HANDOVER MANAGEMENT IN HETEROGENEOUS NETWORKS FOR 4G AND BEYOND CELLULAR SYSTEMS

A Dissertation
Presented to
The Academic Faculty

By

Ravikumar Balakrishnan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
May 2015

HANDOVER MANAGEMENT IN HETEROGENEOUS NETWORKS FOR 4G AND BEYOND CELLULAR SYSTEMS

Approved by:

Dr. Ian F. Akyildiz, Advisor
Ken Byers Chair Professor in Telecommunications, School of ECE
Georgia Institute of Technology

Dr. Matthieu Bloch
Assistant Professor, School of ECE
Georgia Institute of Technology

Dr. Chuanyi Ji
Professor, School of ECE
Georgia Institute of Technology

Dr. Russell Clark
Sr. Research Scientist, College of Computing
Georgia Institute of Technology

Dr. Geoffrey Y. Li
Professor, School of ECE
Georgia Institute of Technology

Date Approved: March 4, 2015

To my family for their unconditional love and support.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my advisor and mentor, Dr. Ian F. Akyildiz, for providing me this wonderful and career-changing opportunity to work under his guidance for my Ph.D. I am profoundly thankful to his boundless energy and encouragement to help me steadily progress in my research direction. I am also thankful for his guidance and mentoring towards shaping my career path.

I would also like to extend my appreciation to all the academic members of the School of Electrical and Computer Engineering at the Georgia Institute of Technology for their insightful and critical reviews throughout my Ph.D. program. In particular, I would like to sincerely thank Dr. Chuanyi Ji, Dr. Geoffrey Ye Li, Dr. Matthieu Bloch and Dr. Russell J. Clark who kindly agreed to serve in my Ph.D. Defense Committee. Their invaluable comments have helped me to build a solid research direction for my thesis.

I would also like to thank the past and present members of the Broadband Wireless Networking (BWN) Lab, for their continued support since the beginning of my Ph.D study and for creating a very friendly atmosphere in the work place. It has been a great learning experience for me to have interacted on a personal and professional level with them over these years.

TABLE OF CONTENTS

ACKNOWLEDGMENT iv

LIST OF TABLES vi

LIST OF FIGURES vii

SUMMARY x

CHAPTER 1 INTRODUCTION 1
- 1.1 Evolution to beyond 4G Cellular Systems 2
- 1.2 Heterogeneous Networks 5
- 1.3 Research Objectives and Solutions 7
  - 1.3.1 Handover Management in Small Cells 8
  - 1.3.2 Handover Management in Multistream HetNets 10
- 1.4 Organization of the Thesis 11

CHAPTER 2 HANDOVER MANAGEMENT IN SMALL CELLS 12
- 2.1 Handover Initiation 15
  - 2.1.1 Traffic-aware Admission Control 15
- 2.2 Handover Execution and Completion 36
  - 2.2.1 Local Anchor based Handover for Clustered Small Cells 36
  - 2.2.2 Group Handover for Mobile Relays 70

CHAPTER 3 HANDOVER MANAGEMENT IN MULTI-STREAM HETNETS 86
- 3.1 Joint Small Cell Discovery and Component Carrier Selection for Multi-stream HetNets 86
  - 3.1.1 Multistream HetNet Architecture 87
  - 3.1.2 Handover Challenges in Multistream HetNets 89
  - 3.1.3 Joint Small Cell Discovery and Component Carrier Selection . . . 92
  - 3.1.4 Performance Evaluation 101
  - 3.1.5 Conclusions 108

CHAPTER 4 CONCLUSIONS 111

REFERENCES 115

PUBLICATIONS 120

VITA 122

LIST OF TABLES

Table 1 Beyond 4G key performance indicators. 2

Table 2 Capacity gain for cellular networks using different approaches. 5

Table 3 Simulation parameters. 32

Table 4 System parameters. 61

Table 5 Group handover cost functions. 83

Table 6 Simulation parameters. 103

LIST OF FIGURES

Figure 1 5G performance requirements in comparison to past generations. [1] . . . 3

Figure 2 B4G deployment scenarios. 4

Figure 3 Heterogeneous networks (HetNets). 6

Figure 4 3GPP LTE-A inter-small cell handover using direct X2 interface. 13

Figure 5 Considered network topology. 19

Figure 6 The proposed admission control framework. 22

Figure 7 Throughput performance under traffic-aware utility based scheduling. . . 34

Figure 8 Mean delay performance of CBR traffic class. 34

Figure 9 Mean delay performance of video traffic class. 35

Figure 10 Mean delay performance of BE traffic class. 35

Figure 11 Delay variance performance of CBR traffic class. 36

Figure 12 Delay variance performance of video traffic class. 37

Figure 13 Delay variance performance of BE traffic class. 38

Figure 14 Local anchor based handover architecture. 42

Figure 15 Local anchor registration table. 43

Figure 16 Local anchor registration table update after handover. 44

Figure 17 LP-based handover mechanism for coordinated small cells. 46

Figure 18 Triangular routing of in-transit data during handover. 47

Figure 19 RO-enhanced handover mechanism for coordinated small cells. 48

Figure 20 DF-enhanced handover mechanism for LA to SC handover. 49

Figure 21 Two-dimensional grid topology for small cell cluster. 51

Figure 22 Grid topology representing aggregated states. 52

Figure 23 Discrete-time Markov model for the aggregated states. 53

Figure 24 Numerical evaluation of handover performance vs number of tiers K and path switching threshold v. 60

Figure 25 Handover interruption time vs number of tiers K and path switching threshold ν. 60
Figure 26 Simulation of a small cell cluster with K = 2. 63
Figure 27 Impact of ν on average signaling cost ratio. 66
Figure 28 Impact of ν on average data forwarding cost ratio. 67
Figure 29 Impact of ν on handover interruption time ratio. 67
Figure 30 Impact of call-to-mobility ratio on signaling cost. 68
Figure 31 Impact of call-to-mobility ratio on handover interruption time. 69
Figure 32 Data forwarding cost under different mobility models 70
Figure 33 Handover interruption time under different mobility models. 71
Figure 34 Group handover scenario in high-speed vehicles with mobile small cells. 72
Figure 35 Relay architecture in 4G WiMAX 802.16m systems. 75
Figure 36 C-plane protocol stack for relay support in 4G WiMAX. 76
Figure 37 U-plane protocol stack for relay support in 4G WiMAX. 77
Figure 38 Proposed group handover mechanism for 4G WiMAX systems. 84
Figure 39 Proposed group handover architecture. 85
Figure 40 Signaling overhead for group handover. 85
Figure 41 Multistream architecture. 88
Figure 42 Inter-band carrier aggregation for beyond 4G systems. 89
Figure 43 Handover scenarios in multi-stream HetNets. 91
Figure 44 Modeling the CC states. 94
Figure 45 Markov model for CC state transition. 96
Figure 46 Average aggregate SCC throughput per user for different values of m_1. . . 104
Figure 47 Average SCell change rate for different values of m_1 and m_2. 105
Figure 48 Average SCell change rate for different user velocities. 106
Figure 49 Average ping-pong rate for different values of m_1 and m_2. 107
Figure 50 Average ping-pong rate for different user velocities. 108

Figure 51 Proposed SCell discovery and handover mechanism for 4G and beyond systems. 110

SUMMARY

New technologies are expected to play a major role for wireless cellular systems beyond the existing 4G paradigm. The need for several orders of magnitude increase in system capacity has led to the proliferation of low-powered cellular layers overlaid on the existing macrocell layer. This type of network consisting of different cellular layers, each with their unique characteristics including transmission power and frequency of operation among others is termed as a heterogeneous network (HetNet). The emergence of HetNets leads to several research challenges and calls for a profound rethinking of several existing approaches for mobility management and interference management among other issues.

The objective of this thesis is to study the handover performance of current and emerging HetNet architectures and to propose novel approaches to achieve seamless mobility for users in multi-layer HetNets. The starting point is to consider the baseline case of inter-small cell handovers. Taking into account of the differentiating features of small cells including the unique access control policies and backhaul limitations, we propose handover management strategies for inter-small cell handovers. Within this, a traffic-aware admission control policy is proposed and evaluated for hybrid-access small cells that governs the handover decisions at the beginning of a handover process. Following the admission control policy, a local-anchor based handover mechanism is proposed considering a cluster of small cells in order to achieve seamless mobility and minimal handover-related costs. For the special case of mobile small cells (or mobile relays) that are suitable for deployment in high-speed vehicles, a group handover scheme is proposed and evaluated where a single group handover procedure can ensure handover of a group of users served by the mobile relay station between neighboring macrocells. Finally, handovers in multistream HetNets is considered where the problem of handover between the smaller cellular layers is transformed into a component carrier selection problem. For this, a joint small cell discovery and component carrier selection approach is proposed and evaluated.

CHAPTER 1

INTRODUCTION

Cellular networks are undergoing a major transformation as they are increasingly expected to handle an enormous number of mobile devices as well as applications and services. The annual visual network index by Cisco forecasts wireless data traffic growth over 190 exabytes in 2018 and over 500 exabytes by 2020 [2]. It is also envisioned that there will be over 50 billion human as well as machine-type devices by the year 2020. Not only that, but there is also a steady growth in the number and types of applications and services that are becoming available for the mobile segment, a large share of which are cloud-based.

In this regard, several incremental enhancements over the existing 4G systems as well as radical technologies envisioned for 5G systems are rapidly evolving. We collectively refer to these as beyond 4G (B4G) systems. Several academic and industry initiatives including the METIS project [1] laid out the key performance indicators for 5G systems envisioned to be available in 2020. The project highlights the increase in performance required over the existing 4G systems. These are shown in Figure 1. According to this, cellular system beyond the 4G will focus on the support for three major communication paradigms including mobile broadband, massive machine-type communication as well as mission-critical machine communication. In order to efficiently realize these diverse applications, the B4G systems will need to achieve user data rates of 10Gbps, $1000x$ increase in system capacity compared to 4G systems and to provide support for ultra-high speed mobile users. In addition, they are also expected to achieve $10x$ energy savings for the devices as well as the network infrastructure. Furthermore, in order to support data communication at ultra-high speeds, the B4G systems should be capable of supporting ultra-low latency communication of the order of 1ms. Table 1 summarizes these requirements.

Table 1: Beyond 4G key performance indicators.

Performance Indicator	Value
Peak Data Rate	10 Gbps
Latency	1 ms
Number of devices supported	100x
Energy Efficiency	10x improvement

1.1 Evolution to beyond 4G Cellular Systems

The Third Generation Partnership Project (3GPP) has been working since 2010 on the definition of Long Term Evolution-Advanced (LTE-Advanced or simply LTE-A) systems and its component technologies. Rel-10 of 3GPP started early in 2010 and was functionally frozen in March 2011 after its approval by the ITU for having met all the requirements for IMT-Advanced. Technologies introduced during that release include carrier aggregation for transmissions in several frequency bands, enhanced multiple input-multiple output (MIMO) techniques, relays, and self-organizing networks (SON).

Then, Rel-11 was started and further enhancements were included to the basic LTE-Advanced technologies developed for Rel-10. The major contribution during Rel-11 was cooperative multipoint transmission and reception (CoMP), which allows different cells to cooperate to serve users. In addition, several important enhancements were introduced for heterogeneous networks (HetNets) such as enhanced inter-cell interference cancellation (eICIC) and mobility management enhancements.

After the recent completion of Rel-11, the standardization work started focusing on Rel-12 LTE-Advanced. Enhancements to CoMP (inter-site CoMP), carrier aggregation (multi-stream carrier aggregation), and MIMO (FullDimension MIMO) are key items in the agenda of 3GPP. Moreover, radically new technologies will also be introduced: Machine-Type Communication (MTC) will enable machines to interact among themselves as part of a large network, and Device-to-device (D2D) communication will allow mobile users to interact with each other without the need to go through the network [3].

Depending on the progress in Rel-12, future releases will start earlier or later in 2014.

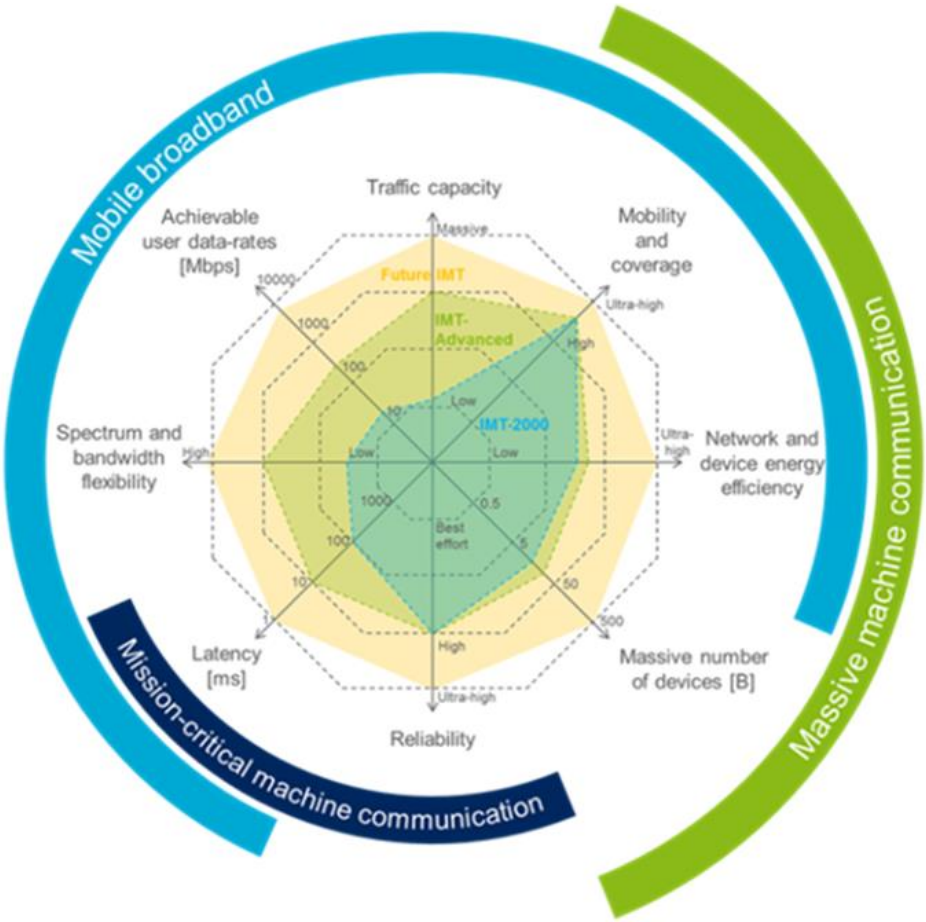

Figure 1: 5G performance requirements in comparison to past generations. [1]

Rel-13 is expected to further enhance LTE-Advanced technologies while Rel-14 and Rel-15 could potentially define a new access technology. Historically, cellular systems have taken considerable leap from one generation to another in terms of technological innovation. This looks increasingly certain as we are getting closer to the 5G standards that are planned to be made available by 2020. Not only the 5G cellular systems will be of a disruptive nature, but also expected to be a first effort in truly integrating the legacy cellular systems as well as other wireless technologies such as WiFi to work in tamdem towards satisfying the growing data needs.

The evolution from 4G to 5G systems will make possible a number of deployment scenarios that haven't existed in 4G systems. These scenarios will play a great role in defining the technologies that will drive the B4G innovation. B4G cellular systems are

expected to utilize new frequency bands including the 5GHz unlicensed bands through the use of intersite carrier aggregation or dual connectivity. More significantly, the use of ultra-high frequencies including mmwave and terahertz to realize peak data rates of over 10Gbps is also being investigated rigorously. These mmwave or terahertz small cells are also expected to support massive multiple-input multiple-output (MIMO) antennas in order to overcome the high pathloss and blocking associated with such high frequencies.

B4G systems are also expected to leverage the different radio access technologies and provide a unified architecture to exploit transmission opportunities across several RATs in a robust and seamless fashion. Another striking scenario will be the deployment of massive number of machine type devices. These devices can further be classified into low-latency, low-power and mission critical devices. Such a wide-range of devices will require the use of a multitude of technologies that will be part of B4G systems. Ad hoc deployments such as device-to-device communications and inter-vehicular and vehicle-to road communications are also envisioned as part of the B4G deployment scenarios. The different deployment scenarios for B4G systems are illustrated in Figure 2.

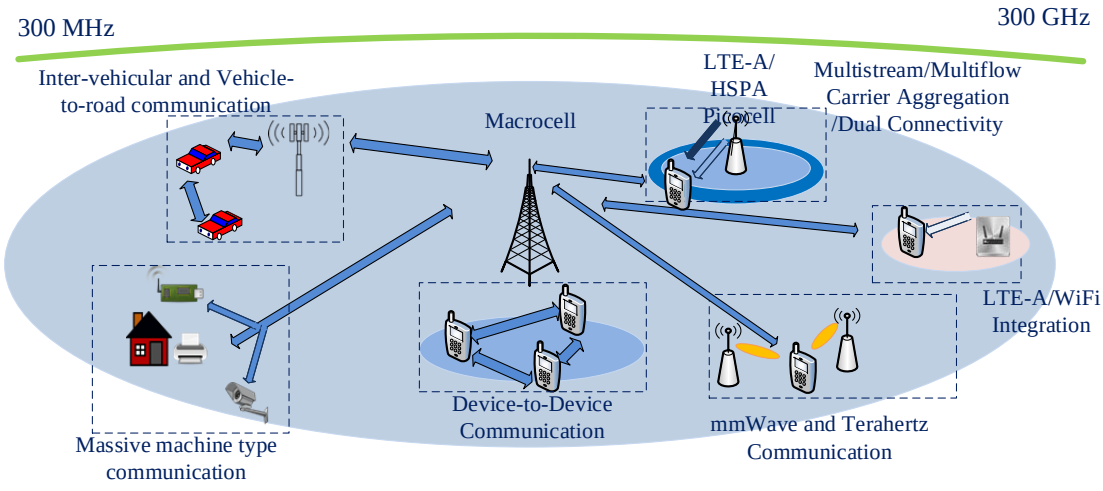

Figure 2: B4G deployment scenarios.

Table 2: Capacity gain for cellular networks using different approaches.

Approach	Capacity Gain
Frequency Division	5
Higher-order Modulation	5
Access to wider range of spectrum	25
Frequency reuse using more cell sites	1600

1.2 Heterogeneous Networks

Table 2 indicates the capacity gain achieved in cellular networks using different approaches [4]. Among these different approaches, it is evident that spatial frequency reuse using more cell sites offers over 3 orders of magnitude capacity gain. In addition, studies also show that more than 50% of voice traffic and 70% of cellular data traffic originate from indoor and enterprise environments [5]. This has led to the proliferation of smaller low-powered cellular layers overlaid on the existing macrocell layer. These low-powered small cells include picocells, femtocells, metrocells, relays among others. Figure 3 showcases a scenario with macrocells overlaid with different small cell types. While picocells, relays and metrocells are utilized for outdoor deployments, femtocells are proposed for deployment in indoor environments such as residential or enterprise buildings. In enterprise deployments, multiple femtocells are typically deployed in a coordinated fashion where the small cell base stations can adaptively self-organize and optimize their transmission parameters.

Such a type of network that includes several overlapping cellular layers, each with their unique characteristics such as transmission power, carrier frequency and backhaul technology is termed as a heterogeneous network (HetNet). HetNets, overall, provide a significant improvement in the network performance and service connectivity by enabling dynamic traffic offloading from macrocell for a number of purposes including network load balancing, capacity boost or coverage extension.

Small cells are, therefore, expected to play a major role in enhancing coverage and capacity of 4G and 5G cellular systems. Small cells are also expected to feature unique access control features which are listed as follows:

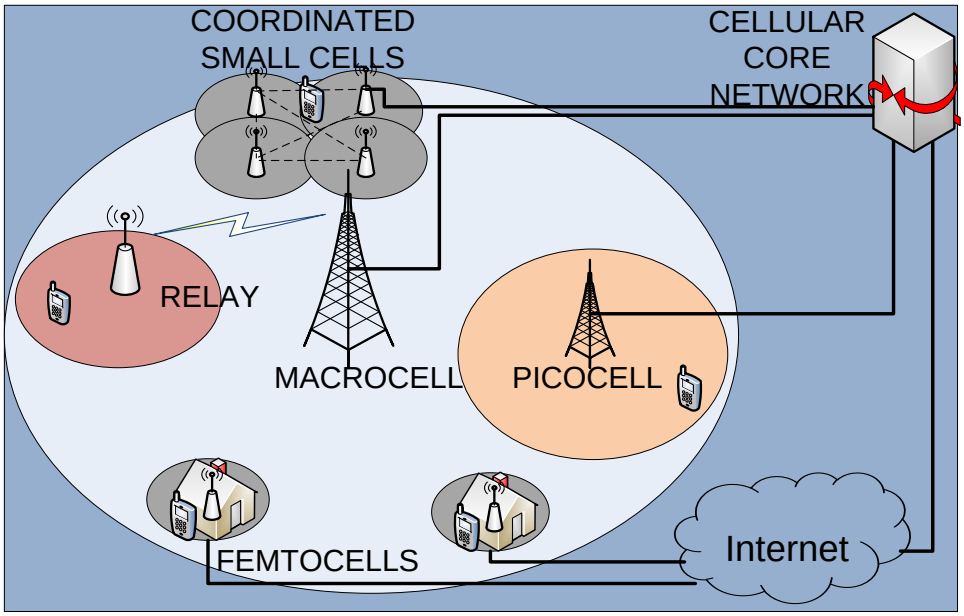

Figure 3: Heterogeneous networks (HetNets).

- **Open Access:** Small cells are open for all users of the cellular network.
- **Closed Access:** Small cell access is reserved for users that obtain a closed subscriber group (CSG) access from the cellular network provider.
- **Hybrid Access:** In addition to providing access to CSG-capable users, small cells are capable of providing limited access to users that do not have CSG capabilities.

However, the heterogeneous nature of the different cellular layers raises several important challenges that impact the actual capacity that can be achieved by the overall system as well as the quality of experience offered to the mobile users. In particular, the key challenges affecting this novel network architecture are inter-cell interference and mobility management. Although these two large problems have been existing in the literature, several essential features of those are still challenging open problems for the community.

Especially, the new deployment scenarios envisioned for B4G systems such as the multi-stream carrier aggregation and mmWave communication significantly alter the cellular network architecture. In addition to the multi-layer HetNet case, 4G systems have focused on the interworking between the different RATs. However, a unified architecture

to exploit all the RATs to achieve maximum gains is only expected to become dominant in B4G systems. Several new architectures are currently being investigated for multi-layer multi-RAT HetNets. Multi-stream carrier aggregation and dual connectivity where the macrocell acts as anchor cell providing coverage and small cells act as capacity boost are promising architectures to achieve seamless mobility performance for users.

1.3 Research Objectives and Solutions

Achieving seamless and robust mobility in HetNets is a major challenge. A detailed study of mobility performance for HetNets show that the handover failure rates and ping-pong effects are far greater in HetNet environments compared to homogeneous deployments [6]. The poor handover performance of users in HetNets is owing to several factors including high handover signaling cost due to large backhaul delays, inefficient small cell detection, high cost for recovery from handover failures. One of the major factors is the need to account for the unique characteristics of HetNets when performing user mobility state estimation in addition to accounting for parameters such as the number of cross-overs. Another factor is the need to perform efficient small cell discovery over different carriers at the base stations. This is essential to enable traffic offload from macro to small cells as well as to reduce the battery consumption for the users and the overall system. Furthermore, since the handover failures can be larger in the HetNet case, more robust recovery procedures are required to mitigate the impact of a failed handover.

The objective of this thesis is to study the handover performance of current and emerging HetNet architectures and to propose novel approaches to achieve seamless mobility for users in a multi-layer HetNet environment. The starting point is to consider the baseline case of inter-small cell handovers. Taking into account the differentiating features of small cells including the unique access control policies and backhaul limitations, we propose handover management strategies for inter-small cell handovers. Within this, a traffic-aware

admission control policy is proposed and evaluated for hybrid-access small cells that governs the handover decisions at the beginning of a handover process. Following the admission control policy, a local-anchor based handover mechanism is proposed considering a cluster of small cells in order to achieve seamless mobility and minimal handover-related costs. For the special case of mobile small cells (or mobile relays) that are suitable for deployment in high-speed vehicles, a group handover scheme is proposed and evaluated where a single group handover procedure can ensure handover of a group of users served by the mobile relay station between neighboring macrocells. Finally, handovers in multi-stream HetNets is considered where the problem of handover between cell layers is transformed into a component carrier selection problem. For this, a joint small cell discovery and component carrier selection approach is proposed and evaluated.

In the following sections, the developed solutions within each topic are summarized.

1.3.1 Handover Management in Small Cells

Unlike handovers between macrocells, several factors need to be accounted when designing handover mechanisms between small cells. First, the varying backhaul link quality of small cells implies that handover between small cells should not involve significant amount of signaling on the backhaul links. Especially, due to the small cell sizes, frequent handovers may occur which could also contribute significantly to the signaling load on the core network. Hence, it is more suitable to anchor handovers locally whenever possible thereby minimizing the load on the core network. Furthermore, the access control features at small cells mean that the admission control decisions at the target small cell during handover need to account for the access privileges available for the mobile user and must not degrade the performance of the users already being served at the target small cell.

The first contribution of this thesis (Chapter 2) is the design, modeling and analysis of admission control and handover for small cells. Within this, we focus initially on admission control for small cells taking into account their unique access control features.

The proposed admission control approach for hybrid access small cells exploits the traffic-awareness in a cross-layer fashion to determine if the incoming users at the small cells can be admitted. In addition, the proposed approach also studies the existing active users and performs pre-emption and QoS decisions in order to satisfy the per-user QoS guarantees. Furthermore, in order to achieve the long-term QoS for the users, a subcarrier assignment strategy for the currently active users is proposed in order to stabilize the user queues for heterogeneous traffic arrivals.

Following this, we provide seamless and low-cost handover mechanisms for two different scenarios. For the first scenario, we consider the case of a cluster of coordinated small cells, also termed as network of small cells. For this case, we propose local anchor based handover strategies where handovers for users between small cells are anchored by a local anchor small cell. A new handover architecture is proposed that can be extended to the current LTE-A systems. We propose an analytical model to evaluate the proposed scheme considering different cluster sizes, user session and mobility dynamics. Our mathematical analysis as well as detailed simulation shows that the proposed schemes significantly minimize the handover interruption time, signaling cost and the core network load associated with handovers.

For the second scenario, we consider the case of mobile small cells (or mobile relays) and propose a group handover strategy to enable the handover of a group of users at once. Our proposed scheme allows the mobile small cells to changes the point of attachment from one macrocell to another while retaining the sessions with the group of users. Therefore, with the handover of the mobile small cell between macrocells, the proposed scheme preserves the user sessions and alleviates the need for separate handover process for each user served under the mobile small cell. Our analysis shows that the proposed group handover strategy can offer significant signaling savings both at the radio access network as well as at the core network.

1.3.2 Handover Management in Multistream HetNets

With the rapid advances in inter-band carrier aggregation, new HetNet architectures are envisioned such as the inter-site or multi-stream aggregation where users can be served by macrocell and small cell simultaneously. The macrocell acts as the primary cell (PCell) providing sufficient coverage for the user while the user can dynamically add or remove small cells that act as secondary cells (SCells) to improve user data rates. Such an architecture requires the macro and small cells to be interconnected using a high-speed backhaul. This scenario is also gaining prominence with the emergence of more centralized architectures involving increased functionality at the macrocells which control a set of remote radio heads deployed under the macrocell coverage through high-speed backhaul links.

The advantage with this architecture is that the macrocell being the larger cell provides reliable coverage for the user. On the other hand, the users can dynamically add, remove or change small cells based on their data rate needs. Therefore, the users do not experience a discontinuity in the service since the macrocell will maintain signaling and possibly data connection with the user while the user undergoes handover between small cells.

Although multistream HetNets significantly improve the handover performance for the users, they pose several challenges in terms of energy consumption and signaling load. Especially, the emergence of carrier aggregation requires the users to perform additional inter-frequency measurements not only for offloading opportunities, but to achieve a higher data rate.The user has to disconnect from the serving cell in order to perform measurements on other component carriers (CCs), and therefore the user cannot be provided any service during this acquisition period. The service interruption time due to this acquisition is called the measurement gap. Achieving the target data rate while minimizing the number of measurements and therefore the measurement gaps is a major challenge. Furthermore, frequently switching between small cells to improve connectivity and user data rates requires constantly setting up and tearing down signaling and data connections for the users and overloads the network.

To this end, we treat the small cell discovery and component carrier selection for multistream HetNets as a coupled problem. We propose and evaluate a joint small cell discovery and component carrier selection approach to minimize the service interruption, energy consumption as well as the signaling load accompanied with these critical tasks.

1.4 Organization of the Thesis

The thesis is organized as follows. In Chapter 2, handover management in small cells is presented. First, a traffic-aware admission control mechanism is proposed, modeled and analyzed. Then, two different handover schemes are presented. In the first one, a local anchor based scheme to provide seamless handover and minimize handover related signaling is proposed, modeled and evaluated. In the second one, a group handover scheme for mobile relays is proposed and evaluated.

In Chapter 3, small cell discovery and handover in multistream HetNets are discussed. A joint small cell discovery and component carrier selection framework is proposed, modeled and analyzed. Finally, in Chapter 4, the research contributions are summarized and future research directions are identified.

CHAPTER 2

HANDOVER MANAGEMENT IN SMALL CELLS

The handover procedure for 3GPP LTE-A systems utilizing the direct interface (X2 interface) between small cells (SCs) is illustrated in Figure 4 [7]. The handover procedure is divided into three phases: (i) handover initiation, (ii) handover execution, and (iii) handover completion. During the *handover initiation* phase, the mobile user (MU) sends a measurement report of its neighboring cells to the serving SC. The serving SC determines the target SC based on the measurement report and sends a handover request to the target SC. The target SC, in turn, executes the admission control algorithm to determine if the user can be admitted with the requested resources. During this process, the target SC also evaluates the resources available for its existing and incoming session and re-allocates resources for all the users. If the new user is admitted, the target SC sends a handover response message to the serving SC. In addition, an uplink data path is established between the target SC and the core network (Serving gateway or S-GW).

During the *handover execution* phase, the serving SC issues a handover command message to the MU to initiate handover execution. The MU now starts attaching to the target SC which includes the establishment of a radio connection. During the same time, the serving SC will forward all the "in-transit" data in its buffer to the target SC. The completion of the radio connection setup is indicated by a handover confirm message from the MU to the target SC.

In the *handover completion* phase, to complete the downlink data path setup from the core network to the target SC, the target SC sends a path switch request message to the core network (MME) after receiving the handover confirm message. While this process is being carried out, the buffered data from serving SC is continued to be forwarded to the target SC. Once the S-GW has completed the downlink path setup with target SC, a path switch request acknowledgement is sent to the target SC from the MME. In addition, an

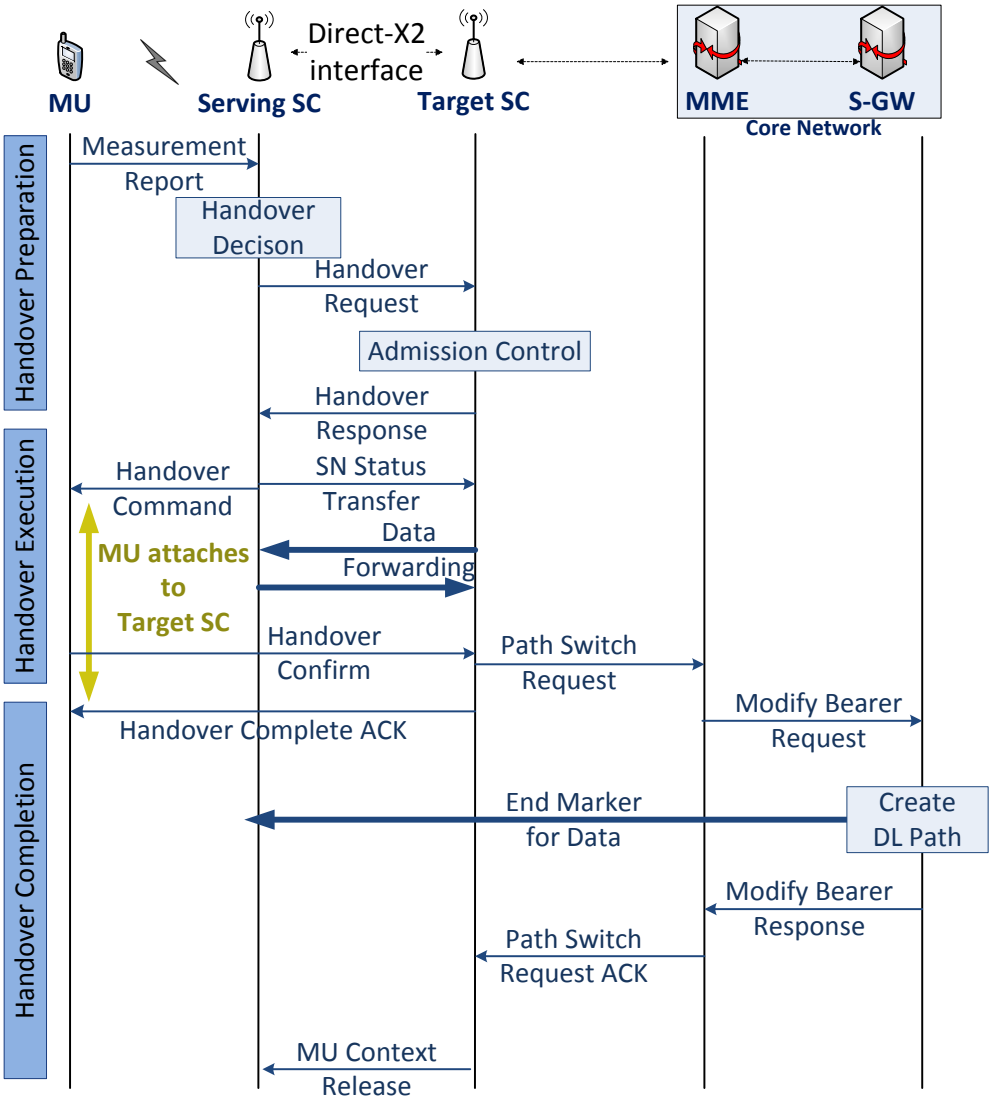

Figure 4: 3GPP LTE-A inter-small cell handover using direct X2 interface.

end marker is sent to the serving SC which is forwarded to the target SC. At this point, the downlink data from the core network for the MU is directly received by the target SC. This also marks the completion of the handover procedure and the serving SC releases all the resources for the MU.

As described earlier, handover is the key mechanism by which users can maintain their session active while moving across several cells. Handovers in a homogeneous cellular network is straight-forward that users arriving at the edge of their current cell will need to handover to the adjacent cell to continue their session in a seamless manner. However, handovers in the case of HetNets are complex and challenging. First, the average number of handovers experienced by a user in a given time is very large, potentially leading to frequent service interruptions and handover failures. This becomes a bigger issue when carrier aggregation is utilized where the users will experience more handover scenarios of cell addition/removal besides the traditional handover [7]. Second, due to the prominent internet-based backhaul connectivity adopted for the small cells, large interruption times and core network load are associated with handovers. Finally, access control features are introduced for the small cells where the small cell owners can provide differentiated services to registered (closed subscriber group or CSG) and unregistered users. This requires careful admission control mechanisms at the target SC taking into account the access control features.

Therefore, a successful handover management strategy for small cells will involve the following attributes. **First**, admission control at the target cell must account for the CSG capability of the incoming users. **Second**, the impact of each handover in terms of the signaling overheads, handover interruption time and load on the radio access and core network must be minimized.

2.1 Handover Initiation

2.1.1 Traffic-aware Admission Control

During the handover initiation phase, the small cell needs to perform admission control to determine if a user can be admitted or not, what resources to allocate to the user, whether existing sessions need to be pre-empted in order to admit the new session. Admission control (or access control) also plays an important role for achieving load balancing across cells. Especially, in small cells where open/hybrid access is enabled, regulating the number of incoming connections plays a key role in achieving QoS to users served by small cells. The application of small cells to a targeted area such as enterprise, private buildings result in unique access control policies to handle different user types. There are several access control policies for small cells. With the open access policy, small cells can service all mobile users within its coverage. With the closed access policy, small cells can reserve exclusive access for pre-registered mobile users, also called as small cell users (SUs). Alternatively, the small cells can also implement a hybrid access policy, where they can use ad hoc schemes to achieve QoS guarantees for the SUs in the presence of unregistered external users (EUs).

Although, the open access policy offers the largest increase in the network capacity, it can degrade the QoS performance of SUs served under the small cells. The QoS degradation is particularly large when the number of EUs increases or when the EUs are running bandwidth-hungry applications [8, 9]. The closed access approach is capable of providing better QoS performance for the SUs but the SU performance can also be significantly affected if there are nearby EUs that cause strong interference to the small cell network. Therefore, the hybrid access scheme can provide differentiable service to SUs and EUs, thus more suited for QoS provisioning.

There are several studies which investigate the hybrid access small cell schemes in different perspective. The authors in [10] consider hybrid access in femtocells where they propose a fixed probability p for EUs to be able to connect to a femtocell based on the

computation of the carrier to interference (C/I) ratio at the location of the EUs. In [11], the authors propose a hybrid access scheme for OFDMA small cells where a limited amount of subchannels v is reserved for EU access. Although, the outage probability is shown to notably decrease for EUs in this scheme, increasing v can affect the throughput achieved by SUs. In addition, lower outage probability does not necessarily equal QoS performance of both SUs and EUs. Further work has been conducted under hybrid access approach, for e.g., in [12], the authors propose an adaptive access control strategy based on the average cellular user density. It is shown that the ergodic rate for EUs is notably increased under low user-density case, whereas under the high user-density case, the rate gain for EUs is not significant.

Unfortunately, the above hybrid access schemes fail to consider the nature of the higher layer traffic in performing scheduling and access control for small cells. One of the fundamental performance metrics is network stability which guarantees the queue size to be bounded for all packet arrivals within the capacity region. Scheduling policies like Maximum Delay Scheduling can stabilize the queues for admissible arrival rates. At the same time, these policies can result in poor delay performance and unfair allocation for the SUs if the EU traffic is bursty.

There are also several scheduling policies in the current literature. The EXP rule proposed in [13] is shown to offer improved QoS performance in terms of throughput and delay over proportionally fair (PF) and Modified Largest Weighted Delay First (M-LWDF) scheduling schemes when there is a mixture of real-time and non-real-time users in the system. Another popular approach to QoS scheduling is the utility based approach. Scheduling rules based on maximizing utility, which represents the amount of satisfaction that can be obtained by scheduling a resource for a user, have been proposed in [14, 15, 16]. The utility functions here are defined as decreasing functions of the packet delay in the queue. In [14, 15, 17], although the scheduling rule is shown to achieve throughput optimality, the utility function does not provide strict bounds on delay. In addition, many of

the above policies only consider subcarrier allocation without any power constraints. Since small cells are limited by hardware on the total transmit power as well as by the interference they cause to the overlapping macrocell layer, power constraints become significant in our problem.

It is also important for an admission control scheme to be traffic-aware and mobility-aware. Traffic-aware admission control approaches were proposed in [18, 19]. However, these approaches are unable to provide strict bounds on delay. In addition, the constraints on the maximum transmission power at the base stations have not been accounted. Mobility-aware admission control mechanisms were proposed in [20, 21] where the user QoS requirements are not explicitly captured for the admisison control decisions.

The ultimate objective of having a hybrid access scheme is that QoS for SUs is provisioned while the EUs will also be served in order to maximize resource utilization. In all the aforementioned studies, the traffic characteristics are no considered while doing the scheduling and access control in LTE-A small cells. However, such considerations are crucially needed to obtain most QoS optimal small cell deployment. Here, one of the fundamental performance metrics is network stability which guarantees the queue size to be bounded for all packet arrivals within the capacity region. The capacity region is defined as the convex hull of the m-element set of all arrival rate vectors $\vec{\lambda}(m)$ defined by

$$\vec{\lambda}(m) = [\lambda_1(m), ..., \lambda_N(m)]; \quad \forall m, \tag{1}$$

for a system containing $N(m)$ users that can be supported without making the queues unstable. If the channel state is represented by a finite number of states M, the capacity region depends on the link transmission rate vector $\vec{r}(m)$ as;

$$\vec{r}(m) = [r_1(m), ..., r_N(m)]; \quad \forall m, \tag{2}$$

under state m where $m \in M$.

Scheduling policies like Maximum Delay Scheduling can stabilize the queues for admissible arrival rates. At the same time, these policies can result in poor delay performance and unfair allocation for the SUs if the EU traffic is bursty. One of the objectives of this work [8] is to propose and evaluate an optimal subcarrier assignment scheme that accounts for the higher layer traffic characteristics for QoS provisioning in hybrid small cells. In addition, when the number of users in the cell increases or when the traffic arrivals are outside the capacity region, the scheduler cannot handle fair allocation towards achieving end user QoS. Therefore, in this chapter, an admission control procedure tightly coupled to the scheduling policy is also proposed where both pre-registered small cell users (SUs) and unregistered external users (EUs) can be serviced.

2.1.1.1 Network Model

In this work, we consider a macrocell base station which orchestrates the External Users (EUs) in a given coverage area as shown in Figure 5. There are some pre-defined hybrid access OFDMA small cells in the network. These small cells serve their pre-registered users (Small cell users or SUs) and some EUs which are under the small cell coverage area. Moreover, we consider a time-varying, bursty and location-dependent wireless channel which also poses a major challenge in achieving optimal QoS performance and scheduling. In order to control all these challenges, we design our proposed mechanism taking into consideration the exchange of system dynamics such as channel conditions, location, queue state, application layer requirements to maintain QoS satisfaction. Under such a setup, the time is slotted and the wireless channel is assumed to be unvarying during the slot length. At the beginning of each slot, the scheduler obtains the channel gain from the lower layers through user feedback. Using this information, the data rates achievable and power required for the user in the time slot is determined. Based on these parameters, the subcarrier assignment algorithm performs resource scheduling to achieve the QoS objectives [8].

The downlink of an OFDMA hybrid small cell overlaid on a macrocell coverage area

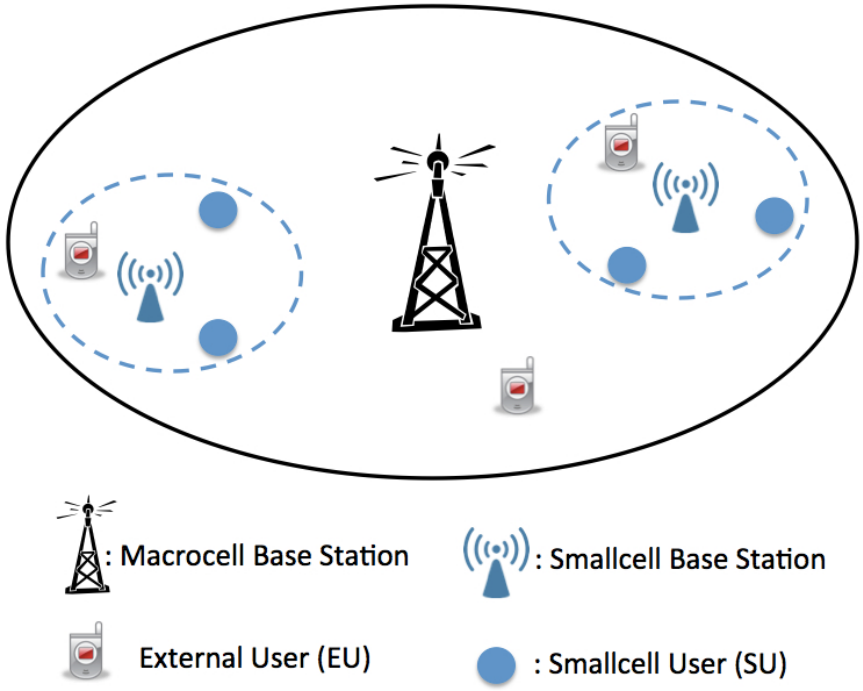

Figure 5: Considered network topology.

is considered as shown in Figure 5 with a small cell access point (SAP) serving $\mathcal{N}$ users $\{1, 2, ...N\}$. Out of this, $\mathcal{F}$ represents the set of all SUs $\{1, 2, ...F\}$ and $\mathcal{E}$ represents the set of all EUs $\{1, 2, ...E\}$, and therefore, $\mathcal{N} = \mathcal{F} \cup \mathcal{E}$. B represents the total system bandwidth consisting of K subcarriers. Hence, the bandwidth of each subcarrier is represented as $\Delta B = \frac{B}{K}$. The time is slotted and each slot has a duration of T_s equivalent to the coherent time of the channel.

The SNR gap [22] is defined as

$$SNR_{gap} = \frac{-1.5}{\ln(5 * BER)}, \tag{3}$$

where BER is a given Bit Error Rate for a user n to transmit on subcarrier k. The transmission rate $r_{n,k}$ for user n on subcarrier k is given as

$$r_{n,k} = \frac{B}{K} * \ log(1 - \frac{1.5 * |h_{n,k}|^2 * p_{n,k}}{\ln(5 * BER) * \sigma^2}), \tag{4}$$

where $h_{n,k}$ represents the channel gain of user n transmitting on subcarrier k. $p_{n,k}$ represents the required power for user n to transmit on subcarrier k for a given bit error rate (BER).

The noise power over a subcarrier is represented by σ^2. Each user can be assigned several subcarriers with the constraint that the same subcarrier cannot be assigned to different users in the same slot. This is represented by the binary variable $s_{n,k}(t)$ indicating whether the subcarrier k is assigned to user n or not in slot t. Hence, the subcarrier assignment constraint is given as $\sum_{n=1}^{N} s_{n,k}(t) = 1$. Therefore, the maximum achievable data rate to user n in slot t is given by the equation

$$\mu_n(t) = \sum_{k=1}^{K} s_{n,k}(t) r_{n,k}(t). \tag{5}$$

Inserting Eq.(4) into Eq:(5), we obtain the maximum achievable data rate per user as

$$\mu_n(t) = \sum_{k=1}^{K} s_{n,k}(t) * \frac{B}{K} * \ log(1 - \frac{1.5 * |h_{n,k}(t)|^2 * p_{n,k}(t)}{\ln(5 * BER) * \sigma^2}). \tag{6}$$

2.1.1.2 Hetergeneous Traffic Model

The SAP has queues corresponding to each of the n user types it serves. The arrival process $\Lambda_n(t)$ represents the number of packet arrivals at queue n in time t. Here, the mean arrival rate is given by $\lambda_n \triangleq E[\Lambda_n(t)]$ and the mean arrival rate vector is given as $\overrightarrow{\lambda} = (\lambda_1, \lambda_2, ...\lambda_N)$. $\overrightarrow{Q(t)} = (Q_1(t), Q_2(t), ...Q_N(t))$ represents the Queue Length vector. The Waiting Time of a packet in the queues is represented by the vector $\overrightarrow{W(t)} = (W_1(t), W_2(t), ...W_N(t))$. The queue evolves according to the Discrete-time Queueing Law as

$$Q_n(t+1) = max(Q_n(t) - \mu_n(t)T_s, 0) + \Lambda_n(t). \tag{7}$$

By Little's Law, the waiting time of user n in slot t is given by

$$W_n(t) = \frac{\overline{Q_n}(t)}{\lambda_n}, \tag{8}$$

where $\overline{Q_n}(t)$ is the average queue length.

The users served under SAP are grouped into three classes using three different queuing disciplines as in [23]. These classes are as follows:

- Constant Bit Rate (CBR) Users: These users have deterministic behaviors, and are modeled by a D/G/1 queuing system. The average waiting time is calculated as:

$$W_{CBR,n} = \frac{\lambda_{CBR,n}\sigma^2_{CBR,\overline{X_n}}}{2(1-\lambda_{CBR,n}\overline{X_n})}, \qquad n \in N, \tag{9}$$

where $\lambda_{CBR,n}$, $\sigma^2_{CBR,\overline{X_n}}$ and $\overline{X_n}$ are the mean arrival rate, the variance of the service time and the mean service time respectively and $\overline{X_n} = E[1/\mu_n]$ for OFDMA.

- Video-Streaming Users: These users are modeled using *Gamma Distribution* with shape parameter s and a G/G/1 queuing system where the average waiting time is

$$W_{Vid,n} = \frac{\lambda_{Vid,n}(\sigma^2_{Vid,\overline{X_n}} + s/\lambda_{Vid,n})}{2(1-\lambda_{Vid,n}\overline{X_n})}, \quad n \in N. \tag{10}$$

- Best-Efffort (BE) Users: The BE users can be modeled using an M/G/1 queueing system where the average queue waiting time is expressed as:

$$W_{BE,n} = \frac{\lambda_{BE,n}(\sigma^2_{BE,\overline{X_n}} + \sigma^2_T)}{2(1-\lambda_{BE,n}\overline{X_n})}, \qquad \forall\, n \in N. \tag{11}$$

where σ^2_T is the variance of the inter-arrival time and $\rho_{BE} = \overline{\lambda_{BE}}/\overline{X_m}$ is the utilization.

Our objective is to stabilize the queues of all SUs and EUs when the arrivals are inside the capacity region. In addition, we want to offer QoS performance for different SU traffic types in terms of maximizing throughput and minimizing delay. These objectives together present an interesting case of QoS provisioning. Scheduling policies such as proportionally fair (PF) scheduling, Modified Largest Weighted Delay First (M-LWDF) are not suitable in the presence of heterogeneous traffic since they do not provide bounded delay performance.

2.1.1.3 Proposed Admission Control Framework

The proposed admission control framework [8, 9] is embedded into each SAP and is illustrated in Figure 6. It has four main parts: (i) *QoS Classification of Heterogeneous Traffic*, (ii) *Computation of Utility Function*, (iii) *Traffic-aware Admission Control*, and (iv) *Power-constrained Subcarrier Allocation*.

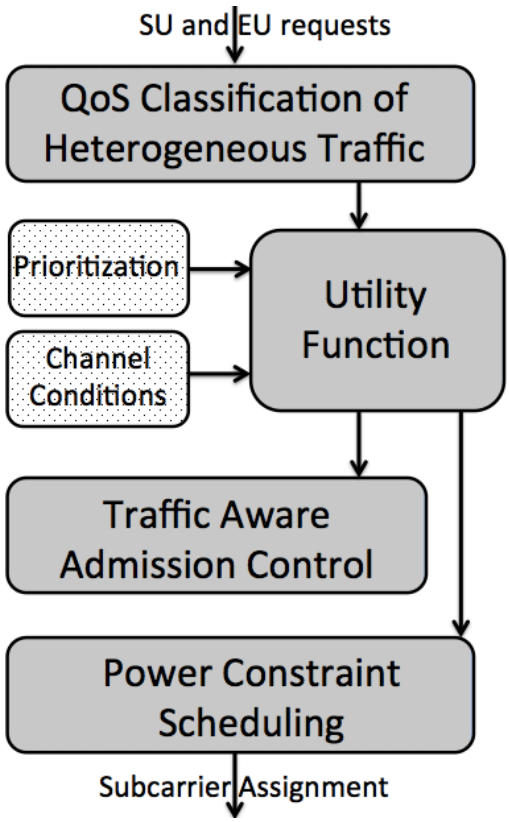

Figure 6: The proposed admission control framework.

The QoS classification of heterogeneous traffic part uses the SU and EU requests, as seen in Figure 6, to calculate the average waiting time of each user types. These calculations are aforementioned in the previous sections as BE, Video-Streaming and CBR traffic in Eqs.(9),(10) and (11). The need for achieving diverse QoS requirements for heterogeneous traffic classes calls for an improved scheduling rule that can deal with their unique attributes. Especially, under the hybrid small cell setup, the packet delay for SUs must be bounded in the presence of EUs. The authors in [24, 25] show that if the message size of users exhibit heavy-tail characteristics with an index α, then the delay has an infinite mean and infinite variance for $\alpha < 1$ and $\alpha < 2$ respectively. The authors also propose a modified maximum weight-α scheduling policy that allocates channels for users based on queue size raised to the power α in order to guarantee bounded delay mean and variance.

In this work, we propose a novel traffic-aware utility based scheduling policy (TA-Utility) for hybrid small cells in order to effectively provision QoS. The scheduler is fed

with the information of the channel state and the traffic information in order to make scheduling decisions at every time slot t based on the computation of the utility function. Our scheduling policy is not only weighted as in [26], but it also considers several heterogeneous traffic types which are modeled by different queuing disciplines. In the original weighted alpha scheduler [26], the only parameter is the weight of the waiting queue size, but in our model, we have also the different channel conditions (meaning different queue-server models) and different utilities. The three different utility functions for BE, Video-Streaming and CBR traffic are derived in the following paragraphs.

The utility function as seen in Figure 6 associated with the allocation of subcarrier k to user n is defined as

$$U_{n,k}(t) = \gamma_n W_n^{\alpha}(t) r_{n,k}(t), \tag{12}$$

where $\gamma_n = \dfrac{a_n}{\bar{r}_n}$. Here, a_n represents the *priority index* and can be tuned for SUs and EUs to achieve the required QoS for each user type. $\bar{r}_n$ represents the average transmission rate for user n over all subcarriers measured over a time window. α is the exponent of the average waiting time W_n for packets in queue n. This is the *traffic coefficient* that takes unique values for different traffic classes.

Finally, by inserting the transmission rate $r_{n,k}(t)$ obtained in Eq.(4) and each of the average waiting times $W_n(t)$ of Eqs.(9),(10),(11), into our utility function Eq.(12), we obtain the utility functions for heterogeneous user types for OFDMA small cells as follows:

The traffic-aware utility function for CBR users is given by

$$U_{n,k}^{CBR}(t) = \gamma_n * \frac{B}{K} * \left(\frac{\lambda_{CBR,n}\sigma^2_{CBR,\overline{X_n}}}{2(1-\lambda_{CBR,n}\overline{X_n})}\right)^{\alpha} * log\left(1 - \frac{1.5 * |h_{n,k}|^2 * p_{n,k}}{\ln(5 * BER) * \sigma^2}\right).$$

The traffic-aware utility function for Video-Streaming users is given by

$$U_{n,k}^{vid}(t) = \gamma_n * \frac{B}{K} * \left(\frac{\lambda_{Vid,n}(\sigma^2_{Vid,\overline{X_n}} + s/\lambda_{Vid,n})}{2(1-\lambda_{Vid,n}\overline{X_n})}\right)^{\alpha} * \ log\left(1 - \frac{1.5 * |h_{n,k}|^2 * p_{n,k}}{\ln(5 * BER) * \sigma^2}\right).$$

Similarly, the traffic-aware utility function for BE users is given by

$$U_{n,k}^{BE}(t) = \gamma_n * \frac{B}{K} * \left(\frac{\lambda_{BE,n}(\sigma^2_{BE,\overline{X_n}} + \sigma_T^2)}{2(1-\lambda_{BE,n}\overline{X_n})}\right)^{\alpha} * \ log\left(1 - \frac{1.5 * |h_{n,k}|^2 * p_{n,k}}{\ln(5 * BER) * \sigma^2}\right).$$

Choice of Utility Function: The utility function defined in Equation (12) is aimed at achieving the heterogeneous objectives of QoS perceived by different user types. In Equation (12), the utility function is proportional to the waiting time of user n's packet raised to the power α. This implies that as the waiting time of the packet of a user becomes large, the QoS requirement of that user is high. Hence, this user has a high priority during admission control and subcarrier assignment.

The parameter α is specified in the exponent to enforce QoS differentiation between services. For the real-time users, the delay performance is critical and they have a strict deadline on the waiting time of the packet. For Constant Bit Rate (CBR) user types, the throughput as well as delay performance are important. Since different users have varying degrees of delay bounds, varying the values of α can impact the strictness of the QoS requirement. In other words, a larger value of α (considering $W > 1$ units) specifies that the queue needs to be served more urgently. It can be observed that by setting the value of α to 1 for all traffic types, the utility function shows similarity to the M-LWDF rule. The relationship between the choice of α and the stability conditions will be discussed later in this section.

In addition to considering packet delay, Equation (12) also captures the transmission rate for user n to transmit on subcarrier k. This enables users that have better channel quality than other users to have higher priority during admission control and subcarrier assignment. In the γ_n parameter, we use the average data rate for user n over all subcarriers. Therefore, $U_{n,k}(t)$ is not simply a function of the instantaneous channel quality but the average channel quality. In order to provide fair allocation for SUs in the presence of bursty EU traffic, the parameter a_n can be set to a higher value for SUs. In other words, a_n is used as a bias factor to increase the priority of the SUs compared to the EUs. In this work, the choice of the priority index a_n we utilize for the simulations is based on the results provided in [27] where $a_n = \frac{-log\,\delta_n}{T_n}$. $\delta_n = Prob\{W_n > T_n\}$ indicates the maximum delay violation probability and T_n is the delay threshold for user n. In order to differentiate between EUs and SUs, we

utilize a higher value of δ_n for the EUs compared to the SUs. The value of T_n depends on the traffic type.

Discussion on Stability and Delay Bounds: The authors in [26] prove that the mean of the queue length under steady-state becomes infinite when scheduling policies such as Maximum Weight Scheduling are utilized where the tail index of the arrival process are not considered in the scheduling decisions considering that at least one of the arrival process follows a heavy-tailed distribution (where the tail coefficient ≤ 2). Under the presence of heavy-tailed distribution, the relation between α parameter used for a maximum weight-α is given in [26] such that if $\alpha_{heavy} + 1 < C_{heavy}$ and $\alpha_{light} + 1 < C_{light}$, then the system of queues containing such a mix of heavy and light-tailed arrivals is stable. Here, C is the tail coefficient. In such a case, it is also shown that the $\bar{Q}_{tail}^{\alpha_{tail}} < \infty$ where $\bar{Q}_{tail}$ is the steady-state mean queue length of queue type *tail* such that $tail \in \{heavy, light\}$.

In our traffic-aware utility based approach, we propose that the utility function depends on the waiting time of the packet raised to the power of α. Using Little's law, the relationship between the average queue size $\bar{Q}_{tail}$ and the (average) waiting time W_{tail} is given as $W_{tail} = \frac{\bar{Q}_{tail}}{\lambda_{tail}}$. Hence with the proper choices of α for the different traffic types, it can be deduced that the bounded queue size and hence bounded average waiting time can be achieved under the proposed utility-based scheduling approach. Such an approach requires the knowledge of the α values of different traffic arrival processes. The mathematical proof for bounded average waiting time for the proposed approach is left for future work.

2.1.1.4 Power Constrained Utility based Scheduling:

The subcarrier allocation with power constraints using the proposed utility function is performed based on the following optimization objective:

$$\max_{S} \quad \sum_{n=1}^{N} \sum_{k=1}^{K} U_{n,k}(t)\, s_{n,k}(t), \tag{13}$$

$$\text{subject to} \quad \sum_{n=1}^{N} s_{n,k}(t) = 1 \ ; \tag{14}$$

$$s_{n,k}(t) \in \{0, 1\} \ ; \tag{15}$$

$$\sum_{n=1}^{N} p_{n,k} \ s_{n,k}(t) \leq P_s \ ; \tag{16}$$

$$\sum_{n=1}^{N} \sum_{k=1}^{K} p_{n,k} \ s_{n,k}(t) \leq P_{tot}, \tag{17}$$

where the optimization variable S is the subcarrier allocation matrix with order $N \times K$. $S_i = \{s_{i,1}, s_{i,2}, ..., s_{i,K}\}$ is the set of subcarriers allocated to node i, P_s is the maximum allowed subcarrier power and P_{tot} is the total transmission power available at the SAP.

2.1.1.5 Minimal Algorithm for Utility-based Subcarrier Assignment

The above optimization problem can be classified into the Multiple Choice Knapsack Problem (MCKP) with additional constraints on the maximum weight of each item. The MCKP is defined as a binary knapsack problem with additional disjoint multiple-choice constraints [28]. The constraints are such that the items are divided into multiple classes and only one item is to be selected from each of the classes. The MCKP has been shown to be NP-hard since the KP problem needs to be solved in the process, nevertheless, through dynamic programming it is shown to be solved in pseudo-polynomial time [29]. A minimal algorithm for solving MCKP is presented in [30]. First, the integrality constraint $s_{n,k}(t) \in \{0, 1\}$ is relaxed to $0 \leq s_{n,k}(t) \leq 1$ to obtain the Linear Multiple-Choice Knapsack Problem (LMCKP). A simple partitioning algorithm is proposed for solving the LMCKP and obtaining a feasible solution. Using the initial solution, dynamic programming is used to solve MCKP. The partitioning algorithm can compute in $O(n)$ time, a small subset of items called as the core of classes, to be considered for the optimal value. New classes are then added to the core by need.

Applied to the utility-based subcarrier assignment problem, the classes correspond to the set of subcarriers $\mathcal{K}$. Each item corresponds to a node n to be assigned for a subcarrier k. In our problem, we have an additional constraint in the form of maximum per-subcarrier power. Only one node among N nodes is assigned for subcarrier k given that it satisfies the global and local power constraints. $U_{n,k}(t)$ indicates the profit obtained when $p_{n,k}(t)$ is the power allocated. The output of the algorithm is the matrix S of dimension $N \times K$ with assignment indicators $s_{n,k}(t)$. The algorithm is presented in Algorithm 1.

The procedure partitionalgo() provides an LP-optimal solution for the relaxed LMCKP problem. The function reduceclass() uses upper-bound computation, dominance tests to prune nodes for each subcarrier while reduceset() checks and updates the *Current Best Solution* if a state improves the lower-bound. The computational complexity of the one-dimensional MCKP is shown to be $O(n + P_{tot} \sum_{R_k \in C} num_k)$ where R_k is the set of subcarriers in the core C and num_k is the number of users considered for each subcarrier in the core. When the number of users considered for the core is small, the algorithm obtains the optimal solution in linear time. For large cores, the optimal solution is obtained in time complexity is pseudo-polynomial. When the number of users is not considerably large, it can also be shown that performing adaptive modulation combined with subcarrier assignment does not increase the algorithm complexity significantly. This is a reasonable assumption since small cells, on average, support few tens of users. Therefore, our optimization framework can also be extended to account for different modulation schemes as done in 4G and beyond systems.

2.1.1.6 Traffic-aware Utility based Admission Control

Admission Control algorithms focus on balancing the load on the system based on a set of rules. It is possible that in a time-varying multi-carrier system, some users experience poor channel conditions for a significant amount of time over all channels. As a result, the scheduler needs to allocate a large amount of resources to these users. The rest of the users experience a significant drop in the data rate achieved. Furthermore, their Head of Line

input : Utility matrix U of dimension $N \times K$, Required Power matrix p of dimension $N \times K$

$\{a, b_k, s_{b_a,a}, s_{b'_a,a}\} \Leftarrow$ `partitionalgo`(U, p, P_{tot}); where $a :=$ fractional subcarrier, $\{b_k\} :=$ LP Optimal soln., $s_{b_a,a}, s_{b'_a,a} :=$ fractional variables in a;

Calculate $\lambda = (U_{b'_a,a} - U_{b_a,a})/(p_{b'_a,a} - p_{b_a,a})$;

Calculate $\lambda_k^+ = \max_{n \in N, p_{n,k} > p_{b_k,k}} (U_{n,k} - U_{b_k,k})/(p_{n,k} - p_{b_k,k})$, $k = 1, ..., K$, $k \neq a$;

Calculate $\lambda_k^- = \min_{n \in N, p_{n,k} < p_{b_k,k}} (U_{b_k,k} - U_{n,k})/(p_{b_k,k} - p_{n,k})$, $k = 1, ..., K$, $k \neq a$;

Gradients $L^+ = \{\lambda_k^+\}$ and $L^- = \{\lambda_k^-\}$ for $k = 1, ..., K, k \neq a$;

`sortascen`(L^+);

`sortdescen`(L^-);

Current Best Solution $z := 0$; Initial Core $C := N_a$;

Current Set of States $Y_C :=$ `reduceclass`(N_a);

Vectors in Y_C represented by states (θ_k, π_k, ν_k); where $\theta_k := \sum_{k \in C} p_{y_k,k} + \sum_{k \notin C} p_{b_k,k}$, $\pi_k := \sum_{k \in C} U_{y_k,k} + \sum_{k \notin C} U_{b_k,k}$, $\nu_k :=$ partial representation of vector $\vec{y_i}$;

repeat

  `reduceset`(Y_C);

  if $(Y_C = \emptyset)$ **then** break;

  Choose *nextsubcarrier* k from L^+ s.t. $R_k :=$ `reduceclass`(k)

  if $|R_k| > 1$ **then** `addclass`(Y_C, R_k);

  Repeat steps 14 - 15 for L^-;

  `reduceset`(Y_C);

  if $(Y_C = \emptyset)$ **then** break;

until *forever*;

Find optimal allocation S;

Algorithm 1: Minimal algorithm for TA-Utility based subcarrier assignment.

(HoL) packet delay which is the delay experienced by the packet at the head of the queue can potentially become unbounded.

Similarly, when a new user requests for resources with the SAP, the admission control procedure must evaluate if scheduling can be performed for the new user without affecting the QoS of the existing users. If the new user is a high-priority user, the algorithm must decide which among the existing connections need to be released or be assigned lesser resources. Due to this dependency between the subcarrier assignment and the admission control procedures, we advocate the joint functioning of scheduling and admission control procedures. The objective of the admission control procedure, therefore, must be to *admit as many users as possible while preserving the feasibility of the scheduler*.

We specify the rules for performing admission control in a hybrid small cell with heterogeneous user traffic. There are two types of priorities: one for the user type, the other for the traffic type. The SUs have a higher priority over the EUs. The traffic types with their decreasing priorities include *CBR*, *Video Conferencing* and *Best Effort*. When a new user with a certain traffic type is requesting for resources, the admission control procedure has broadly three choices:

1. The user is admitted without affecting the QoS of the rest of the users *(or)*
2. The user is admitted provided there is a lower priority user session that can be released or its resources reduced *(or)*
3. The user is not admitted.

The utility function used for the scheduling problem in Eq.(12) is such that a low utility value for a user's session at a given time instant implies that the user has one or more of the following characteristics:

- Deep channel fade for the duration.
- Lower priority (being a EU).
- Lower priority traffic type (Best Effort).
- Lesser packets waiting in the queue.

From the above, it can be observed that the utility function can provide a sense of how the user is performing in a given time slot. At the same time, the instantaneous utility can be misleading because the same user with lesser utility in a given slot may have a greater utility due to better channel conditions in the next slot. Therefore, we define a new performance metric that will be used in making admission control decisions. First, based on the outcome of the *Traffic-aware Utility based subcarrier assignment* procedure, the utility of each user over the assigned subcarriers is computed.

$$U_n(t) = \sum_{k=1}^{K} U_{n,k}(t)\, s_{n,k}(t). \tag{18}$$

Then, we define a *normalized utility* parameter for each user given by

$$\hat{U}_n(t) = \frac{U_n(t)}{U_{max}(t)}, \tag{19}$$

where $U_{max}(t) = \max_n \{U_n(t)\}$. We also define the time-average normalized utility over T slots as

$$E[\hat{U}_n(t)] = \frac{1}{T} \sum_{\tau=t-T}^{t} \hat{U}_n(\tau). \tag{20}$$

The time-average normalized utility provides a long-term view of how the user session is performing. A low value of $E[\hat{U}_n(t)]$ indicates that the user is most likely the candidate to be released as the scheduler is not able to meet the QoS requirements of the user. We represent the utility threshold as U_{th} and is defined in the region [0,1]. The normalized utility is used since the utilities can vary widely over different slots and it is necessary to have a relative value that can be compared with U_{th}. It can be argued that the ratio between the average normalized utilities of different users and the ratio of the average absolute utilities are fairly comparable. Based on this rationale, we utilize the $E[\hat{U}_n(t)]$ to provide a long-term view of the QoS performance. The admission control algorithm is presented in Algorithm 2.

2.1.1.7 Performance Evaluation

The performance of the constrained utility-based scheduling is evaluated where the utility function is modelled based on OFDMA system parameters and queue models using MATLAB. The minimal algorithm routine implemented in C is used through the MATLAB mex file to perform subcarrier allocation decisions.

We simulate the downlink of a small cell with the SAP serving N users. The users distances d_n are modelled randomly with a mean of 10 meters and a variance of 4 meters.

```
for each time slot t do
    for each user n do
        Determine E[Û_n(t)] based on the subcarrier allocation algorithm
        if E[Û_n(t)] < U_th then
            Preempt user n
        end
    end
    if a new user m requests for resources then
        Compute Û_m(t)
        if Û_m(t) < U_th then
            Do not admit the user
        else
            Admit user m and schedule subcarriers according to U_m(t)
        end
    else
        continue
    end
end
```

Algorithm 2: Traffic-aware Utility based Admission Control Algorithm

The path-loss L_n^p for all users are modelled based on the following indoor propagation model,

$$L_n^p = 37 + 30\,log(d_n) + L_n^{wall}, \tag{21}$$

where L_n^{wall} is the penetration loss due to walls for user n and is also modelled randomly. The noise power is obtained from the equation

$$p^{noise} = -174 + 10\,log(\Delta B) + f, \tag{22}$$

where f is the noise figure in dB. All users experience log normal shadowing L_n^{shadow} with a mean of 10dB. The signal to noise (SNR) distribution φ_n for unit power is obtained using the equation

$$\varphi_n = 10^{L_n^p - L_n^{shadow} + p^{noise}}. \tag{23}$$

The SNR distribution is obtained as $\Phi_{n,k} = \varphi_n\, p_{n,k}$. The system parameters considered are shown in Table 3. For the computation of throughput, we have 5 users from each of the

Table 3: Simulation parameters.

Parameter	Value	Parameter	Value
System Bandwidth	1.92MHz	Pathloss Model	Indoor
Number of Subcarriers	64	User distances	Rand., $\mu = 10m$, $\sigma^2 = 4m$
Subcarrier Bandwidth	30KHz	Shadowing	Mean 10dB
BER Required	10^{-3}	λ_{CBR}	75 – 125kbps
Max. SAP Tx. Power	$1W$	λ_{Video}	200 – 256kbps
Max. subcarrier Tx. Power	$0.05W$	λ_{BE}	150 – 180kbps
Total Number of Users	15	Shape parameter	3.066
Slot Length	10ms	Simulation time	5000ms

traffic classes described in the previous section with a mix of SUs and EUs. For this purpose, the arrival rates of CBR users are randomly distributed with range 75 – 125kbps. The Video-Streaming users have arrival rates randomly distributed within range 200 – 256kbps. The BE users have randomly distributed arrival rates within range 150 – 180kbps. The shape parameter for Video-Streaming user is fixed at 3.066. The maximum delay allowed for CBR and Video-Streaming users are fixed at 80ms, 150ms respectively.

In each slot, the utility and power matrices along with the global and local power constraints are computed and fed as input to the algorithm. The output of the algorithm is the subcarrier assignment matrix S. The entire simulation sequence is run for 5000ms which is used to compute the performance metrics. We highlight for this section, the performance results of our scheduling policy in terms of throughput, fairness and delay. The results also highlight how the performance of the TA-Utility scheme can be enhanced with the help of TA-Utility based Admission Control scheme.

Figure 7 shows the time average throughput achieved by the users. As seen in the figure, users 1 to 5 are the CBR users with uniform arrival rate in the range $[75 - 125]kbps$. These users achieve a time-average throughput of approximately $100kbps$. The priority index a_n of the users are set so that the utility of the SUs are higher than that of the EUs and hence the SUs being able to achieve superior throughput. In the simulations, we have utilized $a_s = 2a_e = 0.2$ indicating that the probability of violation $\delta_s = 4\delta_e$, where the index s and e correspond to the small cell user and external user respectively. The delay threshold T_n is

set to the maximum delay values defined earlier based on the traffic type.

In addition to the throughput measurements, the mean of the queueing delay over the simulation time is computed for different arrival rates for each traffic class. The simulation parameters from Table 3 are retained. The mean delay of the worst-case user of each of the traffic classes is plotted for different arrival rates. Two variants of the proposed schemes are implemented. Under Scheme 1, only the TA-Utility scheme is implemented while under Scheme 2, the TA-Utility scheme is enhanced using the proposed admission control procedure. For Scheme 2, the simulation starts with 15 users and then new users of different traffic classes are introduced randomly in the interval [0, 2000] milliseconds. User sessions are released based on the admission control algorithm and for a given fixed utility threshold $U_{th} = 0.05$. Both these results obtained for the mean delay are compared with the baseline M-LWDF scheme for all the three traffic types.

As observed in Figure 8, the mean delay of the worst-case CBR user for both Scheme 1 and Scheme 2 remains within $50ms$ for arrivals below $300kbps$ compared to a mean delay of $80ms$ under M-LWDF scheme. For arrivals beyond $300kbps$, the M-LWDF scheme has the mean delay increasing almost linearly whereas Scheme 1 and Scheme 2 result in a slowly increasing mean delay with the Scheme 2 resulting in up to 50ms less delay than Scheme 1 at arrival rates beyond 650kbps. For the case of video users shown in Figure 9, Scheme 1 has comparable mean queueing delay to M-LWDF for all arrivals below $350kbps$. Beyond $350kbps$, the mean delay is increasing slower than the M-LWDF scheme. Scheme 2 results in further improvement from Scheme 1 with the maximum mean delay reaching up to $190ms$. Therefore, the TA-utility scheduling offers significant delay performance gains for delay-sensitive traffic classes. At the same time, the mean delay for Best Effort users, as shown in Figure 10, is much higher and reaches up to $1s$ for both Scheme 1 and Scheme 2 when the arrival rate is $600kbps$ and above. This is still acceptable given the presence of CBR and $Video$ users with strict QoS requirements.

In addition to mean delay computation, it is also necessary to analyze the delay variance

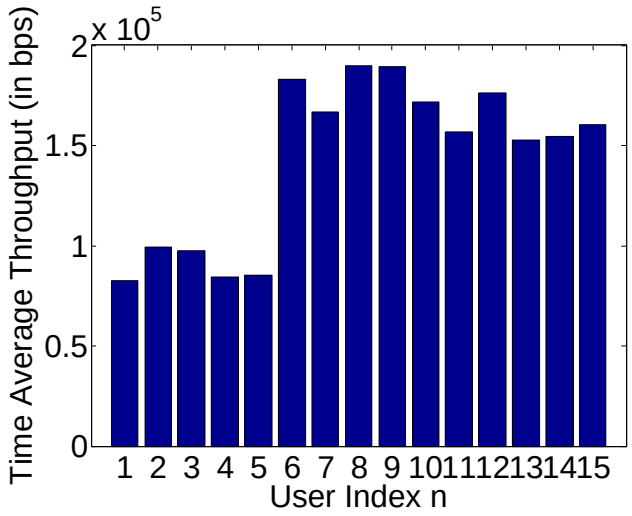

Figure 7: Throughput performance under traffic-aware utility based scheduling.

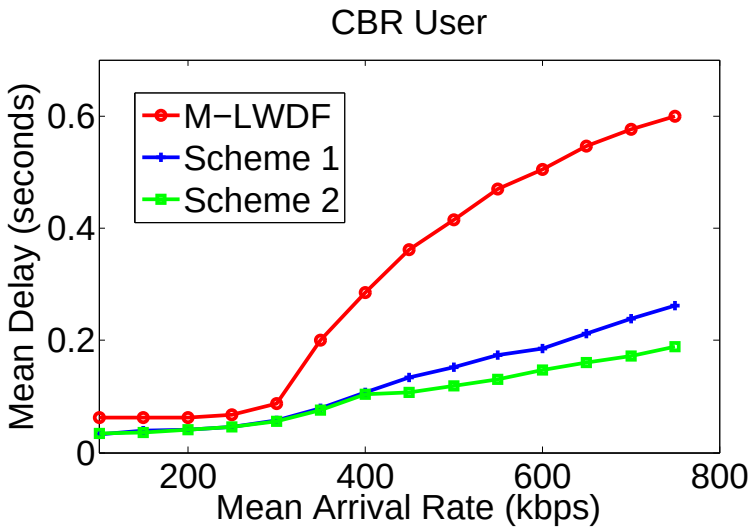

Figure 8: Mean delay performance of CBR traffic class.

performance to determine the fairness of the proposed scheme for different traffic classes. The variance of the delay experienced by the users of a given traffic class is computed and plotted against the mean arrival rates of the users' traffic. The results obtained for Scheme 1 and Scheme 2 are compared with M-LWDF. Figure 11 shows the delay variance of CBR users for different arrival rates. Low values of delay variance for the proposed schemes compared to M-LWDF scheme indicates that a higher degree of fairness is achieved for different CBR users. Similarly, Figure 12 shows that the delay variance of video traffic under both Scheme 1 and Scheme 2 are marginally lower compared to M-LWDF scheme

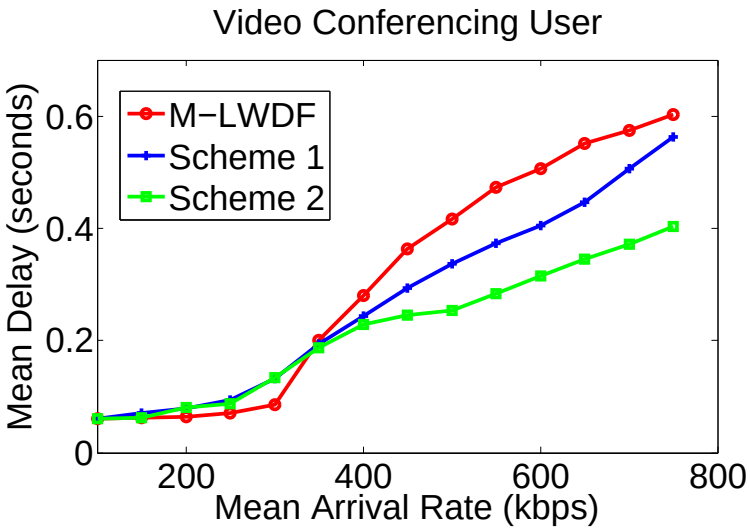

Figure 9: Mean delay performance of video traffic class.

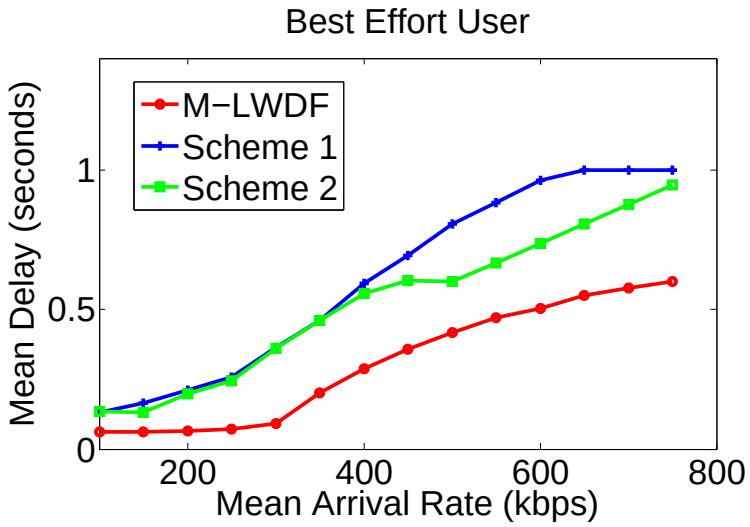

Figure 10: Mean delay performance of BE traffic class.

showing good fairness performance. In Figure 13, the delay variance is significantly higher for low data rates for the proposed schemes, especially for Scheme 1, but as the arrival rates of BE users increases, the delay variance converges towards that of the M-LWDF scheme.

2.1.1.8 Conclusions

In this work, the problem of admission control and QoS provisioning in hybrid small cells with SUs and EUs has been considered. The need for an improved admission control accounting for users' unique access control features to provide stability and QoS performance

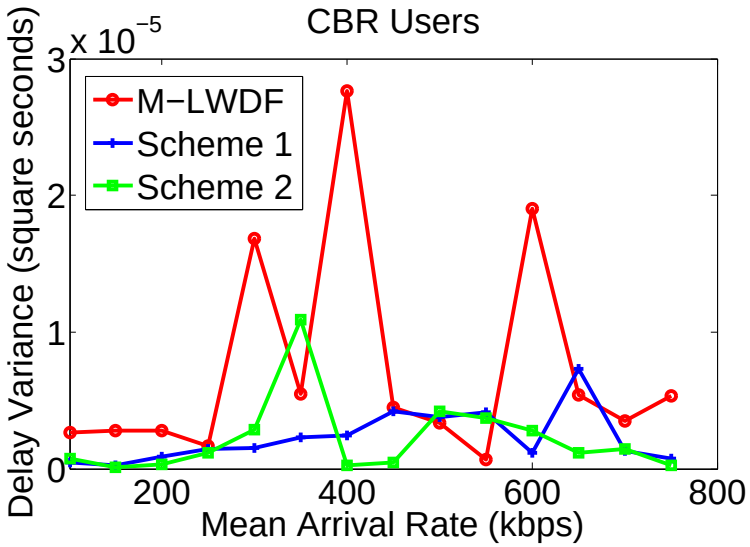

Figure 11: Delay variance performance of CBR traffic class.

in the presence of heterogeneous traffic has been explained and a novel traffic-aware utility function aimed at this problem has been proposed. To this end, a traffic-aware utility maximization approach under power constraints has been proposed and is posed as an optimization objective. In order to obtain the optimal solution, a minimal algorithm that results in a minimal core of allocation vectors is presented. In order to handle the incoming users efficiently, an admission control algorithm based on the proposed traffic-aware utility function is proposed. In the end, the performance of the proposed scheme is illustrated using simulations.

2.2 Handover Execution and Completion

2.2.1 Local Anchor based Handover for Clustered Small Cells

Seamless handover is a major challenge facing the large-scale adoption of the HetNet architecture. For this reason, a 3GPP work item for HetNet mobility improvements in LTE has been laid out in [31]. The 3GPP technical report in [32] provides a summary of the mobility performance study and the proposed mobility enhancements under HetNets. Under this, it is shown that the handover performance for users deteriorates with increase in small

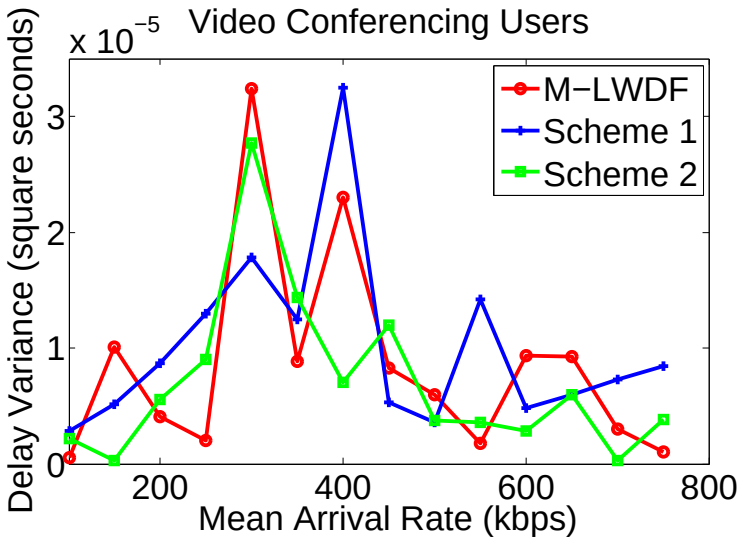

Figure 12: Delay variance performance of video traffic class.

cell density. This is mainly owing to the numerous cell edges resulting from the HetNet architecture. However, a dense deployment of small cells, especially in indoor environments, is deemed essential in order to achieve superior capacity and ubiquitous coverage.

In the case of large-scale small cell deployments such as in airports, malls, large office buildings or auditoriums, coordination among small cells is ideal to achieve optimal performance. In fact, recent standardization and industry efforts target enterprise small cells for the above scenarios [33] where coordination plays a key role. This coordination enables to achieve improved mobility management, interference management as well as self-organizing (SON) functions by utilizing the underlying network infrastructure [34].

Specifically for handover management, coordination plays a significant role in the following ways. First, it can facilitate scalable small cell deployments by potentially minimizing the load on the core network during handovers. Second, since the small cells incorporate different backhaul technologies to connect to the operator's network, coordination can help overcome potential backhaul issues of long latency and operating costs involved in signaling and data forwarding during handovers. However, the use of coordinated small cells also place some constraints such as requiring a network infrastructure with high-speed links between small cells.

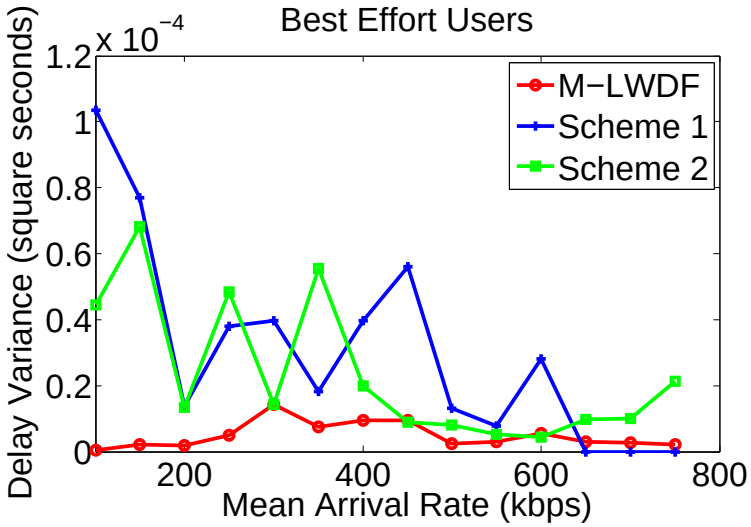

Figure 13: Delay variance performance of BE traffic class.

Handover management in HetNets has been well studied in recent years with a number of solutions aimed to enhance mobility performance. Intelligent handover decision algorithms have been proposed in [35, 36, 37, 38, 39] that, in general, aim to minimize the number of handovers and handover failures in the network. However, when the number of cell crossings are frequent, it is also necessary to minimize the impact of the handover on the network and user. This includes the cost incurred during a handover in the form of signaling and data forwarding costs as well as handover interruption time for the user.

Analogous to the earlier studies of fast handovers in [40], a fast handover scheme is proposed in [41] where the higher layer data is buffered to all the small cells in the proximity of the current cell. Although these approaches reduce the handover interruption time, the signaling load at the core network (CN) is still large which can negatively impact the large scale deployment of small cells. In [42], it is proposed to move the mobility anchor point from the Mobility Management Entity (MME) to the femtocell gateway. Such an approach does not minimize the interruption time or the handover-related costs since the femtocell gateway is also located at the core network.

In [43], a distributed implementation of the Mobility Management Entity (MME),

which is the anchor for handover signaling is proposed. Nevertheless, this type of architecture can raise major issues of security, failure handling and synchronization. For instance, upon an incoming connection request for a mobile user, the MME must be aware of the user contexts including location information, associated base station, etc. If the MME is implemented in a distributed fashion, synchronization is required across other distributed entities in order for the MU related context information to permeate across the network. A lack of synchronization or delay can result in connection failure or failed detection. In terms of security management, all the security keys for the small cell and users are either generated or derived from the MME key using the global information. We argue that a distributed architecture for key management on the one hand requires coordination among the distributed entities in order to promise uniqueness. On the other hand, a revisit to the security handling procedures would be required in order to facilitate this distributed architecture.

In [44], the handover signaling costs are compared for "direct X2" based and "X2-Gateway" based approaches where the direct X2 interface based handover scheme shows significant reduction in core network load and signaling cost. Nevertheless, these schemes involve "path switching" procedure with the core network using the internet-based backhaul causing additional signaling load on the core network. The path switching procedure is a key factor in determining the downlink handover interruption time and delay jitter especially when the number of in-transit packets forwarded from source base station's buffer is not large [45].

Local mobility anchoring is a promising way to achieve overall mobility performance improvement. Local anchor-based mobility management schemes were studied in [46] for optimizing paging and registration updates. Similary, cellular IP was proposed in [47] where some of the mobility management functions are moved to the base stations. In [48], new architectures were proposed to move the mobility anchor point closer to the base stations. However, the proposed approach requires redefining the security key mechanisms

and signaling flow among other major changes. In order to overcome the delay due to path switching, a local anchor based handover was proposed in [49] using X2 data forwarding analogous to the pointer forwarding technique originally proposed in [50]. In addition to increasing the link utilization on the neighbour cells, it is not clear how the intermediate small cells will participate in establishing and maintaining the X2 forwarding chain to enable data forwarding for the user after multiple handovers.

In this work, we consider the case of coordinated small cells and propose a novel local anchor-based architecture to improve the overall handover performance [51, 52]. First, we propose novel handover mechanisms using a local anchor-based architecture for coordinated small cells. Following this, we develop analytical models for a cluster of small cells to study the handover performance. Third, we provide closed-form expressions for handover performance parameters under the proposed handover schemes. Finally, we present numerical and simulation results highlighting the performance gains of the proposed handover schemes.

2.2.1.1 Local Anchor-based Handover Management

We first motivate the need for a new handover architecture for coordinated small cells and then describe our proposed local anchor-based handover architecture.

Motivation: The handover procedure for 3GPP LTE-A systems utilizing the direct interface (X2 interface) between small cells (SCs) is illustrated in Figure 4 [7]. The key observation from this X2-based handover procedure is that the mobility anchor for the handover is the core network (particularly, the MME). Therefore, intuitively, it can be argued that the backhaul for the small cells must have the key characteristics of low-latency and high reliability to achieve better handover performance. This is evident from the handover interruption time computation [45]. The total downlink interruption due to handover τ is determined as

$$\tau = \max(T_r, T_p), \tag{24}$$

where T_r is the time required for the MU to establish radio connection with the target SC where $T_r \approx 18.5ms$ [45]. T_p is the time required by the target SC to perform path switch and depends on the backhaul latency. Therefore, to obtain low handover interruption time, the backhaul latency for the small cell must not exceed few tens of milliseconds. However, the nature of the small cell deployments makes it difficult to achieve this stringent backhaul latency requirements. Therefore, we utilize local mobility anchoring to satisfy the joint objectives of minimizing total handover costs, handover interruption time and core network load.

Local Anchor-based Architecture: To achieve the above core objectives, we propose a local anchor-based (LA-based) architecture for handover in coordinated small cells. This is shown in Figure 14. A large array of small cells in a hotspot area is divided into several clusters where each cluster contains a subset of small cells. The cluster is formed based on a group of neighboring small cells which can form a local network. In the local network, we consider that there is at least one small cell that maintains a link with each of the small cells in the cluster. This small cell will be referred to as the local anchor (LA). Coordination within the cluster is enabled using the local network.

The local anchor is networked to the IP gateway of the local network which interfaces with the core network through a firewall and public internet which is one of the commonly adopted backhaul solutions for small cells. The major benefit of the LA-based architecture is that inter-small cell handover mechanisms can be proposed which utilize the **LA as the mobility anchor** therefore minimizing the handover interruption as well as the handover-related costs.

The role of the local anchor is described as follows:

- *Concentrator of traffic between SCs and CN*: The LA performs proxy or anchor function for the data and signaling traffic flows between SCs in its cluster and the CN. This includes proxy function for the S1-AP tunnel, which is a per-user signaling connection between an SC and CN, as well as proxy for the GPRS tunnelling protocol

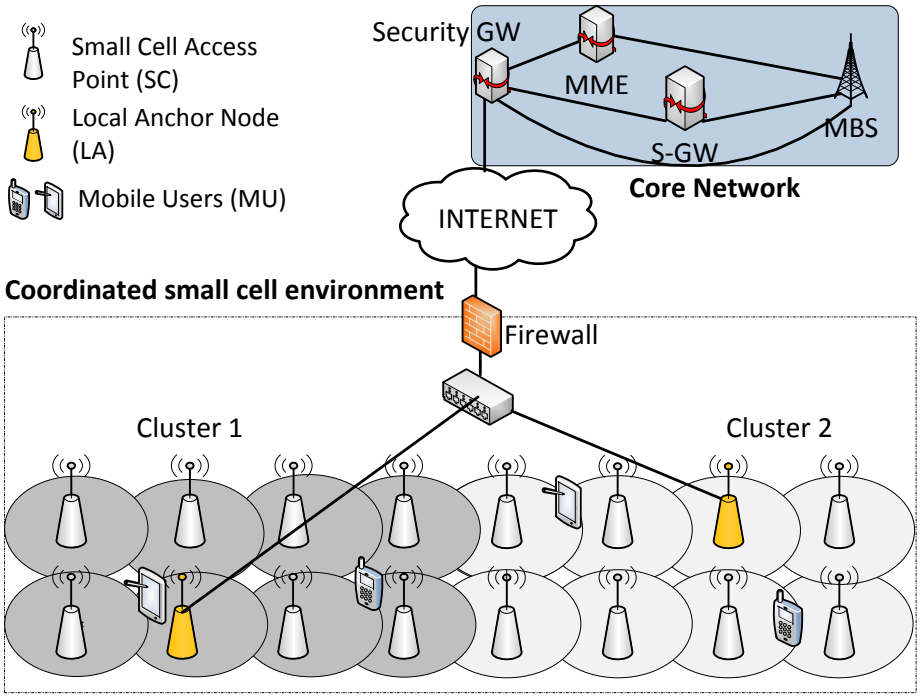

Figure 14: Local anchor based handover architecture.

(GTP), which is a per-user data connection between an SC and CN. Such a proxy function is analogous to the approach utilized in LTE-A relays. To achieve this proxy function, the LA utilizes a local anchor registration table **(LART)** as illustrated in Figure 15. The LART maintains a mapping of the data and signaling plane end point addresses for SC ⇔ LA and LA ⇔ CN links corresponding to each MU session indicated by EPS Bearer ID. The proxy function is made possible by maintaining S1 security over two hops between CN ⇔ LA and LA ⇔ SC links. Furthermore, to support handover to cells outside the cluster including macrocell base stations (MBSs), the LA also supports proxy X2 functions. However, we focus our study on intra-cluster handovers in this work.

- *Local Mobility Anchor for handover between other SCs*: The LA acts as a local mobility anchor for the users performing handover between SCs within its cluster. The LA provides means to avoid the initiation of path switch procedure whenever

EPS Bearer ID	SC Info			CN Info		
	SC S1AP ID	Transport Layer Address	GTP TEID of SC	CN S1AP ID	Transport Layer Address	GTP TEID of CN

Figure 15: Local anchor registration table.

possible without affecting the established procedures.

Using the local anchor based-architecture, the proposed handover mechanisms are described in the next subsections.

Drawbacks of existing handover mechanisms for small cells: In the existing LTE-A handover mechanism, the destination SC sends a path switch request message to the core network to initiate the switching of the downlink path for the MU towards the target SC. Besides the switching of the downlink path, the path switch request message also initiates the generation of new vertical keys for the target SC $\Longleftrightarrow$ MU interface. The newly generated key also enables forward key separation, i.e., once the target SC receives the new key (through vertical key derivation) in path switch request ack from the MME, the previous SC cannot decipher the data from the target SC. A detailed explanation of the key management is provided in [53]. In essence, if the path switching is not performed, then the target SC will continue to use the key derived by the previous SC (horizontal key derivation).

However, we argue that this raises several issues. Due to the large backhaul latency and frequent handovers in small cells, the **path switch request ack** from the MME is not always received in time at the new small cell. *On the one hand*, it can be argued that the 2-hop forward key separation is difficult to achieve anyway for small cells that experience poor backhaul conditions since the user can undergo several handovers before the new small cell receives the necessary information to generate a new vertical key. *On the other hand*, the high likelihood of a delay or in most cases failure in the completion of the path-switching

EPS Bearer ID	Original SC Info			New SC Info			CN Info		
	Original SC S1AP ID	Transport Layer Address	GTP TEID of original SC	New SC S1AP ID	Transport Layer Address	GTP TEID of new SC	CN S1AP ID	Transport Layer Address	GTP TEID of CN

Figure 16: Local anchor registration table update after handover.

mechanism also results in several handover failures since the downlink path is not switched to the new small cell.

Local Path Switching based handover: In this work, we propose a local path switching based (LP-based) handover mechanism for SCs belonging to the same cluster for up to ν number of handovers for a given MU. Due to our proposed local-anchor based architecture, the network still perceives that the user is still attached to the local anchor. Therefore, the handover process can be completed without experiencing handover failures or delay jitters.

To achieve this, the LA maintains a local counter per MU $Count_{MU}$ for the session duration. Path switching is performed with the CN only when $Count_{MU} = n\nu$ where n is a positive integer. This means that, in the *handover completion* phase, the path switch message from the target SC is not forwarded to the CN by the LA. Instead, a new S1 path is created between the target SC and the LA. To support this, the MU session in the LART is updated with the new small cell endpoint address as shown in Figure 16 until the next handover takes place. As a result of this approach, all future downlink data and signaling for the MU is forwarded from the LA to the target SC. For subsequent handovers, the new target SC information will be updated in the LART. Once the target SC receives the "end marker" from the serving SC over the X2 interface, the new DL path between LA and target SC will be utilized for future communication.

For the case of a handover when $Count_{MU} = n\nu$, the actual path switch from the new target SC is forwarded to the core network and the default handover procedure is performed and the backward key separation restored. The counter design, i.e., the choice of ν is based

on how often the full path switching is required by the system. We also show through our results how relaxing this constraint can improve the handover performance. As we have highlighted before, the constraint of 2-hop key separation is difficult to achieve anyway for small cell systems with high-latency backhaul and more often leads to handover failures or delay jitters. The benefit of our proposed scheme is that we propose an alternative handover architecture in order to deal with the practical aspects of small cells systems. Until a major revamp for creating a more distributed MME architecture is proposed, we are suggesting practical solutions within the existing framework. The local path-switching based handover procedure is illustrated in Figure 17.

The proposed LP-based handover mechanism offers the major benefit of minimizing the handover interruption time τ by minimizing the path switching delay T_p during handover. In addition, it offers key benefits of minimizing the handover-related signaling costs. These benefits will be shown quantitatively through a thorough analysis and detailed simulations in later sections.

Route Optimization-enhanced handover: Although the LP-based handover mechanism can potentially offer key handover performance benefits, the handover cost can be further minimized for the LA-based architecture. During an inter-small cell handover, the "in transit" downlink data forwarded from the serving SC's buffer to the target SC follows the path LA $\Rightarrow$ Serving SC $\Rightarrow$ Target SC. This, albeit only for the handover duration, is analogous to the **triangular routing** occuring in mobile IP networks and results in increased handover cost (specifically, data forwarding cost). This is indicated in Figure 18.

In this approach, the triangular routing can be overcome by sending the path switch request message from the target SC before the MU performs radio connection setup at the target SC. Using this modified approach, the LA will be able to establish a new S1 path for the target SC and switch the local data path to the target SC. This means that the downlink data will be directly forwarded to the target SC instead of being forwarded through the X2 link setup between serving SC and target SC. To avoid loss of in-transit data packets,

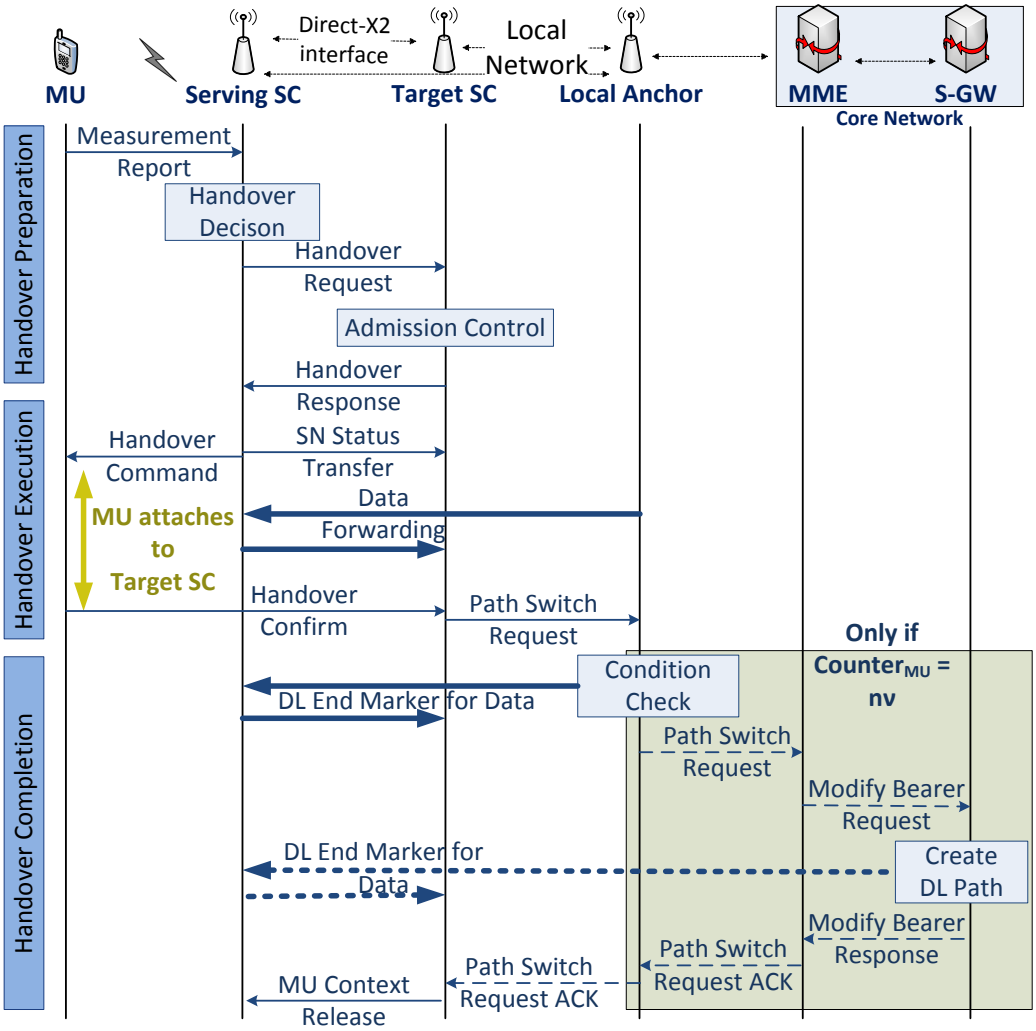

Figure 17: LP-based handover mechanism for coordinated small cells.

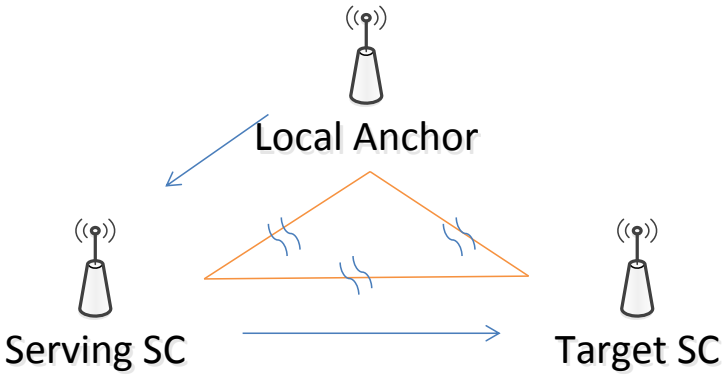

Figure 18: Triangular routing of in-transit data during handover.

the sequence number (SN) of the downlink data packet (delivered over X2 control path over the **SN status transfer message**) is indicated in the path switch request message. However, it must be noted that the uplink data of the user will continue to be forwarded from serving SC's buffer to the target SC until the buffer is empty. We term this approach as route optimization-enhanced (RO-enhanced) handover mechanism. Similar to the case of LP-based handover, the RO-enhanced mechanism involves sending the path switch request message to the core network whenever $Count_{MU} = n\nu$. The RO-enhanced handover scheme is illustrated in Figure 19.

Data Forwarding-enhanced handover: We further investigate a special handover scenario when the MU handovers from a LA to a neighboring SC in the cluster. In this case, it is possible to entirely eliminate the path switching operation. The X2 link created between LA and target SC during the handover preparation phase will be utilized is generally released after the handover procedure. However, in this data forwarding-enhanced (DF-enhanced) approach, the X2 link will be continued to be used to forward data packets from the LA to the target SC in a similar way as proposed in [49]. Without the path switch procedure, the serving SC does not receive an end marker for data transfer and hence the X2 link is continued to be used for data forwarding. Nevertheless, when the MU moves out

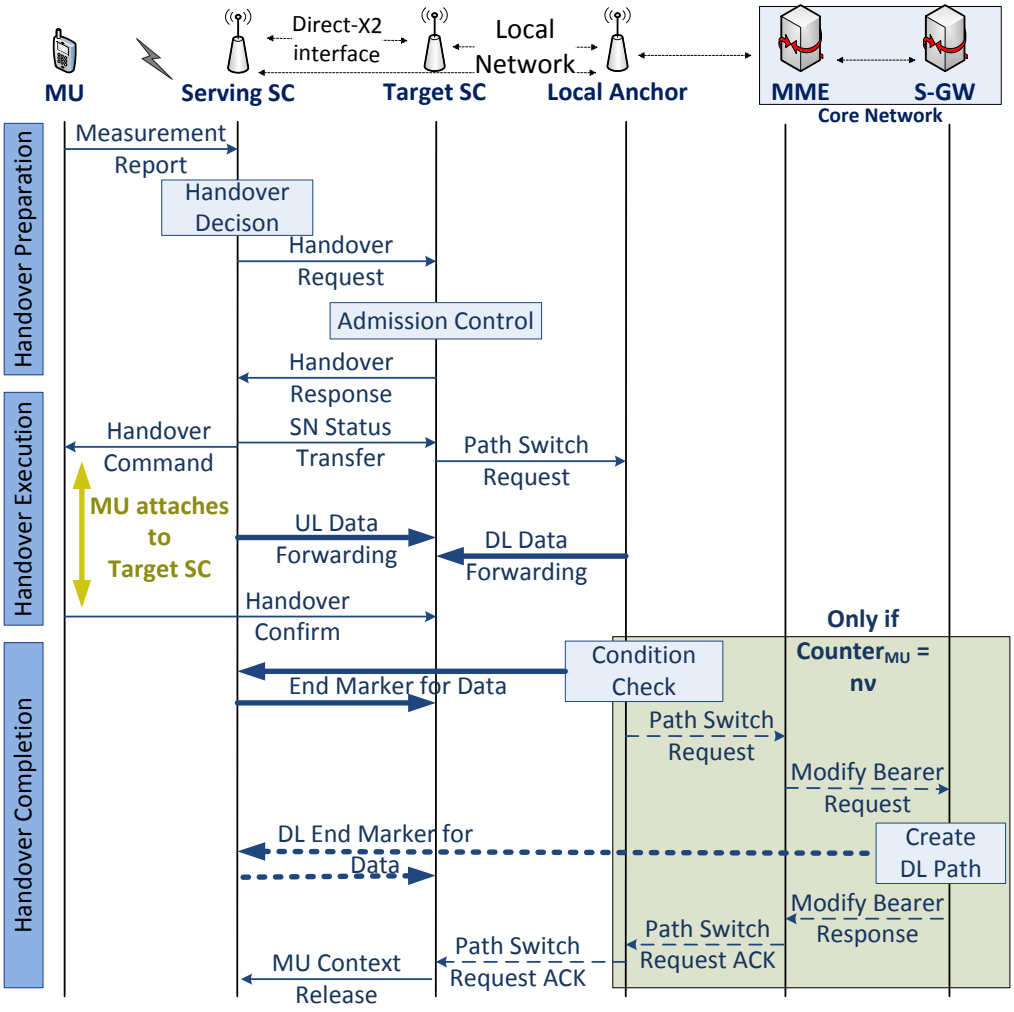

Figure 19: RO-enhanced handover mechanism for coordinated small cells.

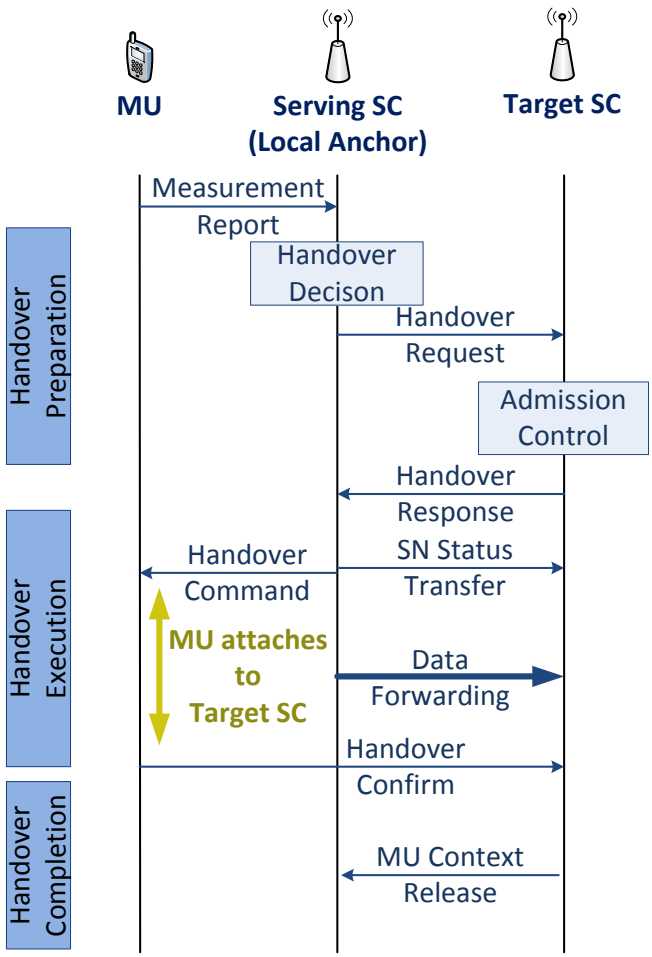

Figure 20: DF-enhanced handover mechanism for LA to SC handover.

of the target SC to a new SC, the RO-enhanced handover mechanism will be applied to obtain maximum savings in terms of handover costs. The DF-enhanced handover mechanism is illustrated in Figure 20.

Key benefits of our proposed handover schemes: Even though the concept of locally anchoring mobility has been proposed in the literature, there is not sufficient work that provides a scalable and practical architecture for supporting handover in high-density small cells. We highlight the key differences of our proposed solutions in this section.

In our proposed schemes, the network still perceives that the mobile user is attached to the local anchor even if it is attached to any small cell in the cluster. This way, there is a single point of control (MME) removing the need for synchronization.

In terms of security key management, the only impact we foresee is that small cell keys will not be refreshed after handover resulting in horizontal key derivation instead of vertical key derivation. At the same time, our argument is that the forward key separation is difficult to achieve anyway for small cells that experience poor backhaul conditions since the users can undergo several handovers before the new small cell receives the necessary information to generate a new virtual key after the user handovers.

In summary, we provide an alternative handover architecture in order to deal with the practical aspects of small cells systems. Until a major revamp for supporting a scalable and distributed MME architecture is proposed, we are suggesting practical solutions within the existing framework.

2.2.1.2 Analytical Model

To evaluate the performance of the proposed handover mechanisms, we need to study the evolution of the user's traffic and mobility behavior in the coordinated small cell network. To achieve this, we utilize a discrete-time Markov model to capture the user behavior in a cluster of small cells containing a local anchor small cell. For the developed analytical model, the stationary probabilities for a MU in each small cell are derived. This result will be later utilized to obtain closed-form expressions for handover performance parameters.

Model Description: We consider a two-dimensional grid topology to model a cluster of small cells as recommended in [54], [55]. This model has also been utilized in [49] mentioned in the prior art. This is shown in Figure 21. Each block in the grid corresponds to a small cell. The local anchor is located centrally in the grid and the other smalls cells within the cluster are deployed surrounding the local anchor in tiers. The topology consists of up to K tiers of small cells around the LA. It is constructed such that each SC has four neighboring SCs except for the SCs in the K^{th} tier. The number of SCs in tier i equals $4i$. State variables are used to indicate the presence of a user with an active session within the area of a small cell in the model. The state S_0 indicates that the MU is associated with the LA. In general, the state variable S_i^j indicates that the MU is associated with the SC in tier

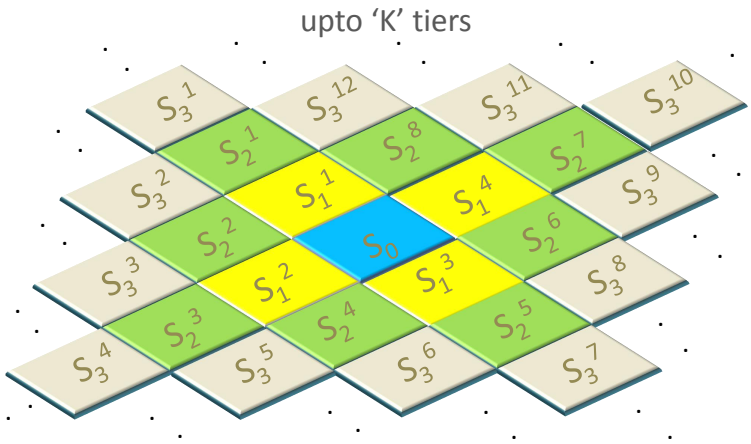

Figure 21: Two-dimensional grid topology for small cell cluster.

i and cell j within the tier i. An additional state variable S_{idle} is used to indicate that the user currently has no active session independent of which cell the user is associated with.

We consider that the MU changes state only at the end of a discrete time slot Δt. The user session is represented using the session arrival and session duration parameters. In this work, we consider that the session arrival follows Poisson distribution with session arrival rate λ. Therefore, the session arrival probability is given by $P_\lambda = \lambda \Delta t$. The session duration follows exponential distribution with mean session duration of $1/\mu$. Therefore, the corresponding probability $P_\mu = \mu \Delta t$ indicates the probability of a session terminating. For the user mobility, a random mobility model is considered where the users can move from an SC to any of its neighbors with equal probability. The cell residence time which represents the time the user remains in its current cell is modeled using exponential distribution. The mean cell residence time is given by $1/r$ and the corresponding probability of a user leaving the current cell is given by $P_r = r \Delta t$. Due to the memoryless properties considered, the evolution of user's traffic and mobility behavior can be modeled using a discrete-time Markov model.

In the Markov model, each cell would be represented by a state variable. However, this

for $m = 1 \rightarrow K$ **do**
  for $b = 1 \rightarrow \lceil (m+1)/2 \rceil$ **do**
    for $a = 0 \rightarrow 3$ **do**
      $S_m^b = S_m^b \cup S_m^{am+b} \cup S_m^{(a+1)m+(2-b)}$
    end
  end
end

Algorithm 3: Algorithm to perform state aggregation for the Markov model.

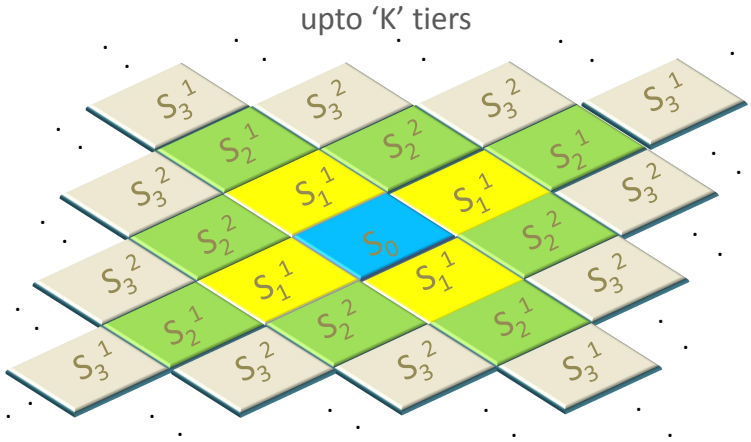

Figure 22: Grid topology representing aggregated states.

will result in state space explosion as the number of SCs in the cluster increases. Therefore, we apply state aggregation for the Markov model making use of location symmetry arising from the considered mobility model. The state aggregation algorithm is given in Algorithm 3. After state aggregation, we have K tiers with M states in each tier such that $M = \lceil (K+1)/2 \rceil$. The topology after performing state aggregation is shown in Figure 22.

The discrete-time Markov model for the aggregated states is represented in Figure 23. In the Markov model, the MU remains in state S_{idle} with a probability of $1 - P_\lambda$. After session arrival, the MU wakes up into any of the S_i^j states based on the cell density in each state. The MU returns to the S_{idle} state from any of the S_i^j states with a probability of P_μ. With a probability of $(1 - P_\mu)(1 - P_r)$, the MU remains in the current cell while having an active session. Also, $P_r(1 - P_\mu)P_{s_{ij}s_{\bar{i}\bar{j}}}$ indicates the transition probability that the MU

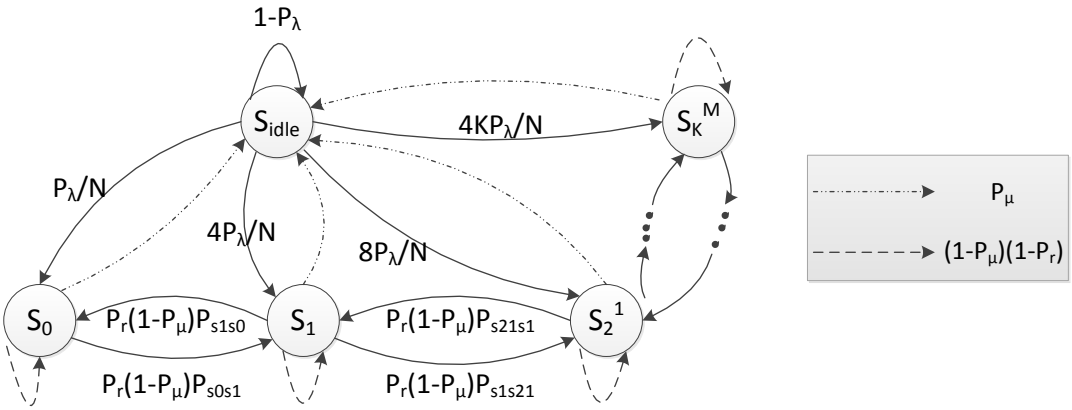

Figure 23: Discrete-time Markov model for the aggregated states.

handovers from state S_i^j to state $S_{\bar{i}}^{\bar{j}}$. Upon the completion of a session, the MU returns to the S_{idle} state. In the model, N represents the total number of small cells in the model. The probability $P_{s_{ij}s_{\bar{i}\bar{j}}}$ is obtained based on the user mobility characteristics and the number of tiers in the model. For example,

$$P_{s_{31}s_{21}} = \begin{cases} \frac{1}{4} & \text{if } K > 3, \\ 1 & \text{if } K = 3. \end{cases} \tag{25}$$

Using the transition probability matrix of the Markov model, the normalization and balance equations are determined. Before providing the balance and normalization equations, we define the following parameters.

$$\{\alpha_i, \beta_i\} = \begin{cases} \{\frac{1}{2}, \frac{1}{4}\} & \text{if } i < K-1, \\ \{1, 1\} & \text{if } i = K-1, \\ \{0, 0\} & \text{if } i = K. \end{cases} \tag{26}$$

$$\pi_{idle} = (1 - P_\lambda)\pi_{idle} + P_\mu \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j, \tag{27}$$

$$\pi_0^1 = \frac{P_\lambda}{N}\pi_{idle} + (1 - P_\mu)(1 - P_r)\pi_0^1 + P_r(1 - P_\mu)\beta_0\pi_1^1, \tag{28}$$

$$\pi_1^1 = \frac{4P_\lambda}{N}\pi_{idle} + (1 - P_\mu)(1 - P_r)\pi_1^1 + P_r(1 - P_\mu)\left\{\pi_0^1 + \beta_1\pi_2^1 + \alpha_1\pi_2^2\right\}, \tag{29}$$

$$\pi_i^1 = \frac{4P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_i^1 + P_r(1-P_\mu)\Big\{\frac{1}{4}\pi_{i-1}^1 + \beta_i\pi_{i+1}^1 + \frac{1}{2}\alpha_i\pi_{i+1}^2\Big\}\ ;\ \forall\ i > 1, \quad (30)$$

$$\pi_2^2 = \frac{4P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_2^2 + P_r(1-P_\mu)\Big\{\frac{1}{2}\pi_1^1 + \frac{1}{2}\alpha_2\pi_3^2\Big\}, \quad (31)$$

$$\pi_3^2 = \frac{8P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_3^2 + P_r(1-P_\mu)\Big\{\frac{1}{2}\pi_2^1 + \frac{1}{2}\pi_2^2 + \frac{1}{2}\alpha_3\pi_4^2 + \alpha_3\pi_4^3\Big\}, \quad (32)$$

$$\pi_i^2 = \frac{8P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_i^2 + P_r(1-P_\mu)\Big\{\frac{1}{2}\pi_{i-1}^1 + \frac{1}{4}\pi_{i-1}^2 + \frac{1}{2}\alpha_i\pi_{i+1}^2 + \frac{1}{2}\alpha_i\pi_{i+1}^3\Big\}\ ;\ \forall i > 3, \quad (33)$$

$$\pi_{2j-2}^j = \frac{4P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_{2j-2}^j + P_r(1-P_\mu)\Big\{\frac{1}{4}\pi_{2j-3}^{j-1} + \frac{1}{2}\alpha_{2j-2}\pi_{2j-1}^j\Big\}\ ;\ \forall j > 2, \quad (34)$$

$$\begin{aligned}\pi_{2j-1}^j =& \frac{8P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_{2j-1}^j \\ &+ P_r(1-P_\mu)\Big\{\frac{1}{4}\pi_{2j-2}^{j-1} + \frac{1}{2}\pi_{2j-2}^j + \frac{1}{2}\alpha_{2j-1}\pi_{2j}^j + \alpha_{2j-1}\pi_{2j}^{j+1}\Big\}\ ;\ \forall j > 2,\end{aligned} \quad (35)$$

$$\begin{aligned}\pi_i^j =& \frac{8P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_i^j \\ &+ P_r(1-P_\mu)\Big\{\frac{1}{4}\pi_{i-1}^{j-1} + \frac{1}{4}\pi_{i-1}^j + \frac{1}{2}\alpha_i\pi_{i+1}^j + \frac{1}{2}\alpha_i\pi_{i+1}^{j+1}\Big\}\ ;\ \forall j > 2, i > 2j,\end{aligned} \quad (36)$$

$$\pi_{idle} + \sum_{i=0}^{K}\sum_{j=1}^{\lceil\frac{i+1}{2}\rceil}\pi_i^j = 1. \quad (37)$$

The balance equations are given in equations (27) - (36). The normalization equation is given in equation (37). Using equations (27) - (37) and using iterative method, we obtain the stationary probability distribution of the Markov model. Here, π_i^j is of the form $\pi_i^j = x_i^j\pi_{idle} + y_i^j\pi_i^j + P_r(1-P_\mu)z_i^j\ ;\ \forall i, j$. For simplicity, in the future sections, we use the following notations to represent the stationary probabilities.

$$\pi_i^j = \begin{cases} \Psi_i^j + \Omega_i^j & \forall i \neq 1, j \neq 1, \\ \Psi_i^j + \Theta_i^j + \Omega_i^j & \forall i = 1, j = 1, \end{cases} \quad (38)$$

where $\Psi_i^j = x_i^j\pi_{idle} + y_i^j\pi_i^j$, $\Theta_1^1 = P_r(1-P_\mu)\pi_0^1$, $\Omega_i^j = P_r(1-P_\mu)z_i^j$ and $\Omega_1^1 = P_r(1-P_\mu)(z_i^j - \pi_0^1)$. In other words, the terms Ψ, Θ and Ω are parts of the stationary probability term π_i^j each indicating a specific action that the user performs. The term Ψ_i^j indicates the probability that the user moves to a cell classified under the same state S_i^j from the previous time while

a session is active or simply wakes up into state S_i^j from an idle state S_{idle} upon the arrival of a session. Ω_i^j indicates the probability that the user undergoes handover from any state other than its current state S_i^j or S_{idle}. For the special case when the user's current state is S_1^1, the term Θ_1^1 indicates the probability with which the user handovers from the local anchor to a cell in state S_1^1.

We utilize the stationary probability distribution π_i^j and π_{idle} to obtain several handover performance metrics.

2.2.1.3 Performance Metrics

Our objective is to utilize the mathematical framework developed above to obtain the closed-form expressions for different handover performance metrics. The key parameters we consider in this work include average handover cost and average handover interruption time.

Average Handover Cost: The average handover cost C^{HO} is a key performance metric for handover schemes and we define this parameter as the mean of the total handover-related costs required per handover in the network. The cost functions are expressed in terms of the link latency and processing time involved for a handover.

$$\begin{aligned} C^{HO} =& \frac{1}{\xi}\Big\lfloor\frac{\xi}{\nu}\Big\rfloor\Big\{\dot{C}_{01}\Omega_0^1 + \dot{C}_{11}\Theta_1^1 + \sum_{i=1}^{K}\sum_{j=1}^{\lceil\frac{K+1}{2}\rceil}\dot{C}_{ij}\Omega_i^j\Big\} \\ &+ \frac{1}{\xi}\Big(\xi - \Big\lfloor\frac{\xi}{\nu}\Big\rfloor\Big)\Big\{\ddot{C}_{01}\Omega_0^1 + \ddot{C}_{11}\Theta_1^1 \\ &+ \sum_{i=1}^{K}\sum_{j=1}^{\lceil\frac{K+1}{2}\rceil}\ddot{C}_{ij}\Omega_i^j\Big\}, \end{aligned} \tag{39}$$

where

- ξ : Mean number of handovers per MU per session.
- ν : Number of handovers before full path switching.
- $\dot{C}_{ij}$: Handover cost to an SC in state S_i^j when $Counter_{MU} = n\nu$.

- $\ddot{C}_{ij}$: Handover cost to an SC in state S_i^j with local path switching when $Counter_{MU} \neq nv$.

Equation (39) is used to compute both the signaling cost C_{ij}^s and the data forwarding cost C_{ij}^D incurred during a handover by using the appropriate cost functions for $\dot{C}_{ij}$ and $\ddot{C}_{ij}$. For example, to obtain the average signaling cost, the cost functions $\dot{C}_{ij}$ and $\ddot{C}_{ij}$ in equation (39) are replaced by the signaling cost functions $\dot{C}_{ij}^s$ and $\ddot{C}_{ij}^s$ defined in Section 2.2.1.3 for each of the proposed handover schemes. Similarly, the average data forwarding cost is obtained by utilizing the data forwarding cost functions $\dot{C}_{ij}^d$ and $\ddot{C}_{ij}^d$ defined in Section 2.2.1.3.

Average Handover Interruption Time: As we discussed before, handover interruption time for a single MU handover is $\tau = \max(T_r, T_p)$ where T_r is the time required for the MU to establish radio connection with the target SC and T_p is the time required to perform path switch at the target SC. Therefore, τ is a measure of the seamlessness of the handover which is a very important metric for QoS performance.

The mean handover interruption time is obtained as follows. Let $\dot{\tau}_{ij}$ and $\ddot{\tau}_{ij}$ indicate the interruption time incurred when a user handovers to a cell in state S_i^j with and without full path switching. Since the interruption times depends on the link delay as shown in equations (41)-(43), different cases of handovers lead to different interruption times. Therefore, we mutiply these interruptions by the probability of the user undergoing each handover case. In addition, the percentage of the user undergoing a handover within the local anchor and with the core network is also considered while computing the expression for the mean handover interruption time. For the proposed model, we obtain the mean handover interruption time τ^{HO} as

$$\begin{aligned}\tau^{HO} =& \frac{1}{\xi}\Big\lfloor\frac{\xi}{\nu}\Big\rfloor\Big\{\dot{\tau}_{01}\Omega_0^1 + \dot{\tau}_{11}\Theta_1^1 + \sum_{i=1}^{K}\sum_{j=1}^{\lceil\frac{K+1}{2}\rceil}\dot{\tau}_{ij}\Omega_i^j\Big\} \\ &+ \frac{1}{\xi}\Big(\xi - \Big\lfloor\frac{\xi}{\nu}\Big\rfloor\Big)\Big\{\ddot{\tau}_{01}\Omega_0^1 + \ddot{\tau}_{11}\Theta_1^1 \\ &+ \sum_{i=1}^{K}\sum_{j=1}^{\lceil\frac{K+1}{2}\rceil}\ddot{\tau}_{ij}\Omega_i^j\Big\}, \end{aligned} \tag{40}$$

where

- $\dot{\tau}_{ij}$: Interruption times for handover to an SC in state S_i^j when $Counter_{MU} = nv$.
- $\ddot{\tau}_{ij}$: Interruption times for handover to an SC in state S_i^j with local path switching when $Counter_{MU} \neq nv$ respectively.

The interruption times are given by $\dot{\tau}_{ij} = max(T_r, \dot{T}_p)$ and $\ddot{\tau}_{ij} = max(T_r, \ddot{T}_p)$ where $\ddot{T}_p$ is given by

$$\ddot{T}_p^{LP} = \ddot{T}_p^{RO} = \begin{cases} C_{la}; & \text{if } i = 0, j = 1, \\ C_{sc} + C_{la} + 2 * C_{s1}; & \text{if } i = 1, j = 1, \\ C_{sc} + C_{la} + 2 * C_{s1}; & otherwise, \end{cases} \tag{41}$$

for the LP-based and RO-enhanced schemes. Here, C_{sc}, C_{la} are the processing costs at small cell and local anchor respectively in milliseconds. C_{X2} is the X2 link cost, whereas C_{s1} represents the S1 link cost with local anchor. For the DF-enhanced scheme, it is given by

$$\ddot{T}_p^{DF} = \begin{cases} C_{la}; & \text{if } i = 0, j = 1, \\ 0; & \text{if } i = 1, j = 1, \\ C_{sc} + C_{la} + 2 * C_{s1}; & otherwise, \end{cases} \tag{42}$$

$\dot{T}_p$ for all the three schemes is given by

$$\dot{T}_p = \begin{cases} C_{la} + C_{scgw} + 2C_{s1*} + C_{ps}; & \text{if } i = 0, j = 1, \\ C_{sc} + 2C_{la} + C_{scgw} + 2C_{s1*} & \\ +C_{ps}; & otherwise, \end{cases} \tag{43}$$

where C_{s1*} is the link cost for the S1 link between the SC and the core network. In addition, C_{scgw} represents the processing cost at the small cell gateway and C_{ps} represents the path switching cost. In order to compare our solution to the 3GPP LTE-A handover approach, we apply $\ddot{\tau}_{ij} = \dot{\tau}_{ij}$.

Average Core Network Load: The average core network load is defined as the average number of signaling messages that involve the core network (both MME and S-GW) resulting from a MU handover. The expression for the average core network load can be obtained from equation (39) by replacing $\dot{C}_{ij}$ and $\ddot{C}_{ij}$ with the per-handover core network load l_{CN}.

Cost Functions: The cost functions are two-fold: (i) Signaling cost C^s_{ij}, and (ii) Data forwarding cost C^d_{ij}. We compute the cost functions as the sum of the link delays due to the transmission of the signaling messages and the processing delay for these signaling messages at different nodes during a handover process.

In our topology, the SC in a cluster can be connected to the LA through multi-hop using the local network. However, since the intermediary SCs only act as IP routers for the SC $\Leftrightarrow$ LA communication, the processing cost at the intermediate SCs mainly originate from the router processing cost. In this case, the intermediate SCs do not require the additional processing cost at the X2 layer performing GPRS tunneling functions and S1 application protocol functions. With this in mind, we have considered that the routing processing cost to be negligible in the computation of the cost function.

LP-based handover: We provide the signaling and data forwarding costs for the LP-based handover mechanism when local path-switching is applied under conditions that $Counter_{MU} \neq n\nu$. For the case of $Counter_{MU} = n\nu$, the 3GPP LTE-A handover cost will be applied which will be presented later in this section. The handover costs can vary depending on whether the handover occurs from LA to SC or SC to SC.

The signaling cost $\ddot{C}^s_{ij}$ is given as

$$\ddot{C}^s_{ij} = \begin{cases} 4C_{sc} + 2C_{la} + 4C_{X2} + C_{s1}; & \text{if } i = 0, j = 1, \\ 5C_{sc} + 2C_{la} + 4C_{X2} + 2C_{s1}; & \text{if } i = 1, j = 1, \\ 5C_{sc} + 2C_{la} + 4C_{X2} + 3C_{s1}; & otherwise. \end{cases} \tag{44}$$

where C_{sc}, C_{la} are the processing costs at small cell and local anchor respectively. C_{X2} and C_{s1} represent the X2 and S1 link costs respectively.

The data forwarding cost $\ddot{C}^D_{ij}$ is given as

$$\ddot{C}^D_{ij} = \begin{cases} C_{sc} + C_{la} + C_{X2} + C_{s1}; & \text{if } i = 0, j = 1, \\ C_{sc} + C_{X2}; & \text{if } i = 1, j = 1, \\ C_{sc} + C_{la} + C_{X2} + C_{s1}; & otherwise. \end{cases} \tag{45}$$

RO-enhanced handover: The signaling cost for this approach is the same as in equation (44). The data forwarding cost is given as

$$\ddot{C}^D_{ij} = \begin{cases} q(C_{sc} + C_{X2}) + C_{la}; & \text{if } i = 0, j = 1, \\ C_{la} + C_{s1}; & \text{if } i = j = 1, \\ q(C_{sc} + C_{X2}) + & \\ (1 - q)(C_{la} + C_{s1}); & otherwise, \end{cases} \tag{46}$$

where q and $1 - q$ are fractions of uplink and downlink data respectively.

DF-enhanced handover: The signaling cost and data forwarding cost under the DF-enhanced handover approach differ from the RO-enhanced handover only for the case when an MU experiences handover from LA to SC. These are given as $\ddot{C}^s_{11} = 4C_{sc} + 4C_{X2}$ and $\ddot{C}^D_{11} = C_{sc} + C_{X2}$.

For the proposed handover approaches, when path switching is applied, the data forwarding cost will still be $\dot{C}^s_{ij} = \ddot{C}^s_{ij}$. However, the signaling cost is given by

$$\dot{C}^s_{ij} = 5C_{sc} + 4C_{X2} + 3C_{s1*} + 2C_{scgw} + C_{ps}. \tag{47}$$

We compare our solution for the case when no local mobility anchoring is utilized. In this case, we apply $\ddot{C}^s_{ij} = \dot{C}^s_{ij}$ to determine the signaling cost. Similarly the data forwarding cost is given by $\dot{C}^D_{ij} = \ddot{C}^D_{ij} = C_{sc} + C_{X2} + C_{la} + C_{s1}$.

2.2.1.4 Analytical and Simulation Results

Numerical Evaluation: First, we numerically evaluate the proposed schemes based on the performance metrics derived in Section **??**. The system parameter values used for the numerical evaluation are provided in Table 4 as recommended in [45]. The handover

performance metrics are provided as ratio of the values achieved by the proposed scheme to the ones achieved by the current 3GPP handover scheme. This means that a smaller ratio corresponds to improved handover performance.

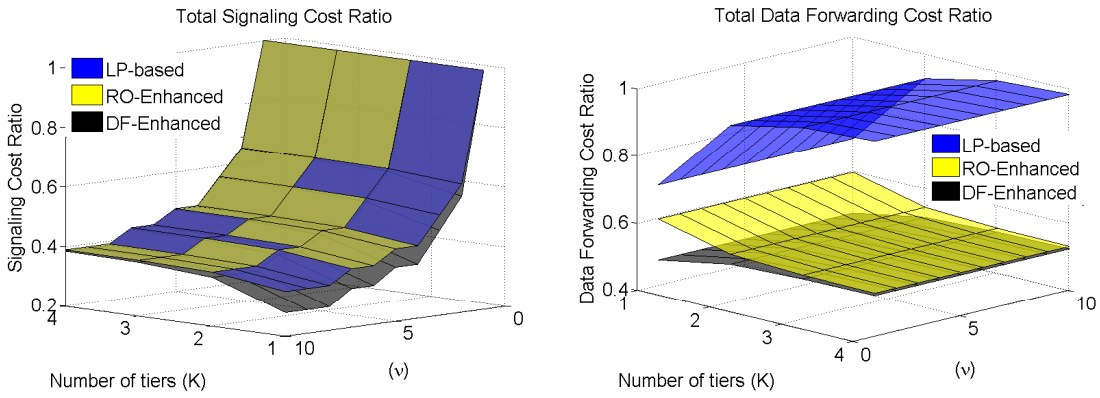

Figure 24: Numerical evaluation of handover performance vs number of tiers K and path switching threshold ν.

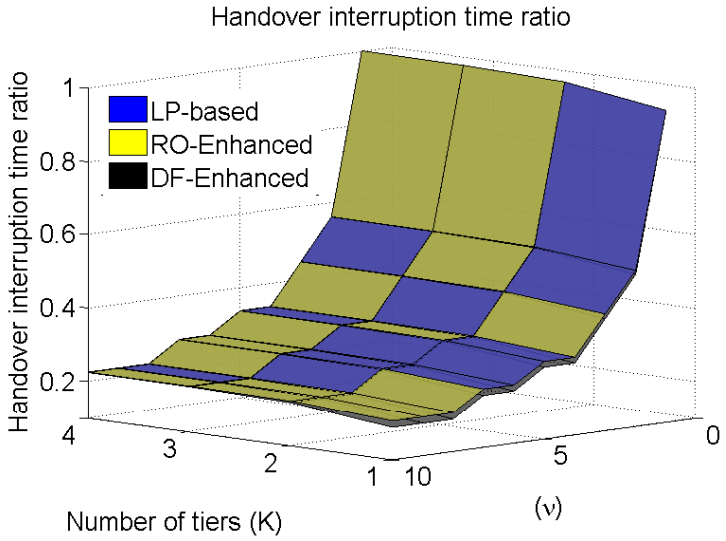

Figure 25: Handover interruption time vs number of tiers K and path switching threshold ν.

In Figure 24(a), the average signaling cost ratio under the proposed handover schemes is plotted as a function of the number of tiers (K) in the cluster and the path switching threshold (ν). We observe that all three proposed schemes vary with K and ν in a similar

Table 4: System parameters.

Parameter	Value	Parameter	Value
Session arrival rate λ	$0.001/s$	C_{sc}	5ms
Session duration parameter μ	$0.01/s$	C_{la}	10ms
Cell residence parameter r	$0.1/s$	C_{s1}	10ms
Number of tiers K	$1, 2, 3, 4$	C_{X2}	5ms
Number of small cells N	$5, 13, 25, 41$	C_{s1*}	50ms
Discrete-time slot duration Δt	$0.01s$	C_{scgw}	10ms
–	–	C_{ps}	5ms

manner. The signaling cost ratio is large for small values of ν. This is expected as low values of ν will trigger more frequent path switching for each handover. However, as ν is increased, we observe that the signaling cost ratio decreases rapidly and reaches to around 40% of the maximum ratio. This corresponds to about 60% of signaling cost savings. This validates our claims that reducing the frequency of full path switching with the network results in signaling cost savings. Among the three mechanisms, the DF-enhanced handover offers the highest savings in signaling cost of about 70% for $K \leq 2$ and $\nu \geq 8$. Even for $\nu = 6$, all the three proposed schemes are able to offer almost 60% signaling cost savings. It is also interesting to note that increase in tier size (K) does not have a major effect on the signaling cost performance. This allows for scalable cluster sizes given that the network infrastructure is capable of supporting coordination among the member SCs.

The average data forwarding cost is plotted as a function of ν and K in Figure 24(b). We observe that the LP-based handover does not offer significant performance gain when $K \geq 3$. This is expected as there is a triangular routing of the "in-transit" data from the serving SC to the target SC during a handover. However, with the support of route-optimization for the "in-transit" data packets, the RO-enhanced and DF-enhanced schemes are able to achieve about 50% gain in the average data forwarding cost. It is also important to note that both these schemes do not vary significantly with ν since the route optimization is independent of the frequency of path switching. It is worthy of observing that the RO-enhanced scheme has a unique behavior compared to other schemes with increasing K. This is caused due to the high cost involved for a MU handover from LA to a neighboring

SC. However, as K increases, the probability of this handover occuring decreases and hence the data forwarding cost becomes low.

Figure 25 highlights the average handover interruption time ratio plotted against ν and K. The parameter affecting the handover interruption time is mainly the path switching threshold. As ν increases, the path switching takes place less frequently and hence we can observe that the handover interruption time decreases and reaches a minimum of up to 20% of the maximum ratio. This corresponds to about 80% reduction in the interruption time. This is a key benefit of our proposed schemes as lower interruption time reduces handover failures and enable seamless mobility for users.

Simulation Setup: In order to validate the performance of the proposed handover schemes, we conduct Monte-Carlo simulation of a small cell cluster with several mobile users using MATLAB. A two-tier small cell cluster is considered consisting of N = 13 small cells (one of which is a local anchor) and 50 mobile users in a 100m x 100m area. This can be visualized as a deployment in a typical office building with a uniform and coordinated deployment. The topology is shown in Figure 26. In the figure, each small cell is considered to provide coverage over a 10m x 10m grid. The users are deployed randomly in the grid indicated by the thick dots under the small cell coverage.

Session arrival for users is modeled using Poisson distribution with mean arrival rate λ. Session duration and cell residence time are modeled using exponential distribution with parameters μ and r respectively. As considered in the analytical model, the users can move to each of their neighboring cells with equal probability at the end of the slot duration. A slot duration of $10ms$ is considered and the simulation is performed for 10000 time slots.

The state evolution is generated for each of the users based on the discrete-time Markov model for the entire simulation duration. Handover occurs in cases when the MUs change their states between any of the S_i^j states in successive time slots. At each time step, the previous and current states are identified to determine if a handover for the MU has occured. When the MU experiences a handover, the handover performance metrics are computed for

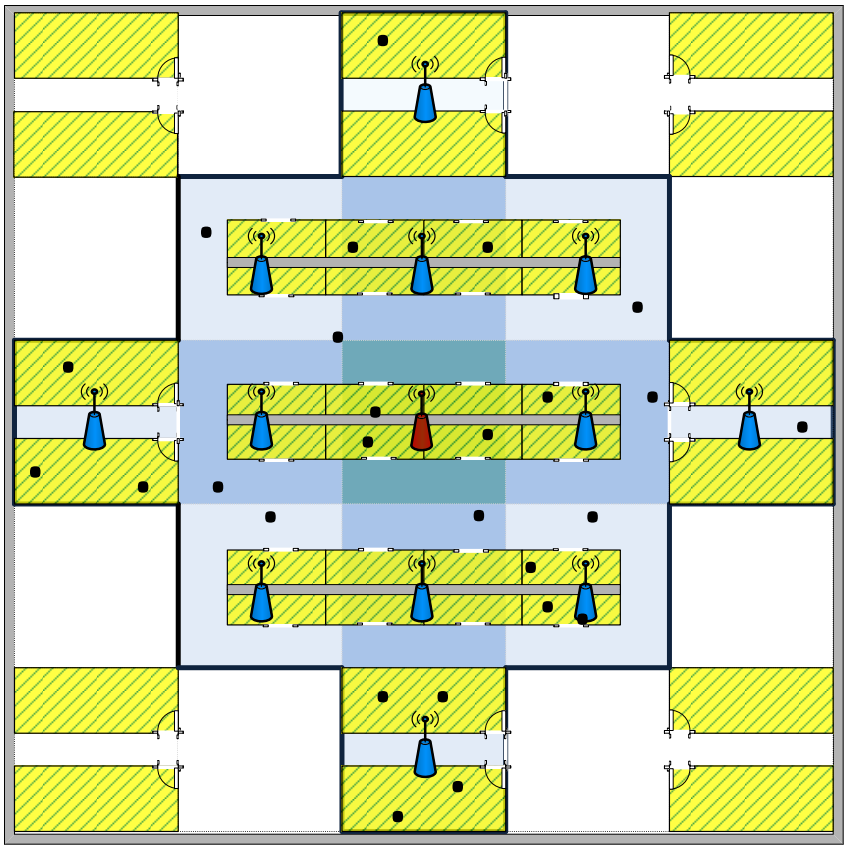

Figure 26: Simulation of a small cell cluster with K = 2.

the proposed schemes and the current LTE-A handover scheme. The entire simulation is conducted for 10^4 Monte-Carlo iterations. At the end of the simulation, the ratio of the handover costs and handover interruption times of the proposed schemes and the existing LTE-A handover scheme are determined.

First, the above simulation is performed for different values of ν and the results are plotted against the path switching threshold ν in comparison to the analytical results obtained earlier. Further, we also study the effect of the session arrival and user mobility on the handover performance metrics. In the latter case, the path switching threshold is fixed but the session arrival parameters and mobility parameters are varied for which the results are obtained.

Simulation of a Large Network with Gauss-Markov mobility model: In order to further validate our results, we simulate a large network with several clusters. For this, we constructed a scenario with a rectangular grid containing 100x100 small cells with each having a coverage area of 10m. The intersection of every third row and column contains a local anchor. This is chosen to resemble a cluster of tier size $K = 2$. The network consists of 500 users with the initial locations of users randomly selected in the 2-D grid. The users experience mobility based on the well-known Gauss-Markov mobility model [56].

Under the Gauss-Markov mobility model, each user chooses a velocity and direction to follow at every discrete time step. The velocity V and direction D of the users at time step t is computed based on the values at time step $t-1$ and a random variable as shown in the following expressions.

$$V_t = \alpha V_{t-1} + (1-\alpha)\bar{V} + \sigma_V \sqrt{1-\alpha^2} w_{V_{t-1}}, \tag{48}$$

$$D_t = \alpha D_{t-1} + (1-\alpha)\bar{D} + \sigma_D \sqrt{1-\alpha^2} w_{D_{t-1}}, \tag{49}$$

where $0 < \alpha < 1$ is a tuning parameter to determine the randomness of the movement. $\bar{V}$ and σ_V indicate the asymptotic mean and standard deviation of the velocity. Similarly $\bar{D}$

and σ_D indicate the asymptotic mean and standard deviation of the direction. $w_{V_{t-1}}$ and $w_{D_{t-1}}$ are uncorrelated Gaussian processes with zero mean and unit variances and are independent of V_{t-1} and D_{t-1}.

Then the (x_t, y_t) coordinates of the user in 2-D space at time t is given by

$$x_t = x_{t-1} + V_{t-1} cos D_{t-1}, \tag{50}$$

$$y_t = y_{t-1} + V_{t-1} cos D_{t-1}, \tag{51}$$

The simulation is run for 1000 time steps such that each time step has a duration of 1s. The session parameters are retained from Section 2.2.1.4. For the Gauss-Markov mobility model, we adopt $\alpha = 0.5$, $\bar{V} = 6m/s$ and $\bar{D} = \pi/2$. The asymptotic variances are set as $\sigma_V = 3$ and $\sigma_D = \pi/8$. $w_{V_{t-1}}$ and $w_{V_{t-1}}$ determined from a standard normal distribution.

Users experience handover if they cross cell boundaries while a session is active between successive time steps. The handover cost functions are computed based on the type of the handover, i.e., if the user handovers from small cell to another small cell or local anchor and so on. For inter-cluster handover, we consider the cost to be equivalent to a SC-SC handover with full path-switching. The simulation is conducted for 10000 Monte Carlo trials and the average handover costs are computed.

Comparison of Numerical and Simulation Results: We provide a comparison of the results from the analytical and the simulation models described above.

Impact of path switch threshold on handover performance: The handover costs and interruption times are plotted against ν for a fixed $K = 2$. Figure 27 plots the average signaling cost ratio against the path switching threshold ν for the proposed schemes. All three proposed handover schemes show consistent behavior in the numerical and simulation evaluation where the average signaling cost decreases sharply at the beginning. However, beyond $\nu = 4$, there is little impact on the signaling cost. This result is very important as it signifies that performance of the proposed schemes is not limited by the path switching threshold setting. This also validates our analytical expressions for the average signaling

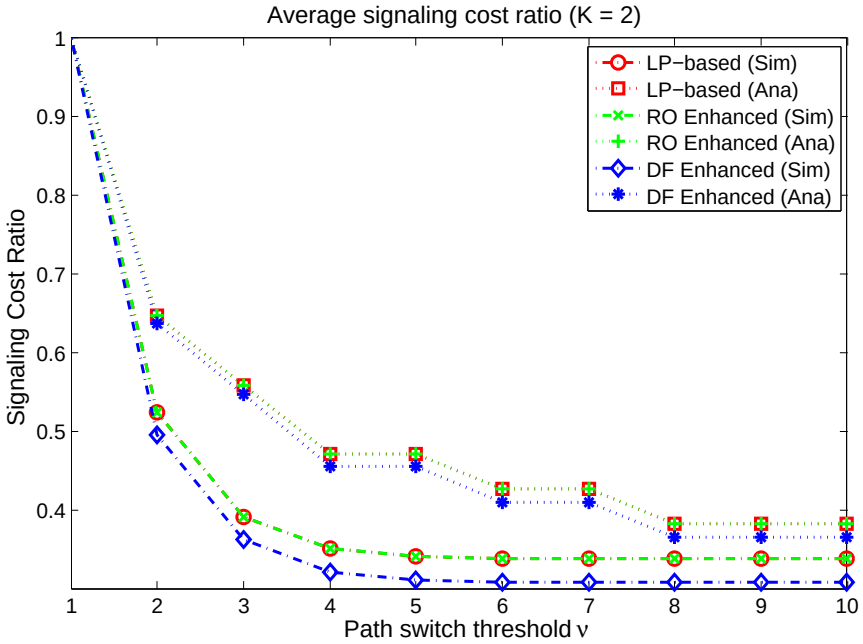

Figure 27: Impact of ν on average signaling cost ratio.

cost derived in section 2.2.1.3. In addition, it is interesting to observe that there is greater signaling cost savings through simulations compared to the numerical results achieving cost saving of over 80% for the LP-based and RO-enhanced schemes. The DF-enhanced scheme is able to achieve over 90% cost savings for $\nu \geq 4$.

In Figure 28, the average data forwarding cost is plotted. As expected, the data forwarding cost is independent of ν, which is in agreement with the numerical results. In addition, the cost values through simulation and numerical analysis follow each other very closely reaching up to 50% cost savings for the DF-enhanced scheme. A similar performance is observed for the handover interruption time as illustrated in Figure 29. All three schemes show similar performance with over 80% reduction in average interruption time as seen from the simulation results.

Impact of call-to-mobility ratio on handover performance: We study the impact of the session arrival rate and cell residence time on the handover performance. To this end, we utilize the call-to-mobility ratio parameter $\rho = \frac{\lambda}{r}$ where λ is the mean session arrival rate and r is the cell residence parameter. A low value of call-to-mobility ratio indicates the user

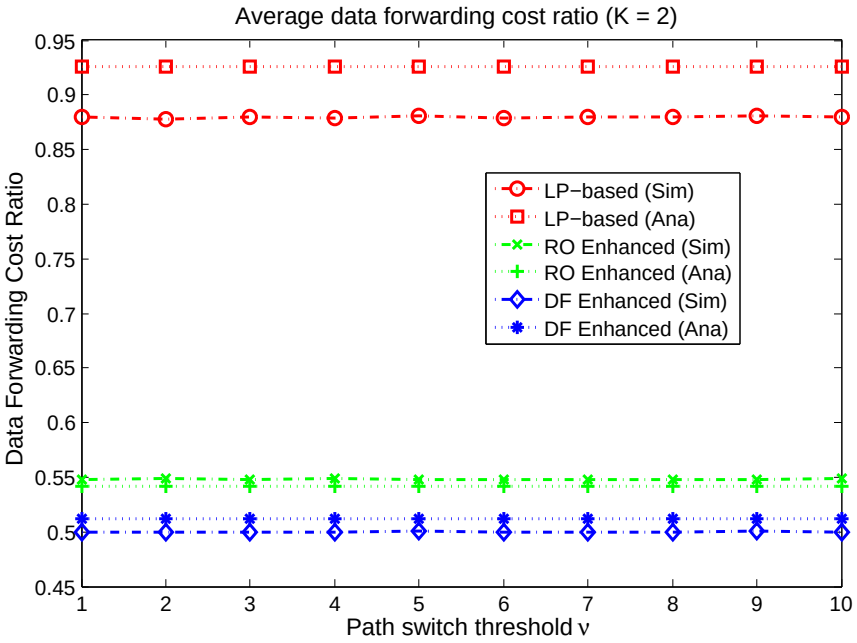

Figure 28: Impact of ν on average data forwarding cost ratio.

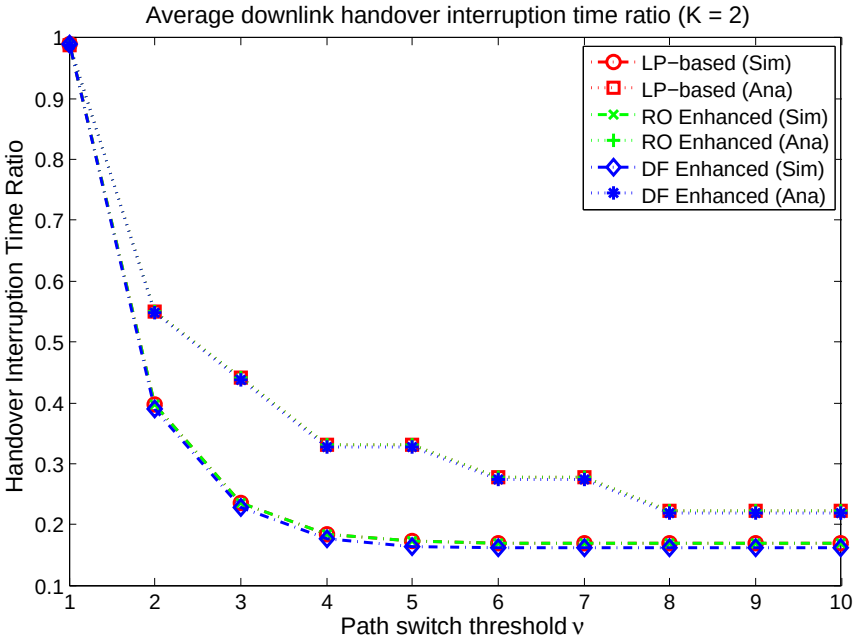

Figure 29: Impact of ν on handover interruption time ratio.

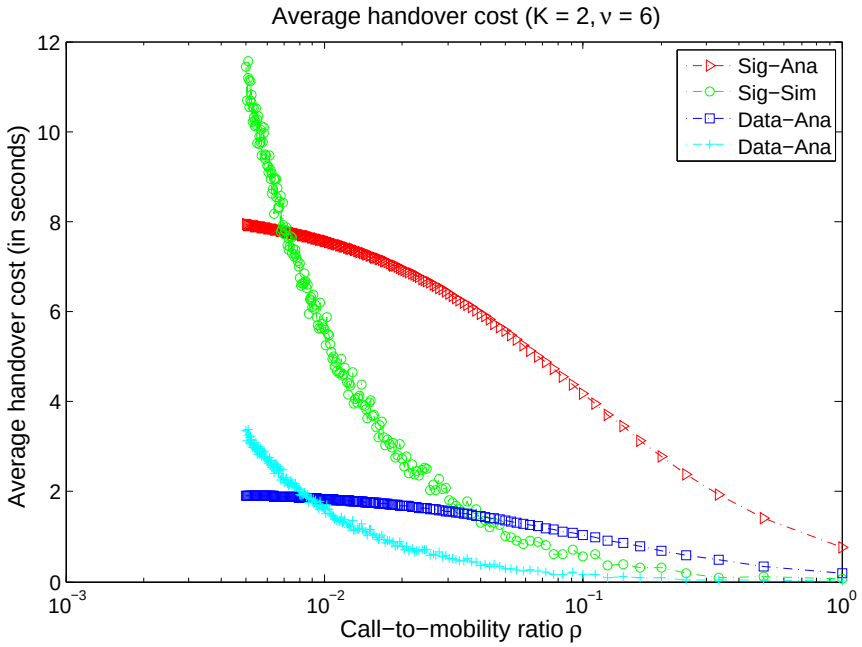

Figure 30: Impact of call-to-mobility ratio on signaling cost.

is highly mobile relative to the mean time between two session arrivals. This implies that the probability of a user undergoing a handover increases leading to high average handover costs. On the contrary, very large values of ρ indicate that the user is less mobile relative to the session arrival. This must lead to a lower probability of user handover leading to low average handover related costs.

The average handover costs for the LP-based scheme are plotted against the call-to-mobility ratio in Figure 30. For low values of ρ, the signaling and data forwarding costs remain very high due to high mobility of users. As ρ increases, the users become less mobile thereby experiencing lower handover probability. This is reflected by the low signaling and data forwarding costs for large values of ρ. Although, both the simulation and analytical results show the handover cost as a decreasing function of ρ, the handover cost experiences a steep fall much earlier in the simulation results than in the case of the analytical results. A similar behavior to the handover cost is observed for the handover interruption time when plotted against ρ as shown in Figure 31. These results also provides validation for the proposed analytical model.

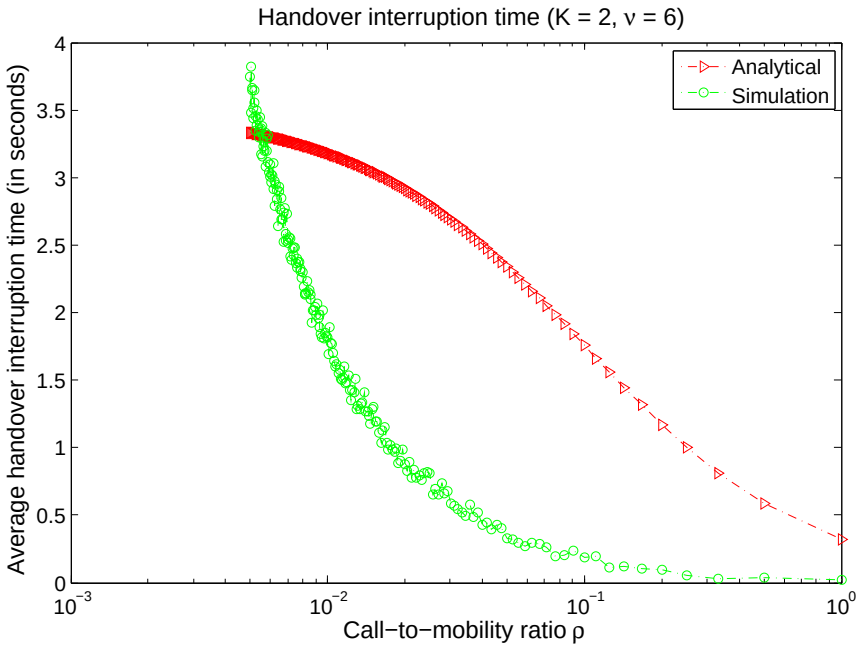

Figure 31: Impact of call-to-mobility ratio on handover interruption time.

Impact of different mobility models on handover performance: The handover costs and interruption times are computed for the different mobility models considered. These results are plotted in Figure 32 and 33. Our results show that the handover performance both in terms of handover costs and interruption time do not vary significantly if the users follow different mobility models. Especially, the Gaussian Markov mobility model, which provides a tradeoff between the randomness and memory follows closely the performance results from our analytical model. This further shows that the proposed analytical model is fairly accurate in capturing the behavior of the handover dynamics in the system.

2.2.1.5 Conclusions

The emergence of small cells requires a rethinking of the traditional cellular system concepts. Coordination is expected to play a key role in smalls cells to achieve improved mobility management and SON features among others. In this work, we utilized coordination among small cells to propose a local anchor-based handover architecture. Based on this, we proposed novel handover schemes employing a local mobility anchor. To evaluate the performance of the proposed schemes, we developed a mathematical framework based

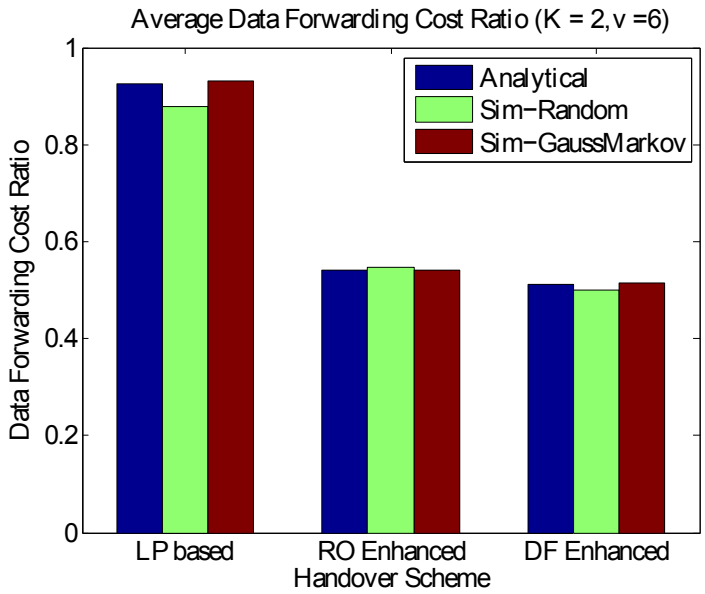

Figure 32: Data forwarding cost under different mobility models

on Markov models. Using this mathematical framework, we derived closed-form expressions for the key handover performance parameters including average handover cost and average handover interruption time. The performance is evaluated using numerical analysis and simulations which indicate savings of at least 60% in the handover cost and over 80% reduction in the handover interruption over existing schemes.

With the expected large-scale adoption of small cells in large offices and hot-spot areas, handover management approaches such as our proposed method will be key to facilitate seamless service for mobile users in a scalable and efficient manner. The proposed mathematical framework have been shown to closely follow more realistic models through simulation and hence can be utilized to analyze the performance of new handover schemes that employ a local anchor.

2.2.2 Group Handover for Mobile Relays

Relays are small cells that have a wireless backhaul to the macrocell base stations. Relay stations (RS) can intelligently relay data between a base station (BS) and mobile stations wirelessly. The communication between BS and RSs takes place in a similar way as the communication between BS and MSs using point-to-multipoint (PMP) connectivity. In

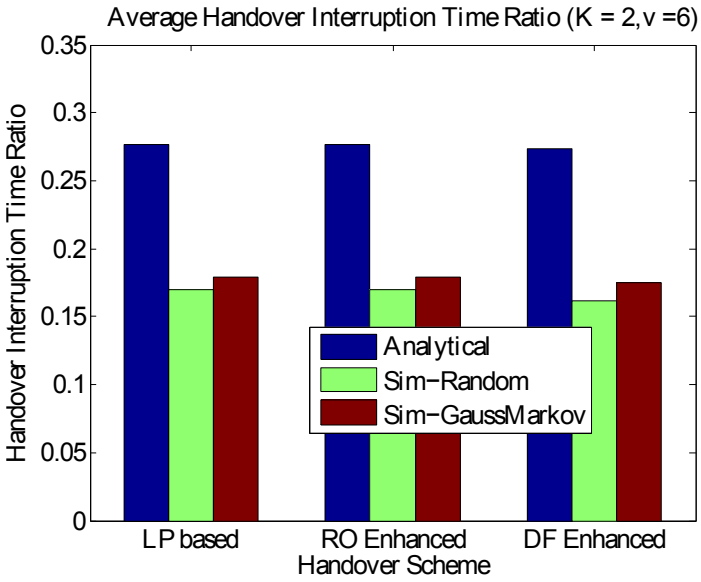

Figure 33: Handover interruption time under different mobility models.

other words, the RS(s) maintains a wireless backhaul connectivity with the BS and hence the network. At the other end, the RS can establish PMP connectivity with the MS. Therefore, the RS can provide both uplink and downlink connectivity for the MS. The BS⇔RS links and RS⇔RS links are usually referred to as relay links whereas the BS⇔MS links and RS⇔MS links are referred to as access links.

Relay station proves beneficial in several ways in the cellular network. First, like other small cells, RS can provide increased capacity through frequency reuse. In other words, capacity increase can be realized when both BS and RS in a given area communicate with different MSs using the same frequency resources. Second, it can provide improved coverage with lesser deployment costs as against femtocells from the fact that relay uses wireless backhaul link with the network. This facilitates adhoc deployment of relays in areas where the BSs cannot provide sufficient coverage (for e.g., at cell edges and coverage holes). Several deployment scenarios are envisioned for relay stations [57].

While fixed relays support several use cases defined for 4G systems, mobile relays can enable more use cases that are not part of the 4G standards yet. We identify that mobile relays can address the key requirements of low latency, reduced handover interruption

time and high spectral efficiency. Prior work on enhanced relay solutions primarily consider multi-hop and cooperative relays but very few take into account the impact on the standards. Similarly, IEEE 802.16j relay standard supports relay mobility. Although, its implementation/deployment is largely limited as it fails to reduce the relay station complexity and has backward compatibility issues with 802.16e [58]. We believe that the mobile relay solution we propose is unique in the way that it provides backward compatibility with legacy systems while providing significant gains in the performance.

The key idea behind the proposed group handover mechanism is that the radio links between the users and the mobile relay station is maintained throughout the handover process. Instead, it is only required for the relay station to change the point of attachment from its old BS to a new BS. This is illustrated in Figure 34. However, the challenge here is about preserving the per-user service flow between mobile users and network in a seamless manner while the backhaul is re-established.

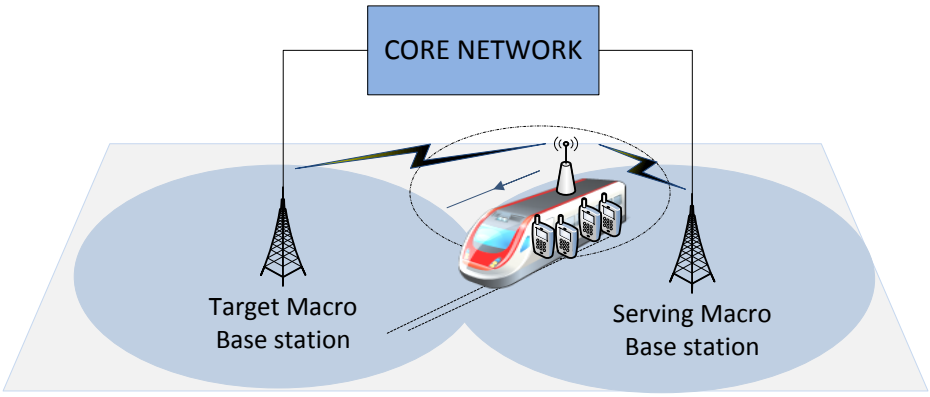

Figure 34: Group handover scenario in high-speed vehicles with mobile small cells.

In this work, we first highlight three important use cases, where mobile relays have the potential to provide performance improvements under high speed user mobility, heavy load conditions. Following this, we propose a group handover mechanism for mobile relays with a strong view of the current relay architecture of 4G systems including IEEE 802.16m as well as LTE-A. Based on this handover procedure, we describe the relay handover architecture for 802.16m Advanced Relay Station. For the proposed group handover scheme, we

highlight the control plane and user plane architectural enhancements required to support relay mobility. Finally, we carry out a performance comparison of mobile relays against the "no relay" and "non-mobile relay" scenarios mainly using the key parameters including overheads on the access and core network.

2.2.2.1 Use Cases of Mobile Relay

The key features of relays such as low equipment cost and flexible deployment options have resulted in strong interest from the industry as well as academia to focus on relay mobility. More intriguing is the fact that mobile relays can potentially be leveraged to achieve new levels of seamless user mobility experience. We highlight three important use cases for leveraging mobile relays to realize some key IMT-Advanced requirements. These use case are discussed as follows.

Group Mobility: First, we highlight group mobility where an improved handover mechanism enables seamless mobility of a group of users. Group mobility makes sense for concurrent handovers to be performed for a group of users in high speed vehicles such as trains and buses. The idea is to have a mobile relay station in the high speed vehicle serving the MSs within. Mobility results in a handover of the relay station to a neighboring BS. At the same time, from the perspective of the MSs, the point of attachment, i.e., RS remains the same. Hence, the PMP connectivity for the MSs is preserved while the RS handover procedure is performed transparent to MSs. Group mobility can significantly reduce the overheads on the radio interface and network thus minimizing latency for all users. In addition, this improved handover mechanism can significantly reduce the handover interruption time which is one of the key requirements for IMT-Advanced systems.

Network Reliability: The mobile relay concept can also be applied to improve the reliability of a relay system. Since RS is attached to the core network with relay link as wireless backhaul, its reliability is less than typical wired backhaul. Reliability may be a major issue for low-cost small-cell RSs. Ideally, radio link failure prevention and recovery at RS should be handled transparent to the attached MSs. With the mobile relay framework,

an RS that is about to experience or currently experiencing radio link failure on relay link is capable of re-establishing backhaul connection with another suitable neighbor BS using a handover procedure, and this just appears as a scheduling glitch for all MSs associated with RS. Therefore, such self-healing wireless backhaul operation significantly improves network reliability without incurring any special handling at MS.

Wireless Backhaul Load Balancing: Similar to the use case of enhancing relay network reliability, the network may even more aggressively switch point of attachment for the RS based on operation status, such as loading of different BS and the associated network gateway. The network may initiate handover for RS to a more suitable BS to ensure load balancing within radio access network as well as core network. Again with mobile relay framework, the network can dynamically perform such operation without impacting the connectivity of MSs associated with the RS. This provides a very attractive feature for network operators.

2.2.2.2 Mobile Relay Architecture

A non-transparent relay standard has been specified for 4G systems by IEEE 802.16m wherein the advanced relay station (ARS) has the features of distinct physical layer cell IDs in each of the sectors it controls. The 16m relay functions using the "decode and forward paradigm where an ARS controls its own cell with a wireless backhaul connection to the access service network gateway (ASN-GW) through the advanced base station (ABS). 802.16m supports both time-division duplex (TDD) as well as frequency-division duplex (FDD) modes for relaying.

Figure 35 shows the interface architecture of a 4G 16m relay [59]. It can be observed that there are several reference points that define the protocols and procedures between the different entities of the 802.16m system [60]. R1 interface provides radio link between base station and mobile stations. R6 and R8 reference points provide the ABS ⇔ ASN-GW and ABS ⇔ ABS interfaces respectively. R3 reference point provides IP connectivity for the ASN-GW from the Connectivity Service Network (CSN). The relay station incorporates

both ABS as well as AMS functions. The AMS part of the relay station has an R1r interface with the ABS whereas the ABS part of the relay station has R8 and R6 logical connections with the ABS and ASN-GW respectively. The ARS ⇔ AMS PMP connectivity is provided by the ABS part of the relay station using R1a interface. The relay station uses R6 interface to communicate with the ASN-GW. Since there is no physical link between the relay station and the ASN-GW, the relay uses the two-hop ARS ⇔ ABS ⇔ ASN-GW physical link to communicate with the network. The access and relay R1 links are referenced using the notations R1a and R1r respectively to underline their functional difference. One of the key features of the IEEE 802.16m relay architecture is that the relay station implements all the control mechanisms of the associated mobile stations such as handoff, security, idle and sleep operations, etc.

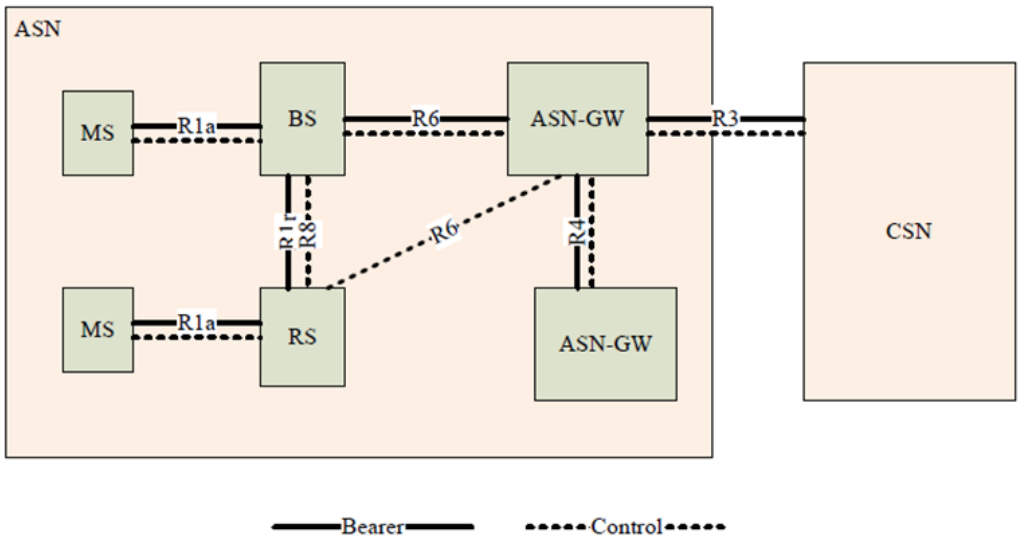

Figure 35: Relay architecture in 4G WiMAX 802.16m systems.

Figure 36 shows the C-plane protocol architecture. On receiving the control messages from ASN-GW over R6, the ABS performs classification to recognize that the received ASN packets are ARS (the ABS part) related control messages. The ABS, then, translates the control messages between the two interfaces by encapsulating them in a “MAC-L2-XFER MAC management message and sends it to the target ARS with FlowID = 1 [61].

Similarly, on the uplink, the ARS sends the control message as MAC-L2-XFER message with FlowID = 1. On the relay link, MAC Control PDUs are used for control message exchange between ARS and ABS. The R8 control messages between the ABS and ARS, similarly, are transferred as MAC-L2-XFER messages over the physical R1r link.

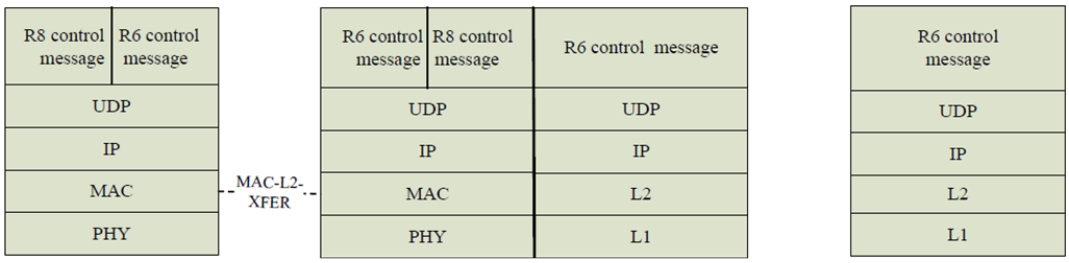

Figure 36: C-plane protocol stack for relay support in 4G WiMAX.

Figure 37 shows the U-plane protocol architecture. The GRE tunnels running frosm the ASN-GWs are terminated at the ABS. The AMSs related packets are then detunneled and encapsulated into a relay MAC PDU (MPDUs) of the ARS where the advanced relay forwarding extended header (ARFEH) may be appended to identify the ASN data traffic. The relay MPDUs encapsulated with the ARFEH header is indicated by GRE* in the figure. The mapping between the GRE tunnel IDs and ARFEH headers is maintained both at the ABS and the relay station so that the R6 data function effectively is running between ARS ⇔ ABS ⇔ ASN-GW. The ARS decapsulates the received relay MPDU and transmits the ASN data traffic to the target AMSs as AMS MPDUs. Similarly, on the uplink, the ARS encapsulates the data traffic from different AMSs into a relay MPDU. Again, the ARFEH may be appended to identify the ASN data traffic. Finally, the ABS maps the data packets to the AMS specific GRE tunnel that runs from the ABS to the ASN-GW.

2.2.2.3 Group Handover Scheme for Mobile Relays

As illustrated in Figure 34, small cells mounted on vehicles with wireless backhauls (for e.g., mobile relay stations) can enable group handovers of such in-vehicle users. Through this approach, a single group handover procedure can ensure handover of a group of users served by the mobile relay station (MRS) between macrocell base stations (BS). Group

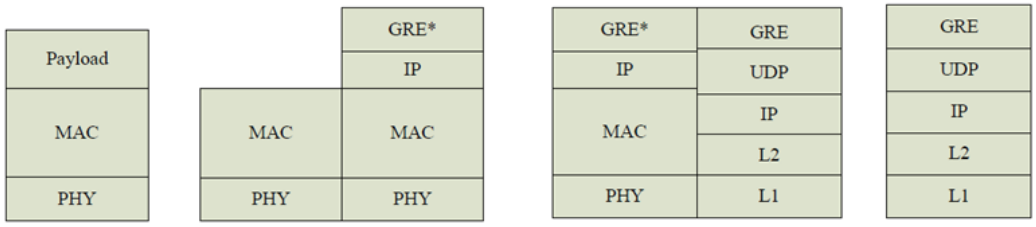

Figure 37: U-plane protocol stack for relay support in 4G WiMAX.

handover can significantly reduce the overheads on the radio interface and network thus minimizing latency for all users.

To this end, a transparent group handover mechanism and handover architecture is proposed in this work for IMT-Advanced systems including IEEE 802.16m (WiMAX) and 3GPP LTE-Advanced. The MRS is considered to be "decode and forward" capable with a distinct physical layer cell ID(s) and able to connect to the core network (CN) through a BS. The MRS also implements all the control mechanisms for the attached mobile stations such as handoff, security and sleep operations. The MRS incorporates both base station and user functions. The user part of the MRS connects to the BS like other MSs. The base station part of the MRS provides point-to-multipoint connectivity for regular MSs.

Group Handover Mechanism: First, we focus on the handover (HO) procedure for the mobile relay between the serving and target ABSs. For simplicity, we only consider intra-ASN handover in this work. The handover mechanism is illustrated in Figure 38 .

The handover and network re-entry process flow for ARS is as shown below:

1. Obtain network topology and neighbor ABS parameters
2. Initiate ARS handover to target ABS (either ARS or ABS initiated)
3. Perform network re-entry at target ABS
4. Configure operational parameters

Except for the final step, where the ARS needs to obtain the configuration to provide PMP connectivity to the AMSs, the handover framework of ARS closely resembles the

handover procedure for a regular mobile station. The access network handover and network re-entry procedures are described below.

Access Network Handover: The handover can be either user-initiated (for e.g., IEEE 802.16m) or network-initiated with user assistance (3GPP LTE-A). We describe the user-initiated group handover process below:

- The ARS sends a Handover Request Message with a list of preferred target base stations to the serving ABS.
- The serving ABS, in turn, sends an R8 Handover Request message to the target ABS(s).
- The target ABS(s) obtains AK Context and initiates data path pre-registration for the ARS with the ASN-GW over R6. The target ABS also classifies the Handover Request message to identify that handover is required for a relay station.
- If handover for the ARS is accepted, the target ABS must send an R8 Handover Response message to the serving ABS. But the target ABS needs the tunnel mapping information to perform GRE Tunnel ID $\Leftrightarrow$ ARFEH mapping to identify the per-AMS ASN tunnels and their corresponding QOS parameters. Hence, the target ABS would piggyback a Tunnel Mapping Context Request message requesting the tunnel mapping table and per-tunnel QOS over the Handover Response message.
- The serving ABS sends a Handover Command message over the relay link to the ARS to inform the ARS about the handover decision.
- The ARS starts ranging at the target ABS to begin the network re-entry procedure. For optimization, if the HO Reentry Mode is set to 1 in the Handover Command message sent to ARS, the serving ABS can ensure that data path is available for the ARS related AMSs until the ARS completes network re-entry at the target base station.

Core Network Re-entry: The core network re-entry phase is described as follows:

- The ARS initiates the Ranging Request/Response message exchange with the T-ABS.
- A Handover Confirmation message is received by the target ABS from the serving ABS. This message includes the piggybacked Tunnel Mapping Context Response message so that the target ABS can perform data path pre-registration with the ASN-GW over R6. The ASN-GW may either set-up a brand new tunnel with the target ABS for the corresponding service flow and break the GRE tunnel with the serving ABS; or it may reuse the same GRE tunnel for the service flow and update its Tunnel forwarding port to be the target base station. The latter case is straight forward where the ARS can establish data path directly with ARS during its re-entry. In the former case, the target ABS needs to perform Data Path Reg/Update with the ARS over R8 for the ARS to update the new tunnel mapping context.
- The data path registration procedure ends when the target ABS receives a Path Registration Response message from the ARS.
- For the ARS to support relay operation at the new serving ABS, a Layer-3 control path from the ASN-GW to the ARS is also established to update configuration from the OAM (Operations, Administration and Maintenance) server.
- The serving ABS, after sending the Handover Confirmation message may discard all connection resource information (ARS and related AMSs) including the MAC state machine and all outstanding buffered PDUs.
- The Handover Complete message from the target ABS indicates the completion of Network Re-entry which prompts the serving ABS to release all MAC context and MAC PDUs associated with ARS. Following this, the serving ABS initiates Data Path De-registration with the ASN-GW.

- Finally the ARS receives the PHY layer operational parameters from the target ABS. Any changes in the operational parameters are communicated to the AMSs by the ARS over the access links.

Re-establishment of Mobile Relay Station Wireless Backhaul: The handover architecture for a mobile relay is illustrated in Figure 39. The serving base station maintains a table containing the mapping between the ARFEH ID and the GRE tunnel ID for each of the per AMS service flows. The figure shows a sample table entry that needs to be exchanged between the base stations during ARS handover. MS Info [=AMS1] includes AMS information such as MSID, service flow info. DP ID [=5B9F9C71] is the 32-bit Data Path identification for the per AMS GRE Tunnel. ARFEH ID [=8FE] is the compressed 12-bit identification for the per AMS service flow between base station and relay station. QOS parameter [=X] contains the QOS description of the service flow.

On the Control plane, the Tunnel Mapping Context Request/ Response over R8 helps the target base station to perform Data Path Registration for ARS associated AMSs. The target base station performs Data Path Pre-Registration Procedure with the ASN-GW. Here, we assume the more common case where the ASN-GW establishes a new GRE tunnel [=A70BF54C] with the target base station for each of the service flows and breaks the GRE tunnels with the serving ABS. The target base station assigns a new ARFEH [=D84] for these per-AMS tunnels for identification. The target ABS also needs to perform Data Path Registration Update with the ARS over R8 for the ARS to update the new tunnel mapping table with the ARS. To enable OAM Configuration at the relay station, a Layer-3 control path is established between the ASN-GW and the ARS.

On the user plane, the ASN-GW establishes a new per-AMS GRE tunnels with the target base station. Once the appropriate tunnel mapping is achieved, data path is established with the relay station. For out-of-band relays, the data path from ASN-GW to the AMSs through the serving ABS is maintained until the network re-entry is complete. For in-band relays, advanced handover mechanisms [62] such as Entry-Before-Break can ensure that

this path is available until the completion of network re-entry. After data path registration and L3 Update from the OAM server, the end-to-end data paths from the ASN-GW to the AMSs are restored through the target base station and relay station. Thus we could observe that the entire handover procedure has been made transparent to the AMSs.

Radio Link Handling for Seamless In-band Relay Mobility: One of the important aspects of mobile relay is to reestablish a wireless backhaul between relay station and the target base station transparent to the mobile stations. We have assumed that the relay station maintains a logical connection with mobile stations during ARS handover. Although for inband relay mobility case, the AMS may experience Radio Link Failure if the relay station handover procedure lasts for a long time. Therefore, the relay station should be notified in the Handover Command message of the superframe structure used by the target base station for the Access and Relay links before handover for ARS takes place so that

- The RS knows the relay zone sub-frames used by the target BS so that it can synchronize its Relay Zone configuration with the target BS to acquire the target BS.
- The RS may continue RS mode operation with the MS using the appropriate Access Zone sub-frames.

Alternatively, the ARS may be notified by the serving ABS about the target ABS frame structure during the handover procedure using R8 messages or directly from the OAM server before handover. The detailed mechanism to enable the above procedure is beyond the scope of this work.

2.2.2.4 Performance Evaluation

We compare the performance of mobile relay with "relay with no mobility support" and "no relay" scenarios. For each of these three scenarios, we perform a comparison of overheads on two levels, the air interface and the network. On the air interface, overheads on R1a, R1r including R6, R8 logical links are considered. On the network side, R6/R4 and R8 overheads are considered for comparison.

For the "No Relay" scenario, all the mobile stations need to perform handover from the serving base station to the target base station in the absence of relay station. For the "Relay with no mobility support" case, first, the relay station has to perform handover of the AMSs to a target base station. Then, it performs de-registration with the serving base station. Once again, when the relay station resumes normal functionality by establishing wireless backhaul with the network, the AMSs will be able to attach to the ARS by performing the routine handover procedure. Hence the overheads are three-fold in this case; handover of AMSs to a target ABS, backhaul connection re-establishment for ARS, AMSs handover back to ARS. Finally, for the "Mobile Relay" case, only the relay station performs handover from the the serving base station to the target base station. After the completion of network re-entry, ARS receives the configuration from the OAM server. Similarly, the target base station updates the PHY configuration at the relay station, which in turn, may update the PHY configuration with the AMSs over the access link.

We consider the overheads due to the mobility of N mobile stations. The number of control messages is scaled for N users. The different Radio and Network Overheads are listed in Table 5. On the air interface, it is very clear that "mobile relay" scenario requires negligibly small number of control messages for handover to take place, especially, as N becomes large. On the network side, again, the overheads are reduced considerably in the case of mobile relay. The establishment of Layer 3 connectivity with the OAM server to configure the ARS as a base station may contribute to overheads but hardly has any impact on the latency. Hence, it becomes less relevant. Figure 40 compares the overhead for the case when N = 100. As an optimization for mobile relay, the serving ABS may also perform de-registration for all ARS-related AMSs in one step. This is possible since the serving ABS is capable of classifying the GRE tunnels that correspond to ARS-related AMSs. Thus, overall, we can observe that mobile relay results in significant overhead reduction both on the radio interface and the network. This also essentially means that the handover latency is largely reduced thereby providing seamless mobility experience for

high speed users.

Table 5: Group handover cost functions.

Parameter	**No Relay**	**No Group handover**	**Group handover**
Radio Connection Setup	N	(2*N)+1	1
Handover Request	N	(2*N)+1	1
Handover Response	N	(2*N)+1	1
Security Context Update	N	(2*N)+1	1
Data Path Setup	N	(2*N)+1	N+1
Tunnel Context Exchange	NA	NA	Required

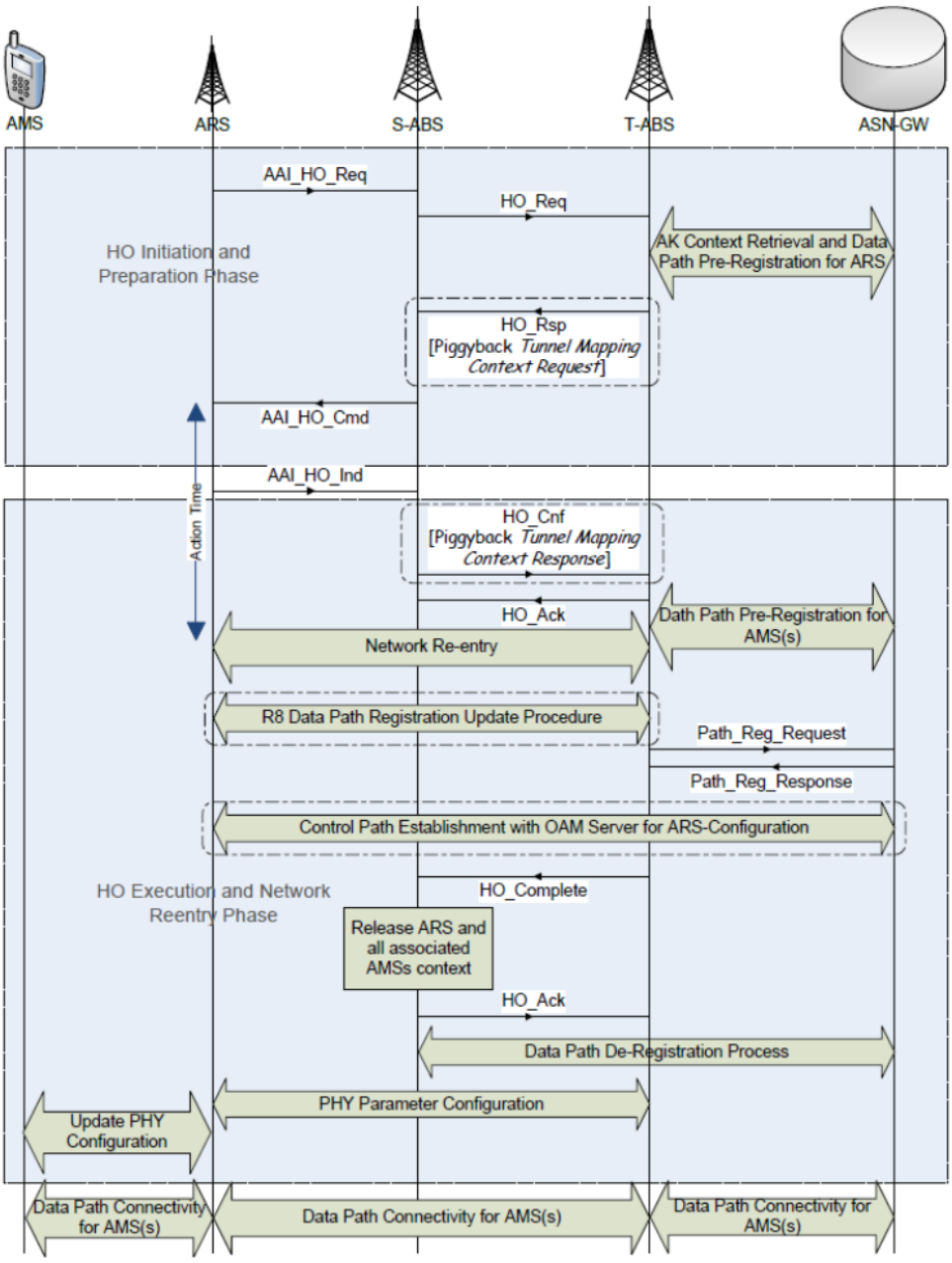

Figure 38: Proposed group handover mechanism for 4G WiMAX systems.

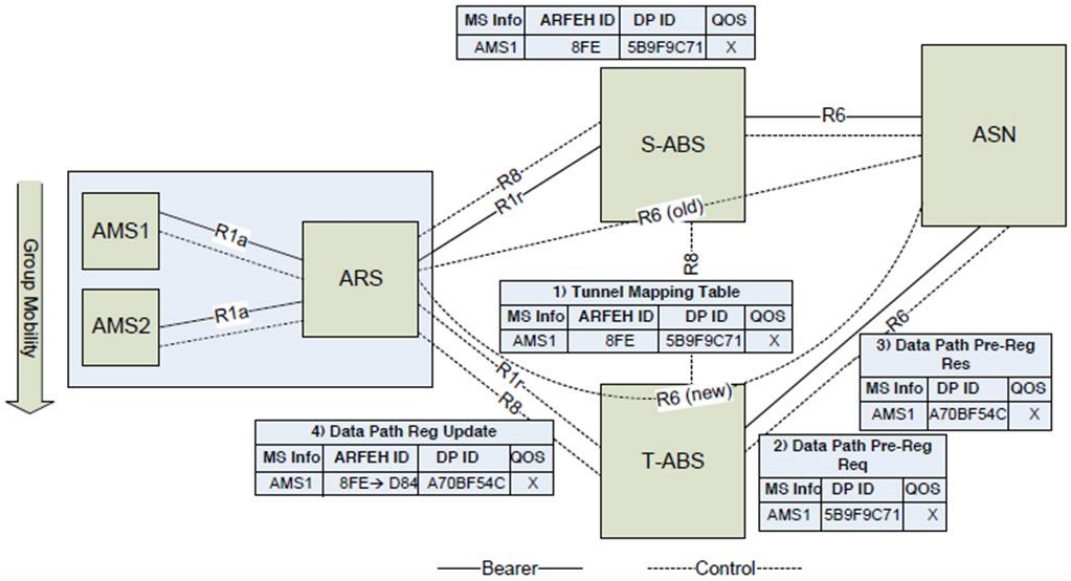

Figure 39: Proposed group handover architecture.

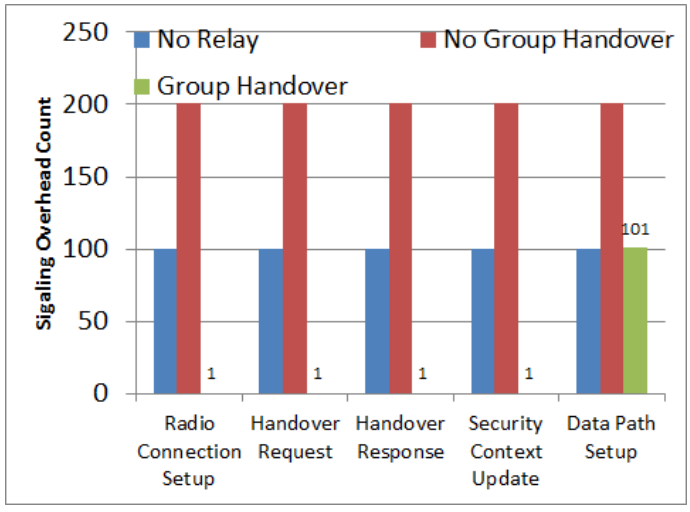

Figure 40: Signaling overhead for group handover.

CHAPTER 3

HANDOVER MANAGEMENT IN MULTI-STREAM HETNETS

3.1 Joint Small Cell Discovery and Component Carrier Selection for Multistream HetNets

HetNets are particularly prone to poor mobility handling due to the number of cell edges arising from the heterogeneous architecture. Recent studies on HetNet handover performance show that the handover failure rates and pingpong effects are far greater in HetNet environments compared to homogeneous deployments [32]. There are several factors that lead to poor handover performance in HetNets. Firstly, the unique characteristics of HetNets such as the cell sizes, number of cross-overs and access control policies are generally not considered when performing user mobility state estimation. In addition, there is a lack of efficient small cell discovery mechanisms to quickly identify inter-frequency small cells for handover. Furthermore, the existing handover recovery procedures are not sufficiently robust to deal with the high number of failed handovers.

In homogeneous deployments, small cell discovery (or measurements) must be performed in specific cases such as when users reach cell boundaries or when load balancing handover is required between carriers. However, in HetNets, these measurements would always have to be performed to effectively utilize the offloading opportunity whenever it occurs. On the other hand, continuously performing intra/inter-frequency measurements will lead to high energy consumption and lower throughput. The key attributes of an efficient small cell discovery approach, therefore, include (i) low-UE power consumption, (ii) minimal interruption due to measurements on the serving cell, and (iii) no degradation of the mobility performance due to measurements.

Several heuristic approaches for inter-frequency small cell discovery have been proposed towards LTE-A Release 11 standardization in [63, 64, 65, 66]. A cache-based measurement scheme was proposed in [67] that maintains a list of recently visited small cells

which will be prioritized for performing measurements. However, this approach does not account for the users mobility pattern while selecting the recently visited list. A user autonomous inter-frequency measurement approach with several small measurement gaps is proposed in [68]. However, this scheme causes burst of small interruptions in addition to packet losses, since the base station is unaware of the measurement gap used by the user.

Intelligent handover decision algorithms for HetNets have been proposed in [35, 36, 37, 38] that, in general, aim to minimize the number of handovers and handover failures. In our earlier work [51, 52], we have considered the case of a cluster of small cells and proposed a novel local anchor-based architecture to improve the overall handover performance. In spite of this, the current HetNet architecture will invariably result in either high handover failure rates or handover related signaling.

Carrier aggregation is another efficient technique for operators to combine and operate separate and often non-contiguous blocks of spectrum in order to achieve the capacity and data rate requirements of 5G systems. LTE-Advanced has considered aggregating several bands from different frequencies up to 100 MHz for 4G systems. Each of these bands are referred to as component carriers (CCs) and LTE-Advanced allows aggregation of up to five CCs with each occupying a bandwidth of up to 20 MHz. Carrier aggregation can utilize both contiguous CCs as well as non-contiguous CCs from different bands. In the case of non-contiguous aggregation, components carriers can be aggregated in an intra-band (from within the same band) or an inter-band (across different bands) fashion. With new frequency bands expected to be included for 5G systems including the 5 GHz band, inter-band non-contiguous will likely play a major role in realizing the performance objectives of 5G systems.

3.1.1 Multistream HetNet Architecture

Multi-stream carrier aggregation (MSCA) is a key enabling technology for providing a seamless and no-edge experience for high-speed users in future LTE-A as well as 5G cellular systems. In this scheme, component carriers can be aggregated for a user over several

contiguous or non-contiguous frequency bands across different base stations simultaneously. The key application of this is that the larger macrocell, also termed as the Primary cell (PCell) or the primary component carrier (PCC), can provide coverage for the users while the smaller cells acting as secondary cells (SCells) or the secondary component carriers (SCCs), can provide significant capacity boost. Such an architecture requires the macro and small cells to be interconnected using a high-speed backhaul.

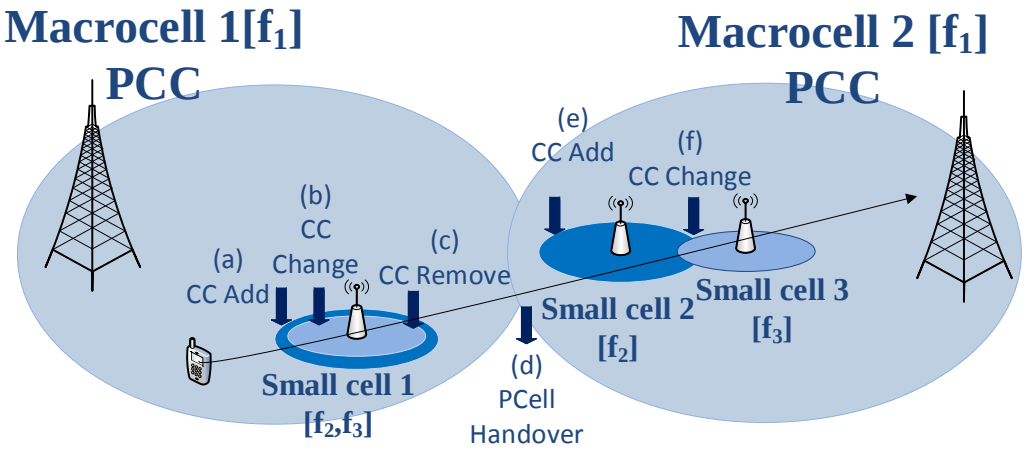

Figure 41: Multistream architecture.

The multistream HetNet architecture is illustrated in 41. The mobile user's movement implies several mobility cases. The macrocell 1 operating on frequency f_1 acts as the PCell or PCC for the user. During the instance (a), the CC f_2 on the small cell 1 is dynamically added for the user with the help of high-speed backhaul link between macrocell 1 and small cell 1 to supplement the PCC link. However, at position (b), the user may change SCC attachment from f_2 to f_3 since it offers a higher data rate or for load balanching purpose. At position (c), the user moves outside the coverage of the SCC on f_3 and hence the SCC is removed. At position (d), the user undergoes a full handover from PCC 1 to PCC 2. Once again, when the user enters the coverage of small cell 2 at position (e), the SCC on f_2 is added which is later changed at location (f) when the user moves to the small cell 3 coverage with carrier f_3.

3.1.2 Handover Challenges in Multistream HetNets

To effectively achieve this seamless or no-edge experience of users, it is important to address the key challenges related to MSCA. As SCCs will have to be activated or deactivated for users on the go, the key challenges in multi-stream carrier aggregation include (i) inter-frequency small cell discovery, and (ii) Component carrier (CC) selection. These are explained as follows:

Inter-frequency Small Cell Discovery: The emergence of carrier aggregation requires the users to perform additional inter-frequency measurements on neighboring small cells not only for offloading opportunities, but to achieve a higher data rate. The user has to disconnect from the serving cell in order to perform measurement of other CCs, and therefore the user cannot be provided any service during this acquisition period. The service interruption time due to this acquisition is called the measurement gap.

The future releases of LTE-Advanced are expected to introduce higher frequency operations including the 3.5GHz band and > 10GHz bands and carrier aggregation can be performed over all of these available bands. This is illustrated in Figure 42. Each band consists of several component carriers. With the latest advances in device hardware, the users will be able to aggregate component carriers not only from contiguous bands, but also across bands that are several GHz of frequencies apart.

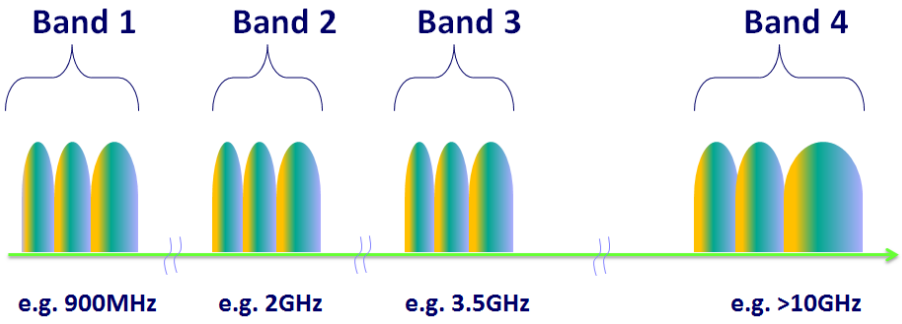

Figure 42: Inter-band carrier aggregation for beyond 4G systems.

As the number of possible component carriers increases, performing inter-frequency measurements of all the possible CCs becomes extremely challenging. The time to identify

inter-frequency carriers $T_{identify}$ is given in [69] as

$$T_{identify} = T_{Basic\ Identify} * 480 * N_{freq}/T_{Inter}, \tag{52}$$

where T_{Inter} is the minimum time available for inter-frequency measurements, N_{freq} is the number of CCs on which inter-frequency measurements are required, $T_{Basic\ Identify}$ is a fixed time duration and approximately equal to 480ms. When T_{Inter} is set to 30ms and for $N_{freq} = 2$, the $T_{identify}$ is $7.68s$. Such long measurement times may result in several failed handovers and missed opportunities for CC selection. Beyond the missed offloading opportunities, frequent measurements also leads to high energy consumption at the user devices.

On the other hand, it is also challenging to determine the best CCs for measurement with the objective of maximizing the throughput for the users. Intuitively, the probability of identifying CCs with higher available capacity increases as the number of CCs measured increases.

Therefore, an optimal measurement or small cell discovery strategy is required to minimize the overall measurement time and maximize offloading opportunities and consequently the user throughput.

Component Carrier selection: It is also an open issue on how to add or remove the SCells (or secondary component carriers) dynamically in the network while incurring the minimum signaling load but maximizing the data rate for the users. The introduction of MSCA results in several SCell handover cases as illustrated in Figure 43. These are itemized as follows:

- **Intra-SCell Intra-band handover (Type A):** Users undergo handover from one CC of the SCell to another CC both within the same band.
- **Intra-SCell Inter-band handover (Type B):** User undergoes handover from one CC of the SCell to another CC on a different frequency band.
- **Inter-SCell Intra-band handover (Type C):** Users are switched from one SCell to another SCell both operating on the same band.

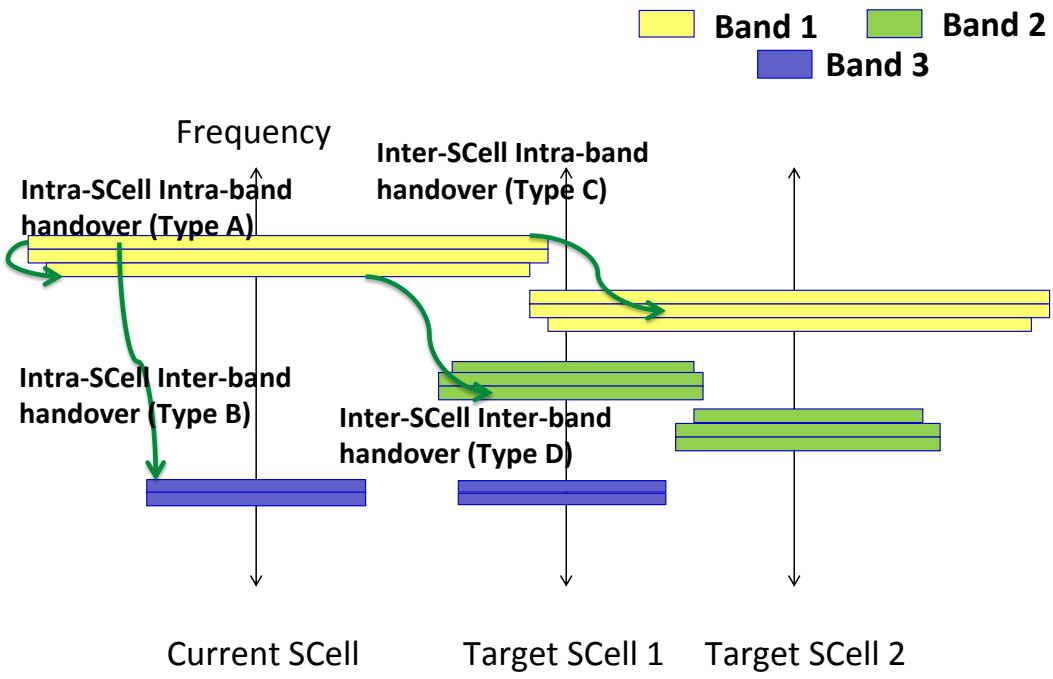

Figure 43: Handover scenarios in multi-stream HetNets.

- **Inter-SCell Inter-band handover (Type D):** Users are switched from one SCell to another SCell both operating on different bands.

As evident from the above SCell handover cases and in contrary to the traditional handover at the cell edges, the users will have to dynamically evaluate component carriers to attach or detach across different frequency bands and SCells. This is because the selection of CCs will impact how frequently users experience handovers, cost and delay associated with handover across CCs. Therefore, an optimal strategy is required to determine the CC selection such that the cost associated with CC handover is minimized. Recently, a user autonomous component carrier addition and removal approach was proposed in [70]. Although a user autonomous approach offers benefits of signaling minimization, the interdependence of inter-frequency measurements and component carrier selection is not exploited to minimize the rate of change of CCs in this approach.

In this chapter, we treat the small cell discovery and component carrier selection for multistream HetNets as a coupled problem. We propose and evaluate a joint small cell discovery/measurement and component carrier selection approach to minimize the service interruption, energy consumption as well as the signaling load accompanied with these critical tasks.

3.1.3 Joint Small Cell Discovery and Component Carrier Selection

The issue of small cell discovery and CC selection will be addressed in this work by continuously learning from the network dynamics and past measurement and CC selection actions. In this work, we only focus on the SCell handover and consider that the user can maintain a connection with a primary cell (PCell) during SCell handovers.

3.1.3.1 Restless Multiarmed Bandit Formulation

The above problem is formulated as a restless multi-armed bandit (RMAB) problem where the objective is for the user to periodically determine the set of CCs $\mathcal{M}_1 \in \mathcal{K}$, such that $|\mathcal{M}_1| = m_1$, for performing measurements and another set of CCs $\mathcal{M}_2 \in \mathcal{M}_1$, such that $|\mathcal{M}_2| = m_2$, for component carrier selection such that the expected total system reward is maximized [71]. The set $\mathcal{K} = \{1, 2, ..., K\}$ is composed of the component carriers supported by the small cells in the system besides the PCC of the user.

Each CC is described as a discrete-time irreducible and aperiodic Markov process with the state space indicated by $\mathcal{N} = \{S_1, S_2, ..., S_{N+1}\}$. At any discrete-time t, $X_k(t)$ indicates the state of CC k. The state evolution of CC k is described by a discrete-time Markov process given by

$$X_k(t+1) = f_{k,t}(X_k(0), ..., X_k(t), A_k^1(t), A_k^2(t)), \tag{53}$$

where $A_k^1(t) \in \{0, 1\}$ and $A_k^2(t) \in \{0, 1\}$ are the control actions indicating if the user performs measurement and access on CC k at time t respectively.

$$A_k^1(t) = \begin{cases} 0 & \text{if CC k not measured in time t,} \\ 1 & \text{if CC k measured in time t.} \end{cases} \tag{54}$$

$$A_k^2(t) = \begin{cases} 0 & \text{if CC k is not selected in time t,} \\ 1 & \text{if CC k in selected time t.} \end{cases} \tag{55}$$

Each state of the system $X_k(t)$ is associated with a stationary and positive reward indicated by $R(X_k(t), A_k^1(t), A_k^2(t))$, for each of the possible action set $A_k^1(t)$ and $A_k^2(t)$.

A joint policy $\Phi := (\Phi_1, \Phi_2, ...)$ is a decision rule such that the control actions $A_k^j(t)$ for $j \in \{1, 2\}$ at time t are random variables taking values in $(d_1, d_2, ..., d_{\binom{K}{m_j}})$ where each d_i is a K-dimensional row vector with m_j ones and $(K - m_j)$ zeros given the current state $X(t)$ of the CC and the associated reward $R(X_k(t), A_k^1(t), A_k^2(t))$. Here, m_j is the number of CCs selected for the control action $A_k^j(t)$.

Then, the objective of the restless MA bandit problem is to determine a joint policy $\hat{\Phi}$ that maximizes

$$J^{\hat{\Phi}} := \lim_{t\to\infty} E^{\hat{\Phi}}\left[\sum_{t=1}^{T} \sum_{k=1}^{K} R(X_k(t), A_k^1(t), A_k^2(t)) \;\middle|\; Z(0) \right], \tag{56}$$

where $Z(0) = [X_1(0), X_2(0), ..., X_K(0)]$ is the initial state of the system. Computing the optimal solution for the above restless multi-armed bandit problem is shown to be PSPACE-hard in [72].

3.1.3.2 Modeling CC States

$X_k(t)$ represents the state of the user with respect to CC k at time t. In our modeling, each state represents the relative distance of the user from the cell with CC k in discrete steps. We utilize the metric effective transmission range as originally defined in [73] that accounts for the stochastic channel variations to determine the CC's transmission range. Therefore, we have a discrete set of states for each user with respect to each CC k.

The effective transmission range is a stochastic metric and is defined as the maximum value of the transmission range which holds the condition of $P_r^k(dB) \geq P_0^k(dB)$ with a very

high probability (typically 0.99). Here, $P_r^k(dB)$ is the received signal power from the CC k and $P_0^k(dB)$ is received power threshold from CC k.

Therefore, the effective transmission range or distance D of CC k is given by

$$D = 10^{\frac{-2.33\sigma_S - 10\log E\chi^2 + c}{10\alpha}} \tag{57}$$

where α is the path loss exponent, σ_s represents the standard deviation of the shadowing correlation and χ^2 represents multipath fading. The constant function $c = P_t(dB) - L_0(dB) - P_0(dB)$ such that $L_0(dB)$ is the average path loss $1m$ away from the transmitter.

Once the effective transmission range is determined, it is divided into N discrete states in steps of ϵ_k meters. The value of ϵ_k is determined as $\epsilon_k = D_k/N$. Now, we characterize the user states corresponding to the CC k. Each user can exist in any of the $N+1$ discrete states indicated by $\{S_0, S_1, ...S_N, S_{N+1}\}$ where S_{N+1} represents the region outside the effective transmission range. Between time steps of duration Δt, the user can moves from location $(x_{k,t}, y_{k,t})$ to $(x_{k,t+\Delta t}, y_{k,t+\Delta t})$. This is illustrated in Figure 44.

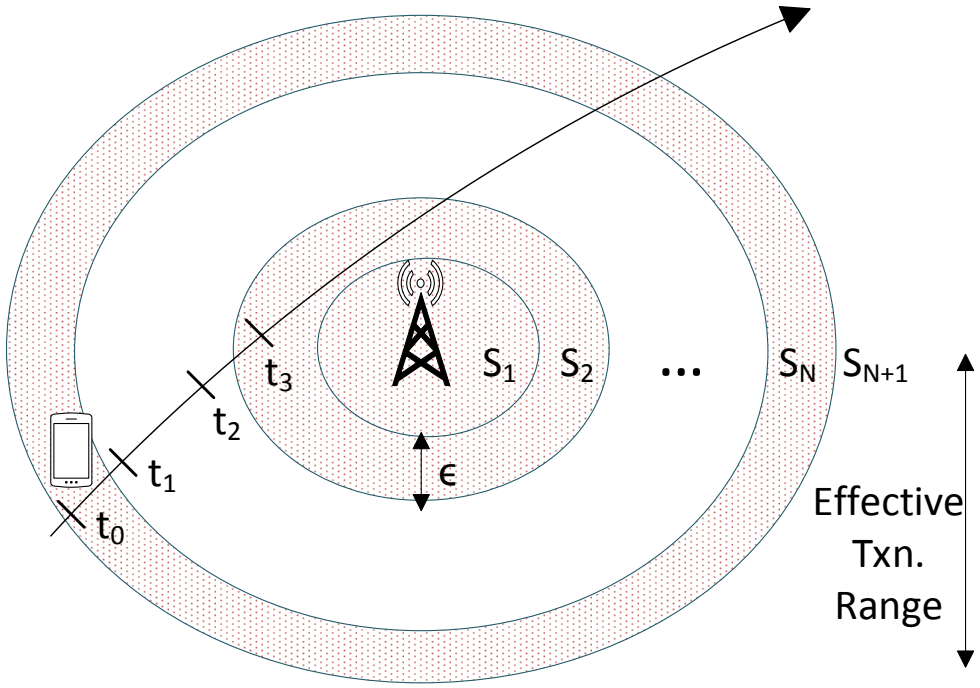

Figure 44: Modeling the CC states.

The state transitions are governed by the stochastic user mobility. User mobility is

described by two random variables, namely velocity and angle. Given $(x_{k,t}, y_{k,t})$ is the location of a user with respect to CC k located at the origin, the user distance $\rho(k,t)$ from CC k is given by $\rho(k,t) = \sqrt{x_{k,t}^2 + y_{k,t}^2}$ where $x_{k,t}$ and $y_{k,t}$ are Gaussian distributed random variables. Then the user velocity $v_k(t)$ is given by

$$v_k(t) = \frac{\sqrt{(x_{k,t} - x_{k,t-\Delta t})^2 + (y_{k,t} - y_{k,t-\Delta t})^2}}{\Delta t}. \tag{58}$$

The pdf of the user velocity $f_V(v))$ is described by the Rayleigh density function

$$f_V(v) = \frac{\pi v}{2\overline{V}^2} e^{\frac{-\pi v^2}{4\overline{V}^2}}, \tag{59}$$

where $\overline{V}$ is the average velocity and Δv is the maximum deviation of $\overline{V}$.

The angle $\psi_k(t)$ with respect to CC k is uniformly distributed in the range $[0, \pi)$ and is given by

$$\psi_k(t) = arccos \frac{\rho_{k,t-\Delta t}^2 + v_t^2 \Delta t + \rho_{k,t}^2}{2\rho_{k,t-\Delta t} v_t \Delta t}. \tag{60}$$

Markov Model for State Transition: For the user velocity $v_k(t)$ and angle $\psi(t)$, the state transition is described by a discrete-time Markov model and is shown in Figure 45. The state transition probabilities are indicated by $P_{ij} = Prob\{\rho(t) \in S_j | \rho(t - \Delta t) \in S_i\}$ and are approximated following the work in [73] as

$$\tilde{P}_{ij} \approx \frac{0.2\epsilon}{\overline{V}} \sqrt{\frac{2j-1}{2i-1}} \log \frac{|(\overline{V} + \delta_V)^2 - \epsilon^2 (j-i)^2 | (i+j-1)^2}{|\epsilon^2 (i+j-1)^2 - (\overline{V} + \delta_V)^2 | (j-i)^2}, \tag{61}$$

$$P_{ij} = \frac{\tilde{P}_{ij}}{\sum_j \tilde{P}_{ij}}. \tag{62}$$

3.1.3.3 Sufficient Statistic

The CC state $X(t)$ cannot be determined exactly unless the user performs measurement over the CC. In other words, the state of CC k can be directly observed only when $A_k^1(t) = 1$. However, when $A_k^1(t) = 0$, it is possible to infer the CC state based on the decision and observation history. Therefore, the current state of information regarding the internal state of the CC can be maintained using the belief vector (or information vector) $\mathbf{\Omega}_k(t) =$

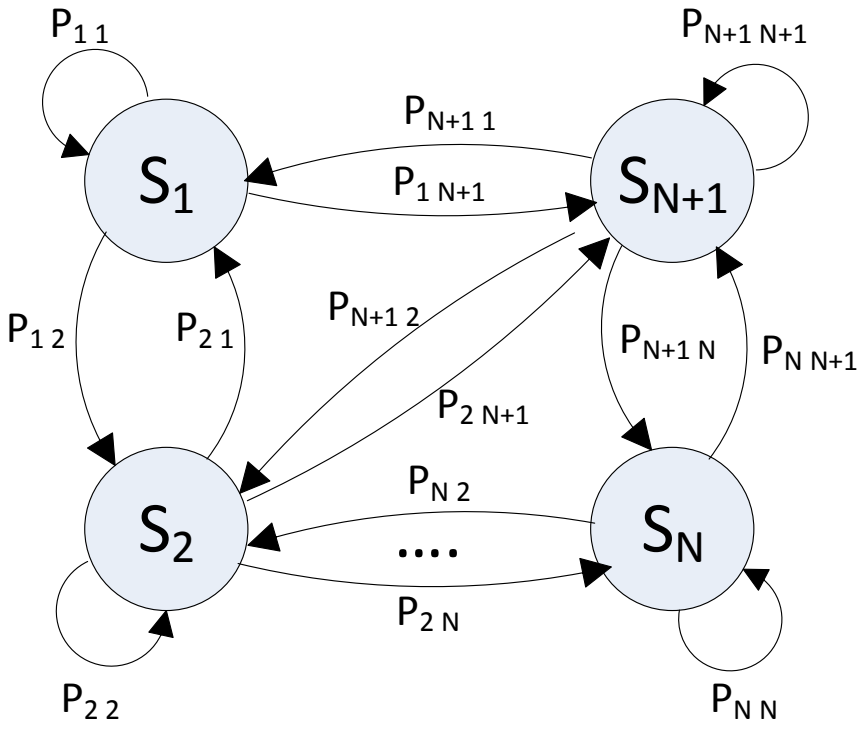

Figure 45: Markov model for CC state transition.

$[\omega_{k,1}(t), ..., \omega_{k,N+1}(t)]$ where $\omega_{k,j}(t)$ is the belief state indicating the conditional probability that the CC is in state j with respect to the user given all past decisions and observations.

An important and intuitive property of the belief vector is that $\sum_{j=1}^{N+1} \omega_j(t) = 1$. Therefore, we can represent the set of all belief vectors as all the points on the hyperplane $\sum_{j=1}^{N+1} \omega_j(t) = 1$ of an N+1-dimensional space. If no information is available on the initial state of CC k, the belief vector at time $t = 0$ indicated by $\mathbf{\Omega}_k(0) = [\omega_{k,1}(0), ..., \omega_{k,N+1}(0)]$ is given as the stationary probability distribution of the underlying Markov chain of the CC.

If the prior belief vector of CC is given by $\mathbf{\Omega}(t)$ and the past control actions are given by $A^1(t)$ and $A^2(t)$, then we have the updated belief vector indicated by

$$\mathbf{\Omega}(t+1) = \mathcal{T}(\mathbf{\Omega}(t)|A^1(t)). \tag{63}$$

The transformation $\mathcal{T}$ transforms a point in the hyperplane of belief vectors for the

given time period into another point on the hyperplane for the next time period. Furthermore, each of the present actions will result in a unique transformation for the next time period.

Therefore, we have a different updated belief vector corresponding to each possible measurement action. When measurement action is performed, i.e., $A^1(t) = 1$, the updated belief vector $\mathbf{\Omega}(t+1)$ is determined from the current state $X(t)$ and the transition probability matrix P as

$$\mathbf{\Omega}(t+1) = \mathcal{T}(\mathbf{\Omega}(t)|A^1(t) = 1) = \mathbf{e_j P},\ for\, X(t) = j, \tag{64}$$

where $\mathbf{P}$ is the state transition probability matrix and $\mathbf{e_j}$ is the standard basis.

Similarly, when the measurement action is not performed, i.e,, $A^1(t) = 0$, the belief update is determined from the past belief vector $\mathbf{\Omega}(t)$ and the transition probability matrix $\mathbf{P}$ as

$$\mathbf{\Omega}(t+1) = \mathcal{T}(\mathbf{\Omega}(t)|A^1(t) = 0) = \mathbf{\Omega}(t)\mathbf{P}, \tag{65}$$

From the above properties, the belief vector $\mathbf{\Omega_k}(t)$ can be readily shown to be a sufficient statistic for the past sequence of observations and decisions for the CC indexed by k. This is obtained by considering perfect observation of the CC state during measurement and using the Bayes' rule as applied in [74].

3.1.3.4 Reward function

The original reward function $R(X_k(t), A_k^1(t), A_k^2(t)$ corresponding to each CC state $X_k(t)$ now becomes the reward function corresponding to each point on the $N+1$ hyperplane of the belief vectors. Given that the knowledge of the CC states is maintained using the belief vector, the reward function indicates the immediate reward or gain obtained by the user for performing actions $A_k^1(t)$ and $A_k^2(t)$ for the CC given that the belief vector is $\mathbf{\Omega}_k(t)$ at that

time. In this work, we define the reward function $R(\mathbf{\Omega}_k(t), A_k^1(t), A_k^2(t)$ as

$$R(\mathbf{\Omega}_k(t), A_k^1(t), A_k^2(t)) = \begin{cases} 0, & A_k^1(t) = A_k^2(t) = 0, \\ 0, & A_k^1(t) = 1, A_k^2(t) = 0, \\ \sum_{j=1}^{N+1}(N+1-j)\omega_{k,j}(t)B_k & A_k^1(t) = A_k^2(t) = 1. \end{cases} \tag{66}$$

The reward function defined in Eqn.(66) shows that the user gains by performing measurement as well as component carrier selection on a CC. A higher value of $\sum_{j=1}^{N+1}(N+1-j)\omega_{k,j}(t)$ indicates that the conditional probability that the user is in the inner tiers of the CC k's coverage is higher compared to that in the outer tiers. The belief vector intrinsically accumulates the information gathered from the past measurement actions and therefore provides a reliable view of the CC's signal strength before performing measurement. Similarly, a greater reward is offered for selecting the CC that has a higher bandwidth to improve the data rate available for the user after attaching to the CC.

Although, a user does not accrue reward for performing measurement alone on a CC, the belief vector is updated taking into account the measurement information which improves the accuracy of the belief vector for future decisions.

3.1.3.5 Index Policy for the Joint Problem

For the class of the restless multiarmed bandit problems, dynamic programming that is based on backward induction suffers from dimensionality problem. Index policies, on the other hand, tackle the dimensionality problem by reducing an N dimensional problem into N independent 1 dimensional problems. These policies are also referred to as strongly indexable index policies.

When applied to the joint measurement and component carrier selection problem, a strongly decomposable index policy assigns an index to each state of a CC which is used to measure how rewarding it is to activate the CC for each of the N+1 states. Therefore, the index depends only on the attributes of the particular CC such as the state transition probabilities and the reward structure of the CC while decoupling the other CCs from the decision process.

Myopic Index based Joint Policy: For the proposed reward function and the belief vector, we propose a myopic index based joint policy which focuses on maximizing the expected immediate reward and ignores the impact of the current measurement and handover decisions on future rewards. Thus, the myopic index based policy is suboptimal and may show optimal performance when certain conditions are met. The proposed myopic index based joint policy for small cell discovery and component carrier selection is described as follows:

- **Step 1:** Each user u is associated with a primary cell that is typically the macrocell providing a large coverage area. The PCell maintains the belief vector $\mathbf{\Omega}_k^u(t)$ for user u. If no prior measurement information is available, the initial belief vector for CC k can be the stationary probability distribution of the CC k's Markov model described in Section 3.1.3.2.

- **Step 2:** Based on the belief vector, the PCell determines the expected immediate reward $\sum_{j=1}^{N+1}(N+1-j)\omega_{k,j}^u(t)B_k$ for all the K CCs in the network where B_k represents the bandwidth of the CC k. The previous measurement and access decisions are embedded on the belief vectors. The expected immediate rewards become the myopic indices for each of the K CCs.

- **Step 3:** The PCell assigns the myopic index to the K CCs based on the decreasing expected immediate reward they offer. The m_1 CCs that have the highest value of the expected immediate reward are selected for measurement reporting. Then the PCell sends a *measurement configuration message (RRC signaling)* using which it indicates the list $\mathcal{M}_1$ together with the measurement scheduling information.

- **Step 4:** The user upon receiving the control message performs measurement on the m_1 CCs as indicated by the PCell. At the end of the measurement period, the user sends the received signal strength information back to the PCell using the *measurement report* control message.

- **Step 5:** The PCell receives the measurement information from the user and uses this to update the belief vectors on the m_1 CCs as shown in Eqn. (64). For the rest of the $(K - m_1)$ CCs, the PCell uses Eqn. (65) to update the belief vector.

- **Step 6:** The PCell uses the updated belief vector to recompute the immediate reward functions for the m_1 CCs and determine the best m_2 CCs for the user to add as SCells. The selected CCs are indicated to the user using RRC Reconfiguration message. During this process, the PCell also indicates the CC removal information to the user to remove those SCells that were not included in m_2.

3.1.3.6 Proposed Joint Myopic Policy based Handover Scheme for B4G systems

The handover mechanism for the joint small cell discovery and component carrier selection applied for LTE-Advanced and beyond systems is described in Figure 51.

In this scheme, the primary cell which hosts the primary component carrier determines the set of CCs $\mathcal{M}_1$ based on the joint myopic policy. This set of CCs is sent to the MU through a Measurement Configuration message over the radio link. This message will also include the measurement scheduling information for the MU. Upon receiving the measurement configuration from the PCC, the MU performs measurement on the $\mathcal{M}_1$ set of CCs and reports the measured values to the primary cell.

Using the measurement information, the primary cell determines the updated belief vectors for the $\mathcal{M}_1$ set of CCs. From this, the PCC can determine the best set of $\mathcal{M}_2$ as defined in the joint myopic policy. Once the new set of secondary CCs are selected, the PCell uses the RRC Reconfiguration message to indicate the MU to handover to the selected SCCs. If the set $\mathcal{M}_2$ includes one or more of the CCs already used by the MU, no RRC Reconfiguration message is sent to the user corresponding to these CCs. If a CC is expected to be removed, this is indicated in the RRC Reconfiguration message. The MU acknowledges by sending an RRC Reconfiguration Complete message to the PCC marking the completion of the SCC handover process.

3.1.3.7 SCell Handover Failure Probability

For the propesed handover scheme, we derive the expression for the SCell handover failure probability in multistream HetNets. The SCell handover failure probability is defined as the probability that the user leaves the coverage of a target SCell during handover from another SCell.

To this end, we first define the parameters handover signaling delay τ and time for user to leave the cell t_{out}. Since we classify the user location into CC states, t_{out} indicates the time the user takes to move from its current CC state S_j to state S_{N+1}. We define the random variable t_{out} as

$$t_{out} = \sup_{\rho_m \in \mathcal{N}; \rho_m \notin S_{N+1}} \{(m+1).\Delta t : m \geq 0, \} \tag{67}$$

where $\rho_m = \rho(t + m.\Delta t)$ is the distance of user from the CC at the m^{th} time interval. The handover signaling delay τ depends on the type of SCell handover. We can represent τ for the generalized case as $\tau = \gamma.\Delta t$ for $\gamma > 0$.

Then the probability of SCell handover failure is defined as the probability that the time taken by user to move to the CC state S_{N+1} of the target SCell is less than the SCell handover signaling delay. Therefore, the SCell handover failure probability p_f is given by

$$p_f = Prob.(t_{out} \leq \tau) = Prob.(t_{out} \leq \gamma.\Delta t), \tag{68}$$

$$= Prob.(\rho_\gamma \in S_{N+1} | \rho_0 \in \mathcal{N}, \rho_0 \notin S_{N+1}), \tag{69}$$

$$= (\mathbf{\Omega}_\gamma).\mathbf{e}_{N+1} = (\mathbf{\Omega}_0.\mathbf{P}^\gamma).\mathbf{e}_{N+1}, \tag{70}$$

where $\mathbf{\Omega}_0$ is the probability distribution of the user's state for the CC before handover was initiated and $\mathbf{e}$ is the standard basis. The stationary probability distribution of the CC can be used to determine $\mathbf{\Omega}_0$.

3.1.4 Performance Evaluation

We evaluate the performance of the proposed scheme using several performance metrics. These metrics are divided into two categories. In the first category, we use the metrics such as the aggregate throughput, energy consumption and measurement gaps to determine the

performance of the CC discovery process. For the second category, we evaluate the SCC handover performance through metrics including SCC change rate, ping-pong rate, SCC handover costs. Different SCC handover types will incur different handover costs. We will study all of these performance metrics using Monte Carlo simulation of a large network of SCCs overlaid on a macrocell layer.

3.1.4.1 Simulation Setup

In order to evaluate the performance of the proposed joint policy, we conduct Monte-Carlo simulation of a large network using MATLAB. The network consists of a macrocell with coverage of 2km radius. The macrocell acts as the primary cell for the users providing the primary component carrier. Throughout the simulation time, we consider that the users remain within the macrocell coverage area and do not undergo a PCell handover.

As an overlay on the macrocell layer, several small cells of different cell radii and supporting one or more SCCs are deployed. The bandwidths of the CCs are selected uniformly from the set {3, 5, 10, 15, 20} MHz which are the possible bandwidths of LTE carriers. Each CC has $N = 20$ possible states. The step distance of each state depends on the CC's coverage range. The SCC radii are determined using a normal distribution with mean of 60m and variance of 40m. The network consists of 100 users. The initial positions of users are selected randomly. The relative velocities of users with respect to each SCC is determined from the distribution in Eqn. (59). For our simulations, we consider that the user buffers are infinitely backlogged with traffic throughout the simulation time.

The PCell maintains the belief vector for each user for the K possible SCCs in the network. Since no measurement information is available at the beginning of the simulation, the PCell uses the stationary probability of the CC k for user u to determine its initial belief state distribution.

The simulation is run for 100 time intervals with each interval having a duration of 10 seconds. During each time interval, the PCell determines the set $\mathcal{M}_1$ and also $\mathcal{M}_2$ based on the measurement results on set $\mathcal{M}_1$. Following this, the handover is occured to the $\mathcal{M}_2$ set

of CCs. The simulation uses different values of $\mathcal{M}_1$ and $\mathcal{M}_2$ to study their impact on the performance metrics defined above. The simulation parameters are summarized in Table 6.

Table 6: Simulation parameters.

Parameter	Value
Number of SCCs K	20
Number of users	100
Number of SCC states N	20
ΔT	$10s$
SCC coverage radius	Rand., $\mu = 60m$, $\sigma^2 = 40m$
Simulation Time	$1000s$
$\overline{V}^{\mu}$	Uniform in range [5,15] m/s
$\delta_{v^{\mu}}$	Uniform in range [1,3] m/s
Number of Monte Carlo trials	$2e5$
m_1	5 to 15
m_2	1 to 5

The performance of the proposed joint policy is compared with a maximum received signal strength (max-RSSI) based scheme. In this case, the first set of $\mathcal{M}_1$ are selected randomly. However, after the measurements are performed, the set $\mathcal{M}_2$ is selected based on the CCs in the set $\mathcal{M}_1$ that have the highest received signal strength for the user as indicated on the measurement report by the user. The simulation results are plotted and discussed in Section 3.1.4.2.

3.1.4.2 Simulation Results

Average throughput vs Number of CCs measured: For the preliminary analysis, we first plot the throughput experienced by users on the SCCs for different values of m_1 given a fixed m_2 of SCCs that the users connect to. The number of SCCs measured plays a major role in determining the aggregate throughput experienced by users over the selected SCCs. It is intuitive that as the number of CCs measured increases, the users have higher chance of detecting CCs that offer greater bandwidth to the users. However, performing measurements on several CCs will invariably increase the energy consumption for the user terminals. The results are compared between our proposed scheme and the max-RSSI based scheme and plotted in Figure 46. For this case, we set the length of m_2 to 3.

From the results, it can be observed that the proposed policy requires fewer measurements compared to the max-RSSI based policy to achieve the same aggregate SCC throughput. This is a key benefit in terms of SCC discovery of our proposed scheme since lower number of CC measurements indicate lower energy consumption for users.

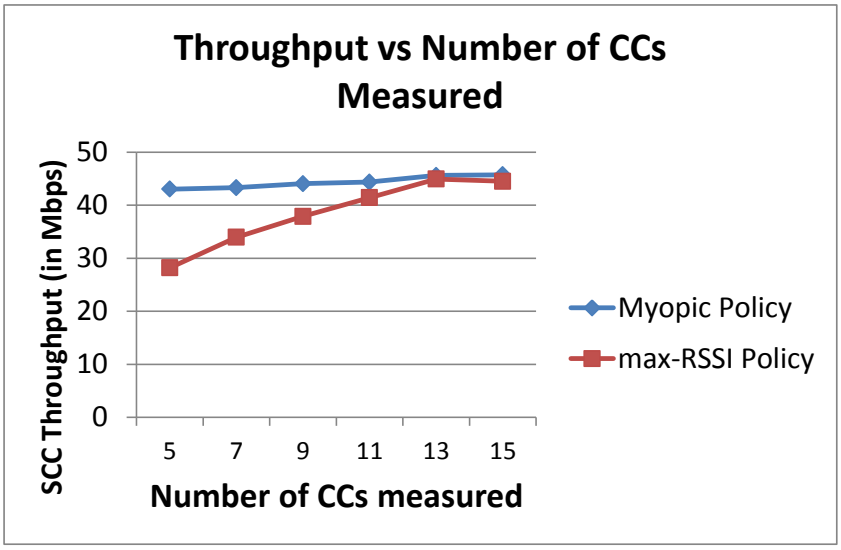

Figure 46: Average aggregate SCC throughput per user for different values of m_1.

SCell Handover Rate: Following the throughput, we evaluate the SCC handover performance for the users. This is measured in terms of the SCC change rate and the ping-pong rate. SCC rate indicates the rate at which the users undergoes handover to a new SCC such that none of the SCCs from the previous time interval overlaps with the SCCs serving the user currently. A large SCC change rate indicates a greater number of handovers for the user leading to increased signaling costs.

The SCC change rate depends on the number of CCs that the user is attached to, i.e., m_2, at any given time. The value of m_2 depends on the transceiver capabilities of the mobile terminal. The SCC change rate for the two schemes are compared in Figure 47.

Given that the users are highly mobile and are in search for CCs with higher available bandwidth in our simulation environment, the users experience frequent SCC handover for low values of m_1. For low values of m_1, the SCC handover rate is still substantially lower for the proposed joint myopic policy in comparison to the max-RSSI scheme. Since the

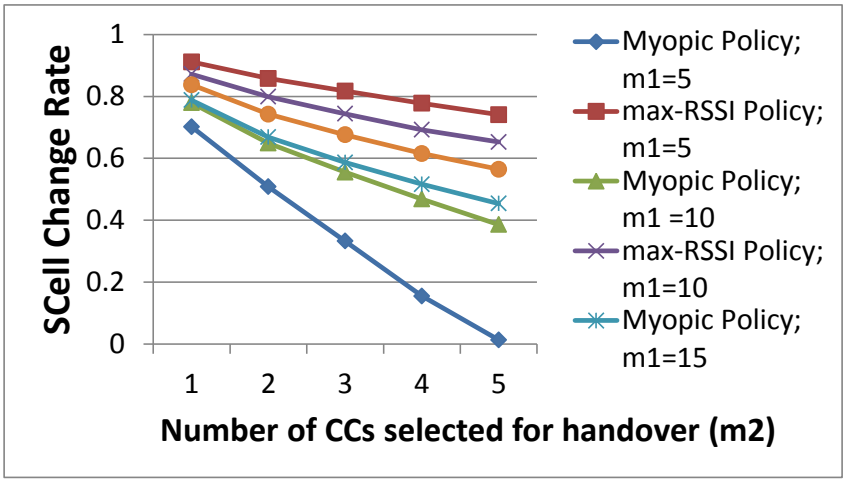

Figure 47: Average SCell change rate for different values of m_1 and m_2.

myopic policy determines the best set of CCs based on long-term accumulated information and not simply based on the CCs with the best RSSI in the current time, the CCs that promise higher long-term rewards are preferred over the current strongest CC.

The benefit of the proposed scheme becomes increasingly evident as the value of m_2 increases. The proposed scheme selects the CCs such that the aggregate reward offered by the CCs is maximized. Since the set of CCs that have higher probability of being visited will likely offer greater aggregate reward is exploited in the proposed scheme. Therefore, the SCell change rate is quite low for the myopic policy. This also implies that the amount of RRC signaling required by the users to complete the SCell handover process is minimized.

It is also interesting to observe the impact of m_1 on the SCell change rate. Performing measurements on more number of CCs will invariably create more handover opportunities for the users. Especially, in the case of the max-RSSI scheme, there is a steady increase in the SCell change rate as m_1 increases since the PCell is able to identify CCs with better signal strength with a greater probability. On the contrary, for the case of the joint myopic policy, the SCell change rate increase initially with increasing m_1. However, this increase is very minimal as m_1 is increased from 10 to 15. This shows that for a given SCell change rate, the proposed scheme requires lesser number of CCs on which measurements are conducted, reducing the energy consumption and measurement gaps.

Impact of User Velocities: It is important to also compare the impact of user velocities for the existing and the proposed schemes. The average velocity $\overline{V}$ for the users is varied from 2.5 m/s to 20 m/s and the SCC change rates are determined corresponding to different values of the velocities. The parameters m_1 and m_2 are fixed to study the impact of the user velocities. The SCell handover/change rate is plotted in Figure 48.

Even for low user velocities, the max-RSSI policy results in high SCell change rate. The SCell change rate increases up to 0.7 at $\overline{V} = 10$ m/s after which it remains more or less the same up to $\overline{V} = 20$ m/s. For the proposed scheme, on the other hand, the SCell change rate is under 0.4 for velocities ≤ 5 m/s. The increase in the change rate is very smooth reaching 0.55 at velocities of 20m/s. It is also interesting and slightly counter-intuitive to observe that the change rate is lower for large values of m_2. However, the reasoning is that the proposed joint myopic policy selects the SCells such that the aggregate reward offered by the CCs is maximized. With future cellular systems allowed to aggregate several CCs for the users, the proposed scheme significantly minimizes the handover rate.

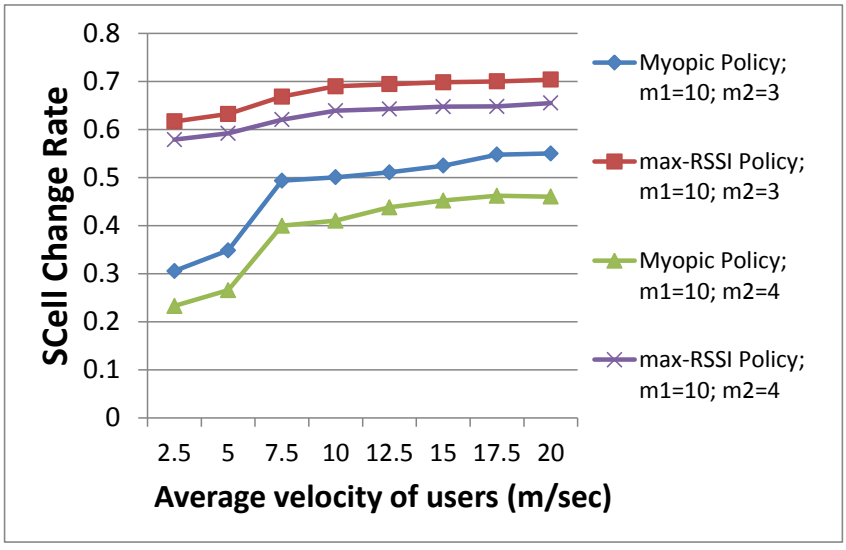

Figure 48: Average SCell change rate for different user velocities.

Ping-pong Rate: As the next step, we study the rate of ping-pongs experienced by the users in the network. Ping-pong rate represents the rate at which a user undergoes SCC handover at the current time interval only to go back to the previous SCC in the successive

time interval. Similar to the SCC change rate, a high ping-pong rate also results in high-signaling costs to perform the SCC handovers frequently for the users.

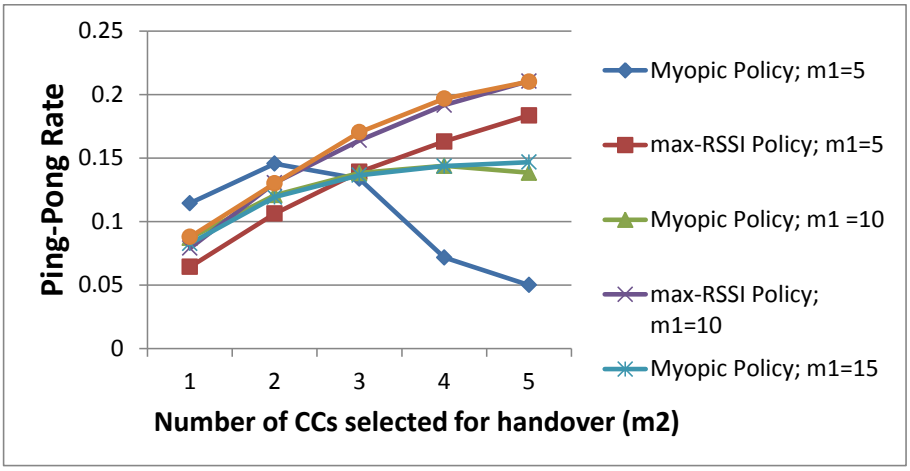

Figure 49: Average ping-pong rate for different values of m_1 and m_2.

As shown in Figure 49, the ping-pong rate for users under the joint myopic policy is comparable to that of the max-RSSI scheme (≈ 0.1) for low values of m_2 irrespective of the number of CCs measured. However, as m_2 increase, the ping-pong rate increase under the max-RSSI scheme reaching up to 0.2 for $m_2 = 5$ for all values of m_1.

However, in our proposed scheme, the ping-pong rate saturates at a value of 0.1 for all values of m_1 and m_2. The same reasoning for the reduced SCC change rage can be used to explain the low ping-pong rate for the joint myopic policy. As m_2 increases, the users are able to attach to more CCs at the same time. It becomes increasing likely that the selected set of CCs need to be changed less frequently than in the case of the max-RSSI policy since the joint policy maximizes the aggregate reward over a long time instead of selecting the strongest CCs during every time interval.

Impact of User Velocities: As in the case of the SCell change rate, we varied the velocities of the users from 2.5 m/s to 20 m/s and observed the change in the frequency of observing ping-pongs in the network. For low-velocities up to 5 m/s, the ping-pong rate under the max-RSSI policy (≈ 0.22) is double that of the proposed policy ($\lessapprox 0.1$). For the medium velocity range from 7.5 m/s to 12.5 m/s, the ping-pong rate is still $1.5x$ under

the max-RSSI policy compared to the proposed approach. For high velocities, the averate ping-pong rate reaches a value of 0.18 under the proposed scheme compared to 0.23 under the max-RSSI scheme.

The choice of m_2 also does not have a major impact on the ping-pong rate for the proposed policy. However, in the case of the max-RSSI policy, the ping-pong rate increases at least by 0.02 as m_2 is increased from 3 to 4.

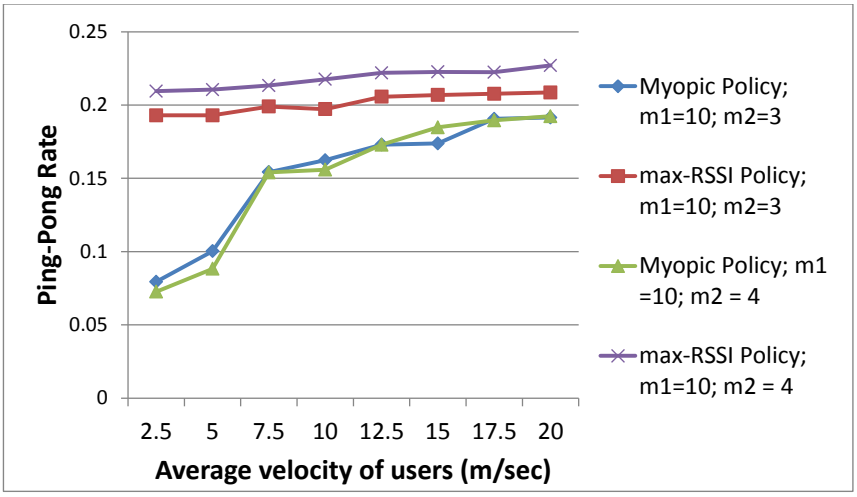

Figure 50: Average ping-pong rate for different user velocities.

3.1.5 Conclusions

In this chapter, we have focused on multistream HetNets, which is an evolving HetNet architecture for beyond 4G systems. Although multistream HetNets promise an improvement in handover performance by retaining coverage for users with a macrocell stream, it also leads to new challenges in terms of CC discovery and CC handover. We focused on these two major issues in multistream HetNets and proposed a joint approach to optimize the CC discovery and handover processes with the help of maintaining per-user belief states for the CCs in the system. Our joint myopic policy provides a practical approach to the problem formulation by breaking down the N-dimensional problem into N 1-dimensional problems. We modeled the multistream HetNets with random coverage and used Monte Carlo simulations to evaluate the performance of the proposed scheme using key performance metrics.

Our results indicated that the proposed scheme shows a significant improvement in the system performance by delivering increased throughput for a given number of measurements, reduced CC handover and ping-pong rates as well as reduced signaling cost.

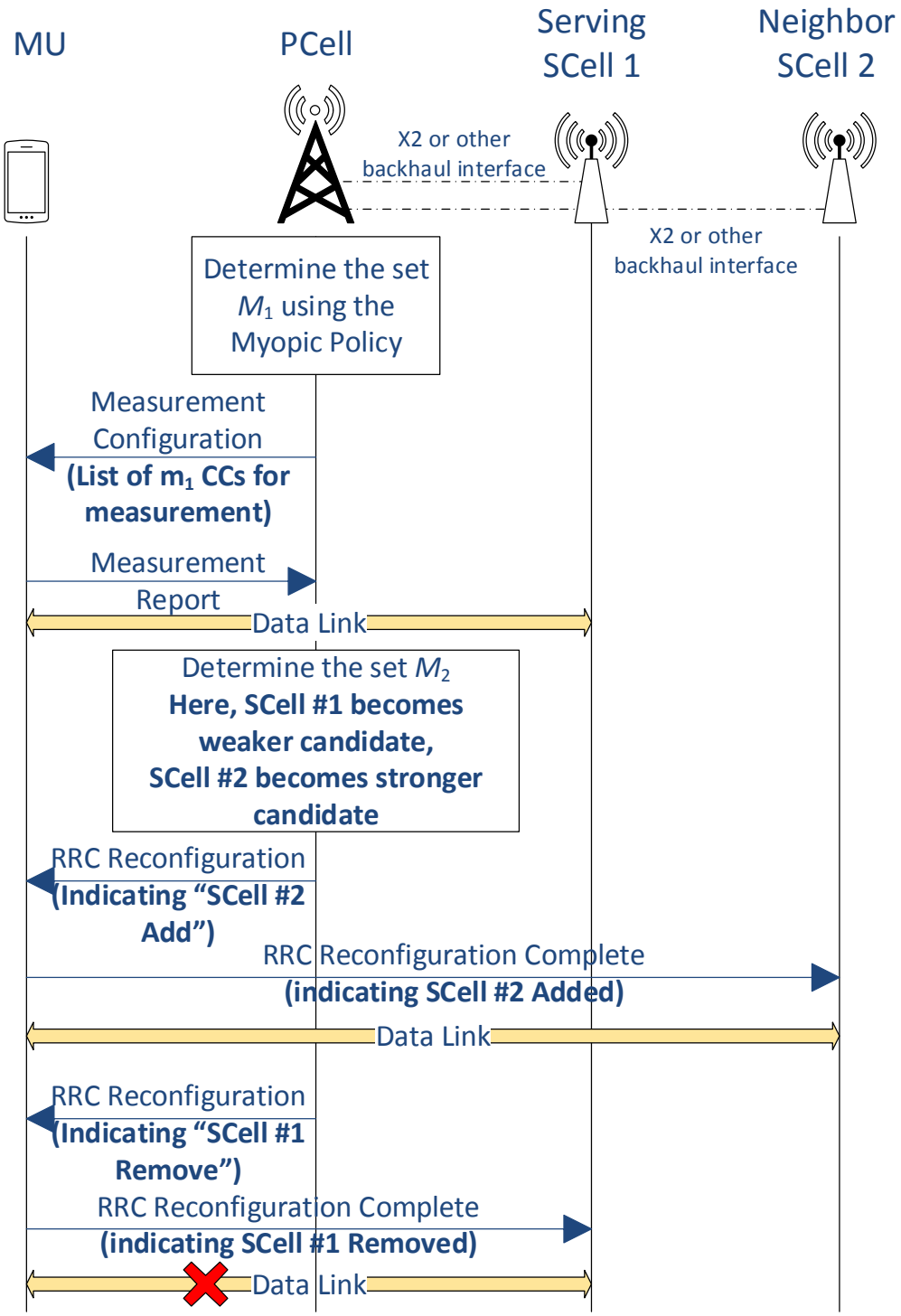

Figure 51: Proposed SCell discovery and handover mechanism for 4G and beyond systems.

CHAPTER 4

CONCLUSIONS

Cellular networks beyond the fourth generation promise significant performance gains through the use of a multitude of techniques. Among these, HetNets are expected to play a dominant role in increasing the overall network capacity as well as providing ubiquitous coverage. HetNets, however, face big challenges that impact the actual capacity gains achieved by the overall system as well as the quality of experience offered to the mobile users. In particular, HetNets raise several challenges in terms of cell association, handover management and energy management that need to be effectively addressed. Furthermore, new architectures for supporting HetNets to tackle some of these problems also need to be evaluated for not only their effective functioning but also in terms of their ease of being integrated into the evolving beyond 4G systems.

In this thesis, we have investigated one of the key issues in HetNets which is handover management. Our starting point has been studying handovers in small cells taking into account their unique features such as user-specific access control, dynamic capacity and reliability of backhaul. We have identified that small cells require advanced admission control functions in view of these unique features and proposed a traffic-aware admission control and subcarrier assignment approach to improve the QoS performance of the existing and incoming users. Furthermore, in view of the backhaul limitations of small cells, we have advocated the use of a local anchor for a cluster of small cells that act as the mobility anchor for user undergoing handovers between small cells. As our final work for handover in small cells, we have focused on mobile small cells supporting a group of users such as in a high-speed vehicle. For this case, we have proposed a group handover scheme to seamlessly handover the users as the mobile relay switches its point of attachment from one fixed macrocell to another.

As a further step, we have studied new architectures envisioned for HetNets such as

the inter-site or multistream carrier aggregation HetNets where users can receive multiple streams of data from a primary cell which is typically the macrocell and also from one or more secondary cells which are typically small cells. In this case, we have particularly focused on two of the handover related issues. First, we have the problem of the SCell discovery where measurements on the secondary cells operating on different component carriers need to be performed in a timely manner with the objective of maximizing the offloading opportunities for the users while minimizing the energy consumption. Second, we have looked at the problem of SCell handover, also termed as component carrier selection where the objective is to dynamically add, remove or change SCells based on user mobility dynamics and bandwidth requirements.

The overall contributions of this PhD thesis can be broken down into two major thrusts. In the first thrust, we have focused on handover management approaches for fixed and mobile small cells. All of these approaches are presented within Chapter 2. In the second thrust, we have focused on novel handover approaches for multistream HetNets which are summarized in Chapter 3. In the following, we recapitulate the contributions of each of the chapters of the thesis.

In Chapter 2, we have addressed the handover related issues in small cells. In particular,

- First, we proposed, modeled and evaluated a traffic-aware admission control and subcarrier assignment approach for provisioning QoS to both the existing as well as the newly arriving users at the small cells.

  - Our proposed model incorporates the heterogeneous traffic characteristics of the users while also taking into account the unique access control features in order to provide differentiated service for users with different privileges on accessing the small cell.

  - The simulation results show that our proposed scheme provides an efficient way to admit new users at the small cells. In addition, it also utilizes the inter-working

between the admission control and scheduling modules to provision QoS to users with three different traffic types and varying access privileges.

- Second, we proposed, modeled and evaluated local anchor based handover schemes where handovers for users between small cells are anchored by a local anchor small cell.

  - For the local anchor based handover schemes, we proposed a novel handover architecture that can be extended to current LTE-Advanced systems.

  - We developed an analytical model to evaluate the proposed scheme for different cluster sizes, session and mobility dynamics.

  - Our detailed results showed that the proposed handover schemes significantly improves the handover performance by minimizing the handover interruption time and signaling costs.

- Third, we proposed a group handover strategy for users served under a mobile relay.

  - Our proposed scheme enabled seamless re-establishment of radio links for a group of users with the mobile relay as the relay changes its point of attachment from one fixed macrocell to another.

  - Our analysis showed that the proposed group handover strategy can offer significant signaling cost savings both at the radio access network as well as at the core network.

In Chapter 3, we have addressed the handover related issues in multistream HetNets. In particular, we have treated the small cell discovery and component carrier selection for multistream HetNets as a coupled problem.

- First, we proposed and evaluated a joint small cell discovery and CC selection approach with the help of maintaining a per-user belief state for each CC in the network. A detailed SCell handover mechanism is proposed to achieve the above tasks.

- Our performance evaluation indicated that the proposed scheme minimizes the energy consumption and signaling load associated with CC discovery. Furthermore, our approach reduced the CC change rate and ping pong rate, thereby also reducing the signaling load when selecting the SCells for handover.

In the future, we intend to focus on emerging HetNet architectures for 5G systems involving mmWave small cells. While the adoption of mmWave small cells for 5G systems definitely promise a significant increase in per-user throughput as well as overall system capacity, handover will be a major issue due to the unique propagation characteristics of the mmWave signal. In particular, high shadowing associated with mmWave signal propagation necessitates the need for a strong LoS component for efficiently receiving the mmWave signal at the user devices. For this reason, mmWave cells are expected to use a large antenna array and support directional beamforming to allow directivity gains. However, the presence of a varied range of blockages including human obstacles, antenna orientation, walls of buildings will all result in radio link failures. Handover, therefore, becomes a major issue if mmWave small cells are adopted for 5G systems. Therefore, in the future, we will explore more heterogeneous architectures that involve microwave and mmWave cells and propose handover schemes that can leverage the high data rate capabilities of mmWave cells for high-speed users while maintaining session continuity for the users.

REFERENCES

[1] Metis, "Scenarios, requirements and kpis for 5g mobile and wireless system," tech. rep., ICT-317669 METIS project, 2013.

[2] Cisco, "Visual netnetwork index," tech. rep., Cisco.com, Feb. 2014.

[3] I. Akyildiz, D. Gutierrez-Estevez, R. Balakrishnan, and E. Chavarria-Reyes, "LTE-Advanced and the Evolution to Beyond 4G Systems," *Physical Communication*, 2013.

[4] "Small cells-What's the big idea? Femtocells are expanding beyond the home." Small Cell Forum whitepaper, Feb. 2012.

[5] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, 2008.

[6] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility enhancements in heterogeneous networks (Release 11);," tech. rep., TS 36.839, Dec. 2012.

[7] 3GPP, "Evolved universal terrestrial radio access and evolved universal terrestrial radio access network; overall description; stage 2," tech. rep., TS 36.300, Mar. 2013.

[8] R. Balakrishnan, B. Canberk, and I. Akyildiz, "Traffic-aware Utility based QoS Provisioning in OFDMA Hybrid Smallcells," in *IEEE ICC*, Jun. 2013.

[9] R. Balakrishnan and B. Canberk, "Traffic-aware qos provisioning and admission control in ofdma hybrid small cells," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2013.

[10] D. Choi and et al, "Dealing with loud neighbors: The benefits and tradeoffs of adaptive femtocell access," in *IEEE GLOBECOM*, 2008.

[11] A. Valcarce, D. López-Pérez, G. De La Roche, and J. Zhang, "Limited access to OFDMA femtocells," in *IEEE PIMRC*, pp. 1–5, IEEE, 2009.

[12] P. Xia, V. Chandrasekhar, and J. G. Andrews, "Femtocell access control in the TDMA/OFDMA uplink," in *IEEE GLOBECOM*, pp. 1–5, IEEE, 2010.

[13] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in *17th International Teletraffic Congress*, 2000.

[14] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Communications Magazine*, vol. 43, no. 12, pp. 127–134, 2005.

[15] G. Song, Y. Li, and L. J. Cimini, "Joint channel-and queue-aware scheduling for multiuser diversity in wireless OFDMA networks," *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 2109–2121, 2009.

[16] W.-H. Park, S. Cho, and S. Bahk, "Scheduler design for multiple traffic classes in OFDMA networks," *Computer Communications*, vol. 31, no. 1, pp. 174–184, 2008.

[17] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 66–71, 2009.

[18] A. Antonopoulos, C. Skianis, and C. Verikoukis, "Traffic-aware connection admission control scheme for broadband mobile systems," in *IEEE GLOBECOM*, pp. 1–5, 2010.

[19] R. Singoria and et al, "Reducing Unnecessary Handovers: Call Admission Control Mechanism between WiMAX and Femtocells," in *IEEE GLOBECOM*, 2011.

[20] R. Ramjee, D. Towsley, and R. Nagarajan, "On Optimal Call Admission Control in Cellular Networks," *Wireless Networks*, vol. 3, no. 1, pp. 29–41, 1997.

[21] S.-E. Elayoubi, T. Chahed, and G. Hébuterne, "Mobility-aware admission control schemes in the downlink of third-generation wireless systems," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 1, pp. 245–259, 2007.

[22] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital Communication*. Kluwer Academic Publishers, third ed., 2004.

[23] B. Canberk, I. Akyildiz, and S. Oktug, "A qos-aware framework for available spectrum characterization and decision in cognitive radio networks," in *IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications, IEEE PIMRC*, pp. 1533 –1538, sept. 2010.

[24] P. Wang and I. Akyildiz, "On the origins of heavy-tailed delay in dynamic spectrum access networks," *IEEE Transactions on Mobile Computing*, vol. 11, pp. 204 –217, Feb. 2012.

[25] P. Wang and I. Akyildiz, "Network stability of cognitive radio networks in the presence of heavy tailed traffic," in *IEEE SECON*, Jun. 2012.

[26] M. G. Markakis, E. H. Modiano, and J. N. Tsitsiklis, "Scheduling policies for single-hop networks with heavy-tailed traffic," in *47th Annual Allerton Conference on Communication, Control, and Computing*, pp. 112–120, 2009.

[27] M. Andrews and et al, *CDMA data QoS scheduling on the forward link with variable channel conditions*. Bell Laboratories, Lucent Technologies, 2000.

[28] P. Sinha and A. A. Zoltners, "The multiple-choice knapsack problem," *Operations Research*, vol. 27, no. 3, pp. pp. 503–515, 1979.

[29] K. Dudzinski and S. Walukiewicz, "Exact methods for the knapsack problem and its generalizations," *European Journal of Operational Research*, vol. 28, pp. 3 – 21, 1987.

[30] D. Pisinger, "A minimal algorithm for the multiple-choice knapsack problem.," *European Journal of Operational Research*, vol. 83, pp. 394–410, 1994.

[31] "RP 110438: HetNet mobility improvements for LTE," TSG-RAN Meeting #51, Nokia Siemens Networks, Nokia Corporation, Alcatel-Lucent, Mar. 2011.

[32] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility enhancements in heterogeneous networks (Release 11);," tech. rep., TS 36.839, Dec. 2012.

[33] "Release Two - Enterprise: Overview." Small cell Forum, Nov. 2013.

[34] J. Ferragut and J. Mangues-Bafalluy, "A self-organized tracking area list mechanism for large-scale networks of femtocells," in *IEEE International Conference on Communications (ICC)*, pp. 5129–5134, June 2012.

[35] B. Jeong, S. Shin, I. Jang, N. W. Sung, and H. Yoon, "A Smart Handover Decision Algorithm Using Location Prediction for Hierarchical Macro/Femto-Cell Networks," in *IEEE Vehicular Technology Conference (VTC Fall)*, pp. 1 –5, sept. 2011.

[36] W. Shaohong, Z. Xin, Z. Ruiming, Y. Zhiwei, F. Yinglong, and Y. Dacheng, "Handover Study Concerning Mobility in the Two-Hierarchy Network," in *IEEE Vehicular Technology Conference, VTC Spring*, pp. 1–5, Apr. 2009.

[37] J.-M. Moon and D.-H. Cho, "Efficient handoff algorithm for inbound mobility in hierarchical macro/femto cell networks," *IEEE Communications Letters*, vol. 13, pp. 755–757, Oct. 2009.

[38] D. Xenakis, N. Passas, and C. Verikoukis, "An energy-centric handover decision algorithm for the integrated LTE macrocell-femtocell network," *Computer Communications*, vol. 35, no. 14, pp. 1684 – 1694, 2012.

[39] I. M. Bălan and et al, "Signalling minimizing handover parameter optimization algorithm for LTE networks," *Wireless Networks*, vol. 18, no. 3, pp. 295–306, 2012.

[40] E. Ha, Y. Choi, and C. Kim, "A New Pre-handoff Scheme for Picocellular Networks," in *IEEE International Conference on Personal Wireless Communications*, 1996.

[41] A. Rath and S. Panwar, "Fast Handover in Cellular Networks with Femtocells," in *IEEE ICC*, pp. 2752–2757, 2012.

[42] L. Wang, Y. Zhang, and Z. Wei, "Mobility Management Schemes at Radio Network Layer for LTE Femtocells," in *IEEE VTC Spring*, pp. 1–5, IEEE, 2009.

[43] X. An, F. Pianese, I. Widjaja, and U. G. Acer, "dMME: Virtualizing LTE mobility management," in *IEEE 36th Conference on Local Computer Networks (LCN)*, pp. 528–536, 2011.

[44] H. Zhang, W. Zheng, X. Wen, and C. Jiang, "Signalling overhead evaluation of HeNB mobility enhanced schemes in 3GPP LTE-Advanced," in *IEEE VTC Spring*, 2011.

[45] 3GPP, "Feasibility study for evolved Universal Terrestrial Radio Access and Universal Terrestrial Radio Access Network," tech. rep., TR 25.912, Sep. 2012.

[46] J. S. Ho and I. F. Akyildiz, "Local anchor scheme for reducing signaling costs in personal communications networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 4, no. 5, pp. 709 – 725, 1996.

[47] A. Campbell, J. Gomez, and A. Valko, "An overview of cellular IP," in *IEEE Wireless Communications and Networking Conference, 1999*, pp. 606–610 vol.2, 1999.

[48] F. Zdarsky, A. Maeder, S. Al-Sabea, and S. Schmid, "Localization of Data and Control Plane Traffic in Enterprise Femtocell Networks," in *Proc. IEEE 73rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, May 2011.

[49] T. Guo and et al, "Local Mobility Management for Networked Femtocells Based on X2 Traffic Forwarding," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 1, pp. 326–340, 2013.

[50] R. Jain and Y.-B. Lin, "An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS," *Wireless Networks*, vol. 1, no. 2, pp. 197–210, 1995.

[51] R. Balakrishnan and I. Akyildiz, "Local Mobility Anchoring for Seamless Handover in Coordinated Small Cells," in *IEEE GLOBECOM (to appear)*, Dec. 2013.

[52] R. Balakrishnan and I. F. Akyildiz, "Local Anchor Schemes for Seamless and Low-cost Handover in Coordinated Small Cells," *in revision, IEEE Transactions on Mobile Computing*, 2014.

[53] 3GPP, "3GPP System Architecture Evolution (SAE); Security architecture," tech. rep., TR 33.401, Mar. 2013.

[54] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); TDD Home eNode B Radio Frequency requirements analysis," tech. rep., TR 36.922, Sep. 2013.

[55] "A Small Cell Forum Whitepaper: Enterprise Femtocell Deployment Guidelines." Small cell Forum, Feb. 2012.

[56] B. Liang and Z. Haas, "Predictive distance-based mobility management for PCS networks," in *IEEE INFOCOM*, vol. 3, pp. 1377–1384, Mar 1999.

[57] J. Sydir and R. Taori, "An evolved cellular system architecture incorporating relay stations," *Communications Magazine, IEEE*, vol. 47, pp. 115–121, June 2009.

[58] "IEEE 802.16 WG SG, IEEE 802 Tutorial: 802.16 Mobile Multihop Relay." IEEE 802.16 standard.

[59] K. Loa, C.-C. Wu, S.-T. Sheu, Y. Yuan, M. Chion, D. Huo, and L. Xu, "Imt-advanced relay standards [wimax/lte update]," *Communications Magazine, IEEE*, vol. 48, pp. 40–48, August 2010.

[60] "WiMAX End-to-End Network System Architecture - Stage 3: Detailed Protocols and Procedures."

[61] "Advanced Air Interface for Broadband Wireless Access Systems."

[62] R. Kim, I. Jung, X. Yang, and C.-C. Chou, "Advanced handover schemes in imt-advanced systems [wimax/lte update]," *Communications Magazine, IEEE*, vol. 48, pp. 78–85, August 2010.

[63] "R2 115745: Inter-frequency Pico cell measurements for Hetnet deployments," 3GPP TSG-RAN WG2 #76, NTT DOCOMO, Inc, Nov. 2011.

[64] "R2-123102: Background search for small cell detection," 3GPP TSG-RAN WG2 #78, Nokia Siemens Networks, 2012.

[65] "R2-120652: On UE-speed-based methods for improving the mobility performance in HetNets," 3GPP TSG-RAN WG2 #77, Alcatel-Lucent, 2012.

[66] "R2-122368: Enhanced MSE based small cell detection," 3GPP TSG-RAN WG2 #78, Nokia Siemens Networks, 2012.

[67] H.-Y. Lee and Y.-B. Lin, "A cache scheme for femtocell reselection," *Communications Letters, IEEE*, vol. 14, no. 1, pp. 27–29, 2010.

[68] C.-H. Lee and J.-H. Kim, "System Information Acquisition Schemes for Fast Scanning of Femtocells in 3GPP LTE Networks," *IEEE Communications Letters*, vol. 17, pp. 131–134, January 2013.

[69] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for support of radio resource management (Release 12)," tech. rep., TR 36.133, 2013.

[70] K. Pedersen, P. Michaelsen, C. Rosa, and S. Barbera, "Mobility enhancements for LTE-advanced multilayer networks with inter-site carrier aggregation," *IEEE Communications Magazine*, vol. 51, no. 5, pp. 64–71, 2013.

[71] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*, pp. 121–151, Springer, 2008.

[72] C. H. Papadimitriou and J. Tsitsiklis, "The Complexity of Optimal Queueing Network Control," in *Proceedings of the Ninth Annual Structure in Complexity Theory Conference*, pp. 318–322, 1994.

[73] W. Wang and M. Zhao, "Joint Effects of Radio Channels and Node Mobility on Link Dynamics in Wireless Networks," in *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, pp. –, 2008.

[74] R. D. Smallwood and E. J. Sondik, "The Optimal Control of Partially Observable Markov Processes over a Finite Horizon," *Operations Research*, vol. 21, no. 5, pp. 1071–1088, 1973.

PUBLICATIONS

Journal Papers

1. R. Balakrishnan and I.F. Akyildiz, "Joint Small cell Discovery and Component Carrier Selection for Multistream HetNets," manuscript in preparation, Feb. 2015.

2. R. Balakrishnan and I.F. Akyildiz, "Local Anchor Schemes for Seamless and Low-cost Handover in Coordinated Small Cells," submitted to IEEE Transactions on Mobile Computing, Feb. 2014, in revision, Aug. 2014, in revision, Feb. 2015.

3. I. F. Akyildiz, D. Gutierrez-Estevez, R. Balakrishnan and E. Chavarria-Reyes, "LTE-Advanced and the Evolution to Beyond 4G (B4G) Systems," Physical Communication (Elsevier) Journal, vol. 10, pp. 31-60, Mar. 2014.

4. I. F. Akyildiz, E. Chavarria-Reyes, D. Gutierrez-Estevez, R. Balakrishnan and J. R. Krier, "Enabling Next Generation Small Cells through Femtorelays," Physical Communication (Elsevier) Journal, vol. 9, pp. 1-15, Dec. 2013.

5. R. Balakrishnan and B. Canberk, "Traffic-aware QoS Provisioning and Admission Control in OFDMA Hybrid Small cells," IEEE Transactions on Vehicular Technology, vol. 63, no. 2, pp. 802-810, Feb. 2014.

6. I. F. Akyildiz, B. F. Lo and R. Balakrishnan, "Cooperative Spectrum Sensing in Cognitive Radio Networks: A Survey," Physical Commununications, vol. 4, no. 1, pp. 4062, Mar. 2011.

Conference Papers

1. R. Balakrishnan, M. Rashid, K. Sivanesan and R. Vannithamby, "On Co-existence of LTE-U and WiFi through LTE-A/LTE-U Dual Connectivity," submitted for publication, Nov. 2014.

2. R. Balakrishnan, and M. Rashid, "LTE on Unlicensed Spectrum for 5G Systems," Intel Labs Open House, Aug. 2014. [Best Poster Award]

3. R. Balakrishnan, and I. F. Akyildiz, "Local Mobility Anchoring for Seamless Handover in Coordinated Small Cells," in Proc. of IEEE GLOBECOM, Dec. 2013.

4. R. Balakrishnan, B. Canberk and I. F. Akyildiz, "Traffic-aware Utility based QoS Provisioning in OFDMA Hybrid Smallcells," in Proc. of IEEE International Conference on Communications, ICC, Jun. 2013.

5. R. Balakrishnan, X. Yang, M. Venkatachalam, I. F. Akyildiz, "Mobile Relay and Group Mobility for 4G WiMAX Networks," in Proc. of IEEE Wireless Communications and Networking Conference, WCNC, Mar. 2011.

Patents

1. R. Balakrishnan, X. Yang, M. Venkatachalam, "Method and Apparatus to Facilitate Mobile Relay and Group Mobility,", Intel Corporation, US Patent Application Number: 20120082084, Publication Date: Apr. 2012. (Patent Pending)

VITA

Ravikumar Balakrishnan received his Bachelor's degree in Electronics and Communications Engineering from SSN College of Engineering, Anna University, India in May 2009 and Master of Science degree from the School of Electrical and Computer Engineering, Georgia Tech, Atlanta, GA in December 2011. He is a Ph.D. Candidate at the Broadband Wireless Networking Lab under the supervision of Pr. Dr. Ian F. Akyildiz. During the year 2010, he held an internship position with Intel Corporation where he specialized in next generation cellular networks focusing on system architecture and protocols. He also held an internship position at ORB Analytics, MA during the summer of 2013 working on a DARPA project for overlay cognitive radio networks. His focus was to propose and develop an optimal spectrum sensing and access approach for cognitive radio networks. From May 2014 until December 2014, he held another internship position at Intel Labs, USA focusing on research and development for LTE on unlicensed bands for beyond 4G systems. He is a member of the IEEE and the IEEE Communications Society. His current research interests are in next generation cellular systems and cognitive radio networks.