes

Different energy efficient RRM schemes are proposed in this work and discussed in Chapters 4 to 7. All the RRM schemes have the same objective of allocating resources in the system in a manner to reduce the number of active ABSs. This is in contrast to the baseline scheme which allocates resources to achieve high data rates. The idle ABSs resulting from the energy efficient allocation of resources by the proposed schemes are switched off to save energy by a TM scheme. The TM scheme utilised is an enhancement of the one proposed in [31] to fit the partially centralised RRM and TM approach followed in this work.

A generic procedure of implementation of the energy efficient RRM and TM in the modified BuNGee Architecture is illustrated by the flow chart in Figure 3.4. The simulation is started with the BSs and MSs positioned in the service area and configured with the required number of antennas and frequency channels. The simulation is an event based type, which implies that it advances only when an event occurs rather than advancing in a continuous time manner (time-based type). The events that take place are user arrivals and user departures. The user arrivals are determined by the inter-arrival time distribution and the frequency of user arrivals depend on the magnitude of the mean inter-arrival time. The mean inter-arrival time is the reciprocal of the mean arrival rate. Variation of offered traffic has been achieved by varying the mean arrival rate.

When a user arrives in the system, a decision is made to allocate resources to it or not depending on the SINR. In addition, the resources of which ABS is utilised depends on the RRM scheme applied. The ABSs and their resources are under the control of the ZBSs. Users depart from the system after they have successfully completed their file transfers. The resources utilised by departed users are then released and cleared to be used by newly arriving users. After each user arrival or departure, attempts are made to switch off or switch on ABSs depending on traffic load and QoS performance. This switching on or off is based on TM rules. These

RRM and TM processes are repeated for several events large enough to guarantee that sufficient statistics are considered to achieve reliable results. The reliability of the results is verified with confidence interval. This verification approach is discussed in detail in the next section. A threshold is defined for the number of events to end the simulation and when this threshold is reached the simulation is ended and the performance metrics are evaluated.

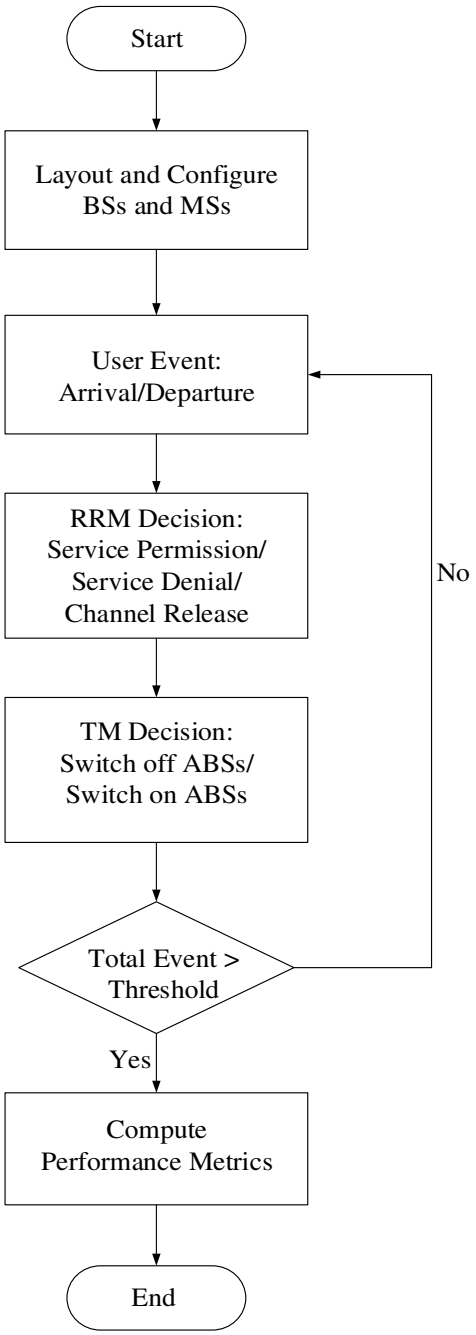


Figure 3.4 Flow Chart of the Simulation

3.6 Verification of Results

A large telecommunication network may be difficult to analyze directly with analytical models because of its complexity; however small sections of the network may be analyzed with these models and the results, possibly, extended to the whole network. The analytical models can serve as performance bounds and can be used to verify the simulation algorithms. The verification can prove that the simulation model is a true representation of the scenario investigated and that the results are reliable. Two approaches for result validation are considered in this work: analytical performance bound derivation and confidence interval estimation.

3.6.1 Analytical Bounds

Queuing theory provides analytical tools for evaluating systems which attempts to serve randomly arriving user requests with limited system resources [135]. It is a quite popular analytical model and has been applied in telecommunication systems as early as 1917 by the Danish mathematician, Erlang, who proposed the Erlang B and Erlang C formulae [129]. In telephone systems, these models provide mathematical relationship between the traffic load offered to the system, the desired grade of service, and the number of channels needed to achieve this grade of service [134].

The Erlang B model assumes that a user's call blocked by the system is lost completely and a reattempt from the same user is seen as a new call [133]. In the work presented in this thesis, the assumption is that users will reattempt their request after waiting for a period of time which follows an exponential distribution with the mean equivalent to the mean arrival rate of users into the system. The Erlang B model is a best case approximation of this approach since blocked calls are cleared and would not contribute to further congestion. It is important to note that although the Erlang B model was originally derived and used in the context of voice calls and telephone channels, it is relevant to the data service considered in this work when utilised in the light of channel occupancy by transmitted data.

The Erlang B model is used to derive a lower bound on the number of active ABSs required to serve a particular traffic load. However, interference is not considered

with this model but interference has been considered in all the simulations in this thesis to approach the practical conditions in wireless systems as much as possible. Nevertheless, the lower bound derived based on the Erlang B model provides a theoretical limit on the minimum number of active ABSs required for a given traffic load. Further details about the Erlang B lower bound and the derivation are provided in Chapter 7. The derived bound is used to verify the performance of an enhanced version of the proposed RRM and TM schemes which mitigates interference properly across all traffic loads.

3.6.2 Confidence Intervals

The confidence interval estimates a population parameter, such as mean, with a range of values the parameter will most likely fall within at a predefined probability of success referred to as confidence level [136, 137]. Also, a confidence interval usually constitutes a range of value above or below a point estimate of the population parameter. In addition, it provides a means of specifying the level of accuracy that should be expected from the method used for estimation of the population parameter. Estimation of a confidence interval at high confidence level makes it possible to state that the population parameter has been evaluated with a high degree of confidence.

In this work, the blocking probability and average file transfer delay have been evaluated after repeated sampling of the user population in order to obtain parameters that are truly representative of the system considered. The average file transfer delay is analogous to the mean of a population and the confidence interval is estimated for the average file transfer delay based on the assumption of normal distribution of this parameter. This is explained in the following.

When the sample size is large enough, usually greater than 30, the point estimate of mean given by $\bar{X}$ is an approximate normal distribution with mean, μ, and variance, σ^2/n [138]. The confidence interval estimation, c, for such large sample case at a confidence level of $(1-\alpha)$, $0 \leq \alpha \leq 1$, is given by [138]:

$$c = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{3.17}$$

where $\bar{x}$ is a particular point estimate of the mean, n is the size of the sample and σ is the variance of the population which can be approximated with the variance of the collected sample without significant loss of accuracy [138]. $z_{\alpha/2}$ is the upper 100 $(\alpha/2)$ percent point of the standard normal distribution [136]. The standard normal distribution is obtained by normalizing the mean random variable, $\bar{X}$.

In the case where the outcome of an evaluation (or experiment) has a binary outcome e.g. failure or success, the probability of success based on several trials can be modelled with a binomial distribution. The binomial distribution is characterised by the probability of success, p, and the number of trials or samples considered, n [139]. The blocking probability is assumed to have a binomial distribution and confidence interval estimation for this case is as follows.

The confidence interval estimate, c, of p for a large sample case at a confidence level of $(1-\alpha)$ is given by [138]:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{3.18}$$

where $\hat{p}$ is a point estimate of the population parameter, p. $z_{\alpha/2}$ varies across confidence levels, high confidence value ranging from 90% to 99.9% can be used to provide high degree of accuracy of estimation of population parameters. The values of $z_{\alpha/2}$ for difference confidence levels are shown in Table 3.1. The confidence interval has been evaluated at 99% for the delay in Chapter 4 to validate the reliability of the results. The confidence interval has also been used in a novel way to develop an adaptive RRM and TM scheme in Chapter 6. As a result, confidence interval and its novel application are explained in greater detail in Chapter 6.

Table 3.1 Confidence Interval Parameters

Confidence Level	90%	95%	99%	99.9%
$\boldsymbol{z_{\alpha/2}}$	1.65	1.96	2.58	3.29

3.7 Conclusion

In this chapter, the system modelling and the performance evaluation techniques utilised in this work are discussed. A simulation model is chosen as the major tool for modelling the system and evaluating performance on the modified BuNGee Architecture considered in this work due to the complexity associated with analytical modelling. The choice of MATLAB as the programming language for the simulation model is as a result of its flexibility, ease of coding and rich database of built-in mathematical and graphical tools. Furthermore, blocking probability, average file transfer delay and throughput are considered as the QoS performance metrics for evaluation in the subsequent chapters. The energy efficiency metrics are energy reduction gain (ERG) and effective energy saving (EES). Finally, the derivation of a performance bound and confidence interval estimation as the approaches used to validate simulation results are also presented.

Chapter 4. Clustering and Interference Mitigation for Energy Saving in a Separation Architecture

4.1 Introduction

Significant growth in mobile subscribers and increasing demand for mobile traffic with the consequent great upsurge in energy consumption [7] requires future wireless networks to be ultra-high capacity and energy efficient. The Separation Architecture is a framework that can meet the high capacity demand and energy efficiency goals prescribed for future wireless networks. Hence, as mentioned earlier in Chapter 3, the separation architecture is utilised for the access network tier of the modified BuNGee network studied in this work. QoS Aware, Energy Efficiency in the separation architecture based BuNGee Network is the focus of the studies in this chapter and the next three chapters.

The separation architecture is based on the concept of separation of the data and control planes, with high power macro BSs handling the control while low power small cell BSs serve user data only [21, 22] . As a result, the high power BSs provide coverage of the service, while the low power BSs meet the capacity needs. Therefore, at low traffic load most low power BSs can be deactivated without creation of coverage holes in the network. In addition to handling coverage, the macro BSs can be configured to handle low-data rate user requests; while small cell BSs handle high-data rate requests [20, 23]. In a network of macro BSs and dense small cell BSs, overhead signalling (including control and reference signal transmission) represents a significant amount of the overall traffic if each small cell BS transmits its own overhead signals [17]. This constitutes high energy consumption since up to 20% of the maximum transmission power may be used to transmit overheads [18]. However, the separation architecture can significantly reduce the overhead signalling as well as optimise the resource utilization and improve energy efficiency [140]. The overhead signalling reduction is explained in detail in the next chapter.

It is proposed in this work that significant energy savings can be achieved in future wireless network through partially centralised Radio Resource Management (RRM) and Topology Management (TM). This implies that nodes of different hierarchy in

the network share the tasks of RRM and TM in such a manner that the system can benefit from the gains of a highly central node without being paralyzed when the node fails. This is explained in more detail in a later section of this chapter.

In this chapter, three RRM schemes are proposed and combined with a TM scheme, an enhancement of the TM scheme in [31] implemented in this study, to reduce energy consumption in the access network of the modified BuNGee Architecture. The RRM schemes are the Normalized Clustering Capability Rating (NCCR), Controllable Quality Clustering Capability Rating (CQ-CCR), and Interference Aware Clustering Capability Rating (IA-CCR). The NCCR scheme clusters or concentrates Mobile Stations (MSs) on to a few active small cells (i.e. access base stations (ABSs)) so that energy saving can be achieved by switching off idle ABSs. The clustering process involves MSs connecting to distant ABSs to reduce the number of active ABSs. Only the traffic load and location information of ABSs are considered under the NCCR scheme for the clustering process. The CQ-CCR scheme clusters MSs as well, however, the relative magnitudes of the SINRs of ABSs are considered in addition to the traffic load and location information of ABSs to improve QoS.

However, inter-cell interference among small cells may lead to power wastage especially under high traffic load conditions if not mitigated. This is because in the bid to cluster MSs to a few ABSs, MSs clustered to distant ABSs may cause high interference to closer ABSs (which could not be turned off or recently turned on). Thus such interfered ABSs may serve their associated MSs at low signal to interference plus noise ratios (SINRs), taking a longer duration to serve user data and consuming more power. If this condition is prevalent network wide, the energy saving gains of MS clustering may be lost. It is shown in this chapter that by allowing MSs to be served by ABSs of high order choices but low SINR value rather than lower order choices and higher SINR value, higher inter-cell interference among small cells is introduced. Hence, the IA-CCR scheme which mitigates inter-cell interference by limiting the choices of ABSs permitted for MSs is proposed. This scheme is an enhancement of the NCCR scheme.

The switching off of idle ABSs resulting from any of the RRM schemes is handled by the enhanced TM scheme, which also switches sleeping ABSs back on to support active ABSs if traffic conditions demand. The energy efficiency and QoS performance of the combination of each RRM scheme and the enhanced TM scheme is compared with schemes previously applied in the literature on the BuNGee Architecture. The previous schemes include the Highest SINR [31, 80], Highest SINR with One Neighbour On [31] and the Capacity Based Channel Assignment scheme [81]. These schemes are explained in more detail later in this chapter.

In this initial study, the highest node (central node) among the RRM nodes only sets static values (policies) for QoS enhancements. Adaptive behavior of this node and the processes leading to the adaptation are presented in Chapter 6. Specifically, a novel approach which conforms to the partially centralised paradigm and adaptively modifies policies in accordance with online system performance is studied in Chapter 6. This results in an adaptive scheme presented in that chapter. Furthermore, the power consumption of ABSs associated with overhead signalling is not considered in this chapter; however, overhead signalling contribution is explained and considered in detail in the next chapter. In addition, a single power model is used for estimating power consumption. In contrast, six power models are considered in the next chapter in order to study the impact of the BS generation utilised in the separation architecture on energy saving. This provides a more comprehensive understanding of energy saving in the architecture and additional insights from this study is utilised in Chapter 6 in the development of the adaptive scheme.

The rest of the chapter is organised as follows. The system model, which is based on the BuNGee Architecture introduced earlier in Chapter 3, is described in more detail in section 4.2. This is followed by the description of the first two clustering capability rating based RRM schemes in section 4.3, and the TM scheme in section 4.4. The energy efficiency and QoS performance evaluation of the combined RRM and TM schemes is presented in section 4.5. Next, the relationship between the choice of ABSs and inter-cell interference is examined in section 4.6, while the IA-CCR scheme is proposed in section 4.7. The energy efficiency and QoS performance evaluation of the IA-CCR scheme relative to earlier schemes is presented in section 4.8. Finally, the chapter is concluded in section 4.9.

4.2 System Model

4.2.1 Network Architecture

As mentioned earlier in Chapter 3, the access network of BuNGee, which has been modified in this work to include high power control BSs in each zone (i.e. Zone Base Stations (ZBSs)) is considered. The BuNGee network topology consisting of the ABSs and ZBSs for access, and the BHSSs and HBSs for backhauling is shown in Figure 4.1. Each ABS is co-located and interfaced to a BHSS as shown in Figure 4.2.

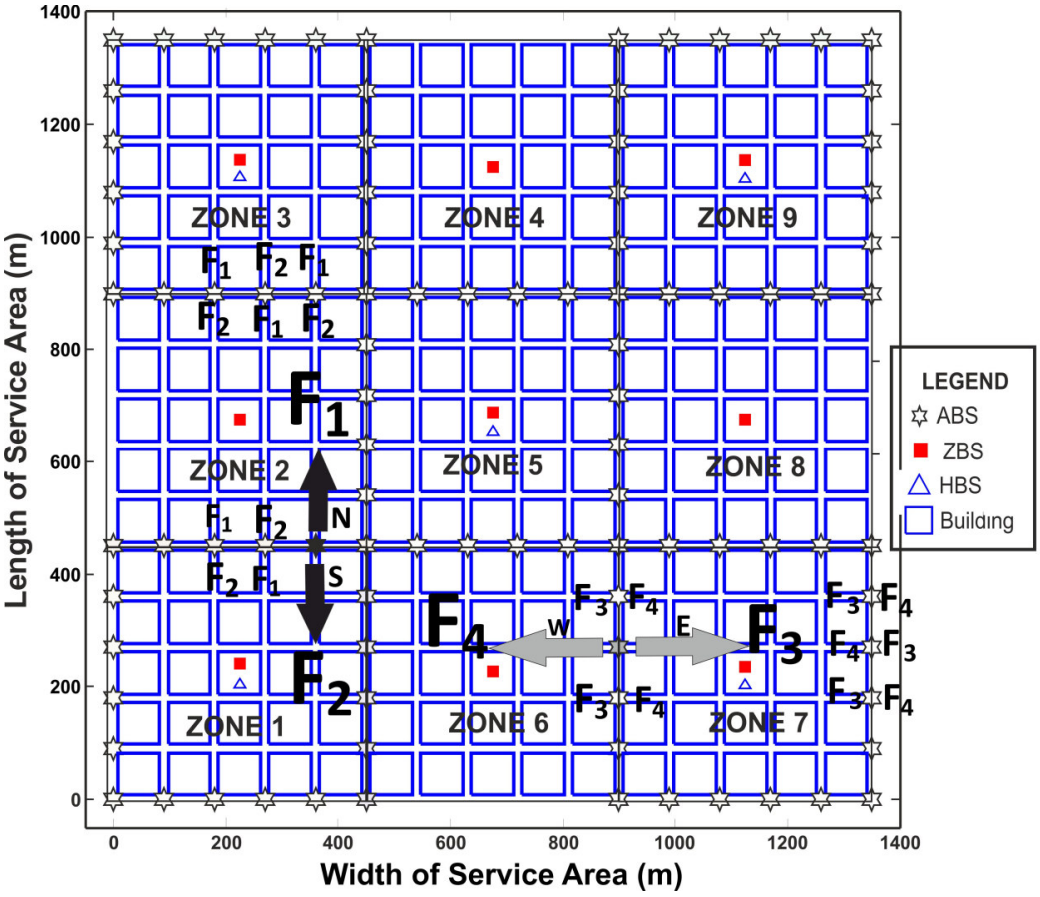


Figure 4.1 BuNGee Topology

As mentioned in Chapter 3, each ZBS co-ordinates the ABSs in its zone to provide broadband access to MSs. MS data is backhauled to HBSs through the BHSSs. Ultra high capacity and reliable backhaul is provided between the HBSs and BHSSs through a combination of in-band and millimetre wave (mmWave) backhauling [141], which will enable low latency communication in the network. It is assumed

that the backhaul frequency band is different from the access frequency band, hence there is no interference between the access and backhaul links.

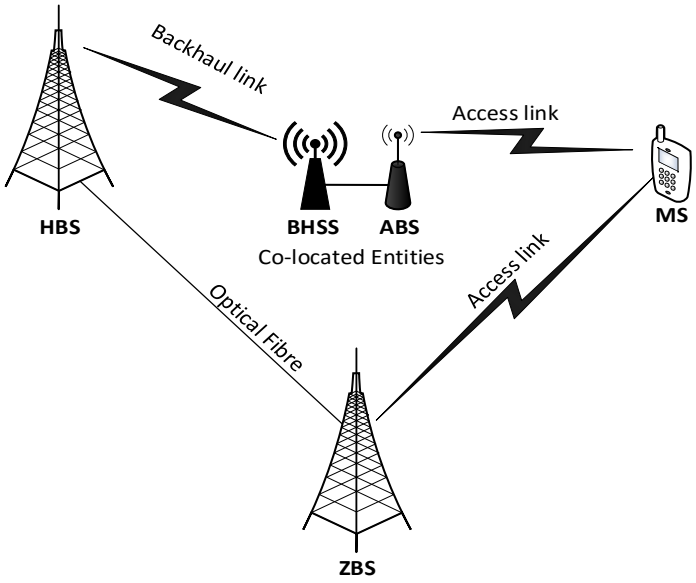


Figure 4.2 BuNGee Access and Backhaul Tier

The fixed frequency plan [125] specified in the BuNGee project is used here. According to this plan, the two antennas of an ABS operate in different frequency bands. Four directions are considered - north, south, east and west - and an ABS can have north and south pointing antenna beams or east and west pointing antenna beams. In Figure 4.1, the four frequency bands in the different directions are shown. ABSs located along the east-west streets have antenna beams in the north and south directions like the ABS shaded in black in the Figure 4.1. ABSs located along the north-south streets have antenna beams in the east and west directions like the ABS shaded in grey. As shown in Figure 4.1, in order to mitigate interference, the antennas of adjacent ABSs facing the same direction operate in different frequency bands and two antennas belonging to different ABSs but pointing along the same street also operate in different frequency bands. Four unique frequency bands are assigned to the small cell layer and each frequency band has a bandwidth of 10 MHz. In this work, each frequency band is further divided into 10 unique subchannels and each subchannel on a particular ABS can be assigned to only one MS. Each MS is assigned only one subchannel at a time for uplink transmission.

The MSs are distributed uniformly outdoors in the service area and each MS is equipped with an omnidirectional antenna with a gain of 0 dBi [80]. The service area is divided into nine square zones as shown in Figure 4.1 and also mentioned earlier in Chapter 3. ABSs can be associated with up to a maximum of four zones. ABSs can only communicate directly with their adjacent neighbours through the co-located BHSSs while they can communicate with the ZBSs through the HBSs. The ZBSs share a 10 MHz frequency band that is out of band to the ABS bands.

In line with the separation architecture paradigm, the ZBSs deployed in the zones are always on to provide universal coverage for the MSs while the ABSs, which can be switched on and off, provide data services. Hence, an MS is always connected to the ZBS in its current zone while it can utilise resources on any ABS in the zone depending on the channel quality and RRM scheme adopted. Although, the ZBS can be configured to serve low data services, the case where only the ABSs provide data services is considered in this chapter and subsequent chapters. Also, the control and data plane separation that makes it possible for separation of universal coverage from data services is explained in detail in the next chapter.

As shown in Figure 4.2, each ZBS is connected to an HBS through optical fibre and information exchange is possible with low delay between a ZBS and an ABS through the backhaul links between a HBS and a BHSS. Whenever a MS has to be served in the DL or UL, the serving ZBS requests the ABSs in its zone to send channel quality measurement in respect of the concerned MS. The channel quality measurement used here is the signal to noise plus interference ratio (SINR). The ZBS will then determine the ABS to serve the MS depending on the objective of the RRM scheme adopted. This co-ordination procedure between the ZBSs, ABSs and MSs before data transmission is similar to the one in [19] and it is illustrated in Figure 4.3 for the UL.

The power consumption at small cell ABSs when they are serving MS uplink traffic is considered in this chapter and subsequent chapters. This is important because the full load downlink power consumption of small cell BSs is of similar order of magnitude as the uplink and no load conditions in existing systems. This is due to the lower share of the power amplifier consumption in small cell BSs [40].

The WINNER II B1 propagation model [127] for Urban micro-cell described in Chapter 3 is used in the system evaluation to determine the path loss and shadowing between an ABS and a MS. This is because ABSs and MSs are deployed outdoors. The uplink data transmission rate, R, is determined from the Truncated Shannon Bound (TSB) [132] as follows:

$$R = \begin{cases} 0\,; & for\ SINR < SINR_{min} \\ \alpha \log_2(1 + SINR)\,; & for\ SINR_{min} < SINR < SINR_{max} \\ R_{\max}\,; & for\ SINR > SINR_{max} \end{cases} \quad (4.1)$$

where α is the attenuation factor, $SINR_{min}$ is the minimum SINR required for reception, $SINR_{max}$ is the SINR at which the maximum throughput, R_{max} can be achieved. The parameters of the BuNGee-specific TSB [80] are α = 0.65, $SINR_{min}$ = 1.8dB, $SINR_{max}$= 21dB and R_{max}=4.5bps/Hz.

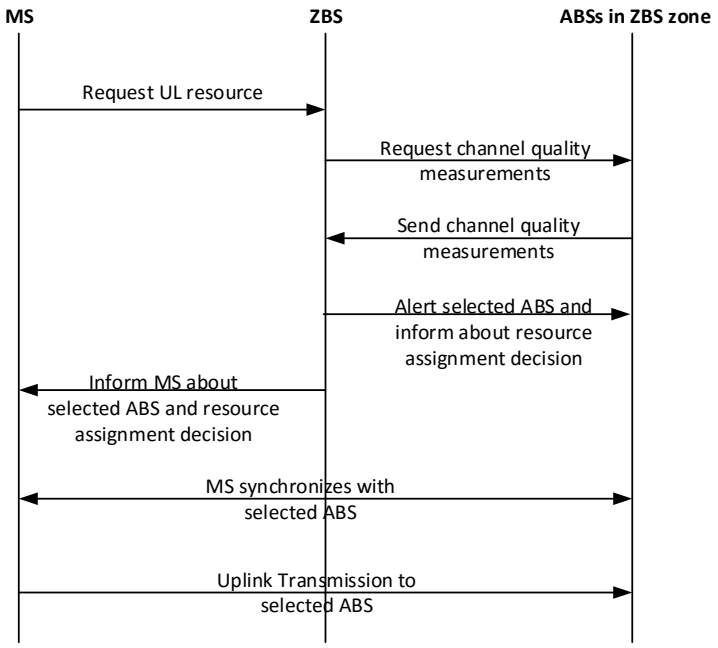


Figure 4.3 Co-ordination Procedure for UL Data Transmission

4.2.2 Power Model

The power model proposed by Han et. al. [31], described earlier in Chapter 2, is used to evaluate the overall energy consumption of the ABSs in the network. The ABS energy consumption, E_{ABS} is given by the following equation:

$$E_{ABS} = \sum_{j=1}^{n_{abs}} \left(t_{sleep,j}\, P_{sleep,j} + t_{Rx,j} \frac{P_{Rx,j}}{\mu_{RF}} + t_{Tx,j} \frac{P_{Tx,j}}{\mu_{RF}} \right.$$
$$\left. + t_{idle,j}\, P_{idle,j} + n_{wakeup,j}\, E_{wakeup} \right) \left(\frac{1}{1-\mu_c} \right) \quad (4.2)$$

where n_{abs} is the number of ABS, P_{sleep} is the power consumed by an ABS when in sleep state. P_{idle} is the power consumed when an ABS is on but not receiving or transmitting, instead it is waiting to serve users. This power is due to the non-radio-frequency components. P_{Rx} and P_{Tx} are the power consumed in the receiving and transmitting states respectively. However, P_{Tx} is assumed to be zero here, because the focus is on the uplink traffic and power consumption due to signalling is not considered. t_{sleep}, t_{idle}, t_{Rx}, and t_{Tx} are the total time the ABS spends in sleep, idle, receiving and transmitting states respectively. However, the time taken to switch from one state to another is assumed to be negligible. μ_{RF} is the efficiency of the power amplifier while μ_c represents the losses in the power supply and battery. n_{wakeup} is the number of times the ABS switches from the sleep state to the idle state. Finally, E_{wakeup} is the energy consumed in the process of waking up the ABS. The values of the different parameters of the equation are given in Table 4.1.

Table 4.1 ABS Energy Consumption Parameters

Parameter	Value
Power in receiving state	5W
Power in sleep state	250mW (assumed 5% of receiving state)
Efficiency of RF	20%
Efficiency of supply loss	10%
ABS max transmit power	5W
Wakeup Energy	50J

The backhaul network tier comprising of HBSs, BHSSs and backhaul links are always active and provide several alternate routes for transferring MS data from/to the access network to/from the core network. Backhaul energy saving is beyond the scope of this work and it is not consider in this chapter and subsequent chapters. The ZBSs are typical macro BSs with maximum transmission power of 40W [131] that operate at 2.6 GHz, while the ABSs operate at a higher frequency of 3.5 GHz. However, the energy consumption of the always on ZBSs is assumed constant because the case of data service support by ABSs alone is considered. The focus is on the energy saving possible with dynamic control of ABS status and the

interference mitigation among the small cell ABS tier of the architecture. Specifically, the goal is to determine the energy saving that can be achieved by switching as many ABSs as possible into sleep state (or off) under a given QoS constraint. It is important to note that switching ABS off is used interchangeably with switching ABSs to sleep state. This should not be confused with turning off the ABS completely such that it consumes no power at all, rather the ABS operates in a sleep state and consumes non-negligible power.

4.3 Energy Efficient Radio Resource Management Schemes

Cellular networks can be operated in an energy efficient manner if MSs can be concentrated, or clustered, on as few as possible BSs and the idle BSs are switched off. This is in contrast to the high data rate approach which spreads MS over as many BSs as available to achieve high capacity and high data rates. Energy efficient RRM schemes based on the MS clustering concept rather than high data rate approach are proposed in this work.

Furthermore, a partially centralised approach is applied in managing resources in this chapter and subsequent chapters. ABSs and ZBSs are involved in the RRM decisions. In addition, it is assumed that the ZBSs are connected through HBSs to a higher node, which provides additional information to enhance the RRM task. However, if this node fails the network functions are still maintained based on the most recent information obtained from the node until the fault is cleared or the node is replaced. Hence, the disadvantage of a single point of failure associated with a fully centralised approach is avoided whilst still benefitting from the global (or wider) view of a central controller. The higher node is termed a Quality Enhancing Processing Unit (QEPU) rather than a central controller or central processing unit since it provides added benefit rather than just being the sole or major decision maker in the network. Hence, mobile services are provided through a hierarchical interconnection of nodes as shown in Figure 4.4.

A similar partially centralised approach is utilised for the TM scheme. Although, the QEPU and the ZBSs are involved in RRM decisions, their energy consumption is not considered as the energy saving benefits studied are the type derived from dynamically switching small cell ABSs on and off. Information such as blocked

requests, ongoing transmissions and successful transmissions about each zone is stored by the ZBS in the zone. This information is readily available at the ZBS in each zone since each ZBS is aware of each MS request and decides whether to permit or block the request based on channel estimation information from the ABSs. Furthermore, the QEPU gathers the information from the ZBSs and provides enhancement parameters to improve network wide QoS. The role of each node is explained in each RRM scheme and the TM scheme.

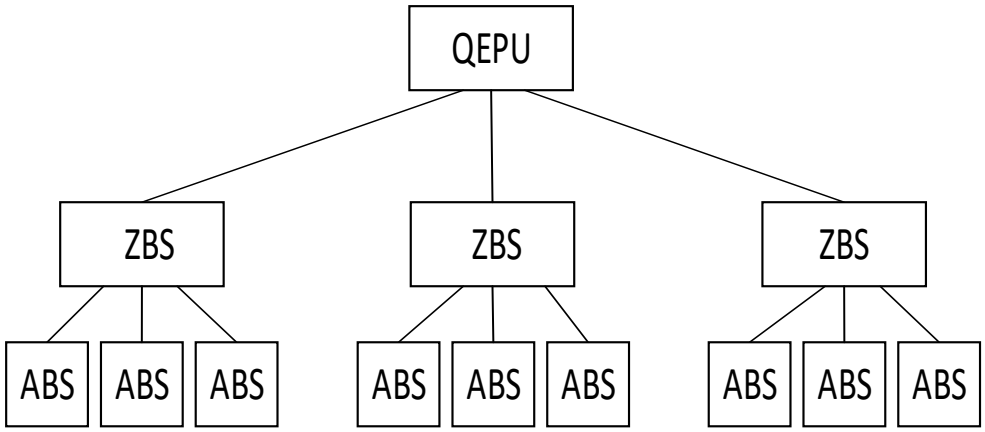


Figure 4.4 Hierarchical Interconnection of Nodes for Resource Management

4.3.1 Normalized Clustering Capability Rating (NCCR) Scheme

This RRM scheme ensures ABSs with more useful coverage area are preferred to serve MSs (i.e. prioritised for MS clustering) over those with a lower useful coverage area. This is because of the potential of the more central ABSs to cluster more users based on their location in the service area. Therefore, more central ABSs can be kept on much longer than ABSs at the edges of the service area.

Specifically, the clustering process is achieved by computing a clustering capability rating (CCR) for each ABS considered as a candidate serving ABS. The CCR value measures the instantaneous capability of an ABS as a good clustering node based on its location and its current load. The CCR value is a linear combination of two ABS location parameters (zone association weight, Z_a, and location weight, L_w) and one ABS load parameter (loading ratio, L_r) and it is evaluated as follows:

$$CCR = aZ_a + bL_w + cL_r \tag{4.3}$$

The zone association weight, Z_a, is obtained by normalizing the number of zones the ABS is associated with by the maximum possible zone association (which is four for the BuNGee Architecture). Location weight, L_w, is obtained by first normalizing the distance of the ABS to the centre of the zone by the distance of the most central ABS to centre of the zone and then finding the reciprocal. The reciprocal is used so that the most central ABS would have the highest value. The loading ratio, L_r, is obtained by normalizing the current ABS load capacity by the maximum possible ABS load capacity. The ABS load capacity is measured in terms of the number of ongoing file transmissions supported by an ABS. The value of each of the ABS parameters can only be between 0 and 1 due to the normalization.

a, *b*, and *c* are constants, and $a = 100$, $b = 10$, and $c = 1$. These constants are used to achieve hierarchical scaling among the ABS parameters. The zone association weight is the strongest indicator of how central an ABS is in the service area and it is assigned the highest hierarchy in the CCR computation. Next in the hierarchy is the location weight and it is a measure of centrality within a zone. The loading parameter is added to ensure resources of an ABS with traffic load are prioritised over equally central ones with no traffic load or lower traffic load. Therefore, a highly central ABS can cluster users from some or all of the zones it is associated with. The final NCCR value is obtained from the normalization of the CCR value by the maximum possible CCR value as follows:

$$NCCR = \frac{CCR}{CCR_{max}} \; ; \; CCR_{max} = 111 \tag{4.4}$$

It is assumed that the uplink SINR for the subchannels at each ABS can be estimated from uplink reference (or pilot) signals transmitted by MSs. Hence, when an MS requests an uplink subchannel, the ZBS requests the highest uplink SINR subchannel magnitude and its position in the frequency band from each ABS in the zone for the MS in question. Only ABSs with a subchannel higher than the call admission SINR need to reply. Finally, the ZBS selects the ABS with the highest NCCR that satisfies the call admission SINR condition to serve the MS. It is important to note that when several ABSs have the same value of NCCR, the ABS with the highest SINR subchannel is selected.

It is also important to note that a different set of values that can maintain the hierarchy among the ABS parameters, other than $a = 100$, $b = 10$, and $c = 1$, can be used as well. Furthermore, the selection of the serving ABS through a three stage process of elimination using the ABS parameters achieves exactly the same result as the approach based on the expression in (4.3). The expression in (4.3) is only a quicker way to achieve this hierarchical decision making.

4.3.2 Controllable Quality Clustering Capability Rating (CQ-CCR) Scheme

In the prioritization of ABSs for clustering, the NCCR scheme focuses mainly on the position and load of ABSs but less on the distance between MSs and ABSs. This would ensure clustering with very few ABSs or equivalently serving MSs with a small number of ABSs and thus, providing opportunity to switch off a lot of ABSs. However, the choice of lower SINR ABSs is encouraged due to the connection of MSs to more distant ABSs rather than closer ABSs with potentially higher SINR. The trade-off of high degree of clustering with lower SINR ABSs for lower degree of clustering with higher SINR ABSs is considered with the introduction of a QoS parameter, termed Quality Factor (QF) into the computation of the CCR. The QF is introduced to control the level of clustering and QoS (since higher SINR ABS choices translate to lower file transfer delays). This results in enhancement of the NCCR scheme and the enhanced scheme is referred to as Controllable Quality Clustering Capability Rating (CQ-CCR).

The QF is obtained as follows:

$$QF = \left(\frac{SINR_{ABS}}{SINR_{Highest}}\right)^{Q} \tag{4.5}$$

where $SINR_{ABS}$ is the SINR of the MS signal at the ABS in question while $SINR_{Highest}$ is the highest SINR among ABSs in the zone. Q is the quality factor power and it is a positive real number. The CQ-CCR value is computed from the product of the NCCR value and the QF for an ABS.

$$CQ\text{-}CCR = NCCR \times QF = \frac{CCR}{CCR_{max}} \times \left(\frac{SINR_{ABS}}{SINR_{Highest}}\right)^{Q} \tag{4.6}$$

It is possible to achieve high QoS and avoid connection to distant ABSs by giving high priority to high SINR ABSs. The CQ-CCR scheme therefore incorporates clustering, proximity and QoS information. In order to emphasise the importance of clustering over QoS, a low value of Q is used; conversely, to emphasise the importance of QoS over clustering a high value of Q is used. In the limit, as $Q \to 0$, the CQ-CCR value approaches the NCCR value. In contrast, as $Q \to \infty$, the NCCR value becomes irrelevant except for the highest SINR ABS. Two values of Q are considered in the system level simulation to demonstrate the clustering and QoS tradeoff. The first value, $Q = 0.01$, introduces the QF into the computation of the CCR but still allows the NCCR value to have significance. The second value, $Q = 10$, deemphasises the NCCR value, but places priority on high SINR ABSs. Q is a network wide parameter set at the QEPU and relayed to ZBSs through the HBSs.

4.4 Topology Management Scheme

The topology management scheme used in this work is a modification of the scheme used in [31] to suit a partially centralised approach to network node and resource management. In [31], a fully distributed approach is used with all TM decisions made by the ABSs without the involvement of a central entity coordinating their actions. In the modified TM scheme used in this work, in addition to the decision making at ABSs, zone level decisions by ZBSs are also included. This is done to harness the higher level zonal observation of performance and control that is not readily available with a fully distributed approach. Furthermore, the QEPU can set network wide TM policies that are passed on to the zone level. Such policies are not considered in this chapter, rather it is considered in Chapter 6. The TM rules utilized under the distributed approach in [31] is explained next. Subsequently, the modification introduced under the partially centralised paradigm applied in this work and the TM rules under this modified TM scheme is discussed.

4.4.1 Existing Distributed Topology Management Approach

In [31], the decision to switch on or switch off an ABS is dependent on the instantaneous traffic load served by the ABS, the average traffic load served by the neighbouring (or adjacent) ABSs and the operating states of these ABS neighbours. Thresholds are set for these traffic load parameters to trigger switching on or off

decisions. Five fully distributed strategies are proposed and the impact of the strategies on system performance is evaluated. C_{s_off} is the traffic load capacity threshold set for an ABS considered for switching off, while C_{n_off} is the threshold set for the neighbours of such an ABS. C_{n_on} is the threshold set for the neighbours of an ABS considered for switching on. The traffic load on an ABS permitted to switch off is handed over to its neighbours. Lists of rules, which contain the conditions under which an ABS can be switched off and switched on, are defined. The lists of rules for switching off and switching on ABSs according to the first strategy are as follows:

ABS switch off rules (that must all be satisfied to switch off an ABS):

1. All ABS neighbours on and
2. The traffic load capacity being served by the ABS itself $< C_{s\_off}$ at time t_{off} when switching off decision is evaluated and
3. The average traffic load capacity of the ABS $< C_{s\_off}$ between time t_{off} and an earlier time t_{off}^{e} and
4. The traffic load capacity being served by all ABS neighbours $< C_{n\_off}$ at time t_{off} and
5. The average traffic load capacities of all ABS neighbours $< C_{n\_off}$ between time t_{off} and an earlier time t_{off}^{e}.

ABS switch on rules (that must all be satisfied to switch on an ABS):

1. The traffic load capacity being served by all ABS neighbours $\geq C_{n\_on}$ at time t_{on} and
2. The average traffic load capacities of all ABS neighbours $\geq C_{n\_on}$ between time *t* and an earlier time t_{on}^{e}.

The monitoring duration for the average traffic load capacities of ABS neighbours, $\Delta t = t_{off} - t_{off}^{e} = t_{on} - t_{on}^{e} = 10s$; while $C_{s\_off} = 30\%$, $C_{n\_off} = 50\%$, and $C_{n\_on} = 90\%$.

Under the second and third strategies C_{s_off} is set to 20% and 40% respectively while all other rules remain unchanged. The fourth strategy requires at least one ABS neighbour to be left on for an ABS to be switched off and C_{s_off} is set to 30%, while all other rules remain unchanged. The fifth strategy does not consider the traffic load and working states of ABS neighbours at all, an ABS local traffic condition alone is considered when switching off ABSs. C_{s_off} is set to 30% as well in this case.

The values of $C_{n\_off} = 50\%$ and $C_{n\_on} = 90\%$ used in [31] are based on the first strategy with $C_{s\_off} = 30\%$ and all ABS neighbours on. In the BuNGee Architecture, as can be seen in Figure 4.1, an ABS has at least two adjacent neighbours that it can forward traffic to or receive traffic from. Hence, if an ABS lies between two ABS neighbours that want to transit into the sleep state, it will have to support its own traffic load and half of the traffic load from its two neighbours transiting to the sleep state. Hence, in the limit the maximum traffic load the ABS needs to support is $C_{n\_off} + \frac{C_{s\_off}}{2} + \frac{C_{s\_off}}{2} = 50\% + \frac{30\%}{2} + \frac{30\%}{2} = 80\%$. This leaves extra 10% to serve newly arriving users before the ABS attains a traffic load capacity of 90% required for switching on ABSs (*ABS switch on rule 1*).

4.4.2 Partially Centralised Based Modified Topology Management Scheme

In the partially centralised based modified TM scheme developed in this work, $C_{n\_off} = 50\%$ and $C_{n\_on} = 90\%$ is retained. However, ABSs are switched off only when they have no load (i.e. idle), thus C_{s_off} is not relevant and no hand over between ABSs is associated with ABS switch off. This leaves extra 40% to serve newly arriving users compared to the 10% in [31] before the $C_{n\_on} = 90\%$ mark.

In this modified scheme, the average load of ABSs and their neighbours are not considered before switching ABSs on or off as done in [31] to reduce delays in the turning on of ABSs. Also, a rule is introduced that requires the ABSs with the highest CCR values in sleep state in a zone to be turned on when blocking in the zone exceeds a threshold. The decision to turn on ABSs as a result of blocking is made by the ZBS, which estimates blocking periodically in the zone. ABSs with the top two no load CCR values in each zone are also kept permanently on to ensure availability of data services at very low load. This is because the ZBS is not used for

data services, rather the ZBS maintains the connection of MSs to the zone and the MSs are free to use resources from any ABS in the zone. In addition, variable monitoring durations are used rather than the fixed duration ($\Delta t = 10s$) irrespective of the traffic load used in [31]. The lists of rules for switching ABSs off and on under the modified TM scheme are as follows:

ABS switch off rules:

1. ABS traffic load capacity = 0 consistently for a period of T_{off} and
2. All ABS neighbour load capacities < C_{n_off}

ABS switch on rules:

1. ABS neighbour load capacity ≥ C_{n_on} or
2. Blocking in zone = N_b in a period, $t \leq T_{on}$

$C_{n\_off} = 50\%$, and $C_{n\_on} = 90\%$ of maximum traffic load capacity of ABSs, C_{max} as in [31] is used. As explained earlier, these values allow for sufficient resources on ABS neighbours to handle traffic when switching off some ABSs.

An idle ABS is switched off (ABS switch off rule 1) if no MS is assigned to it within the time required for at least one neighbour ABS to reach the switch off load threshold, C_{off}, based on the average zonal MS inter-arrival time, T_{int}, in the zone. Thus the waiting time, T_{off}, before switching off is given by:

$$T_{off} = C_{off} \,.\, C_{max} \,.\, T_{int} \tag{4.7}$$

The blocking probability target of 5% is assumed. Specifically, the blocking is counted and ABSs with the highest CCR are switched on if the blocking equals a total of 5 blocked attempts in a duration (based on the average zonal MS inter-arrival time, T_{int}) required to have 100 MS requests or less (ABS switch on rule 2). Thus, $N_b = 5$ and the blocking duration, T_{on} is given by:

$$T_{on} = 99.T_{int} \tag{4.8}$$

This blocking duration measures the expected time on average between the 1st MS request and the 100th MS request.

4.5 Joint Energy Efficient Radio Resource Management and Topology Management Schemes Performance Evaluation

The system model described earlier in section 4.2 is implemented in MATLAB and a Monte Carlo simulation is carried out to evaluate the performance of the schemes. 5 HBSs, 9 ZBSs and 112 ABSs are deployed in a service area of 1.35km by 1.35km and 6,000 MSs (users) are uniformly distributed outdoors along the streets. All users upload single files of a fixed size of 2 MB and the inter-arrival times between users are exponentially distributed constituting a Poisson process. The simulation parameters used are specified in Table 4.2. The simulations for all performance evaluations have been carried out long enough (≥100,000 iterations) to obtain reliable results in this chapter and subsequent chapters. This is validated by the small confidence interval (with all errors $< \pm 0.15s$ for the recorded average file transfer delay values ranging from 3.7s to 7.7s) observed for the delay performance of Figure 4.6 even at a high confidence interval of 99%.

In some previous system performance evaluation on BuNGee [31, 80], the objective of the RRM scheme is to achieve high system capacity and high data rates. Thus, MSs are served by the closest ABSs that can offer them the highest uplink SINR. The RRM scheme is termed "Highest SINR scheme" in this chapter and in subsequent ones. A strategy which has the highest SINR scheme determining the subchannels of which ABSs is used to serve MSs and all ABSs always left on (i.e. no TM) is used as the baseline. Energy savings of other strategies are measured with reference to this baseline. Also, retransmission of blocked requests follows an exponential distribution for all schemes.

Table 4.2 Simulation Parameters [125]

Parameter	Value
Carrier Frequency	3.5GHz
MS Transmit Power	23dBm
ABS Maximum Gain	17dBi
Noise Floor	-114dBm/MHz
Call Admission SINR	10dB
Minimum SINR	1.8dB
Maximum SINR	21dB

In [31], the highest SINR scheme is used in conjunction with the different TM strategies, described earlier, to save energy. The One Neighbour On approach (where at least one adjacent or neighbour ABS must be on for an ABS to be switched off) was regarded as the best TM strategy in terms of balancing energy saving with QoS. The performances of the NCCR and CQ-CCR schemes when combined with the modified TM scheme are compared with the Highest SINR with One Neighbor On scheme. In all cases energy saving is measured relative to the baseline strategy (i.e. highest SINR scheme without TM).

The acceptable range of operation for the system is the region where the blocking probability is less than 5% and the energy reduction gain (ERG) is above zero. ERG [59] is an energy efficiency metric (described in Chapter 2) which measures the energy gains of the new schemes relative to the baseline scheme. The effective energy saving (EES) is another energy efficiency metric, described earlier in Chapter 3, which is considered in this evaluation. The EES shows the performance of the schemes in terms of the balance between the delay and the energy efficiency with a single plot. In this case, the EES is the difference between the ERG of the schemes relative to the highest SINR without TM scheme and the percentage increase in delay also relative to the highest SINR without TM scheme. The energy reduction and delay increase percentages are compared on a one-to-one basis without scaling.

The QoS is evaluated in terms of the blocking probability and the average file transfer delay and is shown in Figure 4.5 and Figure 4.6 respectively. The blocking probability performance of the clustering capability based schemes (NCCR and CQ-CCR) relative to existing schemes is shown in Figure 4.5. The Highest SINR without TM scheme achieves the best blocking probability performance since all the ABSs are available to serve MSs and interference is minimised by MS connection to the closest and highest SINR ABSs rather than the distant and lower SINR ABSs. All the schemes with TM have similar blocking probability below 50 files/s. However, above 50 files/s the blocking probability of the *Highest SINR with One Neighbour On* scheme is significantly higher than that of CQ-CCR (up to 175 files/s with $Q = 0.01$ but in all cases for $Q = 10$) and NCCR (up to 125 files/s) because it puts more ABSs in sleep state than other schemes as shown in Figure 4.7. As the offered traffic load increases, the blocking probability of NCCR increases at a faster

rate than for CQ-CCR even though fewer ABSs are in sleep state than any other scheme above 5 files/s. Blocking in this case can be attributed to relatively high interference resulting in bad SINR on unoccupied channels in the system because MSs do not usually connect to the closest ABSs. Furthermore, the blocking probability of the energy efficient schemes except CQ-CCR at Q = 10 exceed the 5% target before the baseline scheme, highest SINR without TM scheme. Hence, their range of operation is lower compared with the baseline scheme.

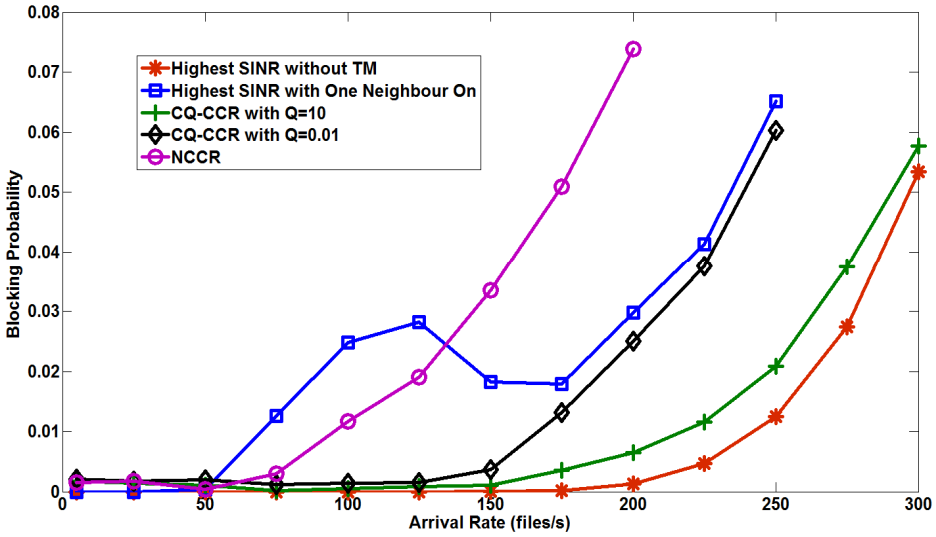


Figure 4.5 Blocking Probability Performance of Clustering Capability Based Schemes Relative to Other Schemes

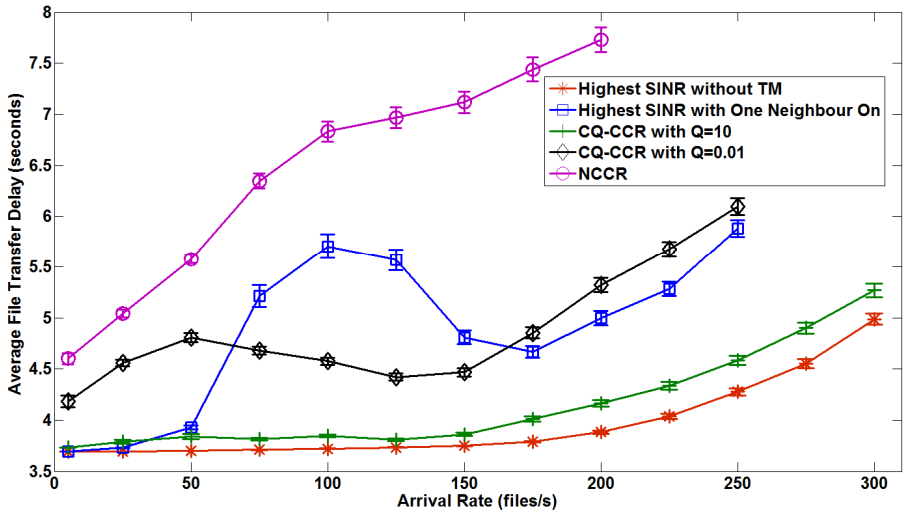


Figure 4.6 Average File Transfer Delay Performance of Clustering Capability Based Schemes Relative to Other Schemes

Figure 4.6 shows that the delay performance of NCCR is the poorest of the schemes considered; this is because relatively low SINR and hence lower data rates are employed for transmission since MSs can connect to distant ABSs. The *Highest SINR with One Neighbour On* scheme has good delay performance at low offered traffic (below 50 files/s) but beyond the low traffic region the performance is poorer compared with the delay performance possible with the CQ-CCR (up to 150 files/s with $Q = 0.01$ and in all cases for $Q = 10$). This is due to the higher number of ABSs in sleep state under the One Neighbour On scheme than the CQ-CCR scheme at higher offered traffic loads. Hence, higher blocking of user requests occur leading to file retransmission and higher file transfer delay. Also, the QoS performance of CQ-CCR scheme is better at the higher Q value, because a high Q value emphasises QoS over clustering. The Highest SINR without TM scheme achieves the best delay performance since the full range of ABSs is available and MSs are served by resources from the closest and highest SINR ABSs. This leads to higher data rates and thus lower delay than the other schemes.

Figure 4.7 and Figure 4.8 show the percentage of ABSs in the sleep and idle states for different schemes respectively. These figures show how the applied TM scheme responds to the behavior of RRM schemes by adapting the number of active ABSs according to traffic load. Generally, it can be seen from Figure 4.7 and Figure 4.8 that the *Highest SINR with One Neighbour On* Scheme clearly keeps more ABSs in the sleep state and few or no ABSs in the idle state, particularly above low offered traffic levels, because of the TM scheme used in [31]. The average load consideration of this TM scheme delays the turning on of ABSs and ensures that once an ABS is turned on it will be used. In contrast, the TM scheme used with NCCR and CQ-CCR promptly turns on ABSs once the neighbour exceeds the capacity threshold or when the blocking threshold is reached in order to ensure good QoS. Hence, NCCR and CQ-CCR have fewer ABSs in sleep state and more ABSs in idle state than the *Highest SINR with One Neighbour On* Scheme. However, this approach is justified by the better blocking probability and delay performance possible with the CQ-CCR scheme.

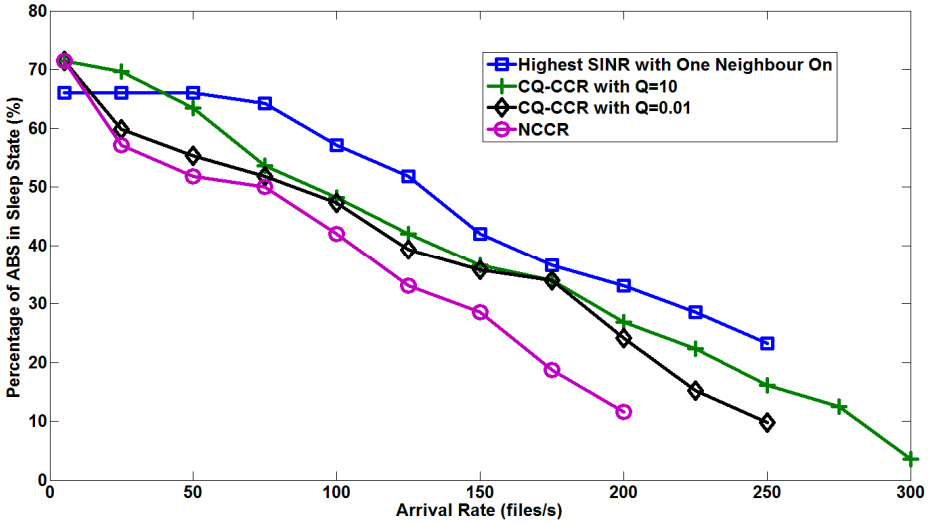


Figure 4.7 Percentage of ABS in Sleep State vs Offered Traffic

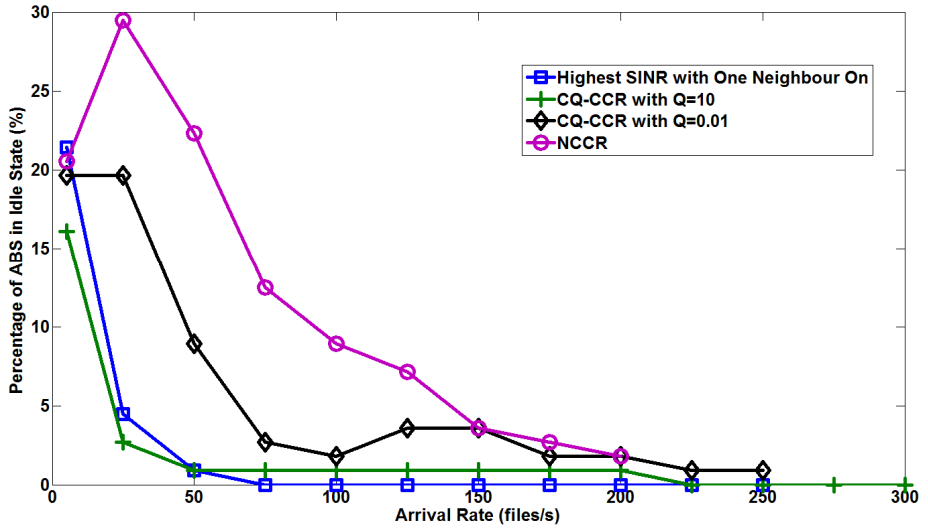


Figure 4.8 Percentage of ABS in Idle State vs Offered Traffic

The energy efficiency performance of the schemes in terms of ERG and EES are presented in Figure 4.9 and Figure 4.10 respectively. The ERG is calculated relative to the energy consumption ratio (ECR) of the Highest SINR scheme without TM which ranges from 11.87µJ/bit at 5files/s to 0.63µJ/bit at 300 files/s. Figure 4.9 shows that the NCCR scheme achieves the highest energy reduction of 67% at 25 files/s but at the expense of relatively high delay. It is important to note that at this traffic load, the NCCR has fewer ABSs in sleep state than other schemes. If energy consumption reduction is dependent only on keeping more ABSs in sleep state, NCCR should have the lowest energy reduction at this point. However, as can be

seen in Figure 4.8, a significant percentage of ABSs (nearly 30%) are in idle state rather than actually receiving traffic. The energy saving is thus not only dependent on the number of ABSs in sleep state but also on the status of the ABSs that are on (i.e. receiving or idle state). The CQ-CCR scheme achieves up to 60% ERG (with $Q = 0.01$) at better QoS than the NCCR. The *Highest SINR with One Neighbour On* Scheme has the best energy reduction performance over the range of traffic load considered.

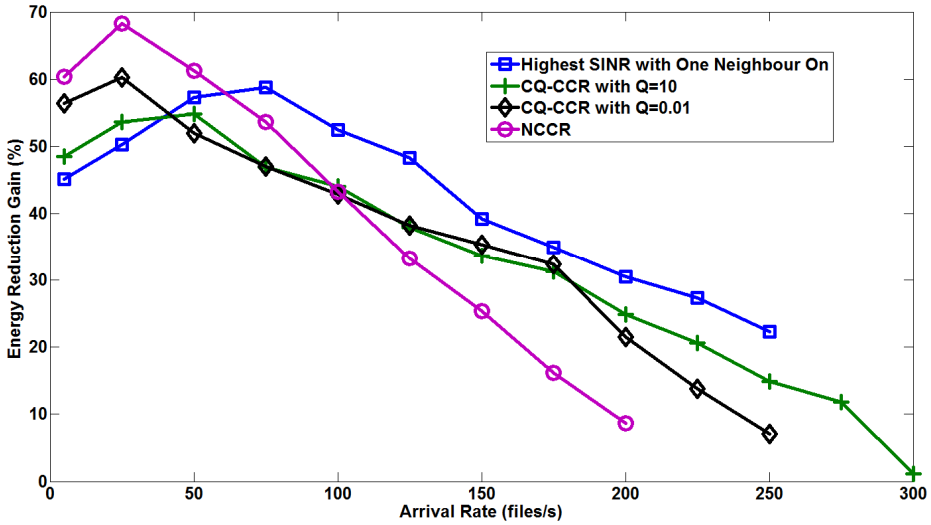


Figure 4.9 Energy Reduction Gain Performance of Clustering Capability Based Schemes Relative to Other Schemes

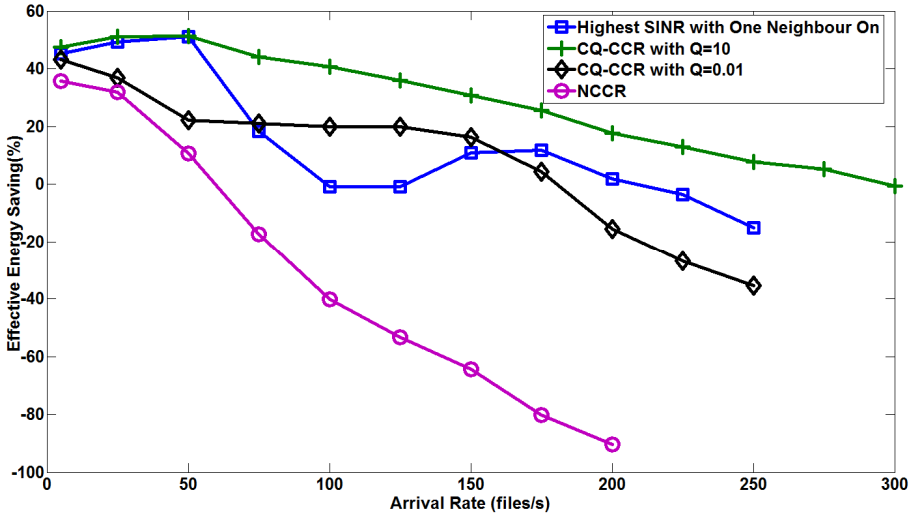


Figure 4.10 Effective Energy Saving of Clustering Capability Based Schemes Relative to Other Schemes

However, a trade-off between delay and energy efficiency (measured in terms of ERG) can be observed by comparing Figure 4.6 and Figure 4.9 respectively. A balance between the delay and energy efficiency is therefore important and the EES is used to estimate this balance. In Figure 4.10, the EES shows that by reducing the clustering with the Quality Factor Power, Q, from a high level (equivalent to $Q = 0$ in NCCR) to a lower level (at $Q = 0.01$ *in CQ-CCR*) and even further lower (at $Q = 10$), better balance can be achieved between the delay and ERG. Furthermore, CQ-CCR can actually provide better balance between QoS and energy saving than the *Highest SINR with One Neighbour On* when a high enough value of Q is chosen as with CQ-CCR with $Q = 10$.

4.6 ABS Choice and Inter-cell Interference

The CQ-CCR scheme regulates the high interference associated with the NCCR scheme with the aid of the quality factor, QF, making it possible for MSs to select higher SINR ABS choices than under the NCCR schemes. This approach is extended further by being more definite about the choices of ABS to permit. For example MSs may be permitted to be served by only the first choice and second choice ABSs in terms of SINR but not third, fourth or higher choices. Firstly, the relationship between the choice of ABSs and inter-cell interference among ABSs is illustrated; then, the Interference Aware Clustering Capability Rating (IA-CCR) scheme, based on this relationship, is proposed. Finally, the relationship between ABS choices and energy efficiency is evaluated by comparing the IA-CCR scheme with the NCCR, Highest SINR and Capacity Based Channel Assignment (CBCA) schemes.

CBCA is an energy efficient scheme proposed for the BuNGee Architecture in [81]. Under the control of CBCA scheme, an MS is served by resources from the ABS with the highest traffic load in its vicinity regardless of its ranking in terms of SINR amongst suitable ABS choices. However, all prospective ABS choices must satisfy the call admission SINR condition. It is important to note that when all ABSs have no load, the MS is served by resources from the ABS with the highest SINR value. The same is true when several ABSs have the highest traffic load.

In full scale dense small cell deployments with high frequency reuse, the proximity of BSs makes it possible for MSs to have several choices of ABSs to connect to. In

Figure 4.11, a snapshot of streets with MSs and ABSs operating at various frequencies is shown to explain the relationship between the base station choices and inter-cell interference in a SCN with fixed frequency plan. As a result of proximity between ABSs, MS1 and MS2 can connect to more than one ABS. However, the SINR achievable at different ABSs will be different due to different path loss conditions. MS1 may have the option of connecting to ABS1, ABS2, ABS3 and ABS6. If it connects to ABS1 which is the closest line of sight ABS it should have the highest SINR possible (i.e. first choice ABS) and the interference to the most interfered ABS (which is ABS3) should be smaller than the interference to ABS6 which is the most interfered one if it connects to ABS2. It is important to note that ABS2 is a lower SINR choice further away than ABS1. This trend can be extended to connection to ABS3 which is even further away and also at lower SINR than ABS2. In this case the interference to ABS1 will be higher than the ABSs in the previous cases. Thus, by allowing MSs to be served by ABSs of high order choices but low SINR value rather than lower order choices and higher SINR value, higher inter-cell interference among small cells is introduced.

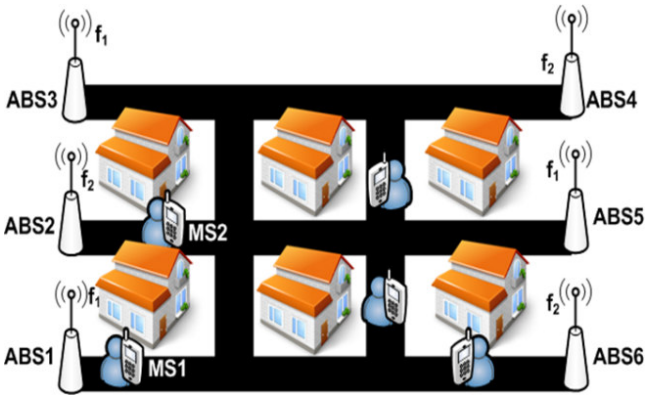


Figure 4.11 BuNGee Streets with ABSs and MSs

The deviation from connection to the closest and highest SINR ABS may be tolerable at low traffic loads as most of the ABSs being interfered with may be switched to the sleep state but beyond the low traffic region these ABSs may be active. The interference already created towards the ABSs can make them unusable in the future as a result of the low SINR choices made earlier on. Intuitively, it is expected that the highest SINR scheme will provide the lowest future interference since MSs always connect to the first choice ABS. On the contrary, only

instantaneous interference is reflected in the decisions made under CBCA and NCCR schemes, but future interference impact of ABS choices is ignored since both schemes permit any ABS choice (which would include very low SINR choices).

On one hand the NCCR and CBCA schemes are strictly energy efficient schemes with the goals of clustering or concentrating MSs with as few as possible ABSs but without the interference mitigation aspect. On the other hand, the highest SINR scheme has a goal of delivering high data rate and avoids the higher inter-cell interference associated with selection of lower SINR choices of ABS (i.e. second, third and higher choices) than the first choice. However, since the highest SINR scheme spreads MS over comparatively larger number of ABSs instead of concentrating on few ABSs, it is not energy efficient. The IA-CCR scheme provides a compromise between the highly energy efficient approach of the NCCR scheme and the high date rate focused approach of the highest SINR scheme. The IA-CCR scheme is presented in detail in the next section.

4.7 Interference Aware Clustering Capability Rating (IA-CCR) Scheme

The IA-CCR scheme mitigates or reduces interference by restricting the choice of ABS for an MS to pre-defined top ranking ABSs in terms of SINR. In addition, energy efficiency is achieved by applying the clustering concept in the final decisions about which ABS an MS connects to among this group of ABSs. This is determined based on the NCCR values of the ABSs. Specifically, ABSs that satisfy the call admission SINR threshold and also fall within the choice range are first selected. Then, the MS connects to the ABS with the highest NCCR value in the subgroup and is served by the highest SINR uplink subchannel on this ABS. The algorithm can be implemented as follows: let L_i and L_{max} represent the current load and maximum load capacity of an ABS i respectively, so that the normalised load, x_i on ABS i is given by:

$$x_i = \frac{L_i}{L_{max}} \tag{4.9}$$

Let $X = [x_i]$ represent the vector of the normalised load of all ABSs, while $S = [h_i]$ denote the vector of the uplink SINR of ABSs in a zone for the MS in question. h_i

represents the highest uplink SINR subchannel of ABS *i*. Let $C = [c_i]$ represent the NCCR vector for all ABSs in the zone, where c_i represents the NCCR value for ABS *i*. Also, let s_{th} represent the call admission SINR threshold. If an n^{th} order restriction is used, then MSs are restricted to connect to the n^{th} choice ABS or lower order choices (i.e. $(n\text{-}1)^{th}$ choice up to the 1^{st} choice). The ABS selection can be determined as follows:

1. All elements of X, S and C are set to zero initially.
2. When MS requests for uplink resource, ZBS requests h_i from all active ABS in the zone.
3. Each ABS verifies the condition: $h_i \geq s_{\text{th}}$ and sends h_i to ZBS only if the condition is satisfied.
4. ZBS updates S with all received h_i, C with c_i for each ABS that responds, and arranges the set of ABSs in descending order of SINR (h_i). However, if no ABS responds, the MS is blocked.
5. If the total number of ABSs in the set is *m*, ZBS will instruct the MS to connect to an ABS with highest c_i value among the set of ABS, if $m \leq n$. However, if $m > n$ ZBS selects the highest ranking *n* ABSs based on h_i, and instruct the MS to connect to the ABS with highest c_i value in the high ranking ABS subset.
6. ZBS updates X to account for the increase in traffic load and resets all the elements of S and C in preparation for a new MS request. Step 1 is not repeated after the first MS request.

It is assumed that the restriction rule (i.e. the value of *n*) is set at the QEPU and made available to the ZBSs via the HBSs. The IA-CCR scheme is evaluated with first, second and third choice restrictions, and the modified TM is applied to turn off idle BSs and turn on neighbours of overloaded ABSs. These modified TM is also applied with the CBCA. The QoS and energy efficiency of the IA-CCR scheme is compared with the previous NCCR scheme, the CBCA and the Highest SINR without TM scheme. The energy savings of the IA-CCR, and CBCA are measured relative to the Highest SINR scheme without TM as before.

4.8 Performance Evaluation of the IA-CCR Scheme

In the same vein as the evaluation of the NCCR and CQ-CCR schemes, the IA-CCR and CBCA schemes are implemented in MATLAB and evaluated over the system model with the same system and user configurations as in the previous evaluation. The simulation parameters in Table 4.2 are used as in the previous case. The performance of the NCCR and Highest SINR without TM schemes in the earlier evaluation is included for the purpose of comparison.

The QoS performance of the schemes is evaluated in terms of the blocking probability and average file transfer delay and presented in Figure 4.12 and Figure 4.13 respectively. In Figure 4.12, the blocking probability performance at low offered traffic (below 100 files/s) is similarly low for all the schemes. This is because few users are active simultaneously and interference is low, hence most ABS channels have suitable SINR values to admit users. However, as the traffic load increases beyond 100 files/s, the blocking probability of the NCCR and CBCA schemes rise at faster rates than the highest SINR without TM scheme and the IA-CCR scheme with different choice restrictions (i.e. at first, second and third choice restrictions).

Furthermore, higher order choice restrictions with permission of lower SINR choices have increasingly higher blocking probabilities than lower order polices permitting higher SINR ABS choices. For example, the blocking probability of the third (3^{rd}) choice restriction is worse than the second (2^{nd}) choice. This trend can be attributed to higher interference introduced when lower choices of SINR are used over and over again beyond low traffic load conditions. At this stage, a lot of users are active at the same time and more ABSs have to be active to serve users unlike under low traffic conditions. The highest SINR without TM performance is the best since all ABSs are on at all times and first choice SINR keeps interference low.

The delay performance shown in Figure 4.13 follows a very similar trend as the blocking probability. The schemes without choice restrictions (i.e. NCCR and CBCA) have poorer delay performance than the IA-CCR scheme. This is because with the IA-CCR scheme, ABSs with higher SINR are the preferred choices relative to NCCR and CBCA and higher SINR results in higher data rates and lower file

transfer delay. The CBCA scheme performs slightly better than the NCCR in terms of blocking probability and delay at high traffic load because of the selection of highest SINR ABSs when all suitable ABSs have no load unlike the NCCR scheme which selects the highest CCR which may not be the highest SINR choice. Also, the delay increases with the order of choice restriction used; hence, the lower the order of choice restriction used the better the delay performance.

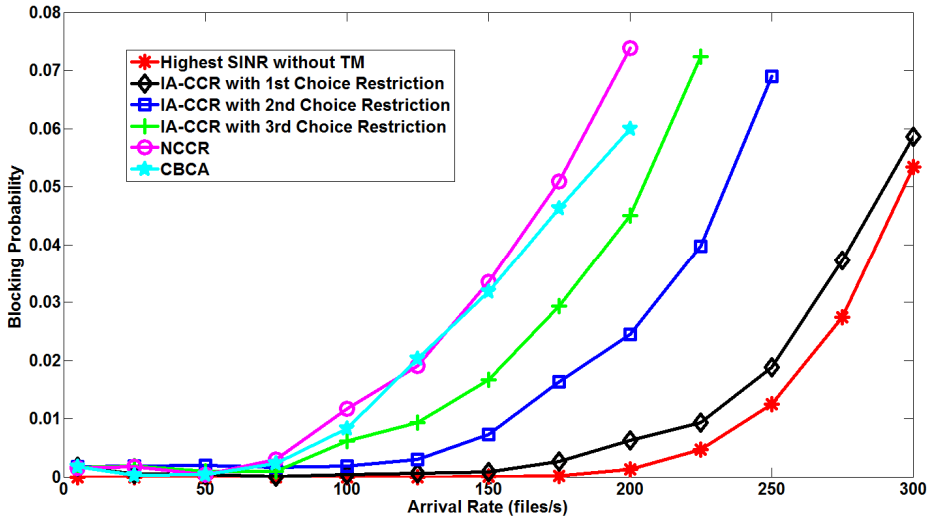


Figure 4.12 Blocking Probability Performance of IA-CCR Scheme Relative to Other Schemes

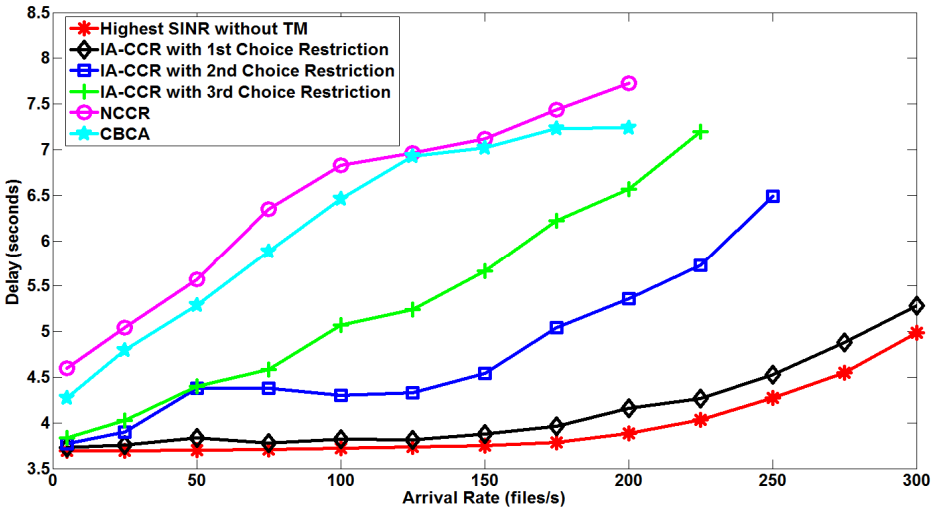


Figure 4.13 Average File Transfer Delay Performance of IA-CCR Scheme Relative to Other Schemes

The energy efficiency performance is evaluated in terms of ERG and EES and presented in Figure 4.14 and Figure 4.15 respectively. In Figure 4.14, at low traffic load (below 50 files/s) the CBCA scheme has the highest ERG (69%) among the schemes. It is closely followed by the NCCR scheme which has ERG of up to 67%, while the proposed scheme, IA-CCR, achieves energy reduction of up to 65% with the third choice restriction. This trend is experienced because as there is change from any choice to increasing higher SINR choice requirements and lower order choice restrictions, the high degree of clustering or concentration of mobile users is traded for interference reduction through limitation of distant MS to ABS connection and emphasis on higher SINR for connection. At low traffic loads, the higher the SINR choice requirement (or equivalently the lower the order of choice restriction), the lower the ERG.

On the contrary, for medium and high traffic loads, more ABSs have to be turned on to serve MSs and the effect of interference becomes more prevalent. Thus, in Figure 4.14 beyond the low traffic regions, the ERG of higher SINR choice schemes improves relative to lower ones so that at some point the third choice IA-CCR becomes more energy efficient than the NCCR and CBCA schemes, then the second choice restriction gets better than the third choice restriction. Finally, at even higher offered traffic, the first choice is better than all other schemes and energy efficiency performance is in the order of choice restrictions. This trend is as result of interference getting more and more severe as the traffic load increases for the schemes; so that eventually the first choice scheme with the best inter-cell interference mitigation supersedes the other schemes at high offered traffic where interference in the system will be highest.

Figure 4.15 shows that for low traffic load (below 50 files/s) the EES for all the IA-CCR policies have close performance in terms of balancing QoS and energy saving. In addition, their performance is better than both the NCCR and CBCA schemes in this region and beyond. This is as result of the poorer delay performance of the NCCR and Most Loaded Scheme as seen in Figure 4.13. Beyond the low traffic region, the EES performance of the third choice restriction becomes increasingly poorer relative to the second and first choices. In a similar manner, the EES performance of the second choice IA-CCR which was initially close to the first

choice policy also gets poorer and the first choice policy is clearly the best option at medium and high traffic load region. Since, the third and second choice IA-CCR policies have better ERG at low load than the first choice restriction; they will be better choices in this region.

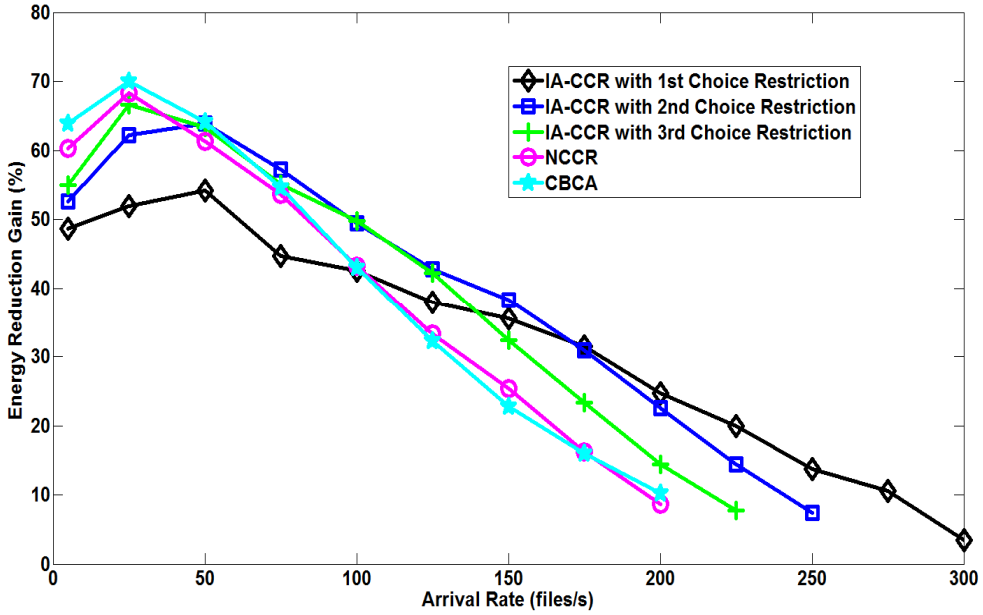


Figure 4.14 Energy Reduction Gain Performance of IA-CCR Scheme Relative to Other Schemes

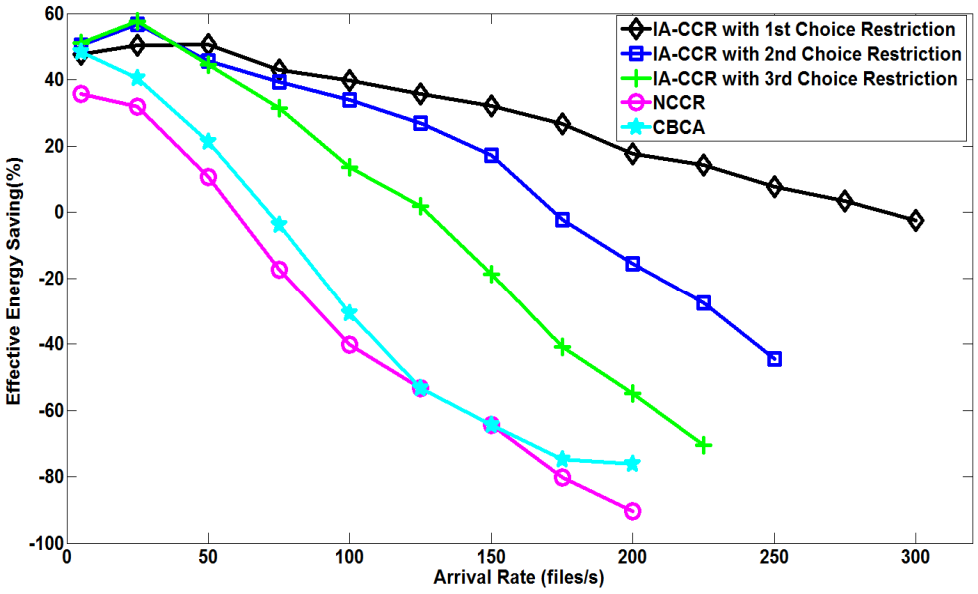


Figure 4.15 Effective Energy Saving Performance of IA-CCR Scheme Relative to Other Schemes

It is therefore clear that there is a need to adapt the choice restriction in accordance with the offered traffic load in order to both save energy and balance energy saving with QoS. Online dynamic adaptation of choice restriction to traffic load is examined in more detail in Chapter 6. The NCCR may not be suitable like the IA-CCR scheme for delay sensitive applications, however it can still be utilised for applications without stringent delay requirements at low traffic loads where they provide higher energy savings.

4.9 Conclusion

In this chapter, it is shown that energy savings can be achieved in a dense small cell based separation architecture through a partially centralised control paradigm, where resource and topology management decisions are shared among different hierarchy of network nodes. Specifically, more central ABSs are prioritised over their less central counterpart by using the novel NCCR algorithm for resource management. This algorithm clusters or concentrates MSs around few ABSs and provides the opportunity to switch off idle ABSs through TM schemes. Also, the benefit of reducing the high degree of clustering of NCCR is explored by the introduction of a QoS parameter into the computation of the CCR. This QoS parameter makes it possible to achieve better balance between the energy saving and QoS as demonstrated with the CQ-CCR scheme.

Furthermore, clustering management possible with the CQ-CCR scheme is implemented with more precision by restricting MSs to predefined ABS choices. It is shown that selection of ABSs that are higher order choices with lower SINR leads to higher interference in the system than selection of ABSs that are lower order choices but with higher SINR. Thus, an enhancement of the NCCR, IA-CCR, is proposed which restricts the choice of ABSs to only top ranking ABSs in terms of SINR and performs NCCR based clustering only with permitted choices. It is shown that a high order choice restriction (with permission of low SINR choices) can lead to higher energy saving and acceptable QoS at low traffic load than a lower order choice restriction (with permission of higher SINR choices). However, beyond low traffic load region, the lower order choice restriction outperforms the high order choice restriction with respect to both energy saving and QoS. Thus, it is necessary to adapt

choice restriction in accordance with traffic load. This adaptation of policy is investigated in Chapter 6 whilst still maintaining the partially centralised approach to energy efficient RRM.

All proposed schemes are able to achieve significant energy savings with the NCCR achieving up to 67% ERG but has poor QoS performance. The introduction of clustering management in both the CQ-CCR and IA-CCR schemes, makes it possible to achieve energy savings and maintain good QoS. The CQ-CCR and IA-CCR are shown to achieve ERG of up to 60% and 65% respectively. Energy saving of these schemes is based on a single power model, more comprehensive understanding of energy savings for the separation architecture is provided in Chapter 5 where six power models are considered. This provides insights about the impact of BS generation on energy saving and the knowledge obtained is applied in the development of the adaptive scheme in Chapter 6.

Chapter 5. Energy Saving in a Separation Architecture under Different Power Model Assumptions

5.1 Introduction

The potential energy savings at the BSs of a wireless network are dependent to a large extent on the length of time the BSs can be made to operate in lower power consumption states rather than higher power consumption states. Furthermore, the magnitude of the energy saving depends on the difference between the higher power consumption states and lower power consumption states of the BSs studied. Therefore, since the power consumption of a state would vary from one BS generation to another, the energy saving of a wireless architecture, including the separation architecture, would change if the BSs were changed to a newer generation. Hence, energy savings would also vary with different power models specified for different BS generations in a separation architecture.

In this chapter, the impact of the choice of power models on energy saving is studied for the separation architecture based modified BuNGee Network (presented in Chapter 4). Unlike in the previous chapter where energy saving was evaluated with one power model (Han model [31]) and the power used for signalling is neglected, six power models are considered and signalling power is also put into consideration in this chapter. This sort of study is important because it facilitates the understanding of why and how energy saving varies with respect to the power model assumptions. This knowledge can aid the design of suitable energy saving strategies or the enhancement of existing ones. The insight gained from this study, for example, is applied in the development of the adaptive IA-CCR scheme presented in the next chapter, which is an enhancement of the IA-CCR scheme proposed in Chapter 4.

Specifically, the following questions are considered for the separation architecture:

1. Is there significant benefit, with regard to energy saving, in operating small cell BSs in the idle state and the sleep state for each power model?
2. What is responsible for the differences in energy saving across these models?
3. How does the energy saving of these models compare with one another?

These questions are addressed by the development of a framework for evaluating the power saving achievable when BSs are operated in lower power consumption states rather than high power consumption states (termed Low Power State Saving (LPSS)). The framework comprises generic equations derived for estimating LPSS over very short timescales for both single and multiple BS scenarios. The short timescale LPSS provides the basis for evaluating energy saving over a long timescale. Furthermore, a small scale separation architecture model comprising of a high power BS and four small cell BSs is also included in the framework to swiftly identify the BS state changes in the small cell layer that contribute significantly to energy savings for each power model. Finally, system level simulation is included to evaluate energy saving and QoS performance of the IA-CCR scheme under consideration of the different power models in the full scale architecture.

The remainder of this chapter is organised as follows. The system model used is discussed in section 5.2, the control and data plane separation that facilitates overhead reduction is described in section 5.3, while the LPSS framework is presented in section 5.4. The simulation results are presented and discussed in section 5.5 and the chapter is concluded in section 5.6.

5.2 System Model

The modified BuNGee Architecture based on the separation architecture paradigm described and evaluated for energy saving in Chapter 4 is also the one considered in this chapter. Therefore, the positions and functions of the ABSs, HBSs, ZBSs, BHSSs, and QEPU remain the same as before. The MSs are distributed outdoors as well in this case, therefore the Winner II B1 propagation model for Urban Micro-cell is still used for modelling the channel between an ABS and MS. Also the data rates over the channels are evaluated with the Truncated Shannon Bound (TSB) as before.

Furthermore, as in Chapter 4, the ZBSs are used only for control signal transmission in this Chapter and are always kept on to ensure network wide coverage. Hence, the energy saving considered is the type that can be achieved in the access network by switching some of the ABSs into low power consumption states when uplink data transmissions from MSs are being served. The power consumption is evaluated for

the ABSs under consideration of six different power models which are presented in a later section.

Unlike in Chapter 4, the control signalling contribution is considered in the computation of the power consumption of the ABSs in this chapter. The control and data plane separation that results in the reduction of signaling overhead transmission by ABSs is discussed in the next section. Detailed power consumption equations for the separation architecture with consideration of the signalling power and effect of reduction in signaling overhead is provided in a subsequent section.

5.3 Control and Data Plane Separation

There are some initial procedures that a freshly switched on mobile device needs to perform before it can begin to use resources in a cellular network. These procedures constitute the initial attachment in cellular systems. In LTE, for example, it begins with the time and frequency synchronization of the MS with a cell and decoding of the cell identity through the detection and utilization of synchronization signals transmitted by the BS (or cell) [142] . Once the MS is synchronised with the cell, it will then access the Master Information Block (MIB) transmitted by that cell [143]. The MIB contains essential information for initial access to a cell [142] such as downlink system bandwidth and the configuration of the physical hybrid automatic request indicator channel (PHICH in LTE) [144]. The PHICH is used to transmit the Hybrid Automatic Request (HARQ) indicator described later. Reference signals (or pilot signals) are also detected by the MS and used to perform the received signal strength measurement for the cell and the decision to select and camp on the cell or not is based on this measurement [142] .

Apart from the synchronization signals (Primary Synchronization Signal (PSS) and Secondary Synchronization Signal (SSS) in LTE) and reference signals (Cell-Specific Reference Signal (CRS) in LTE), some downlink control signals are also transmitted by the cell to facilitate effective data communication. Using LTE as an example, the Downlink Control Information (DCI) provides information about the uplink and downlink resources specified for MSs, Control Format Indicator (CFI) specifies the number of OFDM symbols used for relaying DCI, while the HARQ

acknowledge (ACK) / negative acknowledge (NACK) Indicator (HI) signal is used to indicate whether uplink data transmission is correctly received or not [144, 145].

In addition to its use in cell selection, the reference (or pilot) signal is also used by the MS for downlink channel quality estimation and detection of downlink information [20]. Paging signals are also sent to MSs to inform them of impending downlink data [146]. Apart from the MIB, other information blocks can be transmitted to several users (MSs) or specific users. These information blocks are referred to as System Information Blocks (SIBs) [146]. These signals (synchronization, reference, and downlink control etc.) constitute an overhead in the downlink physical layer of LTE systems and they are transmitted by specific physical channels. Similar overheads are also transmitted in previous generations of cellular systems.

In this study, the control and data plane separation proposed in [20] is adopted. In [20], the plane separation is achieved through network functionality separation of network wide user access and high data rate user information. The network functionalities involved are synchronization, paging, broadcast, multicast (i.e. general user information like mobile TV) and unicast (i.e. user specific data). High power, wide coverage BSs are responsible for transmission of control signal to provide network wide user access. They also handle low data rate user information. Thus, the synchronization, paging, broadcast and multicast network functionalities are associated with these high power BSs. Furthermore, all the overhead signals described earlier have to be transmitted by the high power BSs to fulfill the functions assigned to them.

On the other hand, low power small cell BSs are primarily responsible for high data rate user information. Thus, the synchronization and unicast network functionalities are assigned to them. In addition, the synchronization, reference and downlink control signals are the overhead signals transmitted by the low power BSs. The downlink control signals and reference signals are the main contributors to physical layer overhead [20] as they are transmitted more frequently than other physical layer overhead signals. Nevertheless, because small cells are targeted for hotspot areas such as city centres and stadia with low mobility speed, less frequent transmission of

reference signals can be supported [20]. Hence, control signalling is mainly handled by the high power BSs called control BSs as a result, while the critical high data rate burden is handled by the low power BSs and are called data BSs. This approach has been shown in [147] to result in up to 53% reduction in overhead signal transmission relative to 4G systems at the small cell BSs. Hence, a reduction of 50% is adopted in this study for convenience during analysis, since it involves halving the portion of the power consumption contributed by the overhead signal transmission. Also, it is assumed that the downlink control signals for data BSs (ABSs in this case) are transmitted by the associated control BSs (ZBSs) on their behalf. This makes it possible for the resources of the ABSs to be allocated by the ZBSs as described in Chapter 4.

5.4 Framework for LPSS Evaluation

A BS can achieve some power saving when it operates in a low power state rather than a higher one. The magnitude of the savings is a function of the power model considered, which in turn depends on the BS generation deployed. The power saving due to BSs operating in low power states rather than higher states (termed Low Power State Saving (LPSS)) under different power model assumptions is examined for uplink transmission, and expressions for LPSS are derived for both single and multiple BS scenarios.

5.4.1 Power Models

The power consumption of any type of BS can be approximated with the linear functions of (2.2) and (2.3) in Chapter 2 when a BS is in an active or idle state and when it is in sleep state respectively. Even in the idle state, when no user data is transmitted by a BS, between 10% and 20% of the maximum transmission power is used to transmit reference and control signals (which constitutes overhead) in State-of-the-Art (SotA) cellular systems (e.g. LTE) [18]. Hence, the instantaneous output transmission power, P_{out}, is a combination of the power needed for signalling and power for user data. According to [131] for SotA cellular systems, the instantaneous output transmission power is given by:

$$P_{out} = (1 - p_{OH})P_{max}\rho_b + p_{OH}P_{max}k \tag{5.1}$$

p_{OH} is the fraction of the transmit power required for transmission of fixed overhead signals ($0 \leq p_{OH} < 1$), ρ_b is the fraction of the total bandwidth used for data transmission, and k is a weighting factor that indicates the level of overhead transmitted depending on the state of the BS ($0 \leq k \leq 1$). The value of k for different states are as follows [131]:

$$k = \begin{cases} 0 & sleep \\ 0.5 & idle \\ 1 & active \end{cases} \tag{5.2}$$

In the sleep state, both overheads and user data are not transmitted, partial overheads are transmitted in the idle state, while the complete set of overheads is transmitted in the active state. As stated earlier in Chapter 4, up to 20% of the maximum transmission power may be used to transmit overheads [18], hence p_{OH} is assumed to be 0.2 in this study.

For uplink transmission, BS transmit power is expended on overhead only since no user data is transmitted by the BS. Hence, ρ_b = 0 and the output transmit power, P_{out}, is as follows:

$$P_{out} = p_{OH} P_{max} k \tag{5.3}$$

Furthermore, the values of k equivalent to the 50% overhead reduction adopted for the separation architecture (as explained in section 5.3) for the different states are as follows:

$$k = \begin{cases} 0 & sleep \\ 0.25 & idle \\ 0.5 & active \end{cases} \tag{5.4}$$

The linear functions in (2.2) and (2.3) can be combined as a single function and expressed in terms of static and dynamic parts with the incorporation of the three possible BS states as follows:

$$P_{in} = P_{st} + \Delta p . P_{dy} \ , \quad 0 \leq P_{out} \leq P_{max} \tag{5.5}$$

$$P_{st} = \begin{cases} N_{TRX} . P_0, & Other\ states \\ N_{TRX} . P_{sleep}, & Sleep\ state \end{cases} \tag{5.6}$$

$$P_{dy} = \begin{cases} N_{TRX} \,.\, P_{out}\,, & Other\ states \\ 0\,, & Sleep\ state \end{cases} \tag{5.7}$$

P_{st} is considered to be the static power consumption. This is because P_0 is a parameter that represents BS power consumption at zero output but which is usually measured at 1% of the maximum output transmit power [148]. P_{dy} is considered as the dynamic part since the output transmit power varies with the load. The variation could be due to reduction in occupied subcarriers and/or subframes [41].

Each ABS is classified as a 2x2.5W Microcell BS with a maximum transmission power of 5W. Since power model parameters for the 2x2.5W ABS are not available explicitly in the literature, they have been derived. The maximum power consumption of any generic SotA 2010 base stations can be obtained from the linear function specified in [49]. According to [49], the power consumption of any type of BS at full load, P_{in}, is a linear function of the full load (or maximum) transmission power, P_{max}, as follows:

$$P_{in} = a.P_{max} + b \tag{5.8}$$

P_{in} and P_{max} are dBm power values, while a and b are constants; a = *0.618* and b = *26.1*. From (5.8) the maximum power consumption of each 2x2.5W ABS is 102.6W. Other relevant power model parameters of the ABSs can be obtained from (2.2).

Six power models ranging from a state-of-art 2010 (SotA 2010) model to a future model are considered. Four of the models – SotA 2010, Improved DTX, Market 2014 and Future Models – were previously considered in [54] for single macrocell scenario. The Han model proposed in [31] (and considered in Chapter 4) and Beyond 2020 model (proposed in this study) complete the set of power models. The ABS specific linear model parameters for the different models are provided in Table 5.1 for a single transceiver chain.

The SotA 2010 model is based on the linear power modelling of SotA BS types in 2010 proposed in [40]. The SotA 2010 model is specified for a 2x6.3W microcell, using (2.2) and the relationship between the sleep and no load consumption in [40],

the parameters for the ABS is obtained for this model. The improved DTX model proposed in [44] assumes that significant power consumption reduction can be achieved through cell DTX which is a procedure that switches the BS to a sleep state. This is because sleep state consumption, P_{sleep}, is only approximately 6% of the no load consumption, P_0. However, the no load consumption is unchanged since enhancement of BS hardware is not considered in this model. The Market 2014 model (so called in [54]) suggested in [55] assumes that in addition to the sleep mode capability, BSs are designed with more power efficient components in the future. Hence, substantial reduction in no load consumption (approximately 40%) is assumed in addition to the sleep mode saving.

Table 5.1 Linear Model Parameters for Different Power Models

Models	P_0 **(W)**	Δp	P_{sleep} **(W)**
SotA 2010	44.8	2.6	31.4
Improved DTX	44.8	2.6	2.7
Market 2014	17.9	2.6	2.7
Beyond 2020	1.2	7	1
Han	2.78	4.44	0.14
Future	0.1	7	0.1

A beyond 2020 model is also proposed to reflect the expected design of BSs to have nearly perfect load dependency and very low sleep state and no load power consumption. Hence, a sleep mode consumption that is much lower than the 2014 status is assumed. In addition, rather than a 100 percent increase in power from the sleep to no load consumption assumed for a 2020 small cell model in [149], a much lower increase of 20% is assumed in this case. The Han model, proposed in [31], assumes a relatively low no load power consumption and nearly zero sleep state consumption. In addition, it accounts for power consumed in reactivating a BS in the sleep state referred to as *wake up energy*. However, the contribution of *wake up energy* has been observed to be trivial. As observed in Figure 5.1, the total energy consumption of the IA-CCR scheme with third choice restriction policy with or without the *wake up energy* consideration is similar. Hence, *wake up energy* is not considered in the linear adaptation of this model. Extra power cost incurred by signalling is not accounted for in all states and ABSs transmit at maximum power when active.

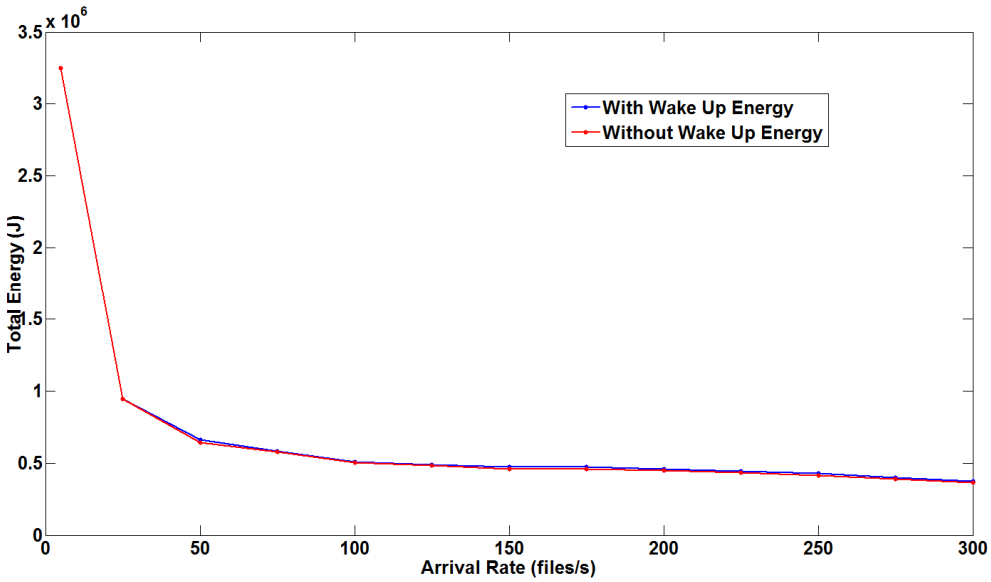


Figure 5.1 Total Energy Consumption of IA-CCR scheme with third choice restriction policy with or without Wake Up Energy

Finally, the Future model proposed in [54] is an ideal model that provides theoretical limit for power consumption. This model results in a near perfect load dependent power consumption; also, the no load consumption is exactly the same as the sleep state consumption. It is assumed here that no extra power cost is incurred for overheads ($k = 0$) in the idle state. This is possible with overhead transmission completely disabled in the idle state. However, overhead power is included in the active state (uplink/downlink). Hence, power is mainly utilised when users are being served.

5.4.2 Short Timescale LPSS in Single BS Scenario

An ABS can be in any of the three possible states at a given time and may make transition to a different state after a period of time (as shown in Figure 5.2). In the same vein, an ABS may operate in a particular state under a certain resource management scheme but operate in a different state under another scheme for a similar observation period and system settings. When the ABS is monitored over a very short timescale of the order of magnitude of the time between user arrivals or departures (a few seconds), it is possible to observe single state changes. Over longer timescales (a couple of minutes or hours), the ABS may undergo several state changes. The short timescale is first considered and the LPSS concept is developed. Subsequently LPSS over the long timescale is considered.

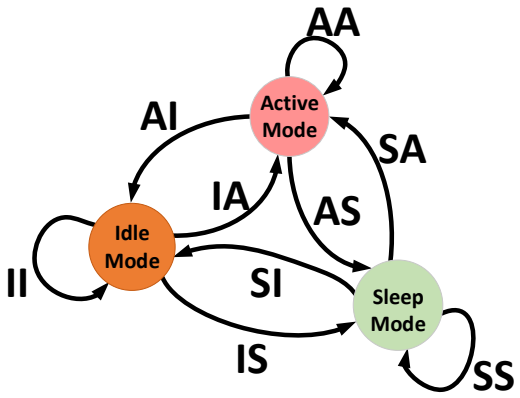


Figure 5.2 BS Possible State Changes

If an ABS changes state under the control of a resource management scheme from an initial state in which its power consumption is P_1 to a new state where it consumes P_2 and remains in this new state for a time period, t; some power saving (LPSS) will be achieved as a result of this state change for the considered period, t, if $P_1 > P_2$. Similarly, if an ABS is monitored over a fixed period of time and fixed system setting (e.g. fixed traffic load and distribution) under two different resource management schemes, one scheme can achieve power saving relative to the other if the ABS effectively operates in different states under the different schemes. The equations derived subsequently are applicable to both the state change and the state difference cases and both are used interchangeable.

Whenever there is a state change, there is a potential increase or decrease in power consumption relative to the initial state. However, only the Active to Idle (AI), Active to Sleep (AS) and Idle to Sleep (IS) state changes lead to power saving. From the linear power model function of (5.5), we can obtain a generic expression for the LPSS due to any of the state changes above. If the BS was initially in a state 1 and changes to a new state 2, the LPSS, P_{12}, can be expressed as follows:

$$P_{12} = P_{st,1} - P_{st,2} + \Delta p(P_{dy,1} - P_{dy,2}) \tag{5.9}$$

$P_{st,1}$, and $P_{st,2}$ are the static power consumption in state 1 and state 2 respectively, while $P_{dy,1}$ and $P_{dy,2}$ are the dynamic power consumption in state 1 and state 2 respectively. Δp is the slope of the dynamic part. The LPSS gain, $P_{G,12}$, which

expresses the LPSS, P_{12}, as a ratio of the power consumption in the initial state, P_1 , can be defined as follows:

$$P_{G,12} = \frac{P_{12}}{P_1} = \frac{P_{st,1} - P_{st,2} + \Delta p(P_{dy,1} - P_{dy,2})}{P_{st,1} + \Delta p.P_{dy,1}} \quad (5.10)$$

The LPSS gains for different state changes under different power model assumptions for a single BS are shown in Figure 5.3. The uplink is considered in this study; thus, the active state represents the periods an ABS is receiving data transmission from MSs. It can be observed from Figure 5.3 that the state changes that are significant with regard to power saving and the potential for power saving varies from model to model. The SotA 2010 model shows the lowest potential for power saving because it has both comparatively high no load and sleep mode consumption, resulting in a lower range of power saving. Other models show better potential for power saving because of the significantly lower sleep state consumption and in some cases low no load consumption. The SotA 2010, Improved DTX and Market 2014 show almost no benefit for operating the ABS in idle state. On the other hand, the remaining models, show appreciable power savings when an ABS is operated in idle state instead of active state. The Future model alone does not benefit from switching idle BS to sleep state since consumption is the same in both states.

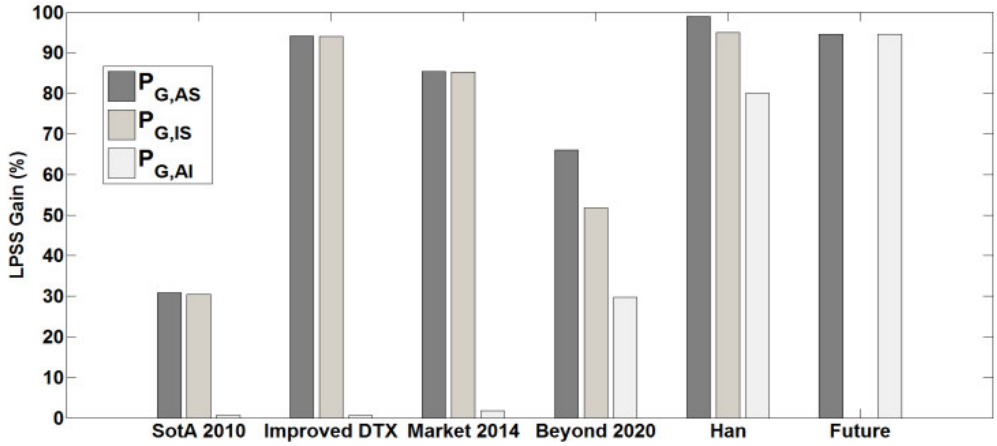


Figure 5.3 LPSS Gains for Single BS

5.4.3 Short Timescale LPSS in Multiple BS Scenarios

The LPSS concept is also extended to multiple BS scenarios. This is typical of the small cell layer of the separation architecture. In this case, the LPSS gains are

expressed in terms of the relative importance of the different state changes on a global scale (i.e. multiple state changes and multiple BSs). The equations are derived based on state differences and the assumption of different resource management schemes.

It is assumed that the first scheme is a baseline resource management scheme that requires all ABSs to be always on (i.e. either in an active or idle state). This is similar to conventional always-on resource management schemes with the goal of spectral efficiency rather than energy efficiency. The second scheme, which is the test scheme, is an energy efficiency driven scheme that can switch off (sleep state) BSs under favourable conditions. With the baseline scheme as the reference scheme, the LPSS gains can be evaluated for the test scheme.

The total power consumption of the baseline and test schemes are $P_{baseline}$ and P_{test} respectively and expressed as follows:

$$P_{baseline} = \sum_{j=1}^{n}\left(P_{st,b,j} + \Delta p.P_{dy,b,j}\right) \tag{5.11}$$

$$P_{test} = \sum_{j=1}^{n}\left(P_{st,t,j} + \Delta p.P_{dy,t,j}\right) \tag{5.12}$$

n refers to the total number of BSs, b represents a BS state under consideration of the baseline scheme while t similarly represents a BS state under consideration of the test scheme. Therefore, $P_{st,b,j}$ and $P_{dy,b,j}$ are the static and dynamic power consumption of the jth BS in state b with consideration of the baseline scheme. Similarly, $P_{st,t,j}$ and $P_{dy,t,j}$ are the static and dynamic power consumption of the jth BS in state t with consideration of the test scheme.

Therefore, following from (5.11) and (5.12) the total power saving of the test scheme with respect to the baseline scheme, P_{saving} is as follows:

$$P_{saving} = P_{baseline} - P_{test}$$

$$= \sum_{j=1}^{n}\left(P_{st,b,j} + \Delta p.P_{dy,b,j}\right) - \sum_{j=1}^{n}\left(P_{st,t,j} + \Delta p.P_{dy,t,j}\right) \tag{5.13}$$

Two types of LPSS gain is defined for the multiple BS case: absolute and comparative. On one hand, the absolute LPSS gain, $P_{A,bt}$, measures the actual saving

due to a particular state difference between the test and baseline scheme for the period of observation with respect to the baseline power consumption. On the other hand, the comparative LPSS gain, $P_{C,bt}$, measures the saving of a particular state difference with respect to the total saving. Thus, the comparative LPSS gain shows explicitly the share of a particular state difference combination in the total saving whereas the absolute LPSS gain just shows its actual value. If P_{bt} is a LPSS of a single BS (kth BS) as defined in (5.9), then the absolute LPSS gain, $P_{A,bt}$, and comparative LPSS gain, $P_{C,bt}$, can be expressed as follows:

$$P_{bt} = P_{st,b,k} - P_{st,t,k} + \Delta p(P_{dy,b,k} - P_{dy,t,k}) \tag{5.14}$$

$$P_{C,bt} = \frac{P_{bt}}{P_{saving}} = \frac{P_{st,b,k} - P_{st,t,k} + \Delta p(P_{dy,b,k} - P_{dy,t,k})}{\sum_{j=1}^{n}(P_{st,b,j} + \Delta p.P_{dy,b,j}) - \sum_{j=1}^{n}(P_{st,t,j} + \Delta p.P_{dy,t,j})} \tag{5.15}$$

$$P_{A,bt} = \frac{P_{bt}}{P_{baseline}} = \frac{P_{st,b,k} - P_{st,t,k} + \Delta p(P_{dy,b,k} - P_{dy,t,k})}{\sum_{j=1}^{n}(P_{st,b,j} + \Delta p.P_{dy,b,j})} \tag{5.16}$$

$$\text{Hence, } P_{A,bt} = P_{C,bt} \times \frac{P_{saving}}{P_{baseline}} = P_{C,bt} \times P_{gain} \tag{5.17}$$

where P_{gain} is the overall power saving gain of the test scheme relative to the baseline. Subsequently, comparative LPSS gain is evaluated over a snapshot of the BuNGee network using the highest SINR scheme as the baseline scheme and the combination of the IA-CCR and the modified topology management (TM) schemes presented in Chapter 4 as the test scheme. It is important to note that MS arrivals are modelled by flow level dynamics, which constitutes a random arrival of MSs into the network each with file transfer request and departure from the network when the file has been successfully transferred [150]. The case where MSs send one file at a time and are assigned one subchannel for this purpose is considered.

5.4.4 Comparative LPSS Gain in BuNGee Snapshot

A multiple BS scenario comprising four ABSs of the BuNGee Architecture is considered and it is assumed that a short timescale exists that contains all three power saving state differences (i.e. AI, AS, IS) of the baseline scheme relative to the test scheme as shown in Figure 5.4. Only one case of each power saving state difference is observed across all ABSs and each ABS is associated with only the

indicated state difference during this short timescale. The combinations in Figure 5.4 are considered in order to compare all three power saving state differences under an equal weighting regime of one occurrence per state difference. This is used to obtain an evenly weighted comparative LPSS gain for all the power models and to show the significance of each state difference to power saving on a more global level than the single BS case. This way the BS state differences that are significant with respect to energy saving can be identified. The absolute LPSS gains can be obtained from comparative gains using (5.17).

The state difference combinations of Figure 5.4 are illustrated in a real service area with a snapshot of BuNGee streets with MSs, ABSs and ZBSs in a zone as shown in Figure 5.5. Only the frequency band of antennas pointing in the zone is shown for the ABSs and the energy calculation is carried out for the single transceiver chain of each ABS serving the zone. Thus, each ABS is modelled as a single transceiver ABS.

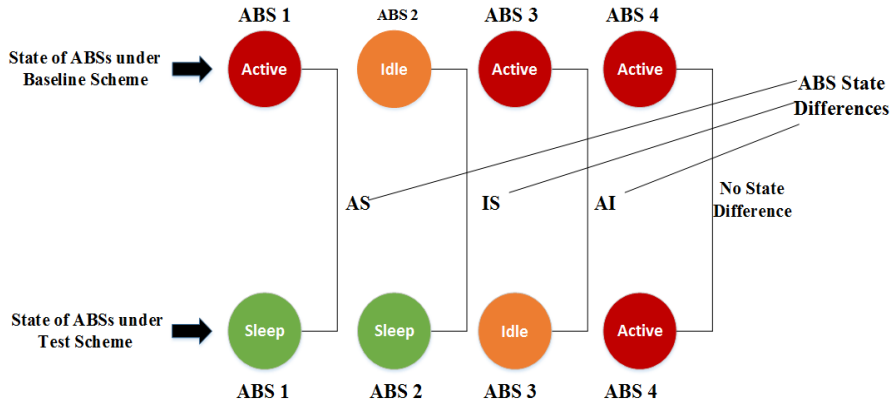


Figure 5.4 Multiple BS State Change Saving Concept

It is assumed that the six MSs are the only active users that arrived with uplink requests and are allocated resources prior to the short timescale considered. In addition, it is assumed that no MS departure occurs during the short timescale. As mentioned earlier each MS is assigned one subchannel out of 10 subchannels configured on each ABS antenna, so the level of ongoing traffic is low when compared to capacity of 40 MSs that can be supported theoretically by the four ABS antennas actively serving the zone.

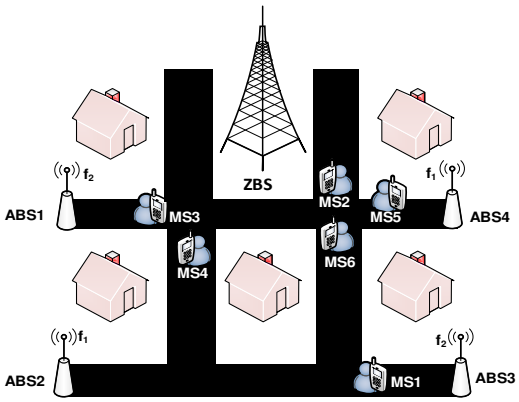


Figure 5.5 BuNGee Snapshot of Streets with ABSs and MSs

Since the traffic is low, interference will be low and a single ABS can serve all six MSs. However, based on the baseline scheme, the highest SINR scheme, the ZBS assigns MSs to the closest ABSs that can give the highest SINR. Hence, it is expected that ABS 1 will serve MS 3 and MS 4; ABS 3 will serve MS 1; while, ABS 4 will serve MS 2, MS 5 and MS 6. Therefore, during the short timescale considered, ABS 1, ABS 3 and ABS 4 will be in active state while ABS 2 will be in idle state since ABS deactivation (sleep state) is not supported under the baseline scheme.

On the other hand, the test Radio Resource Management (RRM) scheme (IA-CCR) is based on the concept of clustering MSs around a few ABSs as long as the ABSs are permitted choices. Assuming that MSs are permitted to use resources from up to the fourth choice ABS, then all the MSs can be served by ABS 4 or ABS 1 which is the most central ABS. MSs are clustered using ABS 4 as an example in this scenario. Since, ABS 4 will be serving six MSs with greater than 50% of its resources (60% precisely), this will prevent the switching off its neighbour, ABS 3, according to the topology management ABS switch off rules. However, ABS 1 and ABS 2 are both idle without overloaded neighbours, thus, the switch off rules are satisfied for both of them and they can be switched off (sleep state). Thus for the test scheme, ABS 1 and ABS 2 will be in sleep state, ABS 3 in idle state and ABS 4 in active state. Therefore, comparing the baseline and test schemes, there are three power saving state differences in three ABSs and no state difference in one ABS as shown in Figure 5.4.

From (5.15) and (5.17) the comparative LPSS and overall power saving gain, $\boldsymbol{P_{gain}}$, can be calculated under consideration of the different power models for this scenario. These are shown in Figure 5.6. In terms of the overall power saving gain, $\boldsymbol{P_{gain}}$, Figure 5.6 shows that the SotA 2010 model has the lowest potential for power saving as in the single BS case. Furthermore, since the comparative LPSS gains are evenly weighted, only the state differences that end in sleep states (i.e AS and IS) are significant for energy saving with regard to SotA 2010, Improved DTX and Market 2014 models. On the other hand, all state differences contribute to energy savings to varying degrees under Beyond 2020 and Han models. However, AS is the most significant in both cases. For the Future model, IS is of no benefit to energy saving while AS and AI are equally significant.

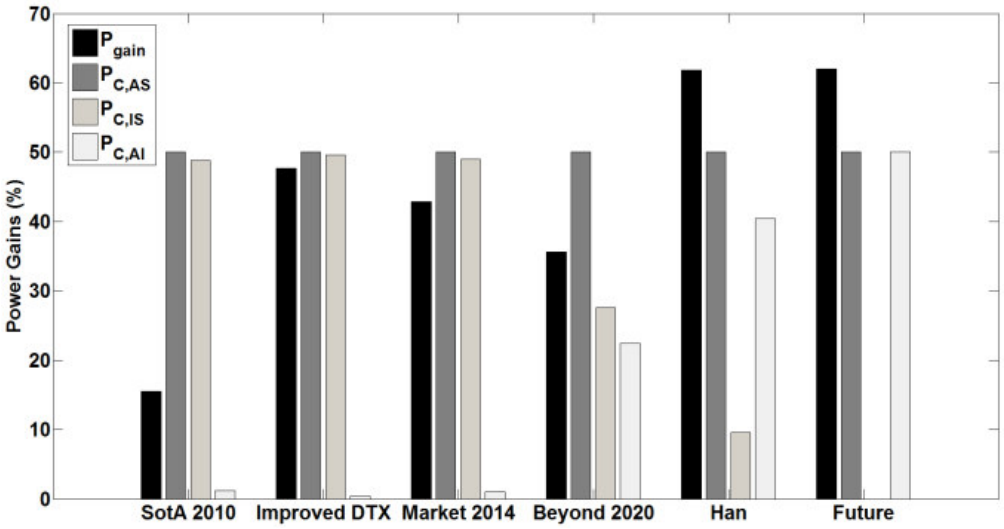


Figure 5.6 Power Saving Gain and Comparative LPSS for BuNGee Snapshot

5.4.5 Long Timescale LPSS

In the previous sections, short timescale LPSS is the focus with only one power saving state difference occurring. However, when an ABS is observed over a long period under flow level dynamics, apart from state differences associated with power saving, state differences associated with power losses can also be observed and also no state differences at all. Assuming that an ABS is observed over a long timescale that is divided into very short timescales, then energy saving will be achieved if and only if the total saving from power saving state differences exceeds the total losses from power loss state differences. The long timescale LPSS is developed based on

this approach and energy saving in the long timescale is expressed in terms of the sum of short timescale power saving state differences. The energy saving, E_i, of any ABS i achieved by a test scheme relative to a baseline scheme over a period T, divided into n short timescales, can be given by:

$$E_i = P_{j_1 k_1} t_1 + P_{j_2 k_2} t_2 + \cdots + P_{j_n k_n} t_n = \sum_r^n P_{j_r k_r} t_r \tag{5.18}$$

where j_r and k_r are the states of the ABS under the baseline and test schemes respectively during the r^{th} timescale, while t_r is the duration of the r^{th} timescale. $P_{j_r k_r}$ is the difference between the power consumption of the baseline state, P_{j_r},and test state, P_{k_r}, in r^{th} timescale and thus, it is equivalent to the LPSS state difference of (5.9). Since $P_{k_r j_r}$ is also a state difference term, then,

$$P_{j_r k_r} = P_{j_r} - P_{k_r} = -P_{k_r j_r} \tag{5.19}$$

$$P_{j_r k_r} = 0 \; ; \;\; if \; j_r = k_r \tag{5.20}$$

From Figure 5.2 in section 5.4.2, there are 9 state difference combinations of the three possible states (active (A), idle (I), and sleep (S)) i.e. II, IA, AI, SI, IS, SS, AS, SA, AA. II, SS, and AA lead to zero power saving; IA, SI, and SA lead to power losses; while, AI, IS, and AS lead to power saving. Since from (5.19) a power loss state difference can be expressed as a negation of a power saving state difference, the energy saving (or loss) can be expressed as a function of the three power saving state differences. Hence the energy saving (or loss), E_i, of any ABS i can be expressed as:

$$E_i = P_{AI} A_i + P_{IS} B_i + P_{AS} C_i \tag{5.21}$$

A_i is summation of timescales associated with AI or IA, B_i is total timescales associated with IS or SI, and C_i is total timescales associated with AS or SA respectively; thus, A_i, B_i, $C_i \in$ R . For a scenario comprising m ABSs, from (5.21) the total energy saving (or loss) over all m ABSs, E_{Total}, can be expressed as:

$$\begin{aligned} E_{Total} &= \sum_i^m (P_{AI} A_i + P_{IS} B_i + P_{AS} C_i) \\ &= P_{AI} A_T + P_{IS} B_T + \; P_{AS} C_T \end{aligned} \tag{5.22}$$

$A_T = \sum_i^m A_i$; $B_T = \sum_i^m B_i$; $C_T = \sum_i^m C_i$. Hence, from (5.22) energy saving over a long timescale will be dependent on how well the test scheme makes decisions that emphasises the significant positive power saving LPSS state differences in short timescales during system operation, since some LPSS state differences are almost negligible.

As long timescales may include a large number of short timescales, which might be tedious to analyze in practice, a less cumbersome approach is used which estimates the energy saving based on the total time spent by the ABSs in each state. Assuming the sum of the individual duration spent by the ABSs in the active, idle and sleep states under the baseline scheme are $T_{A,b}$, $T_{I,b}$, and $T_{S,b}$ respectively while $T_{A,t}$, $T_{I,t}$, and $T_{S,t}$ are the equivalent durations for the test scheme; then the total baseline energy consumption, E_b, and the total test scheme energy consumption, E_t, are as follows:

$$E_b = P_A T_{A,b} + P_I T_{I,b} + P_S T_{S,b} \tag{5.23}$$

$$E_t = P_A T_{A,t} + P_I T_{I,t} + P_S T_{S,t} \tag{5.24}$$

P_A, P_I, and P_S are the power consumption of an ABS in the active, idle, and sleep states respectively. Therefore the energy saving, E_{Total}, in terms of the duration in the different states is given by:

$$E_{Total} = P_A\left(T_{A,b} - T_{A,t}\right) + P_I(T_{I,b} - T_{I,t}) + P_S(T_{S,b} - T_{S,t}) \tag{5.25}$$

If N_{abs} is the number of ABSs in the network, then the average energy saving per ABS, $\overline{E_{Total}}$ is therefore:

$$\overline{E_{Total}} = P_A \frac{(T_{A,b} - T_{A,t})}{N_{abs}} + P_I \frac{(T_{I,b} - T_{I,t})}{N_{abs}} + P_S \frac{(T_{S,b} - T_{S,t})}{N_{abs}} \tag{5.26}$$

If $\overline{T_{A,b}}$ and $\overline{T_{A,t}}$ are defined as average duration of an ABS in the active state under the baseline and test schemes respectively, then a difference term can be defined that indicates how effective the test scheme is in reducing the active duration of ABSs relative to the baseline. We refer to this term as the *net average active duration*, $\overline{T_{A,net}}$, therefore:

$$\overline{T_{A,net}} = \frac{(T_{A,b} - T_{A,t})}{N_{abs}} \tag{5.27}$$

Similarly, the *net average idle duration,* $\overline{T_{I,net}}$, and *net average sleep duration,* $\overline{T_{S,net}}$, are given by:

$$\overline{T_{I,net}} = \frac{(T_{I,b} - T_{I,t})}{N_{abs}} \tag{5.28}$$

$$\overline{T_{S,net}} = \frac{(T_{S,b} - T_{S,t})}{N_{abs}} \tag{5.29}$$

The net average durations of (5.27), (5.28), and (5.29) cannot all be positive if there is some energy saving since the test scheme will prioritise ABS operation in some states over the other states relative to the baseline scheme to achieve this saving.

Furthermore, the energy reduction gain (ERG) used to measure energy efficiency in Chapter 4 can be approximated as a function of the energy saving, E_{Total}, if the throughput is equivalent or almost equivalent in both baseline and test schemes. In chapter 2, the ERG is expressed in terms of the Energy Consumption Rating (ECR) and Energy Consumption Gain (ECG) in (2.9) and (2.10) respectively. Using ECG and ECR, the ERG is expressed in terms of E_{Total} as follows.

If S_b and S_t are the successfully delivered bit under the baseline and test scheme respectively while E_b and E_t still stand for the energy consumed under the baseline and test scheme respectively.

Hence, from (2.8) $ECG = \frac{E_b S_t}{E_t S_b}$ (5.30)

Then, $ECG = \frac{E_b}{E_t}$; if $S_b \cong S_t$ (5.31)

Therefore, from (2.10) $ERG \cong \left(1 - \frac{E_t}{E_b}\right) \times 100\% = \frac{E_{Total}}{E_b} \times 100\%$ (5.32)

where $E_{gain} = \frac{E_{Total}}{E_b} \times 100\%$ (5.33)

E_{gain} is defined in a similar way as the power saving gain, P_{gain}, as the overall energy saving gain of the test scheme relative to the baseline.

It is shown in the next section that the throughput of the baseline and test schemes are nearly equal. Since the throughput is measured for the same duration for both schemes, the successfully delivered bit in both cases are nearly equal as well. Therefore, just the energy saving plots are shown and used to measure energy efficiency; the ERG plots are not shown since they are similar to the energy saving plots for the reasons explained earlier.

5.5 Simulation Results and Discussion

In order to understand how the energy saving and QoS varies with the power model on a large scale, the complete BuNGee separation architecture is modelled in MATLAB in a similar way as in Chapter 4. 5 HBSs, 9 ZBSs and 112 ABSs are deployed in the network, while 6,000 MSs are distributed uniformly outdoors along the streets. Monte Carlo simulations are performed to evaluate the energy savings of the test scheme relative to the baseline scheme, (i.e. highest SINR) under different power model assumptions representing different BS generations. The simulation parameters are specified in Table 5.2. The simulations are carried out over traffic load ranging from low traffic load to the high traffic load. Simulation for each traffic load is ended after a time period required to request upload of a total of 100,000 files in the network. Two cases of the test scheme are compared with the baseline scheme. The first one is implemented without the TM scheme in order to evaluate the effect of idle state saving only on both QoS and energy saving. The second case involves both the IA-CCR scheme and the TM scheme and shows the added effect of sleep state operation.

In all instances, each user arrives into the system with a fixed file size of 2MB to upload and the arrival rate of users into the system has a Poisson distribution with a mean λ. Also, the third choice restriction, which has been shown to achieve significant energy saving at low load in Chapter 4, is set as the restriction level for MSs in the IA-CCR scheme. This rule is set at the Quality Enhancement Processing Unit (QEPU), the higher decision making node described in Chapter 4. The acceptable range of operation of the system is the region where the blocking probability is less than 5% and the energy saving is above zero. The QoS is measured in terms of blocking probability and average file transfer delay. It is

important to note that the QoS performance will be the same irrespective of the power model because the same maximum transmission power and system bandwidth is considered for each ABS for all the power models. The major difference is how the static and dynamic power consumption is proportioned to achieve this transmit power and bandwidth requirement in different power models.

Table 5.2 Simulation Parameters [125]

Parameter	Value
Deployment area dimension	1350m×1350m
Street width	15 m
Building block size	75m×75m
ABS antenna height	5m
MS antenna height	1.5m
Carrier Frequency	3.5GHz
MS Transmit Power	23dBm
ABS Maximum Gain	17dBi
Noise Floor	-114dBm/MHz
Call Admission SINR	10dB
Minimum SINR for Reception	1.8dB
SINR for highest throughput	21dB

The throughput performance of the schemes is shown in Figure 5.7. The throughput of the schemes is nearly the same across all traffic loads. This is because when uplink file requests are blocked, they are permitted to be retransmitted later until they eventually get through. Therefore, when the throughput is measured over a long period of time as done in this study, nearly equal numbers of files are successfully delivered. Thus, although the throughputs are similar across all traffic loads, the delay performance varies from scheme to scheme as explained later.

The blocking probability performance of the schemes is shown in Figure 5.8. It is observed that the baseline scheme, i.e. highest SINR scheme, has the best blocking probability performance. This is because all ABSs are always on and MSs are served by their closest and first choice ABSs. The test scheme without TM, i.e. IA-CCR without TM, has blocking probability comparable to the highest SINR below medium traffic load (less than 200 files/s). However, as the traffic load increases the interference becomes more and more significant because of the permission of connection to other choices apart from the first choice. Thus, it has much poorer blocking probability with respect to the baseline scheme at high traffic load (above

200 files/s). The blocking probability is further worsened by allowing idle ABSs to be switched off (i.e. IA-CCR with TM). This is because when ABSs are switched off, options of ABSs available for data services reduce and alternatives are in sleep state when active ABSs have no suitable channel.

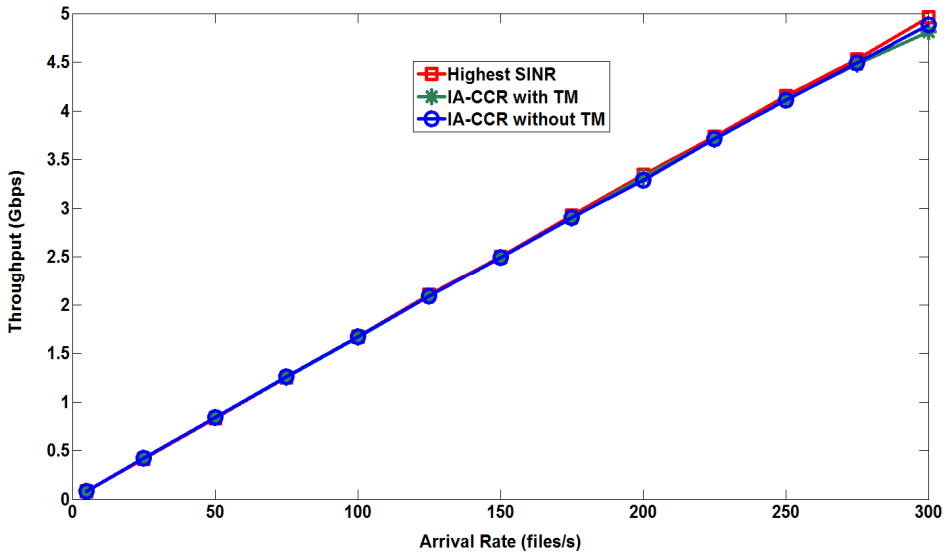


Figure 5.7 Throughput of Baseline scheme and Test scheme without TM and with TM

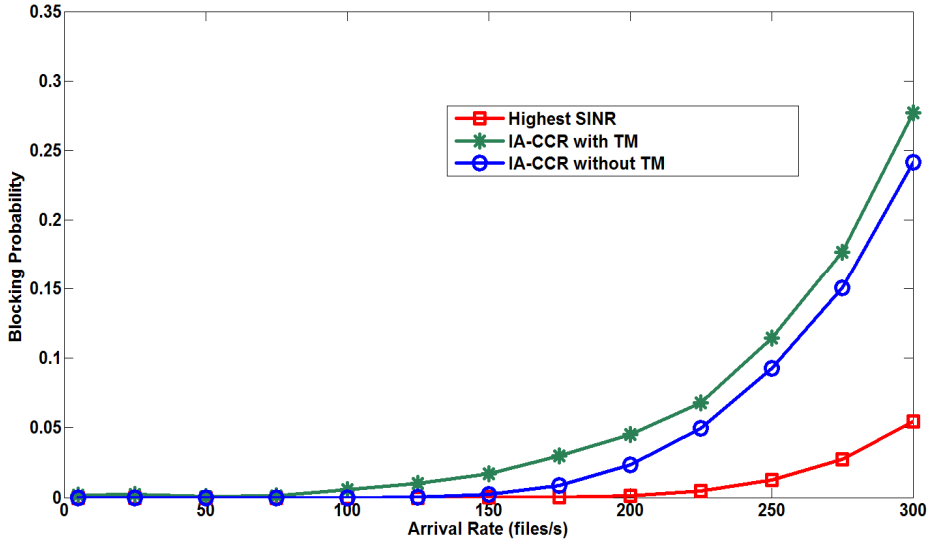


Figure 5.8 Blocking Probability of Baseline scheme and Test scheme without TM and with TM

The average file transfer delay performance shown in Figure 5.9 follows a similar trend as the blocking probability. Again, the highest SINR scheme has the best performance and this is because MSs use the highest SINR possible for transmission

and therefore they complete transmission faster. The “IA-CCR without TM” causes many MSs to operate at lower SINRs than the highest SINR due to permission of second and third choice connections. Thus, higher average file transfer delay is experienced under the test scheme. The delay is further increased when idle ABSs are allowed to sleep. This is because more distant MS to ABS connections will be experienced and even lower SINRs will be utilised than when idle ABSs are not put into sleep state.

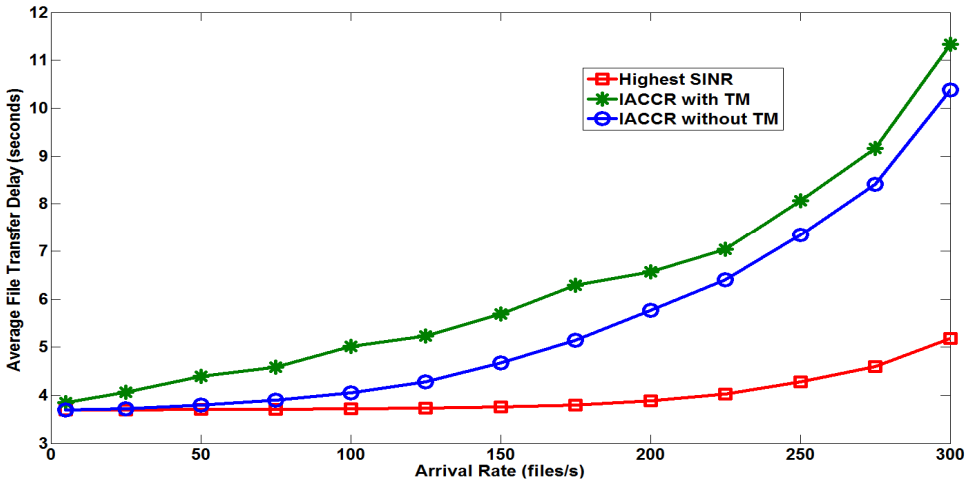


Figure 5.9 Average delay of baseline scheme and test scheme without TM and with TM

The net average duration of an ABS in different states for the baseline scheme relative to the test scheme when sleep state is not permitted is shown in Figure 5.10. At the lowest traffic load (5 files/s) the net average duration is nearly zero for both active and idle states. This is because very small number of ABSs is required simultaneously at such a very low traffic load to serve users and there is hardly any room to benefit from clustering MSs with few ABSs. However, at higher traffic loads (up to 125 files/s) the highest SINR scheme serves MSs with a higher number of ABSs (in the active state) relative to the test scheme. Hence, the net average active duration increases positively while the net average idle duration is increasingly negative. Thus, the state difference in this case is only active to idle (AI) and it is a power saving LPSS type. As the traffic load increases further the trend is reversed and eventually both net average durations reach zero at 300 files/s. This is as a result of increasing load and interference which makes it increasingly difficult for the test scheme to cluster MSs with few ABSs.

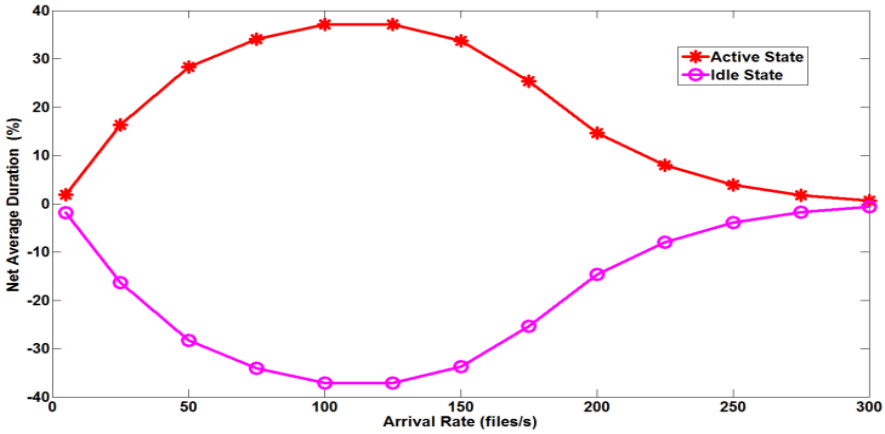


Figure 5.10 Net average duration of baseline scheme relative to test scheme without TM in different ABS states

The energy saving for the test scheme without TM is shown in Figure 5.11 for the different power models. SotA 2010, Improved DTX and Market 2014 are nearly zero because the LPSS gain for AI is negligible for these power models. However, significant energy saving is achieved with the Beyond 2020, Han and Future models, as the LPSS gain for AI is significant for these set of models. Also, the Future and Han models have higher savings than Beyond 2020 model. This is because the Future model assumes ABSs have very low idle state consumption like sleep state; while the Han model assumes ABSs have high uplink active consumption. Beyond 2020 model assumes more moderate idle and active state consumptions, thus the lower saving noticed. Furthermore, the AI is negligible in SotA 2010, Improved DTX and Market 2014 because the active and idle state consumption are of the same order of magnitude. This is not the case for the other three models, as the active state is of higher order of magnitude than the idle state, which further explains the reason for the appreciable saving in these models and not in the other models. It can be concluded that significant idle state saving is possible only when the active state is of a higher order of magnitude than the idle state.

The net average duration of an ABS in different states for the baseline scheme relative to the test scheme with sleep state supported (with TM) is shown in Figure 5.12. It is observed that the net average active duration and net average idle duration are both positive while the net average sleep duration is negative. Therefore, the state differences in this case are active to sleep (AS) and idle to sleep (IS) and both state

differences are LPSS state differences. The net average active duration (like the no TM case) rises from a near zero value at the lowest traffic load to a peak value (at 50 files/s) and then gradually decreases until it reduces to zero at the highest traffic load. Comparatively, while unused ABSs are left in the idle state in the baseline scheme, in the test scheme these unused ABSs would be put to sleep alongside those that would have been in the active state to serve users. Hence, at low load the net average idle duration is high but as traffic load increases, the baseline scheme requires more ABSs to be in active state. Thus, the net average idle duration is negligible beyond 100 files/s. As a result below 100 files/s power saving state differences are both AS and IS but mainly AS after 100 files/s.

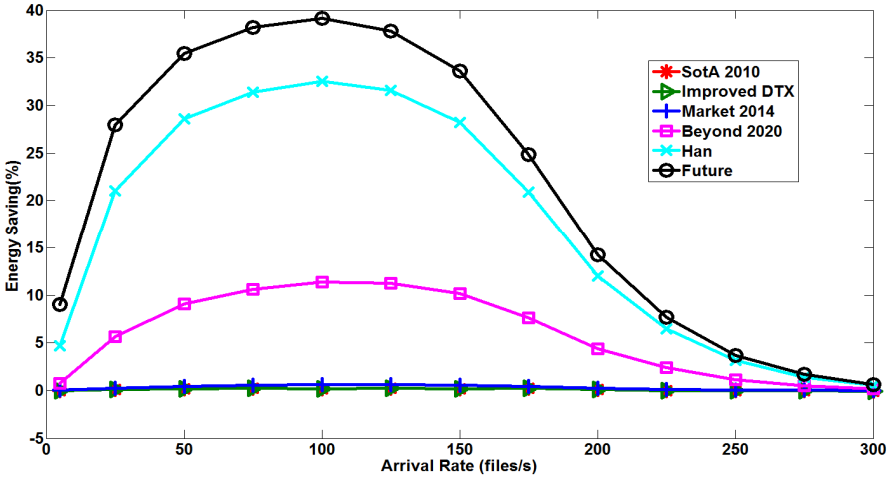


Figure 5.11 Energy saving of test scheme without TM for different power model assumptions

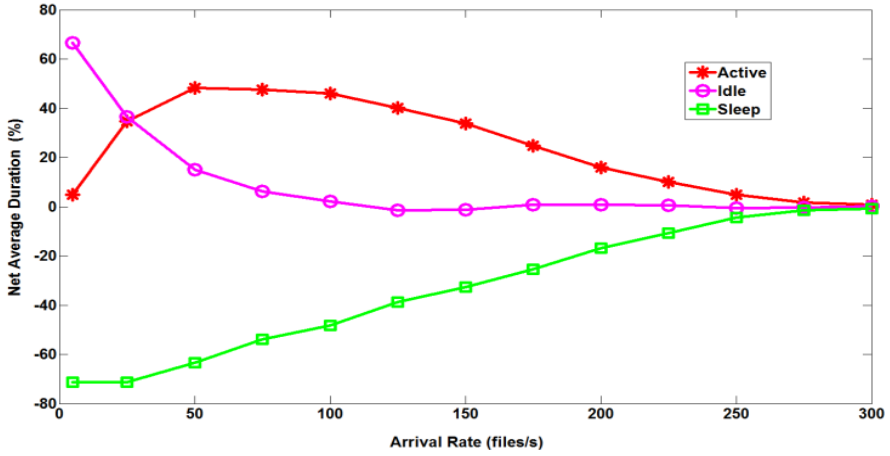


Figure 5.12 Net average duration of baseline scheme relative to test scheme with TM in different ABS states

The energy saving for the different power models when TM is applied in the test scheme is shown in Figure 5.13. At the lowest load of 5 files/s, as can be observed in Figure 5.12, the IS state difference is the predominant power saving state difference. Since, the IS state difference gives zero savings for the Future model, it has relative low energy saving obtained for the small AS state difference. However, as the AS state difference duration increases the energy saving increases and reaches its peak when the AS duration also reaches its peak. Beyond 2020 and Han Models benefit from IS state difference but to lower degree compared to the AS state difference (as observed under Comparative LPSS in Figure 5.6). Therefore, both models only reach their peak values after the AS state difference becomes significant. The other models (SotA 2010, Improved DTX and Market 2014) benefit nearly equally from both AS and IS state differences and are therefore at peak values at the lowest traffic load. Energy saving reduces to zero at the highest load when there are no state differences. The sleep state saving provides additional energy saving benefits to the idle state saving for all models.

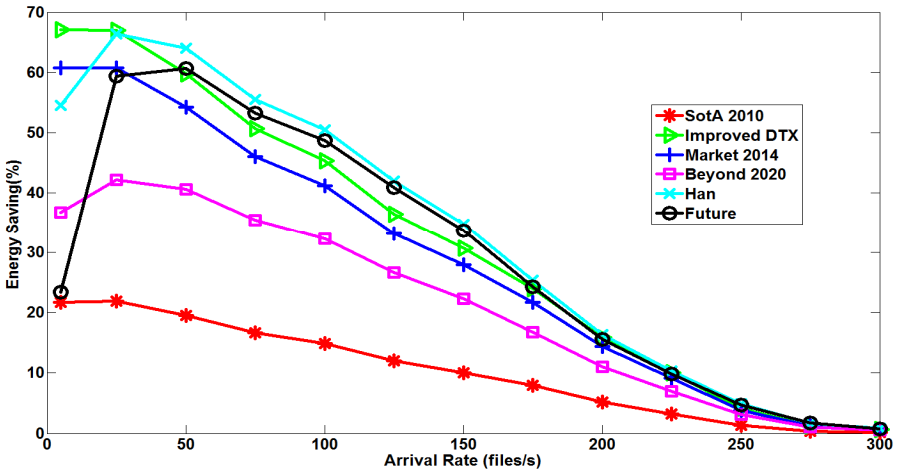


Figure 5.13 Energy saving of test scheme with TM for different power model assumptions

The SotA 2010 model achieves the lowest energy saving (up to 21%) because of the high sleep consumption while the massive reduction of sleep state consumption in the Improved DTX and Market 2014 models lead to high energy savings (up to 61% and 67% respectively). Although, the no load consumption is reduced significantly for both Han and Future Models, their energy saving are comparable with the

Improved DTX because sleep mode consumption is almost zero in both cases. More conservative sleep mode consumption is assumed for Beyond 2020 model and the energy saving is lower (up to 42%). Furthermore, it is important to note that sleep state saving (i.e. energy saving with TM) provides additional energy saving benefits over the idle state saving (i.e. energy saving without TM) for all models. This is clearly demonstrated by the higher energy saving under the sleep state energy saving regime (in Figure 5.13) relative to the idle state saving case (in Figure 5.11). This is because if an ABS is put to sleep it will not be available to serve an MS that cannot be served by other active ABSs whereas if it is only idle it can immediately serve the MS. Thus the AS state difference and the corresponding energy saving is higher for sleep state saving than the AI state different and corresponding energy saving for idle state saving. Moreover, energy saving due to AS state difference is usually higher than the AI state difference case except for the ideal future model where both have equal energy saving potential.

5.6 Conclusion

In this chapter, the impact of power model assumptions on the achievable energy saving is studied in the BuNGee Separation Architecture by operating small cell BSs in low power states rather than higher power states (referred to as Low Power State Saving (LPSS)). The BS state changes or state differences that results in LPSS are identified and LPSS gain over very short timescales are computed for different power models. It is shown that these short timescale LPSS gains determine energy saving performance in multiple BS scenarios and over long timescales. Simulation results show that energy saving of an energy efficient resource management scheme relative to a baseline, high capacity density focussed scheme varies across different power models as a function of model-specific significant LPSS state differences.

Also, if the separation architecture is based on existing small cell BSs , modelled by state of the art power models, which have idle state consumption of the same order of magnitude as the active state consumption, significant idle state energy saving is not possible. Significant energy saving is only possible through sleep mode activation in this type of scenario. On the contrary, if the separation architecture is assumed to be based on future small cell BSs, modelled by more advanced power

models, significant energy saving is possible through both idle and sleep state BS operation. In this case, energy saving through idle state operation at high traffic load is still possible even if longer BS waiting time before sleep are introduced for better QoS or even if BSs are not switched off anymore. This insight is applied to enhance the modified TM scheme presented in Chapter 4 in the next chapter, where an adaptive joint RRM and TM scheme is presented. Peak energy savings ranging from 21% to 67% are obtained across the power models. More importantly, up to 42% energy saving is obtained for the Beyond 2020 model which is based on less than the ideal assumptions of the Future model. Furthermore, the IA-CCR scheme when combined with the modified TM scheme can achieve significant energy saving regardless of the small cell BS type or power model assumption.

Chapter 6. Confidence Level Based Adaptive Resource and Topology Management

6.1 Introduction

Significant energy saving can be realised in a cellular network comprising a dense deployment of small cell BSs if a large number of these BSs can be switched to the sleep state. This can be achieved at low traffic, as demonstrated in Chapter 4, by clustering or concentrating MSs on a few small cell BSs which are selected to serve MS data requests, while remaining BSs are put to sleep. However, aggressive clustering and a high percentage of sleeping ABSs can drive QoS beyond operating limits at medium and high traffic load levels. This is the experience with some of the previous energy efficient schemes - Highest SINR with One Neighbour On, NCCR and CBCA schemes. As shown in Chapter 4, the blocking probability of these schemes exceeds the 5% threshold at traffic load levels lower than the Baseline Highest SINR without TM scheme. This implies that these energy efficient schemes have lower traffic load range of operation relative to the baseline schemes. A scheme that can adapt to different traffic loads such that significant energy is saved at low traffic loads while at higher traffic loads, moderate energy saving and QoS targets are achieved is desirable. The development of such an adaptive scheme is the focus of this chapter.

In this chapter, a novel confidence level based adaptive joint resource and topology management scheme is proposed. This scheme utilises traffic data from the different zones of the network to estimate QoS parameters and modifies radio resource management (RRM) and topology management (TM) policies when QoS thresholds are exceeded. A range of values which the QoS parameters will likely fall within, referred to as the confidence interval, is estimated at a predefined confidence level as described earlier in Chapter 3. The decision to modify RRM and TM policies are taken only when the confidence interval of the estimated QoS parameter does not exceed a defined maximum value criterion at the chosen confidence level.

Specifically, the IA-CCR scheme is enhanced to incorporate the capability to modify restriction policies, while the modified TM introduced in Chapter 4 is also enhanced to support the capability to prohibit or permit the sleep state TM decisions in the

network so as to maintain target QoS levels. The restriction policies are the RRM policies, while permission and prohibition of sleep state constitute the TM policies. These policies are decided and modified at the QEPU, which is the most central and highest node among the RRM nodes. The QEPU decision is relayed to the ZBSs, which then implement the policies at the zone level. As before the partially centralised paradigm is maintained, since RRM and TM tasks can go on when there are no new instructions from the QEPU.

Furthermore, this adaptive approach is in line with the Self-Organising Network (SON) paradigm that has been introduced by the Third Generation Partnership Project (3GPP) for the operation, administration and maintenance (OAM) of LTE/LTE-A networks (the most adopted 4G system) [151]. The objective of SON is to bring about the automatic and dynamic management of an LTE network with limited human intervention [152]. This is also important for future radio access technologies because of the complexity that will be associated with tuning several parameters and managing large numbers of nodes comprising conventional macrocells and dense small cell deployments [153]. SON functionalities can be classified under three categories: self-configuration, self-optimization and self-healing [154]. Self-configuration involves the automatic configuration of new nodes in a network, self-optimization involves the automatic adaptation of the network parameters to improve performance; while self-healing involves the automatic detection and correction of network failures [155].

According to [28], an RRM technique that utilises SON functionality is referred to as a SON-RRM. This chapter presents a novel SON based joint RRM and TM scheme which can be termed a SON-RRM utilizing self-optimization functionalities. Furthermore, several SON use cases, such as mobility load balancing, coverage and capacity optimization, and energy saving have been proposed by 3GPP [156]. The work presented in this chapter utilises traffic statistics and a confidence level based adaptive approach to modify policies in order to maintain QoS targets while still saving energy. Thus, it falls under the SON energy saving use case domain.

The remaining section of this chapter is organised as follows. The system model is discussed in section 6.2, while the novel confidence level based adaptive strategy is

presented in section 6.3. The simulation results of the system level evaluation of the proposed strategy are presented and discussed in section 6.4 and the chapter is concluded in section 6.5.

6.2 System Model

The BuNGee Separation Architecture considered in the previous chapters is still the focus here and the functions of the ABSs, HBSs, ZBSs, BHSSs and QEPU still remain the same. The MSs are uniformly distributed outdoors in the service area as in previous cases. However, the QEPU and its relationship with the OAM system (described subsequently) and the interaction with other nodes for the dynamic adaptation of policies is explained in detail herein. This is because the adaptation decision is made at the QEPU.

The function of the QEPU makes it possible for its implementation at different levels of the network and as one of different possible modules e.g. in software, as a firmware or as a unique system. It could be implemented as a software module in a policy control node like the Policy and Charging Rules Function (PCRF) in the core network of an LTE network, as a network node on its own with connection to the OAM or as a subsystem of the OAM. The case of a subsystem of the OAM is considered in this work with the advantage of quick access to traffic statistics from BSs as explained subsequently.

The OAM system of a mobile network is responsible for the operation, administration, management, and maintenance of the network [156]. Network management can be carried out at mainly three different levels of management with reference to the 3GPP OAM architecture; these are the network element (NE) layer, the network element management (NEM) layer, and the network management (NM) layer (or the operation support system (OSS) layer) [157]. The NE layer is the lowest layer and may consist of base stations and base station controllers, the NEM layer consists of different vendor-specific element managers, each for similar set of network elements; the NM or OSS layer is the topmost layer with a 3GPP standardised interface (Itf-N) to element managers, allowing management of the overall network [152].

A SON algorithm can be classified as centralised, distributed or hybrid depending on the management layer(s) where it is implemented [151]. Centralised SON algorithms are those implemented solely in the NEM or OSS layer and involve the control of NEs such as BSs by a central entity. Distributed SON algorithms are those implemented at NE layers, in this case for example BSs co-ordinate through exchange of local information without involving a central entity. Hybrid SON algorithms are those implemented at two or more layers (e.g. OSS and NE layers) such that SON functions are shared between the different layers.

In this work, a hybrid SON approach is considered for the energy saving use case. The QEPU is assumed to be a subsystem in an OSS system and it is responsible for coordinating energy saving under QoS constraints. It can however be assigned other RRM related responsibilities, such as determining policies for admission control. In carrying out its responsibilities the QEPU can take advantage of the rich traffic statistics and Key Performance Indicators (KPIs) available at the OSS [30], since these data are available normally for daily operation of the network anyway [152]. The traffic statistics and KPIs are aggregated from measurements recorded by different event and performance counters of network elements (such as base stations) and conveyed to the higher OAM layers [152, 158].

The hierarchical architecture of the nodes involved in the hybrid SON based algorithm presented in this chapter is shown in Figure 6.1. It is assumed that network operation starts with default RRM and TM policies set by a human expert at the QEPU associated with the OSS system. The network subsequently adapts policies automatically to suit different traffic loads. The default RRM and TM policies are conveyed to the ZBSs and based on these policies each ZBS makes RRM decisions concerning resources on which ABS should be used to serve an MS requesting service as described in Chapter 4. The ZBS also makes TM decisions. Specifically, when the blocking probability in its zone deteriorates, it requests activation of the highest CCR ABSs in sleep state (as also initially stated in the TM rules in Chapter 4). As shown in Figure 6.1, an active ABS can receive both RRM and TM messages from the ZBS but a sleeping ABS can receive only TM messages. ABSs can also make TM decisions on their own. They can go into the sleep state when they are idle for a specific period of time and their neighbors are not

overloaded as specific in the TM rules outlined in Chapter 4. In addition, an overloaded active ABS can request its sleeping ABS neighbour to switch on to share the burden of serving MSs.

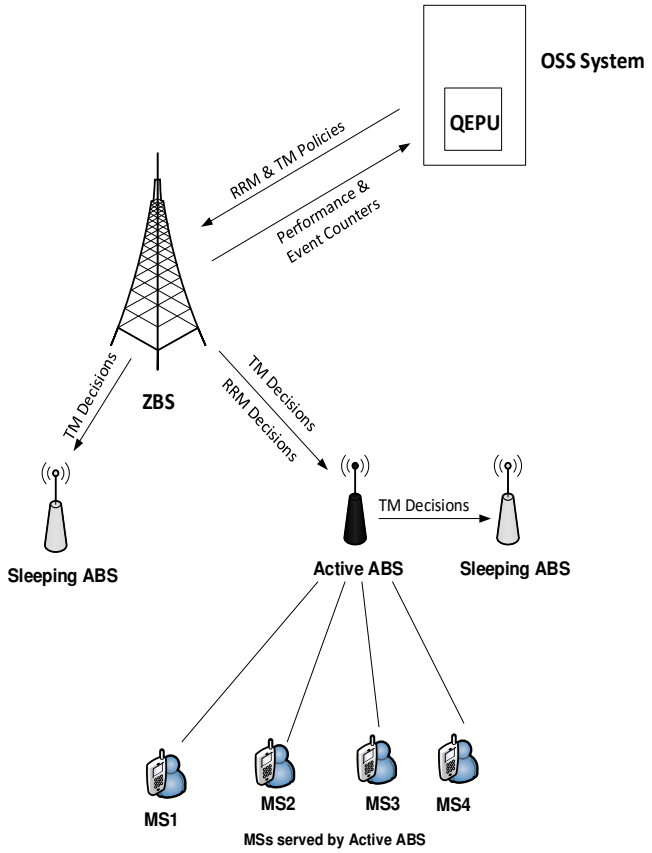


Figure 6.1 Hybrid SON Nodes

Although, ZBSs take action to mitigate poor QoS, the default RRM and TM policies might be unsuitable such that the action of the ZBSs might not be sufficient to attain the desired QoS target. In such situations the QEPU steps in, it detects the poor QoS condition and modifies policies for the service area based on traffic statistics and KPIs calculated from performance and event counters collected from the ZBSs. In this work, the instantaneous number of users being served by the system, the total number of blocked requests and the overall number of user requests (including blocked and processed requests) are collected as traffic statistics. The blocking probability, calculated from the ratio of total blocked request to overall requests, is defined as the QoS parameter (and KPI as well). A blocking probability target is predefined and when the QEPU estimates system blocking probability to be higher than the target, it modifies the combination of RRM and TM policies. Although,

blocking probability is used as QoS parameter in this study, other QoS parameters can be used as well such as delay and throughput.

The RRM policy used here is the ABS choice restriction put in place to limit resource utilization to a set of high SINR choices as described in Chapter 4. Hence, at a given time the RRM policy can be for example 1st choice, 2nd choice or an even higher choice restriction. On the contrary, the TM policy can be Sleep State ON or Sleep State OFF. While Sleep State ON implies the sleep state is permitted, Sleep State OFF means sleep state is not permitted so that ABSs cannot go to the sleep state. Therefore, full transition from any ABS state to any other state, as in Fig 5.2, is no longer possible under the Sleep State OFF condition. Transitions from the sleep state to other states are possible but transitions from other states to the sleep state are not possible as shown in the modified BS state changes of Fig 6.2 below. Sleep State OFF is particularly relevant in high traffic load situations. The complete policy message sent by the QEPU is a combination of a RRM policy and a TM policy.

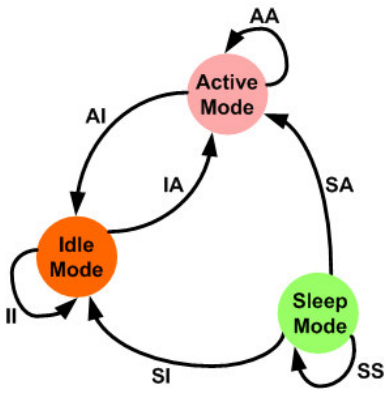


Figure 6.2 BS Possible State Changes under Sleep State OFF condition

Each RRM policy and TM policy combination is defined as a policy level in the database of policy levels that the QEPU can select from. In this model, the QEPU performs the adaptation task by switching step by step from one level to another in order to drive the network to a state that meets the QoS target. This approach is utilised for the RRM and TM policies considered for complexity reasons and the way the adaptation algorithm proposed exploits available information. This is explained in detail in the subsequent section. For another algorithm, the QEPU might not need to take a stepwise approach. For example, when nothing is known about

the relationship between the policy level and the traffic load, a reinforcement learning approach based on less ordered progression through the policy levels may be considered.

As shown in Figure 6.3, the QEPU can make (N-1) adaptation decisions before getting to Level 1 based on a stepwise approach and starting at the highest level defined as Level N. This choice of adaptation direction will be the case if it is already known that the QoS performance improves with decreasing level of the policy selected. The time interval between adaptation decisions needs to be short enough to enable practicality but not too short to ensure the collected traffic statistics are sufficient enough to make good estimation of the QoS parameters (or KPIs). The confidence interval and the associated confidence level are used in this study as a quantitative measure of the precision expected from the parameter estimation procedure [139]. It is also used for controlling the time interval between adaptation decisions and thus the size of traffic statistics collected between time intervals for QoS parameter estimation. This is demonstrated in detail in the next section.

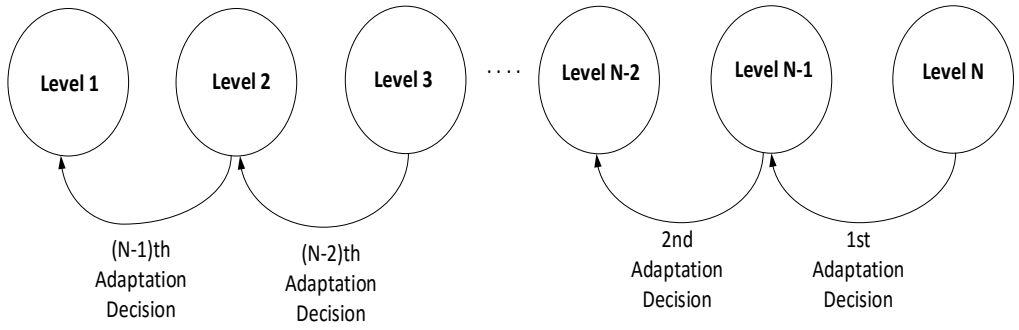


Figure 6.3 Policy Levels and Adaptation Procedure

It is important to note that ZBSs and ABSs make TM and/or RRM decisions in timescales of user arrivals and departures which can be of the order of few seconds or lower. On the contrary, the QEPU is designed to make RRM and TM policy decisions in longer timescales related to the average offered traffic with slower variations of the order of minutes or hours. However, near real time decisions are possible at the QEPU when a limited coverage area is monitored, since performance and event counters collection interval between 10 seconds to 5 minutes is possible with such a limited scope [152]. A distributed implementation of the OAM at different locations as suggested in [18] can support this type of limited coverage.

Hence, the policy modifications can be made near real time while zonal RRM and TM decisions can be made in real time [152]. The detailed process of policy modification based on the novel confidence level approach is presented in the next section.

6.3 Confidence Level Based Adaptive Radio Resource Management and Topology Management

The previous energy efficient schemes proposed for the BuNGee Architecture, as shown in Chapter 4, exceed the 5% blocking probability threshold at traffic load levels much lower than the Baseline Highest SINR without TM scheme. However, no provision is made to address QoS deterioration in these schemes and thus, they have a lower range of suitable operation compared to the baseline scheme in terms of QoS. The confidence level based adaptive scheme proposed in this chapter permits the automatic adjustment of RRM and TM parameters to enable the network to be restored within a defined operating QoS limit. The IA-CCR and the modified TM schemes are enhanced to produce an adaptive joint RRM and TM scheme using the QEPU to adapt RRM and TM policies to suit different traffic load conditions. This adaptation task is achieved by utilizing performance and event counters provided by the ZBSs as traffic statistics for making decisions.

The high level information already known about the relationships between ABS choice restriction, traffic load, QoS and ERG is exploited for the RRM policy modification. It is demonstrated in Chapter 4 that beyond the low traffic load region (i.e. medium and high traffic load) the lower the order of choice restriction the lower the blocking probability and delay. For example, a second order choice restriction will result in lower blocking probability and delay than a third order choice restriction at a specific medium or high traffic load. Furthermore, with respect to ERG, at low traffic loads, the lower the choice restriction the lower the ERG. However, the performance of a low order choice restriction decreases with increasing traffic load at a lower rate than higher order choice restrictions. Therefore, although the performance of a low order choice restriction is lower than counterpart higher order choice restrictions at low traffic load, it eventually performs better than them beyond the low traffic load region. Thus, a high order choice restriction may be

preferable at lower traffic loads as long as the QoS is acceptable, while a lower option might be more appropriate at higher traffic loads to maintain QoS targets.

Also, the information regarding sleep state and idle state saving is exploited for the TM policy modification. In Chapter 5, it is shown that sleep state saving provides additional savings to idle state saving but it is associated with additional depreciation in QoS over the idle state saving. In spite of disabling sleep state, energy saving can still be achieved relative to a high data rate centric RRM scheme (like the Highest SINR scheme) in a network based on advanced and ideal small cell BSs modelled by Beyond 2020 and Future Models respectively. At low traffic loads sleep state saving and idle state saving have comparable QoS and QoS targets are generally satisfied in both cases. However, at higher traffic loads the idle state saving QoS is clearly better, thus if QoS targets are breached under sleep state saving regime, sleep state operation can be disabled to improve QoS.

A sequential policy level database is created based on the information known about the RRM policy and TM policy. Each policy level is a combination of a RRM policy and a TM policy as stated previously. As already established, QoS improves when order of choice restriction is reduced and when sleep state is disabled, while at low traffic loads more energy saving is achieved with higher order restriction and sleep state enabled. However, a low order restriction will be better than a higher order restriction at some higher traffic load but the exact traffic load is not known. Thus, the policy level database is constructed with higher order restriction policies plus Sleep State ON above corresponding lower order restriction policies plus Sleep State ON. Furthermore, Sleep State OFF is only considered at the lowest policy level so that the higher energy saving potential of the sleep state saving over idle state saving is maintained until the final restriction policy modification is made without QoS benefit. Therefore, QoS improvement is achieved by moving from a high policy level to low policy level, while energy saving is still possible since both sleep state saving and idle state saving achieves energy saving.

The policy level adaptation for a four level database is shown in Figure 6.4 with the topmost level being the third choice restriction plus Sleep State ON while the lowest level is the first choice restriction plus Sleep State OFF. This four level database is

used in this study and the system starts with a default setting of level 4 (i.e. third choice restriction plus Sleep State ON) and the QEPU switches from higher levels to lower ones until the QoS target is achieved.

Such a stepwise adaptation of policies as described above is considered because determining the right combination of policies beforehand requires a mathematical model of the relationship between the traffic load, RRM and TM policies. Such a model is difficult to come by because of the different choice relationship an ABS may have with the MSs being served. For instance, under a third choice restriction policy, an ABS may be serving some users as their first choice ABS, whereas it is a second or third choice ABS for other users. In addition, there is mutual effect of RRM and TM on each other resulting in further complexity. Hence, the high level relationships established in previous chapters are utilised and the QEPU adapts policies to meet QoS target for each traffic load experienced in the network.

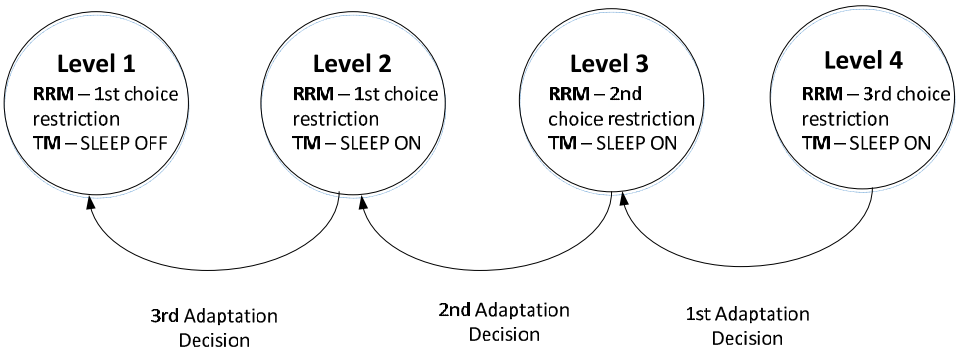


Figure 6.4 Adaptation Procedure for four policy levels

In implementing the algorithm, three unique statistics are created from the performance and event counters. These are: the total number of users served concurrently by the network (referred to as carried traffic herein), the total number of blocked requests and the overall number of user requests (including blocked and processed requests). The carried traffic data is used to estimate the average traffic load and decisions are not taken until the mean carried traffic is deemed fairly stable. This is done to allow the system to reach a stable condition after a change in policy. It is assumed that the carried traffic counter at each ZBS is modified at each user arrival or departure and the time of each change is also recorded. Normally, this sort of information is important to accurately charge users for service provided by a

mobile network operator. The QEPU, however, utilises this information to create a record of the carried traffic at each event, it eventually evaluates the mean of the resulting series of values and decides the system is stable if the confidence interval is within a specified range at a predetermined confidence level.

The blocked request and overall request counters are also modified each time there is an event related to these counters. The time associated with each event is also recorded so that the blocked or overall request at each instance relative to an earlier reference point can be obtained. The records of these counters can then be used to evaluate the overall blocking probability from an earlier reference point to a particular time instance or the current policy level blocking probability measured from the time the current policy level is applied. If current policy level blocking probability is higher than the threshold, the policy level is modified. The blocking probability is estimated with some level of certainty also using the confidence interval. The confidence interval is expected to fall within a specified range and at the predetermined confidence level before the blocking probability is considered for making policy level change decisions.

It is important to note that the blocking probability is considered after the mean carried traffic converges within the defined confidence interval. Furthermore, the higher the confidence level, the more statistics are required to achieve the desired confidence interval. Also, a higher confidence level leads to higher waiting period for collection of statistics and consequently, longer decision epochs. This is explained in more detail in the subsequent section.

6.3.1 Confidence Interval

The use of the confidence interval as a means of demonstrating that performance metrics have been estimated with high degree of confidence is discussed in Chapter 3 and applied to validate the simulation results in this thesis. In this Chapter, it is used in a different manner to ensure sufficient traffic statistics are collected to estimate QoS and also to regulate the time adaptation decisions are made. Specifically, the confidence interval is used for estimating the mean of the carried traffic and also for the blocking probability. While, the mean of the carried traffic is assumed to have a normal distribution, the blocking probability is assumed to have a

binomial distribution since it involves binary outcomes i.e. failure or success. Therefore, confidence intervals relating to normal and binomial distributions are discussed in the following.

The confidence interval estimates a population parameter (such as mean) with an interval and it is usually accompanied by a confidence level [136]. The confidence level (also referred to as degree of confidence) is the probability that the population parameter falls within the confidence interval [137]. For a random sample, X_1, X_2, …, X_n, taken from a population with an unknown mean, μ, and known variance, σ, the confidence interval evaluated with a confidence level of $(1-\alpha)$ for the mean is such that [136]:

$$P(L \leq \mu \leq U) = 1 - \alpha \tag{6.1}$$

L and U are random variables of the lower and upper limits of the interval estimate for the mean respectively and $0 \leq \alpha \leq 1$.

If the sample considered is large ($n \geq 30$) and it has a point estimate of mean given by $\bar{X}$, then $\bar{X}$ has an approximate normal distribution with mean, μ, and variance, σ^2/n [138]; the confidence interval limits can be obtained starting with standardizing $\bar{X}$ as follows as shown in [138]:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \tag{6.2}$$

Z has an approximately standard normal distribution. For any given Z, there is a value $z_{\alpha/2}$ (which is the upper 100 $(\alpha/2)$ percent point of the standard normal distribution [136]) such that

$$P(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1 - \alpha \tag{6.3}$$

Substituting (6.2) in (6.3) and expressing inequalities relative to μ instead of Z gives:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{6.4}$$

If $l, u, and\ \bar{x}$ are the values of the random variables L, U, and $\bar{X}$ respectively, then the lower and upper confidence limits (l and u) are given as follows:

$$l = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.5}$$

$$u = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.6}$$

For the large sample case, if the variance of the population, σ, is unknown it can be replaced by the estimated variance, *s*, of the sample, without a significant effect on the accuracy [138].

$z_{\alpha/2}$ can be determined from (6.3) and the graphical representation of the probability on the left hand side (LHS) of the equation. Since Z is a standard normal distribution it has a mean of 0 and the area under the associated integral is evenly distributed into symmetric negative and positive sides as shown in Figure 6.5. Therefore, the LHS of (6.3) can be expressed as the summation of two probabilities, one for the positive side and the other for the negative side of the area of the integral:

$$P\left(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}\right) = P(-z_{\alpha/2} \leq Z \leq 0) + P(0 \leq Z \leq +z_{\alpha/2}) = 1 - \alpha \tag{6.7}$$

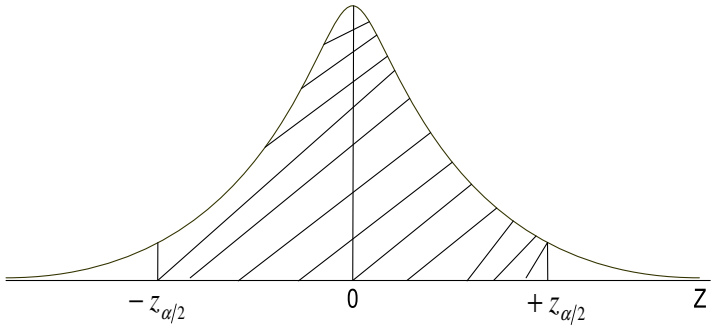


Figure 6.5 Standard Normal Curve for *Z*

Then, due to the symmetry of the positive and negative areas:

$$P\left(-z_{\alpha/2} \leq Z \leq 0\right) = P(0 \leq Z \leq +z_{\alpha/2}) \tag{6.8}$$

Hence, from (6.7)

$$P\left(0 \leq Z \leq +z_{\alpha/2}\right) = \frac{1-\alpha}{2} \tag{6.9}$$

Since α is known, $z_{\alpha/2}$ can be evaluated from a standard normal integral table which gives numerical values for integrals of the form [138]:

$$P(0 \leq Z \leq z) = \int_0^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \tag{6.10}$$

Four confidence levels (30%, 60%, 90% and 99.9%) are considered in this work to demonstrate the impact of quick decision making and sample size of collected statistics on the performance of the algorithm. The confidence levels considered here also include low confidence levels unlike in Chapter 3 where typical high confidence levels used in confidence interval estimations are provided. The associated $z_{\alpha/2}$ value for each confidence level is shown in Table 6.1.

Table 6.1 Confidence Interval Parameters

Confidence Level	30%	60%	90%	99.9%
$z_{\alpha/2}$	0.39	0.84	1.65	3.29

The maximum deviation from the point estimate of mean, $\bar{X}$, based on the $(1-\alpha)$ confidence level is used in this work to decide when the average carried traffic has converged. The estimated maximum deviation for average carried traffic is represented by D_c and from the expressions for the lower and upper confidence limits, (6.5) and (6.6) respectively, D_c is given by:

$$D_c = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{6.11}$$

$D_c < K_c$ is used as criterion for determining that the average carried traffic has converged , where K_c is a real number and is a desired maximum deviation threshold for deciding the carried traffic convergence. This implies that the maximum deviation must be lower than the prescribed value, K_c, at the specified confidence level, $(1-\alpha)$, before it is accepted that the average carried traffic has converged.

The use of the confidence interval for convergence determination is hinged on the fact that the traffic statistics are collected over time and the sample size grows with time rather than all sample sizes being available at any time. Hence, after a transition time such as a policy level change, initial individual carried traffic entries collected will vary during this unstable initial phase and confidence intervals will be wide due to the high variance. This is because the estimated maximum deviation, D_c, is directly proportional to the standard deviation (which is the square root of the

variance) as can be observed from (6.7). However, as the system stabilises by adjusting to the new policy level, the variation in the carried traffic will reduce and so will the confidence intervals estimated.

Also from (6.7) and the values of $z_{\alpha/2}$ for the different confidence levels from Table 6.1, for a given value of K_c, the higher the confidence level the higher the sample size that will be required to satisfy the criterion. This is because the value of $z_{\alpha/2}$ increases with the confidence level and the estimated maximum deviation, D_c, is directly proportional to $z_{\alpha/2}$ while it is inversely proportional to the sample size, n. Since, the traffic statistics sample size increases as time passes; then, higher confidence levels will lead to longer decision epochs but will produce estimates with better degree of confidence.

As mentioned earlier, the confidence interval for the binomial distribution is considered to handle the case of blocking probability. The binomial distribution is characterised by a parameter, p, which is the probability of success and the number of trials, n [139]. Only two outcomes are possible for each trial, i.e. failure or success. Assuming that $\hat{p}$ is the probability of success estimated from a large sample ($n \geq 30$), the confidence interval for p at a confidence level of $(1-\alpha)$ can be estimated in a similar way as the mean, μ, for the population considered earlier as follows [138]:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{6.12}$$

This is the case because the random variable, $\hat{\boldsymbol{p}}$ with outcome $\hat{p}$, is an unbiased estimator of p and has a normal distribution with mean p and variance, $p.(1-p)/n$ [139].

The blocking probability evaluated from the data collected from the different zones is defined in a similar way as $\hat{p}$ in this study, with blocking considered as the successful event being counted. Also, like the carried traffic case, the maximum deviation criterion has to be satisfied before decisions are taken. However, in this case, the focus is on the use of the criterion to estimate the blocking probability with

some degree of confidence not just convergence. The estimated maximum deviation for blocking probability, D_p, is given by:

$$D_p = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{6.13}$$

The criterion in this case is given by $D_p \leq K_p$, where K_p is the desired maximum deviation threshold for blocking probability. How this criterion is applied in adapting the RRM and TM parameters to traffic load are discussed in detail under the algorithm implementation in the next section.

6.3.2 Algorithm Implementation

The processes involved in achieving the adaptation of the RRM and TM parameters are described in the following. This is an enhancement of the IA-CCR and modified TM scheme presented in Chapter 4 to enable the network to respond to poor QoS that could not be resolved at the zone level with the initial RRM and TM policies (jointly referred to as a policy level). It is important to note that once the policy is updated by the QEPU, the IA-CCR scheme (which is the RRM scheme) is implemented with the new RRM policy and the same process of allocating resources as in the initial case described in Chapter 4 is followed. In contrast, with the TM policy if the policy changes to SLEEP OFF then the rule governing ABSs going to the sleep state is suspended while the other rules are observed as usual. Hence, when SLEEP OFF is set as the new TM policy, ABSs in the sleep state before or after this new policy is implemented can still be switched on based on TM rules stated in Chapter 4. However, ABSs cannot be switched to the sleep state after it is implemented.

There is a mutual relationship between the policy adaptation module and the zonal RRM and TM module. The zonal RRM and TM module are the IA-CCR scheme and modified TM scheme presented in Chapter 4 without the dynamic parameter or policy adaptation. On one hand, the policy adaptation module feeds the zonal RRM and TM module with RRM and TM policies to govern its decisions; on the other hand, the zonal module feeds the policy adaptation module with statistics to monitor performances and update policies when necessary. The combination of these two

modules produces an adaptive joint RRM and TM strategy. This is a partially centralised strategy since the central entity, QEPU, only makes the decisions about policies but does not specifically decide which ABS to switch on or off and which resources on which ABS to allocate. These decisions are left to the ZBSs and ABSs. For this same reason, it is a hybrid SON strategy. Only the policy strategy module part is outlined here as the zonal module part has been discussed earlier in Chapter 4.

Consider C, O, and B as three unique sets representing the instantaneous carried traffic, overall user requests and total blocked requests respectively. $C = \{c_1, c_2, \dots, c_i, \dots, c_n\}$ where c_i is the carried traffic at the i^{th} event (which can be a user arrival or departure) in the system, while c_n is the carried traffic at the n^{th} event which is also the last event observed before records are available at the QEPU. Similarly, $O = \{o_1, o_2, \dots, o_i, \dots, o_n\}$ and $B = \{b_1, b_2, \dots, b_i, \dots, b_n\}$, while o_i and b_i are the overall user requests and total blocked requests at the i^{th} event respectively. Furthermore, o_n and b_n are the overall user requests and total blocked requests respectively at the n^{th} and last observed event.

Let P_t represent the target blocking probability that the system is required to maintain, while P_q is a control blocking probability ($P_q < P_t$) set by the QEPU such that the maximum deviation threshold for blocking probability, K_p, has a magnitude of $P_t - P_q$. Let $\widehat{P_o}$ and $\widehat{P_l}$ represent the sample estimates of overall and current policy level blocking probabilities respectively. $\widehat{P_l}$ is evaluated when $D_p \leq K_p$ at a predetermined confidence level. Hence, when $\widehat{P_l} > P_q$ the actual blocking probability of the current policy, P_l, will fall within the range: $\widehat{P_l} - K_p \leq P_l \leq \widehat{P_l} + K_p$. The target blocking probability, P_t, as a function of the control blocking probability and maximum deviation threshold, K_p is given by:

$$P_t = P_q + K_p \tag{6.14}$$

Since $\widehat{P_l} > P_q$, therefore, the upper bound of the actual blocking probability of the current policy, $P_{l,u}$, exceeds the target blocking probability, P_t. This is shown as follows:

$$\left(\widehat{P_l} + K_p\right) > P_t \tag{6.15}$$

$$P_{l,u} = \widehat{P_l} + K_p \tag{6.16}$$

$$P_{l,u} > P_t \tag{6.17}$$

Hence, the actual blocking probability of the current policy approaches the target blocking probability or probably exceeds it. At this stage, the policy level is changed to improve QoS. The control blocking probability is used to define decision points to enable the start of adaptation before the target blocking probability is reached. Also, as stated earlier D_c is the desired maximum deviation threshold for the carried traffic that is used to determine that the average carried traffic has converged and the system has stabilised after a change of policy level.

Furthermore, let $L = \{l_1, l_2, \ldots, l_i, \ldots, l_n\}$ be the set of policy levels available to the QEPU such that l_i is the i^{th} policy level and l_i leads to better QoS than l_{i+1} as explained earlier. $l_d \in L$ is defined as the default policy level the system starts with, while $l_q \in L$ is the current system policy level chosen by QEPU and actively used at the zone levels. The event counter is represented by e and it increases in magnitude after every user arrival or departure.

If the m^{th} event is the last event just before the policy level changes to the current policy level while the n^{th} event ($n > m$) is the last event at the moment a new policy level modification decision is considered, the sample estimates of overall blocking probability, $\widehat{P_o}$, and the current policy probability, $\widehat{P_l}$, are calculated differently as follows:

$$\widehat{P_o} = \frac{b_n}{o_n} \tag{6.18}$$

$$\widehat{P_l} = \frac{b_n - b_m}{o_n - o_m} \tag{6.19}$$

However, $\widehat{P_o} = \widehat{P_l}$ at the first policy change decision point, since $m = 0$, $b_m = 0$ and $o_m = 0$ at this stage. It is assumed initially that the records become instantly available at the QEPU without the constraint of periodic upload. This allows the

effect of each confidence level on performance to be evaluated first before considering the impact of periodic upload on the performance of the scheme. Each evaluation is ended when the total number of new user requests (excluding retransmissions), *Utotal*, equals to a defined maximum value, *Umax*. Starting with the default policy level, l_d, the adaptation process is implemented as follows:

Algorithm 1: Policy Level Adaptation Process based on Traffic Statistics

---

1. Initialise $C = \emptyset$, $O = \emptyset$, $B = \emptyset$, $l_q = l_d$, $e = 0$, $n = 0$, $m = 0$, *Utotal* = 0
2. **While** *Utotal* ≤ *Umax* **do**
3. At user arrival or departure: $e = e + 1$
4. **Update** set $C, O,$ and B with new elements c_{e+1}, o_{e+1}, and b_{e+1} respectively
5. For user arrival: *Utotal* = *Utotal* + 1
6. n=e
7. **if** $(n - m) \geq 100$ **then**
8. Calculate D_c
9. **if** $D_c < K_c$ **then**
10. Calculate $\widehat{P_o}$
11. **if** $\widehat{P_o} > P_q$ **&** $(o_n - o_m) \geq 100$ **then**
12. Calculate $\widehat{P_l}$ and D_p
13. **if** $\widehat{P_l} > P_q$ **&** $D_p \leq K_p$ **then**
14. **if** $d > 1$ **then**
15. $l_q = l_{d-1}$
16. m=n
17. **end if**
18. **end if**
19. **end if**
20. **end if**
21. **end if**
22. **end while**

The sample size is considered large enough to assume approximate normal distribution when $n \geq 30$ [138]; 100 is used in this work rather than 30, to be above the limiting condition. The effectiveness of the algorithm combined with the initial zone based RRM and TM module is evaluated in a system level simulation. This is presented in the next section.

6.4 Simulation Results and Discussion

System level evaluation of the adaptive scheme in the BuNGee Separation Architecture used in previous chapters is carried out in MATLAB. 6,000 MSs are uniformly distributed along the streets. User arrival is based on a Poisson process with inter-arrival times exponentially distributed and each user arrives with a request to upload a fixed file size of 2MB. The focus is on the medium and high traffic load region where QoS deteriorates significantly with energy saving. The simulation parameters are provided in Table 6.2.

Table 6.2 Simulation Parameters

Parameter	**Value**
Carrier Frequency	3.5GHz
MS Transmit Power	23dBm
ABS Maximum Gain	17dBi
Noise Floor	-114dBm/MHz
Call Admission SINR	10dB
Minimum SINR	1.8dB
Maximum SINR	21dB
Maximum Number of Iteration, Umax	400,000
Confidence Level	30%, 60%, 90%, 99.9%
The blocking probability threshold, P_t	0.05
QEPU control blocking probability, P_q	0.045
The maximum deviation threshold for blocking probability, K_p	0.005
The maximum deviation threshold for the carried traffic, K_c	1 file
Default policy level	Third choice restriction + SLEEP ON

The blocking probability and energy efficiency is evaluated at 30%, 60%, 90% and 99.9% confidence levels using the corresponding upper 100 ($\alpha/2$) percent point, $z_{\alpha/2}$. The blocking probability threshold, P_t, is set at 0.05 (5%). The maximum deviation threshold for the blocking probability, K_p, is 0.005, hence the control blocking probability, P_q, is 0.045. This permits operation of the network just below the 5% target. The carried traffic is estimated in terms of total users served by the network at the same time and the maximum deviation threshold for the carried traffic, K_c, is set as one (1) user or file. Hence, a confidence interval of less than two (2) users is believed to be sufficient to conclude that the mean carried traffic has

converged for a sample size of at least 100 measurements of the carried traffic. The default policy as mentioned earlier is the combination of third choice restriction and SLEEP ON. Maximum number of user requests (iteration), *Umax*, is set at 400,000.

The performance at different confidence levels are compared with an upper bound decision which waits until the end of 100,000 iterations to decide whether to switch from the current policy level to a new one. As mentioned in Chapter 4, estimation of performance measures at 100,000 iterations and above is good enough to achieve estimates with small errors. A total of 400,000 iterations enables the upper bound decision case to adapt the system if necessary from the policy level 4 to policy level 1 by going through three switching stages. The energy efficiency is evaluated using the Beyond 2020 model described in Chapter 5. The Beyond 2020 model is chosen in order to investigate performance on advanced small cell based networks based on better load dependence and moderate static power consumption. The energy efficiency is evaluated in terms of the energy reduction gain (ERG) and the Highest SINR without TM (i.e. sleep state transition prohibited) is used as the baseline scheme for evaluating ERG.

The temporal blocking probability at 280 files/s is shown in Figure 6.6. It shows the typical blocking probability of the scheme for each confidence level from the default policy until the final policy level settled on by the QEPU. The points marked on each confidence level plot are the points at which policy levels are modified. This figure confirms that the lower the confidence level, the quicker the policy level is modified. This is because the maximum deviation criteria for carried traffic and blocking probability are met sooner with lower confidence level since $z_{\alpha/2}$ decreases as the confidence level decreases as shown in Table 6.1. The upper bound takes longer because it waits until the end of 100,000 iterations before considering policy changes. The maximum number of policy changes (which is 3) starting with level 4 and ending at level 1 are made in less than 10 minutes for all confidence levels. More importantly, the blocking probability decreases when policy level modifications are carried out, and for this traffic load the blocking probability is reduced below the 5% threshold at all confidence levels. This implies that QoS deterioration can be addressed promptly using traffic statistics and with a high degree of confidence in the decisions made.

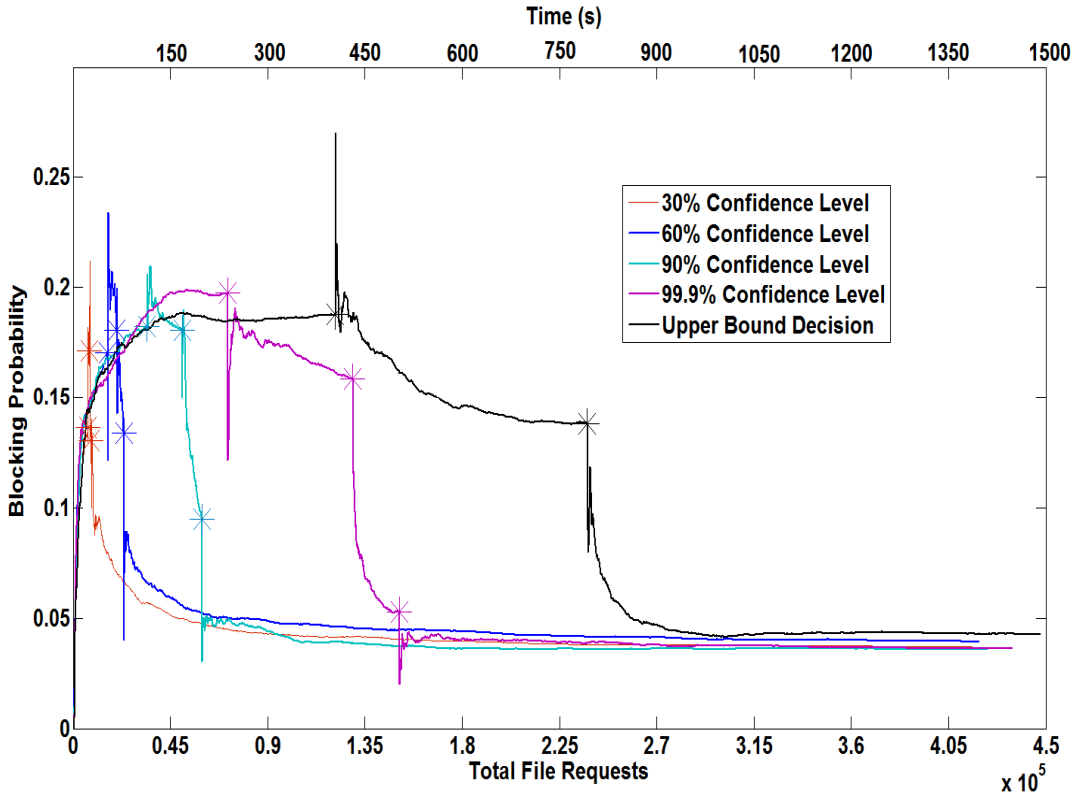


Figure 6.6 Temporal Blocking Probability at 280 files/s

The temporal estimate of the mean carried traffic is shown in Figure 6.7. This shows the progression of the estimated mean carried traffic as the policy levels are changed. The points marked are the same as with the blocking probability plot in Figure 6.6. As mentioned earlier in section 6.3, the mean carried traffic should converge before the blocking probability is estimated and policy level is changed so that the change is made when the system has adapted to the current policy level. It is shown in Figure 6.7 that the mean carried traffic is deemed to have converged quicker for a low confidence level than for the higher ones. This is also because the $z_{\alpha/2}$ decreases as confidence level decreases, so that the deviation constraint is satisfied quicker with lower confidence levels. The sharp, straight downward slopes before the marked points occur because the mean traffic convergence happens before these points and measurements are resumed after the policy change. Hence, the final point before and after the reset are wide apart resulting in the sharp slope. The 99.9% confidence level shows better convergence than the lower confidence levels.

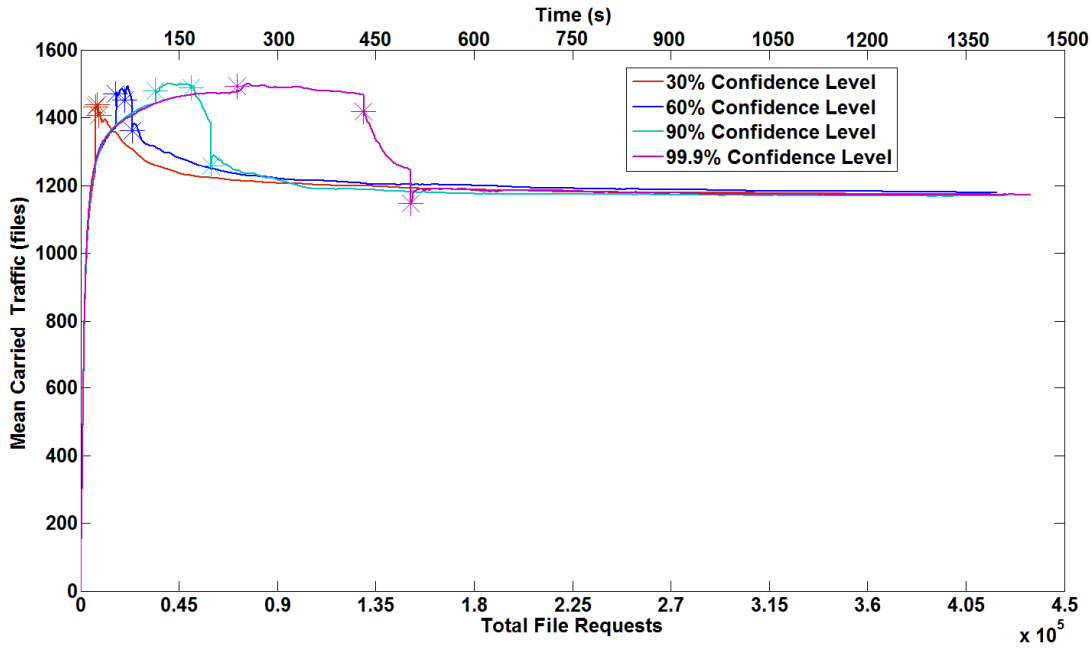


Figure 6.7 Temporal Estimate of Mean Carried Traffic

The final policy level the QEPU settles at over the range of traffic load is shown in Figure 6.8. The general trend is that the QEPU when set at a low confidence level settles at a policy level the same or lower than higher confidence level options. Also at the higher confidence level (90% and 99.9%), the QEPU does not change policy levels as quickly with increasing traffic loads as the lower confidence levels (30% and 60%) do. This is also associated with the comparably lower $z_{\alpha/2}$ values of these lower confidence levels. The upper bound decision takes longer to change policy levels with increase in traffic loads than all the confidence level based decisions. This is because the endpoint evaluation of the blocking probability is used in this case whereas the decisions are made as soon as estimated blocking probability exceeds the threshold in the case of the confidence level approaches. However, the 99.9% confidence level decisions are not far off from the upper bound outcomes.

The blocking probability after the QEPU has settled at a final policy level is shown in Figure 6.9. The blocking probability is evaluated for the final 100,000 fresh user requests of the final policy level. Comparison of Figure 6.9 with the policy level plot of Figure 6.8 shows that the blocking probability depends on the final policy level the QEPU settles at. The blocking probability is lower than the 5% threshold up to 280 files/s for all confidence levels even when the confidence interval of ± 0.005 is

considered. This is not far from the 290 files/s supported by the Highest SINR scheme at 4.4% blocking probability and is better than the maximum range of the previous schemes which is less than the 250 files/s with regard to QoS as observed in blocking probability evaluation of Chapter 4.

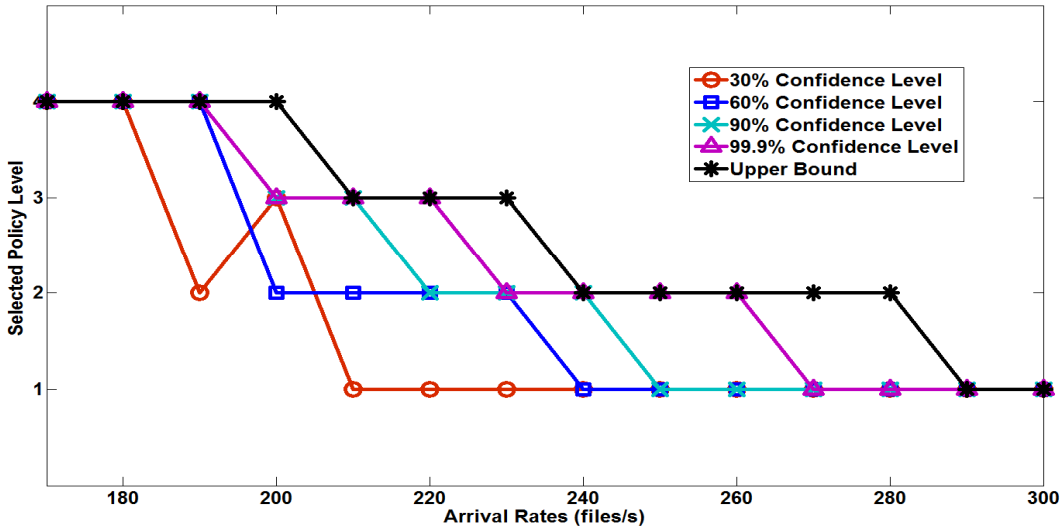


Figure 6.8 Final Policy Level at Different Traffic Load

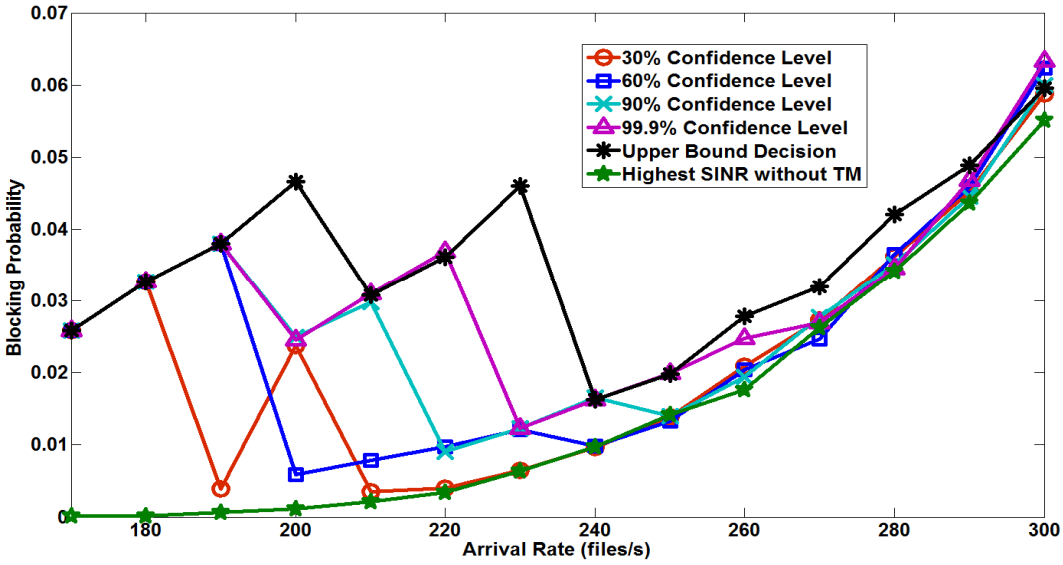


Figure 6.9 Final Policy Level Blocking Probability

The policy level modification plot of Figure 6.8 together with the blocking probability plot of Figure 6.9 show that at 200 files/s and 230 files/s, the 99.9% confidence level based strategy changes policy level because the blocking

probability approaches 0.045 which is used as the control value at the QEPU. It is evident that the blocking probability approaches 0.045 at these traffic load values as the Upper Bound decision which makes changes at endpoints of 100,000 iterations ends at around 0.045 in these cases. This demonstrates that the algorithm executes instructions as initially designed.

Furthermore, still using Figure 6.8 and Figure 6.9 together, it is clear that blocking probability increases with traffic load but by changing the policy level, this trend can be reversed at medium load (170 to 240 files/s). Policy level changes have predominantly involved Levels 4, 3 and 2 and thus involve changes in the RRM choice restrictions as shown in the policy adaptation process in Figure 6.3. Beyond 240 files/s, the modification of policy level does not dampen rising blocking probability significantly. The policy level changes in these high traffic regions are all from Level 2 to Level 1, which is 1st choice restriction plus SLEEP ON to 1st Choice restriction plus SLEEP OFF. This implies that the prohibition of sleep has lower impact on improving QoS than changes in choice restriction. The lower confidence levels (30% and 60%) generally have better blocking probability performance overall than the higher ones (90% and 99.9%) because they generally settle at the same policy levels as the higher ones or at lower policy levels.

The ERG based on the Beyond 2020 model is shown in Figure 6.10. The ERG is estimated relative to the Highest SINR without TM over the 100,000 fresh user requests of the final policy level. Comparison of Figure 6.10 and the final policy level plot of Figure 6.8 show that a confidence level approach achieves higher ERG than a second option whenever it settles at a policy level above level 1 but lower than the policy level chosen by the second option. The lower confidence levels (30% and 60%) generally settle at the same or lower policy levels than the higher ones (90% and 99.9%). Thus, for example the 60% confidence level achieves higher ERG at 200 files/s and 210 files/s than the higher confidence level since it settles at policy level 2 while higher confidence levels settle at higher policy levels.

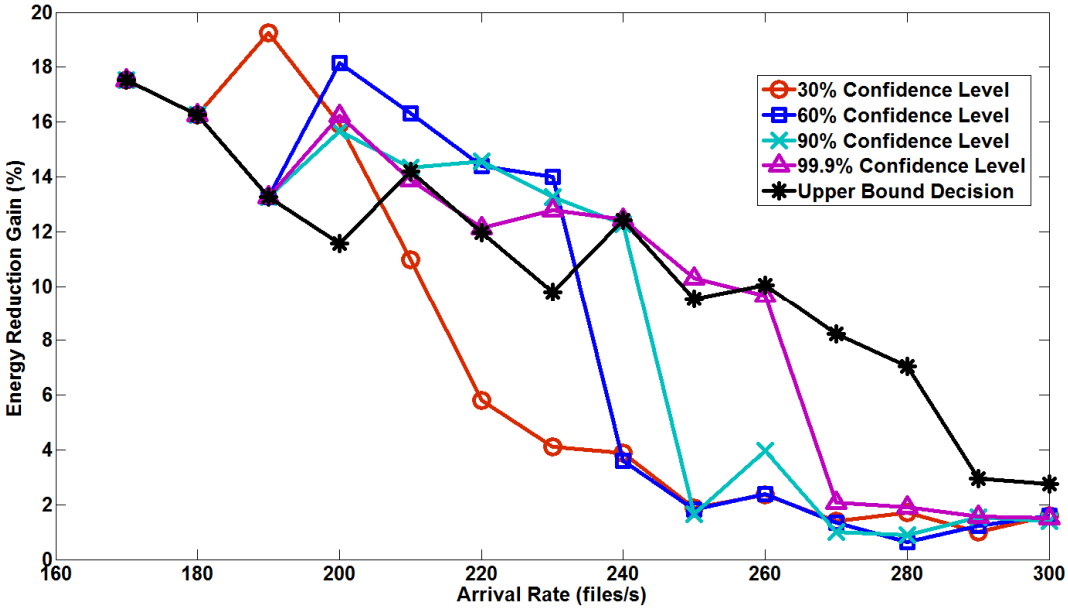


Figure 6.10 Energy Reduction Gain based on Beyond 2020 model

The increase in ERG when policy level modifications involve a change from higher order restriction to lower order restriction (i.e. Policy levels 4, 3, and 2) is because of the faster rate of depreciation in energy saving of higher order restrictions relative to lower order restrictions at medium and high traffic loads. However, it is not known beforehand at a random medium and high traffic load which policy level gives the best energy saving. The information that is available is that changing to lower policy level generally leads to better QoS while at low traffic load higher policy level leads to better ERG. However, a low order policy level may provide better ERG than higher policy levels at some higher traffic loads but the exact point where this occurs is not known beforehand. The available information is exploited here with the aim of restoring QoS within specified targets while achieving moderate energy saving. Extra information can be acquired to achieve better energy efficiency. The benefit of utilizing such extra information for improved energy efficiency is evaluated in the next chapter.

ERG up to 17% is achieved with all the confidence level options. Even when the confidence level choice results in the QEPU settling at the lowest level where sleep state is prohibited, energy saving is still achieved. This is in line with the available information that energy saving relative to the baseline approach is possible even without sleep state in a network based on advanced small cell BSs. Furthermore, the

99.9% confidence level based strategy is the best choice in terms of the energy efficiency when the whole traffic load range is examined. The upper bound decision only performs better than the rest in high traffic region above 260 files/s. This is not a contradiction, as the upper bound is related to the decision making regarding estimating the blocking probability, and not the balance of QoS and energy saving.

The 99.9% confidence level option achieves good QoS and energy efficiency, and also estimates the blocking probability better than the lower confidence level options. However, the waiting time before decision making is a constraint to settling for this choice. When quick adaptation decisions and good QoS is highly necessary, the 90% confidence level option will be the better candidate at the cost of lower energy efficiency.

The impact of periodic update of traffic data from ZBSs to the QEPU is also evaluated. The case with the QEPU making decisions based on the 90% confidence level is used as a typical example. When traffic data are received from the ZBSs, the adaptive algorithm is evaluated and decisions to change the policy level are made if the necessary maximum deviation criteria for carried traffic and blocking probability are satisfied. The QEPU is constrained to make decisions at regular intervals albeit with the defined confidence level (or degree of confidence). If the confidence level based criteria are not met at the latest event counters update, the QEPU waits for the next update.

Figure 6.11 shows the temporal blocking probability at 280 files/s and 90% confidence level under consideration of different periodic traffic data upload intervals. As mentioned earlier in section 6.2, collection of performance and event counters is possible for intervals between 10 seconds to 5 minutes [152]; thus this range is considered for the periodic traffic data upload to the QEPU. The marked points represent the moment policy level changes are made. Although, the update times are a few seconds, the confidence level constraint decisions from being made until the necessary conditions are met. The lower upload intervals of 10s, 30s, and 60s adapt to the 90% confidence level behaviour while the longer update intervals of 180s and 300s leads to longer decision times than the initial 90% confidence level strategy without periodic traffic data uploads. The longer decision times translate to

higher confidence levels than the defined 90% confidence level, since more statistics than necessary to achieve the 90% confidence level will be available at the point of decisions. Hence, irrespective of the upload intervals, the decision making is constrained to take place at instances when collected statistics can provide confidence intervals with the defined confidence level or higher.

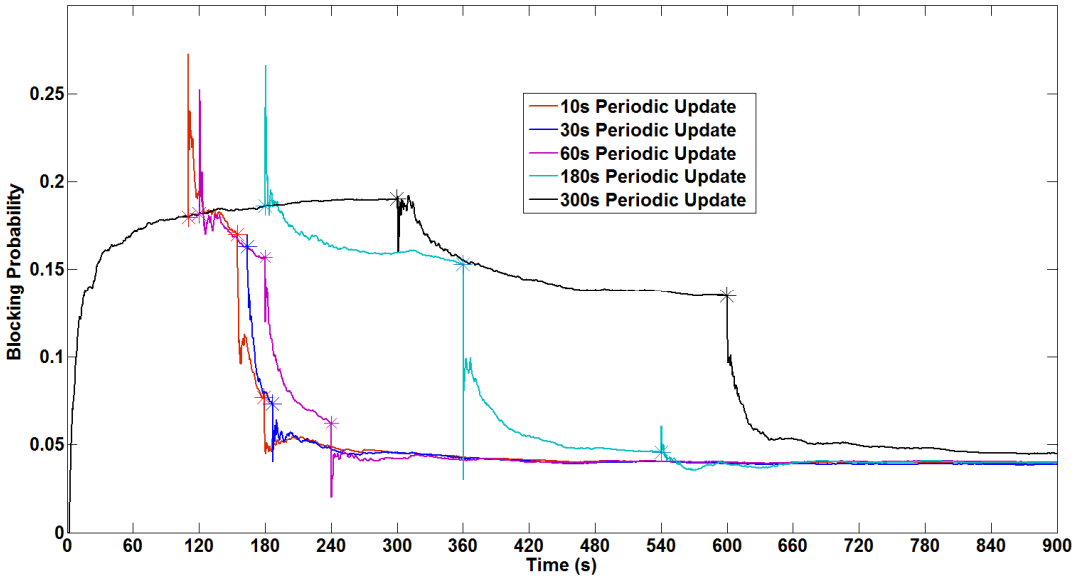


Figure 6.11 Temporal Blocking Probability Under Different Periodic Update Settings

6.5 Conclusion

In this chapter, a novel confidence level based adaptive joint resource and topology management strategy is presented. The partial centralised paradigm of managing resources is maintained with the QEPU modifying the RRM and TM policies using traffic statistics collected from ZBSs to maintain system blocking probability within specified threshold. It is proposed that the confidence level based criteria can both regulate the interval at which RRM and TM policies are modified and estimate the blocking probability, which is the QoS parameter considered, with a definite confidence level (e.g. 90% confidence level). More importantly, the confidence level based adaptive algorithm when combined with initial RRM and TM schemes developed in Chapter 4, can drive QoS within specified limits in near real time while still achieving moderate energy saving at medium and high traffic load regions.

The performance evaluation of the adaptive scheme on the BuNGee Architecture in terms of speed of response to QoS deterioration, blocking probability and energy efficiency confirms the projection about the scheme. The adaptive strategy can indeed regulate when decisions are made by variation of the confidence level. Also, decisions are shown to be made about blocking probability estimation satisfactorily by the scheme.

Furthermore, the scheme achieves energy efficiency even at medium and high load while maintaining the QoS within operating limits. The scheme extends range of operation to up to 280 files/s which is very close to 290 files/s achieved by the Highest SINR scheme and above the operating limits of previous schemes. The 90% confidence level is shown to be suitable for quick decision making, good QoS and moderate energy efficiency. In addition, the 99.9% confidence level can achieve equally good QoS and even better energy efficiency, however, the interval between decision epochs is longer.

Also, it is shown that when decision making are done in a periodic fashion, the confidence level criteria constrains decisions from being made unless statistics are deemed sufficient. Finally, this algorithm allows the network to restore QoS within defined limits in a self-organised manner without requiring further intervention from a human expert after the default setting.

Chapter 7. Linear Search and Database Aided RRM and TM Policy Selection

7.1 Introduction

Self-Organising Networks (SON) are autonomous networks that should be able to automatically adapt network parameters to improve performance including QoS [155]. Such automatic adaptation of parameters has been achieved in Chapter 6 through application of a confidence level based technique to improve blocking probability performance in a separation architecture heterogeneous network. QoS deterioration are detected and rectified online by adaptively modifying the RRM and TM policies at the central node. These policies are used to guide RRM and TM decisions of macrocells and small cells. This approach has been shown to effectively rectify QoS deterioration at all confidence levels considered while achieving moderate energy efficiency at medium and high traffic loads.

The online, adaptive scheme of Chapter 6 has been developed based on the known information that QoS improves with the reduction in the order of choice restriction and with the prohibition of sleep state transition rather than its permission. Furthermore, the insight that higher energy efficiency is achieved by a high order choice restriction at low traffic load but lower energy efficiency is achieved at some traffic load beyond the low traffic load region relative to lower order choice restrictions is also considered. However, the information regarding the location of such traffic loads where a high order choice restriction is a poorer energy efficiency choice is not exploited. Hence, as shown in Chapter 6, adaptation with a low confidence level can sometimes rectify QoS targets and achieve better energy efficiency than preferred higher confidence level based cases. It is envisaged that improvement in energy efficiency and better balance between QoS and energy efficiency can be achieved by exploiting addition information regarding the traffic load where different policies are more suitable options.

In this chapter, a linear search and database aided scheme is proposed for joint RRM and TM policy selection. The scheme searches through previously stored records of QoS and energy efficiency performance of the system offline and maps a combination of RRM policy and TM policy that best balances QoS and energy

efficiency to each traffic load previously experienced in the network. This mapping is stored in a database and it is subsequently utilised to select policies for new traffic load encountered by the system. The scheme is implemented at the central node, QEPU, which decides the RRM and TM policies to utilise at the zones as explained in previous chapters. Also, a lower bound on the number of active ABSs is derived based on the Erlang B queuing model. This is compared with the number achieved by the proposed scheme and used to verify the performance of the scheme.

The rest of the chapter is organised as follows. The system model utilised is described in section 7.2 while the proposed scheme is discussed in section 7.3. The lower bound derivation is presented in section 7.4, while simulation results are presented and discussed in section 7.5. Finally, the chapter is concluded in section 7.6.

7.2 System Model

The separation architecture considered in the previous chapters is still utilised in this chapter. Furthermore, the hybrid SON approach is still utilised with RRM and TM decisions involving the QEPU, ZBSs and ABSs like in Chapter 6. The partially centralised paradigm for RRM and TM is maintained here as well for this same reason. Specifically, the central node, QEPU, sets the RRM and TM policies while the ZBSs enforce the policies in their zones. A ZBS decides the resources on which ABS in its zone to utilise for serving an MS request as before. A ZBS also activates ABSs when QoS is not satisfied in the zone. Furthermore, overloaded ABSs activate neighboring ABSs in sleep state as before.

Furthermore, like in Chapter 6, a combination of the RRM policy and TM policy selected by the QEPU constitutes a policy level. However, unlike in Chapter 6, the QEPU does not progress through policy levels adaptively rather it goes through a training phase where it utilises the previous QoS and energy efficiency performance metrics to map the policy levels to traffic loads. Traffic load levels ranging from low to medium to high are considered across the range supported by the system. This is done to train the QEPU and enhance the policy level selection for better QoS and energy efficiency performance.

In chapter 6, it is demonstrated how QoS deterioration can be corrected by adapting policy levels in a step by step manner. However, in this chapter, the focus is selecting the best policy level that can balance QoS and energy efficiency. Thus, choice restrictions up to the reasonably highest level beyond the third choice restriction considered in Chapter 6 are included. In this case, up to the 4th choice restriction is considered but not beyond. This is because for the BuNGee architecture an MS can be line of sight with at most four ABSs and can achieve better QoS with a line of sight ABS than a non-line of sight type. In addition, the higher the order of choice restriction utilised, the poorer the QoS experienced in the network especially beyond low traffic regions.

A total of five policy levels are available to the QEPU. The policy levels and typical mapping to the traffic load are shown in Figure 7.1. As before, level 1 is the only one that involves SLEEP OFF which implies prohibition of sleep state transitions. SLEEP ON and SLEEP OFF constitute the TM policies as described in Chapter 6. It is important to note that policy level 1 is equivalent to the highest SINR scheme without topology management. This is because 1st choice ABSs serve MSs always under 1st choice restriction and sleep state transitions are not permitted. As shown in the Figure 7.1, a policy level can be mapped to more than one traffic load level or not mapped to any one at all depending on its performance with respect to the traffic load considered.

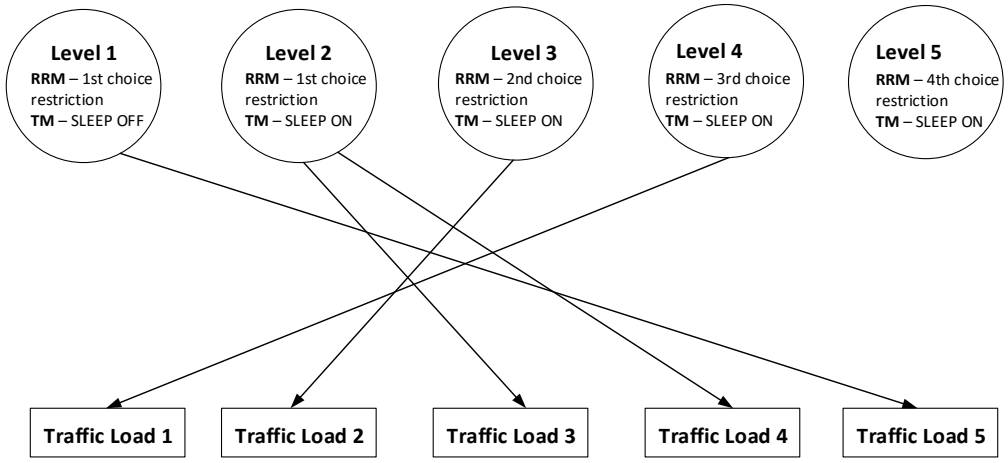


Figure 7.1 Policy level Mapping to Traffic Load

7.3 Linear Search and Database Aided RRM and TM Policy Selection

Linear search, also called sequential search, is a simple method for searching sequentially in a list for an item until the desired one is found [159]. [160] used it for optimal power allocation in underwater wireless sensor networks while [161] used it to determine the device to device (D2D) communication links that lead to the lowest total power consumption. Linear search can be time consuming, since the time requirement increases linearly with the number of items in the list [162]. However for a small list of items it is both simple [159] and practical [163]. Hence, it is appropriate for the selection of a suitable policy level out of five possible candidates considered in this chapter.

Specifically, linear search is used to select the policy level that leads to low energy consumption while keeping the average file transfer delay low as well. Unfortunately, a choice restriction that leads to the lowest energy consumption often leads to higher delay and vice-versa. This is due to the tradeoff between energy efficiency and delay. Hence, a suitable RRM policy (and consequently policy level) that balances these tradeoff is desirable.

The linear search method is first used to map policy levels to different traffic loads offline across the range supported by the system based on previously stored energy consumption rating (ECR), delay and blocking probability measurements. Subsequently, a database of this offline mapping is created and used to decide policies for traffic load experienced in the system in the future. The proposed linear search method is discussed in more details next and then followed by an explanation of the application of the mapping database for policy selection.

7.3.1 Proposed Linear Search Method

The objective of the linear search method presented in this chapter is to enable the selection of a policy level that balances QoS and energy efficiency satisfactorily. In order to achieve this, the energy consumption rating (ECR), average file transfer delay and blocking probability performances for different traffic loads are estimated for different policy levels. This represents a training phase for the network, which can be carried out at the earlier stages of operation of the network. The goal is to

meet a target blocking probability while keeping ECR and average file transfer delay low. More than one policy level might satisfy the blocking probability at a given traffic load, however some might lead to low delay but high ECR while others might lead to high delay but low ECR. The search is done to find a policy level that avoids these two extremes but instead balances the delay and ECR such that it is not the lowest ranked in either delay or ECR performances. As much as possible, it should be close to the top in ranking in both delay and ECR performances. The blocking probability, ECR and average file transfer delay performance metrics can be obtained from traffic statistics and key performance indicators stored at the Operation Support System (OSS) which are obtained from event and performance counters at the base stations in a similar manner as explained in Chapter 6.

The average file transfer delay increases with the order of policy level because delay increases with the order of choice restrictions as shown in Chapter 4, whereas, sleep state prohibition is shown to lead to better delay than sleep state permission in Chapter 5. The ECR is stored in the database rather than the energy reduction gain (ERG) or effective energy saving (EES) since it can be evaluated without considering a baseline scenario like the other two. Hence, the database creation and consequently selection of a policy level that balances QoS and energy efficiency can be done independently of a baseline scheme. The ERG is proportional to the ECR as earlier stated in (2.9) in Chapter 2.

Therefore, for a given baseline value, $ECR_{baseline}$, the ERG depends only on the ECR value of the test scheme. In this case, evaluation of the system under different policy levels constitutes consideration of different test schemes. From (2.9) it can be deduced that the higher the ECR of the test scheme the lower the ERG and vice versa. At low traffic load, the ERG increases with the order of policy level since ERG increases with the order of choice restriction and sleep state permission leads to better ERG than its prohibition as shown in Chapters 4 and 5 respectively. Therefore, at low traffic load, ECR decreases with the order of policy level due to the inverse relationship between ERG and ECR. Beyond low traffic load, the ERG does not exhibit this linear relationship with the order of policy level, since as also established in Chapter 4, a low order choice restriction is eventually better than a higher order

choice restriction at some medium traffic load. Hence, the ECR does have a linear relationship with policy level beyond low traffic region.

For each traffic load, a set of measured delay and ECR values, $\mathrm{D} = \{ d_1, d_2, \dots, d_i, \dots, d_{n-1}, d_n\}$ and $\mathrm{E} = \{ e_1, e_2, \dots, e_i, \dots, e_{n-1}, e_n\}$ are created. n is the number of policy levels that satisfy the blocking probability target. The policy levels that do not satisfy the blocking probability target are not considered in the selection process. The set of considered policy level, L, in increasing order of level is defined as $\mathrm{L} = \{ l_1, l_2, \dots, l_i, \dots, l_{n-1}, l_n\}$. Furthermore, policy levels are sorted in ascending order of delay, D, and ECR, E, to create delay and ECR sets of policy levels respectively. The delay policy level set, Q, and the ECR policy level set, Z, are defined as $\mathrm{Q} = \{ q_1, q_2, \dots, q_i, \dots, q_{n-1}, q_n\}$ and $\mathrm{Z} = \{ z_1, z_2, \dots, z_i, \dots, z_{n-1}, z_n\}$ respectively; where $q_i, z_i \in$ L. Since L, Q, and Z are all ordered set (ascending order), When L is compared with Q or with Z, for $i, j \in [1, n]$, if $q_i = l_j$ then i = j always; while if $z_i = l_j$ then i = j or i ≠ j. This is because the delay policy level set, Q, will always have the same element in the same position as the considered policy level set, L, since delay increases with increasing order of policy level. Whereas the ECR policy level set, Z, may or may not have the same element in the same position as the considered policy level set, L, since ECR does not always increase with the order of policy level.

The linear search is conducted along the elements of ECR policy level set, Z, which is not always sequential. This is done one element at a time and at each step the current element and the preceding elements are checked for matches over the same range in the delay policy level set, Q. For each step *m* of the search, corresponding to the m^{th} element of Z, a subset of Z (Z^*) is created such that:

$$Z^* \subseteq Z \tag{7.1}$$

$$Z^* = \{ z_1, z_2, \dots, z_i, \dots, z_{m-1}, z_m\} \; ; m \leq n \tag{7.2}$$

This is compared to the same range of the delay policy level set to find a policy level with low delay and low ECR (which is equivalent to high energy efficiency). Similarly, a subset of Q (Q^*) is created such that:

$$Q^* \subseteq Q \tag{7.3}$$

$$Q^* = \{ q_1, q_2, \dots, q_i, \dots, q_{m-1}, q_m \} \ ; m \leq n \tag{7.4}$$

As mentioned earlier, policy levels in both delay and ECR policy level sets are arranged in ascending order of the magnitude of their delay and ECR measurements. Hence, if a match between the two sets is found at an early stage of the linear search, the choice of policy level will lead to low delay and good energy efficiency. The matched policy level set, P, at the mth stage of the linear search is given by:

$$P = Z^* \cap Q^* \tag{7.5}$$

If $n(P) = \emptyset$ that is no match found, the search progresses onto to the $(m+1)^{th}$ step . However, if $n(P) = 1$ then one match is found and the final policy level selected, P^*, is given by:

$$P^* = P = l_i; \ \ l_i \, \epsilon \, L \tag{7.6}$$

If $n(P) > 1$ more than one match is found. Hence the matched policy level set, P, is a set of policy levels as follows:

$$P = \{l_a, l_b, l_c, \dots, l_i, \dots, l_k\} \, ; \ a, b, c, \dots, i, \dots, k \, \epsilon [1, n] \, ; \, k \leq n \tag{7.7}$$

The equation in (7.7) implies that the policy levels in P can be any of the policy levels, however the number of element of P cannot exceed the total number of policy levels considered. In this situation, where more than one match is found, the policy level with the lowest delay measurement is selected to achieve the best QoS as explained below.

Assuming I is set of the order of the matched policy levels in (7.7), then $I = \{a, b, c, \dots, i, \dots, k\}$ and $k \leq n$. Since the delay increases with the order of policy level, the policy level with the lowest order should be selected. Hence, the order of the selected policy level, I^*, is given by:

$$I^* = \min_i (I); \ \ \forall i \, \epsilon \, I \tag{7.8}$$

Hence, the selected policy level under multiple matches, P_u^*, is given by :

$$P_u^* = l_{I^*}; \qquad l_{I^*} \epsilon L \tag{7.9}$$

In Figure 7.2 shown below, a typical mapping scenario is shown for illustration purposes. The policy levels are arranged in ascending order of the delay and ECR measurements in the delay and ECR rows respectively. The search is started with the first element of the ECR row, which is policy level 4. When this is compared with the first element of the delay row no matching is found since this element is policy level 1. Thus, the search is continued and in the next step the second element and the first element of the ECR row, i.e. policy levels 3 and 4 respectively, are checked for a match in the delay row considering similarly the second and first elements of this row. Also, no match is found at this at this step and the search is continued. At the third step, the third, second and first elements of the ECR row are compared with the third, second and first elements of the delay row. In this case two matches are found as the policy level 3 and policy level 2 are both found in the delay row at this stage. Since policy level 2 has lower delay it is selected as the suitable policy for the traffic load for which the mapping is being carried out and the mapping is stored in the database.

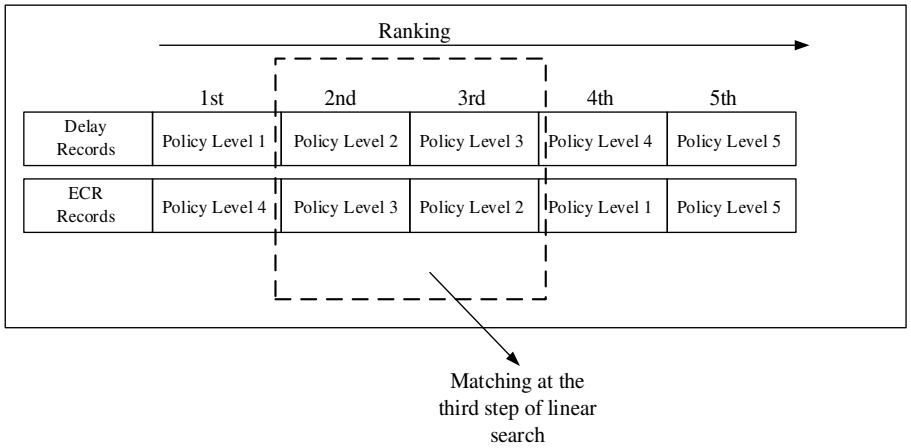


Figure 7.2 Policy Selection with Linear Search

The flowchart for the implementation of this approach is shown in Figure 7.3 for any given traffic load with related average file transfer delay, blocking probability and ECR measurements already in the database. The database is assumed to be available at the Operation Support System (OSS) and readily accessible to the central node,

QEPU, as explained in Chapter 6. It is important to note that the partially centralised paradigm of RRM and TM is still maintained. The QEPU only sets policies for long timescale average traffic load while ZBSs make RRM and TM decisions at short timescale of user arrivals and departures. The mapping of policy levels to traffic load is done over a range of traffic load and utilise for policy selection for different traffic load not already characterised by the system. How this is achieved is explained next.

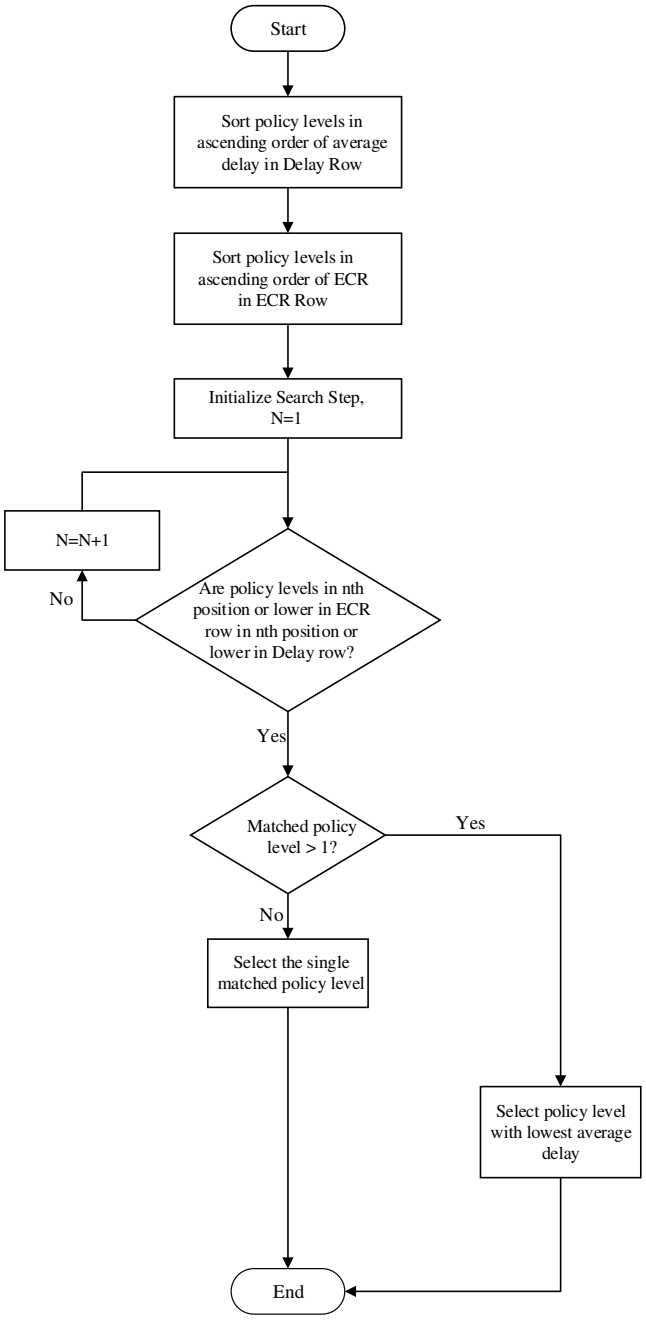


Figure 7.3 Flowchart of the Linear Search Scheme for Policy Level Selection

7.3.2 Application of Policy Level Mapping for Future Policy Level Selection

The outcome of the offline mapping stage is a series of traffic loads mapped to suitable policy levels, which is then stored in a database in the OSS. Policy levels are selected for new traffic loads encountered by the system by utilizing this mapping. It is expected that new traffic loads not already mapped to a policy level will fall before, after or between characterised traffic loads in terms of magnitude as shown in Figure 7. 4.

For a new traffic load falling before or after all characterised load like "new traffic load 1" and "new traffic load 3" in Figure 7.4, there is only one policy level option to choose, the traffic load after or before it respectively. However, for a new traffic load that falls in between already characterised traffic loads like "new traffic load 2" two policy level options can be selected, the one before it or the one after. Intuitively, to maintain good QoS the policy level matched to the traffic load after it should be selected. This is because it is expected that if the policy level maintains a good QoS for a higher load it will be suitable for the lower traffic load. Before evaluating the performance of the proposed scheme, a lower bound on the number of active ABSs required for a given traffic load based on the Erlang B model is derived.

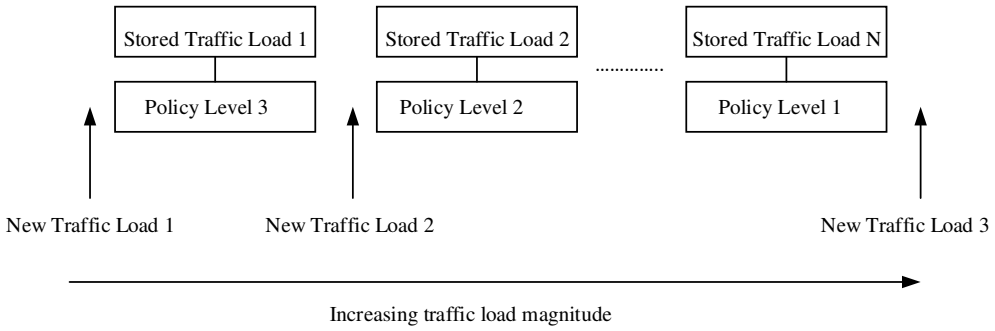


Figure 7.4 Matched Policy Levels for Future Policy Selection

7.4 Erlang B based Lower Bound for Number of Active ABSs

The Erlang B model, as explained earlier in Chapter 3, is an analytical model based on queuing theory. Queuing theory provides analytical tools for evaluating performance of systems where user requests arrive randomly and are served with limited resources [135]. The random processes that describe user arrivals and the

nature of the service facility must be known in order to fully describe a queuing system [164]. Furthermore, queuing systems are usually described in terms of arrival process, service process, number of servers, system capacity (queue length plus the number of servers) and maximum number of potential users [129]. The arrival process is specified in terms of the distribution of inter-arrival times between users while the service process is described in terms of the distribution of the service time (time taken to serve a user request) [129].

The Erlang B model is based on a queuing system with exponentially distributed inter-arrival times, exponentially distributed holding times, and infinite number of potential users [133]. In addition, the number of servers and system capacity are equal and therefore there is no queue for users who cannot be served immediately [129]. The exponential inter-arrival times leads to a Poisson distribution of the number of users within a period of time with an average arrival rate λ and the exponential service time distribution has an average of $1/\mu$ [165]. The system can be represented by a birth-death process captured by a state transition rate diagram as shown in Figure 7.5. The state of the system is defined in terms of the number of users currently being served (or equally the number of busy servers) [129]. The birth rate is the transition rate, λ, from a current state to the next state with higher number of busy servers while the death rate, $k\mu$ (where k is the number of busy servers), is the transition rate to a previous state with lower number of busy servers.

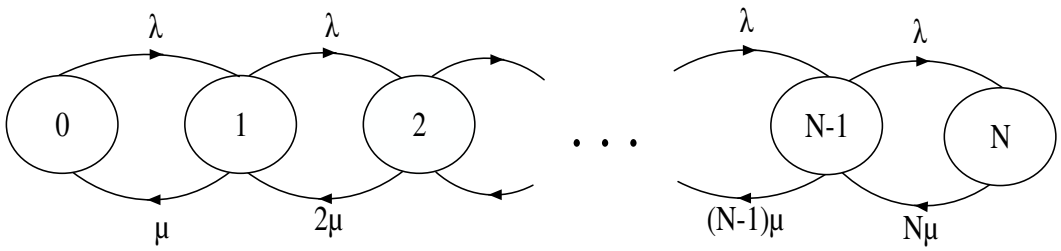


Figure 7.5 State transition diagram for the Erlang B Model [164]

The steady state probability of a user arriving and meeting a given number of busy servers can be obtained when the system is in equilibrium. As the number of servers is N, a new user arriving into the system when there are already N busy servers will be denied service or blocked. Hence, the probability of being blocked is the same as

the probability of all servers being busy. This system has been applied in dimensioning telephone exchanges and the probability in this case is equivalent to the probability of all available telephone channels being busy, P_B. This probability is also referred to as the grade of service and is given by the Erlang B formula as follows [133]:

$$P_B = \frac{\frac{A^N}{N!}}{\sum_i^N \frac{A^i}{i!}} \tag{7.10}$$

where A is the average offered traffic in Erlangs and $A = \lambda/\mu$. N is the total number of available channels. Tables and graphs based on the Erlang B formula have been created [164]. This has been used to determine the Erlangs of traffic that can be carried by a number of channels at a specific probability of blockage or grade of service [133]. Although, data transmission is considered in this work and not voice, the Erlang B model is utilised from the perspective of the time the transmitted file occupies the frequency channels in the network i.e. the service time. This channel occupancy perspective is used to determine the number of ABSs that needs to be active in the network to support a given average arrival rate.

Furthermore, interference is not considered and once users are connected, it is assumed that they can operate at the highest data rate supported by the ABSs. This models a perfect channel but with user proximity in small cell scenarios and interference mitigation, performance approaching this perfect model is possible. The Erlang B formula assumes that when a user's call is blocked he will wait for a long period before reattempting the call; hence, the retrial is seen as a new call [133]. Furthermore, the Erlang B formula has been derived based on exponential service time distribution. Nevertheless, the Erlang B model is valid for constant service time and gives very similar performance at low blocking probability as its enhanced version, the Extended Erlang B (EEB) model, which is proposed for retrials [166]. Hence, the Erlang B model is suitable for this work where constant service time and retrials are considered under low blocking probability regime.

In the separation architecture considered in this work, the ZBS in each zone serves the user requests in the zone with resources of the ABSs in that zone. Furthermore, some ABSs belong to more than one zone and are more central in the service area than those associated with one zone. For example in Figure 7.6, the ABSs highlighted in “black” in Zone 1 of the architecture are more central in the service area compared to the ones in “white”. The clustering concept used in this work prefers utilization of such more central ABSs to the ones closer to cell edge. This enables the ABSs associated with more than one zone to be available to serve users in the associated zones while those associated with only one zone can be switched off (i.e. switch to sleep state). This will enable energy saving at low and medium traffic loads.

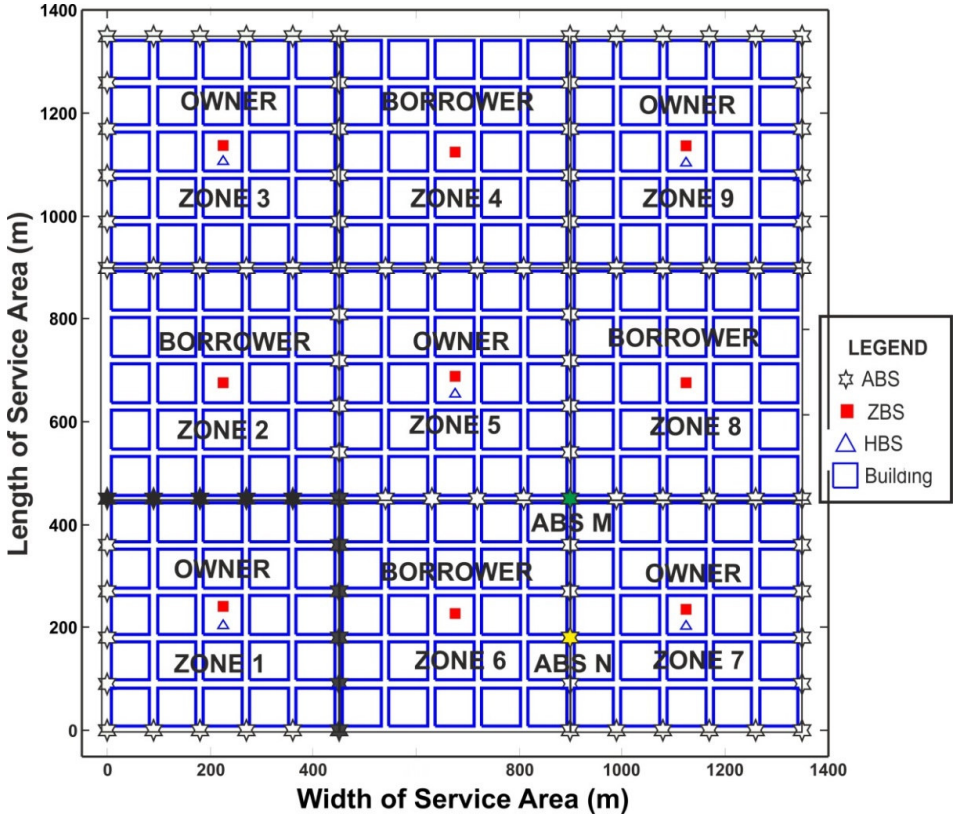
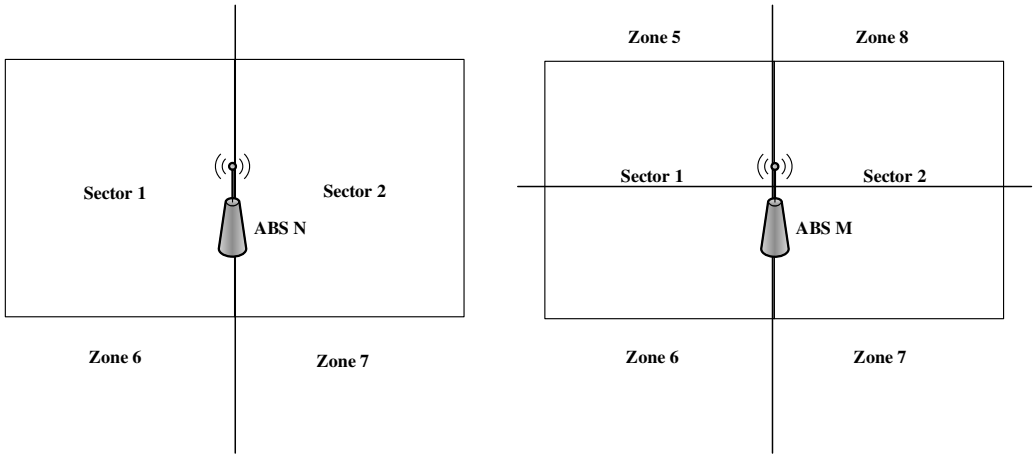


Figure 7.6 Central ABSs in BuNGee Architecture

Furthermore, each ABS has two directional antennas pointing in opposite directions. ABSs serving two zones have two sectors, one in each zone, supported by the directional antennas as shown in Figure 7.7 (a) for ABS N (shaded in yellow) in Figure 7.6 as an example. While those at the boundaries of four zones still have two

sectors but half of each sector covers a single zone, this is illustrated in Figure 7.7 (b) for ABS M (shaded in green) in Figure 7.6 as an example. The coverage of each sector is approximately represented by a rectangle due to the shape of the antenna pattern in each direction as shown in Chapter 3. If it is assumed that ABSs are owned by the zones with HBSs, then zones 1, 3, 5, 7, and 9 own ABSs and can be termed *owner zones* as shown in Figure 7.6. Zones 2, 4, 6, and 8 can be considered as borrowing resources from the other zones in the form of the sector antennas serving these zones and can be termed *borrower zones* also shown in Figure 7.6.



(a) ABS associated with two zones (b) ABS associated with four zones

Figure 7.7 ABSs and zones served

In terms of energy saving, for a given traffic load in the worst case scenario, the sectors used to serve users in an *owner zone* will belong to a different ABS from those used in a *borrower zone* to serve users as much as possible. In the best case scenario, the ABSs used in an *owner zone* to serve users will be used in a *borrower zone* to serve users as much as possible. On one hand, the worst case scenario leads to the highest number of active ABSs and few ABSs can be switched off to save energy. On the other hand, the best case scenario leads to the lowest number of active ABSs and a lot of ABSs can be switched off to save significant energy especially at low load. The best case scenario is therefore the lower bound on the number of active ABSs and highest energy saving.

For a given average arrival rate, λ, in the whole network, the average arrival rate in each zone is λ/m_T (where m_T is the total number of zones) since users (MSs) are uniformly distributed. As stated earlier the Erlang B Formula Table can be used to

determine the number of Erlangs, A, that can be carried by a number of channels at a given blocking probability, P_B. Under the best case scenario, the number of active ABSs in the *owner zones* determines the number of active ABSs in the network. This is because it is assumed that a *borrower zone* utilises ABSs also being utilised to serve MSs in an *owner zone* at all times.

In [105], the number of ABSs that need to be active is determined based on the maximum level of traffic that can be carried by individual ABSs to meet a target blocking probability or lower. A similar approach is considered here, however, individual sectors of ABSs are considered instead. For a sector of an ABS with a total of n channels and serving an *owner zone*, the maximum Erlangs of traffic, A_s, that can be carried by this sector at a given blocking probability can be determined from the Erlang B Table. Since, each *owner zone* experiences an average arrival rate of λ/m_T, the total traffic, A_z, carried by each *owner zone* (which is the same as in a *borrower zone*) is given by:

$$A_z = \frac{\lambda}{m_T \mu} \tag{7.11}$$

Hence, the number of sectors, N_s, similarly with n channels required to carry a total of A_z Erlangs of traffic in an *owner zone* is given by:

$$N_s = \frac{A_z}{A_s} = \frac{\lambda}{m_T \mu A_s} \tag{7.12}$$

For simplicity, each sector is assumed to serve only one zone, so the coupling between zones in terms of the sectors of ABSs associated with two zones (see Figure 7.6 (b)) is neglected. Hence, the total number of sectors, $N_{s,o}$, that will be active in all *owner zones* when the network serves an average arrival rate of λ is given by:

$$N_{s,o} = \frac{\lambda\, m_o}{m_T \mu A_s} \tag{7.13}$$

where m_o is the number of *owner zones.* Since, it is assumed that the *borrower zones* will utilise the active ABSs in *owner zones* in the best case scenario, therefore, the total number of active ABS required is the same as the total number of sectors

required in the *owner zones.* Hence, the total number of active ABSs, N_{ABS}, is given by:

$$N_{ABS} = \frac{\lambda\, m_o}{m_T \mu A_S} \tag{7.14}$$

The average service time, $1/\mu$, is obtained from the from the file size, F, and transmission rate, R, as follows:

$$\frac{1}{\mu} = \frac{F}{R} \tag{7.15}$$

This is used to obtain the period of occupancy of the channel by a file.

Significant energy saving can be achieved by reducing the number of active ABSs whilst meeting the target QoS conditions. This bound on number of active ABS is compared with the performance of the proposed linear search and database aided scheme.

7.5 Simulation Results and Discussion

The performance of the proposed scheme is evaluated with Monte Simulation in the BuNGee Separation architecture which is implemented in the same way as in previous chapters. 5 HBSs, 9 ZBSs and 112 ABSs are deployed in the network, while 6,000 MSs are distributed uniformly outdoors along the streets. Users arrive into the system based on a Poisson process with inter-arrival times exponentially distributed. Each user arrives with a request to upload a fixed file size of 2MB. The acceptable range of operation for the system is the region where the blocking probability is less than 5% and the energy reduction gain (ERG) is above zero.

The training phase is first evaluated. This involves operating the system over a range of traffic loads with different policy levels one at a time in order to match policy levels to traffic load using the linear search method. Then, the system is evaluated with a different set of traffic loads to test the suitability of the matched policy levels stored in the database. The performance of the scheme in terms of reducing the active set of ABSs is verified with the derived Erlang B based bound. The simulation parameters are shown in Table 7.1. It is important to note that the RRM and TM

decision in the local zones still follow the same procedure as in previous chapters, only the selection of the RRM and TM policies (equivalently selection of policy levels) is modified.

Policy level 1 is equivalent to the baseline scheme used in this work, which is the Highest SINR scheme without TM. Thus the Energy Reduction Gains (ERGs) of other policy levels are evaluated relative to the policy level 1 in the training phase. The ERG is calculated from the ECR measurements saved in the database. The training phase QoS evaluation is similar to the IA-CCR evaluation done in Chapter 4 except that in addition the results are now utilised to map RRM and TM policies. The QoS performances (blocking probability and delay) are similar because the ABS bandwidth and transmit power are the same as before. However, the energy efficiency performance is different, as the Beyond 2020 model is considered rather than the Han Model. The Beyond 2020 model is used to enable prediction of performance in future systems after the Year 2020 with the assumption of better load dependency, moderate static power consumption and more conservative, lower sleep state power consumption than the Han model.

Table 7.1 Simulation Parameters

Parameter	**Value**
Deployment area dimension	1350m×1350m
Street width	15 m
Building block size	75m×75m
ABS antenna height	5m
MS antenna height	1.5m
Carrier Frequency	3.5GHz
MS Transmit Power	23dBm
ABS Maximum Gain	17dBi
Noise Floor	-114dBm/MHz
Call Admission SINR	10dB
Minimum SINR for Reception	1.8dB
SINR for highest throughput	21dB
Training phase average arrival rates	5, 25, 50,75, 100, 125, 150, 175, 200, 225, 250, 275, 300
Validation phase average arrival rates	17, 39, 69, 83, 113, 136, 160, 189, 234, 289
File size	2 Mbytes
Target Probability	0.05 (5%)

The training phase QoS performance of the system in terms of the blocking probability and the average file transfer delay are shown in Figures 7.8 and 7.9 respectively. Generally, both blocking probability and average file transfer delay improves as the policy level is decreased from level 5 to level 1. This is due to improvement in QoS with reduction of the order of choice restriction and with change from sleep state permission to prohibition. As mentioned earlier in section 7.3.1, for each policy level when the target QoS is exceeded, the policy level is no longer evaluated for the traffic load in question. The target blocking probability is 5% and this is exceeded at different traffic loads under the different policy levels. This implies some policy levels are not even usable at some traffic loads, for example policy level 5 is not usable at an arrival rate of 200 files/s and beyond.

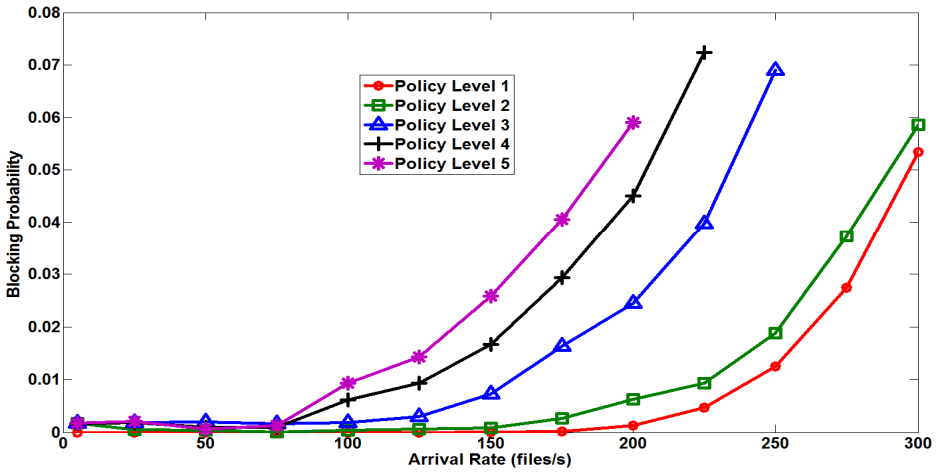


Figure 7.8 Training Phase Blocking Probability for Different Policy Levels

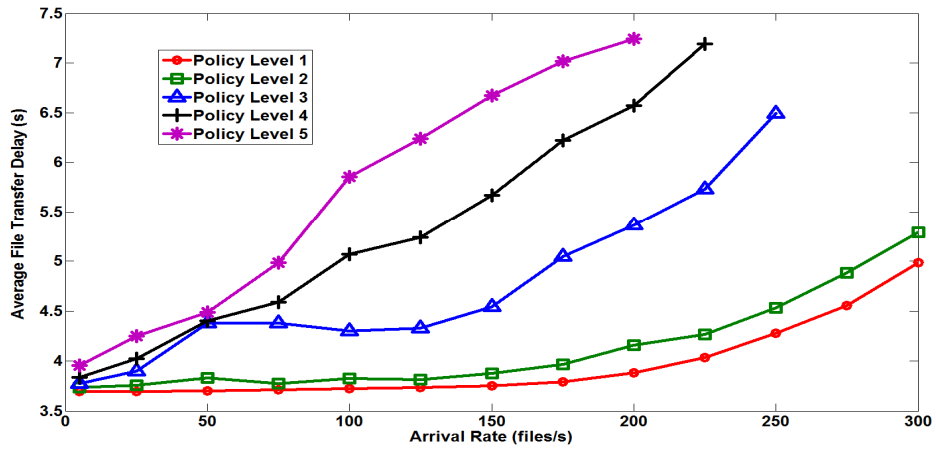


Figure 7.9 Training Phase Average File Transfer Delay for Different Policy Levels

The energy efficiency performance in terms of ERG, of policy levels 2, 3, 4, and 5 relative to the policy level 1 is shown in Figure 7.10. Generally, a high policy level achieves higher ERG at low traffic load than a lower policy level but beyond low traffic load the lower policy level provides better ERG eventually at some traffic load. This is similar to the performance with the Han Model in Chapter 4 except that the magnitude of the ERG is lower due to the more conservative, lower sleep state power consumption of the Beyond 2020 model. It can be seen from the comparison of Figures 7.9 and 7.10 that the policy level that achieves the lowest average file transfer delay, at a particular traffic load, may not achieve the highest energy efficiency, ERG. This is why the linear search scheme is used to select a policy level that balances the delay and ERG as stated earlier.

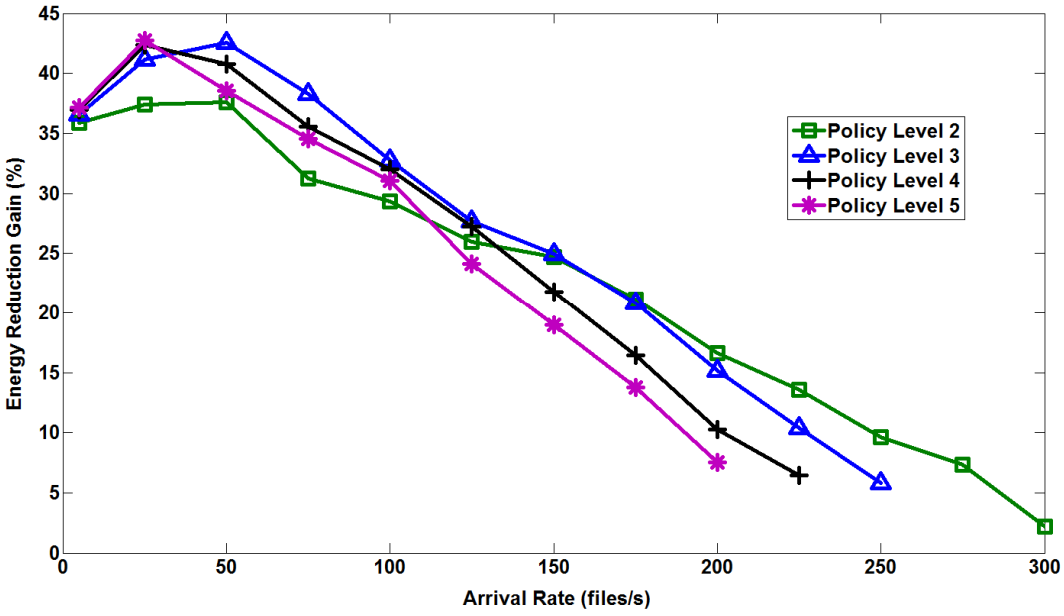


Figure 7.10 Training Phase Energy Reduction of Policy levels relative to Policy Level 1

The policy levels selected for each training phase traffic load under the proposed scheme, Linear Search and Database Aided Selection, is shown in Figure 7. 11. The optimal selection of policy levels based on the adaptation concept of Chapter 6 is also considered. In Chapter 6, the blocking probability is estimated with a specified confidence interval at a given confidence level. For very high confidence levels the estimation is close to the actual value. Whenever the estimated blocking probability exceeds the target value, the policy level is changed to a lower one. For the optimal

case, it is assumed that the estimation is perfect. In Figure 7.11, Different policy levels are selected in most cases under the proposed scheme and optimal adaptation method. In contrast to the optimal adaptation approach, the proposed scheme selects lower policy level even when higher policy levels meet the target blocking probability. This is done to achieve a better balance QoS and energy efficiency.

The performance of the proposed scheme in terms of QoS and energy efficiency are compared to the performance of the optimal adaptation based selection and previous energy efficient RRM and TM schemes in the literature for the training phase traffic load. The previous schemes in the literature are Highest SINR with One Neighbour On and the Capacity Based Channel Allocation (CBCA) schemes. As explained in Chapter 4, the Highest SINR with One Neighbour On scheme assigns resources of the ABS with the highest SINR to users. Also, it turns ABSs on or off if traffic load threshold criteria are satisfied within a period of time (10s) and at least an adjacent ABS is kept on. The CBCA scheme assigns resources of the ABS with the current highest traffic load to users. The initial modified TM scheme in this work without the adaptive feature is used to turn off ABSs for the CBCA scheme. The baseline scheme, Highest SINR without TM, assigns users to the ABS with the highest SINR but keeps all ABSs on always. The ERG is evaluated relative to the baseline scheme.

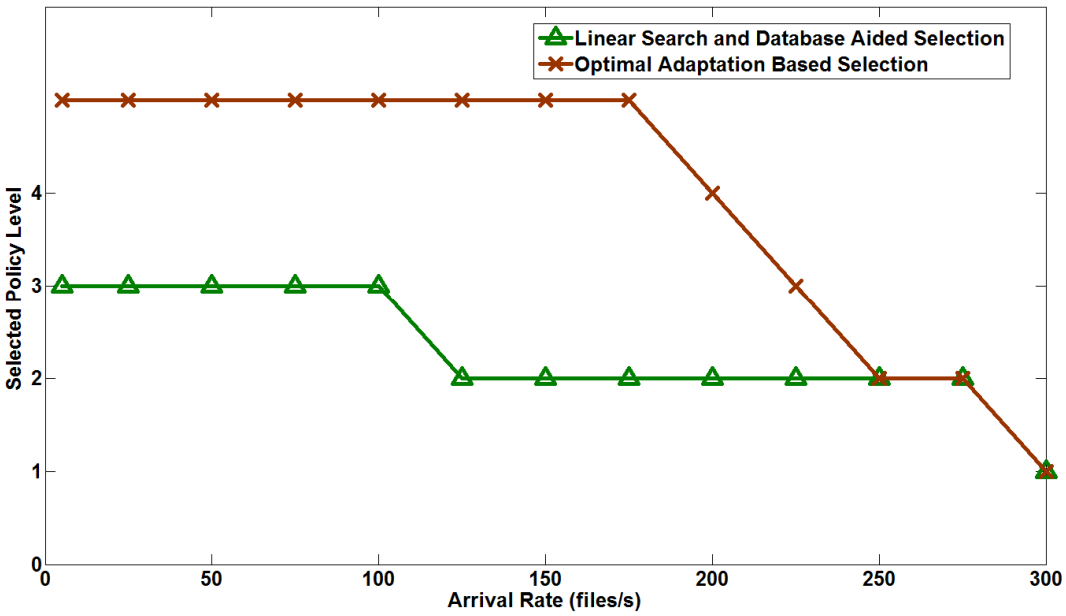


Figure 7.11 Selected Policies for Training Phase Traffic Loads

The blocking probability and average file transfer delay performance of the schemes with respect to the training phase traffic load are shown in Figures 7.12 and 7.13. The baseline scheme achieves the best blocking probability and average file transfer delay performance since all ABSs are always available and MSs connect to the highest SINR choice. The optimal adaptation based selection keeps the blocking probability below the defined 5% target just like the proposed scheme as shown in Figure 7.12 up to 275 files/s supported by the baseline scheme. However, the Highest SINR with One Neighbour On and the CBCA schemes exceed the target blocking probability at 250 files/s and 200 files/s respectively. This is because the policy level selection strategies modify policy levels (i.e. RRM and TM policies) with traffic load experienced to maintain target blocking probability whereas the other schemes provide no means to do this once it is exceeded. CBCA exceeds blocking probability earlier than other schemes because it permits any choice of ABS to be selected. This leads to higher interference in the network than other schemes especially at high traffic load.

In terms of average file transfer delay, the proposed scheme achieves better performance overall than other schemes except the baseline scheme as shown in Figure 7.13. This is because unlike the other schemes, keeping delay low is jointly considered with energy saving while maintaining target blocking probability. This consideration of low delay is also responsible for the lower blocking probability of the proposed scheme relative to the other schemes (as shown in Figure 7.12). The optimal adaptation approach focuses on meeting target blocking probability while achieving energy saving. As long as the target blocking probability is not exceeded no policy change is done. This is the reason for the higher delay performance and also blocking probability relative to the proposed scheme. The selection of first choice ABSs always by the Highest SINR with One Neighbour scheme keeps the delay low below 50 files/s. This is because sufficient ABSs are kept on based on the requirement to have at least an adjacent ABS on before switching off any ABS. However, at higher traffic loads more ABSs are needed, but the traffic load threshold and time criteria for switching on ABSs prevents switching on sufficient ABSs. Hence, higher blocking of user requests occur leading to file retransmission and higher file transfer delay. The CBCA scheme leads to higher delay than other

schemes due to the permission of any choice of ABS which results in lower SINR choices and lower data rates.

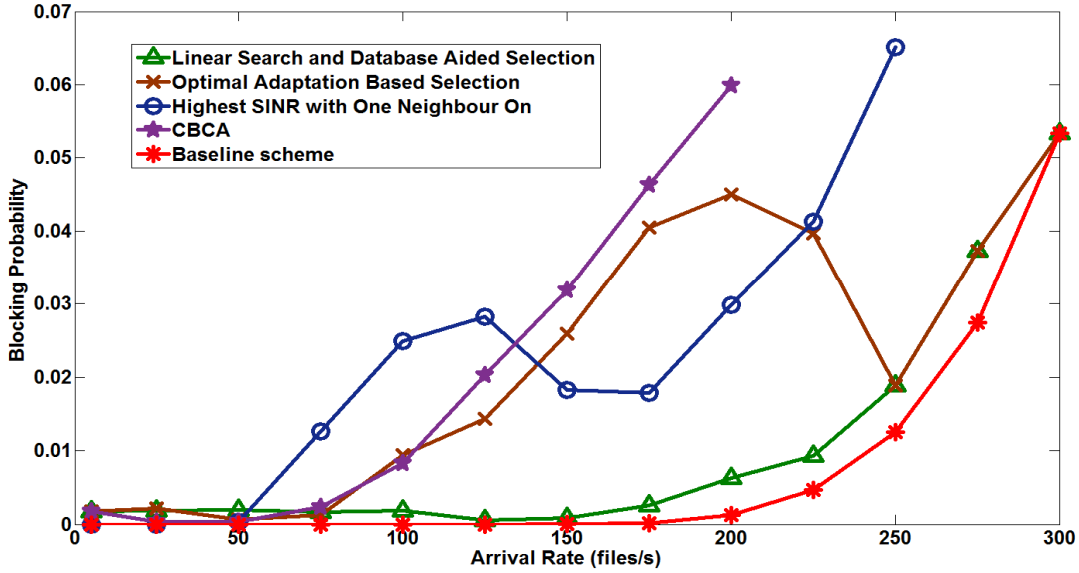


Figure 7.12 Blocking Probability for Training Phase Traffic Loads

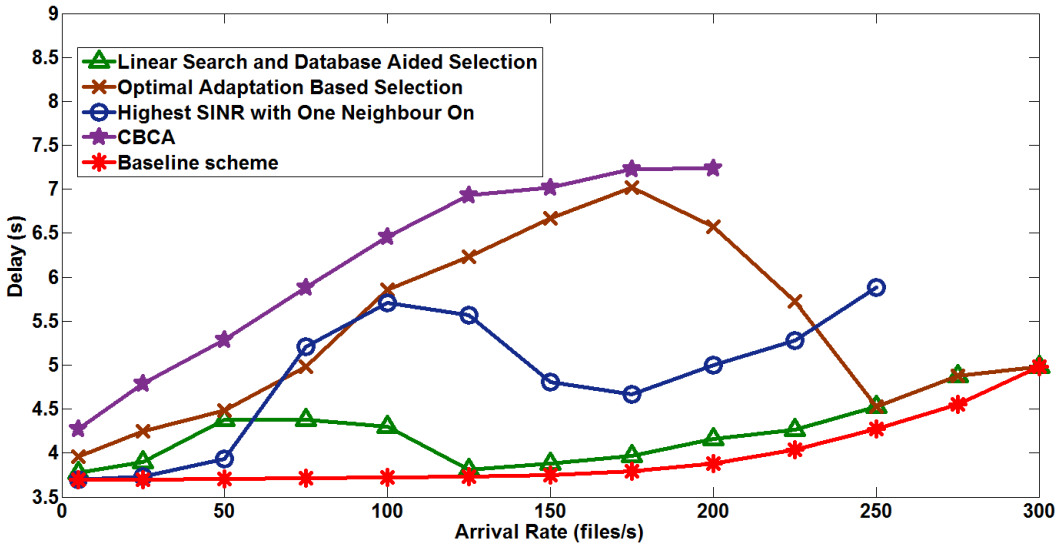


Figure 7.13 Average File Transfer Delay for Training Phase Traffic Loads

In terms of energy efficiency, the proposed scheme achieves higher ERG than the Highest SINR with One Neighbour scheme at low traffic load of less than or equal to 50 files/s as shown in Figure 7.14. However, the Highest SINR with One Neighbour scheme outperforms the proposed scheme beyond 50 files/s. This is because the proposed scheme aims to always balance the energy efficiency with QoS. Hence,

high energy saving is achieved at low traffic load where QoS deterioration is not a challenge. Whereas beyond low traffic levels high energy saving is traded off for QoS to prevent deterioration of QoS. Compared to the optimal adaptation approach, the proposed scheme in most cases selects lower policy levels even when higher policy levels guarantees target blocking. This decision leads to better ERG overall and also better QoS performance as shown in Figures 7.12 and 7.13. The CBCA scheme achieves the lowest ERG due to the higher interference which causes lower SINR choices, lower data rates, higher file transfer delay and consequently higher active periods for ABSs.

The effective energy saving (EES) performance of the scheme is shown in Figure 7.15. As explain in Chapter 3, EES measures how well the energy efficiency is balanced by the delay. EES is the difference between the ERG relative to the baseline scheme and percentage increase in delay also relative to baseline scheme. Therefore, the proposed scheme achieves better balance between energy efficiency and delay than other schemes while maintaining target probability as well over the traffic load supported by the system. The Highest SINR with One Neighbour scheme outperforms the optimal adaptation approach due to the poorer delay and ERG performance of this strategy in most cases. The CBCA scheme has the poorest EES performance since it achieves the poorest delay and ERG relative to other schemes in most cases.

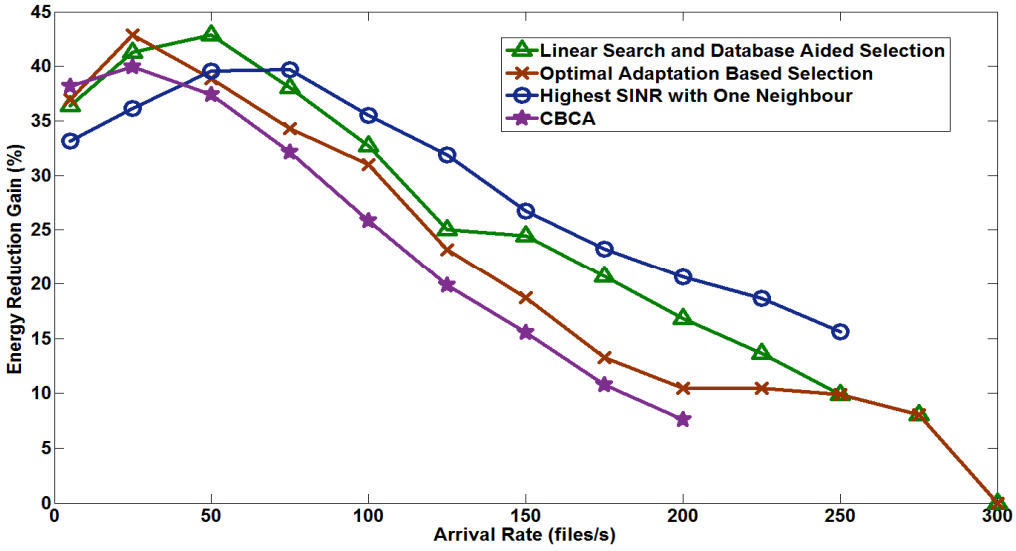


Figure 7.14 Energy Reduction Gain for Training Phase Traffic Load

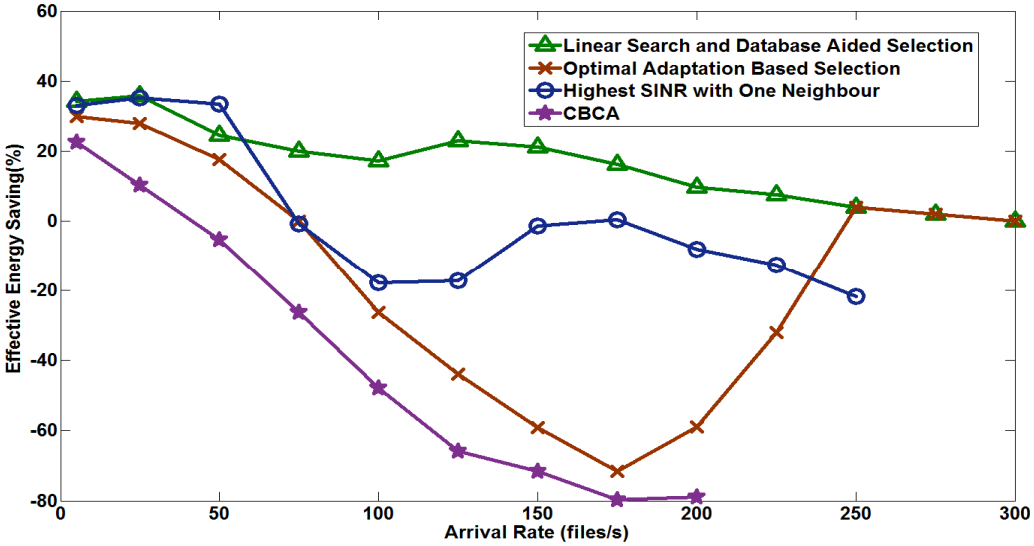


Figure 7.15 Effective Energy Saving for Training Phase Traffic Load

The proposed scheme is also evaluated with a set of traffic load completely different from the training phase set. It is assumed that the expected traffic load is known beforehand through accurate traffic load prediction and the appropriate policy level is selected at the start of each traffic load. The capability of the proposed scheme to effectively reduce the number of active ABS is compared with the Erlang B lower bound derived in section 7.4 for the validation phase traffic loads. This is shown in Figure 7.16. For the lower bound calculation, each ABS antenna serving a sector supports 10 subchannels of 1MHz each as explained in Chapter 4. The number of erlangs carried by 10 subchannels of each sector is evaluated at the 5% target blocking probability. The average service time is determined at the maximum data rate achievable on a subchannel based on the Truncated Shannon Bound as explained in Chapter 3. As shown in Figure 7.16, the gap between the lower bound and the proposed scheme is wider at lower traffic loads but reduces as traffic load increases. This is because the lower bound evaluates the lowest possible number of active ABSs assuming a lossless channel, free from interference, path loss or shadowing at blocking probability of less than or equal to 5%. Furthermore, while the blocking probability achieved at low load by the proposed scheme is less than 1% (in Figure 7.18), requiring more ABSs to achieve such high QoS than the lower bound. However, as the blocking probability increases with traffic load the performance approaches the lower bound prediction.

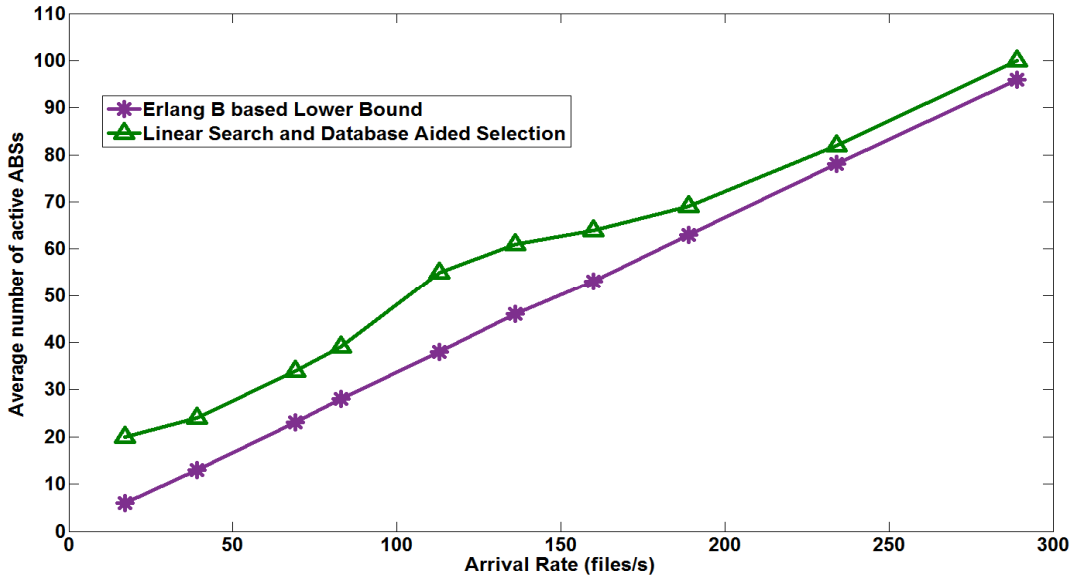


Figure 7.16 Average Number of Active ABSs

The selected policy levels of the proposed scheme under the new set of traffic loads are shown in Figure 7.17. In a similar way as under the training phase traffic loads (Figure 7.11), the proposed scheme selects higher policy level at low traffic load but the order of selected policy level decreases with the traffic load. Also, the policy levels selected are between policy level 1 and policy level 3 for all traffic loads over the range supported by the system as with the training phase traffic loads.

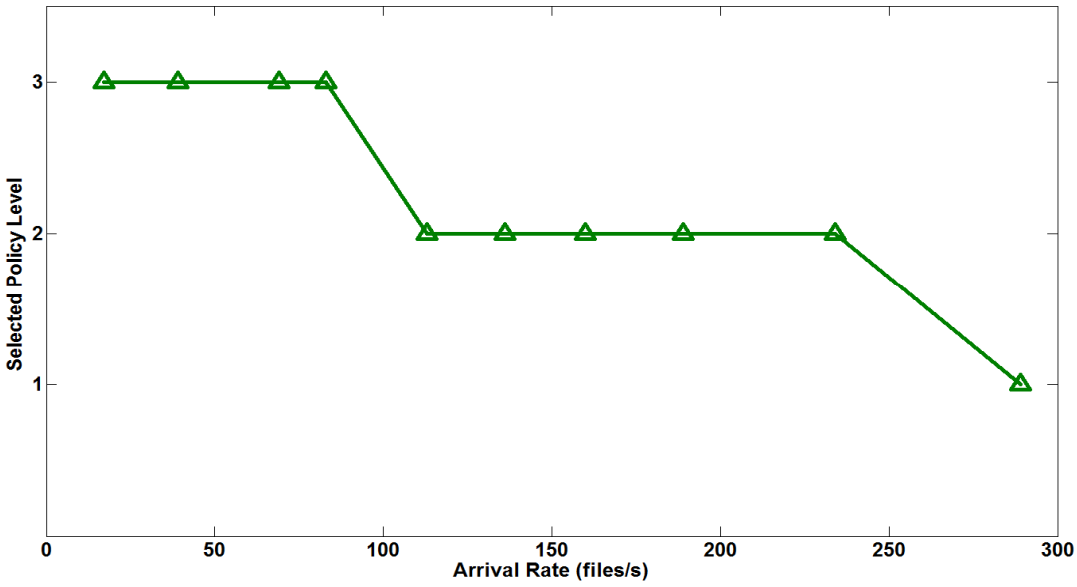


Figure 7.17 Selected Policy Levels during Validation Phase

The blocking probability and delay performances are shown in Figures 7.18 and 7.19 respectively. It is shown in Figure 7.18 that the proposed scheme keeps the blocking probability below 5% in all cases. In addition, the blocking probability achieved by the proposed scheme is close to the performance of the baseline scheme. This is because by seeking to keep the delay low, the system also keeps blocking probability even well below the target value in most cases.

Figure 7.19 shows that at low traffic load, the difference in delay between the proposed scheme and the baseline scheme is higher below 113 files/s than above it. This is because for the low traffic load cases, high energy saving is possible and the higher delay is justified by the high energy efficiency achieved. However, for the medium and high traffic load situation lower energy saving is possible since fewer ABSs can be switched to the sleep state and hence, the energy saving achievable cannot justify high delay. It is important to note that higher policy levels could have been selected at low traffic load to achieve higher energy efficiency but the further deterioration of delay cannot be balanced by the resulting increase in energy efficiency.

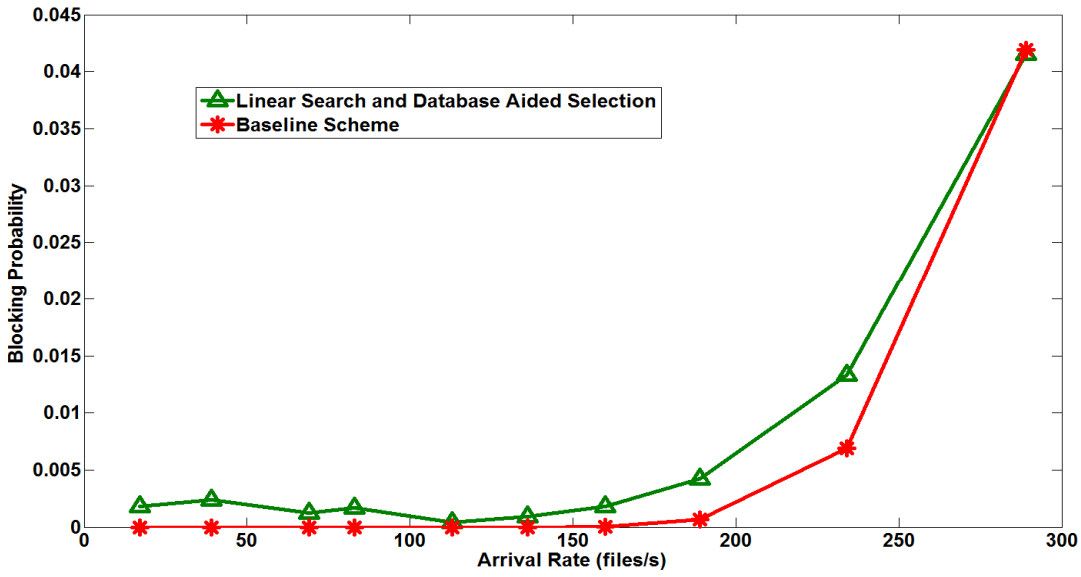


Figure 7.18 Blocking Probability for Validation Phase Traffic Loads

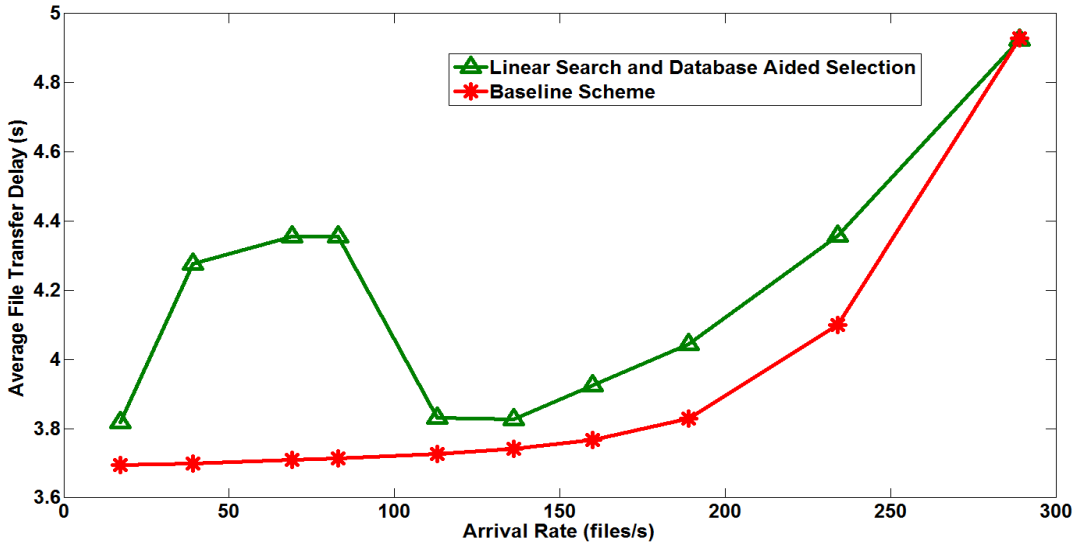


Figure 7.19 Average File Transfer Delay for Validation Phase Traffic Loads

The ERG and EES performance under the new traffic load set is shown in Figures 7.20 and 7.21 respectively. The proposed scheme achieves ERG of about 42% at low traffic load as shown in Figure 7.20. It is important to note that higher energy saving is possible in existing systems where higher gains can be achieved by switching off ABSs. Furthermore, the proposed scheme achieves balance between energy efficiency and QoS since the EES, as shown in Figure 7.21, is never below zero over the range of traffic supported by the system.

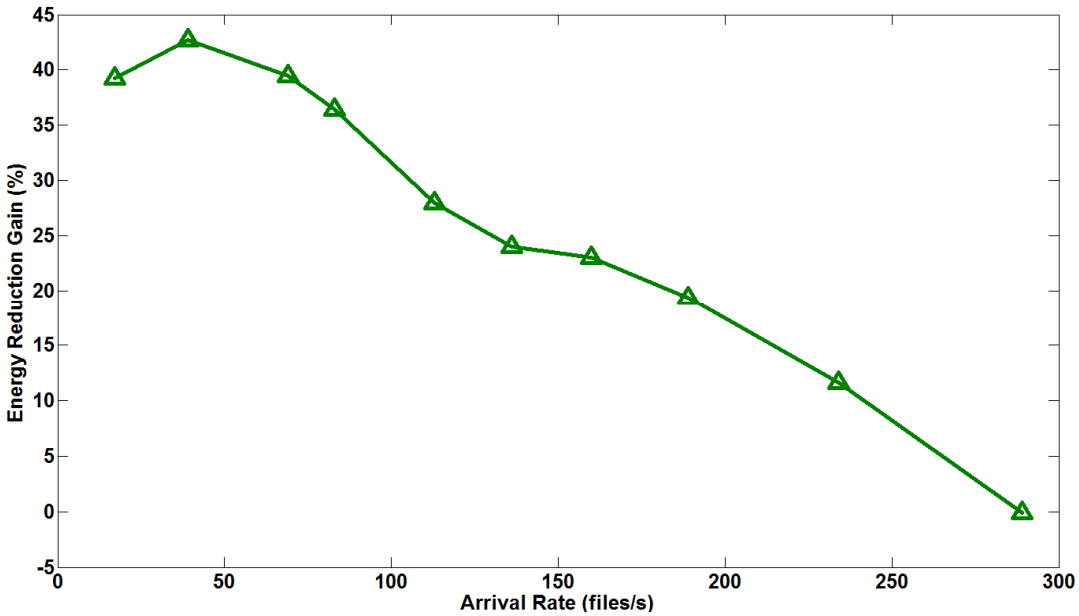


Figure 7.20 Energy Reduction Gain for Validation Phase Traffic Loads

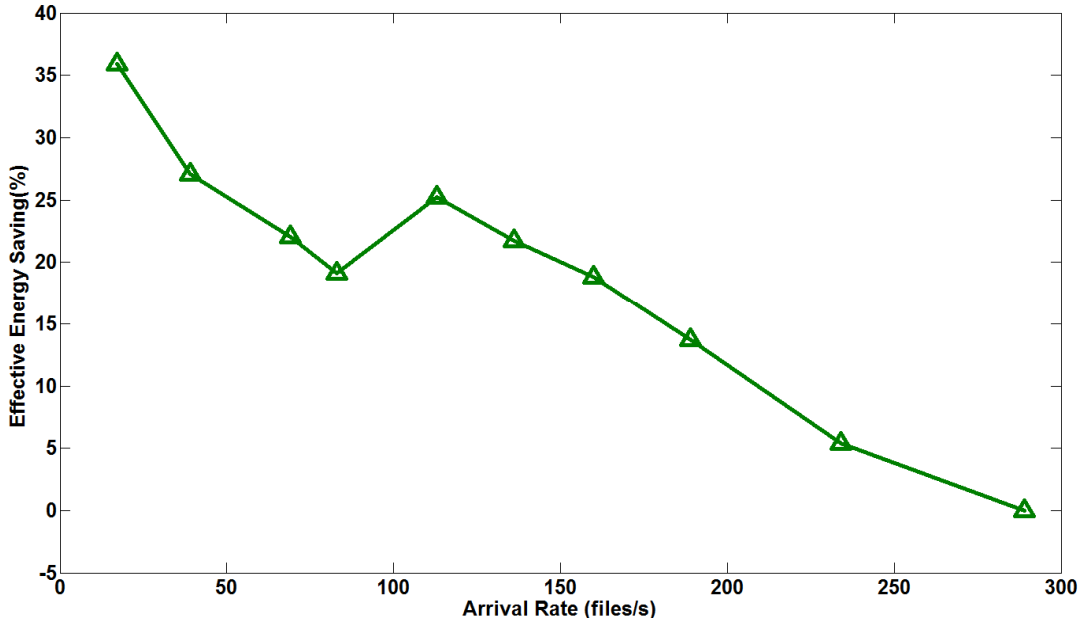


Figure 7.21 Effective Energy Saving for Validation Phase Traffic Loads

7.6 Conclusion

A novel linear search and database aided RRM and TM policy selection scheme is proposed in this chapter. Performance metrics of the network in terms of blocking probability, average file transfer delay and ECR at different traffic load are saved in a database at the OSS during an initial training phase. A linear search method is then used to map a policy level, which is a combination of RRM and TM policies, to each training phase traffic load to achieve low file transfer delay, blocking probability and ECR. This is done to achieve a good balance between energy efficiency and QoS. The policy level to traffic load mapping is saved in the database and used for RRM and TM policy selections for new traffic load experienced by the network. The scheme is implemented at the QEPU, which is responsible for the policy level selection decisions while the selected policy levels are applied locally at the zone level by the ZBSs. Initial RRM and TM responsibilities of the ZBSs and ABSs are unchanged and thus the partially centralised paradigm of RRM and TM is maintained.

It is shown that the proposed scheme achieves better QoS and energy efficiency than the adaptation strategy proposed in Chapter 6 because they allow selection of lower policy levels even when higher policy levels can achieve target blocking probability.

In addition, although the Highest SINR with One Neighbour On scheme achieves better ERG than the proposed scheme beyond low traffic, however the proposed scheme achieves better balance between QoS and energy efficiency. Unlike previous schemes in the literature, once the policy mapping is done and applied, the system can be operated over the same range of traffic loads supported by the baseline scheme at good energy efficiency and QoS using the proposed scheme.

Chapter 8. Summary and Conclusions

This thesis has investigated how a combined application of radio resource management (RRM) and topology management (TM) strategies can be used to achieve significant energy efficiency and QoS in heterogeneous cellular networks. A partially centralised concept for RRM and TM has been developed. RRM schemes have been proposed to cluster mobile stations on a fraction of small cells based on decisions at the macrocells. Also, a TM scheme has been developed to switch idle (sleeping) small cells off (on) according to the traffic load and QoS based on decisions from the central node, the macrocells and the small cells. The combined application of RRM and TM schemes has been shown to achieve significant energy efficiency over a wide range of traffic load without compromising the QoS in the network.

Furthermore, in order to mitigate interference and improve QoS performance, a choice restriction technique that restricts mobile stations to utilise resources from only a subset of their suitable small cell BS options has been proposed. The choice restriction is applied at the central node and shown to improve QoS performance without significant impact on energy efficiency. Also, a framework, low power state saving, was developed and used to investigate energy saving in the network under different power models. It is shown that significant energy saving is possible even when BSs are not switched off in separation architecture networks based on advanced small cell BSs with low idle state power consumption.

In addition, an adaptive joint RRM and TM scheme was developed with the ability to detect and rectify QoS deterioration problems in the network. RRM and TM policies are modified at the central node when the QoS estimated by a novel confidence level approach violates target QoS. This scheme has been shown to rectify QoS deterioration faults while maintaining reasonable energy saving at medium and high traffic loads. Further improvement in energy efficiency and QoS has been achieved by utilizing a database of past performance metrics and a linear search technique for selection of appropriate RRM and TM policies. A summary of conclusions from the main chapters of the thesis is provided as follows.

Chapter 1 provides important background information and discussed the main purpose and motivation for this thesis. In addition, a summary of the novel contribution of this thesis is provided. A literature review of energy efficiency studies in cellular networks with particular emphasis on heterogeneous networks (HetNets) is provided in Chapter 2. Specifically, energy models and energy efficiency metrics proposed in the literature are discussed. Also, proposed enhancements to base station components are reviewed. Furthermore, different network deployment approaches for improving energy efficiency are discussed. Finally, studies that proposed energy efficient resource and topology management schemes for heterogeneous cellular networks are discussed.

Chapter 3 discussed the simulation techniques, system modeling approaches, performance metrics and verification techniques used in this thesis. Simulation models have been extensively utilised to evaluate the performance of the proposed schemes. MATLAB is the simulation tool utilised throughout this thesis. Blocking probability, delay, throughput and energy reduction gain and effective energy saving are the performance metrics evaluated in this thesis. A confidence level method and Erlang B based bound have been utilised to validate the capability of the simulation models developed to generate reliable results.

Clustering and interference mitigation are investigated for achieving energy saving in the Separation Architecture based HetNet in Chapter 4. The partially centralised concept for RRM and TM, which is based on the principle of shared RRM and TM responsibilities among the network nodes, is introduced and explained. RRM schemes have been proposed to cluster MSs on more centrally located small cells in order to reduce the number of active BSs based on decision at the macrocells. Furthermore, a RRM choice restriction has been proposed to mitigate interference and improve QoS by restricting mobile users to utilise resources from only a subset of suitable small cell BSs. In addition, a TM scheme has been developed that switches small cell BSs on and off depending on the instantaneous traffic load and QoS.

Significant energy reduction gain of up to 67% has been achieved by clustering MSs onto few ABSs without particular focus on the SINR rank of the base station choices

utilised, while idle ABSs are switched off according to the TM rules. However, this resulted in poor blocking probability and average file transfer delay especially beyond low traffic loads. It has been shown that changing from a high order choice restriction to lower order choice restriction leads to better QoS. An RRM scheme with choice restriction permitting up to selection of the third choice restriction has been shown to achieve an energy reduction gain of up to 65% with better QoS performance than the approach without consideration of the ranking of base station choices.

The impact of different power model assumptions on energy saving in a separation architecture has been investigated in Chapter 5 to understand how improvement in idle and sleep state power consumptions of small cells affect energy saving. A framework, termed low power state saving (LPSS), has been developed to evaluate energy saving in the network due to operating small cells in lower power consumption state rather than higher power consumption state for different power models assumptions. The framework has also been used to examine how and why energy saving varies across power models.

It has been shown that energy saving is possible even when small cells are not switched off (i.e. no sleep state transitions) in a separation architecture based on future small cell BSs with low power consumption in idle state. In addition, an RRM scheme incorporating the third choice restriction is shown to achieve significant energy saving regardless of the power model assumption (futuristic or state-of-the-art). This implies the scheme is applicable in current and future systems for energy efficient operation.

An online, adaptive joint RRM and TM scheme with the capability to detect and rectify QoS deterioration is investigated in Chapter 6. The knowledge regarding the relationship between choice restriction and QoS in the RRM domain (Chapter 4) and the possibility of energy efficiency with or without sleep state transitions in the TM domain (Chapter 5) has been exploited to create different joint RRM/TM policies. The central node has been configured to continuously modify the joint RRM/TM policies until the QoS target is satisfied. A novel confidence level approach has been proposed to estimate QoS at a predefined confidence level using traffic statistics

collected from macrocell BSs. The confidence level approach both defines the degree of confidence in statistics collected and regulates when decisions are made at the central node.

It has been shown that when modified policies are adopted by the macrocell and small cell BSs, QoS deterioration can be corrected and QoS targets maintained subsequently, while still achieving moderate energy efficiency at medium and high traffic levels. However, further improvement in energy efficiency is possible when extra information about the quantitative performance of policies is available beforehand.

The exploitation of extra information apart from the relationship between choice restriction and QoS or energy saving and sleep state configurations is investigated for balancing QoS and energy efficiency better in Chapter 7. A database has been created and utilised to record previous QoS and energy efficiency performance of the system. A linear scheme method has been proposed to map offline the best combination of RRM and TM policy to different traffic loads encountered across the range supported by the system.

It has been shown that improvement in energy efficiency is possible by utilizing the offline mapping to inform RRM and TM policy selection at the central node for both mapped and unmapped traffic load. More importantly, better balance between QoS and energy efficiency is possible over the range of traffic load supported by the system relative to the online adaptation approach and previous schemes proposed in the literature.

Chapter 9. Future Work

This chapter discusses future studies that can be carried out following on from the work done in this thesis. Some of these studies centre mainly on further enhancements to the resource and topology management schemes to improve performance. In addition, enhancements required for these schemes to be suitably applied when there is a modification in the network architecture or the nature of services supported is also discussed.

9.1 Energy Efficiency under Non-neutral Regimes

The Internet is based on an open architecture that facilitates innovations and permits applications of different types to be provided to connected users anywhere [167]. Traditionally, a neutral approach has been taken with regard to handling different applications over the Internet [168]. This approach has often been termed net neutrality and implies that all traffic, users or applications should be treated equally without discrimination of some over others [169, 170]. However, net neutrality debates centred on switching to a non-neutral regime have been ongoing over the last decade by law and policy makers [168]. The impact of the net neutrality conflict on the energy efficiency of cellular network has not been considered in the literature.

The studies carried out in this thesis have all been based on a neutral approach to traffic handling without prioritization of any user over another based on the price paid or ownership of traffic. However, it is important to evaluate energy efficiency under a non-neutral regime as this might become a feature of future cellular network partially or fully. The resource and topology management schemes developed in this thesis can be adapted to function in a non-neutral regime. In a non-neutral regime with prioritization of content providers based on price paid for content delivery, users may be grouped according to the content they intend to download or upload.

Unlike, the approach in this thesis which treats user requests equally, the resource management scheme will have to be modified to prioritise the requests of the users (or subscribers) of higher paying content providers over lower paying or non-paying counterparts. A content provider might have paid for a certain guaranteed QoS, hence the resource management scheme will treat users utilizing this application in a

manner different from users utilizing an unpaid application. Rather than assign resources from a distant small cell base station permitted under the choice restriction applied, resources from a closer base station that will guarantee the desired QoS may be utilised. There may be need to specially protect the paid users from severe interference as well by denying access to would be interfering non-paying user requests.

Furthermore, with respect to the topology management scheme, there may be a need to consider the services of the paid users especially when making decisions about switching off or on small cell base stations in a manner different from the initial approach. In this thesis, the small cell base stations in sleep state are turned on when the load on neighbouring base stations become too high or when the blocking exceeds the target as explained in Chapter 4. However, it is envisaged that under the non-neutral regime, base stations may need to be turned on even when the load and blocking targets are not exceeded to meet the negotiated QoS, such as minimum delay and data rate, for a paid service. This is the case when the active base stations cannot meet these conditions and some of the sleeping base stations are closer to the users and suitable to meet the negotiated QoS. The energy efficiency obtained under such modified resource and topology management schemes when compared with the results in thesis will provide an understanding of the impact of the non-neutral regime with consideration of price discrimination.

9.2 Energy Efficiency under Hybrid Power Sources

Renewable energy such as solar and wind are becoming increasing popular as a means of improving energy efficiency. Although, the initial cost of deployment might be high, the operational cost compared to traditional power grid fed approach is smaller. It also has the additional advantage of being environmentally friendly. Some energy efficiency evaluations based on the joint utilization of traditional grid power and renewable have been considered for the heterogeneous network without control and data plane separation [171-174]. Similar evaluation is also relevant to the separation architecture considered in this work as energy consumption is even lower for small cells in this architecture compared to the basic heterogeneous network.

In such an evaluation, different renewable energy sources can be investigated to understand to what extent they can be utilised to support the energy required in the network. For example with solar energy, the length of time the energy can be harvested, how long the energy harvested will last, the portion of the time that the grid power is needed are important parameters to determine. A combination of different renewable energy sources can be investigated for both the small cell base stations and the control base stations as well. The low power consumption of the small cell base stations implies that they can be easily supported by renewable energy sources [171]; but it is still worth considering the benefits of renewable energy sources to the high power base stations as well.

Furthermore, the resource and topology management schemes will have to be modified to account for the times base stations operate under renewable sources. It is envisaged that the base stations can be assumed to operate at zero energy when renewable energy are being utilised and more base stations can be left on under this condition. In addition, when renewable energy is available at a base station it can be turned on to increase the resources available in the network to serve the users and achieve better QoS than the specified targets. This would lead to improved QoS and higher energy efficiency as well since the energy consumed from the renewable sources are not added to the energy budget of the network.

9.3 Power Control for Improved Energy Efficiency and QoS

Power control is a resource management technique that is used to regulate the power levels at which MSs and BSs transmit over assigned resources (like frequency or time slot) [175]. It is applied in both downlink [176, 177] and uplink [178, 179] directions to mitigate inter-cell interference among nodes utilizing similar time slots and/or frequency channels. The restriction of the choice of small cell BSs to serve MSs is used to mitigate interference in the uplink direction in this work while all MSs transmit at the same power level. Further reduction in interference can be achieved by introducing uplink power control.

The uplink power control can be co-ordinated by the ZBSs which assign the resources (frequency channels in this work) of the small cell BSs in their zones to the MSs. The ZBSs can in addition to assigning frequency channels determine the power

level each MS should transmit at on the assigned frequency channels. The aim is to limit interference to other MSs transmitting on similar frequency channels in the zones served by the ZBSs. The assigned power levels would be chosen such that they are sufficiently high to achieve good QoS for newly admitted MSs without compromising the transmission of already active MSs. Further reduction in interference will translate to higher SINR and data rates for the MSs. The higher data rate will improve file transfer delay performance since user data transmission will be completed sooner. In addition, the energy efficiency can also be improved if the small cell BSs can go into the sleep state sooner with user transmissions being completed quicker.

9.4 Data Handling Control Base Stations

In this thesis, the ZBSs have been utilised to transmit overhead signals and to manage the resources of the ABSs within their zones. However, they have not been utilised to receive user data. This responsibility has been assigned exclusively to the ABSs in this work. This makes it possible to compare the proposed schemes fairly to the previous schemes which are evaluated on the initial BuNGee Architecture which does not include the ZBSs. Data transmissions are completely handled by the ABSs under these previous schemes as well and energy efficiency evaluation has been focussed on the small cell BSs tier. At very low traffic load, if the ZBSs are permitted to transmit and receive user data, they can handle all the traffic in the network. Thus, all ABSs can be switched off to achieve even higher energy saving than achieved in this work.

This is also relevant to a mixed traffic scenario comprising low data rate and high data rate services. In this case, the ZBSs can be required to handle low data rate services while the ABSs handle high data rate services. In such situations if low data rate services dominate or high data rate services are completely non-existent in the network, the ZBSs which are always on anyway for coverage reasons can handle majority of the traffic. Significantly higher energy savings can be achieved since more ABSs can be switched off under this arrangement than the previous case used in this work because no ABS needs to be active for the purpose of ensuring availability of data services at very low load.

Glossary

2G	Second Generation
3G	Third Generation
3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
ABS	Access Base Station
ACK	Acknowledge
AI	Active to Idle
AS	Active to Sleep
ASE	Area Spectral Efficiency
APC	Area Power Consumption
BB	Baseband
BS	Base Stations
BLER	Block Error Ratio
BHSS	Backhaul Subscriber Station
BuNGee	Beyond Next Generation
CBCA	Capacity Based Channel Assignment
CBS	Coverage Base Station
CCR	Clustering Capability Rating
CFI	Control Format Indicator

CQ-CCR	Controllable Quality Clustering Capability Rating
CRS	Cell-Specific Reference Signal
CSG	Closed Subscriber Group
CSI	Channel State Information
D2D	Device to Device
DCI	Downlink Control Information
DD	Delay Degradation
DTX	Discontinuous Transmission
ECG	Energy Consumption Gain
ECR	Energy Consumption Rating
EEB	Extended Erlang B
EER	Envelope Elimination and Restoration
EES	Effective Energy Saving
ERG	Energy Reduction Gain
GB	Gigabytes
GSM	Global System for Mobile Communications
HARQ	Hybrid Automatic Request
HBS	Hub Base Station
HetNets	Heterogeneous Networks
HI	HARQ Indicator
IA-CCR	Interference Aware Clustering Capability Rating

ICT	Information and Communications Technology
IS	Idle to Sleep
ISD	Inter-Site Distance
KPI	Key Performance Indicator
LOS	Line of Sight
LPSS	Low Power State Saving
LTE	Long Term Evolution
LTE-A	Long Term Evolution Advanced
Mbps	Megabits per second
MeNB	Macro eNodeB
MIB	Master Information Block
MIMO	Multiple Inputs Multiple Outputs
mmWave	millimeter Wave
MS	Mobile Station
NACK	Negative Acknowledge
NCCR	Normalized Clustering Capability Rating
NE	Network Element
NEM	Network Element Management
NLOS	Non-Line of Sight
NM	Network Management
OAM	Operation Administration and Maintenance

OFDM	Orthogonal Frequency Division Multiplex
OSS	Operation Support System
PA	Power Amplifier
PBCA	Priority Based Channel Assignment
PCC	Phantom Cell Concept
PCRF	Policy and Charging Rules Function
PEC	Per-Energy Capacity
PHICH	Physical Hybrid Automatic Request Indicator Channel
PSR	Partial Spectrum Reuse
PSS	Primary Synchronization Signal
QEPU	Quality Enhancing Processing Unit
QoS	Quality of Service
RAN	Radio Access Network
RF	Radio Frequency
RRM	Radio Resource Management
SCN	Small Cell Networks
SeNB	Small cell eNodeB
SIBs	System Information Blocks
SINR	Signal to Interference plus Noise Ratio
SNR	Signal to Noise Ratio
SON	Self-Organising Network

SotA	State-of-the-Art
SPM	Saturation Proximity Metric
SSS	Secondary Synchronization Signal
TBS	Traffic Base Station
TM	Topology Management
TSB	Truncated Shannon Bound
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
ZBS	Zone Base Station

References

[1] M. W. Arshad, A. Vastberg, and T. Edler, "Energy efficiency gains through traffic offloading and traffic expansion in joint macro pico deployment," in *IEEE Wireless Communications and Networking Conference*, pp. 2203-2208, 2012.

[2] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Communications Magazine,* vol. 49, pp. 55-62, 2011.

[3] Ericsson, "Mobile Data Traffic Surpasses Voice," http://www.ericsson.com/news/1396928, 2010, accessed January 2013.

[4] S. Tombaz, A. Vastberg, and J. Zander, "Energy- and cost-efficient ultra-high-capacity wireless access," *Wireless Communications, IEEE,* vol. 18, pp. 18-24, 2011.

[5] S. Landström, A. Furuskär, K. Johansson, L. Falconetti, and F. Kronestedt, "Heterogeneous networks (hetnets)–an approach to increasing cellular capacity and coverage," *Ericsson Review,* 2011.

[6] F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. M. Aggoune*, et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine,* pp. 123, 2014.

[7] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green Cellular Networks: A Survey, Some Research Issues and Challenges," *Communications Surveys & Tutorials, IEEE,* vol. 13, pp. 524-540, 2011.

[8] G. Fettweis and E. Zimmermann, "ICT energy consumption-trends and challenges," in *Proceedings of the 11th International Symposium on Wireless Personal Multimedia Communications*, 2008, pp. 6.

[9] M. H. Sunil Vadgama, "Trends in Green Wireless Access Networks," *IEEE International Conference on Communications Workshops (ICC),* pp. 1-5, 2011.

[10] M. Webb, "SMART 2020: Enabling the low carbon economy in the information age," http://www.smart2020.org/_assets/files /03_Smart2020Report_lo_res.pdf, accessed January 2013.

[11] Gartner, "Gartner estimates ICT industry accounts for 2 percent of global CO2 emissions," ed: Gartner Press Release (Apr 2007) http://www.gartner.com/it/page.jsp?id=503867, accessed January 2013.

[12] F. Richter, A. J. Fehske, and G. P. Fettweis, "Energy Efficiency Aspects of Base Station Deployment Strategies for Cellular Networks," in *IEEE Vehicular Technology Conference Fall*, pp. 1-5, 2009.

[13] J. Hoydis, M. Kobayashi, and M. Debbah, "Green Small-Cell Networks," *Vehicular Technology Magazine, IEEE,* vol. 6, pp. 37-43, 2011.

[14] H. Tabassum, M. Z. Shakir, and M.-S. Alouini, "Area green efficiency (AGE) of two tier heterogeneous cellular networks," in *IEEE Globecom Workshops*, pp. 529-534, 2012.

[15] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi Offloading: Trading Delay for Cellular Capacity," *IEEE Transactions on Wireless Communications,* vol. 13, pp. 1540-1554, 2014.

[16] M. Bennis, M. Simsek, A. Czylwik, W. Saad, S. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," *Communications Magazine, IEEE,* vol. 51, pp. 44-50, 2013.

[17] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking*, pp. 25-30, 2013..

[18] O. Blume, H. Eckhardt, S. Klein, E. Kuehn, and W. M. Wajda, "Energy savings in mobile networks based on adaptation to traffic statistics," *Bell Labs Technical Journal,* vol. 15, pp. 77-94, 2010.

[19] T. Zhao, P. Yang, H. Pan, R. Deng, S. Zhou, and Z. Niu, "Software defined radio implementation of signaling splitting in hyper-cellular network," in *Proceedings of the second workshop on Software radio implementation forum*, pp. 81-84, 2013.

[20] X. Xu, G. He, S. Zhang, Y. Chen, and S. Xu, "On functionality separation for green mobile networks: concept study over LTE," *Communications Magazine, IEEE,* vol. 51, pp. 82-90, 2013.

[21] I. Chih-Lin, C. Rowell, H. Shuangfeng, X. Zhikun, L. Gang, and P. Zhengang, "Toward green and soft: a 5G perspective," *Communications Magazine, IEEE,* vol. 52, pp. 66-73, 2014.

[22] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *Communications Magazine, IEEE,* vol. 52, pp. 74-80, 2014.

[23] A. Capone, I. Filippini, B. Gloss, and U. Barth, "Rethinking cellular system architecture for breaking current energy efficiency limits," in *Sustainable Internet and ICT for Sustainability*, pp. 1-5, 2012.

[24] N. Zhisheng, W. Yiqun, G. Jie, and Y. Zexi, "Cell zooming for cost-efficient green cellular networks," *Communications Magazine, IEEE,* vol. 48, pp. 74-79, 2010.

[25] A. Fisusi, D. Grace, and P. Mitchell, "Energy efficient cluster-based resource allocation and topology management for beyond next generation mobile broadband networks," in *IEEE International Conference on Communications Workshops*, pp. 576-580, *2013*.

[26] D. Cao, S. Zhou, and Z. Niu, "Improving the energy efficiency of two-tier heterogeneous cellular networks through partial spectrum reuse," *Wireless Communications, IEEE Transactions on,* vol. 12, pp. 4129-4141, 2013.

[27] S. Bu, F. R. Yu, and H. Yanikomeroglu, "Interference-Aware Energy-Efficient Resource Allocation for OFDMA-Based Heterogeneous Networks With Incomplete Channel State Information," *Vehicular Technology, IEEE Transactions on,* vol. 64, pp. 1036-1050, 2015.

[28] W. Pramudito and E. Alsusa, "A Hybrid Resource Management Technique for Energy and QoS Optimization in Fractional Frequency Reuse Based Cellular Networks," *Communications, IEEE Transactions on,* vol. 61, pp. 4948-4960, 2013.

[29] A. Fisusi, D. Grace, and P. Mitchell, "Interference aware, energy efficient resource allocation for beyond next generation mobile networks," in *IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC),* pp. 2197-2202, *2013*.

[30] I. de la Bandera Cascal, R. Barco, P. Munoz, and I. Serrano, "Cell Outage Detection Based on Handover Statistics."

[31] H. Yunbo, D. Grace, and P. Mitchell, "Energy efficient topology management for beyond next generation mobile broadband systems," in *International Symposium on Wireless Communication Systems*, pp. 331-335, 2012.

[32] A. Prasad and A. Maeder, "Backhaul-aware energy efficient heterogeneous networks with dual connectivity," *Telecommunication Systems,* vol. 59, pp. 25-41, 2015.

[33] S. Chakraborty, J. Peisa, T. Frankkila, and P. Synnergren, *IMS Multimedia Telephony over Cellular Systems: VoIP Evolution in a Converged Telecommunication World*: Wiley, 2007.

[34] M. F. Hossain, K. S. Munasinghe, and A. Jamalipour, "An eco-inspired energy efficient access network architecture for next generation cellular systems," in *IEEE Wireless Communications and Networking Conference (WCNC),*pp. 992-997, 2011.

[35] S. Palat and P. Godin, "Network Architecture," in *LTE-The UMTS Long Term Evolution: From Theory to Practice*, I. T. Stefania Sesia, Matthew Baker, Ed., : John Wiley & Sons Ltd, United Kingdom, pp. 21-50, 2009.

[36] L. Budzisz, F. Ganji, G. Rizzo, M. A. Marsan, M. Meo, Y. Zhang*, et al.*, "Dynamic resource provisioning for energy efficiency in wireless access networks: a survey and an outlook," *Communications Surveys & Tutorials, IEEE,* vol. 16, pp. 2259-2285, 2014.

[37] A. R. Elsherif, W.-P. Chen, A. Ito, and Z. Ding, "Resource Allocation and Inter-cell Interference Management for Dual-Access Small Cells," *Selected Areas in Communications, IEEE Journal on,* vol. 33, pp. 1082 - 1096, 2015.

[38] Wikipedia, "Embodied energy," http://en.wikipedia.org/wiki/Embodied_energy, accessed January 2013.

[39] B. Badic, T. O"Farrrell, P. Loskot, and J. He, "Energy Efficient Radio Access Architectures for Green Radio: Large versus Small Cell Size Deployment," in *IEEE Vehicular Technology Conference Fall*, pp. 1-5, 2009.

[40] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson*, et al.*, "How much energy is needed to run a wireless network?," *Wireless Communications, IEEE,* vol. 18, pp. 40-49, 2011.

[41] G. Auer, V. Giannini, I. Godor, P. Skillermark, M. Olsson, M. A. Imran*, et al.*, "Cellular Energy Efficiency Evaluation Framework," in *IEEE Vehicular Technology Conference,*pp. 1-6, 2011.

[42] S. Luis, N. Loutfi, and B. Jean-Marie, "An overview and classification of research approaches in green wireless networks," *EURASIP Journal on Wireless Communications and Networking,* vol. 2012, pp. 142-142, 2012.

[43] M. Imran, E. Katranaras, G. Auer, O. Blume, V. Giannini, I. Godor*, et al.*, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," Tech. Rep. ICT-EARTH deliverable, https://bscw.ict-earth.eu/pub/bscw.cgi/d71252/EARTH_WP2_D2.3_v2.pdf, accessed January 2015.

[44] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing Energy Consumption in LTE with Cell DTX," in *IEEE Vehicular Technology Conference* , pp. 1-5, 2011.

[45] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-Efficient Scheduling and Power Allocation in Downlink OFDMA Networks With

Base Station Coordination," *Wireless Communications, IEEE Transactions on,* vol. 14, pp. 1-14, 2015.

[46] W. Jian, Z. Sheng, and N. Zhisheng, "Traffic-Aware Base Station Sleeping Control and Power Matching for Energy-Delay Tradeoffs in Green Cellular Networks," *Wireless Communications, IEEE Transactions on,* vol. 12, pp. 4196-4209, 2013.

[47] H. Tao and N. Ansari, "Enabling Mobile Traffic Offloading via Energy Spectrum Trading," *Wireless Communications, IEEE Transactions on,* vol. 13, pp. 3317-3328, 2014.

[48] A. Akbari, F. Héliot, M. A. Imran, and R. Tafazolli, "Energy efficiency contours for broadcast channels using realistic power models," *Wireless Communications, IEEE Transactions on,* vol. 11, pp. 4017-4025, 2012.

[49] Z. Wang and W. Zhang, "A Separation Architecture for Achieving Energy-Efficient Cellular Networking," *Wireless Communications, IEEE Transactions on,* vol. 13, pp. 3113-3123, 2014.

[50] M. A. Yigitel, O. D. Incel, and C. Ersoy, "QoS vs. energy: A traffic-aware topology management scheme for green heterogeneous networks," *Computer Networks,* vol. 78, pp. 130-139, 2015.

[51] S. Tombaz, P. Monti, W. Kun, A. Vastberg, M. Forzati, and J. Zander, "Impact of Backhauling Power Consumption on the Deployment of Heterogeneous Mobile Networks," in *IEEE Global Telecommunications Conference*, pp. 1-5, 2011.

[52] T. Nakamura, S. Nagata, A. Benjebbour, Y. Kishiyama, T. Hai, S. Xiaodong, *et al.*, "Trends in small cell enhancements in LTE advanced," *Communications Magazine, IEEE,* vol. 51, pp. 98-105, 2013.

[53] S. Tombaz, S. Ki Won, and J. Zander, "On Metrics and Models for Energy-Efficient Design of Wireless Access Networks," *Wireless Communications Letters, IEEE,* vol. 3, pp. 649-652, 2014.

[54] H. Holtkamp, G. Auer, and H. Haas, "On Minimizing Base Station Power Consumption," in *IEEE Vehicular Technology Conference*, pp. 1-5, 2011.

[55] L. M. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, Go, *et al.*, "Challenges and enabling technologies for energy aware mobile radio networks," *Communications Magazine, IEEE,* vol. 48, pp. 66-72, 2010.

[56] E. Initiative, "Network and Telecom Equipment— Energy and Performance Assessment," http://www.ecrinitiative.org/pdfs/ECR_3_0_1.pdf, accessed June 2012.

[57] T. H. Congzheng Han, Simon Armour, Ioannis Krikidis, "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Communications Magazine,* vol. 49, pp. 46 - 54 2011.

[58] H. K. Tao Chen, Yang Yang, "Energy efficiency metrics for green wireless communications," *International Conference on Wireless Communications and Signal Processing,* pp. 1 - 6, 2010.

[59] H. Jianhua, P. Loskot, T. O'Farrell, V. Friderikos, S. Armour, and J. Thompson, "Energy efficient architectures and techniques for Green Radio access networks," in *International ICST Conference on Communications and Networking in China (CHINACOM),* pp. 1-6, 2010

[60] C. Turyagyenda, T. O'Farrell, and W. Guo, "Energy efficient coordinated radio resource management: a two player sequential game modelling for the

long-term evolution downlink," *IET communications,* vol. 6, pp. 2239-2249, 2012.

[61] A. D. Maleki and B. Abolhassani, "New scheduling scheme for green communications in long term evolution networks," *Communications, IET,* vol. 8, pp. 2438-2444, 2014.

[62] T. Keenan and R. Villing, "Optimizing area power consumption in LTE type cells," in *IET Irish Signals and Systems Conference*, pp. 1-8, 2013.

[63] F. H. Raab, P. Asbeck, S. Cripps, P. B. Kenington, Z. B. Popovic, N. Pothecary*, et al.*, "Power amplifiers and transmitters for RF and microwave," *Microwave Theory and Techniques, IEEE Transactions on,* vol. 50, pp. 814-826, 2002.

[64] S. C. Cripps, *RF Power Amplifiers for Wireless Communications* 2nd Edition : Artech House Microwave Library, 2006.

[65] J. Jingon, H. Chin Keong, K. Adachi, and S. Sumei, "A Survey on Power-Amplifier-Centric Techniques for Spectrum- and Energy-Efficient Wireless Communications," *Communications Surveys & Tutorials, IEEE,* vol. 17, pp. 315-333, 2015.

[66] L. Cheng-Po, J. Je-Hong, W. E. Stark, and J. R. East, "Nonlinear amplifier effects in communications systems," *Microwave Theory and Techniques, IEEE Transactions on,* vol. 47, pp. 1461-1466, 1999.

[67] R. Darraji and F. M. Ghannouchi, "Digital Doherty Amplifier With Enhanced Efficiency and Extended Range," *Microwave Theory and Techniques, IEEE Transactions on,* vol. 59, pp. 2898-2909, 2011.

[68] M. Vasić, O. Garcia, J. A. Oliver, P. Alou, D. Diaz, J. A. Cobos*, et al.*, "Efficient and Linear Power Amplifier Based on Envelope Elimination and Restoration," *Power Electronics, IEEE Transactions on,* vol. 27, pp. 5-9, 2012.

[69] N. Saitou, Y. Endo, and Y. Shibuya, "Mobile WiMAX Base Station Architecture and RF Technology," *Fujitsu Sci. Tech. J,* vol. 44, pp. 325-332, 2008.

[70] S. Vadgama, "Trends in green wireless access," *Fujitsu Sci. Tech. J,* vol. 45, pp. 404-408, 2009.

[71] N. Tuan-Duc, O. Berder, and O. Sentieys, "Cooperative MIMO Schemes Optimal Selection for Wireless Sensor Networks," in *IEEE Vehicular Technology Conference,* pp. 85-89, 2007.

[72] T. Rappaport, "Wireless Communications - Principles and Practice, Second edition," : Prentice Hall, 2002.

[73] L. Hansung, B. Seon Yeob, and S. Dan Keun, "The Effects of Cell Size on Energy Saving, System Capacity, and Per-Energy Capacity," in *IEEE Wireless Communications and Networking Conference,* pp. 1-6, 2010.

[74] S. Tombaz, K. W. Sung, and J. Zander, "Impact of densification on energy efficiency in wireless access networks," in *IEEE Globecom Workshops,* pp. 57-62, 2012.

[75] J. Lorincz and T. Matijevic, "Energy-efficiency analyses of heterogeneous macro and micro base station sites," *Computers & Electrical Engineering,* vol. 40, pp. 330-349, 2014.

[76] S. Tombaz, S.W. Han, K. W. Sung, and J. Zander, "Energy efficient network deployment with cell DTX," *Communications Letters, IEEE,* vol. 18, pp. 977-980, 2014.

[77] Real-Wireless, "An assessment of the value of small cell services to operators based on Virgin Media Trials," http://www.realwireless.biz/realwireless/wp-content/uploads/2012/10/Small-Cells-Value-for-operators-based-on-Virgin-Media-Trials-V3.1.pdf, accessed January, 2013.

[78] BuNGee, "Beyond Next Generation Mobile Broadband," http://cordis.europa.eu/fp7/ict/future-networks/projectsummaries/bungee.pdf, accessed May 2016.

[79] Z. E. Roth, M. Goldhamer, N. Chayat, A. Burr, M. Dohler, N. Bartzoudis*, et al.*, "Vision and architecture supporting wireless GBit/sec/km2 capacity density deployments," in *Future Network and Mobile Summit,* pp. 1-7, 2010.

[80] T. Jiang, P. Li, C. Liu, N. Khan, D. Grace, A. Burr*, et al.*, "BuNGee Deliverable: D4.1.2 Simulation Tool(s) and Simulation Results", http://cordis.europa.eu/docs/projects/cnect/7/248267/080/deliverables/001-BuNGeeD412UoYv1022052012.pdf, accessed May 2016.

[81] S. Rehan and D. Grace, "Combined green resource and topology management for beyond next generation mobile broadband systems," in *International Conference on Computing, Networking and Communications (ICNC),* pp. 242-246, 2013.

[82] F. Richter, G. Fettweis, M. Gruber, and O. Blume, "Micro base stations in load constrained cellular mobile radio networks," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications Workshops,* pp. 357-362, *2010*.

[83] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Future Network and Mobile Summit,* pp. 1-8, 2010.

[84] J. Weitzen, L. Mingzhe, E. Anderland, and V. Eyuboglu, "Large-Scale Deployment of Residential Small Cells," *Proceedings of the IEEE,* vol. 101, pp. 2367-2380, 2013.

[85] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, Present, and Future," *Selected Areas in Communications, IEEE Journal on,* vol. 30, pp. 497-508, 2012.

[86] C. Fengming and F. Zhong, "The tradeoff between energy efficiency and system performance of femtocell deployment," in *International Symposium on Wireless Communication Systems,* 315-319, 2010.

[87] H. Claussen, L. T. W. Ho, and F. Pivit, "Effects of joint macrocell and residential picocell deployment on the network energy efficiency," in *International Symposium on Personal, Indoor and Mobile Radio Communications, 2008*, pp. 1-6, 2008.

[88] M. W. Arshad, A. Vastberg, and T. Edler, "Energy efficiency improvement through pico base stations for a green field operator," in *IEEE Wireless Communications and Networking Conference*, pp. 2197-2202, *2012*.

[89] H. Ishii, Y. Kishiyama, and H. Takahashi, "A novel architecture for LTE-B: C-plane/U-plane split and phantom cell concept," in *IEEE Globecom Workshops,* pp. 624-630, *2012*

[90] A. Capone, A. F. d. Santos, I. Filippini, and B. Gloss, "Looking beyond green cellular networks," in *Annual Conference on Wireless On-demand Network Systems and Services*, pp. 127-130, *2012*.

[91] 3GPP, "Overview of 3GPP Release 12 V0.2.0 ", ed. http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/, accessed November 2015.

[92] G. Americas, "Understanding 3GPP Release 12 Standards for HSPA+ and LTE-Advanced Enhancements," http://www.4gamericas.org/files/6614/2359/0457/4G_Americas_-_3GPP_Release_12_Executive_Summary_-_February_2015.pdf, accessed January 2015.

[93] T. Asai, "5G radio access network and its requirements on mobile optical network," in *International Conference on Optical Network Design and Modeling,* pp. 7-11, *2015*.

[94] H. Wang, C. Rosa, and K. I. Pedersen, "Dual connectivity for LTE-advanced heterogeneous networks," *Wireless Networks,* pp. 1-14, 2015.

[95] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE release 12 and beyond [accepted from open call]," *Communications Magazine, IEEE,* vol. 51, pp. 154-160, 2013.

[96] Z. Chen, L. Qiu, and X. Liang, "Energy-Efficient Combination of Small Cells and Multi-Antenna Under Separation Architecture," *Communications Letters, IEEE,* vol. 19, pp. 1572-1575, 2015.

[97] P. Soldati, "Interference coordination and resource allocation in dense small cell networks," in *IEEE Global Communications Conference,* pp. 3742-3747, *2014*.

[98] J. B. Rao and A. O. Fapojuwo, "A Survey of Energy Efficient Resource Management Techniques for Multicell Cellular Networks," *Communications Surveys & Tutorials, IEEE,* vol. 16, pp. 154-180, 2014.

[99] E. Yaacoub and Z. Dawy, "A Survey on Uplink Resource Allocation in OFDMA Wireless Networks," *Communications Surveys & Tutorials, IEEE,* vol. 14, pp. 322-337, 2012.

[100] C. Patra, A. Mondal, P. Bhaumik, and M. Chattopadhyay, "Topology Management in Wireless Sensor Networks," *Wireless Sensor Networks and Energy Efficiency: Protocols, Routing and Management: Protocols, Routing and Management,* pp. 14, 2012.

[101] L. Saker, S.-E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *Selected Areas in Communications, IEEE Journal on,* vol. 30, pp. 664-672, 2012.

[102] E. Chavarria Reyes, I. Akyildiz, and E. Fadel, "Energy Consumption Analysis and Minimization in Multi-Layer Heterogeneous Wireless Systems," *Mobile Computing, IEEE Transactions on,* vol. 14, pp. 2474-2487, 2015.

[103] L. Jorguseski, A. Pais, F. Gunnarsson, A. Centonza, and C. Willcock, "Self-organizing networks in 3GPP: Standardization and future trends," *Communications Magazine, IEEE,* vol. 52, pp. 28-34, 2014.

[104] V. Monteiro, K. M. S. Huq, S. Mumtaz, C. Politis, and J. Rodriguez, "Energy efficient load balancing for future self-organized shared networks," *Telecommunication Systems,* vol. 59, pp. 123-135, 2015.

[105] S. Rehan and D. Grace, "Macro-cell overlaid green topology management for beyond next generation mobile broadband systems," in *Future Network and Mobile Summit (FutureNetworkSummit),* pp. 1-9, *2013*.

[106] S. Mukherjee and H. Ishii, "Energy efficiency in the phantom cell enhanced local area architecture," in *IEEE Wireless Communications and Networking Conference (WCNC),*pp. 1267-1272, *2013*.

[107] E. Ternon, P. Agyapong, L. Hu, and A. Dekorsy, "Database-aided energy savings in next generation dual connectivity heterogeneous networks," in *IEEE Wireless Communications and Networking Conferenc*, pp. 2811-2816, *2014*.

[108] E. Ternon, P. K. Agyapong, and A. Dekorsy, "Performance evaluation of macro-assisted small cell energy savings schemes," *EURASIP Journal on Wireless Communications and Networking,* vol. 2015, pp. 1-23, 2015.

[109] N. Zhisheng, G. Xueying, Z. Sheng, and P. R. Kumar, "Characterizing Energy-Delay Tradeoff in Hyper-Cellular Networks With Base Station Sleeping Control," *Selected Areas in Communications, IEEE Journal on,* vol. 33, pp. 641-650, 2015.

[110] S. Zhang, J. Gong, S. Zhou, and Z. Niu, "How Many Small Cells Can Be Turned off via Vertical Offloading under a Separation Architecture?," *Wireless Communications, IEEE Transactions on,* vol. 14, pp. 5440-5453, 2015.

[111] X. Zhang, J. Zhang, W. Wang, Y. Zhang, Z. Pan, G. Li*, et al.*, "Macro-assisted data-only carrier for 5G green cellular systems," *Communications Magazine, IEEE,* vol. 53, pp. 223-231, 2015.

[112] A. J. Fehske and G. P. Fettweis, "On flow level modeling of multi-cell wireless networks," in *International Symposium on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks*, pp. 572-579, *2013*.

[113] S. B. Fred, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level," in *ACM SIGCOMM Computer Communication Review*, 2001, pp. 111-122.

[114] R. Patachaianand and K. Sandrasegaran, "System-Level Modeling and Simulation of Uplink WCDMA," in *Fifth International Conference on Information Technology: New Generations*, pp. 1071-1076, *2008*..

[115] J. Gozalvez and J. Dunlop, "Link level modelling techniques for analysing the configuration of link adaptation algorithms in mobile radio networks," in *Proceedings of European Wireless*, pp. 325-330, 2004.

[116] K. Etemad and M.Y. Lai, *WIMAX technology and network evolution* vol. 6: Wiley-IEEE Press, 2011.

[117] J. C. Ikuno, M. Wrulich, and M. Rupp, "System Level Simulation of LTE Networks," in *IEEE Vehicular Technology Conference,* pp. 1-5, *2010*.

[118] S. Hämäläinen, H. Holma, and K. Sipilä, "Advanced WCDMA radio network simulator," in *Personal, Indoor and Mobile Radio Communications*, pp. 951-955, 1999.

[119] B. L. Jones, P. Aitken, and D. Miller, *C Programming in One Hour a Day, Sams Teach Yourself*, 7th Edition ed.: Sams Publishing, 2013.

[120] G. Perry and D. Miller, *Beginning Programming in 24 Hours, Sams Teach Yourself*, Third Edition ed.: Sams Publishing, 2013.

[121] I. Bisio, S. Delucchi, F. Lavagetto, M. Marchese, G. Portomauro, and S. Zappatore, Eds., *Hybrid Simulated-Emulated Platform–HySEP* (Handbook of Research on Next Generation Mobile Communication Systems. IGI Global, 2015.

[122] E. N. Ekwem and K. Nisar, "An Experimental Study: Using a Simulator Tool for Modelling Campus Based Wireless Local Area Network," *International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC),* vol. 6, pp. 35-53, 2014.
[123] C. S. Lent, *Learning to Program with MATLAB: Building GUI Tools: Building GUI Tools*: Wiley Global Education, 2013.
[124] M. Goldhamer, P. Hemphill, O. Marinchenco, and A. Burr, "BuNGee Deliverable: D5.4 Standardisation Report," http://cordis.europa.eu/docs/projects/cnect/7/248267/080/deliverables/001-BuNGeeD54ARTv11.pdf, accessed May 2016.
[125] T. Jiang, "Reinforcement learning-based spectrum sharing for cognitive radio", PhD Thesis, 2011.
[126] M. Hunukumbure, R. Agarwal, and S. Vadgama, "Handover Mechanisms for Planned Cell Outage in Twin State Green Wireless Networks," in *IEEE Vehicular Technology Conference*, pp. 1-5, *2011*.
[127] P. Kyosti, Juha Meinilä, Lassi Hentilä, Xiongwen Zhao, Tommi Jämsä, Christian Schneider*, et al.*, "IST-4-027756 WINNER II D1.1.2. WINNER II channel models," http://www.ist-winner.org/WINNER2-Deliverables/D1.1.2v1.1.pdf, accessed June 2012.
[128] X. Li, *Radio Access Network Dimensioning For 3G UMTS*: Springer, 2011.
[129] I. Glover and P. M. Grant, *Digital communications*, Third Edition ed.: Pearson Education, 2010.
[130] A. Schmitz, M. Schinnenburg, J. Gross, and A. Aguiar, "Channel Modeling," in *Modeling and Tools for Network Simulation*: Springer, 2010, pp. 191-234.
[131] K. Adachi, J. Joung, S. Sun, and P. H. Tan, "Adaptive coordinated napping (conap) for energy saving in wireless networks," *Wireless Communications, IEEE Transactions on,* vol. 12, pp. 5656-5667, 2013.
[132] 3GPP, "TR 36.942: "Evolved Universal Terrestrial Radio Access (EUTRA); Radio Frequency (RF) system scenarios" version 8.2.0 Release 8," ed. http://www.etsi.org/deliver/etsi_tr/136900_136999/136942/08.02.00_60/tr_136942v080200p.pdf, accessed October 2015.
[133] R. L. Freeman, *Telecommunication system engineering*, 4th edition ed.: John Wiley & Sons, 2015.
[134] B. Black, P. DiPiazza, B. Ferguson, D. Voltmer, and F. Berry, *Introduction to wireless systems*: Prentice Hall Press, 2008.
[135] D. Gross and C. M. Harris, *Fundamentals of queuing theory*: Wiley New York, 1998.
[136] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*: John Wiley & Sons, 2010.
[137] Y. Wang, J. Li, Y. Li, R. Wang, and X. Yang, "Confidence Interval for Measure of Algorithm Performance Based on Blocked 3 2 Cross-Validation," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 27, pp. 651-659, 2015.
[138] R. L. Scheaffer and J. T. McClave, *Probability and statistics for engineers*, Third ed.: PWS-KENT Publishing Company, 1990.
[139] M. J. Panik, *Statistical Inference: A Short Course*: John Wiley & Sons, 2012.
[140] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C. Soong*, et al.*, "What will 5G be?," *Selected Areas in Communications, IEEE Journal on,* vol. 32, pp. 1065-1082, 2014.

[141] O. Marinchenco, A. Tipograff, L. Eller-Shein, M. Goldhamer, N. Bartzoudis, M. Dohler*, et al.*, "BuNGee Deliverable: D5.3 BuNGee Exploitation Action Plan," http://cordis.europa.eu/docs/projects/cnect/7/248267/080/deliverables/001-BuNGeeD53v04.pdf, accessed May 2016.

[142] F. Tomatis and S. Sesia, "Synchronization and Cell Search," in *LTE: The UMTS Long Term Evolution*, S. Sesia, I. Toufik, and M. Baker, Eds., Second Edition: John Wiley & Sons, Ltd, 2009, pp. 141-157.

[143] P. Chaturvedi and K. Gupta, "Detection and Prevention of various types of Jamming Attacks in Wireless Networks," *IRACST International Journal of Computer Networks and Wireless Communications (IJCNWC),* pp. 2250-3501, 2013.

[144] T. L. Antti Toskala, Esa Tiirola, Kari Hooli, Mieszko Chmiel, Juha Korhonen, "Physical Layer," in *LTE for UMTS Evolution to LTE-Advanced* A. T. Harri Holma, Ed., Second Edition: John Wiley & Sons Ltd, pp. 83-139, 2011.

[145] R. P. Jover, "LTE PHY Fundamentals ", http://www.ee.columbia.edu/~roger/LTE_PHY_fundamentals.pdf, accessed December 2015.

[146] W. H. Antii Toskala, Colin Willcock, "LTE Radio Protocols," in *LTE for UMTS Evolution to LTE-Advanced* A. T. Harri Holma, Ed.: John Wiley & Sons Ltd, pp. 141-184, 2011.

[147] A. Mohamed, O. Onireti, Y. Qi, A. Imran, M. Imran, and R. Tafazolli, "Physical Layer Frame in Signalling-Data Separation Architecture: Overhead and Performance Evaluation," in *Proceedings of European Wireless*, pp. 1-6, *2014*.

[148] S. Landstrom, H. Murai, and A. Simonsson, "Deployment aspects of LTE pico nodes," in *IEEE International Conference on Communications Workshops (ICC),*pp. 1-5, *2011*.

[149] GreenTouch, "GreenTouch Green Meter Research Study: Reducing the Net Energy Consumption in Communications Networks by up to 90% by 2020," ed. http://www.greentouch.org/uploads/documents/GreenTouch_Green_Meter_Research_Study_26_June_2013.pdf, accessed January 2015.

[150] H. Kim and G. De Veciana, "Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals' energy," *IEEE/ACM Transactions on Networking (TON),* vol. 18, pp. 802-815, 2010.

[151] GreenTouch, "GreenTouch Green Meter Research Study: Reducing the Net Energy Consumption in Communications Networks by up to 90% by 2020," http://www.greentouch.org/uploads/documents/GreenTouch_Green_Meter_Research_Study_26_June_2013.pdf, accessed October 2015.

[152] J. Ramiro and K. Hamied, *Self-Organizing Networks (SON): Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE*: John Wiley & Sons, 2012.

[153] J. Moysen and L. Giupponi, "A Reinforcement Learning Based Solution for Self-Healing in LTE Networks," in *IEEE Vehicular Technology Conference,* pp. 1-6, *2014*.

[154] H. Hu, J. Zhang, X. Zheng, Y. Yang, and P. Wu, "Self-configuration and self-optimization for LTE networks," *Communications magazine, IEEE,* vol. 48, pp. 94-100, 2010.

[155] K. Tsagkaris, G. Poulios, P. Demestichas, A. Tall, and Z. Altman, "An open framework for programmable, self-managed radio access networks," *Communications Magazine, IEEE,* vol. 53, pp. 154-161, 2015.

[156] S. Hämäläinen, H. Sanneck, and C. Sartori, *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*: John Wiley & Sons, 2012.

[157] P. J. M. Johansson, "Method for Centralizing MDT User Involvement," Google Patents, 2012.

[158] C. Schafer and B. Muller, "LTE Impact on OSS Landscape," ed. https://www.7p-group.com/fileadmin/redaktion/portfolio/leistungen/mobile_solutions/7P_Whitepaper_LTE_Impact_on_OSS_Landscape.pdf, accessed October 2015.

[159] W. Dos Passos, *Numerical methods, algorithms and tools in C#*: CRC Press, 2009.

[160] H. Yang, F. Ren, C. Lin, and B. Liu, "Energy efficient cooperation in underwater sensor networks," in *International Workshop on Quality of Service,* pp. 1-9, *2010*.

[161] C. Gao, X. Sheng, J. Tang, W. Zhang, S. Zou, and M. Guizani, "Joint mode selection, channel allocation and power assignment for green device-to-device communications," in *IEEE International Conference on Communications (ICC),* pp. 178-183, *2014*.

[162] M. Dixit, B. Barbadekar, and A. B. Barbadekar, "Packet classification algorithms," in *IEEE International Symposium on Industrial Electronics,.*, pp. 1407-1412, *2009*.

[163] Y. Ma, H. Xie, Z. Chen, Q. Dai, Y. Huang, and G. Ji, "Fast Search of Binary Codes with Distinctive Bits," in *Advances in Multimedia Information Processing*: Springer, pp. 274-283, 2014.

[164] L. Kleinrock, "Queueing Systems, Volume I: Theory,": John Wiley & Sons, 1975.

[165] R. M. Buehrer, "Code division multiple access (CDMA)," *Synthesis Lectures on Communications,* vol. 1, pp. 1-192, 2006.

[166] R. Parkinson, "Traffic engineering techniques in telecommunications," *INFOTEL SYSTEMS CORP. http://www. infotel-systems. com,* accessed November 2015.

[167] K. Rafique, C. Yuan, and M. Saeed, "Net Neutrality Paradox: Regulator's Dilemma," in *International Conference on Wireless Communications, Networking and Mobile Computing,* pp. 1-5, *2011*.

[168] R. T. Ma and V. Misra, "The public option: a nonregulatory alternative to network neutrality," *Networking, IEEE/ACM Transactions on,* vol. 21, pp. 1866-1879, 2013.

[169] S. Jordan, "Traffic management and net neutrality in wireless networks," *Network and Service Management, IEEE Transactions on,* vol. 8, pp. 297-309, 2011.

[170] J. Domżał, R. Wójcik, and A. Jajszczyk, "Qos-aware net neutrality," in *First International Conference on Evolving Internet*, pp. 147-152, *2009*.

[171] G. Piro, M. Miozzo, G. Forte, N. Baldo, L. A. Grieco, G. Boggia*, et al.*, "Hetnets powered by renewable energy sources: Sustainable next-generation cellular networks," *Internet Computing, IEEE,* vol. 17, pp. 32-39, 2013.

[172] B. Wang, Q. Kong, and L. Yang, "Context-Aware User Association for Energy Cost Saving in a Green Heterogeneous Network with Hybrid Energy Supplies," *Mobile Networks and Applications,* pp. 1-15, 2015.

[173] H. Tao and N. Ansari, "Green-energy Aware and Latency Aware user associations in heterogeneous cellular networks," in *IEEE Global Communications Conference*, pp. 4946-4951, *2013*.

[174] M. Miozzo, L. Giupponi, M. Rossi, and P. Dini, "Distributed Q-learning for energy harvesting Heterogeneous Networks," in *IEEE International Conference on Communication Workshop,* pp. 2006-2011, *2015*.

[175] Z. Han and K. R. Liu, *Resource allocation for wireless networks: basics, techniques, and applications*: Cambridge university press, 2008.

[176] J. Chen, C.-C. Yang, and S.-T. Sheu, "Downlink Femtocell Interference Mitigation and Achievable Data Rate Maximization: Using FBS Association and Transmit Power-Control Schemes," *Vehicular Technology, IEEE Transactions on,* vol. 63, pp. 2807-2818, 2014.

[177] Y. Jin, F. Cao, and R. A. Dziyauddin, "Inter-cell interference mitigation with coordinated resource allocation and adaptive power control," in *IEEE Wireless Communications and Networking Conference*, pp. 1797-1802, *2014*.

[178] Y. Wang, M. Qian, X. Han, Y. Zhou, and J. Shi, "Game-Theoretic Power Control for Interference Mitigation in Two-Tier Small Cell Networks," in *IEEE Vehicular Technology Conference (VTC Spring),* pp. 1-5, *2014*.

[179] E. Biton, A. Cohen, G. Reina, and O. Gurewitz, "Distributed inter-cell interference mitigation via joint scheduling and power control under noise rise constraints," *Wireless Communications, IEEE Transactions on,* vol. 13, pp. 3464-3477, 2014.