RESOURCE MANAGEMENT SOLUTIONS IN OFDMA-BASED 4G CELLULAR NETWORKS

by

Jongwon Yoon

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: 12/15/2014

The dissertation is approved by the following members of the Final Oral Committee:

Suman Banerjee, Professor, Computer Sciences
Aditya Akella, Associate Professor, Computer Sciences
Paul Barford, Professor, Computer Sciences
Parameswaran Ramanathan, Professor, Electrical and Computer Engineering
Xinyu Zhang, Assistant Professor, Electrical and Computer Engineering

To Hoewook and Nakyung

ACKNOWLEDGMENTS

I have finally closed my graduate study and am about to start a new chapter of my life. During my graduate study, I have been fortunate enough to work with many great people. First and foremost, I am deeply grateful to my advisor, Suman Banerjee, whose guidance, encouragement and support have been a great motivation for me to successfully finish my PhD. He always encouraged and helped me to work on practical problems in necessary resources that could be beneficial in the wireless research community as well. He offered me great support. I have learnt how to guide students as an advisor from him, and following his footsteps will be greatly useful for me in my faculty career. I would also like to thank Aditya Akella, Paul Barford, Parameswaran Ramanathan and Xinyu Zhang for their profound feedback and priceless comments all of which greatly improved this thesis.

At the early stages of my research, I worked closely with Vivek Shrivastava on a wireless network driver. He was a good friend and colleague who helped me to establish myself as a researcher pursuing interesting problems in wireless networks. I also appreciate the time spent with my fellow labmates: Sayandeep Sen, Tan Zhang, Mike Griepentrog, Shraven Rayanchu, Ashish Patro, Peng Liu, Theophilus Benson, Ashok Anand, and Derek Myer. They were always reachable with any matter and were also good friends with whom I share great memories, traveling to conferences, attending the Packers game, enjoying life, and etc. We went through many good and tough times in both research and life, and all of these experiences bound us together like a brotherhood. They were all a big part of my journey.

Besides graduate school, I have enjoyed working in NEC Labs during my long-term internships. I would like to thank Sampath Ragarangan who gave me opportunities to collaborate with great mentors on many interesting topics in wireless networks. Sampath was very supportive and did not hesitate to acquire new (and expensive!) equipment. He provided me with great freedom to work on research problems. Especially, I would like to thank Karthik Sundaresan and Honghai Zhang for mentoring me with great guidance and

patience. Their constant support and guidance helped me to be improved to the next step. I will never forget the insightful discussion, hard-work and volleyball games with Mustafa Arslan, Rajesh Mahindra, Amir Khojastepour and Srikanth Krishnamurthy during my internships.

Being apart from family, living and studying in a foreign country, was not easy for me. Everytime I went through a tough time, my priest, Father Joseph Yoo, encouraged me and raised me emotionally. With his dedication and care, I was able to build my mental toughness. What I learnt from him is fundamental and the most valuable asset for my life.

I have received great love and support from my family during every step of this journey. My parents, Ohjun Yoon and Kwangjun Lee, have been proud of me throughout my graduate study and loved me from the deepest of their hearts. I would like to give all of my love to my family – Jonghyun Yoon, Sookiel Chung, Ockhyung Lee, and Woongsup Chung. To my wife, Hoewook Chung, I would not have completed the PhD process without your love and support. You have never stopped believing in me and always encouraged me with love. I dedicate this dissertation to Hoewook and my lovely daughter, Nakyung.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

RESOURCE MANAGEMENT SOLUTIONS IN OFDMA-BASED 4G CELLULAR NETWORKS

Jongwon Yoon

Under the supervision of Professor Suman Banerjee
At the University of Wisconsin-Madison

With the proliferation of mobile devices, the demand for high bandwidth in wireless networks rapidly increases. Small-cells are employed for increased spatial reuse of spectral resources with reduced cell sizes and dense deployments. In the dense deployment of small-cell networks, however, intercell interference is one of the performance limiting-factor in harnessing their potentials of reusing spectrum resources. In addition, unique design features of OFDMA systems pose critical challenges for developing resource management solutions. This motivates us to designing and building practical resource management systems that improve the spectrum efficiency in OFDMA-based 4G cellular networks.

Considering network architecture (marco-cell and femto-cell) used in OFDMA networks, we take different approaches for designing resource management schemes. More specifically, we design an efficient video multicast algorithm, MuVi, that optimizes resource utilization while accommodating heterogeneous multiple users in the same OFDMA resource frames. MuVi differentiates video frames and incorporates an efficient radio resource allocation algorithm to optimize the overall video quality across all users in the multicast group. In multi-cell deployment, the interference from neighboring cells is pervasive and severely impacts the performance of interfering cells. Thus, a resource management solution should handle the interference across the femto-cells properly and maximize the spatial reuse to improve capacity. Towards this goal, we propose two solutions, RADION and ProBeam. RADION is a distributed resource management framework that maximizes resource reuse while providing resource isolation when neighboring cells are within transmission range of each other in the time and frequency domains. ProBeam is a practical multi-cell

beamforming system that leverages smart antenna to increase spatial reuse in the space domain without sacrificing spectrum resources for resource isolation.

We believe this dissertation was successful in developing some of the important building blocks to improve the performance of multi-cell networks. These contributions can be carried out for other resource management schemes in next-generation cellular networks (i.e., 5G) that try to improve the spectral efficiency in multi-domains.

1 INTRODUCTION

Mobile device sales show strong growth worldwide, and smartphone shipments overtook PCs in 2011 [12]. With the rapid increase in mobile device usage (e.g., smartphones, tablets), the amount of traffic destined for mobile devices is also exponentially growing. For instance, a recent industry report predicted that mobile data traffic will increase nearly 11-fold between 2013 and 2018 [38]. In consequence, the demand for high bandwidth in wireless networks continues to grow. 3G cellular networks, however, cannot properly handle demands for increased capacity driven by ever-increasing data traffic. This necessitates the deployment of 4G mobile networks with new technology, Orthogonal Frequency Division Multiple Access (OFDMA) to keep pace with the increased demand for higher data rates.

The emergence of 4G cellular networks (e.g., WIMAX, LTE and LTE-A) provides much higher capacity, extended coverage and greater robustness to mobile users compared to 3G cellular networks. Besides the new technology, cellular networks employ two-tier network architecture (depicted in Figure 1.1) to increase spatial reuse and capacity. The macrocell provides continuous radio access while covering a large area with high power transmission. In addition, the broadband network deploys more femtocells as a cost-effective means of offloading data traffic from the macrocell network. For example, femtocells provide increased spatial reuse and capacity in crowded city areas where the demand for high capacity (i.e., data rate) rapidly increases. Two-tier architecture strikes a good balance; while the macrocell covers a large area controlled by a centralized network operator, femtocells offer improved coverage and energy-efficient deployment options for a smaller area [45, 102]. Operating both a macrocell and femtocells, OFDMA-based cellular networks provide a satisfactory quality of service to users while covering both large and small areas more efficiently.

OFDMA system is evolving and efficient resource management is still needed to meet the high bandwidth and stringent delay requirements of applications. However, resource management solutions in OFDMA systems have not

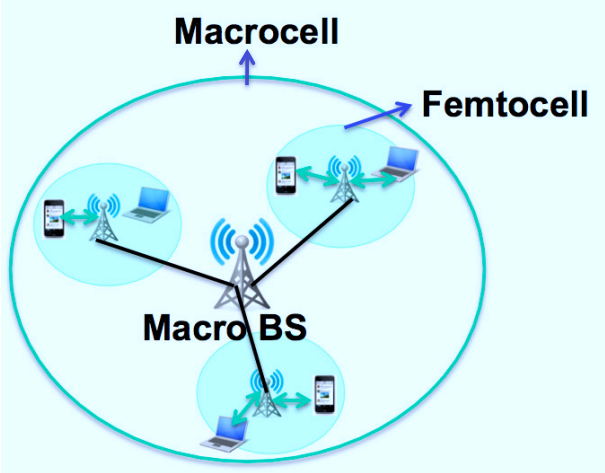


Figure 1.1: Example of macrocell and femtocells deployment: Femtocells are connected to the macrocell and reuse macro spectrum.

been studied in depth. Prior studies [95, 104] have been restricted to theory and include several simplifying assumptions that prevent their adoption in practice. Several solutions proposed in the WiFi domain cannot be directly adopted in OFDMA systems because of systematic differences. For instance, OFDMA systems require tight synchronization between the transmitter and receiver, whereas WiFi is based on asynchronous random channel access. Moreover, OFDMA standard does not employ channel sensing, and hence, this limits their ability to sense and manage interference effectively.

This challenge motivates us to design and implement practical resource management solutions that improve the spectrum efficiency in OFDMA networks. We look into this problem in two domains (single macro-cell and multi femto-cells) according to the two-tier architecture used in OFDMA-based cellular networks. More specifically, we need (a) to determine efficient resource allocations for heterogenous clients in marco-cell networks, (b) to mitigate the interference between neighboring cells, and (c) to increase the spectrum reuse in multi femto-cell networks.

Before we design resource management solutions, we need to understand

the unique design features of OFDMA systems. We summarize several major design features of OFDMA that are different from other OFDM-based systems (e.g., WiFi) as follows.

- OFDMA system has the ability to dynamically assign a subset of sub-carriers to individual users for multiplexing multiple users in the same resource frame. To realize multiple access, OFDMA system divides the frame into multiple sub-frames and provides data transmissions simultaneously while all clients associated with a certain BS receive the same resource frame concurrently.
- In multi-cell deployment, OFDMA small-cells (femto, pico) reuse macro-cell spectrum to increase bandwidth efficiency. Thus, resources allocated to one small-cell directly impact the resources for the interfering cells. In addition, intercell interference is one of the performance limiting-factor if it is not handled well.
- MAC-PHY protocols are required to follow the same synchronous channel access methods. More specifically, rigid frame structure and timing synchronization across neighboring small-cells are essential for operating multi-cells. Moreover, standard OFDMA base stations do not employ carrier sensing to infer interference.

The above features and characteristics should be carefully considered when designing resource management solutions in OFDMA networks.

1.1 FOCUS OF THESIS

Thesis statement

As described, the unique design features of OFDMA systems pose critical challenges for developing resource management/utilization solutions. Considering such characteristics and two-tier architecture employed in OFDMA systems, this thesis tries to address the following question:

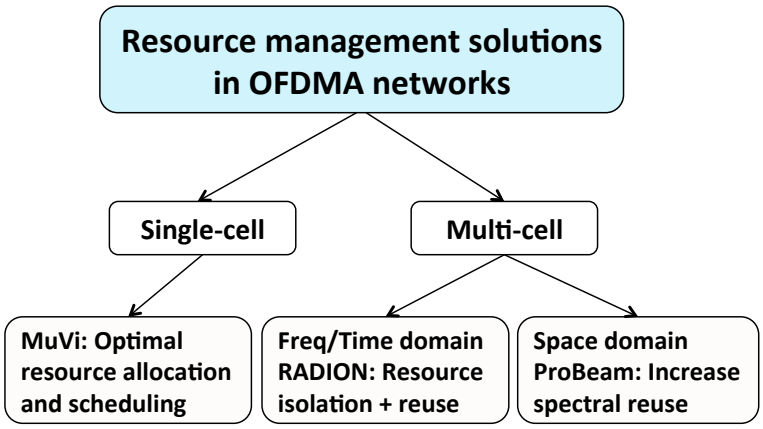


Figure 1.2: Our solutions in this thesis. We mainly focus on several solutions of utilizing spectrum resources in OFDMA networks. We propose three resource management solutions with respect to the deployment of OFDMA small-cells.

How do we design and build practical resource management solutions that improve spectral efficiency in OFDMA networks while mitigating interference and maximizing resource reuse?

More specifically, we would like to build practical resource utilization schemes in both macro-cell and femto-cell networks that efficiently use the available spectrum resources. Figure 1.2 summarizes our approaches to building efficient resource managements in OFDMA networks. Broadly, we explore two-tier, macro-cell and femto-cell, approach. In macro single-cell networks, we propose efficient video multicast scheme that optimizes resource allocation and scheduling in 4G cellular network. In multi femto-cell deployments, we propose two resource management solutions that apply in time/frequency/space domains. First, we design a distributed resource management framework that leverages resource isolation and reuse in time and frequency domains. Second, we develop a coordinated beamforming system to improve spatial reuse in multi-cell networks. We discuss the challenges and the overview of our solutions in detail below.

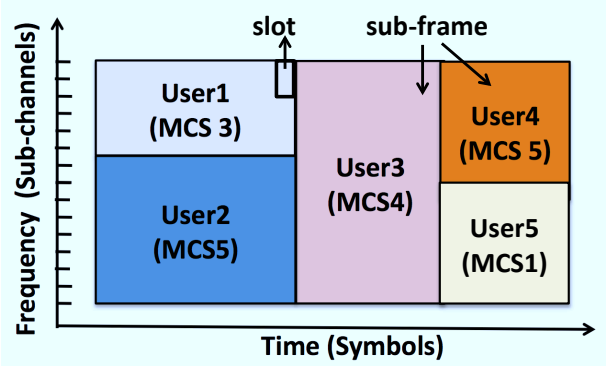


Figure 1.3: Example of resource allocation and MCS selection for scheduling multiple users in OFDMA frame. Both the optimal resource allocation and MCS selections for each user are critical to maximize resource utilization.

Challenge 1: Scheduling heterogeneous clients in an OFDMA resource frame

In a typical wireless environment, clients experience different channel conditions depending on their deployment. For instance, clients near the macro base station may have good channel conditions and be able to successfully decode transmissions with higher PHY rates. On the other hand, others deployed near the macrocell edge may experience poor channel conditions, and hence robust wireless transmission (i.e., low PHY transmission rate) is required to guarantee the successful delivery of data. In addition, these are prone to experience from interference from neighboring base stations as well.

In OFDMA broadband networks, a single OFDMA frame carries data for multiple clients and provides data transmissions simultaneously. For instance, the OFDMA frame is divided into multiple sub-frames (which consists of slots) to schedule multiple clients in TDD (time division duplexing) or FDD (frequency division duplexing) mode. Then, each of the sub-frames are modulated and delivered with various PHY transmission rates (MCS) to designated users as depicted in Figure 1.3. Naturally, when the client's channel condition is good enough to successfully decode data transmission with higher MCS, then more bits can be delivered within resource slots assigned to that user. This eventually

leads to higher resource utilization. In contrast, if the client experiences poor wireless connectivity, then it would hinder higher resource utilization because of the fact that more slots are required to be allocated to users to deliver the same amount of data with lower MCS.

Heterogeneous channel conditions across the clients make it hard to schedule them in the same OFDMA frame. Therefore, optimal resource allocation for users accounting for wireless channel condition is essential to efficiently utilize the resources. Specifically, efficient resource utilization can be achieved through both *optimal resource allocation* and *PHY rate adaptation*. Given heterogeneous conditions across multiple clients, decisions about resource assignment and PHY rate selection have to be client-specific for efficient resource utilization in the OFDMA network.

There are several approaches focus on efficient resource utilization in WiFi network [31, 78, 79]. Unfortunately, the solutions used in the WiFi domain can not be applied to OFDMA systems due to systematic differences (e.g., different channel access mechanism), therefore an OFDMA specific resource allocation solution accounting for each clients' channel condition is necessary to maximize the utilization of given resources. New system design for resource management solution must be discussed in the context of OFDMA systems.

Challenge 2: Pervasive interference in multi-cell deployment

The amount of mobile traffic originated from indoor is significantly proliferated with the rapid growth of handheld devices [38]. To meet the demand for increasing traffic, small-cells are intended to deploy to improve capacity and spectral efficient by reusing same spectrum resources in small area such as home and enterprise indoor environment as opposed to deploying macro base stations. Unlike macrocell base stations managed by the cellular operators, small-cells are deployed in an unplanned fashion within enterprise and residential complexes. Thus, as small-cell deployments continue to grow, interference (between neighboring cells) will inevitably be a performance-limiting factor in a manner similar to that occurs in residential WiFi networks [61].

Consider an example of multi small-cell deployment as shown in Figure

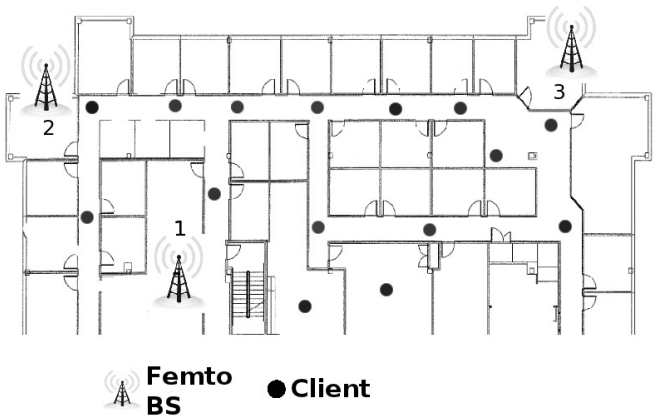


Figure 1.4: Client is associated with BS1 and receives UDP downlink traffic while other two BSs project interference.

1.4. We deploy three WiMAX small-cells in an indoor environment and select 15 locations for the client that represent typical indoor deployment (synchronization across the BSs is achieved through GPS). At each location, the client is associated with the BS1 and receives downlink UDP traffic from BS1 with and without the operation of neighboring BSs (BS2 and 3). In multi-cell deployment, the presence of neighboring BSs causes significant interference because of transmitting preamble and control signals (if they are within the tx/rx range).[1] To quantify the impact of that interference, we measure the client's throughput while varying the number of interferers from 0 to 2.

Figure 1.5 shows the CDF of throughput performance obtained from multiple runs with respect to the number of interferers. We notice that median throughput is around 11 Mbps without any interference, however it drops to 5 Mbps with 1 interferer. Even worse, in 70% of cases the client barely receives data transmission when 2 interferers are presented (throughput < 0.2 Mbps). The presence of preambles and controls from neighboring BSs causes severe interference and hinders data transmission between BS1 and the client. Simi-

[1]Even without data transmission, WiMAX BS periodically (every 5 msec) transmits preamble and control signals to maintain connections (and synchronization) with its clients.

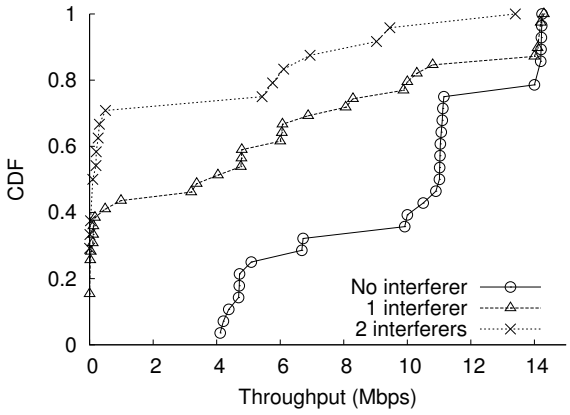


Figure 1.5: CDF of throughput with respect to various number of interferer in three small-cell deployment. Mitigating interference is critical for operating multi cells simultaneously.

larly, impact of interference is summarized in our previous work [23]. There are several cases where the client is able to achieve some throughput (> 4 Mbps) when two BSs project interference, but this is due to the fact that the client is located very close to its own BS and hence, the client receives a strong signal from BS1 that can nullify the inference.

This clearly shows the significance of interference for operating multi small-cells and motivates us to examine the challenges of interference when multi-cells are in range of each other. Thus, resource management solution in multi-cell deployment should define strategies to mitigate the interference and to maximize resource reuse as well.

Our solutions

First, to build practical resource management solutions in OFDMA cellular networks, we examine the challenges by breaking them into the following sub-problems in the context of deployments (macrocell and femtocell) as shown in Figure 1.2. The reason we take different approaches is that the goal of providing a resource management solution with respect to deployment is different in each

case. We then propose solutions that address the problems in various domains such as time, frequency and space.

Since OFDMA systems include multiple heterogeneous clients in the same frame, resource allocation and the selection of transmission rate determine the utilization of resources in a single-cell network. Hence, optimal decisions considering clients' profiles for scheduling could lead to the increased performance of the system. Towards this end, we develop and implement an efficient video delivery scheme, MuVi, in OFDMA 4G cellular networks that optimizes resource utilization when the clients experience various channel conditions (Chapter 2). Specifically, we design a joint optimal resource allocation and MCS selection algorithm to maximize the overall video quality across all users in the multicast group.

In multi-cell deployment, the interference from neighboring cells is pervasive and severely impacts the performance of interfering cells. Thus, a resource management solution should handle the interference across the small-cells properly and maximize the spatial reuse to improve capacity. To meet these two purposes, we design and implement RADION, a distributed resource management framework that addresses the challenges delineated above in the time and frequency domains (Chapter 3). RADION maximizes resource reuse while providing resource isolation when neighboring cells are within transmission range of each other and would otherwise cause interference. A self-organizing resource management framework works well, especially in uncoordinated and unplanned multi-cell deployment where centralized control is not feasible, however, this approach has to scarify resources in either the time or frequency domain to completely mitigate the interference.

This motivates us to investigate the problem from a different angle, in the space domain, where the resource isolation is not required at all. Therefore, we can maximize the resource reuse in each small-cell. For this, we develop ProBeam, a practical multi-cell beamforming system that leverages smart antenna to increase spatial reuse in the space domain (Chapter 4). In ProBeam, each cell employs a directional antenna to create beamforming that is intended for its clients in order to realize the throughput gains from spatial reuse. We also consider a flexible client association approach that jointly addresses client

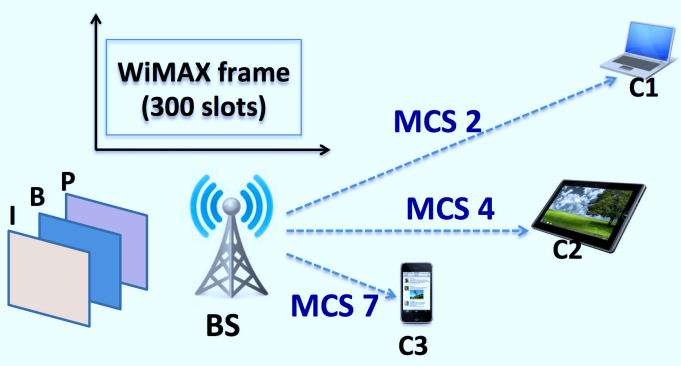


Figure 1.6: Three clients deployment example where each client experiences a different channel condition (in terms of supported MCS).

association with beam selection for small-cells, whereby client association can be effectively used to maximize the spatial reuse potential of beamforming.

In the next few sections, we expand on each of our solution in detail.

1.2 VIDEO MULTICAST WITH OPTIMAL RESOURCE ALLOCATION AND MCS SELECTION

Since a single OFDMA frame carries data for multiple clients, resource allocation for each client has to be carefully handled to maximize resource utilization. Besides resource allocation, transmission rate selection for sub-frames is also important because different clients experience different wireless channel conditions in their deployments. For example, some clients might have good channel conditions and be able to successfully decode transmission with higher PHY rates (MCS), while others have poor channel conditions and hence robust wireless transmissions are required to guarantee successful delivery of data. Higher resource utilization can be achieved when using higher transmission rate, however clients with bad channel can not receive OFDMA frames modulated and delivered with high MCS. Therefore, both resource allocation and selection of transmission rate have to be carefully optimized to efficiently utilize the frame resources considering client's profile. Resource allocation in single-cell networks has two goals: 1) utilizing resource frames efficiently and 2) providing

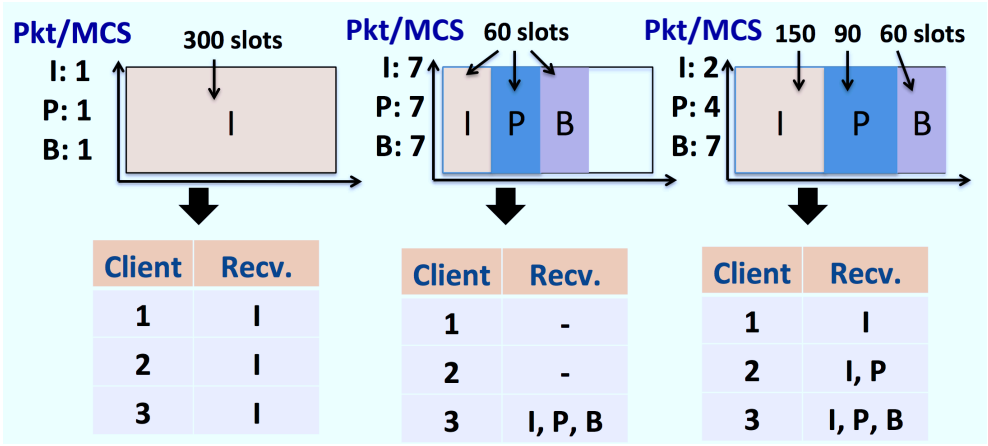


Figure 1.7: Decisions of resource assignment and MCS selection result in different receptions of video frames.

satisfactory services to heterogeneous clients simultaneously.

Consider an example wherein three clients are associated with BS and receive video traffic, as depicted in Figure 1.6. The deployment of these clients inherently gives them different channel conditions; each client can receive sub-frames modulated and transmitted with MCS up to 2, 4, and 7, respectively. The BS intends to send three video frames, I, P, and B packets, over a single resource frame which consists of 300 slots. BS needs to decide which modulation scheme (MCS) to use for transmitting video frames to all three clients. Note that the selection of MCS affects the number of slots required for delivering data. Higher MCS can deliver more data within the same number of slots compared to lower MCS. In other words, more slots are required to transmit the same video packet when lower MCS is adapted. The clients channel condition also determines whether the client can receive delivered WiMAX sub-frames according to the MCS used to send such sub-frames.

In Figure 1.7, we present three cases that use various MCS and allocated a different number of slots for sending video packets over WiMAX frame. We also include three results that present the receptions of video frames for each client with respect to the three cases. First, when BS uses the lowest MCS (1)

to guarantee the delivery of sub-frames, only I-frame is scheduled within the given WiMAX frame (300 slots). All three clients receive I-frame, however P and B-frames are not scheduled because of limited slots in a frame. Second, if BS picks the highest MCS (7) to maximize the resource utilization, then all three I, P and B-frames can be delivered in a single WiMAX frame, but only client 3 will receive all of them. The other two clients are not able to decode transmitted WiMAX sub-frames due to their poor channel conditions. The ideal case can be achieved when using different MCS for each video packet considering the priority of video frames. Specifically, lower MCS is used for important video frames to guarantee its delivery to all clients and higher MCS is used to transmit less important video packets such as P and B-frames in order to maximize resource utilization.

Both *resource allocation* and *MCS selection* play an important role in determining resource utilization in OFDMA systems. These two decisions have to incorporate the clients' profiles for scheduling. As shown in the above example, wireless video multicast is an application that requires both efficient resource assignment and MCS selection to provide satisfactory services to multiple users in single-cell networks. For this, we develop and implement an efficient Multicast Video delivery scheme (MuVi) in OFDMA 4G cellular networks. MuVi differentiates video frames based on their dependency in reconstructing the video and incorporates an efficient radio resource allocation algorithm to optimize the overall video quality across all users in the multicast group. In Chapter 2, we present the detailed description of MuVi that jointly consider optimal resource allocation and adaptive modulation and coding scheme for efficiently multicasting video traffic in 4G cellular networks.

1.3 DISTRIBUTED RESOURCE MANAGEMENT FRAMEWORK IN MULTI FEMTOCELL NETWORK

Femtocells present an alternative approach for increasing spectral efficiency (reuse all resources) by reducing transmission ranges, as opposed to deploying macro base stations. Unlike the planned deployment of macrocell base stations by the cellular operators, femtocells are deployed in an unplanned manner within enterprise and residential complexes. As small-cell deploy-

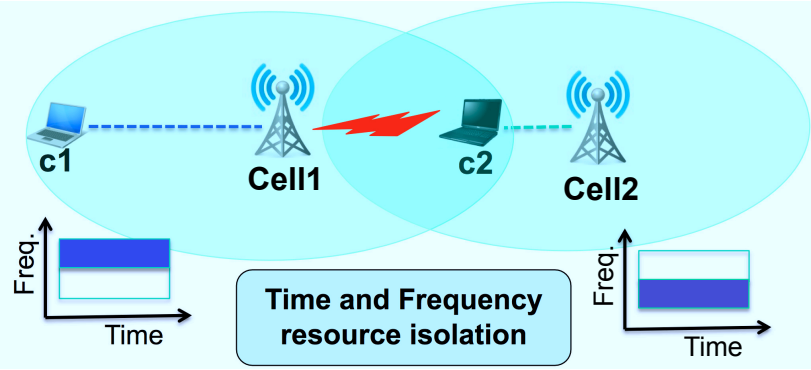


Figure 1.8: Resource (frequency) isolation: Cell1 is operating on the upper half of frequency resources while Cell2 is on the lower half of frequency. Interference can be alleviated by allocating orthogonal resources to interfering femtocells.

ments continue to grow, interference between neighboring cells is pervasive as shown in the experiment result in Challenge 2, thus interference will be a performance-limiting factor.

The problem of interference mitigation through efficient resource management has been studied in different domains. However, the unplanned deployments of femtocells make solutions designed for the well-planned macrocells inapplicable; for example, the fractional frequency reuse (FFR) [35] approaches for macrocells are ineffective in femtocells where interference is pervasive and unpredictable. Similarly, solutions designed for WiFi cannot be directly applied, since femtocells have to operate synchronously to maintain compatibility with cellular standards. More specifically, standard OFDMA BSs do not employ carrier-sensing-based deferral and they operate on a licensed spectrum; hence, they cannot sense the spectrum occupied by other femtocells and tune themselves onto orthogonal frequencies in order to resolve interference. Thus, resource management solutions for macrocells or WiFi cannot be directly applied in the femtocell context. This motivates us to design a new resource management approach that is tailored to the characteristics of OFDMA femtocells.

In OFDMA multi-cell networks, two main approaches can be taken to mitigate interference: (i) in the time and frequency domains and (ii) in space domain. As a part of the first approach, we propose RADION, a framework for distributed management of time-frequency resources in OFDMA femtocell networks. Figure 1.8 shows an example of frequency management when two femtocells are in the same transmission range, therefore, they are not able to reuse all frequency resources. By allocating orthogonal frequency resources to interfering cells, we can allow them to operate simultaneously without causing interference. In contrast, femtocells can reuse all resources if they do not introduce interference to neighboring cells. *Resource isolation* and *reuse* are the main ideas behind this resource management solution in a multi-cell network, a solution that both alleviates severe interference and maximizes resource reuse.

This framework, RADION, is designed for a scenario where nearby femtocells cannot explicitly coordinate or interact with each other, and hence is suitable for unplanned residential deployments. Our solution requires each femto base station (BS) to intelligently probe the availability of resources and use them in an opportunistic and distributed manner. Hence, the solution in RADION is stylized such that each residential femto BS will try to optimize for its own local objective without explicitly exchanging information with neighboring femtocell BSs. More importantly, the design of the framework, which can integrate with various distributed resource management solutions, provides more flexibility to account for various deployment scenarios. RADION's core building blocks enable femtocells to opportunistically determine the available resources in a completely distributed and efficient manner. In Chapter 3, we elucidate resource management challenges arise when operating multi-cells and how we address them in RADION.

1.4 COORDINATED BEAMFORMING AND CLIENT ASSOCIATION IN SMALL-CELL NETWORK

Small-cells form a critical component of next generation OFDMA cellular networks, where spatial reuse is the key to higher spectral efficiencies. Beamforming adjusts transmissions at the antenna to enhance the signal strength received at the receiver [85]. For instance, by radiating transmit signal energy

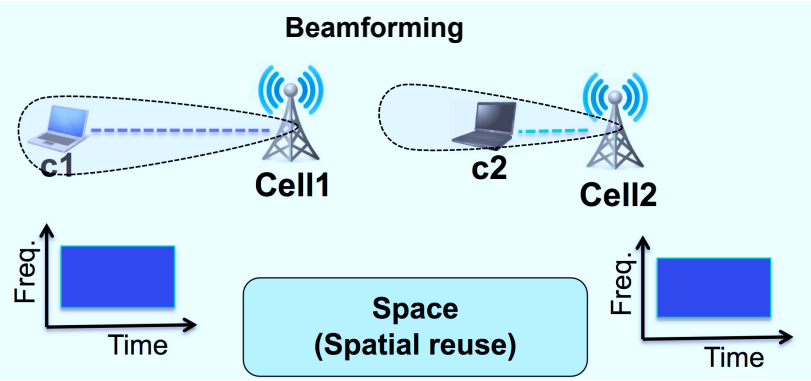
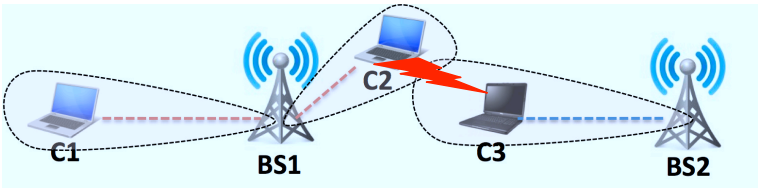


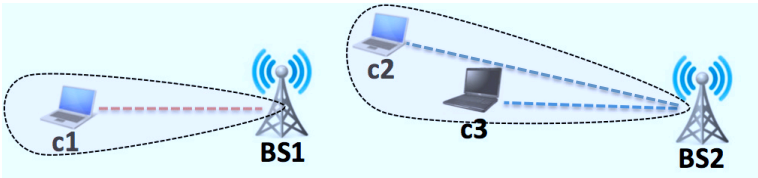
Figure 1.9: Example of beamforming in multi-cell networks. The small-cells radiate transmit signal energy towards its client. Both Cell1 and Cell2 can be operating utilizing all frame resources without causing interference.

in a specific direction using a directional antenna, we can increase spatial reuse so that neighboring femtocells can be operating on all available resources at the same time as described in Figure 1.9. Interference management in the spatial domain through beamforming allows for increased reuse without having to sacrifice resources in the time or frequency domain.

Client association is also an important factor in efficient resource utilization, given that OFDMA systems provide simultaneous services to multiple clients by allocating sub-frame resources. Thus, decoupling beamforming from client scheduling is critical for managing multi-cells because different client association leads to different beam selections across BSs. To see this, consider the illustration in Figure 1.10. In conventional association, where SNR is used as a metric for client association, clients C1 and C2 will be associated to BS1, while C3 will be associated to BS2 based on high SNR. BSs will then determine the best beams to communicate with their respective clients. When BS1 is employing a beam to communicate with C2, this will receive interference from the beam used by BS2 to communicate with its client C3. By fixing the client association, the ability of beamforming to effectively suppress interference between cells is limited. In contrast, by allowing for flexible association shown in Figure



(a) C2 is associated with BS1



(b) C2 is associated with BS2

Figure 1.10: (a) BS2 causes interference to C2. (b) Both BS1 and BS2 can be operating simultaneously without causing interference.

1.10b, C2 can be associated with BS2 even though it has a lower SNR than BS2. This would allow BS2 to schedule C2 and C3 jointly on a beam that suffers no interference from the beam employed by BS1, thereby resulting in a potentially higher SINR for all clients.

Existing beamforming techniques for spatial reuse, being coupled with client scheduling, face a key limitation in practical realization especially with OFDMA small-cells. In this context, we argue that in order to create a practical spatial reuse system with beamforming, it is important to decouple beamforming from client scheduling. Towards this, we propose *ProBeam* – a system for multi-cell beamforming and a client scheduling algorithm to maximize the network performance in multi femtocell networks. Considering the clients' deployment, a central controller jointly adapts beam patterns for each femtocell along with client association in order to improve link quality while increasing resource reuse. ProBeam incorporates two key components, SINR estimation module and joint client association and beam selection algorithm. In Chapter 4, we describe the details of ProBeam.

1.5 THESIS CONTRIBUTIONS

Although wireless broadband technologies have significantly evolved to provide higher bandwidth over the past decade, they are still insufficient to support fast-growing mobile traffic. The broadband network operators intend to deploy smaller cells to maximize the reuse of the spectrum to satisfy the demand for increased data rates. However, unmanaged deployment of small-cells introduces pervasive interference and hence, interference mitigation and efficient resource utilization are critical to improve the spectrum efficiency in wireless networks. The major contribution of this thesis lies in designing and implementing practical resource management solutions for efficient resource utilization in OFDMA-based cellular networks. We take various approaches in multiple domains (e.g., time, frequency and space domains) for developing efficient resource management solutions where they can be adapted to in OFDMA systems. We summarize our contributions of this thesis as follows:

- We design an algorithm for both allocating resource and selecting MCS in a single-cell OFDMA network. Based on proposed algorithm, we implement video multicasting scheme, MuVi. We adapt dynamic programming approach to jointly determine resource allocation and PHY rate selection for optimizing multicast video delivery. MuVi is the first practical video multicasting solution running on WiMAX testbed and is a lightweight solution with most of the implementation in the gateway along with slight modification in the base-station. Experimental results show that MuVi improves the average video PSNR by up to 13 dB and 7 dB compared to the Naive and the Adaptive schemes, respectively. MuVi does not require modification to the video encoding scheme or the air interface, thus it allows speedy deployment in existing systems.

- We design and implement RADION, a distributed resource management framework running on real WiMAX testbed, consisting of three femtocells in an unplanned indoor environment. To the best of our knowledge, this is the first implementation based design and evaluation of a self-organizing framework for OFDMA femtocells. We first identify challenges

of operating multi small-cell network and design several building blocks to address them. The underlying notion of building blocks can be adapted to other OFDMA system, LTE and LTE-A. We wish to point out that RADION is modular. In particular, depending on the context and the objectives, different resource allocation solutions can be easily incorporated within the RADION framework.

- We propose and implement ProBeam, a system for multi-cell beamforming and client association in OFDMA small-cell networks. ProBeam incorporates two key components - a low complexity, highly accurate SINR estimation module that helps determine interference dependencies for beamforming between small-cells; and an efficient, low complexity joint client association and beam selection algorithm for the small-cells that accounts for scheduling at the small-cells without being coupled with it. Note that accurate SINR estimation would require measurement with respect to all possible combination of beam choices at small-cells, thus it results in exponential complexity. We highlight that ProBeam's SINR estimation module requires only linear number of measurements with an estimation error less than 1 dB for 95% confidence. In addition, our flexible client association does not follow conventional approach for scheduling client in multi APs or small-cell deployments and we show its efficacy. We have prototyped ProBeam on a WiMAX network of four small-cells and evaluations reveal the reuse gains from joint client association and beamforming to be as high as 115% over baseline strategy.

1.6 OUTLINE

The rest of the thesis is organized as follows. In the first part of this thesis, we focus on an optimal resource allocation in a single-cell network. We present MuVi, a multicast video delivery solution through joint optimal resource allocation and adaptive modulation and coding scheme in 4G cellular networks (Chapter 2). In the second part of the thesis, we focus on resource management solutions in a multi small-cell deployment where interference is a major factor in network

performance (Chapters 3 and 4). Specifically, in Chapter 3, we present RADION, a distributed resource management framework in multi small-cell networks. RADION effectively manages interference across small-cells while maximizing spectrum reuse in both the time and frequency domains. In Chapter 4, we present ProBeam, a practical multicell beamforming system in small-cell networks. ProBeam leverages beamforming for both mitigating interference and improving spatial efficiency by reusing resources across neighboring small-cells. In Chapter 5, we compare our works with other approaches to resource management and utilization in both cellular and WiFi networks. We conclude and describe future research topics in Chapter 6.

2 VIDEO MULTICAST WITH OPTIMAL RESOURCE ALLOCATION IN OFDMA-BASED CELLULAR NETWORKS

2.1 INTRODUCTION

OFDMA networks multiplex multiple clients in the same resource frame. In this sense, both the frame utilization and client scheduling are major factors that impact the performance of network. Therefore, resource allocation in single-cell networks has two goals: 1) utilizing resource frames efficiently and 2) providing satisfactory video services to heterogeneous clients simultaneously. Wireless video multicast is one good example of an application that requires both efficient resource allocation and modulation and coding scheme (MCS) adaptation to provide satisfactory services to multiple users in single-cell networks. Here, we develop and implement an efficient multicast video delivery scheme (called MuVi) in OFDMA mobile broadband wireless networks.

Motivation

Mobile video streaming (e.g., Netflix, Youtube) is one of the most popular applications in recent years and the amount of video traffic destined for mobile devices is increasing rapidly as the number of hand-held devices (smart phones, tablets) grows. According to a recent survey [38], video accounted for about 40 percent of consumer Internet traffic in 2010 and will reach 62 percent by the end of 2015. Global mobile data traffic is expected to increase 26 times from 2010 to 2015. Moreover, high-definition (HD) video surpasses standard definition video and will becomes the dominant form of video traffic. Current WiFi systems can not provide satisfactory quality of video streaming services due to the small coverage and relatively limited bandwidth as the number of mobile users increases. Even worse, WiFi networks are not robust enough to sustain user mobility. 3G mobile networks such as CDMA and UMTS can provide more robust wireless connections to mobile users. However, their bandwidth is not adequate to support applications with high bandwidth requirements (HD video). Limitations of current WiFi and 3G systems naturally turn our

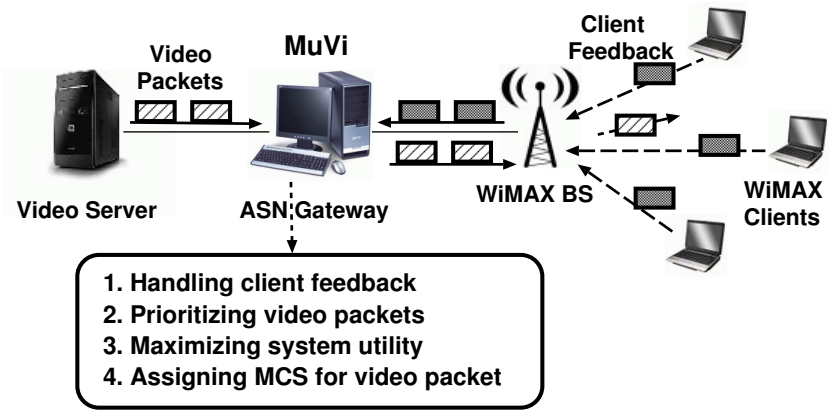


Figure 2.1: MuVi performs four operations while receiving video packets from the media server, and then it delivers video packets to the WiMAX base-station.

attentions to emerging 4G cellular networks.

4G cellular networks, such as WiMAX and LTE, have emerged as alternatives that can provide much higher bandwidth, spectrum efficiency, and extended coverage. 4G networks are more robust to user mobility compared to WiFi systems, so that they can provide seamless real-time video streaming services. More specifically, 4G technologies can provide peak data rate of 100 Mbps for high mobility applications and 1 Gbps for nomadic applications [40]. Despite the much higher bandwidth provided by 4G technologies, efficient resource utilization is still needed for meeting the high bandwidth and stringent delay requirements of video applications, because the wireless spectrum is shared by multiple users. To meet the requirements, we develop a multicast video delivery scheme.

MuVi Overview

MuVi is designed as a proxy in the Access Service Network (ASN) gateway as depicted in Figure 2.1. There are four key design elements in MuVi. (i) MuVi collects the feedback of wireless channel quality information of all clients in a multicast group through the base-station and obtains the supportable MCS for each client. (ii) MuVi prioritizes video frames by setting a utility value for each

frame based on the frame type. The utility may also depend on the user profile. Thus, MuVi supports user differentiation. (iii) MuVi performs efficient resource allocation to maximize the system utility considering both the available radio resources and the video packets to be transmitted. (iv) MuVi assigns MCSs for the video packets and hands them over to the base-station, which sends the video packets over the air using the assigned MCSs.

Different from several recent works on video streaming [16, 58] which provide graceful video delivery, MuVi does not modify the video encoding nor the wireless transmission schemes. It performs resource allocation in operational wireless systems and video frame prioritization with existing video encoding technologies to optimize the resource utilization and video delivery with the commercial off-the-shelf clients (e.g., WiMAX card). MuVi is also different from several WiFi-specific video optimizations which utilize the notion of differential value of data packets, e.g., Medusa [90], in that 1) MuVi is designed for OFDMA-based systems, and 2) MuVi does not require client side modifications. Since it does not modify the air interface or the clients, MuVi is standards compatible and allows immediate deployment in commercial WiMAX systems.

In summary, we have the following findings.

- MuVi improves the average video quality by up to 13 dB and 7 dB in terms of PSNR compared to the Naive scheme and the Adaptive scheme, respectively.
- MuVi reduces the inter-packet arrival delay by up to 80% and 55%, compared to the Naive and the Adaptive schemes, respectively. This indicates significantly fewer glitches and video stalls for MuVi.
- MuVi is a gateway solution and does not require any modification on the client side. It can be easily incorporated in cellular networks with only minimum modifications on the base-stations.
- We design and implement MuVi, arguably the first *practical* video multicast scheme for OFDMA systems. Although it is implemented in WiMAX, MuVi is also applicable to other OFDMA-based systems, e.g., LTE and LTE-Advanced systems.

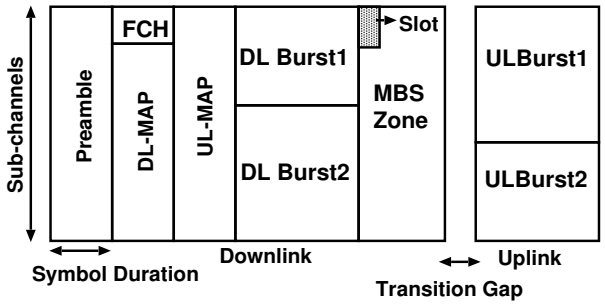


Figure 2.2: WiMAX frame structure.

2.2 WIMAX PRELIMINARIES

While our work applies to Orthogonal Frequency Division Multiple Access (OFDMA)-based networks in general, we implement MuVi and conduct experiments on a WiMAX (802.16e [8]) testbed due to its availability. Here, we give a brief overview of OFDMA and WiMAX networks.

In OFDMA-based WiMAX system, the spectrum is divided into multiple orthogonal sub-carriers and several sub-carriers are grouped to form a sub-channel. Grouping of sub-carriers determines the channel diversity in the system. There are two modes of grouping: distributed and contiguous grouping. In *distributed grouping*, sub-carriers are picked as per a pre-determined permutation to compose a sub-channel. This allows a user to see uniform gain and interference across different sub-channels. It requires a single channel feedback from the user. In *contiguous grouping*, a contiguous set of sub-carriers is grouped to form a sub-channel. While this preserves channel diversity, it also requires per sub-channel feedback overhead from each user. Distributed grouping, the default mode in 802.16e standard, is considered in our work.

WiMAX has a two-dimensional frame (see Figure 2.2) that carries data across time (symbols) and frequency (sub-channels). A combination of a symbol and a sub-channel constitutes a *slot*, which is the basic unit of resource allocation. Data to multiple mobile stations (MSs) are scheduled as rectangular bursts of tiles in a frame and frames are sent out periodically (every 5 milliseconds). In

Index	Modulation and Coding Rate
0	BPSK
1	QPSK(1/2)
2, 3	16 QAM(1/2, 3/4)
4, 5, 6, 7	64 QAM(1/2, 2/3, 3/4, 5/6)

Table 2.1: MCS indices available for data modulation and coding rates.

multi-cell OFDMA systems, frames are synchronized both between the BS and MSs and across BSs.

A frame consists of a preamble, control data and payload data. The preamble is transmitted with power boost and allows mobile clients to lock on to the base-station. The control signaling consists of a Frame Control Header (FCH) and a MAP. The FCH contains the system control information and the information about the MAP. The DL-MAP indicates where each burst is placed in the frame, which client it is intended for, and what modulation and coding scheme (MCS) decodes it. Table 2.1 shows the different modulation levels employed in a WiMAX base-station. Similarly the UL-MAP indicates where the client should place its data on the uplink frame. The uplink sub-frame has dedicated sub-channels for HARQ, which are used by clients to explicitly acknowledge (ACK/NACK) the reception of each data burst. Clients also use the uplink sub-frame to report the instantaneous channel state information (CSI) to the base-station. Base-stations use a dedicated data burst zone, called MBS zone (in Figure 2.2), for multicast and broadcast services. While data bursts can be modulated using any MCS, the control parts (FCH and DL/UL MAPs) are always transmitted using QPSK and typically with multiple repetitions. The preamble is transmitted with a higher power compared to the other parts of the frame.

2.3 DESIGN AND OPERATION OF MUVI

Our proposed system, MuVi, comprises a media server, media-aware gateway, base-station, and multiple clients in the multicast session as depicted in Figure

2.1. The basic operations of MuVi are as follows:

1) The WiMAX clients send the channel state information (CSI) to the base-station periodically. Providing CSI is a standard mechanism in 802.16e (and all mobile broadband systems). Upon receiving the channel feedback, the base-station computes appropriate moving average for each client, aggregates the averaged feedback, and forwards them to the ASN gateway. The ASN gateway then determines the supportable MCS for each client based on their average channel conditions, along with an MCS table.

2) The media server sends the video packets to the ASN gateway, which performs packet scheduling and possibly drops some video packets based on the available radio resources and the frame types. The packet scheduling includes packet re-ordering and determining the MCS, i.e., the PHY rate, employed at the base-station for each video packet.

3) After receiving the video packets, the base-station applies the MCS selected by the ASN gateway and transmits them over the air using multicast.

4) When the clients receive the packets, they perform packet re-ordering and then pass the received packets to the video player (e.g., VLC [14]). Packet re-ordering is typically implemented in upper-layer protocols such as RTP (Realtime Transport Protocol).

To execute the above operations, we modify the ASN gateway and the base-station. Our major modifications lie at the ASN gateway, which consist of 1) collecting the supportable MCS for each client, 2) classifying packet types (i.e., video frame types), and 3) determining the PHY transmission rate for each video packet.

The ASN gateway can perform deep-packet-inspection to find the type of video frames, but this may cause extra overhead at the ASN gateway. Hence, we modify the video player application (VLC player) running on the media server to ease the job at the ASN gateway. The customized VLC player inserts the video frame type information in the IP header (e.g., DiffServ field) before

sending video packets to the ASN gateway. The ASN gateway simply looks up the IP header to categorize received packets.

The base-station needs to make small modifications to 1) forward the client channel feedback to the ASN gateway, and 2) transmit the packets using the MCS instructed by the ASN gateway. We believe these are within the scope of what BS vendors are willing to do. The clients are not required to change given that re-ordering can be performed by upper-layer protocols such as RTP, which is typically employed in real-time multicast streaming.

In the following subsections, we describe the details of each design element.

Handling Client Feedback

The wireless channel suffers from packet losses and errors caused by interference and noise. It is hard to determine the target MCS (i.e., PHY transmission rate) in a wireless multicast system because of the different channel conditions amongst multiple users in the multicast group. Client channel condition is an important factor to determine the target PHY rate to multicast while avoiding resource under-utilization and packet loss. To guarantee data delivery over wireless channels, traditional wireless multicast systems use the most robust MCS (i.e., the lowest PHY rate). However, this approach leads to under-utilization of available radio resources when most multicast users can support higher MCS than the lowest one. A carefully selected MCS can improve the goodput and the video quality of a client. The most appropriate MCS for each client depends on its channel conditions. MuVi leverages the information of the supportable MCS for each client to make efficient use of radio resources for multicasting.

To find the relationship between the client channel condition and the supportable MCS, we measure the client throughput from various locations where the channel conditions vary. We deploy the clients on several locations where clients experience different levels of channel conditions in terms of Carrier to Interference plus Noise Ratio (CINR). For each location, base-station sends UDP traffic using *iperf* to the client while trying out all possible MCSs and the client records the obtained UDP throughput with respect to each MCS used. We plot

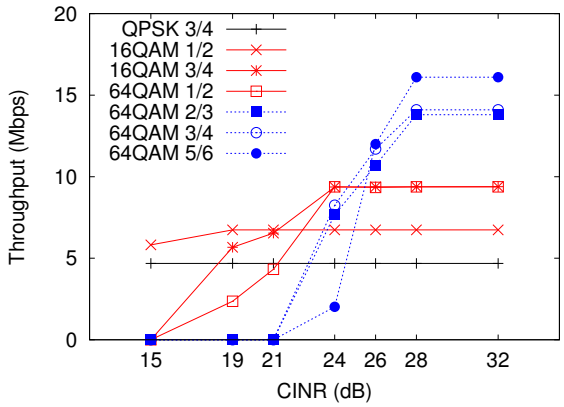


Figure 2.3: MuVi can find the supportable MCS for a client based on its CINR feedback.

MCS index	CINR range (dB)
7	(28, ∞)
6	(26, 28]
5	(24, 26]
3, 4	(20, 24]
2	(15, 20]
1	(-∞, 15]

Table 2.2: CINR range and the corresponding MCS. MuVi selects MCS based on CINR feedback.

the measured throughput for all locations where each location is represented as a point in Figure 2.3. Based on this result, we build an MCS-CINR table as shown in Table 2.2. For example, a client can successfully decode packets transmitted with MCS 4 (64-QAM with 1/2 coding) or lower when its observed CINR is 24 dB. We comment that constructing the MCS-CINR table only needs to be done once.

For clients, providing channel feedback (CINR report) to base-station is part of the IEEE 802.16e (WiMAX) standard [8]. While the clients use dedicated uplink channel to report feedback to the base-station, it does not interfere

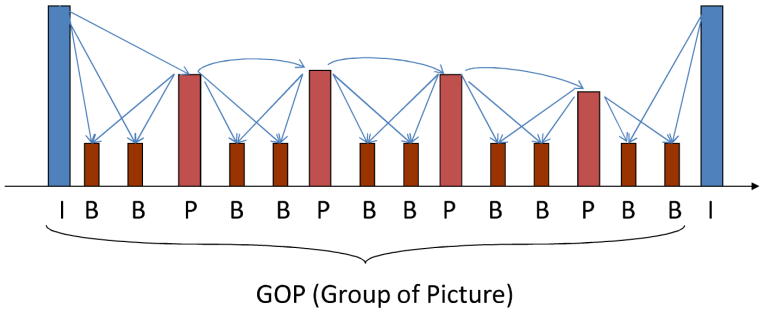


Figure 2.4: An example video frame sequence in a GOP. An arrow indicates the frame dependency.

with the downlink video traffic. We modify the MAC code of the WiMAX base-station to compute the moving average of the CINR values for each client and forward them to the ASN gateway periodically. Once the ASN gateway receives the clients' CINR from the base-station, it can identify the highest MCS that can be decoded by the client using the Table 2.2. We point out that obtaining CINR from clients is critical for optimizing utility and enhancing video quality, however it does not reduce the scalability of MuVi. Typically in the multicast session, groups of client form clusters based on their wireless channel conditions. Given that, it is not necessary to gather feedback from all clients in the multicast group. In other words, the fraction of client for different MCS level would be sufficient to calculate the utility function for system as the clients have similar channel feedback.

Packet Values

We describe the characteristics of MPEG4/H.264 video frames, since we use MPEG4 as example video encoding schemes. In MPEG4 or H.264, a video sequence is partitioned into Group of Pictures (GOP). Figure 2.4 shows an example of frame sequence in a GOP. Each GOP consists of a certain number of I, P, and B video frames, which are then further divided into packets typically

with fixed length except for the last packet of each video frame. I frames are intra-coded pictures and it can be decoded independently. P frames are forward predicted pictures and require their preceding I or P frames to be decoded. B frames are bidirectionally predicted pictures and require both preceding and succeeding I or P frames for decoding. The arrows in Figure 2.4 represent the dependencies between frames. Each video frame contains different video/audio information. For example, P and B frames carry only difference information so they depend on some other frames (i.e., I and P frames) to be decoded successfully. A typical GOP consists of 30 frames of repeated sequences (e.g., IBBPBBPBB...PBB).

MuVi prioritizes the packets based on their dependency and assign the priorities as a function of number of packets depending on them. The more packets depend on it, the higher priority is assigned to a packet. Typically, I frames have the highest priority and B frames have the lowest priority. P frames may have different priorities depending on their positions in a GOP. For example, a P frame in the earlier part of a GOP has more number of dependent frames compared to that in the later part of a GOP, hence the former has higher priority than the latter.

In addition, users may have different priorities. For example, a high-profile user may pay higher subscription fee and expect high video quality. So we allow a video frame (i.e., the packets) to have different priority values for different clients.

Utility Maximization

The main media optimization engine lies in the ASN gateway. It collects the channel feedback for each client in the multicast group from the base-station and determines the resource allocation for the video frames in order to maximize the total system utility.

The problem setup is as follows. There are Q available slots in a WiMAX frame for the video multicast session, where a slot is a two-dimensional minimum allocation unit in the time and frequency domains in an OFDMA frame. The OFDMA frame duration is τ ms ($\tau = 5$ in most WiMAX systems) therefore

there are W OFDMA frames per second ($W = 200$). There are M Modulation and Coding Schemes (MCS) representing different transmission rates. For MCS m, one OFDMA slot can deliver R_m bytes and the minimum required CINR for decoding MCS m is $\bar{\gamma}_m$, where $1 \leqslant m \leqslant M$. There are K clients in the multicast session and their CINR values are $\gamma_1, \gamma_2, \cdots, \gamma_K$.

We consider the transmission of a Group of Picture (GOP) with J video frames. A GOP includes an I frame, a sequence of P frames, and a group of B frames as depicted in Figure 2.4. The jth video frame has length L_j. Receiving frame j at client k obtains a utility u_j^k given that all frames it depends on are also received by client k. By including a superscript k in the utility function, we allow different priorities among different users. We assume that the inter-frame time is Δ second and thus the total number of slots available for transmitting the GOP is $T = \lfloor \Delta J Q / \tau \rfloor$. The objective is to maximize the total utility received by all clients subject to the total slot constraint.

The dependency relationship between I/P frames and B frames is different. All frames in the GOP need to refer I frame when decoding. P frames depend on their preceding P frames and are needed by both intermediate B frames and succeeding P frame. B frames are not referred by other frames. Considering different frame dependencies, we first schedule groups of I/P frames and B frames separately and then optimize the overall utility across the two groups.

I/P Frame Scheduling

We first consider the problem of scheduling only I and P frames in a GOP. For notational convenience, we assign a special sequence number P_0 to the only I frame in the GOP and hereafter, we treat I frame as a special P frame P_0. So we have a sequence of P frames P_j, where $j = 0, 1, \cdots, G$, such that frame P_{j+1} depends on P_j ($0 \leqslant j \leqslant G-1$), where G is the number of P frames. Let the total number of available slots for P frames be $t \leqslant T$ and the objective is to maximize the total utility subject to the total slot constraint t for all P frames.

To ensure minimum video quality at every client, we require that the first $j_0 \geqslant 0$ P frames be received by all clients, and set aside t_1 slots for transmitting these P frames using the highest MCS that can be decoded by all clients. Thus

the available number of slots for the remaining P frames is $t' = t - t_1$. Without loss of generality, we assume $j_0 = 0$ (i.e., no P frame is required by all clients) in the following.

Let binary variable x_{jm} indicate that frame P_j is transmitted with MCS m and z_j^k indicate that frame j is valid for client k, which means that all frames on which frame j depends, including j, can be decoded by client k, so

$$z_j^k = \begin{cases} 1, & \text{if for all } 0 \leqslant l \leqslant j, \text{ there exists } m \\ & \text{such that } x_{l,m} = 1 \text{ and } \bar{\gamma}_m \leqslant \gamma_k, \\ 0 & \text{otherwise.} \end{cases}$$

Now the problem of utility maximization can be written as

$$\begin{aligned} \max \quad & \sum_{j=0}^{G} \sum_{k=1}^{K} z_j^k u_j^k \\ \text{s.t.} \quad & \sum_{m=1}^{M} x_{jm} \leqslant 1, \\ & \sum_{m} \left\lceil \sum_{j} \frac{x_{jm} L_j}{R_m} \right\rceil \leqslant t, \end{aligned} \tag{2.1}$$

where in the last equation, we assume that frames transmitted using the same MCS can be bundled together for the allocation of radio resources (i.e., slots). We slightly abuse the notations by using u_j^k to denote the utility of frame P_j for client k.

Interestingly, problem (2.1) is similar to the problem of maximizing sum of video quality when multicasting SVC-encoded videos by viewing the jth P frame as the jth layer in SVC video. In deed, the dependency relationship between P frames in non-scalable videos are equivalent to that between different layers in SVC videos. Therefore, we can leverage a dynamic programming algorithm similar to the one used in SVC video multicasting [70] to solve problem (2.1).

Theorem 2.1. *The optimization problem (2.1) is NP-hard*

Proof. Our optimization problem (2.1) is a special case of one-group WiMRA problem in [42] considering P frame dependencies in our problem as multi layers in SVC video. The goal of the one-group WiMRA problem is to maximize the system utility $\sum_{j\in N} \log(\lambda_b + h_j\lambda)$ under the resource constrain R' where h_j is the number of enhancement layers that user j can receive, λ_b is the transmit rate of the base layer and λ is the rate of the enhancement layers. The one-group WiMRA problem is a special case of our problem (2.1) with the number of P frames $G = K$, MCS for each frame $m = \lambda$ and the utility $z_j^k u_j^k = \log(\lambda_b + h_j\lambda)$. Hence, the problem (2.1) is also NP-hard. □

We now develop an efficient algorithm to solve problem (2.1). We first have the following lemma.

Lemma 2.2. *There exists an optimal solution to (2.1) such that the MCS used for frame* P_{j-1} *is lower than or equal to that for frame* P_j*, if both are transmitted, where* $0 < j \leqslant G$.

Proof. Suppose in an optimal solution, frame P_{j-1} and P_j is transmitted with m' and m, respectively, where $m' > m$. Then frame P_j, along with all its succeeding P frames using MCS m, can be transmitted with MCS m' without decreasing the total utility while the total number of used slots does not increase. Therefore, this new transmission scheme is also optimal while satisfying the requirement in the Lemma. □

This partial ordering of the MCS for video frames is crucial for developing dynamic-programming-based solution to problem (2.1). From this lemma, if the frame P_j is transmitted with MCS m, all the preceding frames $P_l, 0 \leqslant l < j$ should only use MCS m or lower. Therefore, the MCSs employed for all P frames are non-decreasing with respect to the P frame indices. This property is crucial to develop a dynamic programming based algorithm.

Define $U_P(j, m, t)$ as the optimal utility with the P frames $P_l, l = 0, \cdots, j$ with MCS up to m and at most t slots. Let $\tau_{j_1,j_2,m} = \lceil \sum_{l=j_1}^{j_2} L_l/R_m \rceil$ be the number of slots required to deliver frames $P_l, l = j_1, \cdots, j_2$ using MCS m. To compute the optimal utility $U_P(j, m, t)$, only the last few frames ending at P_j may choose MCS m. Assuming that frames $P_l, l = i + 1, \cdots, j$, are transmitted

with MCS m (note that if $i = j$, no frame is transmitted with MCS m), the utility $U_P(j, m, t)$ is then the summation of the optimal utility of the first i frames using MCS up to $m-1$ with $t-\tau_{i+1,j,m}$ slots and the utility obtained by transmitting frames $i+1$ to j using MCS m. Maximizing over all possible i, we obtain the recursive equation for $U_P(j, m, t)$ as follows.

$$U_P(j, m, t) = \max_{0 \leqslant i \leqslant j} \left[U_P(i, m-1, t-\tau_{i+1,j,m}) + \sum_{l=i+1}^{j} \sum_{k \in S_m} u_l^k \right] \tag{2.2}$$

$$q(j, m, t) = \text{argmax}_i \left[U_P(i, m-1, t-\tau_{i+1,j,m}) + \sum_{l=i+1}^{j} \sum_{k \in S_m} u_l^k \right] \tag{2.3}$$

where S_m is the set of users who can decode MCS m. In Equation (2.3), $q(j, m, t)$ keeps track of the best parameter i that maximizes Equation (2.2) and it is used for finding the optimal resource allocation later.

The initial conditions for $U_P(j, m, t)$ are

$$\begin{aligned} U_P(j, m, t) &= -\infty, \text{ if } t < 0 \\ U_P(j, 0, t) &= -\infty, \text{ if } j \geqslant 0, t \geqslant 0 \\ U_P(-1, m, t) &= 0, \text{ if } m \geqslant 0, t \geqslant 0 \end{aligned} \tag{2.4}$$

The first two equations state that $t < 0$, or $m = 0$ and $j \geqslant 0$ is not a valid choice for the utility function $U_P(j, m, t)$. We note that the valid MCS choices are 1 to M and $m = 0$ is a dummy MCS used for initialization. In the last equation above, $j = -1$ with MCS $m \geqslant 0$ indicates that all frames ($j \geqslant 0$) have been considered and the dummy frame $j = -1$ has zero utility.

From Equation (2.2), we can see that $U_P(j, m+1, t) \geqslant U_P(j, m, t)$. Therefore, the optimal utility is always achieved at $m = M$. For each t, we can then compute the optimal utility U_P^* for each available time slot t:

$$\begin{aligned} U_P^*(t) &= \max_{j \geqslant 0} U_P(j, M, t) \\ j^*(t) &= \text{argmax}_{j \geqslant 0} U_P(j, M, t) \\ m^*(t) &= \min\{m : U_P(j^*(t), m, t) == U_P^*(t)\} \end{aligned} \tag{2.5}$$

where $j^*(t)$ achieves the optimal utility, indicating that P frames $j > j^*(t)$ are dropped. The dynamic-programming procedure for I/P frame scheduling is illustrated in Algorithm 1. We now discuss the complexity of Algorithm 1; for the worst case of the recursion procedure (steps 2 to 4), we would compute the value of $U_P(j, m, t)$ for all $0 \leqslant j \leqslant G$, $1 \leqslant m \leqslant M$, $0 \leqslant t \leqslant T$ and every step of the recursion has a maximum complexity of $O(G)$ as shown in Equation (2.2). Thus, the worst case complexity is $O(G^2MT)$ and it is pseudo-polynomial given that the running time of obtaining Equation (2.2) is bounded by a polynomial on the size of input parameters, j, m, t, and the number of slot resource $t \leqslant T$ may be exponential. In practice, T is limited by the frame size, therefore the complexity of the algorithm is acceptable in most practical systems. Indeed, it causes very minimal overhead to the system (details in subsection 2.5).

Algorithm 1 I/P frame scheduling

1: Use Equation (2.4) to compute the utility $U_P(j, m, t)$ at the boundary.
2: **for** all j, m, t **do**
3: Compute $U_P(j, m, t)$ iteratively using Equation (2.2).
4: **end for**
5: Find the optimal utility $U_P^*(t)$ for all $0 \leqslant t \leqslant T$ using Equation (2.5).

B Frame Scheduling

Given that t out of total T slots are allocated for transmitting I and P frames, $T - t$ slots are left for B frames. Since some P frames may be dropped after the P frame scheduling, transmitting the B frames that depend on those dropped P frames does not produce any benefit. Hence, we only consider the set of B frames that are still useful. Let

$\mathcal{B}(t) = \{b :$ B-Frame b does not depend on P frame $j > j^*(t)\}$. As B frames are less important than P frames, we use higher (or equal) MCS for B frames than the last transmitted P frame $j^*(t)$ for efficient use of given resources (lower MCS requires more number of slots for transmission). Assume that B frames in $\mathcal{B}(t)$ are naturally ordered by their decoding time (or display time).

The problem of B frame scheduling can be formulated as, selecting a subset of B frames and an appropriate MCS $m \geqslant m^*(t)$ for each B frame, to maximize

the total utility, such that the total number of slots does not exceed $T - t$. The problem can easily be seen as a multiple-choice knapsack problem and is NP-hard. Nevertheless, most B frames have relatively small sizes and small utility compared to I and P frames.

We use the following algorithm to obtain a sub-optimal solution by imposing the requirement that all B frames use the same MCS. To be specific, for each MCS $m \geqslant m^*(t)$, we find maximum number b_m of B frames that can be transmitted using MCS m under the slot constraint. Finally, we pick the MCS that maximizes the total utility. B frame scheduling algorithm is described in Algorithm 2.

Algorithm 2 B frame scheduling

1: Sort all B frames in $\mathcal{B}(t)$ in the decoding-time order
2: **for** all $m^*(t) \leqslant m \leqslant M$ **do**
3: Find the maximum b_m such that the first b_m B frames in $\mathcal{B}(t)$ can be transmitted with MCS m in $T - t$ slots.
4: The resulting utility is $U_B(m, T - t) = \sum_{b=1}^{b_m} \sum_{k \in S_m} u_b^k$.
5: **end for**
6: Find the optimal $m_0 = \text{argmax}_m U_B(m, T - t)$ and obtain the utility $U_B^*(T - t) = U_B(m_0, T - t)$.

Joint I/P/B Frame Scheduling

To perform the joint scheduling, we find the optimal resource allocation between I/P frames and B frames. Let

$$\begin{aligned} U^* &= \max_t \{U_P^*(t) + U_B^*(T - t)\} \\ t^* &= \text{argmax}_t \{U_P^*(t) + U_B^*(T - t)\} \end{aligned} \tag{2.6}$$

be the optimal total utility and the optimal number of slots allocated to I/P frames, respectively. $T - t^*$ is the optimal number of slots allocated to B frames. Equation (2.6) can be optimally solved by enumerating all possible values $0 \leqslant t \leqslant T$ (given that we consider a discrete time system, which is employed in most mobile networks).

MCS Assignment

Finally, MuVi determines the number of I, P and B frames transmitted and the MCS for each transmitted video frame based on the result of utility optimization obtained in the previous subsection. The following steps are executed in order. Note that I frame is viewed as the special P frame with index 0.

1. t^* and $T - t^*$ (obtained from Equation (2.6)) are the number of slots allocated to P frames and B frames, respectively.
2. The first $j^*(t^*)$ P frames (from Equation (2.5)) are transmitted and the rest P frames are discarded.
3. $t = t^*, j = j^*(t^*), m = m^*(t^*), i = q(j, m, t)$.
4. P frames $P_l, l = i + 1, \cdots j$ are transmitted with MCS m (if $i == j$, no frames are transmitted with MCS m).
5. If $i < 0$, go to Step 6. Otherwise, $t = t - \tau_{i+1,j,m}$, $j = i, m = m - 1, i = q(j, m, t)$, go to step 4.
6. $m_0 = \text{argmax}_m U_B(m, T - t^*)$. The first b_{m_0} B frames are transmitted with MCS m_0 and the rest B frames are dropped.

Once the MCS for each video frame is determined, the packets belonging to each frame are marked with the assigned MCS index in the DiffServ field of the IP header.

Transmitting Video Packets through WiMAX Frames

Since MuVi off-loads the sophisticated scheduling algorithm from the MAC layer of the WiMAX base-station to the ASN gateway, the role of the base-station is simply packing the forwarded video packets to the WiMAX frames. Once the video packet is passed onto the MAC layer at the base-station after applying our algorithms, the DiffServ field of the IP packet is checked to identify the MCS assignment for each packet. Then it packs the received video packets

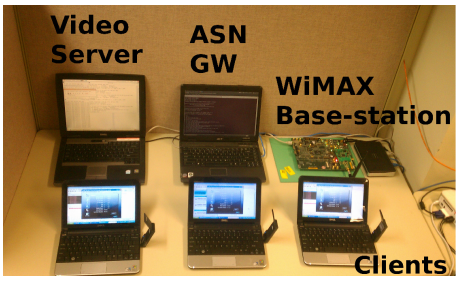


Figure 2.5: WiMAX testbed consists of femto base-station, ASN gateway, a video server and several clients.

into the WiMAX DL subframe (i.e., constructing DL burst), considering the amount of available slots in a WiMAX frame and the MCS. The burst packing component implements a rectangular alignment of the video packets into the WiMAX frame. If the video packet is too large to send in a single WiMAX frame, then the packet will be fragmented and this will be indicated in the DL-MAP so that the client will reconstruct the segmented video packets once they are received. On the other hand, multiple video packets can be packed in a single WiMAX frame when the resources are sufficient. After packing, the base-station hands it to the PHY layer to transmit the WiMAX frame to the clients with appropriate PHY rate.

2.4 WIMAX TESTBED AND PROTOTYPE IMPLEMENTATION

Testbed

Our WiMAX testbed consists of four components: a WiMAX femto base-station, an Access Service Network (ASN) gateway, a video server, and several WiMAX clients as depicted in Figure 2.5. The base-station is a PicoChip [10] WiMAX platform based on IEEE 802.16e standard [8]. The PicoChip base-station is tuned to operate in a 10 MHz bandwidth with the center carrier frequency of 2.59 GHz, for which we have obtained an experimental license to transmit WiMAX signals over the air. Both the ASN gateway and the video server run on

typical Linux machines with a 2 GHz processor and 1 GB memory. The WiMAX clients are Windows laptops with commercial USB dongle WiMAX cards [1] or Beceem PCMCIA [3] interfaces. Since MuVi does not require changes on the client side, MuVi can also support WiMAX mobile phones such as HTC EVO 4G phone as the clients without any modifications.

The clients can be associated with the WiMAX base-station through the ASN gateway. The ASN gateway controls and maintains both uplink and downlink connections between the WiMAX base-station and the clients through configuring service flows which are unidirectional flows of data traffic. All uplink and downlink traffic between the base-station and the video server are tunneled through the ASN gateway.

The base-station manages the scheduling for both downlink and uplink traffic. Since multiple users share the OFDMA wireless resources (i.e., the WiMAX frames), the base-station incorporates a scheduler for efficient resource management. The downlink and uplink scheduler assigns a certain number of slots to each flow and informs the clients about the resource allocations through DL/UL MAP.

Implementation

Our prototype implementation includes modifications at the video server, the ASN gateway, and the WiMAX base-station. The main components of MuVi (handling client feedback, optimizing resource allocation, and frame re-ordering) are implemented on the ASN gateway using the Click Modular Router [80] as a user-level module to handle all video frames to the base-station. We significantly extend and modify the click module (about 2000 lines of C++ code) to realize MuVi's building blocks.

The modified version of VLC media player [14] runs on the video server to send media traffic to the group of multicast users. The VLC module first inserts the frame information in each video packet's IP header (i.e., DiffServ field) and then sends them to the ASN gateway. This marking process helps the ASN gateway classify the video packets easily without deep packet inspection.

After classifying the packet types (i.e., video frame types), the ASN gateway puts them in different queues based on their types. Then, ASN gateway schedules video frames to the base-station after applying the packet re-ordering and rate selection algorithms as described in Section 4.4. In broadband cellular networks, gateways are typically sophisticated servers managing hundreds of base-stations. Since all downlink and uplink traffic go through the gateway, assigning an MCS for all packets at the gateway causes very little overhead, therefore, the overhead of MuVi at the gateway is minimal (evaluation in subsection 2.5). As a result of the scheduling algorithm, each packet is marked with the assigned MCS index in the DiffServ field in the IP header and is delivered to the base-station.

Now that the scheduling process is offloaded from the base-station to the ASN gateway, the base-station simply packs incoming video frames into the WiMAX data bursts and transmits them to the clients in a multicast group. The reference design of PicoChip platform does not involve sophisticated scheduling routines and provides just a working link between the base-station and the clients. We modify the base-station in two aspects to incorporate MuVi. First, it reads the DiffServ Field of the IP header to extract the MCS information determined by the scheduling process in the ASN gateway and transmits the packets using the specified MCS. Second, we modify the base-station MAC code to provide the client feedback to the ASN gateway once every 100 milliseconds. Client channel conditions are periodically reported from the clients to the base-station as part of the standard operation, so the base-station just aggregates the feedback and forwards to the ASN gateway. The size of each channel feedback is relatively small (less than 40 Bytes), so this routine causes very little overhead at the base-station.

2.5 EVALUATION

We evaluate the efficacy of MuVi on our WiMAX testbed by comparing it with existing wireless multicast schemes. We first describe our experiment setup, two reference schemes, evaluation metrics, and test video.

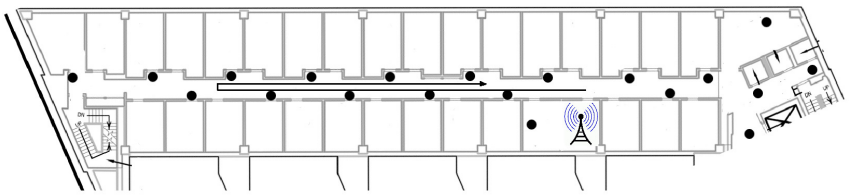


Figure 2.6: Dots represent the client locations for experiments. The line shows the path of a mobile client under the mobility experiment.

Experiment Setup: Our WiMAX testbed (Figure 2.5) is deployed in a typical indoor environment (3rd floor, CS building) as depicted in Figure 2.6. To generate various topologies, we deploy clients in multiple locations, where the channel quality varies in terms of CINR. The clients are exposed to a different level of channel condition. All clients have the same priority and the utility of each frame is set to the number of frames depending on it (including itself). For the confidence results, we repeat each set of experiments more than 5 times and present the averaged results with a 95% confidence interval, except for the microscopic and mobility results, where we present the results of a single run with finer granularity. We maintain the same client topologies while running different video multicast schemes to ensure fair comparisons.

Reference Schemes: We compare the performance of MuVi with that of Adaptive and Naive approaches. The Naive scheme is a traditional WiFi/WiMAX multicast approach that uses the most robust (lowest) MCS for delivering data to the group of clients. This approach does not consider the client channel conditions and only guarantees successful deliveries without considering the real-time requirement. For the Adaptive scheme, we only adapt transmission bit-rates for video frames like DirCast [32]. Based on the client channel conditions, Adaptive scheme uses the highest MCS which can be supported by all clients in the multicast group. The MCS used in Adaptive scheme to deliver the multicast data reflects instantaneous client channel conditions.

Evaluation Metrics:

- *PSNR:* PSNR (Peak Signal-to-Noise Ratio) is a standard metric of video quality and is a function of the mean square error between the original and the received video frames. If a video frame is dropped or past the deadline, it is considered lost and is concealed by copying from the last received frame before it.

- *Inter-packet delay:* We measure inter-packet delay of received packets to quantify the jitter of delivered video stream. To successfully display a streaming video on the client, all video packets belonging to the same video frame need to be received within a given deadline. High jitter values between packets cause bad visual quality.

- *Ratio of packets past the deadline:* PSNR measures video quality based on the received video frames regardless of the deadline of video frames. However, the video frames past the deadline can not be used for playing in real-time. We measure the ratio of packets which miss the deadline to reflect the real-time streaming video quality.

- *MCS:* MCS selection for each video frame is important to satisfy the guaranteed delivery to all clients that are exposed under different channel conditions. MCS selection is even more important to prevent the under-utilization of given resources (i.e., number of slots in WiMAX frame). We measure the average MCS used for delivering video frames which reflects the efficacy of WiMAX frame resources usage.

Test Video: We use MPEG4-encoded [9] video streaming for evaluations. In the experiments, the video is encoded at 2.5 Mbps and is multicasted to the clients. Although we consider non-scalable video sequences in the experiments, our scheme can also be applied to scalable videos by assigning packets with different utility based on the layers they belong to.

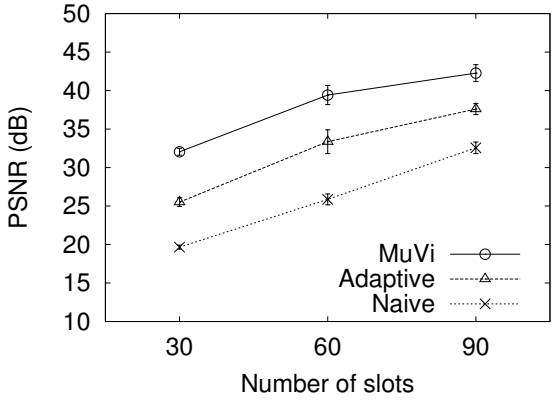


Figure 2.7: MuVi provides best video quality regardless of available resources.

Performance under Different Resource Constraints

We evaluate the performance of MuVi and compare it to Adaptive and Naive schemes while varying the number of available resources (i.e., slots) in a WiMAX frame. We use 2.5 Mbps video stream for all experiments. The data throughput depends on the number of slots allocated in each WiMAX frame and the MCS level chosen. For example, allocating 60 slots per frame with the highest MCS (64-QAM with 5/6 coding) can yield 2.6 Mbps, theoretically. In practice, MCS used for multicast may be less than the highest one considering various channel conditions of the users. Hence, 60 slots are insufficient for successfully delivering the video encoded with 2.5 Mbps. We generate different resource constraints by varying the number of slots in a WiMAX frame for video traffic. In the experiments, 30 and 60 slots per frame represent very tight resource constraints and 90 slots per frame represent just enough resources.

PSNR and Throughput

The average PSNR and 95% confidence intervals for each multicast scheme obtained from 10 experiments are presented in Figure 2.7. MuVi significantly outperforms the other two schemes under all resource constraints. Even when the radio resources are severely insufficient (e.g., 30 slots per frame), MuVi still

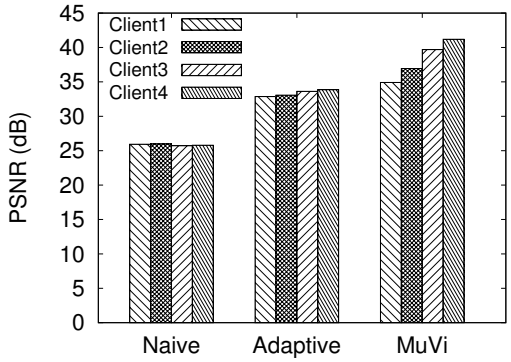


Figure 2.8: MuVi provides differentiated service to the clients depends on their channel conditions.

provides good video quality while the other two schemes suffer. In average, MuVi improves the video quality by up to 13 dB and 7 dB, compared to the Naive scheme and the Adaptive scheme, respectively. Similar subjective results are observed. We notice more frequent glitches and stalls with Adaptive and Naive schemes, while enjoying much smoother streaming with MuVi. The reason that MuVi outperforms the other two schemes is mainly due to the packet scheduling (e.g., resource optimization) and MCS assignment discussed in Section 2.3. Using the most robust MCS can guarantee successful deliveries, but it leads to very low efficacy of resource utilization and eventually lots of frames cannot be delivered before their display or decoding deadlines. Instead, MuVi selectively drops some relatively less important video frames, and judiciously assigns the MCS of packets based on their priorities in order to achieve the maximum overall system utility.

We deploy the clients in different locations to create distinct channel conditions. Specifically, clients 1, 2, 3, and 4 are placed in the CINR range of 19-21 dB, 22-24 dB, 25-27 dB, and 28-32 dB, respectively. Figure 2.8 shows the PSNR values of each client. The Naive scheme uses the most robust MCS (i.e., MCS index 0) for delivering all video frames, hence, the PSNR values for clients are almost equal. Similarly, the Adaptive scheme uses a single MCS limited by the

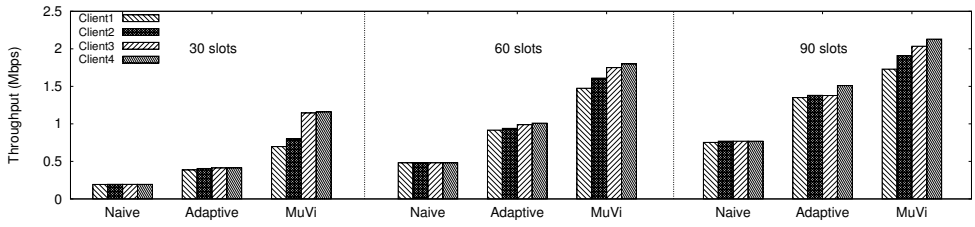


Figure 2.9: Measured throughput for all clients under different resource constraints. MuVi provides at most 4.9× and 2.35× throughput improvement comparing to Naive and Adaptive schemes, respectively.

client who has the worst channel conditions. On the other hand, MuVi uses different MCS for different video frames, and hence the client's PSNR values highly depend on their channel conditions. The client 4, which has the best channel condition, experiences the highest video quality amongst all clients.

During the same experiments, we measure the throughput for each client and present them in Figure 2.9 with respect to the different resource constraints. The throughput pattern is very similar to that of PSNR (Figure 2.8), since throughput is highly related to the video packet receptions. The more packets a client receives, the higher throughput and PSNR value are observed. The throughput results for both Naive and Adaptive schemes are linearly correlated to the resource constraints. By contrast, MuVi provides differentiated service to the clients and guaranties relatively high throughput even under very limited resource conditions (i.e., 30 slots). Specifically, MuVi achieves 2.35–4.9 times higher throughput than the other schemes and utilizes these resources for enhancing video quality, therefore it provides 32.5–43 dB video quality in terms of PSNR. Given that the higher throughput achieved in MuVi, it can support 2–4 times the number of multicast groups for a given amount of resources (30–90 slots per WiMAX frame) if it assigns the slot resources for supporting more number of multicast groups which have different channel conditions. Even in this case, MuVi still provide a reasonably good PSNR range, 26–38 dB with respect to the amount of resources (30–90 slots).

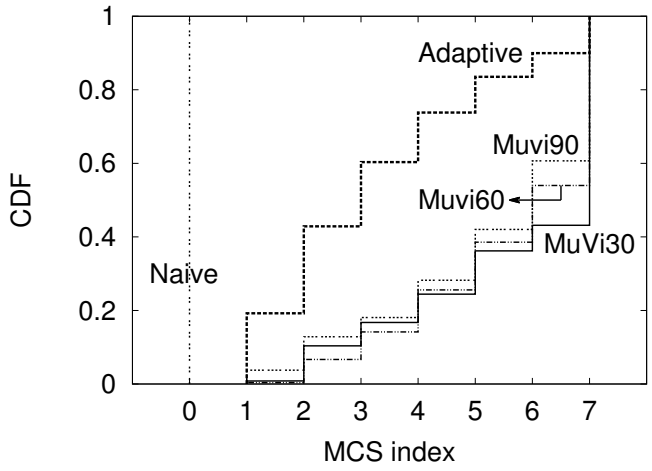
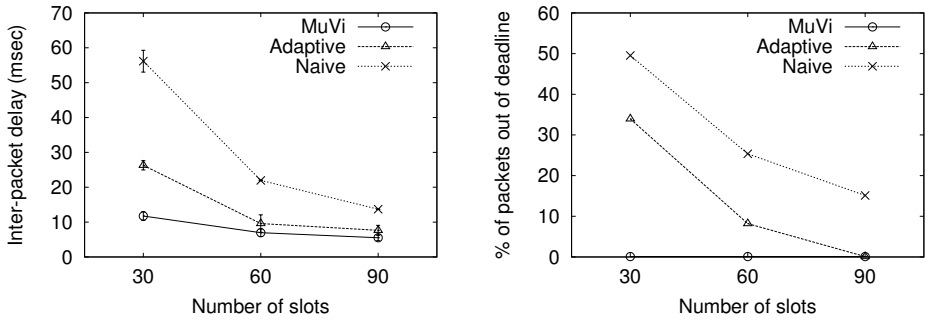


Figure 2.10: MuVi uses higher MCS compared to the Adaptive and Naive schemes.

MCS Selection

We have seen that the PSNR gains for MuVi mainly come from differentiated MCS selection based on the frame priority. To understand the MCS usage, we present the CDF of selected MCSs for all video frames obtained from multiple runs of various topologies in Figure 2.10, where the curve MuVi30 (60, 90) represents the case using MuVi with 30 (60, 90) slots per WiMAX frame. MuVi uses higher MCSs than the other schemes. MuVi transmits 50% of frames using MCS 6 or 7 while the Adaptive scheme transmits 50% of frames using MCS 3 or higher. Recall that, both MuVi and Adaptive schemes tune the MCS for frames instantaneously based on the clients' feedback of their channel conditions. The Adaptive scheme keeps the same MCS for all video frames, however MuVi assigns different MCS to each frame reflecting its system utility. Some clients cannot receive all video frames due to their bad channel conditions and the higher MCSs employed with some frames, but it helps to improve the video quality for other clients with good channel conditions. This is especially beneficial when the resource constraint is very tight (30 slots per frame). This explains why MuVi selects higher MCSs than the Adaptive and the Naive



(a) Average inter-packet delay (b) Ratio of packets missing a deadline

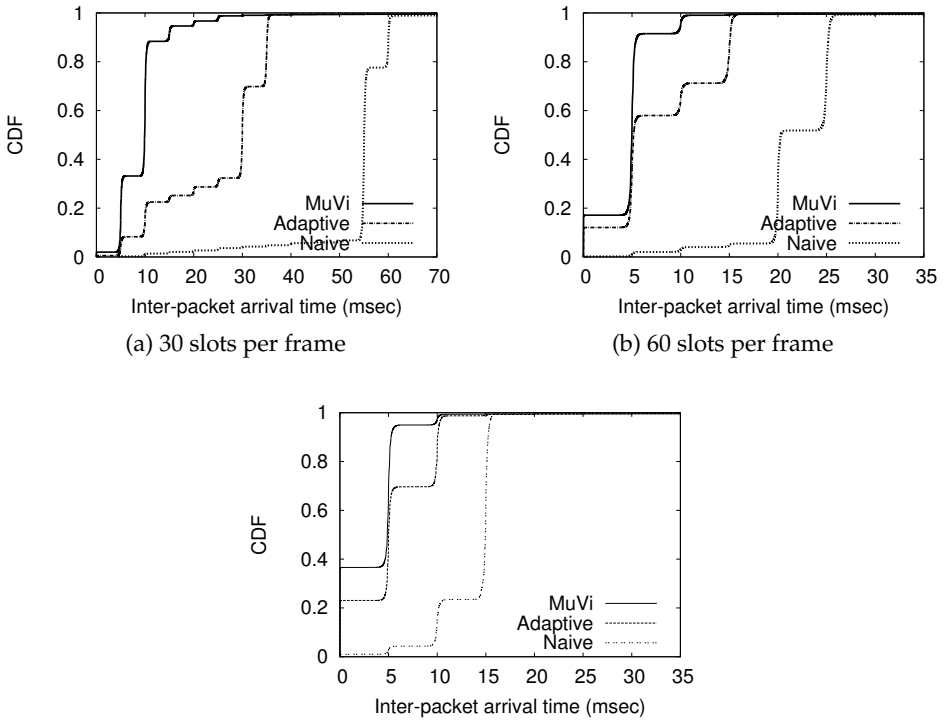
Figure 2.11: MuVi outperforms the Adaptive and the Naive schemes in terms of inter-packet delay and percentage of packets missing the deadline under various resource constraints.

schemes and provides differentiated service to the clients.

The other interesting fact we notice is that MuVi aggressively uses higher MCS when there are insufficient resources (e.g., MuVi30 and MuVi60 curves in Figure 2.10). The reason behind this is that when the resources are severely insufficient, MuVi tends to sacrifice some users with weak channel conditions by selecting higher MCSs and reducing the transmission time of selected video packets.

Inter-packet Delay

The average inter-packet delay for the three schemes is presented in Figure 2.11a. MuVi keeps the inter-packet delay low compared to the other schemes regardless of the number of available slots. The inter-packet delay depends on the MCSs used for transmitting packets; lower MCS leads to longer transmission time, which in turn results in larger inter-packet delay. The gap between MuVi and other schemes is larger under tight resource constraints (30 slots), and it becomes smaller as the number of slots increases. Increasing radio resources significantly increases the number of packets packed in WiMAX frame and reduces the delay for the Naive and the Adaptive schemes. However, the improvement (reduced inter-packet delay) due to the increased resources is



(a) 30 slots per frame

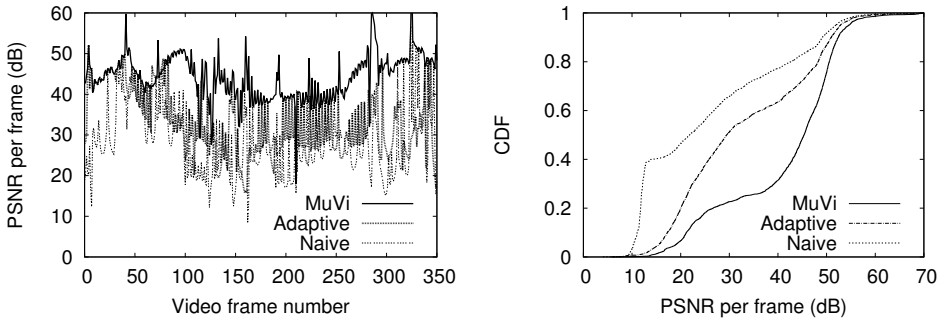
(b) 60 slots per frame

(c) 90 slots per frame

Figure 2.12: CDF of inter-packet arrival time under different resource constraints. MuVi keeps inter-packet time close to 5 milliseconds (same as WiMAX frame interval) regardless of the available resources.

marginal for MuVi because, MCS selection for video frame and packet scheduling are already optimized.

Figure 2.12 shows the CDF of inter-packet arrival time from the same set of experiments as shown in Figure 2.11a. The results of MuVi in Figure 2.12b and 2.12c show that more than 94% of packets have inter-packet arrival time less than or equal to 5 msec, which is the WiMAX frame duration. Even when the radio resources are scarce, MuVi delivers 90% of the frames within 10 msec (Figure 2.12a). On the other hand, the inter-packet delay for the Adaptive and the Naive schemes highly depends on the number of slots available (delay decreases as the number of slots increase). Lower inter-packet delay naturally



(a) PSNR value of each video frame (b) CDF of per frame PSNR

Figure 2.13: (a) MuVi's per frame PSNR is higher than the other two schemes. (b) MuVi's per frame PSNR is higher than the other two schemes.

leads to the higher chance to deliver packets before playback deadline. Figure 2.11b shows the percentage of received packets that miss deadlines. MuVi guarantees delivery of video frames within deadlines regardless of resource constraints.

Micro-benchmark

For further understanding, we present the microscopic results. All results are obtained from the experiments with 60 slots for multicast traffic (we observe similar behavior for 30 and 90 slots and omit them for the sake of brevity).

Per frame PSNR

The video frame rate is 30 frames per second and each frame consists of multiple video packets. Per frame PSNR value is determined by the reception of video packets. In Figure 2.13a, we plot the per frame PSNR values (first 350 video frames) from a single run. We pick a client in the multicast group whose CINR values are in the range of 25-27 dB and present the per frame PSNR value for three schemes. We observe similar patterns from all other clients which have different channel conditions. The frame level PSNR of MuVi is consistently higher than that of the Adaptive and the Naive schemes. Moreover, MuVi

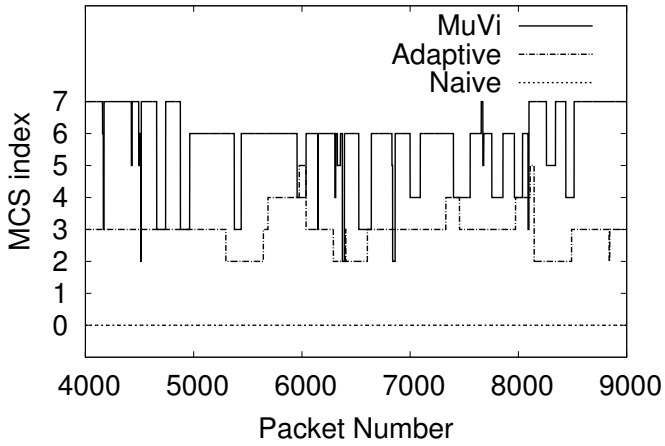


Figure 2.14: MCS for each packet. MuVi uses higher MCS for transmitting frames, however it provides better video quality.

provides very stable and high PSNR (at least 40dB) across video frames, while the other two schemes obtain much lower PSNR and very high variation.

In Figure 2.13b, we plot the CDF of per frame PSNR from same experiment. We can see that the per frame PSNR of MuVi is significantly higher than that of the other two schemes. The median of per frame PSNR for MuVi, Adaptive and Naive schemes is 45, 29 and 21 dB, respectively. Moreover, the PSNR for 75% of video frames in MuVi is greater than 33 dB.

Per packet MCS

In Figure 2.14, we present the MCS used for each video packet from the same experiment as above. MuVi uses higher MCS than the Adaptive and the Naive schemes. We also notice that the variation of chosen MCS for MuVi is higher than that of the other two schemes because MuVi adaptively changes the MCS for each video frame based on the system utility and the client's feedback. Typically, higher MCSs (64-QAM with 2/3 coding or higher) are used for transmitting B frames and lower MCSs are used for transmitting I and P frames. The Adaptive scheme uses MCS 3 for most of the packets while adapting it for some

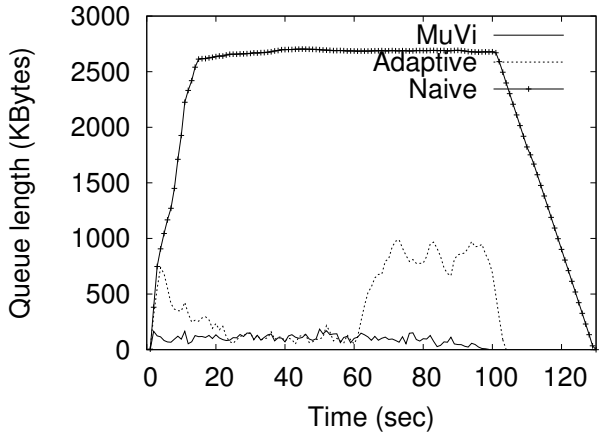


Figure 2.15: Queue length at the WiMAX BS. MuVi keeps the queue size significantly smaller than the other two schemes.

packets based on the clients' channel conditions. The Naive scheme uses the robust MCS (i.e., index 0) for all packets to guarantee the successful delivery regardless of wireless channel variation.

MCS used for MuVi is higher than the other schemes, but MuVi provides better video quality as presented in Figure 2.13b. Although the Adaptive and the Naive schemes use lower MCSs for robust transmission, this requires longer time to deliver the video packets. As a result, many packets miss their deadlines and are discarded at the client, which results in lower PSNR values. In other words, Adaptive and Naive schemes waste the given resources while focusing on robust transmission rather than resource optimization.

Queue Length

Figure 2.15 shows the instantaneous queue size (in KBytes) at the base-station for the three schemes. The queue length of MuVi is constantly below 100 KBytes and is significantly smaller than the other two schemes. The queue length highly depends on the MCS (i.e., the transmission rate) used for transmission. When the base-station uses higher transmission data rate, it can quickly send data

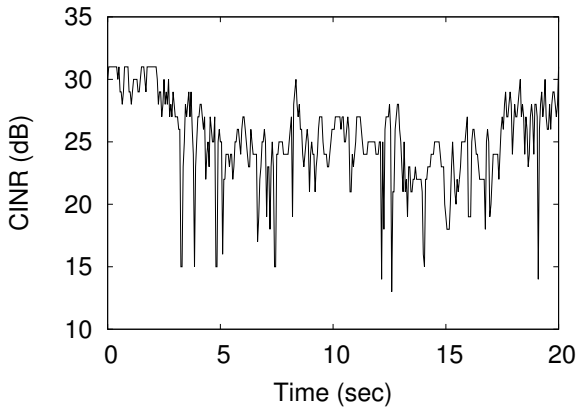
and empty the queue[1]. As we see in Figure 2.14, MCS used for MuVi is higher, and hence it leads to smaller queue length than Adaptive and Naive schemes. The duration of video we used for the experiments is 100 seconds. With given resources (60 slots), MuVi can deliver all incoming video traffic in time, while Adaptive scheme completely empties its queue at around 104 seconds. Even worse, Naive scheme finishes at 125 seconds.

Client Mobility

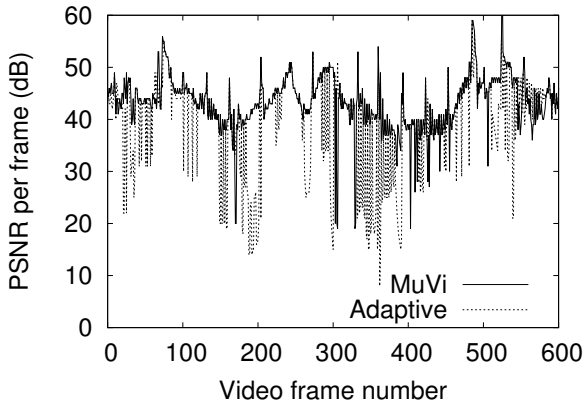
Here, we evaluate the performance of MuVi with client mobility. In this set of experiments, the multicast group contains two types of clients; one stationary and the other moving at walking speed. The mobile client starts from a location close to the base-station, moves away from it and then moves back as the route depicted in Figure 2.6.

Figure 2.16a shows the observed instantaneous CINR of the mobile client. We see that the CINR decreases gradually till 14 seconds as the client moves away from base-station and then increases. The CINR increase at 8 second and multiple downside spikes are probably caused by random channel fading. Figure 2.16b plots the PSNR value of each frame for both MuVi and the Adaptive scheme under the same mobile trajectory. We observe the PSNR values for both MuVi and the Adaptive scheme are highly correlated to the CINR of the mobile client, however, the Adaptive scheme obtains much lower PSNR values and higher variation than MuVi. Although MuVi obtains low PSNR values for some frames, most of them are due to dropped B frames, which do not affect I or P frames for decoding. On the contrary, some P frames may get lost in the Adaptive scheme, which in turn affect several other dependent frames. This indicates that instantly adapting MCS based on the weakest client's feedback is not robust enough to sustain user mobility. Simultaneously optimizing resource allocations and differentiating packet priorities by adapting MCS can provide more robust performance under client mobility. WiMAX base-station receives the CINR report (which is part of the WiMAX standard) from all clients

[1]In this work, we do not consider re-transmissions, which are typically not employed for multicast traffic to avoid feedback explosion.



(a) Instantaneous CINR of the mobile client



(b) Per frame PSNR values of first 600 frames

Figure 2.16: The video frame rate is 30 frames per second, so the total number of video frames generated in 20 seconds is about 600.

periodically (once every 5 msec) through the dedicated uplink channel. MuVi leverages this information to determine the highest MCS that each client can successfully decode.

Overhead

MuVi's core engine lies at the ASN gateway like Opal [66]. Typically, a single ASN gateway manages multiple base-stations in a cellular network. The overhead at the ASN gateway could be problematic considering the number of base-stations associated with it. Hence, any implementations or additional work load on the ASN gateway will require appropriate provisioning.

To investigate the computational overhead of MuVi's operations (optimizing resource allocation, determining MCS for each packets and processing packet re-ordering), we measure the execution time of MuVi while delivering video frames to the multicast group. The total execution time for MuVi's algorithm is 83.63 milliseconds while the total experimental duration (the same as the video length) is 100 seconds. It amounts to less than 0.1% of the total CPU time. Therefore, the overhead of MuVi is almost negligible for handling a single base-station and would be small even if several hundred base-stations are managed by a single gateway.

The process of handling client's CINR report for getting supportable MCS for each client will be done by referring the MCS-CINR table we summarized in Table 2.2. The base-station has the MCS-CINR table and hence whenever base-station receives the CINR report from the clients it can easily read the supportable MCS value accordingly. The complexity of this process is $O(1)$. The uplink channels are used to send the client's CINR report to the base-station, it would not affect the downlink resources for video traffic to the multicast clients.

2.6 SUMMARY OF MUVI

In this chapter, we first motivated the necessity of optimizing frame resources when multiple users are scheduled in the same WiMAX frame. Considering the importance of frame utilization and client scheduling in OFDMA system, we

designed an efficient multicast video delivery scheme (MuVi) in 4G broadband wireless network. Specifically, we developed joint optimal resource allocation and adaptive modulation and coding scheme and implemented it using a WiMAX testbed to show its efficacy in real systems. MuVi differentiates video frames based on their importance in reconstructing the video and incorporates an efficient radio resource allocation algorithm to optimize the overall video quality across all users in the multicast group. MuVi is a lightweight solution with most of the implementation in the gateway, slight modification in the base-station, and no modification at the clients. We implement MuVi on a WiMAX testbed and compare its performance to a Naive wireless multicast scheme that employs the most robust MCS, and an Adaptive scheme that employs the highest MCS supportable by all clients. Experimental results show that MuVi improves the average video PSNR by up to 13 dB and 7 dB compared to the Naive and the Adaptive schemes, respectively. MuVi does not require modification to the video encoding scheme or the air interface, thus it allows speedy deployment in existing systems.

3 DISTRIBUTED RESOURCE MANAGEMENT FRAMEWORK IN OFDMA MULTICELL NETWORKS

3.1 MOTIVATION

With the rapid growth of mobile devices and increased data usage, the demand for higher data rate in wireless network continues to grow. Next generation wireless networks (i.e., WiMAX, LTE, LTE-A) provide higher bandwidth and spectrum efficiency leveraging smaller (femto, pico) cells with orthogonal frequency division multiple access (OFDMA). The uncoordinated, dense deployments of femtocells however, pose several unique challenges relating to interference and resource management in OFDMA femtocell networks compared to macrocells (traditional cell towers) and WiFi: **(i)** OFDMA schedules multiple clients in the same frame. The clients experience different levels of interference and hence, a resource management framework has to account for the characteristics of each client. Specifically, clients with strong interference need to operate on orthogonal resources (i.e., frequency isolation), while clients with weak interference can operate on all frequency resources (i.e., reuse). For the efficient use of resources in OFDMA frames, we need to first differentiate the clients. **(ii)** Given that multiple clients share frame resources in OFDMA networks, the next challenge is "how to accommodate multiple clients of various classes in the same frame?". The frame structure has to be carefully managed for various clients considering their interference levels and demands. In particular, each cell needs to determine how much resources can be reused without causing interference to the neighboring cells and how much frequency resources need to be isolated to mitigate the interference from other cells. The frame structure can impact the network wide resource reuse as well, where multiple contention domains are involved. **(iii)** The resources allocated to one femtocell directly impact the resources for the interfering cells. We need to determine the time and frequency resources of operation for the clients in each cell while accounting for the resources used by neighboring cells. Resource allocation for each femtocell should be adaptive to the network changes, but without explicit coordination

due to lack of any central component in the system.

Towards addressing these challenges, we propose RADION, a *distributed* resource management framework that effectively manages interference across femtocells. RADION's core building blocks enable femtocells to opportunistically find the available resources in a completely distributed and efficient manner. In RADION, we address the above challenges through the specific design of the following four building blocks: *i) client throughput estimation, ii) client categorization, iii) resource decoupling and iv) two-phase adaptation and allocation*.

3.2 RESOURCE MANAGEMENT CHALLENGES

The problem of resource management is for each femtocell to distributively determine the frame resources (slots) that it can use to schedule its clients. Since the resource allocation decisions of one cell impact multiple other cells, efficient adaptive and iterative mechanisms are needed to quickly converge to a network-wide resource allocation. We first discuss the challenges in achieving this objective and briefly describe how RADION addresses each of them.

Client Categorization

OFDMA schedules multiple clients in the same frame. Different clients may experience different levels of interference from neighboring femtocells; clients subject to strong interference need to operate on orthogonal sub-channels partitioned across cells (i.e., frequency isolation), while clients with weak interference can use the entire spectrum (i.e., reuse) and still tolerate interference via link (rate) adaptation. Unlike in WiFi, interference avoidance comes at the cost of a reduced set of slots per femtocell, which in turn leads to under-utilization if not properly exercised. Specifically, not all clients in a femtocell require resource isolation; link adaptation may suffice to cope with interference for clients that are in close proximity to the BS. Differentiating the clients is key in realizing good spectral efficiencies. A client *categorization* is included within RADION to accurately differentiate between such clients (which can reuse the spectrum *reuse clients*) and those that need spectral isolation (*isolation clients*). Since sens-

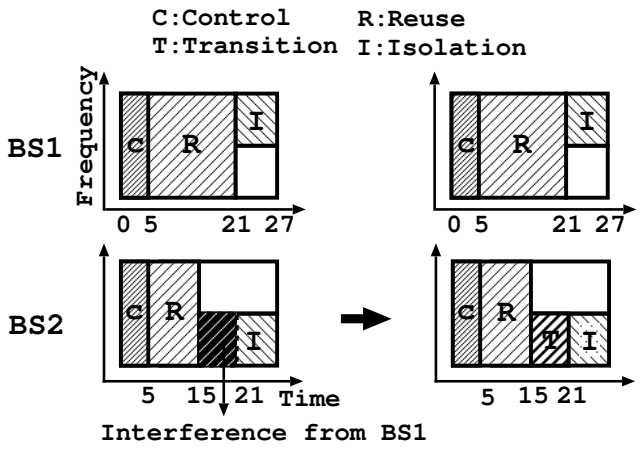


Figure 3.1: Usage of transition zone. BS2 sets transition zone to mitigate interference from BS1's reuse zone.

ing the medium is not possible with standard OFDMA femtocells, RADION uses an intelligent active probing technique to achieve client categorization with high accuracy. Note that the accuracy of categorization has a direct impact on how efficiently the slots are utilized.

Resource Decoupling Among the Clients

Once the clients are categorized, next question is "how to accommodate multiple clients of different categories (*isolation* and *reuse* clients) in the same frame?". Handling heterogeneous clients in the same frame directly impacts on the network resource usage hence, it has to be carefully addressed, otherwise it in turn leads to under-utilization if not properly addressed. Each frame can be segmented into two zones, the reuse and the isolation zone, to accommodate the two types of clients (BS1 in Figure 3.1); clients in the reuse zone will receive data encoded on the entire spectrum, while clients in the isolation zone will receive on a subset of frequencies (determined by an allocation algorithm). The static FFR approaches [35], using pre-determined sizes of reuse and isolation zone across cells, will however not work in femtocell deployments where interference

is pervasive (i.e., not localized). Further, the load of a femtocell as well as interference from other BSs also need to be taken into account in adapting the zone sizes of a particular cell. By splitting frames into multiple zones (reuse and isolation zones), femtocell can schedule heterogenous clients based on their requirements (reuse all frequencies or resource isolation) in the same frame. RADION uses a novel *three-zone* frame structure to address these issues as will be described in section 3.3; in a nutshell the additional third zone is a *transition zone* that prevents resource coupling across cells (discussed next).

Resource Decoupling Across the Femtocells

When two interfering femtocells have different reuse zone sizes (based on their loads), the larger reuse zone will interfere with the isolation zone of the other cell (BS2 in Figure 3.1). Having a common reuse zone for the two cells is essential to avoid this. However, irrespective of whether the maximum or the minimum of the reuse zone sizes is chosen as this “common" zone size, it is easy to see that there is under-utilization in one of the cells (either part of the reuse zone or part of the isolation zone is not utilized). Specifically, if the maximum (minimum) reuse zone is chosen as the common reuse zone, the cell with the smaller (larger) reuse zone will not be able to use any (all) of the resources in the region between its desired and common reuse zones for its isolation (reuse) clients as it will receive (cause) interference from (to) the other cell. While this coupling and resulting resource under-utilization is inevitable to alleviate interference, the challenge is to avoid this coupling from propagating to the entire network, which would result in significant under-utilization that is unwarranted. RADION’s *transition zone* intelligently localizes this resource coupling and prevents such propagation.

Resource Allocation

Each femtocell has to determine its zone sizes and resource usage in a completely distributed manner. RADION uses an iterative, joint time (zone sizes) and frequency (sub-channels in the isolation zone) resource allocation algorithm

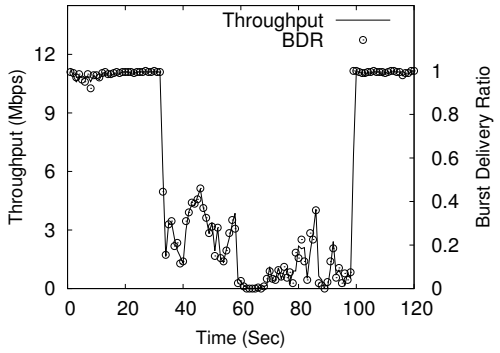


Figure 3.2: Burst Delivery Ratio (BDR) provides an accurate representation of throughput.

that converges to efficient allocations and adapts to network dynamics quickly and efficiently.

3.3 RADION AND ITS BUILDING BLOCKS

We next describe RADION in detail by elaborating on the functionalities of its building blocks. We explain how building blocks work and present experimental results of the components to validate their efficacy.

Client Throughput Estimation

The interference from the neighboring BS can be measured only from the client by accounting the throughput. While a femto BS does not have direct access to the throughput at a client, it receives feedback about the reception of each data burst via ACKs on the uplink frame. We define Burst Delivery Ratio (BDR) to be the ratio of successfully delivered bursts to the total number of transmitted bursts. The BS can estimate BDR by taking the ratio of the number of ACKs received to the total number of feedbacks received for a given client. We perform experiments to understand if the BDR estimate at the BS can provide an understanding of the throughput at the client. While transmitting data burst

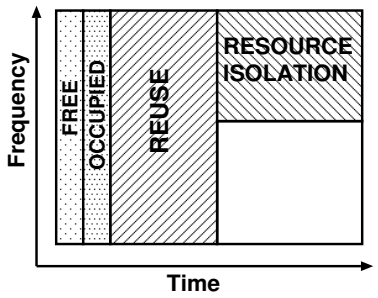


Figure 3.3: Free and Occupied zones are used for client categorization.

to a client, we move the location of the client to sample varied Signal to Noise Ratio (SNR). Figure 3.2 shows that indeed the BS can very accurately track the client throughput using the BDR estimates. We use BDR as a good estimation of a client throughput to measure the interference level in RADION. The clients use existing (dedicated) uplink ACK channels on robust MCS for transmitting feedback to BS, thereby not imposing additional overhead.

With the help of BDR estimates, the femto BS can now compensate for the lack of sensing capability by clients through active probing, whereby the BS schedules its clients on specific resource regions in the frame and estimates their corresponding BDR to detect resource availability. Such probing can be done by transmitting small amounts of data to specific clients on specific frame resources for a very short time interval (e.g., order of 100 msec) while accounting ACK burst receiving every 5 milliseconds through uplink channel. In this manner, probing (using BDR) can indeed detect the resource usage quickly.

Client Categorization

Initial Step

The second component in RADION categorizes clients into two classes; the first needs just link adaptation (class 1) and the second needs resource isolation (class 2). To determine the client's class, we need to estimate the throughput that a client receives with and without interference. One can expect that if the

throughput with interference is comparable to that without interference, the client is class 1; else it is class 2. This is achieved through active probing and BDR estimation. To facilitate client categorization, the frame structure contains two measurement zones of equal size (only 2 symbols); *free* and *occupied* zones as depicted in Figure 3.3. All BSs are required to transmit in the *occupied* zone constantly, therefore scheduling a client in the *occupied* zone enables BS to calculate the BDR in the presence of interference from other BSs. The *free* zone will be used for obtaining the interference-free BDR estimation with probabilistic scheduling. However, it is possible that more than one BS would simultaneously schedule its client in the *free* zone and it would yield an inaccurate BDR estimation corresponding to the case of sampling collision. In order to avoid this, a random access mechanism with probability $\frac{1}{n}$ is used where n is the number of interfering BSs (the interfering BS set can be inferred without cooperation by leveraging signal strength measurements from clients, e.g., each client reports the signal strength from different cells called the automatic neighbor reports in LTE). Each BS performs the following two steps towards categorizing a client leveraging these two zones:

1) Schedule the client's data in the *occupied* zone and schedule it in the *free* zone probabilistically. Keep track of the resulting BDR in both zones over K frames.

2) Determine the normalized throughput per tile in the two zones T_{occ} and T_{free} corresponding to their BDRs. If $T_{free} \geqslant (1+\alpha)T_{occ}$ then the client is categorized as class 2; otherwise class 1.

Through exhaustive measurements (Table 3.1), we set $\alpha = 0.25$ and $K = 25$ frames. This parameter setting yields very high accuracy (>90%) of client categorization and do not depend on the network size or load. Class 1 clients are scheduled in the *reuse* zone using all sub-channels while class 2 clients are scheduled in the *isolation* zone using subset of sub-channels (*reuse* and *isolation* zones are depicted in Figure 3.3).

	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$
K = 25	0	0	0	13%	29%
K = 50	0	0	0	12%	28%
K = 125	0	0	0	12%	27%
K = 250	0	0	0	12%	22%
K = 500	0	0	0	10%	20%

Table 3.1: Categorization error with respect to α and K. We set α=0.25 and K=25 frames that yield almost negligible categorization error.

Refinement Step

After initial categorization, we refine it with a sub-classification of class 2 clients that is suited for distributed operation. In the above step, while clients in class 1 are identified accurately, not all clients categorized as class 2 may need resource isolation. With resource isolation, a BS allocates only a subset of sub-channels to a client. For clients with low-moderate interference, link adaptation to cope with interference may be a better option than sacrificing resources through isolation. Thus, to further refine the categorization of class 2 clients, we factor in the loss of resources due to isolation. This was missing in the step **2**) in the initial categorization, since equal resources were used in the *occupied* and *free* zones. The amount of isolated resources available to a cell depends on the resource allocation algorithm. If the resources assigned to the isolation zone is a fraction f of that of the reuse zone, the BS refines the status of a client in class 2 by scheduling the client on resources in the *isolation* zone and determining its normalized per tile throughput in this zone, T_{isol}. It retains the client in class 2 only if $f \cdot T_{isol} \geqslant (1+\beta)T_{occ}$; otherwise the client is reverted to class 1. Here, β (0.05, experimentally obtained) is used to avoid oscillations in categorization. RADION further sub-classifies clients in class 2 as those that benefit significantly from resource isolation (class 2h) and those that benefit marginally from it (class 2l);

- Class 2h client, when $\frac{f \cdot T_{isol}}{T_{occ}} \geqslant (1+\alpha)$
- Class 2l client, when $(1+\beta) \leqslant \frac{f \cdot T_{isol}}{T_{occ}} < (1+\alpha)$

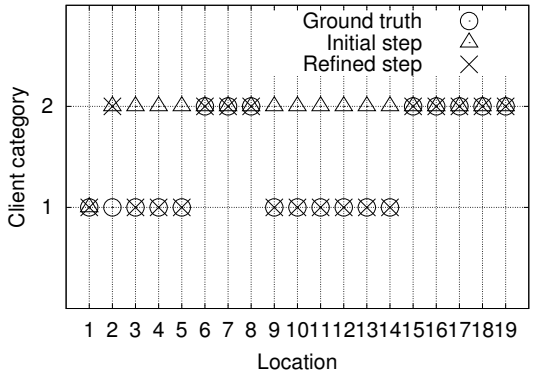


Figure 3.4: Refinement step yields accurate categorization.

The benefits of such a sub-classification will be discussed in section 3.3.

Evaluation of Client Categorization

We consider two cells (1 and 2); clients 1 and 2 belong to the two cells respectively. We generate multiple topologies with different levels of interference by varying the location of client 1 in the presence of interfering cell 2. The isolation zone of each cell has the same number of symbols as the occupied zone but operates on an orthogonal half the sub-channels. First, for every location of client 1, its ground truth is determined by scheduling data to the client in the *occupied* and *isolation* zones and determining it as class 2 if $0.5 \cdot \mathrm{T}_{\mathrm{isol}} \geqslant \mathrm{T}_{\mathrm{occ}}$. Then our three step (two initial steps and refinement step) categorization algorithm is executed. The categorization results of the initial and the refinement steps are shown in Figure 3.4. It is seen that the initial step wrongly categorizes clients in ten locations, however, the refinement step corrects most (nine) of them. The only erroneous classification (location 2) was due to a change in channel conditions during the process. The clients who need to be in class 2 are correctly classified even with the initial step, while refinement only adds more clients from class 2 to class 1. The client categorization, both the initial and refinement steps, yields 94.7% accuracy (comparing ground truth and refined result in Figure 3.4).

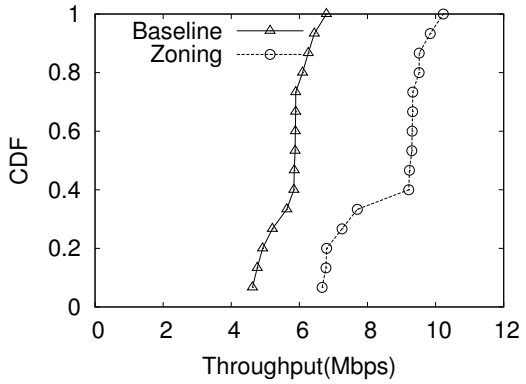


Figure 3.5: Two-zone structure can achieve 35% throughput gain over baseline (no-zone) strategy.

Resource Decoupling

Two-zone Structure

To schedule clients of both classes in the same frame, we use two variable size data of *reuse* and *isolation* zones. Zones are virtual partitions of frame for transmitting data burst to different clients. The reuse zone operates on all sub-channels and schedules class 1 clients, while the isolation zone schedules class 2 clients on only a contiguous subset of *sub-channels*. To understand the benefits of the two-zone structure, we conduct experiments with two BSs. Each BS has two clients (one in each class), and interferes with the class 2 client of the other BS. The baseline scheme operates the two BSs on two orthogonal sets of sub-channels, with each BS scheduling both its clients within its own subset. This is compared against a two-zone scheme where a BS schedules its class 1 client on all sub-channels, while the isolation zone uses the other half of the frame for its class 2 client on half the sub-channels. We generate various interference topologies and the CDF of the net throughput is plotted for the two schemes in Figure 3.5. We see that with the two zones, class 1 clients can be scheduled in tandem to reuse sub-channel resources effectively, yielding over a 35% throughput gain over the conservative isolation scheme.

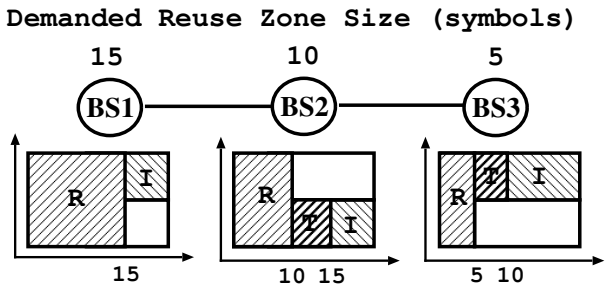


Figure 3.6: Use of *reuse*, *transition* and *isolation* zones in three cells topology.

Drawbacks of Two-zone Structure

A two-zone structure enables resource reuse, but is insufficient in a multi-cell context. Different cells will have different reuse zone sizes based on the load generated by the class 1 clients. If two interfering cells were to operate independently, the larger reuse zone will overlap with the isolation zone of the other cell and hence, interfere with the isolation clients of that cell (Figure 3.1). Consequently, it is important to use a common reuse zone between interfering cells. For the example topology in Figure 3.6, the result in Figure 3.7 indicates that the throughput of BS2 (with the smaller reuse zone) is degraded when its isolation zone starts right after its reuse zone without accounting for the larger reuse zone of BS1 (2Z curve). This is in comparison to even a simple scheme (2Z-CR curve) that starts the isolation zone of BS2 only after the end of the reuse zone of BS1, leaving the region between the reuse zones of the two BSs unused in BS2.

Need for Common Reuse Zone

Each cell has to determine the common reuse zone in a distributed manner. The common reuse zone size can either be the minimum or maximum of the reuse zone sizes of its neighboring interfering cells. RADION uses the maximum of the reuse zone sizes within the interference neighborhood as the common reuse zone for two reasons: (i) since the deployment is un-coordinated and non-cooperative (e.g., residential complexes), there is no incentive for a cell to

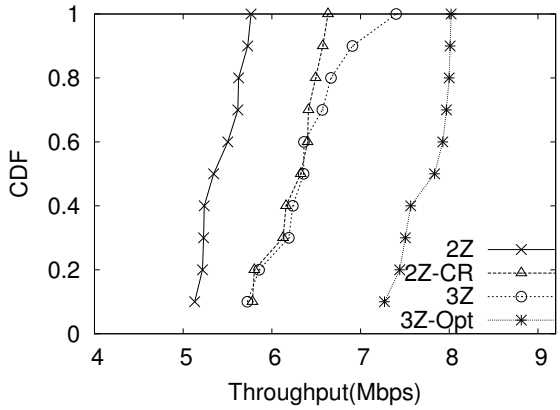


Figure 3.7: Three-zone structure yields a 35% throughput improvement over two-zone structure.

decrease its reuse zone size (only other cells suffer); (ii) given the absence of sensing by the clients, active probing by the BS to determine resource availability can be employed to *only* determine the reuse zone of a cell with a larger zone size effectively (elaborated in section 3.4).

Three-zone Structure

RADION uses a three-zone transmission structure as shown in Figure 3.6. While a BS's *reuse zone* remains the same (for class 1 clients), its *isolation zone* only begins from the end of the common (maximum) reuse zone in its neighborhood. The region between its reuse zone and common reuse zone is the *transition zone*. Blindly scheduling class 2 clients in the transition zone is harmful due to interference from neighboring cells with larger reuse zone sizes. Leaving the transition zone empty (2Z-CR curve in Figure 3.7) will under-utilize resources. RADION intelligently picks class 2 clients (using the sub-classification in section 3.3) to be scheduled in the transition zone on the same subset of sub-channels as the isolation zone. Specifically, class 2$\mathfrak{l}$ clients are scheduled in the transition zone, while class 2$\mathfrak{h}$ clients operate in the isolation zone. There are two benefits to such an approach. First, the transition zone allows selected class 2 clients

to opportunistically reuse resources without incurring significant interference. Second, operating the transition zone on the same subset of sub-channels as the isolation zone prevents the common reuse zone from propagating to the entire network (across multiple contention domains), thereby eliminating under-utilization due to resource coupling. To see this, in Figure 3.6, if the transition zone is allowed to operate on all sub-channels, then the common reuse zone that terminates at 15 symbols in the contention domain between cells 1 and 2, will also propagate to cell 3. This allows cell 3's class 2h clients to be scheduled only after 15 symbols, while the common reuse zone in its own contention domain (between cells 2 and 3) terminates at 10 symbols.

Evaluation of Three-zone Structure

The benefits of RADION's three-zone structure are evident in Figure 3.7. While both 3Z and 3Z-Opt employ a three-zone structure, 3Z-Opt employs intelligent scheduling of class 2 clients in the transition and isolation zones, 3Z randomly schedules class 2 clients in the two zones. It is clear that the three zone structure with sub-classification yields a 35% throughput improvement.

3.4 DISTRIBUTED ALLOCATION FRAMEWORK

We now describe the distributed resource allocation process. The goal of each cell is to determine the size of the reuse (s_r) and transition (s_t) zones as well as the specific contiguous set of sub-channels ($\mathcal{C}$) for operation in the isolation zone. Each BS first classifies its clients into classes 1, 2l and 2h. The preamble used by each BS is chosen from a standardized set (of orthogonal sequences); using this, the clients measure the signal strength to various BSs and hence lock on to a BS with strong signal strength. The same process is used by each client in class 2h to determine its set of strong interferers (received signal strength over a threshold, S_{th}). Using feedback from these clients, the BS determines the super-set of strong interferers. The cardinality of this set including itself (n) determines the fair share allocation of sub-channels ($m = \frac{N}{n}$, where N is total

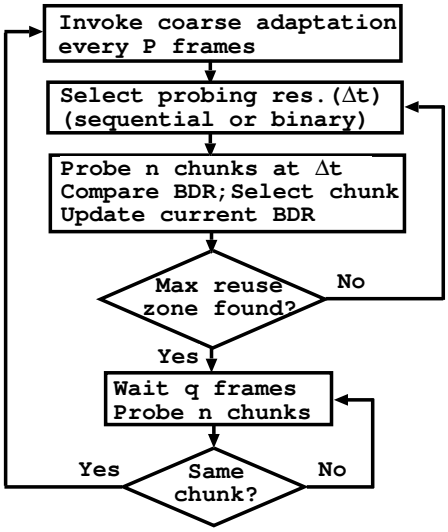


Figure 3.8: Flow chart of RADION's two-phase adaptation.

number of sub-channels) for the clients in the isolation zone (class $2h$)[1]. Then, based on data availability and the set of clients chosen for transmission by the MAC scheduler (using proportional fair scheduling), the BS determines the desired size of its reuse zone (s_r). This is proportional to the relative traffic load from the clients in the two classes. Next, the BS determines the common reuse zone (in time, s_t) in its interference neighborhood as well as the specific set of m sub-channels (in frequency, $\mathcal{C}$) for operation in the isolation zone (details to follow). After determining the resource allocation parameters, clients in classes 1, $2l$ and $2h$ are scheduled in the reuse, transition and isolation zones respectively. Determination of the resource allocation parameters in RADION is accomplished with the following joint time-frequency *probe-adapt* mechanism.

[1]The solution can be extended to weighted allocations, albeit at the cost of information exchange between femtocells.

Two-phase Adaptation

RADION employs a combination of both coarse and fine time scale adaptations (Figure 3.8). Every BS picks a period P of coarse adaptation (order of several seconds; thousands of frames). P is picked from a set of large prime numbers to reduce the frequency of overlap of adaptation (and probing) periods across cells. The goal of coarse adaptation is to track coarse network dynamics such as (de)activation of cells/clients, load changes, etc., that happen at the granularity of several seconds. Every P frames, each BS triggers a series of fine time scale adaptations automatically (can also be based on changes detected in interference conditions as indicated by its clients). Once triggered, the goal of fine adaptation is to quickly converge to the right set of allocation parameters $(s_t, \mathcal{C})$, for a given set of network conditions, while that of the coarse adaptation is to track coarse network dynamics such as traffic load changes, (de)activation of clients, etc. that happen at the granularity of several seconds. During fine adaptation, the BS performs a *probe-and-adapt* procedure every q frames till convergence, where q is randomly selected from $[1, 0.1P]$ and operates at the granularity of hundreds of milliseconds. The randomness of q minimizes probing collisions.

Employing coarse adaptation in isolation will result in long recovery times in the event of probing collisions across femtocells. This leads to large periods of degraded performance. On the other hand, employing fine adaptation in isolation will require continuous probing to track network dynamics, thereby resulting in large overhead. RADION strikes a good balance between coarse and fine adaptations; fine adaptation is suspended after quick convergence to an efficient resource allocation and invoked again only in the next coarse adaptation period. Thus, the BS spends a large fraction of its P frames operating on an efficient allocation with its probing and adaptation mechanism constituting only a small portion of it. Further, even during the probing-adaptation procedure, data scheduling to clients is not interrupted and is seamlessly incorporated into it.

Probe and Adapt

The goal of fine adaptation is to ensure quick convergence in the determination of resource parameters in both time (common reuse zone) and frequency (sub-channels in isolation zone) domains. This is achieved with a joint time-frequency adaptation algorithm. Each BS probes a vertical strip of resources in the frame, of size $\Delta t \times N$ (i.e., encompassing all sub-channels in time Δt), where Δt is the granularity of probing in the time domain (few symbols). The frequency domain is further probed in *chunks* of Δf contiguous sub-channels ($\Delta f = m$). Δt and Δf can be varied to tradeoff fine grained allocation and convergence time. Note that if clients are capable of sensing (e.g., cognitive clients), then multiple sub-channels can be sensed simultaneously to allow fine grained allocation without impacting convergence time. Again, RADION uses K=25 frames to probe resource and it only takes less than 600 msec to probe entire frame resources. The probing mechanism is fast enough to detect the bursty-ness nature of interference.

Algorithm 3 RADION: Distributed Resource Allocation Framework

$m = \frac{N}{n}$, $c_i \in \mathcal{C}$, $s_t \leftarrow s_r$, b_i: BDR, e_i: counter, $\forall e_i$=0
Probe c_c, update b_c, s_t+=Δt, f_c: current frame
while $(f_c \mod P) \equiv 0$
  for $i = 1 : n$
    Probe c_i, update b_i /*probe n frequency chunks*/
  $b_u : u = \max_{i:c_i \in \mathcal{C}}(b_i)$ /*find the max BDR*/
  if $(b_u > b_c \cdot \alpha) \parallel (b_c > \beta)$
    Select c_f : call Algorithm 4 or 5
    $c_c \leftarrow c_f$, update b_c (we found s_t)
    Pick $q \in [1, 0.1P]$, wait q frames /*fine adaptation*/
    for $i = 1 : n$
      Probe c_i, update b_i
    Select c_f : call Algorithm 4 or 5
    if $c_c \equiv c_f$
      $\exists c_i \in \mathcal{C} \setminus \{c_f\}$, s.t $b_i > \beta$, e_i++
      $\exists e_i \geqslant 2$, pick one of the c_i, e_i=0; otherwise c_i=$\emptyset$
      $c_c \leftarrow c_c \cup c_i$, goto step 2 /*frequency converged*/
    else
      $c_c \leftarrow c_f$, goto step 10 /*re-do fine adaptation*/
  else
    s_t+=Δt, $c_c \leftarrow c_u$, update b_c, goto step 4

Algorithm 4 Gibbs Sampler: Frequency Resource Selection

Temperature parameter: $T = 0.05$
$\forall c_i \in \mathcal{C}$ compute the probability:
$\pi(c_i) = (e^{\frac{b_i - 1}{T}})/(\sum_{i=1}^{n} e^{\frac{b_i - 1}{T}})$
Sample a random variable $rand$ with law π
Select c_f according to $rand$

Algorithm 5 Greedy: Frequency Resource Selection

index: $i = \arg\max_{i:c_i \in \mathcal{C}}(b_i)$, $c_f \leftarrow c_i$

Joint Probing in Time and Frequency

When coarse adaptation is triggered, the BS probes resource regions after its own reuse zone to determine the common reuse zone in its neighborhood (step 2 in Algorithm 3). The intuition is that since the interfering cell with the largest reuse zone will use all sub-channels till its reuse zone, when frequency chunks are probed within the largest reuse zone, they will exhibit similar (degraded) BDRs, while when probed beyond the largest reuse zone, there will be at least one frequency chunk, whose BDR exceeds those of the other chunks by α (see inference in section 3.3). This observation is used by every BS to determine the common reuse zone. Specifically, to probe in a vertical resource region when a vertical resource strip in the frame is probed, the BS transmits data to a client in each of n randomly chosen frequency chunks of size m. Since P varies across cells, and the frequency chunk to be probed is chosen at random, probing conflicts across resource regions are avoided (probability of collision is $\frac{1}{\Pi_{i=1}^{B} P_i \times n^B}$, where B is the number of BSs and evaluation is in section 3.5). Each chunk is probed for twenty-five frames and the BDR on each of these chunks is estimated (step 4,5); the maximum BDR (across chunks) is compared to the client's current recorded BDR (step 6,7).

Convergence in Time

We consider two approaches to probing in the time domain: sequential and binary search (Figure 3.9). In sequential probing (outlined in the pseudo-code), the vertical strip to be probed is advanced sequentially by Δt till a gain exceeding α or a high value ($>\beta$=0.8[2]) of BDR (for BS constituting the maximum reuse zone) is seen compared to the current BDR (step7). Otherwise, the current BDR is updated based on the maximum BDR with the recent probing (step 21). In binary search, two adjacent vertical strips are probed and the BDR with the left and right strips are compared. If $BDR_{right} > BDR_{left} \cdot \alpha$ then common reuse zone has been detected (time convergence). Otherwise the maximum value of the BDR (across frequency chunks) is compared with the current BDR to

[2]Considering the wireless loss of ACK bursts, we set β to be lowest but close to 1, while it yields 0 probing error.

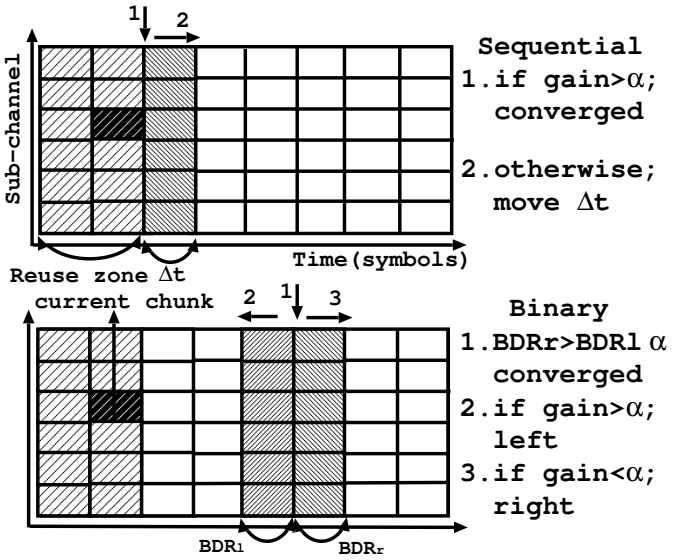


Figure 3.9: Two probing methods of RADION (sequential and binary search).

determine if the direction of adaptation should be to the left ($> \alpha$) or right ($\leqslant \alpha$), and the current BDR is updated only when the region probed is within the maximum reuse zone. If there are multiple clients in class $2h$, they are probed together in each of the chunks and decisions are made with respect to each client. Since different clients may receive interference from different cells, the common reuse zone varies with respect to clients. Time domain probing continues till the common reuse zone for each client in class $2h$ is determined, with the largest common reuse zone determining the termination of the transition zone for the cell.

Convergence in Frequency

Once the common reuse zone is detected, the BS simultaneously has the BDR information on n frequency chunks, with multiple frequency chunks potentially available for operation. We consider two approaches for the selection of a frequency chunk (step 8,9): greedy and Gibbs sampler (Algorithm 4 and 5). While the greedy scheme is deterministic and picks the chunk yielding the

highest BDR, the Gibbs sampler is probabilistic and favors chunks with higher BDR [61]. It has a temperature parameter T, which can be varied with time to provide an annealed version that converges to stable states of low potential (low interference and high BDR). While convergence in the time domain (common reuse zone) can be achieved with high accuracy, frequency domain convergence is sensitive to frequency selectivity and channel errors. Hence, the frequency chunk selected is confirmed by another iteration of fine adaptation, which probes the frequency chunks alone (i.e., no increment of Δt) after q frames (step 10-13). If the same frequency chunk yields the highest BDR, then there is convergence in the frequency domain and the chosen time and frequency parameters are employed for operation till the next coarse adaptation (step 14-17). Otherwise, probing in the frequency domain is repeated every q frames till contention (interference) is alleviated (step 18,19). Thus, by probing vertical strips of frequency chunks, RADION determines both the common reuse zone (s_t) and the set of sub-channels ($\mathcal{C}$ in isolation zone) simultaneously; this leads to quick convergence. Note that BSs adapt both control channel and frequency resource isolation (i.e., FDIctrl isolation [23]) when converging/operating in frequency domain to completely avoid the interference from the control channels caused by neighboring cells.

Handling Network Dynamics

Client (dis)-associations impact the traffic load of a cell and consequently, the resource allocations in the isolation zone. From an ideal (centralized) resource allocation standpoint, every cell has to share the frequency resources in the isolation zone of a frame in the contention domains that it belongs to (cliques in the interference/conflict graph), with its ideal share being determined by the size of the largest contention domain that it belongs to. When a new cell is introduced or an existing cell leaves (or traffic ceases) the contention domain, the existing share of frequency resources decreases or increases, respectively. However, this change has to be detected by each cell in a completely distributed manner in RADION; this allows cells to contract and expand their sub-channel allocations in the isolation zone. Such a feature is also useful in improving the

resource utilization in the network. Note that, since a cell (A) does not have information on its contention domains (requires global knowledge), it computes its fair share (say x, $x \leqslant$ ideal share) based on its interfering neighbors. However, if one of its neighboring cells (B) belongs to a larger contention domain (hence has a lower share $< x$), then some resources (unused by B) will be under-utilized in cell A's contention domain. The ability to probe and expand resource usage will avoid such under-utilization that is a by-product of distributed operations (smaller granularity of chunk sizes ($< \Delta f$) lowers such under-utilization).

RADION allows a cell to adapt its sub-channel usage as follows. Although a cell selects one of the frequency chunks for operation upon convergence, it keeps track of the BDR in other frequency chunks and the potential set of chunks unused by neighboring cells. It continues to monitor such unused chunks for an additional period of P frames, giving its neighbors enough time to detect and use their fair share. If some of these chunks still continue to be available, then the BS decides to expand its resource usage by adding one of the unused chunks to its allocation (step 16). Adding one chunk at a time, allows other cells in its contention domain to also share the unused resources in a fair manner. This expansion of resource usage will address cases when cells switch off or cease to carry traffic. However, if after expansion, the ceased traffic in a cell restarts or a new cell enters the contention domain, this will be detected in the form of degraded BDRs on the frequency chunks or as a new interferer sensed by its clients. In either case, the BS will contract to its conservative share of sub-channels (in the isolation zone) computed based on its updated set of interfering neighbors and re-run its adaptation algorithm. Any resulting under-utilization in its contention domains will be addressed subsequently through its resource expansion mechanism.

3.5 SYSTEM EVALUATION

Testbed: Our testbed consists of three femtocells deployed in an indoor environment reflecting a typical femtocell deployment with walls, doors, and rooms (Figure 3.10). We use PicoChip's femto BSs that run WiMAX (802.16e) [8]. Our clients are laptops with commercial USB WiMAX cards [1]. Each femtocell has

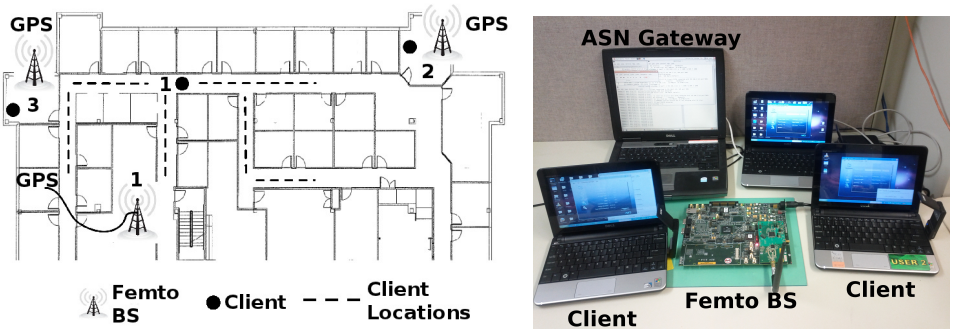


Figure 3.10: Deployment and picture of our WiMAX testbed.

multiple clients (class 1, 2l and 2h clients) and all class 2 clients are located within the transmission range of other femtocells. All the three cells operate on a 8.75 MHz bandwidth with a carrier frequency of 2.59 GHz. We achieve synchronization across femtocells via external GPS modules [13], aligning the start times of frames.

Our testbed is restricted to three femtocells, however, the impact of interference generated by two interferers is sufficient [24] to evaluate the proposed framework. We augment evaluations with simulations, where we study scalability under dense deployments with large number of BSs. Unlike WiFi, femtocells perform synchronous frame transmissions without carrier sensing. Hence, it makes more sense to generate different interference topologies by varying the locations of clients not the BSs. This provides a finer control on the inter-BS interference magnitude as opposed to changing the locations of the BSs, and also covers a wide range of scenarios that include both line-of-sight (LOS) and non-LOS links. We also consider ideal link adaptation where we sequentially run the experiment over all MCS levels and record the one delivering the highest throughput for the given topology.

Implementation: RADION is implemented on the PicoChip platform [10], which provides a *base reference* implementation of the 802.16e standard. We significantly extend and modify the MAC scheduler (≈2000 lines C code) to realize

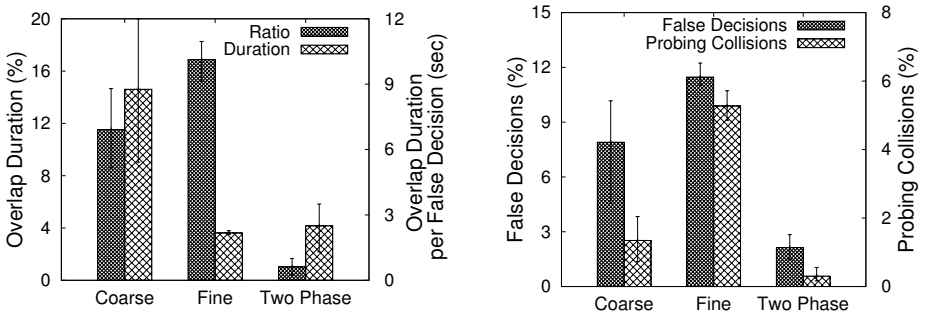
the components in RADION. Our key modifications include: (1) BDR estimator: The BS tracks the BDR status for each client connection and updates the BDR estimates as a moving average. BDR estimation plays an important role in both client categorization and probe-and-adapt mechanisms of RADION. (2) Client categorization: We introduce two measurement zones (i.e., free and occupied zones) for client categorization and schedule data bursts in each measurement zone. (3) Resource decoupling: Each BS tracks its clients' categories to schedule them in the appropriate zone. From the set of clients to be scheduled, the BS determines the traffic demand (data) from clients for each zone and splits the frame to schedule them simultaneously. (4) Probe-and-Adapt module: The two-phase adaptation algorithm is implemented on each BS; adaptations are triggered at frame boundaries. Further, RADION's modular nature allows us to incorporate the two variants of channel selection as outlined in section 3.3.

Prototype Evaluations

Having evaluated the first two components of RADION in section 3.3, here we focus on evaluating the two-phase adaptation algorithm. To understand the efficiency of algorithm we first evaluate the adaptation process in time and frequency domains in isolation, followed by their joint evaluation. We conclude the evaluation of RADION's adaptiveness to network dynamics. We create multiple clique topologies where clients in class 2h are within the transmission range of other BSs by varying client locations. The BSs operate on frames with 30 sub-channels and 22 time symbols for data bursts. Each experiment is run for at least 10 minutes and is repeated multiple times to generate confidence results.

Frequency Domain Convergence

Here, we fix the reuse zone of all the three cells to be the same, thereby eliminating the need for determining a common reuse zone. Hence, the focus is only on frequency domain convergence in the isolation zone for class 2h clients; every cell has to identify and operate on a contiguous set of 10 sub-channels each (3 contending cells). We evaluate RADION's two-phase (coarse + fine) adapta-

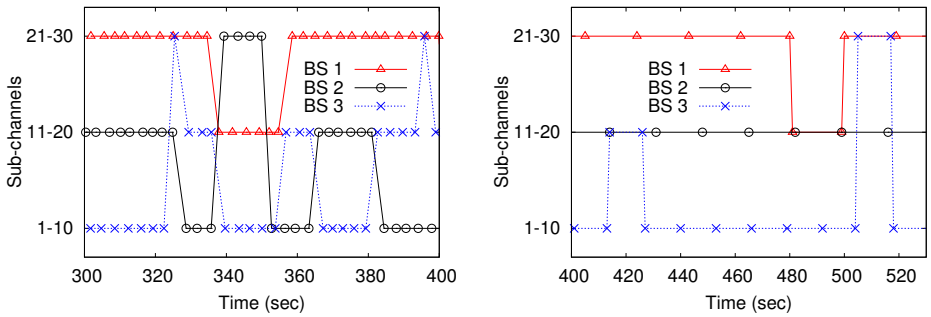


(a) Overhead (b) False decision

Figure 3.11: Two-phase adaptation outperforms others.

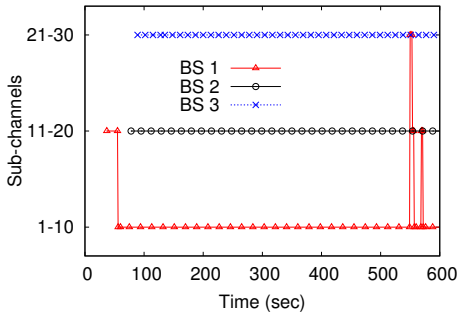
tion against the coarse and fine time scale adaptations in isolation. The coarse adaptation period is chosen to be a prime number of frames in [1000,6000] for each BS and is fixed for subsequent adaptations. Similarly, the period of fine adaptation is chosen randomly in [1,600] for every adaptation. Further, selection of sub-channels after probing follows the greedy approach (Gibbs sampling considered later).

To understand the impact of false decisions on system (throughput) performance, we consider the following metrics: **(**i) *fractional overlap duration*: the fraction of time that cells operate on overlapping resources (leading to collisions); **(**ii) *Overlap duration per false decision*: ratio of the net duration of resource overlap to the total number of false decisions (resource overlap $\equiv$ collisions); **(**iii) *Fractional false decisions*: ratio of net false decisions to total number of decisions; **(**iv) *Fractional probing collisions*: ratio of net probing collisions to total number of decisions. A probing collision occurs if two or more cells probe at the same time and choose the same resources for isolated operations resulting in a false decision. However, a false decision can also occur due to inaccurate BDR estimates and/or asymmetric interference patterns. We omit the throughput results and rather focus on the convergence to orthogonal resources among the BSs. Throughput degradation only occurs when the BSs operate on the same resource therefore, the *fractional overlap duration* reflects the throughput result indirectly.



(a) Fine adaptation

(b) Coarse adaptation



(c) Two-phase adaptation

Figure 3.12: Three convergence patterns. Two-phase adaptation incurs very few false switches (only twice) and provides very fast recovery from false decisions.

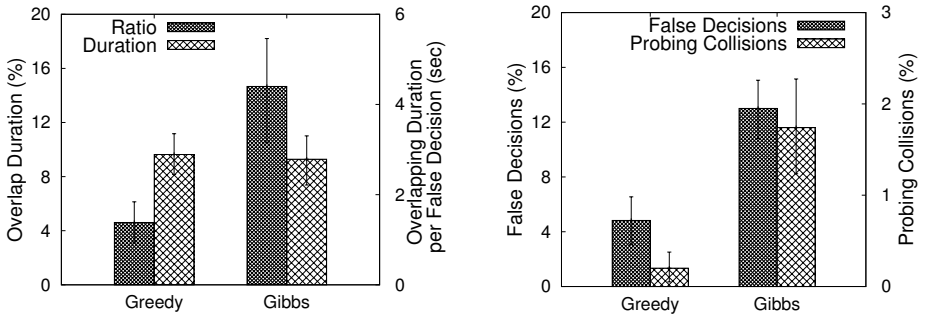
The results of multiple runs of frequency convergence are in Figure 3.11. Coarse adaptation incurs fewer false decisions compared to fine adaptation, but the time to recovery (overlap duration) is large per false decision. In contrast, while the overlap duration per false decision is small for fine adaptation, probing collisions and hence, false decisions increase with the number of active probes. RADION's two-phase adaptation strikes a good balance between coarse and fine adaptations, resulting only in 2% of the frames colliding and hence outperforms others in all metrics.

To illustrate, we present one particular example of sub-channel convergence as a function of time in Figure 3.12. We found similar patterns for all other

runs. While the cells maintain orthogonality of sub-channels for most part of the experiment, several differences are observed between the schemes. With fine adaptation (Figure 3.12a), at 352 seconds, BS2 switches from sub-channels 21-30 to 1-10, which is already occupied by BS3. This is a false decision due to an inaccurate probing estimate. After BS2 switches to sub-channels 1-10, BS3 switches to sub-channels 11-20 when its next adaptation is invoked (after q frames). Thus, each cell reacts to others' decisions to avoid interference and maintains orthogonal sub-channel usage for its class $2h$ clients. While the number of such switches is small with coarse adaptation (Figure 3.12b), recovery from a false decision is also slow. While recovery is fast with fine adaptation, multiple switches result in both increased probing overhead and false decisions. The two-phase adaptation (Figure 3.12c) exhibits the best of both adaptations; it incurs fewer switches while also providing fast recovery from false decisions.

Time Domain Convergence

Here, the three cells are pre-assigned orthogonal sets of sub-channels in the isolation zone. We focus on their convergence to the common reuse zone. Recall that RADION probes multiple frequency chunks and compares their BDRs to determine convergence to this common zone. Hence, estimation of BDR plays an important role in the accuracy of common zone determination. We study both convergence and the impact of the number of probing frames used, for estimating BDR. Sequential search is used for time domain adaptation. The three cells are set different reuse zone demands of 9, 13 and 17 symbols, respectively. BS3, having the maximum reuse zone, will quickly converge to the common reuse zone, while the other two cells will require adaptation. For a given number of probing frames (per chunk), we run the adaptation experiment over 100 times and determine the fraction of cases where the common reuse zone is accurately determined by cells 1 and 2. We repeat this experiment by varying the number of probing frames. We observe that 25 probing frames are sufficient to correctly determine the common reuse zone in over 94.5% of the cases (more frames only provided marginal improvements).



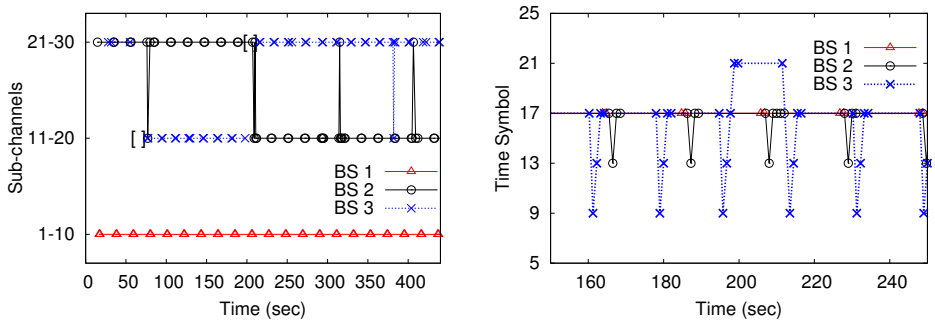
(a) Overhead (b) False decision

Figure 3.13: Greedy selection outperforms Gibbs sampler.

Joint Time-Frequency Convergence

With three BSs (1,2,3) in a clique topology with reuse zone demands set at 17, 13 and 9 symbols respectively, we run the joint time-frequency adaptation process of RADION at each BS. From the results of multiple runs in Figure 3.13, we see that the greedy sub-channel selection yields quick convergence with a low false error rate. The Gibbs sampler incurs a higher fraction of collided frames compared to the greedy. Since both the schemes employ the two-phase adaptation, the average collision duration per false decision is similar with the two schemes (Figure 3.13a). While one might expect the probabilistic nature of Gibbs sampling to yield better convergence [61], this is only true if inferences are based on sensing as opposed to probing; the probabilistic selection increases the number of probing collisions and hence false decisions (Figure 3.13b). Although the greedy approach is deterministic, diversity in sub-channel gain across BSs (and their clients) implicitly results in cells picking different frequency resources for their operation. Since both the schemes employ the same time-domain adaptation procedure, their convergence error in the time domain remains similar and less than 5%.

The time-frequency convergence patterns for the greedy scheme are shown in Figure 3.14. Here we present one example of the multiple runs but we confirm very similar pattens for all runs. The frequency adaptations are directly



(a) Frequency convergence (b) Time convergence

Figure 3.14: Convergence patterns of greedy selection.

indicated. False convergences in the time domain are indicated by brackets to capture the duration for which the cell operates with a wrong common reuse zone. We see that the number of probing collisions and the resulting quick switch in sub-channel resources to maintain orthogonality in the frequency domain. We omit the pattern of Gibbs sampling for the sake of brevity; however, we highlight that it shows many more frequent switches as the results (Figure 3.13a and 3.13b) show. The time convergence pattern for the greedy approach is expanded in Figure 3.14b. We see that BS3 computes the common reuse zone falsely as 21 symbols (instead of 17) at 200 secs; this is corrected in the next coarse adaptation period. Till the reuse zone is corrected, BS3's false decision does not cause interference to BS1's and BS2's class 2h clients since it continues to use its isolated resources in its transition zone. However, BS3's class 2h clients now incur unfairness as their operational resources are reduced by four symbols. Since the number of false decisions is very small, we let these correct themselves in the next coarse adaptation period.

In summary, we find that the joint time-frequency adaptation process in RADION yields quick and accurate convergence at each femtocell to the common reuse zone as well as to provide orthogonal sub-channels in the isolation zone.

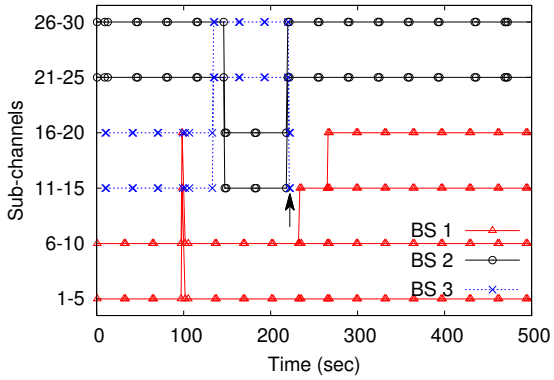


Figure 3.15: Frequency convergence (single contention).

Network Dynamics (Single contention domain)

We now evaluate RADION's ability to adapt to network changes. First, we consider a single contention domain where all BSs cause interference to each other. In the clique topology of three cells, RADION's adaptation algorithm allows each BS to detect its fair share and converge to orthogonal sets of 10 sub-channels in the isolation zone. The frequency convergence pattern is shown in Figure 3.15, where frequency resources are probed at the granularity of five sub-channel chunks. RADION's resource expansion mechanism allows each BS to expand its frequency resources in the isolation zone if it probes empty resources for more than a coarse adaptation period. When BS3 is switched off at about 210 seconds (indicated by the arrow), its sub-channels 11-20 become free. BS1 probes the availability of chunks 11-15 and 16-20 but decides to expand its resources only to 1-15, to allow fair access to other cells in the contention domain. However, since the chunk 16-20 remains available in the next adaptation period as well, BS1 continues to expand its resources to 1-20. BS2 fails to grab the chunk 16-20 due to an inaccurate BDR estimate. Thus, while RADION paves the way for a distributed fair sharing of unused resources, distributed operations without information sharing prevents it from controlling the utilization-fairness tradeoff in the best way possible.

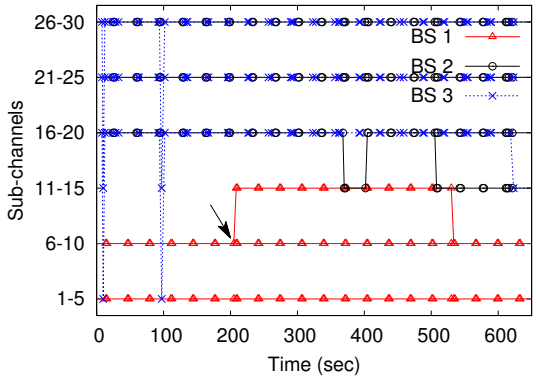


Figure 3.16: Frequency convergence (multiple contention).

Network Dynamics (Multiple contention domain)

Here we evaluate the adaptiveness in a multiple contention domain by creating chain topology BS3-BS1-BS2; the ideal fair shares of all BSs are 15 sub-channels each. However, since BS1 does not have global information, it computes its fair share to be 10 sub-channels (based on two interfering neighbors) without realizing that its neighbors belong to different contention domains. BS2 and BS3 compute their fair shares as 15 sub-channels each. RADION's resource expansion can help BS1 salvage under-utilized resources. The convergence is shown in Figure 3.16. Initially, while BS2 and BS3 converge to operate on sub-channels from 16-30, BS1 converges to sub-channels 1-10. RADION's resource expansion/contraction feature is enabled after 200 seconds. At this point, BS1 probes the available chunk from 11-15 and expands its allocation to 1-15; this remains stable till 370 secs. At 370 secs, BS2 expands its allocation to 11-30 due to an inaccurate probing, which is immediately rectified in the next adaptation. This again happens at around 510 secs. However, this time, before BS2 rectifies its decision, BS1 responds to interference by contracting its allocation back to its initial fair share of 1-10. This in turn prompts BS3 to probe and expand its allocation to 11-30 (similar to BS2).

The above experiment demonstrates that when a new BS joins the contention

domain, it is detected by other cells in the domain; these BSs update (contract) their fair share and run the adaptation process to determine their isolated resources. Stated otherwise, it succinctly captures both the expansion and contraction features incorporated into RADION to track network changes. It also indicates the transition between a fair allocation and one with high utilization. However, to finely control such transitions, information exchange across multiple cells is required, which may not be feasible in residential environments. RADION's *best-effort* utilization-fairness tradeoffs are particularly suited for such environments.

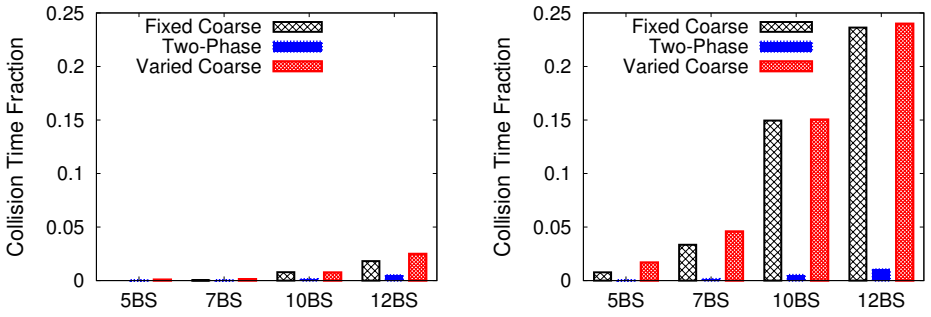
Evaluation through Simulations

The purpose of simulations is to help evaluate the effectiveness of RADION in a large-scale network (severe interference and complex topologies). Given that probing chunks in the frequency domain (sub-channel selection) indirectly controls convergence in the time domain, here we only focus on conflict resolution (i.e., probing) in the frequency domain. We increase the number of BSs to stress test RADION's convergence in the frequency domain.

Simulation Set-up

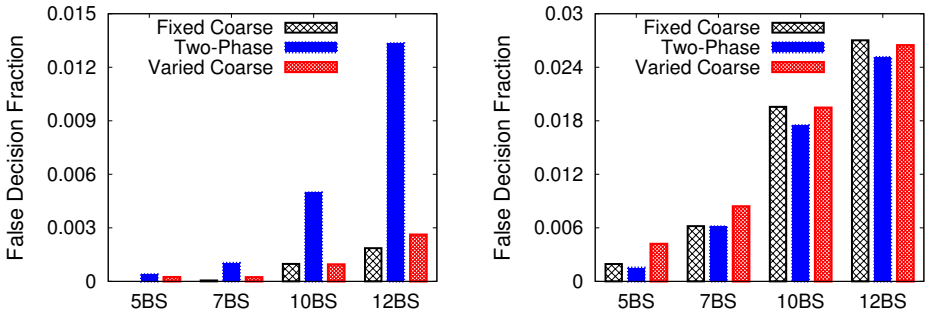
We consider two versions of coarse adaptation: (i) a BS picks a fixed prime number of frames P from $[8000, 16000]$, and (ii) P is varied (randomly) across adaptation periods. A larger range is chosen for P to allow multiple cells to pick the prime periods without inducing excessive probing collisions.

Fine adaptation period (q) is drawn from the range $[n \times s, 600]$, where n is the number of BSs and s is the number of probes sent on each chunk (set to 10 frames in the simulation). The variable range is based on the observation that smaller ranges are unfair to the case where large number of BSs simultaneously probe and adapt. In addition, when we evaluate fine adaptation in isolation, we simulate by changing the maximum period (max. q) to 600, 1600, 2600 and 3600 (min. period of $n \times s$ is still common). We use the same evaluation metrics as in the prototype evaluation and simulate by varying the number of BSs that



(a) Greedy selection

(b) Gibbs selection



(c) Greedy selection

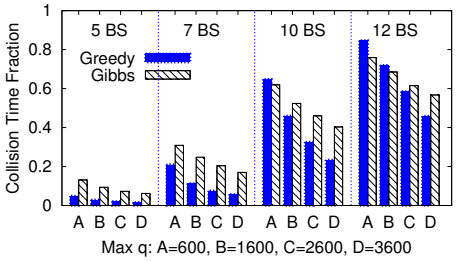
(d) Gibbs selection

Figure 3.17: Greedy selection outperforms Gibbs selection in collision and false decision. RADION's two-phase adaptation provides a significant reduction in the collision time.

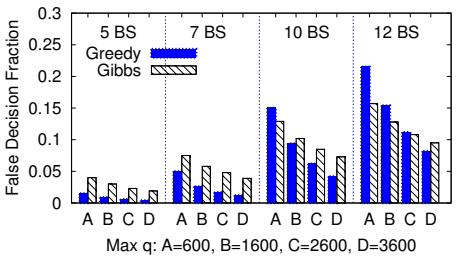
simultaneously sample chunks and decide on the chunk of operation. Each measurement is the average of 25 randomly generated parameters (P and q).

Two-phase Adaptation

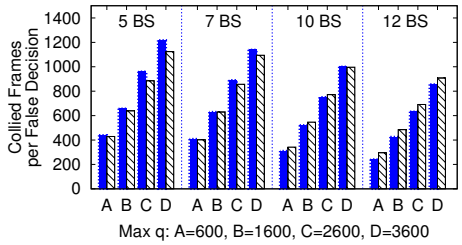
It is seen from Figure 3.17 that RADION's two-phase adaptation results in a lower fractional collisions than coarse (both fixed and varied) and fine adaptation in isolation. These results clearly corroborate our findings from the prototype evaluation, thereby demonstrating RADION's scalability in dense deployments. In addition, the Greedy selection outperforms Gibbs sampling.



(a) Collision time fraction



(b) False decision fraction



(c) Number of collied frames per false decision

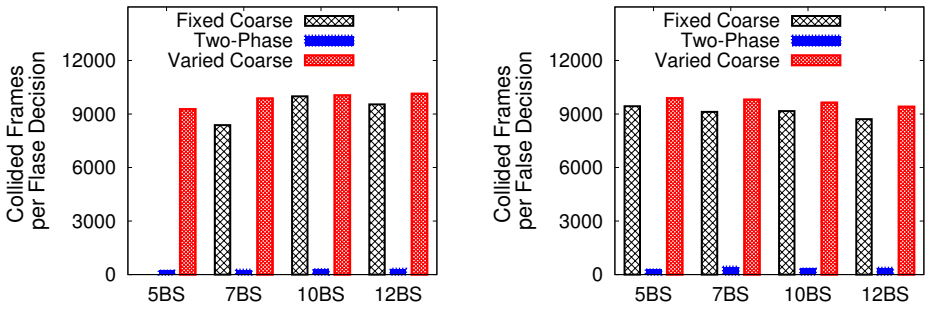
Figure 3.18: Results for the fine adaptation process. Fine adaptation is simulated with different period of range (max q).

Similar to the prototype evaluation, the reason behind this is the fact that since the samples are actively taken (by transmitting bursts) from each chunk, they may not be representative of the state (occupied vs. free) of each chunk due to sampling collisions. This coupled with the probabilistic nature of the frequency selection incurs a higher fraction of false decisions (plotted in Figure 3.17c and 3.17d) and hence a larger number of collisions.

Probing Duration

While fewer false decisions are incurred with coarse adaptations for Greedy selection (Figure 3.17c), the collision duration per false decision is longer than two-phase adaptation. This is mainly attributed to the fact that the utilizing a fine adaptation period reduces the observation time before sampling and making a decision. When a BS observes the state of its own chunk (i.e., not sampling), other BSs also probe the same chunk and result in collisions. If the observation duration is not long enough, these collisions result in an inaccurate BDR average for the current chunk. This results in a few back-to-back false decisions until the BS obtains an accurate measurement from each chunk. However, we show that even with this characteristic, two-phase adaptation still provides significant benefits of reduced fractional collisions. The importance of the length of observation duration can also be seen from Figure 3.18, where a larger range helps in reducing fractional collisions for fine adaptation in isolation. Note that even with larger ranges, fine adaptation alone performs worse than RADION's adaptation.

The contrary is true with fine adaptation (smaller collision duration per false decision but a higher number of false decisions) as seen in Figure 3.18. RADION combines the best of coarse and fine adaptation and provides a significant reduction in the time spent in collisions. As with prototype results, the greedy approach to sub-channel selection outperforms Gibbs sampling; increased probing collisions and hence, false decisions are seen with Gibbs sampling. Figure 3.18 shows the impact of q's range on fine adaptation with respect to the number of BSs. Increasing the range of q provides more time for estimating BDR on the current frequency chunk. Hence, false decisions due to inaccurate BDR estimates and thus, collision durations are reduced. However, adapting q's range is a double-edged sword. While it results in less number of false decisions, it incurs a higher penalty of a false decision since the probing collisions cannot completely be eliminated. In practice, the design of such a system that eliminates probing collisions is hard (without coordination) due to the asynchrony of decisions among BSs. We also show that fine adaptation alone does not perform well since it incorporates a short observation duration which results in many



(a) Greedy Selection (b) Gibbs selection

Figure 3.19: Two-phase adaptation very quickly corrects false decision, and hence minimizes the collisions of operating on the same frame resources.

wrong decisions. Hence, adapting q's range alone is not sufficient; the two-phase process as in RADION is needed.

False Decision

In Figure 3.19, we plot the number of collided frames per false decision. We observe that the penalty of a false decision is significantly smaller in two-phase adaptation due to fine adaptation periods that are more reactive than coarse adaptation. On the other hand, for coarse adaptations (both fixed and varied), the penalty of a false decision is typically on the order of a P value (there is negligible variation between Greedy and Gibbs). This means that once a false decision is made, it is corrected only until the next coarse adaptation period of another BS which is chosen among large primes to address sampling collisions.

Convergence

We evaluate the convergence of our algorithm with large-scale network (12 BSs). The time duration when more than two BSs operating on the same frequency resources is about only 30 seconds during 100 minutes simulation. In other words, 12 BSs stay on orthogonal resources (converged) 99.5% of time. We also

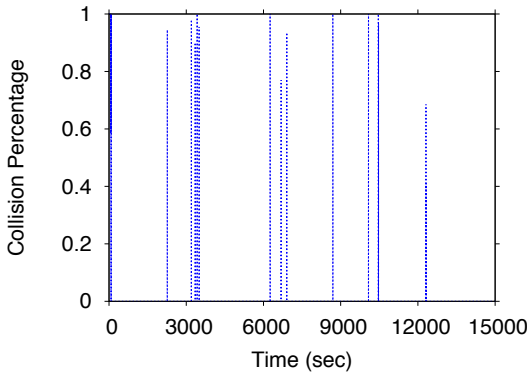


Figure 3.20: Total collision duration is less than 0.7%.

measure how quickly BS converges to the orthogonal resources after collisions and the average time takes to correct wrong decision is 1.62 seconds for two-phase adaptation while coarse adaption takes 50.68 seconds. These results show quick convergence and correction of algorithm.

Stress Test

Here, we perform a stress testing of RADION's frequency selection in the 10 BS scenario and present an example convergence trace. Initially, all BSs start on the same frequency chunk and try converging on orthogonal chunks. If, in a frame, at least two BSs use the same chunk, we mark this frame as a collision frame. We then calculate the number of such frames in a second and get the ratio over 200 frames (in WiMAX, 200 frames are transmitted in a second). The simulation is run for 3,000,000 frames (15,000 seconds). Figure 3.20 plots the collision percentage; spikes indicate that some BS made a wrong decision and resulted in collisions. A collision percentage of 0 means that all BSs operate on orthogonal chunks at that particular instant. It is seen that with RADION, BSs stay on orthogonal chunks with very short interruptions due to false decisions (resolved quickly with fine adaptation). We observed that the total collision duration due to false decisions was $\approx$ 100 seconds - less than 0.7% of the total

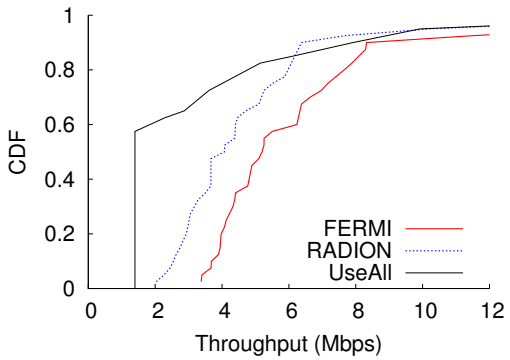


Figure 3.21: Performance comparison.

active time of the BSs.

Comparison to Other Solutions

We compare the performance of RADION against other resource management schemes, e.g., FERMI and UseAll. Again, we only compare the frequency resource allocation since FERMI and RADION adapt different reuse zone convergence strategies. Our purpose is to understand the relative sub-optimality introduced due to the lack of cooperation across BSs in RADION. In FERMI, a central controller generates an interference map between BSs and assigns frequency resource to BSs using clique based allocation [24]. On the other hand, each BS in RADION determines the frequency allocation based solely on its neighboring (interfering) cells. For the baseline strategy, we introduce the UseAll scheme where each BS uses all available resources in a greedy manner. Figure 3.21 shows the CDF of throughput of each BS (simulated with 40 BSs). The median throughputs are 1.4 / 4.1 / 5.1 Mbps for UseAll, RADION and FERMI, respectively. The performance of RADION is superior to UseAll and close to that of FERMI. Since FERMI constructs the global view of interference relations across the BSs, it optimally allocates the resources (maximizing frequency reuse), and hence provides the best performance. However, RADION still provides 36% higher throughput (on aggregate) compared to the base-

line strategy (without the inter-cell coordination) without incurring the central computation overhead of FERMI, thereby balancing both performance and overhead.

3.6 SUMMARY OF RADION

Femtocells are usually deployed in a distributed manner, thus coordinations across the femtocell are hard to achieve. Resource management in multi femtocell network is required to incorporate such characteristic of femtocell deployment. In addition, it needs to address the interference across the neighboring cells and to maximize resource reuse for increasing spectrum efficiency. To achieve this goal, we design and implement RADION, arguably the first self organizing resource management framework for OFDMA femtocell networks. RADION consists of four key building blocks, *client throughput estimation, client categorization, resource decoupling* and *two-phase adaptation and allocation*. RADION allows appropriately chosen clients to opportunistically reuse the spectrum while isolating resources for the other clients in a distributed way. We implement RADION using a WiMAX testbed to show its quick convergence to an efficient resource allocation in real settings. We also demonstrate the scalability and efficacy of RADION in larger scale settings with simulations. We only consider downlink performance, however, a similar approach can be applied to the uplink.

4 A PRACTICAL MULTICELL BEAMFORMING SYSTEM FOR OFDMA SMALL-CELL NETWORKS

4.1 INTRODUCTION

With reduced cell sizes and dense deployments, small cells are geared for increased spatial reuse of spectral resources – a valuable and scarce commodity in next generation wireless networks (e.g., WiMAX, LTE, LTE-A). Given the dense deployment of small cells, interference plays a key limiting factor in harnessing their potential. While the sheer scale limits planned deployment of small cells (similar to WiFi), handling interference is a very different problem in small cells compared to WiFi. This can be attributed to their synchronous access mechanism (borrowed from macrocells), coupled with OFDMA (orthogonal frequency division multiplex access) transmissions, wherein multiple users are served in the same frame. Earlier works on interference management in small cell networks [23, 24] employed interference avoidance in the time or frequency domain by allocating orthogonal resources to interfering small cells, therefore part of the spectral resources have to be sacrificed between interfering cells. In this work, we aim to avoid such sacrifices of spectral resources by exploring interference management for small cells in the spatial domain through beamforming antennas.

Employing beamforming or directional antennas for spatial reuse in a multi-cell set-up has been considered in the context of WiFi [73, 74]. However, such approaches face a key limitation when it comes to practical realization in that a single client is assumed for each AP when computing interference conflicts and determining the spatial reuse schedule. When the client scheduled for an AP changes, the interference conflicts change therefore it requires a re-computation of the schedule which is potentially at the granularity of every packet. This makes it hard to realize such solutions for WiFi networks and more so for small-cell networks, where multiple clients are scheduled in each OFDMA frame. Hence, the goal of this chapter is to leverage beamforming for spatial reuse across small cells but at the same time decouple it from per-frame scheduling

at the small cell base station (BS), thereby allowing for beam selections to be computed only at the granularity of seconds (hundreds of frames).

Executing beam selection at coarser time scales compared to client scheduling allows for tangible spatial reuse benefits across cells. However, the beam chosen for a small cell must deliver good transmission rates to all users that are associated (and hence can be scheduled) with the small cell in order to realize the throughput gains from spatial reuse. Hence, we argue that to realize practical and efficient spatial reuse with small cells, it is important to not just decouple beam selection from scheduling but also integrate beam selection with client association. Towards this goal, we propose *ProBeam* – a practical system that enables joint multi-cell beamforming and client association for increased spatial reuse in small cell networks.

4.2 BEAMFORMING

Beamforming is one of the core features in next generation networks that is adopted to improve SNR at the intended receivers while decreasing interference at unintended receivers. A beamforming system typically uses multiple antenna elements in an array to form various beam patterns. Beam patterns reinforce transmission energy in desired directions by weighting the signal from the antenna array in both magnitude and phase. Beamforming can be either switched (directional) or adaptive. In switched, a pre-determined set of directional beam patterns covering the azimuth are stored and chosen based on coarse feedback (SNR or RSSI) from the client. In adaptive, fine-grained feedback of channel estimation from the client is used to adapt the beam pattern on the fly to reinforce multipath components and maximize the SINR at the client. By adapting to the instantaneous multipath channel, adaptive provides higher gain (at the cost of increased feedback) compared to switched. However, at the same time, it is more sensitive to channel fluctuations and requires timely feedback to track the channel state - a limiting constraint especially during mobility and in multi-cell resource management.

Both switched and adaptive beamforming co-exist in a complementary manner in cellular systems. Macrocells are sectored in operation (e.g., three 120°

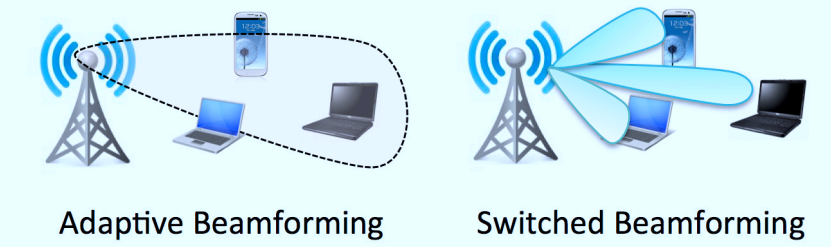


Figure 4.1: Illustration of Adaptive and switched beamforming.

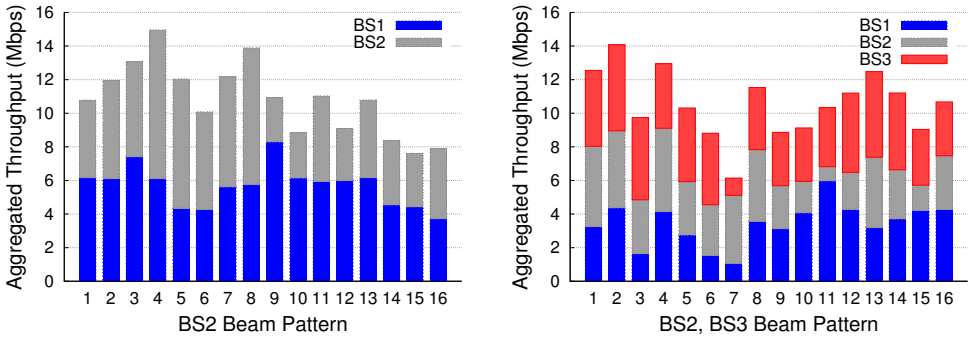
or six 60° directional beams), while adaptive beamforming is enabled to clients within each of the sectors separately. Unlike macrocells, where interference is restricted to cell-edges, thereby allowing for all sectors to operate in tandem, interference is a more pervasive phenomenon in small cells [23]. This requires small cells to select a single sector (switched beam) for operation in a frame (adaptable across frames) so as to avoid interference and maximize reuse among small cells in a dense deployment. Note that adaptive beamforming can still be enabled to clients within the sector of operation at each small cell (see Figure 4.1 for illustration).

4.3 Motivation for Coordinated Beamforming and Client Association

We now motivate the need to couple client association with multi-cell beamforming in order to maximize the benefits of spatial reuse. We present results from an experimental WiMAX-based network of four small cells, each equipped with an eight element phased array antenna (details in Section 4.5) to substantiate our claims.

Need for Coordinated Beamforming

Beamforming in a multi-cell context has two benefits: (i) increased link capacity through improved SNR, and (ii) increased network capacity through reduced interference (higher SINR) and hence higher spatial reuse. The beam choice of one cell impacts the interference seen by the clients of another cell, thereby



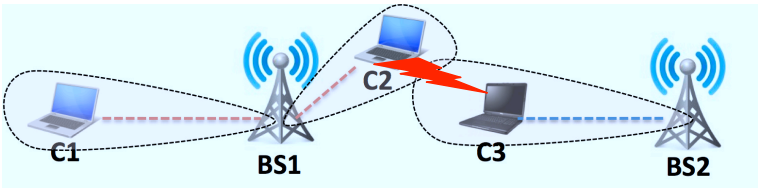
(a) Two cell network

(b) Three cell network

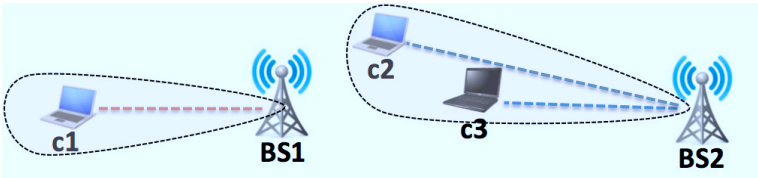
Figure 4.2: The aggregated throughput *w.r.t* the beam pattern. Small-cells need to coordinate the beam patterns to achieve the best performance in the network.

requiring a coordinated approach to beam selection across small cells for maximum reuse benefits. However, given the simplicity of un-coordinated, per-cell beamforming (focusing only on SNR), it is important to understand the benefits from coordination and hence the need for it.

We construct a topology with two small cells, each with one scheduled client. First BS1 cycles through all its sixteen beam patterns to determine the one yielding the best rate to its client (C1) in isolation. BS1 is then fixed to use its best beam to C1. Now, in the presence of BS1, BS2 is made to transfer data to its client (C2) on each of its 16 patterns sequentially. We plot the throughput observed at C1 (blue bars) and C2 (grey bars) as a function of the beam pattern used by BS2 in Figure 4.2a. Two observations can be made: (i) The interference projected by BS2 on C1 depends tightly on the beam chosen by BS2. C1 achieves its highest throughput (8.3 Mbps) when BS2 employs its 9th pattern and its lowest throughput (3.7 Mbps) when BS2 employs its 16th pattern. (ii) The beam maximizing the throughput of one cell does not necessarily maximize the multi-cell network throughput. While the 9th beam pattern maximizes C1's throughput, it is the 4th pattern that maximizes the aggregate network throughput. A similar behavior is also evident in the three cell experiment presented in Figure 4.2b, where the pattern (11th) maximizing throughput for



(a) SNR based association (C2 is associated with BS1)



(b) Flexible association (C2 is associated with BS2)

Figure 4.3: Flexible association could both mitigate the interference and increase the network performance.

C1 differs from the one (2nd) maximizing the aggregate network throughput. The throughput gain of employing the 2nd pattern over the 11th one is almost 40%.

Thus, *a well-coordinated beamforming algorithm across the small cells is indeed important to maximize the aggregate network throughput*.

Need for Joint Client Association

Client association has been traditionally employed to load balance clients between multiple cells so as to effectively utilize the capacity of each cell and network as a whole. However, in the context of multi-cell beamforming, client association has a bigger role to play. Note that, unlike in WiFi systems, where a single client is served by a cell at a time, OFDMA systems multiplex multiple clients in the same frame (diversity scheduling). This requires that the beam selected for the small cell cater effectively to all its associated and scheduled clients. Further, since the beam choice for a cell impacts the interference and hence performance seen by other cells, this naturally results in client association being closely coupled with multi-cell beamforming.

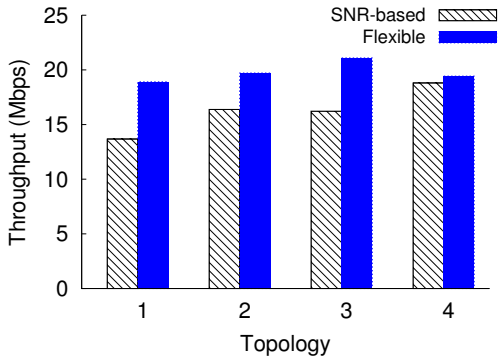


Figure 4.4: Joint association increases throughput by 40% comparing to decoupled (SNR based) association.

To see this, consider the following illustration in Figure 4.3. In conventional association, where SNR is used as a metric for client association, clients C1 and C2 will be associated to BS1, while C3 will be associated to BS2 based on (high) SNR and completely decoupled from beamforming. BSs will then determine the best beams to communicate with their respective clients. Let b1 and b2 be the only beams on which C2 and C3 can receive good signal strength from their respective BSs. Now, when BS1 is employing beam b1 to communicate with C2, this will receive interference from the beam b2 used by BS2 to communicate with its client C3. By fixing the client association, depending on the location of associated clients, the ability of beamforming to effectively suppress interference between cells is potentially limited. In contrast, by allowing for flexible association (Figure 4.3b), C2 can be associated with BS2 even though it has a lower SNR to BS2. This would allow BS2 to schedule C2 and C3 jointly on a beam that suffers no interference from that employed by BS1, thereby resulting in a potentially higher SINR for all clients.

To quantify the benefits of coupling client association with beam selection for small cells, we conduct the following experiment with two small cells and three clients, and generate multiple topologies by varying the client locations. We consider two association strategies: *decoupled association*, where the best coordinated beam (for maximum aggregate throughput) for each small cell is

selected after client association is done based on SNR; *joint association*, where the client association yielding the highest aggregate throughput is computed for all beam combinations between the two cells. The aggregate throughput results for these two strategies in Figure 4.4 indicate that joint association can yield gains as high as 40%, with an average gain of about 25%.

This in turn motivates the *need to jointly address client association with beam selection for small cells, whereby client association can be effectively used to maximize the spatial reuse potential of beamforming*.

4.4 DESIGN OF PROBEAM

In this section, we first present an overview of ProBeam followed by a detailed exposition of its components.

System Overview

Small cell networks can be deployed for enterprises as well as outdoors. A central controller (separate entity or one of the small cells) is designated to perform resource and interference management for a cluster (tens) of small cells jointly with a high speed backhaul available for information exchange between them. We expect ProBeam to reside in this central controller (CC). Self-organizing network (SON) enables the radio and network components to interact among themselves, and to configure and tune the mobile system automatically in real time [15]. We can consider that CC is one particular example of a centralized SON server because of the fact that CC coordinates and configures all small cells in the network. In addition, CC also synchronizes all small cells in the network for adapting beam patterns. The CC exchanges small messages with the BSs via dedicated wired connection, and hence it induces minimal overhead to the system and does not interfere with the BS's functionalities. Collecting SNR measurements follows IEEE 802.16e standard [8]. Note that while our primary focus is small cell networks, our system is equally applicable to WiFi networks as well.

ProBeam's spatial reuse solution operates in epochs which span several

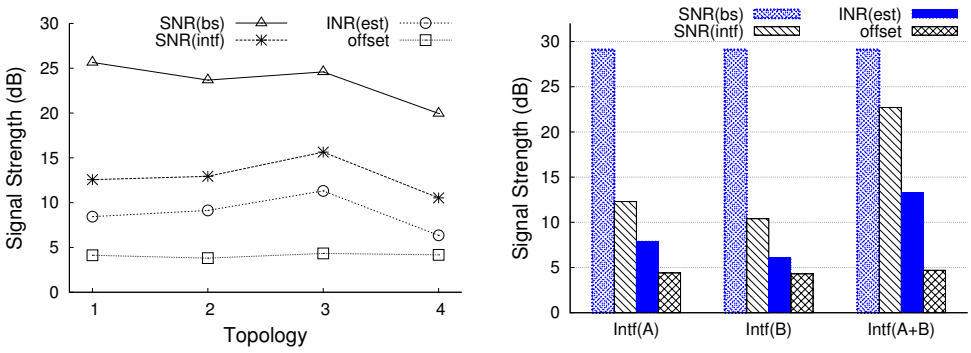
seconds (hundreds of frames). In each epoch, the sequence of operations are as follows. (i) *Interference estimation for beamforming*: The clients measure the average SNR on each of the beams from each of the BSs and forward it to the CC, which then infers their corresponding SINR for various beam combinations at the small cell BSs. (ii) *Joint beam selection and client association*: Based on the interference information collected, the CC runs its spatial reuse algorithm (for a desired objective) to determine the beam choice for each of the small cells as well as the clients that are associated with it for that epoch. (iii) Scheduling: Once each small cell BS receives its beam choice and client set, it begins scheduling its clients locally using its own scheduler (proportional fair, max-min fair, etc.) for each frame in the epoch, while applying the beam selected to the frame transmissions.

Interference Estimation for Beamforming

Reducing Complexity

Measuring the SINR directly at the clients for various beam configurations (interference) used by small cells is the most accurate approach. However, this would entail that each small cell cycle through each beam pattern, while keeping the beam patterns at other cells fixed and measuring the resulting SINR at all clients. This would however result in a total of $O(nm^n)$ measurements, where m in the number of beam patterns and n is the number of small cells. ProBeam measures only the client SNR from each of the small cells in isolation for the various beam choices and then uses this information to estimate the projected client SINR for a given beam configuration at the small cells. By allowing the small cells to operate in isolation during measurements, this significantly reduces the SINR estimation complexity to $O(mn)$. The key question remaining is the accuracy or lack thereof of such an estimation procedure.

Note that SINR can be expressed as $SINR_{ij} = \frac{SNR_{ij}}{\sum_{k \neq i} INR_{kj} + 1}$, where SINR at client j from BS i is related to its SNR and net interference to noise ratio from other BSs ($INR_j = \sum_{k \neq i} INR_{kj}$). Small cells being interference limited, $INR + 1 \approx INR$. In the logarithmic (dB) domain, the relation can be expressed



(a) Interference estimation (b) Estimation with 2 interferers

Figure 4.5: Accurate estimation of SINR from individual SNR estimates.

as SINR (dB) = SNR (dB) - INR (dB). Hence, in principle, the SINR at a client can be estimated from its SNR from the desired BS and its aggregate INR from all interfering BSs. For this to be possible, one needs to estimate each INR_{kj}, which can potentially be approximated as the client SNR when associated with the interfering BS in isolation (i.e., SNR_{kj}). However, in reality, the accuracy of such an estimation may depend on multiple factors such as quantization, offsets, estimation error, etc.

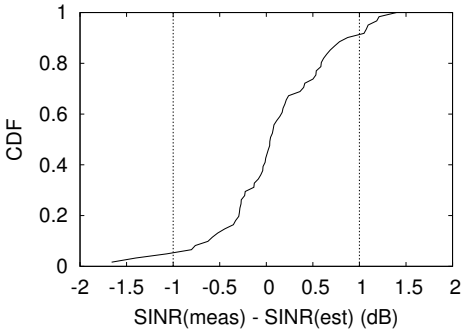
To verify this approximation, we conducted the following experiment with two small cell BSs, whose beam choices are such that they interfere at the client under consideration. The results are presented in Figure 4.5a. First, the client measures the signal strength from the associated BS in the absence (SNR_{BS}) and presence (SINR_{BS}) of interference respectively, from which we estimate $\mathrm{INR}_{est} = \mathrm{SNR}_{BS} - \mathrm{SINR}_{BS}$. Then, the client records the signal strength SNR_{intf} after associating with the interfering BS in isolation. Comparing SNR_{intf} to INR_{est} in Figure 4.5a, we see that there is a consistent 4 dB offset between the estimated interference and its corresponding signal strength and this remains fixed regardless of the topology and client SNR considered (SNR_{BS} varies from 19 to 26 dB). We attribute this constant 4 dB difference to the inherent offset β introduced (during client feedback) by the MAC and its quantization of the signal strength value reported from the PHY layer. β being

platform dependent, can be calibrated by the client and fed back to the Central Controller for its appropriate estimation of INR. Further, note that when SINR is directly measured, there is only one feedback value from the client. However, when SINR is estimated from SNR and multiple INRs, then each of the SNR feedback (corresponding to INR) introduces an offset that needs to be compensated. When appropriately compensated, the resulting estimation reduces to $SINR_{ij}$ (dB) $= SNR_{ij}$ (dB) $- 10\log_{10}(\sum_{k\neq i} SNR_{kj}) + \beta$. Note that since interference is aggregated in absolute units, the offset for the aggregate interference remains to be β in dB. This is observed in Figure 4.5b, where the offset in the presence of one (A or B) and two (A+B) small cell interferers remains to be the same 4 dB. Thus, with the help of isolated measurements from the small cells, it is indeed possible to estimate SINR, thereby resulting in a linear (in n) complexity of only $O(mn)$.

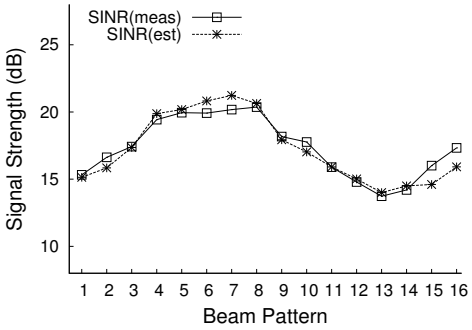
SINR Estimation Procedure

ProBeam initiates a measurement phase at the beginning of each epoch, where it operates each small cell BS in the cluster one after another in isolation. When activated, BS i applies its m beam patterns sequentially, each lasting ten frames. All the clients measure the average received SNR from BS i corresponding to beam pattern m. A client j forwards SNR_{ijm}, i.e., measured SNR from BS i with beam pattern m to the CC in ProBeam through its current associated BS. In WiMAX and LTE, clients automatically send Channel State Information (CSI) to BS periodically via dedicated uplink channel resources in every frame. We use such standard feature for obtaining our desired SNR measurements. Once ProBeam gathers SNR measurements from all the clients, then any desired SINR (in dB) for a given beam configuration ($\pi = \{\pi(i)\}$, $\forall i$, beam choices for small cells) can be estimated as,

$$\begin{aligned} SINR_{ij\pi}(dB) &= SNR_{ij\pi(i)}(dB) \\ &- 10\log_{10}(\sum_{k\neq i} SNR_{kj\pi(k)}) + \beta(dB) \end{aligned} \tag{4.1}$$



(a) CDF of estimation error



(b) Example of SINR measurements and estimations *w.r.t.* the beam patterns

Figure 4.6: (a) 95% of SINR estimates have less than 1dB difference. (b) The maximum estimation error is 1.4dB in case of beam pattern 15 and 16.

Note that SNR measurements can be done within $mn \times 10$ frames. For reasonable values of m (say 10 beams) and n (say 10 cells in a cluster), this would amount to 1 sec in LTE (for 1 ms frames). Also actual data is transmitted during the measurement phase, therefore we do not waste resources for SNR measurements. However, reuse cannot be leveraged, whose overhead (reuse loss) can be amortized as long as the epoch duration is several seconds.

Validation of SINR Estimation

To validate our estimation procedure, we conduct the following experiments with three small cell BSs and a single client. First, the client measures the SNRs from all three BSs for a given beam configuration in isolation and records them. Then, we make the client associate with one of the BSs and measure the SINR in the presence of the other two BSs projecting interference. The beam configuration is chosen so as to project interference to the client under consideration. We repeat the above experiment by changing the beam configuration as well as the topology (i.e., client locations) to obtain confidence in results. Measurements are taken at different client locations to generate plurality of interference scenarios and to also emulate different clients (varying BS deployment is considered

in Section 4.5). We obtain over 100 sets of measurements and present the CDF of the SINR estimation error ($SINR_{meas} - SINR_{est}$) in Figure 4.6a. As we can see, 95% of our SINR estimates have less than 1 dB error ($\leqslant 5\%$), with the highest estimation error being only about 1.65 dB.

Figure 4.6b further presents a microscopic example of SINR measurement and estimation with respect to beam patterns. Similarly, we create a three BS scenario, where two BSs project interference with a fixed beam choice, and measure the $SINR_{meas}$ at the client while applying 16 overlapping beam patterns (45° each) sequentially from the desired BS. For the same scenario, we collect the SNR measurements from all BSs in isolation for the same beam choices and estimate the $SINR_{est}$. As we see from Figure 4.6b, $SINR_{est}$ values are very close to $SINR_{meas}$ across all beam patterns, with the highest estimation error being 1.4 dB in the case of 15th and 16th beam patterns.

Our results clearly indicate the high accuracy of our SINR estimation method, thereby avoiding the complexity of obtaining measurements for all possible combinations of beam patterns at small cells.

Joint Client Association and Beam Selection (CABS)

Similar to other resource management problems, we can formulate our problem as a utility maximization problem in every epoch.

$$\text{Maximize} \sum_{j \in \mathcal{K}} U(t_j)$$

where t_j represents the average throughput received by client j in the epoch, $\mathcal{K}$ represents the set of clients and $U()$ is a function to capture the corresponding utility. Note that the choice of the utility function determines the fairness policy in the system. We assume utility functions to be concave and non-decreasing. This captures proportional fairness (defined by using the utility function $U(t_j) = \log(t_j)$) that is popular in the standards (WiMAX, LTE). While we need to decouple the time scales of operation for CABS from scheduling, it must be noted that the eventual objective is related to throughput and hence dependent on scheduling. Hence, to allow the decoupling, throughput needs

to be modeled as the average throughput received by the client over the epoch for a given scheduling policy. Our problem can be formulated as,

$$\begin{aligned}(\boldsymbol{\pi}^*, \boldsymbol{x}^*) \quad &= \quad \arg\max_{\boldsymbol{\pi}, \boldsymbol{x}} \sum_{j \in \mathcal{K}} \sum_{i \in \mathcal{S}} x_{ji} U(t_{ji}^{\pi}) \\ &\text{s.t.} \quad \sum_{i \in \mathcal{S}} x_{ji} \leqslant 1, \ \forall j \in \mathcal{K} \end{aligned} \tag{4.2}$$

where $\mathcal{K}$ and $\mathcal{S}$ represents the set of clients and small cell BSs respectively. Further, $\boldsymbol{\pi} = \{\pi(i), \ \forall i\}$ denotes the beam selection vector for all BSs, while $\boldsymbol{x} = \{x_{ji}, \ \forall j, i\}$ denotes the association vector for all clients ($x_{ji} = 1$ if client j is associated with BS i and 0 otherwise). t_{ji}^{π} indicates the client j's average throughput when associated with BS i under beam configuration $\boldsymbol{\pi}$ and depends on the SINR ($SINR_{ij\pi}$) seen by the client from BS i in the presence of interference from other BSs under the beam configuration $\boldsymbol{\pi}$ (see Eq. (4.1)).

A note on fairness

While fairness (starvation) among clients is typically achieved (avoided) over a longer time period, instantaneous per-frame decisions may favor clients with good channel conditions (e.g., proportional fairness). In the case of CABS, decisions are made at the granularity of epochs. Hence, if fairness is ensured over much longer time scales ($>>$ epoch), then several clients could be subject to starvation in an epoch (several seconds). This would increase the jitter perceived by such clients – a factor critical for real-time media and is hence not desired. Thus, it is more appropriate to ensure fairness within each epoch. This would allow all clients to be scheduled in every epoch. On the other hand, since beam selection decisions are fixed for the entire epoch, accommodating all clients could potentially limit the amount of reuse that can be leveraged in the epoch. Hence, to strike a balance between throughput performance (reuse) and fairness, an alternative is to restrict the utility functions to be non-negative in addition to concave and non-decreasing. This would account for fairness, while at the same time allowing for a small number of clients to be removed

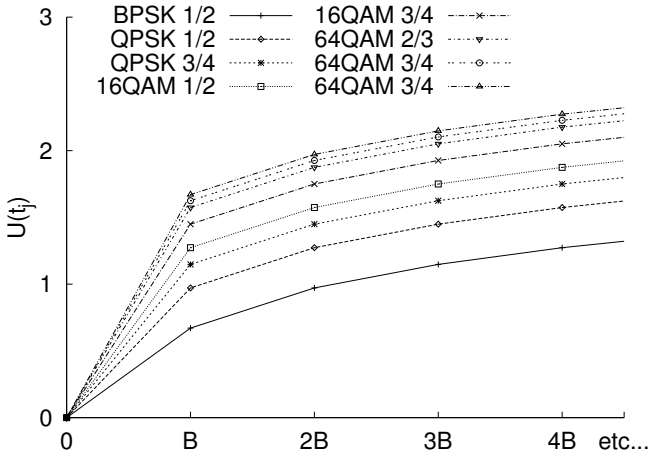


Figure 4.7: Utility as a function of feasible allocation t.

from scheduling in an epoch. By weighting the client utility functions inversely proportional to their throughput received (T_j) thus far, one can avoid starvation for all clients across epochs.

In the case of proportional fairness, we can modify the utility function as $U(t^{\pi}_{ji}) = w_j \log(t^{\pi}_{ji})$; if $t^{\pi}_{ji} \geqslant 1$ and 0 otherwise, where $w_j \propto \frac{1}{T_j}$. In WiMAX and LTE system, resource allocation to each user is done in terms of slots/resource blocks (each resource block is several symbols and each symbol carries at least one bit in lowest MCS). Since t is the throughput (bits allocated) in an epoch and the minimum allocation is a slot/resource block, $t > 1$ bits whenever a user is allocated. Thus, $t = 1$ will not happen if a user is allocated. Hence, we can assume 0 utility for both $t = 1$ and $t = 0$ and can define the utility function for $t \geqslant 1$ and 0 otherwise. In Figure 4.7, we present the utility function while varying the number of slots assigned to a user with respect to various MCS values. The x-axis values are multiple of B bits, where B is the least number of bits that can be sent over a single time-frequency slot using the lowest MCS (BPSK). Given that a single WiMAX slot (= 24 subcarriers $\times$ 2 symbols) consists of 48 modulation symbols, B is 48 bits with the lowest MCS, and hence B is

always > 1.[1] This indicates that $t > 1$ whenever resource is allocated to a user and $t = 0$ if there is no allocation. Figure 4.7 also shows that utility function is concave and non-decreasing.

Further, T_j at current epoch e is updated through an exponentially weighted moving average as $T_j(e) = (1 - \frac{1}{\alpha})T_j(e-1) + (\frac{1}{\alpha})t_j(e)$, where α is the filtering coefficient. Let r_{ji}^{π} be the average transmission rates (MCS) seen by client j in a slot when associated with BS i under beam configuration $\boldsymbol{\pi}$, and N be the total number of time-frequency slots in an OFDMA frame with M frames per epoch. In [62, 63], it has been shown that proportional fairness allocation of network resources is equivalent to the optimization of the utility function, $\max \sum_r w_r \log(x_r)$, where the number of slots are allocated among all the scheduled clients in the proportion of their weights (equal when $w_j = 1, \forall j$). This would in turn result in an average client throughput of $t_{ji}^{\pi} = \frac{NMw_j r_{ji}^{\pi}}{\sum_{k\in\mathcal{K}} x_{ki} w_k}$.

Hardness

For a given client association, the problem of beam selection is itself NP-hard [73, 74]. Hence, it comes as no surprise that our joint CABS problem is NP-hard as well. From the perspective of designing algorithms, it helps to understand if beam selection is the only source of hardness or does client association also contribute to the hardness. In this regard, we have the following result.

Theorem 4.1. *For a given beam selection, the CABS problem remains to be NP-hard.*

Proof. Consider the client association problem for a given beam choice for each of the small cells (π) under proportional fairness. The resulting objective function is then,

$$\begin{aligned} \mathbf{x}^* \quad &= \quad \arg\max_{\mathbf{x}} \sum_{j\in\mathcal{K}} \sum_{i\in\mathcal{S}} x_{ji} w_j \log\left(\frac{NMw_j r_{ji}^{\pi}}{\sum_{k\in\mathcal{K}} x_{ki} w_k}\right) \\ &\text{s.t.} \quad \sum_{i\in\mathcal{S}} x_{ji} \leqslant 1, \ \forall j \in \mathcal{K} \end{aligned}$$

[1] In LTE, there are 72 symbols in a resource block (minimum resource unit that can be allocated to a user), thus B is 72 bits when BPSK is used.

To prove our problem is NP-Hard, we consider a simple instance of it, consisting of just two small cells: Given a set of $\mathcal{K}$ users, find a solution to split the flows among the two small cells, say a and b so as to maximize the overall utility. Let the rate of a user be the same in both cells, $r_{j,a} = r_{j,b} = \frac{1}{w_j}$. Let $\mathcal{K}_a$ and $\mathcal{K}_b$ represent the set of users that are assigned to cells a and b respectively.

The throughput for a particular user on cell a is given by:

$$t_{j,a} = \frac{1}{\sum_{j \in \mathcal{K}_a} w_j} \quad \forall j \in \mathcal{K}_a$$

Hence, the utility U_a for all users assigned to cell a is given as:

$$\begin{aligned} U_a &= \sum_{j \in \mathcal{K}_a} w_j \cdot \log(t_{j,a}) \\ U_a &= \sum_{j \in \mathcal{K}_a} w_j \cdot \log(\frac{1}{\sum_{j \in \mathcal{K}_a} w_j}) \end{aligned} \tag{4.3}$$

Similarly, the utility U_b for all users assigned to cell b is given as:

$$U_b = \sum_{j \in \mathcal{K}_b} w_j \cdot \log(\frac{1}{\sum_{j \in \mathcal{K}_b} w_j}) \tag{4.4}$$

Now, let $X = \sum_{j \in \mathcal{K}_a} w_j$. Applying normalized weights, without loss of generality: $1 - X = \sum_{j \in \mathcal{K}_b} w_j$
Hence, the overall utility of the system U is given by

$$\begin{aligned} U &= U_a + U_b \\ U &= X \log(\frac{1}{X}) + (1 - X) \log(\frac{1}{1 - X}) \end{aligned} \tag{4.5}$$

The solution that maximizes the above utility function is $X = \frac{1}{2}$. Hence, the problem can now be equivalently defined as: Given a set of users $\mathcal{K}$, the solution should return a set of users $\mathcal{K}_a$ and a set of users $\mathcal{K}_b$ such that the sum of the weights of the flows belonging to the two sets are equal. This is an instance of the subset sum problem (partition problem): Given a set of k integers, the solution should return two subsets such that the sum of the integers of the first

set is equal to that of the second set. Our problem can be mapped to a subset sum problem where the input is the set K with elements that have a weight w_j, and the output will be two sets such that the sum of the weights of the elements of each set are equal. Since, subset sum problem is proven to be NP Complete, the proof is sufficient to show that our problem is NP Hard. □

Algorithm 6 CABS Algorithm

INPUT: average SNR ρ_{ji}^b, $\forall i \in \mathcal{S}, j \in \mathcal{K}, b \in \mathcal{B}$
OUTPUT: Beam selection $\pi(i)$ and client association $\mathcal{A}_i$, $\forall i \in \mathcal{S}$
Initialization of beam choices, i.e., $\pi(i), \forall i$
for $i \in [1 : |\mathcal{S}|], b \in [|\mathcal{B}|]$ **do**
  $\mathcal{L} = \emptyset$, $u_{ib} = 0$
  while 1 **do**
    $j^* = \arg\max_{j \in \mathcal{K} \setminus \mathcal{L}} \{\sum_{k \in \mathcal{L} \cup j} U(t_{ki}^b) - u_{ib}\}^+$
    % j^* is arg max of positive incremental utility, it will return only if the incremental utility is greater than or equal to 0, otherwise $\emptyset$
    if $j^* = \emptyset$ **then** break
    $\mathcal{L} \leftarrow \mathcal{L} \cup j^*$; $u_{ib} = \sum_{k \in \mathcal{L}} U(t_{ki}^b)$
  end while
end for
$\pi(i) = \arg\max_b u_{ib}, \forall i$

for $i \in [1 : |\mathcal{S}|]$ **do**
  for $b \in [1 : |\mathcal{B}|]$ **do**
    % Solve client association by varying only one beam element at a time
    $\pi(i) = b$, $\mathcal{A}_i = \emptyset$, $\forall i$
    while 1 **do**
      $(i^*, j^*) = \arg\max_{(i,j)\, \text{s.t. } j \notin \cup_i \mathcal{A}_i} \{\sum_{k \in \mathcal{A}_i \cup j} U(t_{ki}^\pi) - \sum_{k \in \mathcal{A}_i} U(t_{ki}^\pi)\}^+$
      % It will return only if the incremental utility is greater than or equal to 0, otherwise return $\emptyset$
      if $(i^*, j^*) = \emptyset$ **then** break
      $\mathcal{A}_{i^*} \leftarrow \mathcal{A}_{i^*} \cup j^*$; $u_{ib}^\pi = \sum_i \sum_{j \in \mathcal{A}_i} U(t_{ji}^\pi)$
    end while
  end for
  $\pi(i) = \arg\max_b u_{ib}^\pi$
end for

Algorithm

Since both components of our CABS problem are hard, we must carefully choose the interaction between these components in our solution. Unlike the beam selection problem, the client association problem, although hard, can be solved more efficiently. Hence, ProBeam proposes and employs a simple but efficient client association algorithm as the core building block for solving the CABS problem. At a high level, it solves the client association problem for a given beam configuration and the resulting utility is used to manipulate the beam configuration of small cells in an iterative manner till an efficient CABS solution is attained. The algorithm is given in Algorithm CABS.

The input to the algorithm is the average client SNR (ρ_{ji}^{b}) for the epoch with respect to its neighboring small cells when they employ different beams ($b \in \mathcal{B}$) in isolation (step 1). Using the approach in Section 4.4, the CC can then determine the average client rates in the presence (r_{ji}^{π}) and absence (r_{ji}^{b}) of interference. The CC first determines a bootstrap beam configuration for the small cells as follows (steps 3-12). For each of the small cells, it determines the beam that yields the highest utility in the absence of interference, assuming all active clients can be potentially associated with it, i.e., $\pi(i) = \arg\max_{b\in\mathcal{B}}\{\sum_{j\in\mathcal{K}} x_{ji} U(t_{ji}^{b})\}$. Note that t_{ji}^{b} depends on the scheduling policy and is hence coupled with the set of clients associated with the small cell. For example, in proportional fairness, $t_{ji}^{b} = \frac{NMw_j r_{ji}^{b}}{\sum_{k\in\mathcal{K}} x_{ki} w_k}$. Hence, even to determine a beam initialization $\pi(i)$, one needs to determine the set of clients (x_{ji}) that maximize the utility for the given beam in the absence of interference.[2] This can be done optimally by adding users one by one such that incremental utility is maximized (steps 6-10).[3] Specifically, for proportional fairness, the incremental utility (step 7) would correspond to,

[2]Note that accommodating all users can hurt the utility due to fixed frame resources but varying client rates.

[3]Initial client association is done by assuming no interference between cells (just a bootstrap beam configuration). In the absence of interference, the problem is no longer hard and is just a concave optimization problem that can be solved optimally by incrementally picking the best candidate.

$$j^* = \arg\max_{j \in \mathcal{K} \setminus \mathcal{L}} \sum_{k \in \mathcal{L} \cup j} w_k \log(\frac{NMr^b_{ki}}{1 + |\mathcal{L}|}) - \sum_{k \in \mathcal{L}} w_k \log(\frac{NMr^b_{ki}}{|\mathcal{L}|})$$

Note that j^* is arg max of positive incremental utility (it will return only if the incremental utility is greater than or equal to 0, otherwise return $\emptyset$). CABS does not consider scheduling a client if the incremental utility is not positive by adding such client, therefore j^* can be null. Since CABS algorithm takes positive incremental utility, it is possible that the algorithm finds no user that satisfies such condition after an iteration. In this case, the algorithm stops and breaks out from the iteration (step 8). This is why CABS might leave out some users and schedule them in the next epoch.

After the beam initialization, CABS algorithm perturbs the beam choice for each of the small cells, one by one and one beam at a time. For each of the beam choices at a given cell ($\pi(i) = b$), CABS retains the rest of the beam choices for the other cells unchanged and solves the client association problem for all the small cells jointly under the updated beam configuration to determine the new utility (steps 16-22). In each iteration (steps 19-21), CABS computes a client association while getting a utility value that is then used to determine which beam to pick for that cell under consideration. Again, (i^*, j^*) is arg max of positive incremental utility (step 19), and hence CABS might not find the client association and break out from the iteration (step 20). The actual client association, $\mathcal{A}_{i^*}$, is the one that is done in the last iteration (step 21) when the beam for the last cell is fixed and there are no more changes to beams. CABS then fixes the beam choice for the small cell as the one that yields the highest utility among all its choices (step 24). The same process is repeated for updating the beam choice for each of the small cells sequentially (steps 14-25).

Note that, although after one complete round of beam updates for each of the small cells (along with joint client re-association), we cannot guarantee convergence to the optimal solution, our evaluations in Section 4.5 reveal this is sufficient to obtain a performance very close to that of exhaustive search for beam configurations. CABS runs in $O(|\mathcal{K}|^2|\mathcal{S}|^2|\mathcal{B}|)$, with a large portion of the complexity coming from the client association module $O(|\mathcal{K}|^2|\mathcal{S}|)$.

Performance Guarantee

Given the hardness of the joint CABS problem, it is hard to establish an approximation guarantee for the entire algorithm. However, we can establish the following performance guarantee for the core building block in CABS, namely the client association part when the popular proportional fair scheduling policy is considered at the small cells.

Theorem 4.2. *CABS is a $\frac{1}{2}$—approximation algorithm under proportional fairness when beam configuration is given.*

We provide some definitions on matroid and sub-modularity that are relevant to the proof.

Partition Matroid: Consider a ground set Ψ and let S be a set of subsets of Ψ. S is a matroid if, (i) $\emptyset \in S$, (ii) If $P \in S$ and $Q \subseteq P$, then $Q \in S$, and (iii) If $P, Q \in S$ and $|P| > |Q|$, there exists an element $x \in P \backslash Q$, such that $Q \cup \{x\} \in S$. A partition matroid is a special case of a matroid, wherein there exists a partition of Ψ into components, $\phi_1, \phi_2, \ldots$ such that $P \in S$ if and only if $|P \cap \phi_i| \leqslant 1, \ \forall i$.

Sub-modular function: A function $f(\cdot)$ on S is said to be sub-modular and non-decreasing if $\forall x, P, Q$ such that $P \cup \{x\} \in S$ and $Q \subseteq P$ then,

$$\begin{aligned} f(P \cup \{x\}) - f(P) &\leqslant f(Q \cup \{x\}) - f(Q) \\ f(P \cup \{x\}) - f(P) &\geqslant 0, \quad \text{and } f(\emptyset) = 0 \end{aligned}$$

Proof. The sub-optimality of maximizing a sub-modular function over a partition matroid using a greedy algorithm of the form $x = \arg\max_{x \in \phi_i} f(P \cup \{x\}) - f(P)$ in every iteration was shown to be bounded by $\frac{1}{2}$ in [48]. We will now show that CABS is such an algorithm (step 18 being the key step), with our client association objective for a given beam configuration (π) corresponding to a sub-modular function to obtain the desired result.

Consider the ground set to be composed of the following tuples.

$$\Psi = \{(i, j) : \ i \in [1 : |\mathcal{S}|] \cup \emptyset, j \in [1 : |\mathcal{K}|]$$

Now Ψ can be partitioned into $\phi_j = \{(i,j) : i \in [1 : |\mathcal{S}|] \cup \emptyset\}, \forall j$. $i = \emptyset$ allows for the possibility of clients not being scheduled in an epoch. Let R be defined on Ψ as a set of subsets of Ψ such that for all subsets $P \in R$, we have (i) if $Q \subseteq P$, then $Q \in R$; (ii) if element $x \in P \backslash Q$, then $Q \cup \{x\} \in R$; and (iii) $|P \cap \phi_j| \leqslant 1,\ \forall j$. This means that R is a partition matroid. Now, it is easy to see that any $P \in R$ will provide a feasible schedule with at most one feasible association to a small cell for each client ($|P \cap \phi_j| \leqslant 1,\ \forall j$), thereby allowing the partition matroid R to capture our client association problem. Since each client can associate to only one small cell, our client association objective can be given as,

$$\begin{aligned} f(P) &= \sum_{i \in \mathcal{K}} \mu_i(P) \\ \text{where, } \mu_i(P) &= \sum_{j:(i,j)\in P} w_j \log\left(\frac{NMw_j r_{ij}^{\pi}}{\sum_{k:(i,k)\in P} w_k}\right) \end{aligned}$$

It can be seen that if $Q \subseteq P$, then $\mu_i(Q) \leqslant \mu_i(P)$ since the algorithm picks only elements that result in positive incremental utility. Hence, it only remains to be shown that for an element (i,ℓ) such that $P \cup \{(i,\ell)\}$ forms a valid schedule, then $f(P \cup \{(i,\ell)\}) - f(P) \leqslant f(Q \cup \{(i,\ell)\}) - f(Q)$. Now, define incremental utility $\Delta_P(i,\ell) = f(P \cup \{(i,\ell)\}) - f(P)$ and similarly define $\Delta_Q(i,\ell)$. Applying the objective function and simplifying, we can show that,

$$\begin{aligned} \Delta_P(i,\ell) &= w_\ell \log(NMw_\ell r_{i\ell}^{\pi}) - w_\ell \log\left(w_\ell + \sum_{k:(i,k)\in P} w_k\right) \\ &\quad - \sum_{j:(i,j)\in P} w_j \log\left(\frac{w_\ell + \sum_{k:(i,k)\in P} w_k}{\sum_{k:(i,k)\in P} w_k}\right) \\ \Delta_Q(i,\ell) &= w_\ell \log(NMw_\ell r_{i\ell}^{\pi}) - w_\ell \log\left(w_\ell + \sum_{k:(i,k)\in Q} w_k\right) \\ &\quad - \sum_{j:(i,j)\in Q} w_j \log\left(\frac{w_\ell + \sum_{k:(i,k)\in Q} w_k}{\sum_{k:(i,k)\in Q} w_k}\right) \end{aligned}$$

Thus, the difference between $\Delta_P(i,\ell)$ and $\Delta_Q(i,\ell)$ arises in the second (reduction) term, which increases with the number of elements in the allocation thus far. Since $Q \subseteq P$, the reduction term is more for P than for Q, resulting in $\Delta_P(i,\ell) \leqslant \Delta_Q(i,\ell)$. This establishes that the function $f(P)$ is indeed sub-modular. Further, our client association problem aims to maximize this non-decreasing sub-modular function over a partition matroid. Hence, picking the (client, small cell) pair yielding the highest marginal utility for a given beam configuration in CABS (steps 16-19) would correspond to determining

$$(i^*, j^*) = \arg \max_{(i,j)\in R} \{f(P \cup \{(i,j)\}) - f(P)\}$$

Thus, the sub-optimality of $\frac{1}{2}$ would then follow from the result in [47]. □

Scheduling

Once the CC determines the beam configuration and client association for the epoch, the appropriate beam and allowable client set are notified to each of the small cell BSs for configuration. Each small cell BS then locally runs its scheduling algorithm (e.g., proportional fair) among the associated clients for each frame in the epoch, while employing the chosen beam for its transmissions. Further, instantaneous channel rate feedback from clients is used in per-frame scheduling for leveraging multi-user diversity.

Practical Considerations

Mobile Clients

While beamforming algorithms work well for static clients, it is important to understand their limitations with respect to mobile clients. Note that, any adaptive beamforming scheme that relies on fine grained channel state information (CSI) will be highly sensitive to lack of timely and accurate CSI, both of which are hard to obtain during mobility. On the other hand, switched beamforming relies only on coarse grained channel feedback (SNR or RSSI) and hence is

less sensitive to mobility. As long as the epoch duration is not long enough (several seconds is reasonable), pedestrian to moderate vehicular speeds can be accommodated without warranting a completely new beam to be employed for the client.

Epoch Duration

Keeping the epoch duration long is conducive for implementation and overhead. However, it must also be capable of tracking traffic dynamics and client mobility. Allowing for a few seconds of epoch duration strikes a good balance between these objectives.

Average SNR

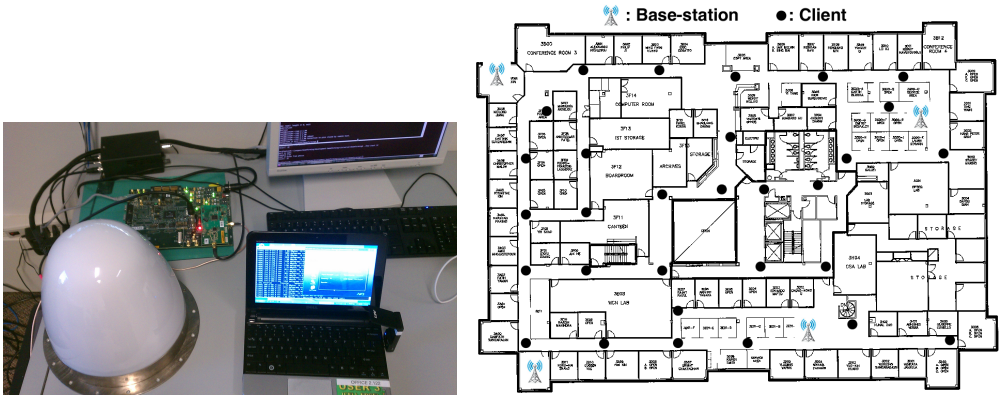
Note that employing average SNR feedback for CABS does not preclude the small cells from leveraging multi-user diversity during scheduling. Using average SNR allows CABS to account for the scheduling policy in its CABS decisions, while at the same time avoiding the need to jointly address scheduling with CABS - the latter not being conducive to a practical implementation. Once the CABS decisions are made at the CC and disseminated to the small cells, local per-frame scheduling at the small cells does leverage multi-user diversity with the help of instantaneous channel feedback from clients.

4.5 SYSTEM EVALUATION

We implement ProBeam on a WiMAX testbed and evaluate its performance through both experiments and large scale simulations. We first describe our testbed, comparing schemes, evaluation metrics and then present system evaluations.

Testbed and Prototype Implementation

Our WiMAX testbed consists of four small cells (deployed in an indoor enterprise environment), clients and a central controller as depicted in Figure 4.8a. The small cell BS is a PicoChip [10] WiMAX platform based on IEEE 802.16e



(a) Set-up: Small cell, beamformer and client (b) Test-bed deployment

Figure 4.8: Picture of one set of small cell and the deployment of four small cells testbed. Dots represent the client locations.

standard [8]. The BS is tuned to operate in a 10 MHz bandwidth with the center carrier frequency of 2.59 GHz, for which we have obtained an experimental license to transmit WiMAX signals over the air. In the absence of a macro cell to coordinate with, we use a GPS module to synchronize the WiMAX frame transmissions across the small cells. Each BS has an eight element (analog) phased array antenna [6] connected to its RF port. The antenna array generates sixteen overlapping beam patterns of 45° each, spaced 22.5° apart to cover the entire azimuth of 360°. The BS controls the antenna array through a serial port application that we have developed in C. There is a delay of one frame (5 msec) before a particular beam pattern is actually applied by the antenna following the command from the application. This is not an issue given the time scale of epoch or the measurement phase.

ProBeam is standards compatible and works with commercial off-the-shelf clients. We use Windows laptops with a WiMAX interface [1] and omni-directional antennas as our clients. Investigating directionality at the clients is part of our future work. We select 30 locations as marked in Figure 4.8b for client deployments. The clients are oblivious to beam selection at BS and simply measure the SNR and report them back to the BS for SINR estimations

through standard feedback mechanisms. Our experiments have verified that the SNR received on each beam is relatively stable over several seconds for static clients. This gives confidence to the SNR measurements reported by clients in the measurement phase.

All algorithms (CABS and reference schemes) are implemented on the CC and do not require any changes or operational overhead to the BS. All BSs are connected to the CC through an ethernet switch in our set-up.

Prototype Evaluations

Topologies and rate adaptation: Each data point in our result is averaged over multiple topologies, which are generated by picking random subsets of client locations (among 30) for a given number of clients. Further, unless otherwise specified, we consider topologies with four small cells and twenty clients. To remove the influence of rate adaptation algorithms, we consider an ideal PHY rate adaptation by trying out all MCS and record the highest throughput (best MCS) for a client given a network configuration.

Reference schemes: We evaluate the performance of our CABS algorithm in ProBeam against the following benchmark algorithms.

- *Decoupled:* Client association is decoupled and first computed based on SNR, followed by determination of coordinated beams for each BS using the same beam selection component as in CABS.
- *CABS-all:* Allows for joint determination of client association and beam selection as in CABS but requires that all clients be associated and scheduled in every epoch.
- *UB-beam:* Employs the same client association component as in CABS but exhaustively searches over all possible beam combinations at BSs - serves as an upper bound for beam selection in CABS.
- *UB-assoc:* Employs the same beam selection component as in CABS but exhaustively searches over all possible combinations of client association

- serves as an upper bound for client association in CABS.

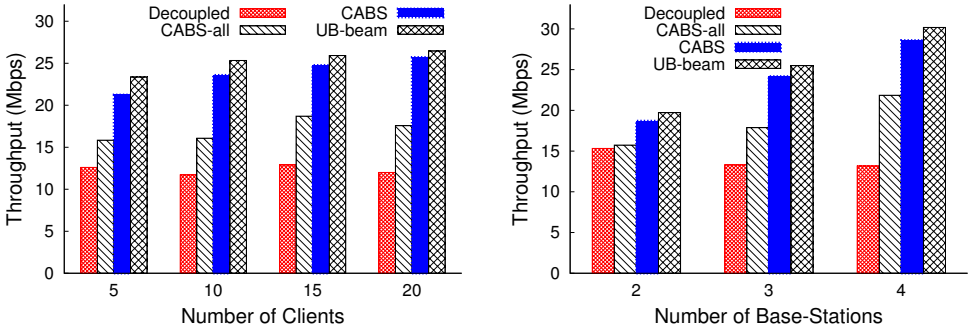
Evaluation metrics: We consider the following metrics.

- *Throughput:* Aggregate throughput of all clients in the network.
- *Utility:* Captures both throughput and fairness; aggregate utility of all clients: $\sum_{j \in K} w_j \log(T_{c_i})$ if $T_{c_i} > 0$, where C is the list of the clients and T_{c_i} is the throughput for client c_i.
- *Fraction of scheduled clients:* Captures the number of clients not scheduled in an epoch to improve spatial reuse (in CABS and upper bounds).
- *Load balancing factor:* Measures Jain's fairness index among the number of clients associated with each BS.

Throughput

Figure 4.9a presents the throughput results as a function of number of clients in the network. Three observations can be made: (i) CABS' performance is within 96% of that of exhaustive beam search and is not impacted by client density. Given the complexity of the latter, CABS provides a fine balance between performance and complexity. (ii) The increased spatial reuse from jointly addressing client association with beamforming (CABS-all) provides gains as high as 50% (over the decoupled approach). Further, the gains are more pronounced at higher client density, where it becomes harder to isolate interference between small cells without a joint optimization that allows for flexible client association. (iii) Interestingly, by going one step further and allowing some clients from not being scheduled in a given epoch provides CABS with an additional 50% gain over CABS-all, resulting in *a net gain of around 115% over the decoupled approach. Removing even a small fraction of bottleneck clients from scheduling in an epoch can greatly improve the spatial reuse configuration between small cells*.

The impact of interference from increased number of BSs is presented in Figure 4.9b. As the number of BS increases, the interference in the network gets severe therefore the throughput improvement from additional BS does



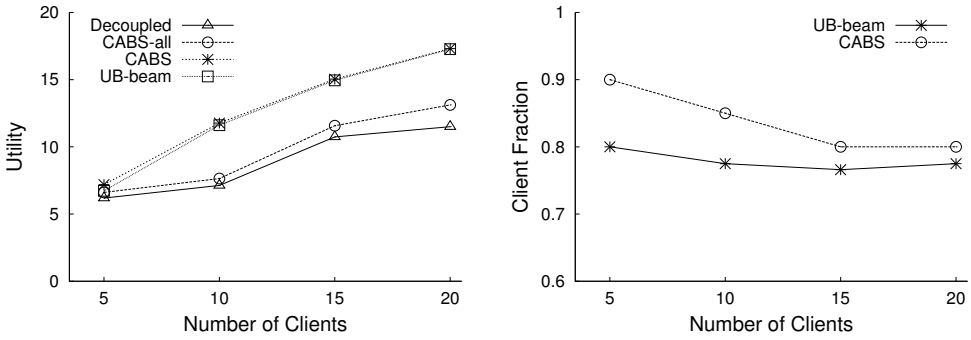
(a) Throughput *w.r.t* the number of clients (b) Throughput *w.r.t* the number of BSs

Figure 4.9: Experimental evaluation of ProBeam with 4 small cells: (a) CABS provides almost 96% of performance with very low computational complexity comparing to upper bound. (b) Even in the severe interference (four BSs case), CABS alleviates the interference efficiently hence it provides significant throughput improvement (115%).

not follow linear increment. In spite of the increased interference, the ability to jointly address client association with beam selection helps CABS handle interference effectively, the benefits of which are more pronounced with larger number of interferers.

Utility and Fairness

Recall that some of the reuse gains in CABS comes from removing a subset of clients from scheduling in a given epoch. While starvation of such clients is avoided across epochs, it is important to understand if the throughput gains of CABS are not realized at the expense of fairness even within an epoch. The utility measure helps account for fairness within an epoch, whose results are presented in Figure 4.10a. It can be clearly seen that CABS' utility is very close to that of its upper bound and outperforms that of the (baseline) decoupled approach. Thus, *adopting a utility based approach to joint CABS, enables ProBeam to bypass some clients from an epoch to maximize reuse gains without compromising on fairness*.



(a) System utility (b) Scheduled client fraction

Figure 4.10: Effective client management in ProBeam: (a) CABS achieves high utilization due to the efficient scheduling. (b) The throughput improvement of the UB-beam comes from the less number of scheduled clients.

Note that if the number of clients bypassed is large, this would automatically reflect in a reduced system utility. Hence, to further verify this, we present the fraction of scheduled clients in an epoch in Figure 4.10b. This clearly shows that only a small fraction of clients (10-20%) are bypassed in CABS. The upper bound is more aggressive in deferring clients to the next scheduling epoch, which in turn contributes to its marginal throughput gains over CABS (Figure 4.9a), however, the gains are minimal.

Load balancing

A by-product of *utility maximization in CABS is that it should automatically lead to load balancing*. This is because, given a fixed amount of frame resources, balancing number of users across cells, provides more resources per user and hence better aggregate utility. The load balancing factor, captured through Jain's fairness index between number of clients associated with small cells, is presented in Figure 4.11. CABS provides very good load balancing as expected. The decoupled approach does not implicitly account for load balancing, but a uniform distribution of clients automatically provides reasonable load balancing, when SNR-based client association is employed. The interesting observation is that

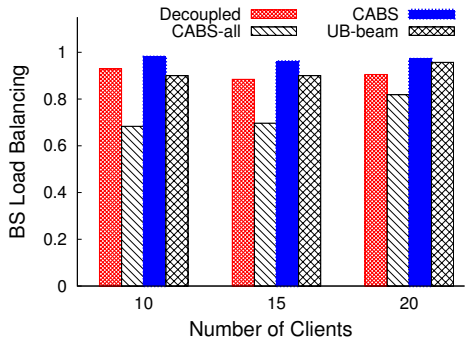


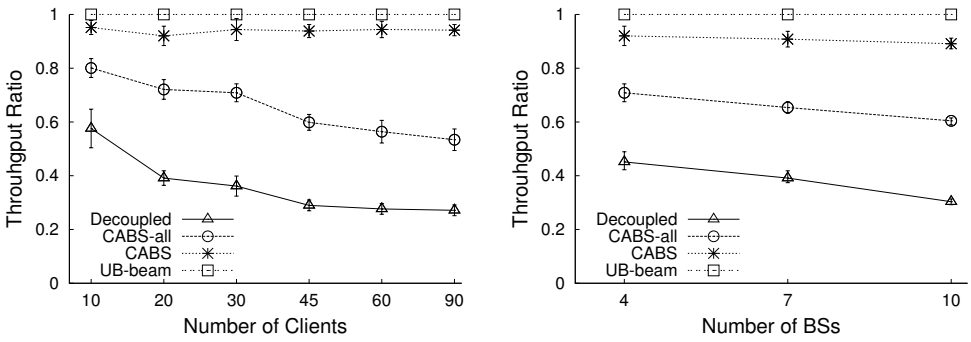
Figure 4.11: Network load balancing: CABS balances network load most efficiently.

CABS-all's load balancing suffers, especially when the number of clients is not high. Recall that CABS-all's throughput gain (over the decoupled approach) from better interference suppression (and hence reuse) through flexible association, comes at the expense of potential load imbalance across cells, especially when all clients are accommodated.

Trace-driven Simulations

Our experimental set-up with few tens of clients and three dominant interferers constitutes a realistic set-up for a cluster of small cells. However, to further understand CABS's effectiveness in much denser deployments (10 BSs and 90 clients), we resort to trace based simulations. We collect SNR traces for clients from our experimental network, feed it into a simulator running ProBeam (SINR estimation and CABS) to evaluate the various algorithms. We place our four BSs in various other locations to emulate more small cell BSs and measure SNR traces at the clients from them on all beams. Similarly, we also vary the client locations to emulate a larger set of clients and obtain corresponding SNR traces. Given the traces, we can generate a topology with a specific number of BSs and clients, by sampling BSs and clients randomly from our SNR trace database.

Our simulation results are presented in Figure 4.12, where throughput is measured as a fraction of that achieved by the upper bound (UB-beam). The

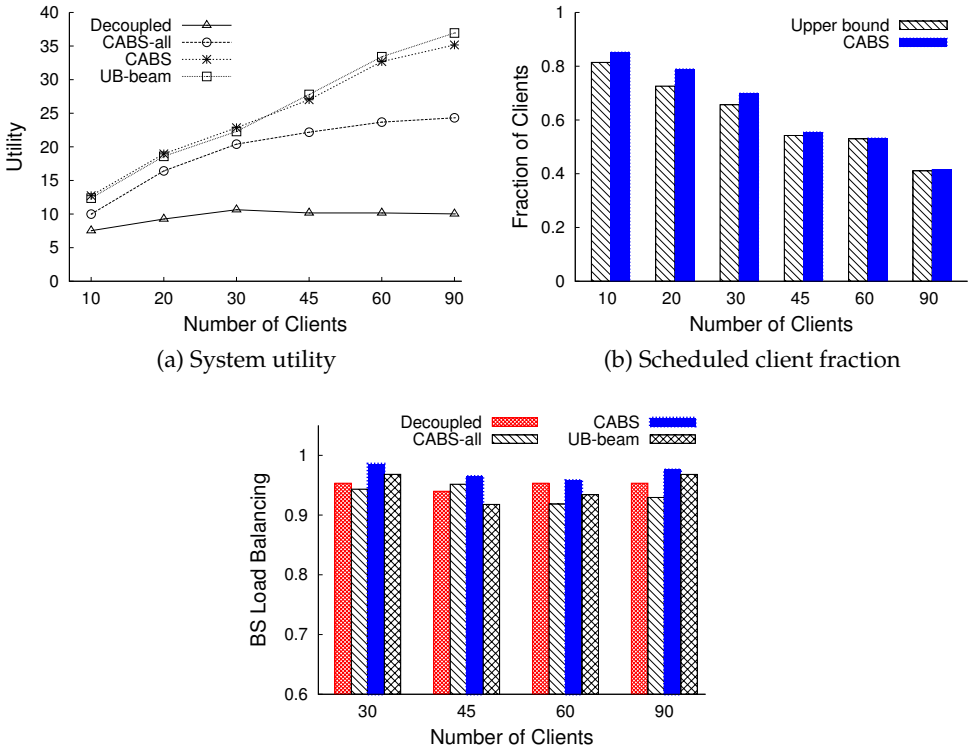


(a) Throughput ratio *w.r.t.* the number of clients (b) Throughput ratio *w.r.t.* the number of BSs

Figure 4.12: Large scale evaluation of ProBeam through trace-driven simulations: CABS yields almost 96% of throughput performance compare to upper bound.

trends in these large scale results, including the magnitude of gains possible with CABS, are very similar to those from the experiments, thereby reinforcing our inferences from the prototype evaluation. For example, the CABS yields almost 96% of throughput of what UB-beam provides and the gain is consistent regardless of number of clients or BSs in the network. We also present the fraction of scheduled clients for CABS and UB-beam in Figure 4.13b. As the network density increases more clients are bypassed for next scheduling epoch because it is hard to schedule the clients who experience interference severely. Therefore, the UB-beam approach aggressively leaves out more clients to the next epoch to achieve higher system utility. CABS schedules more clients than the UB-beam meanwhile providing similar utility (Figure 4.13a). In addition, CABS balances network load most efficiently even in a dense deployment as we see in Figure 4.13c. CABS close performance with respect to its upper bound in these results indicates the efficiency of its beam selection component as both the schemes employ the same client association mechanism.

Given the hardness of computing a tight upper bound for the joint CABS solution, we now evaluate the efficiency of its client association component as well. We compare it against an upper bound for client association (UB-assoc)



(a) System utility

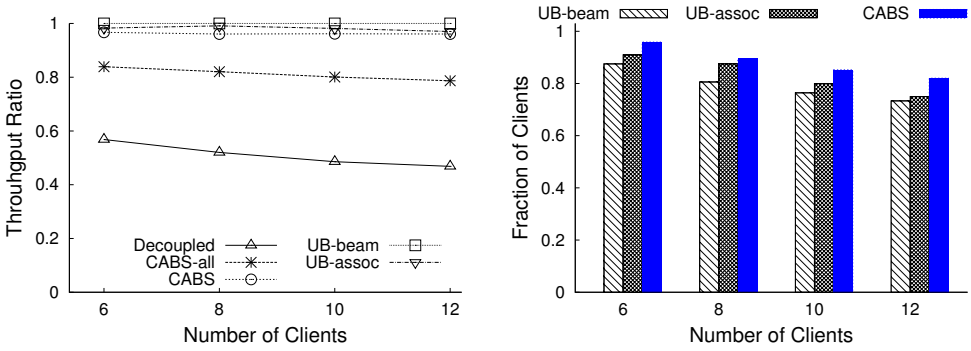
(b) Scheduled client fraction

(c) Network load balancing

Figure 4.13: Large scale evaluation of ProBeam through trace-driven simulations: (a) CABS provides consistent utility regardless of the number of clients. (b) Small throughput improvement of the upper bound comes from the less number of scheduled clients. (c) CABS balances network load most efficiently.

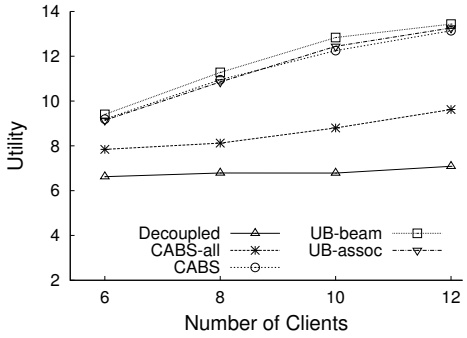
that exhaustively searches over all possible client associations, while employing the same beam selection mechanism as in CABS. The results in Figure 4.14 indicate that, while the sub-optimality of CABS' client association component can at most be within half of the optimal (see Section 4.4) in the worst case, in practice, it yields a performance that is very close to its upper bound (i.e., UB-beam and UB-assoc). Thus, *the high efficiency of the individual components in CABS in turn synergistically contribute to the net gains seen by it*.

Simulation results show that CABS handles the interference effectively,



(a) Throughput ratio

(b) Scheduled client fraction



(c) System utility

Figure 4.14: Evaluation of client association component in ProBeam.

and hence it can provide steady performance regardless of the level of the interference in dense deployments (i.e., number of clients, number of BSs).

4.6 SUMMARY OF PROBEAM

In this chapter, we explored resource management in spatial domain. Interference management through beamforming allows for increased reuse without having to sacrifice resources in the time or frequency domain. We design and implement ProBeam, a practical system for improving spatial reuse through beamforming in OFDMA based small cell networks. We show that decou-

pling beamforming from client scheduling is necessary for practical feasibility. Further, we highlight the need to jointly address client association with beamforming to maximize the reuse benefits from the latter. ProBeam incorporates a low complexity, highly accurate SINR estimation module with less than 1 dB error ($\leqslant 5\%$) to determine interference dependencies between small cells. Prototype implementation in a real WiMAX networks of four small cells shows 115% of capacity gain compared to other baseline reuse schemes. We also demonstrate the scalability and efficacy of our system in larger scale settings through simulations.

5 RELATED WORK

In this chapter, we discuss related works. We partition them into three parts. First, we discuss several studies related to OFDMA multi-cell deployments and operations. In this section, we describe why prior resource management solutions in the WiFi domain are not directly applicable in OFDMA multi-cell systems. We then describe the uniqueness of our solution for multi-cell OFDMA networks. Second, we focus on various video multicast solutions in wireless networks and contrast our solution, MuVi, with these other solutions. Last, we introduce several researches leveraging directional antenna and beamforming to improve the performance of wireless networks. We then highlight how our solution, ProBeam, is different from existing problems and solutions in wireless networks.

5.1 RESOURCE MANAGEMENT IN OFDMA SYSTEM

In this section, we discuss prior research focusing on resource management in OFDMA systems in the context of both multi-cell and single-cell networks. The goal of resource management in multi-cell network is two fold: 1) maximizing resource reuse in the network and 2) mitigating the interference between the neighboring femtocells. Several approaches for resource management are proposed in WiFi network, however, their solutions can not be directly adopted in OFDMA systems. In the following subsections, we discuss why different approaches are needed in OFDMA networks.

Interference Mitigation in OFDMA Multi-cell Network

There are several research works [20, 76] that focus on operating OFDMA multi femtocell networks (e.g., WiMAX, LTE and LTE-A). Several efforts that focus on the macro-femtocells interference and the interference to cell-edge users in OFDMA macrocells exist [99, 104]. The localized (cell edge) interference and planned cell layouts have aided various interference management solutions [75] for both downlink ([30, 65]) and uplink ([60, 93]). These also include

fractional frequency partitioning (FFR) approaches [35, 68], where the spectrum is partitioned into pre-determined static sets that require different levels of coordination between the macrocells.

Femtocells are deployed in an unplanned manner without coordinated operations and hence, vulnerable to interference [27, 64]. This necessitates new design of interference mitigation solutions that incorporate dynamic resource partitioning strategies. There have been recent researches in this direction but are restricted to theory with several simplifying assumptions that restrict their scope and prevent their adoption in practice [87, 95, 104]. Simply operating interfering femto BSs on fixed sized orthogonal bands will result in underutilization of the licensed spectrum. Instead, these systems have to use orthogonal fragments (resources) of the same central frequency to isolate their transmissions from each other. However, this limits the amount of spectrum used, therefore resource isolation in OFDMA must be carefully invoked only when needed. New system design for resource management framework has to be discussed in the context of OFDMA system.

Self-organizing (Distributed) Resource Management

In [24], FERMI, a *centralized* resource management solution for enterprise femtocells was proposed where centralized solution run by a central controller. For efficient resource management, central controller determines resource allocation for each cell after gathering interference map between femtocells. However, centralized approach is not appropriate for unplanned deployment settings (such as residences), where cooperation among femtocells is not realistic due to the complexity of collecting interference map and handling network dynamics (e.g., activating and de-activating of cells). There has been a research proposing distributed algorithm for resource selection of WLANs [61]. The main purpose of this work is to select the operating channels for interferences mitigation without coordinating among the wireless devices. Operating on non-overlapping channels can resolve interference between APs in WiFi system, however, in OFDMA system interfering femtocells can only operate on orthogonal subset of the spectrum [34]. Resource management in macrocell requires synchronous

operation and the interference is localized only at cell-edges. Unlike macrocells, interference is less predictable and more pervasive across femtocells due to their dense un-planned deployments. Different approaches are needed considering unique and specific environment of femtocell networks. We focus on the framework which enables resource management for OFDMA femtocell networks in a distributed manner.

Resource Allocation in OFDMA Single-cell Network

There has been growing interest from the WiFi community to move towards OFDMA. Recent efforts show the benefits of OFDMA by building practical systems that enable dynamic spectrum fragmentation [101, 103] and adaptive channel width sizing [33, 81]. While most of these works relate to the adaptive spectrum access capability of OFDMA, [97] explores 802.11 based random channel access with OFDMA in a single cell. The resource allocation problems in single-cell downlink OFDMA networks has been extensively studied [54, 65, 91, 100] mainly focusing on transmission power allocation, PHY transmission rate adaptation, resource optimization and etc. For instance, [91] proposes efficient rate and power allocation algorithms for OFDMA downlink systems where each tone is taken by at most one user. Huang *et al.* discuss efficient scheduling and resource allocation in OFDMA systems [54]. In [100], the goal is to minimize the total transmit power given target bit-rates for each user to optimize the resource utilization. A resource allocation problem in uplink was formulated and an iterative algorithm was proposed with relatively high complexity in [50]. Pfletschinger *et al.* proposed a heuristic algorithm that tries to minimize each user's transmission power while satisfying the individual rate constraints [86]. However, they are limited to theoretical works and none of them are practically implemented in a real testbed to show their efficacy. We proposed a practical efficient resource allocation solution which incorporates with video traffic. Wireless video multicast is a good example that requires efficient resource utilization to provide satisfactory services to multiple users in the multicast session. Since our resource allocation solution focuses on video multicast, we discuss the prior research efforts in efficient video multicast in

WiFi networks.

Inapplicability of WiFi Solutions in OFDMA System

There has been growing interest from the WiFi community to move towards OFDMA. Recent efforts showcase the benefits of OFDMA by building practical systems that enable dynamic spectrum fragmentation [101, 103], adaptive channel width sizing [33, 81], and etc. While most of these works relate to the adaptive spectrum access capability of OFDMA, [97] explores 802.11 based random channel access with OFDMA in a single cell. Several previous approaches addressing interference in WiFi [31, 79] are not directly applicable in OFDMA system. Since WiFi is based on asynchronous channel access, the main focus in all these efforts is enabling synchronization and alignment between the transmitter and receiver pair (crucial for OFDMA operations). In WiFi, each interfering AP is tuned to a different channel typically using graph coloring approaches [78]. APs use one among multiple 20 MHz channels (frequency chunks) and several conventional distributed channel selection algorithms [61] can be used to configure APs on different channels to avoid interference. However, in OFDMA femtocell networks, the entire spectral chunk (say 20MHz) is available to all the cells where synchronous channel access allows for tight synchronization between the transmitter and receiver is a part of standard operations. While these synchronization benefits are carried over to femtocells, this also limits their ability to sense and manage interference effectively, thereby making resource management even more challenging. Moreover, femtocells need to operate on mutually orthogonal subsets of frame resources (sub-channels and time symbols) to avoid interference. Thus, resource allocation has to adapt to network dynamics (such as traffic, load etc.).

5.2 VIDEO MULTICAST IN WIRELESS

In this section, we discuss some of the related work in the area of video multicast in wireless network. Given the high volume of video streaming traffic, a large number of prior research has focused on improving the performance of video multicast by enhancing MAC/PHY layer design. MuVi is the first piece of work

that tried to design and implement video multicast system in WiMAX network. Prior to MuVi, several scalable video multicast solutions in WiMAX focused on theoretical analysis, thus it prevents their deployments in real system. In the following subsections, we discuss related approaches and contrast our contributions with prior works.

Video Multicast

There are many prior works on wireless multicast that aim to improve the performance through PHY and MAC layer design. One of the approach to enhance the quality of video streaming involves modifying the coding rate based on the wireless channel condition. DirCast [32] applies association control to minimize the total multicast delay, thus within each access point (AP) DirCast chooses the transmission rate based on the channel condition of the "worst" client in the multicast group. However, such decision limits the performance of other clients whose channel capacity is sufficient enough to support higher transmission rates. On the other hand, MuVi does not limit the performance based on the "worst" client in the system and provides the differentiated service to all multicast users.

A recent work Medusa [90] employs a proxy-based solution to improve media streaming performance in wireless LAN. Medusa focuses on a content-dependent PHY rate selection and packet value awareness for WiFi multicast. However, Medusa is designed for asynchronous, WiFi systems where the wireless media is shared via contention-based random channel access. As a result, Medusa does not perform radio resource allocations and only employs a heuristic-based rate adaptation algorithm. Moreover, Medusa requires that clients periodically send reception reports about the packets transmitted previously and performs retransmissions using network coding, and thus it requires modifications at the client side, which are challenging in cellular networks (because it involves modifications of the wireless standards).

In contrast, MuVi is designed for OFDMA-based, synchronous broadband mobile wireless systems (e.g., WiMAX, LTE) where the radio resources are allocated by the base-stations. As a result, MuVi employs a near-optimal re-

source allocation algorithm to maximize the total system utility by intelligent resource allocation and PHY rate adaptation of each video packet. MuVi does not require packet reception reports or perform re-transmissions. Instead, MuVi incorporates CINR reports that are available in current cellular networks to adjust the OFDMA resource allocation and PHY rate adaptation. Thus it does not require modification at the client-side or the air interface, and has minimum control overhead.

Scalable Video Multicast

Scalable video multicast has been extensively studied in both wired [77, 98] and wireless networks [36, 53]. In [77], receiver-driven multicast approach is proposed which receivers decide set of multicast groups to join based on available bandwidth. In [42], Deb *et al.* studied the problem of multicasting scalable video (SVC) in WiMAX cellular networks with the goal of maximizing the system utility. They developed a greedy algorithm to allocate radio resources and adaptively assign the MCS for each transmitted video layer and allocating resources. In [69], the authors use dynamic programming approach to find the optimal system utility and to assign modulation and coding schemes for scalable video traffic in mobile cellular networks. In [70], Li *et al.* considered joint-layer resource allocation to further improve multicast performance and developed approximation algorithms to trade off algorithm complexity with performance. All these works were only evaluated through theoretical analysis and simulations based on fixed layering sizes. The dynamics of video traffic across different video frames were not considered and no system implementation has been conducted.

Channel-unaware Wireless Video Transmission

A couple of recent works proposed wireless video encoding and transmission schemes that need not be aware of the wireless channel conditions. SoftCast [58] proposed a joint channel encoding and video source coding scheme for mobile video transmission. SoftCast takes an analog approach for delivering video over the wireless. Specifically, it represents data in a numerical (analog)

format and converts it to symbols to transmit with certain tx power. There are several differences between SoftCast and MuVi; (i) SoftCast uses customized content format that is different from what is used in popular standards (e.g., H.264). On the other hand, MuVi can utilize any format while differentiating them with its dependency value. (ii) SoftCast does not incorporate feedback from the clients. In contrast, we show that MuVi takes simple feedback, SINR, to provide differentiated services to all clients in multicast group corresponding to their wireless channel conditions.

FlexCast [16] modified the MPEG4 video codec and incorporated rateless coding for efficient video streaming in wireless systems. Neither SoftCast nor FlexCast requires any feedback about the wireless channel conditions. The received video quality automatically adjusts depending on the channel quality at the clients. As a result, these two schemes provide natural support for wireless video multicast although they are not specifically designed for it. Nevertheless, both SoftCast and FlexCast require heavy modifications to video source coding, the air interface, and the mobile clients. By contrast, MuVi does not require any changes in these elements. MuVi requires long-term channel feedback and optimizes the video multicast transmission via efficient radio resource allocation and frame prioritization under the existing video and wireless standards. Thus it allows speedy deployment.

5.3 MULTI-CELL BEAMFORMING

In multi-cell OFDMA-based network, interference has been shown to be a key performance limiting factor across small cells [23]. This necessitates interference mitigation solutions that incorporate dynamic resource partitioning strategies. In Chapter 3, we have proposed a distributed resource management framework for mitigating interference and maximizing resource reuse. The authors in [24] propose centralized resource management scheme respectively for interference mitigation and demonstrate their efficacy in practice. Above-mentioned two solutions allocate orthogonal resources to interfering small cells to avoid interference while reusing resources for the clients that do not incur interference. However, such resource isolation either in time or frequency comes at the cost of

sacrificing resources, which in turn can be avoided by addressing interference in the spatial domain through beamforming.

Using Directional Antenna in Wireless Network

Directional antennas have been widely used in wireless network for many purpose. Several works adopt directional antennas to extend TX/RX range of wireless transmission both in indoor and outdoor wireless network [67, 84]. It has been primarily used for improving system performance (e.g., throughput) while providing higher connectivity (higher SINR) for vehicular in wireless networks [83, 88]. Researches in [37, 96] proposed various MAC protocol designs for directional wireless Ad-hoc network and showed performance gains. Authors in [94] proposed video multicast scheduling algorithm leveraging directional antenna and [106] presented directional transmission and reception algorithm for supporting QoS in wireless network.

Beamforming and Client Scheduling

In the space of beamforming and client scheduling, Dimic *et al.* [43] considered joint problem of downlink beamforming and client association. The main difference from ProBeam is that the algorithm adopt greedy approach thus it could not guarantee the fairness among users, whereas ProBeam took proportional fairness. Authors in [22] designed and implemented beamformer algorithm in multi femtocell network, but the solution is limited to uplink scheduling. Dirc [73] and Speed [74] proposed to increase the capacity of WLANs through spatial reuse by considering directional antennas only at the APs or at both APs and clients. However, client association is assumed and conflicts and reuse schedule are computed with respect to a single client at each AP. This limits the practical applicability of such solutions (especially for OFDMA systems) since conflicts and reuse schedules have to be recomputed (potentially every packet) every time the client scheduled with any of the AP changes. Several theoretical works [41, 51] have looked at adaptive beamforming in a multi-cell context. However, idealized settings are assumed that require fine grained CSI from all transmitters to all clients be made available to the reuse algorithm at

every frame interval. Given the practical feasibility (or lack thereof) of such approaches, experimental works [25] have appropriately focused on adaptive beamforming for SNR improvements within a single cell. Further, none of these works address client association jointly with beamforming.

Resource Allocation and Client Association

There are several recent works on resource allocation and client scheduling problems in OFDMA networks. The authors in [17] investigated the problem of joint user-scheduling and resource allocation under channel uncertainty in downlink OFDMA systems. The authors in [52] considered the problem of resource allocation and network optimization in LTE networks. In particular, they proposed a distributed protocol to achieve weighted proportional fairness among clients under various system models. Similarly in ProBeam, we adopt proportional fairness in our utility function while achieving fairness among clients and maximizing system utility. [29] proposed a distributed optimization approach for jointly allocating power and assigning users to cells in OFDMA cellular network. In contrast, ProBeam is a centralized solution to jointly optimize the client associations and beam selections. Our goal is to mitigate interference purely in the spatial domain without having to sacrifice time/frequency/power resources in each cell.

The focus of ProBeam is to design a *practical* multi-cell spatial reuse system that, decouples client scheduling from beamforming, employs switched beamforming for interference management between small cells, and jointly addresses client association to increase the potential of spatial reuse from beamforming. Being complementary, adaptive beamforming can still be leveraged for SNR improvement within each small cell.

6 CONCLUSION AND FUTURE WORK

In this thesis, we have explored various resource management solutions to efficiently use scarce spectrum resources in OFDMA cellular networks. Considering the two-tier network architecture (marco-cell and femto-cell) widely used in 4G cellular systems, we have taken different approaches for designing resource management schemes. We have also considered various angles (e.g., time, frequency, space domains) to maximize spectrum reuse without causing interference. We have implemented several resource management and utilization schemes in WiMAX multi-cell testbed and shown their efficacy. In addition, we have shown a practical video multicast algorithm in 4G cellular networks. Our video delivery solutions provide good service quality when compared to other existing solutions, especially when the spectrum resources are not sufficient to transmit original video without losses. Last, our findings and inferences in operating multi-cells in OFDMA networks could be a good starting point for designing resource utilization solutions in next-generation cellular networks which are also based on multi-cell architecture for maximizing resources.

In the next section, we summarize the main contributions of this thesis, and then present several problems for future research endeavors.

6.1 CONTRIBUTIONS

Practical Video Multicasting in 4G Cellular Network

Although wireless broadband technologies have evolved significantly over the past decade, they are still insufficient to support the fast-growing mobile traffic, especially due to the increasing popularity of mobile video applications. Wireless multicast, aiming to exploit the wireless broadcast advantage, is a viable approach to bridge the gap between the limited wireless capacity and the ever-increasing mobile video traffic demand. Several video multicast solutions are limited their usage in theory and hence, hinder their deployments. Toward this, we developed a practical video multicast scheme, MuVi, in 4G cellular

network. We first identifies the priority of video packets and clients' wireless link conditions to schedule. We then design an algorithm for both allocating resource and selecting MCS for efficiently scheduling multiple of heterogenous clients in the same WiMAX resource frame. MuVi is the first practical video multicasting solution running on WiMAX testbed and is a lightweight solution with most of the implementation in the gateway along with slight modification in the base-station.

Our experimental results show that MuVi improves the average video PSNR by up to 13 dB and 7 dB compared to the Naive and the Adaptive schemes, respectively. MuVi incorporates the clients' channel conditions in real-time therefore, it provides the best performance even for mobile clients comparing to the other solutions. In addition, MuVi's optimization enables guaranteed video delivery within deadline while having very minimal inter-packet delay. MuVi does not require modification to the video encoding scheme or the air interface, thus it allows speedy deployment in existing systems.

Self-organizing Resource Management Framework in Multi Small-cell Network

4G cellular networks provide higher bandwidth and spectrum efficiency leveraging smaller cells with orthogonal frequency division multiple access (OFDMA). The uncoordinated, dense deployments of small-cells however, pose several unique challenges relating to interference and resource management in OFDMA femtocell networks. We first identify the impact of interference in operating multi small-cells while reusing same spectrum resource and then design several building blocks to address them. Based on our inferences, we design and implement RADION, a distributed resource management framework running on real WiMAX testbed, consisting of three femtocells in an unplanned indoor environment.

RADION's core building blocks enable small-cells to opportunistically determine the available resources in a completely distributed and efficient manner. Further, RADION's modular nature paves the way for different resource management solutions to be incorporated in the framework. Two distributed solutions are enabled through RADION and their performance is studied to

highlight their quick self-organization into efficient resource allocations. To the best of our knowledge, this is the first implementation based design and evaluation of a self-organizing framework for OFDMA small-cells. We wish to point out that RADION is modular. In particular, depending on the context and the objectives, different resource allocation solutions can be easily incorporated within the RADION framework.

Multicell Beamforming System for OFDMA Small-cell Network

Small-cells form a critical component of next generation cellular networks, where spatial reuse is the key to higher spectral efficiencies. Interference management in the spatial domain through beamforming allows for increased reuse without having to sacrifice resources in the time or frequency domain. Existing beamforming techniques for spatial reuse, that are coupled with client scheduling, face a key limitation in practical realization with OFDMA small-cell network. In this context, we argue that it is important to decouple beamforming from client scheduling for a practical beamforming system by showing that jointly addressing client association with beamforming is critical to maximizing the resource reuse. We propose and implement ProBeam, a system for multi-cell beamforming and client association in OFDMA small-cell networks.

ProBeam incorporates two key components - a low complexity, highly accurate SINR estimation module that helps determine interference dependencies for beamforming between small-cells; and an efficient, low complexity joint client association and beam selection algorithm. We highlight that ProBeam's SINR estimation module requires only linear number of measurements with an estimation error less than 1 dB for 95% confidence. In addition, our flexible client association does not follow conventional approach for scheduling client in multi APs or small-cell deployments and we show its efficacy. We have prototyped ProBeam on a WiMAX network of four small cells and evaluations reveal the reuse gains from joint client association and beamforming to be as high as 115% over baseline strategy.

6.2 FUTURE WORK

In this thesis, we found several design choices and inferences for building efficient resource utilization solutions in wireless networks. Specifically, we designed various resource management solutions in multi-cell networks exploring diverse angles in time, frequency, and space domains in OFDMA-based cellular networks. We believe this dissertation was successful in developing some of the important building blocks to improve the performance of multi-cell networks. These contributions can be carried out for other resource management schemes in next-generation cellular networks (e.g., 5G) that try to improve the spectral efficiency in multi-domains. We now describe some of the potential problems for future research in next-generation cellular network.

Efficient Operations in Next-generation Cellular Networks

One of the essential requirements for the next-generation cellular networks (5G) is a higher data rate, a requirement that stems from supporting the mobile data traffic explosion. For instance, it is thought that 5G needs to support a 1000$\times$ higher amount of data compared to 4G. To provide higher capacity, 5G cellular networks not only adapt increased bandwidth by moving toward the mmWave spectrum but also by making better use of WiFi's unlicensed spectrum in the 5 GHz band [21]. The next-generation cellular networks will also need to be able to support a large number of users and a diverse set of devices. To meet this purpose, they shrink cell size to improve the area spectral efficiency and ensure reduction in the number of users that have to share the resources at each cell. This will impact the cellular base-station deployments; as the number of users increases, the density of the cellular base-station is also increasing. For operating multi small-cells, tight coordination across multi-cells is essential. That coordination includes handling client mobility across the small-cells, efficiently managing resources, and mitigating interference across neighboring cells. The *anchor-booster* architecture is proposed as a means to coordinate between macro and small-cells; while the macro cell operates as an *anchor* base station to coordinate the small-cells, the small-cells operate

as a *booster* to offload data traffic for increasing spatial reuse [26]. This is a similar architecture to the one that we used for ProBeam's centralized beam selection and client association approaches. In this context, our inferences and resource management solutions in 4G cellular networks are good fundamentals for operating multiple small-cells in 5G cellular networks in order to increase spatial spectrum reuse.

In addition to the enhancement solutions in the MAC and Network layers, further improvement can be achieved with PHY layer mechanisms. Given the limited usage of channel width, an even higher capacity can be obtained by employing massive MIMO (multiple-input and multiple-output) arrays at the base-station. MIMO systems can create multiples of concurrent signals between the transmitter and the receiver. For instance, MU-MIMO (multi-user MIMO) enables the base-station to concurrently communicate with several users and hence, offers an increased multiplex gain. MU-MIMO has been included in the 3GPP (third generation partner-ship project) LTE-Advanced standard, however, its full potential remains untouched. The performance improvement can be obtained through fast and accurate feedback of channel state information (CSI). MIMO systems require a knowledge of CSI at the transmitter and the receiver. In two-tier architecture, *macro and femto*, the macrocell supports high user mobility and it is very challenging to quickly gather accurate feedback for all mobile users. In contrast, in femto deployment (e.g., indoors, enterprises and residential areas), it is possible to obtain timely and accurate CSI due to the limited and low mobility of users. In this sense, small-cells have great potential that we can explore for our future work in improving MIMO performance by designing even more dedicated mechanisms.

Besides the capacity improvement, the metric for mobile user's performance shifts to quality of experience (QoE). This includes supporting continuous connectivity, easy access, energy efficiency, context-aware user experience, and etc. Consider the following video streaming example where two clients intend to watch video that is encoded at two video bitrates, 400 Kbps and 1400 Kbps, while sharing a network bandwidth of 2000 Kbps. The fair bandwidth share would assign 1000 Kbps to each user and both clients would end up with a selecting bitrate of 400 Kbps for streaming video. In contrast, *context-*

aware coordination could have an asymmetric assignment, 500 Kbps and 1500 Kbps, and hence one of the users could enjoy a higher bitrate (1400 Kbps video streaming instead of 400 Kbps). Context-aware coordination could improve the QoE in such a manner when the network bandwidth is not sufficient to provide equally good QoE to all users. In MuVi, we proposed a *channel-aware* adaptation that instantaneously takes the SNR feedback from the clients to adapts the MCS. We believe that our channel-aware solution presents intriguing possibilities for providing a better QoE to the users for various applications in next-generation cellular networks.

Centralized resource management and beamforming in cloud-based radio access networks

Radio access units (RAUs) are tied with baseband units (BBUs, cellular base stations) in cellular networks. The BBUs mainly process resource scheduling and transmit signal to multiple users (clients) through the RAUs. Small cell BBUs are deployed rapidly to meet the increased demand for resources, however, it would lead to significant interference in the neighboring small cells if they were not managed properly. Even worse, operating many BBUs overburdens the network operators in terms of cost and resources. The mobility of clients in the network makes it hard to configure network load balancing when multi-cells are operating on orthogonal resources. Cloud-based radio access networks (C-RAN) of small cell networks can be employed to address the problem caused by the increased number of small cells (BBUs).

The key idea of C-RAN is to decouple the RAUs from the BBU processing by using radio-over-fiber technology. By separating the RAUs from the BBU processing and integrating the BBUs to a centralized entity, C-RAN enables the RAUs to be deployed on a large scales for small cells. Meanwhile the centralized BBU processing can efficiently handle the interference amongst multi-cells and increase the capacity through spectral reuse. Many network operators (e.g., China Mobile) and service providers (e.g., Alcatel-Lucent, Nokia Siemens) propose C-RAN architecture and show its efficacy [2, 4]. Bhaumik *e*t al. proposed centralized architecture that replaces base stations with antennas

and a few other active RF components; the remainder of the digital processing is carried out in a central server [28]. This research, however, was limited to simulation-based theoretical work. We would like to build a centralized system for efficient resource management in a real C-RAN testbed. Since one-to-one or many-to-one logical mapping may be feasible between RAUs and BBUs, we plan to incorporate an optimal multi-cell beamforming algorithm in RAUs. These two components are expected to enhance the performance in C-RAN.

[1] Accton Wireless Broadband. *http://www.awbnetworks.com/*.

[2] Alcatel Lucent, Light radio network: a new wireless experience, white paper, 2012. *http://www2.alcatel-lucent.com/enrich/en/v6i1/lightradio-network-a-new-wireless-experience/*.

[3] Beceem wimax chips for mobile applications. *http://www.beceem.com/*.

[4] China mobile, C-RAN: the road towards green ran, white paper 2012. *http://labs.chinamobile.com/cran/*.

[5] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017. *http://www.cisco.com/*.

[6] Fidelity Comtech. *http://www.fidelity-comtech.com/*.

[7] HLS, IETF Internet-drafts 2014. *http://tools.ietf.org/html/draft-pantos-http-live-streaming-13/*.

[8] IEEE 802.16e-2005 Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems. *IEEE 802.16e standard*.

[9] MPEG4. *http://mpeg.chiariglione.org/standards/mpeg-4/mpeg-4.htm*.

[10] PicoChip Femtocell Solutions. *http://www.picochip.com/*.

[11] Raspberry Pi, single board computer. *http://www.raspberrypi.org/*.

[12] Smartphone shipments surpass PCs for first time, what's next, Feb. 2011. *http://www.pcmag.com/article2/0,2817,2379665,00.asp/*.

[13] TeraSync GPS solutions. *http://www.terasync.net/*.

[14] VLC media player. *http://www.videolan.org/vlc/*.

[15] WiMAX forum tech report: Architecture, detailed Protocols and Procedures Self-Organizing Networks. *WMF-T33-120-R016v01*.

[16] Aditya, S. T., and S. Katti. 2011. FlexCast: Graceful Wireless Video Streaming. In *Proceedings of ACM MobiCom*.

[17] Aggarwal, R., M. Assaad, C. E. Koksal, and P. Schniter. 2011. Joint Scheduling and Resource Allocation in the OFDMA Downlink: Utility Maximization under Imperfect Channel-State Information. *IEEE Transactions on Signal Process* 59(11):5589–5604.

[18] Akhshabi, S., L. Anantakrishnan, C. Dovrolis, and A. C. Begen. 2012. What Happens when HTTP Adaptive Streaming Players Compete for Bandwidth? In *Proceedings of ACM NOSSDAV*.

[19] Akhshabi, S., A. C. Begen, and C. Dovrolis. 2011. An Experimental Evaluation of Rate-Adaptation Algorithms in Adaptive Streaming over HTTP. In *Proceedings of ACM MMSys*.

[20] Akyildiz, I. F., D. M. Gutierrez-Estevez, and E. C. Reyes. Dec. 2010. The Evolution to 4G Cellular Systems: LTE-Advanced. *Physical Communication* 3(4):217–244.

[21] Andrews, J. G., S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang. Jun. 2014. What Will 5G Be? *IEEE Journal on Selected Areas in Communications* 32(6):1065–1082.

[22] Arslan, M., K. Sundaresan, S. Krishnamurthy, and S. Rangarajan. 2012. Design and Implementation of an Integrated Beamformer and Uplink Scheduler for OFDMA Femtocells. In *Proceedings of ACM MobiHoc*.

[23] Arslan, M., J. Yoon, K. Sundaresan, S. V. Krishnamurthy, and S. Banerjee. 2012. Experimental Characterization of Interference in OFDMA Femtocell Networks. In *Proceedings of IEEE Infocom*.

[24] Arslan, M. Y., J. Yoon, K. Sundaresan, S. V. Krishnamurthy, and S. Banerjee. 2011. FERMI: A FEmtocell Resource Management System for Interference Mitigation in OFDMA Networks. In *Proceedings of ACM MobiCom*.

[25] Aryafar, E., A. Khojastepour, K. Sundaresan, S. Rangarajan, and E. Knightly. 2013. ADAM: An Adaptive Beamforming System for Multicasting in Wireless LANs. *IEEE/ACM Transactions on Networking* 21(5): 1595–1608.

[26] Bangerter, B., S. Talwar, R. Arefi, and K. Stewart. Feb. 2014. Networks and Devices for the 5G Era. *IEEE Communications Magazine* 52(2):90–96.

[27] Barbieri, A., A. Damnjanovic, T. Ji, J. Montojo, Y. Wei, D. Malladi, O. Song, and G. Horn. Apr. 2012. LTE Femtocells: System Design and Performance Analysis. *IEEE Journal on Selected Areas in Communications* 30(3):586–594.

[28] Bhaumik, S., S. Chandrabose, M. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo. 2012. CloudIQ: A Framework for Processing Base Stations in a Data Center. In *Proceedings of ACM MobiCom*.

[29] Borst, S., M. Markakisy, and I. Saniee. 2011. Distributed Power Allocation and User Assignment in OFDMA Cellular Networks. In *Proceedings of IEEE Allerton Conference on Communication, Control, and Computing*.

[30] Boudreau, G., J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic. 2009. Interference Coordination and Cancellation for 4G Networks. *IEEE Communications Magazine* 47(4):74–81.

[31] Broustis, I., K. Papagiannaki, S. V. Krishnamurthy, M. Faloutsos, and V. Mhatre. 2007. MDG: Measurement-Driven Guidelines for 802.11 WLAN Design. In *Proceedings of ACM MobiCom*.

[32] Chandra, R., S. Kandula, Moscibroda T, V. Navda, J. Padhye, R. Ramjee, and L. Ravindrananth. 2009. DirCast: A Practical and Efficient Wi-Fi Multicast System. In *Proceedings of IEEE ICNP*.

[33] Chandra, R., R. Mahajan, T. Moscibroda, R. Raghavendra, and P. Bahl. 2008. A Case for Adapting Channel Width in Wireless Networks. In *Proceedings of ACM SIGCOMM*.

[34] Chandrasekhar, V., and J. G. Andrews. Oct. 2009. Spectrum Allocation in Two-tier Networks. *IEEE Transactions on Communications* 57(10):3059–3068.

[35] Chang, R., Z. Tao, J. Zhang, and C. Kuo. 2009. A Graph Approach to Dynamic Fractional Frequency Reuse (FFR) in Multi-Cell OFDMA Networks. In *Proceedings of IEEE ICC*.

[36] Chi, H., C. Lin, Y. Chen, and C. Chen. 2008. Optimal Rate Allocation for Scalable Video Multicast over WiMAX. In *Proceedings of IEEE International Symposium on Circuits and Systems, ISCAS*.

[37] Choudhury, R. R., X. Yang, R. Ramanathan, and N. H. Vaidya. 2002. Using Directional Antennas for Medium Access Control in Ad Hoc Networks. In *Proceedings of ACM MobiCom*.

[38] Cisco Visual Networking Index. Feb. 2014. Global Mobile Data Traffic Forecast Update, 2013-2018.

[39] Clinch, S., J. Harkes, A. Friday, N. Davies, and M. Satyanarayanan. 2012. How Close is Close Enough? Understanding the Role of Cloudlets in Supporting Sisplay Appropriation by Mobile Users. In *Proceedings of IEEE PerCom*.

[40] Costa, J. M. 2007. More Frequencies Needed for Mobiles - Terrestrial Spectrum Sought for IMT. In *ITU News, No3*.

[41] Dahrouj, H., and W. Yu. 2010. Coordinated Beamforming for the Multicell Multi-Antenna Wireless System. *IEEE Transactions on Wireless Communications* 9(5):1748–1759.

[42] Deb, S., S. Jaiswal, and K. Nagaraj. 2008. Real-time Video Multicast in WiMAX Networks. In *Proceedings of IEEE Infocom*.

[43] Dimic, G., and N. Sidiropoulos. Sep. 2005. On Downlink Beamforming with Greedy User Selection: Performance Analysis and a Simple New Algorithm. *IEEE Transactions on Signal Processing* 53(10):3857–3868.

[44] Dinh, H., C. Lee, D. Niyato, and Ping Wang. Dec. 2013. A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches. *Wireless Communications and Mobile Computing* 13(18):1587–1611.

[45] Ekstrom, H., A. Furuskar, J. Karlsson, M. Meyer, S. Parkvall, J. Torsner, and M. Wahlqvist. 2006. Technical Solutions for the 3G Long Term Evolution. *IEEE Communications Magazine* 44(3):38–45.

[46] Fesehaye, D., Y. Gao, K. Nahrstedt, and G. Wang. 2012. Impact of Cloudlets on Interactive Mobile Cloud Applications. In *Proceedings of IEEE International Enterprise Distributed Object Computing Conference (EDOC)*.

[47] Fisher, M., G. Nemhauser, and G. Wolsey. 1978. An Analysis of Approximations for Maximizing Submodular set Functions-II. *Mathematical Programming Studies* 8:73–87.

[48] Fleischer, L., M. Goemans, V. Mirrokni, and M. Sviridenko. 2006. Tight Approximation Algorithms for Maximum General Assignment Problems. In *Proceedings of ACM-SIAM symposium on Discrete algorithm*, 611–620.

[49] Ha, K., P. Pillai, G. Lewis, S. Simanta, S. Clinch, N. Davies, and M. Satyanarayanan. 2013. The Impact of Mobile Multimedia Applications on Data Center Consolidation. In *Proceedings of IEEE International Conference on Cloud Engineering (IC2E)*.

[50] Han, Z., Z. Ji, and K. Liu. 2005. Fair Multiuser Channel Allocation for OFDMA Networks Using Nash Bargaining Solutions and Coalitions. *IEEE Transactions of Communications* 53(8):1366–1376.

[51] He, S., Y. Huang, L. Yang, A. Nallanathan, and P. Liu. 2012. A Multi-Cell Beamforming Design by Uplink-Downlink Max-Min SINR Duality. *IEEE Transactions on Wireless Communications* 11(8):2858–2867.

[52] Hou, I.-H., and C. S. Chen. 2012. Self-organized Resource Allocation in LTE Systems with Weighted Proportional Fairness. In *Proceedings of IEEE ICC*.

[53] Hu, D., and S. Mao. Apr. 2012. On Medium Grain Scalable Video Streaming over Femtocell Cognitive Radio Networks. *IEEE Journal on Selected Areas in Communications* 30(3):641–651.

[54] Huang, J., V. Subramanian, R. Berry, and R. Agrawal. 2009. In *Book chapter in Orthogonal Frequency Division Multiple Access Fundamentals and Applications*.

[55] Huang, T.-Y., R. Johari, N. McKeown, M. Trunnell, and M. Watson. 2014. A Buffer-Based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service. In *Proceedings of ACM Sigcomm*.

[56] Huang, Z., C. Mei, L. E. Li, and T. Woo. 2011. CloudStream: Delivering High-quality Streaming Videos through a Cloud-based SVC Proxy. In *Proceedings of IEEE Infocom*.

[57] ISO/IEC 23009-1, MPEG Dynamic Adaptive Streaming over HTTP (DASH). http://dashif.org/mpeg-dash/.

[58] Jakubczak, S., and D. Katabi. 2011. A Cross-Layer Design for Scalable Mobile Video. In *Proceedings of ACM MobiCom*.

[59] Jiang, J., V. Sekar, and H. Zhang. 2012. Improving Fairness, Efficiency, and Stability in HTTP-based Adaptive Video Streaming with FESTIVE. In *Proceedings of ACM CoNEXT*.

[60] Jo, H., C. Mun, J. Moon, and J. Yook. 2009. Interference Mitigation Using Uplink Power Control for Two-Tier Femtocell Networks. *IEEE Transactions on Wireless Communications* 8(10):4906–4910.

[61] Kauffmann, B., F. Baccelli, A. Chaintreau, V. Mhatre, K. Papagiannaki, and C. Diot. 2007. Measurement-Based Self Organization of Interfering 802.11 Wireless Access Networks. In *Proceedings of IEEE Infocom*.

[62] Kelly, F. 1997. Charging and Rate Control for Elastic Traffic. *European Transactions on Telecommunications* 8:33–37.

[63] Kelly, F. P., A. Maulloo, and D. Tan. 1998. Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability. *Journal of the Operational Research Society* 49(3):237–252.

[64] Kim, Y., S. Lee, and D. Hong. 2010. Performance Analysis of Two-Tier Femtocell Networks with Outage Constraints. *IEEE Transactions on Wireless Communications* 9(9):2695–2700.

[65] Kittipiyakul, S., and T. Javidi. 2004. Subcarrier Allocation in OFDMA Systems: Beyond Water-filling. *IEEE Signals, Systems, and Computers* 1: 334–338.

[66] Kokku, R., R. Mahindra, S. Rangarajan, and H. Zhang. 2011. Opportunistic Alignment of Advertisement Delivery with Cellular Basestation Overloads. In *Proceedings of ACM MobiSys*.

[67] Lakshmanan, S., K. Sundaresan, R. Kokku, A. Khojestepour, and S. Rangarajan. 2009. Towards Adaptive Beamforming in Indoor Wireless Networks: An Experimental Approach. In *Proceedings of IEEE INFOCOM*.

[68] Lee, T., J. Yoon, S. Lee, and K. G. Shin. 2010. Resource Allocation Analysis in OFDMA Femtocells Using Fractional Frequency Reuse. In *Proceedings of IEEE PIMRC*.

[69] Li, P., H. Zhang, B. Zhao, and S. Rangarajan. 2009. Scalable Video Multicast in Multi-carrier Wireless Data Systems. In *Proceedings of IEEE ICNP*.

[70] ———. 2010. Scalable Video Multicast with Joint Layer Resource Allocation in Broadband Wireless Networks. In *Proceedings of IEEE ICNP*.

[71] Li, Z., Y. Huang, G. Liu, F. Wang, and Z. Zhang. 2012. Cloud Transcoder: Bridging the Format and Resolution Gap between Internet Videos and Mobile Devices. In *Proceedings of ACM NOSSDAV*.

[72] Liu, C., I. Bouazizi, M. Hannuksela, and M. Gabbouj. Apr. 2012. Rate Adaptation for Dynamic Adaptive Streaming over HTTP in Content Distribution Network. *Image Communication* 27(4):288–311.

[73] Liu, X., A. Sheth, M. Kaminsky, K. Papagiannaki, S. Seshan, and P. Steenkiste. 2009. DIRC: Increasing Indoor Wireless Capacity Using Directional Antennas. In *Proceedings of ACM Sigcomm*.

[74] ———. 2010. Pushing the Envelope of Indoor Wireless Spatial Reuse Using Directional Access Points and Clients. In *Proceedings of ACM MobiCom*.

[75] Lopez-Perez, D., G. Roche, A. Valcarce, A. Juttner, and J. Zhang. 2008. Interference Avoidance and Dynamic Frequency Planning for WiMAX Femtocells Networks. In *Proceedings of IEEE ICCS*.

[76] Lopez-Perez, D., A. Valcarce, G. Roche, and J. Zhang. 2009. OFDMA Femtocells: A Roadmap on Interference Avoidance. *IEEE Communications Magazine* 47(9):41–48.

[77] McCanne, S., V. Jacobson, and M. Vetterli. 1996. Receiver-Driven Layered Multicast. In *Proceedings of ACM SigComm*.

[78] Mishra, A., S. Banerjee, and W. Arbaugh. 2005. Weighted coloring based channel assignment for WLANs. In *Proceedings of ACM SIGMOBILE Mobile Computing and Communications Review*.

[79] Mishra, A., V. Brik, S. Banerjee, A. Srinivasan, and W. Arbaugh. 2006. Client-driven channel management for wireless lans. In *Proceedings of IEEE Infocom*.

[80] Morris, R., E. Kohler, J. Jannotti, and M. F. Kaashoek. 1999. The click module router. *SIGOPS Operating System Rev.* 33(5):217–231.

[81] Moscibroda, T., R. Chandra, Y. Wu, S. Sengupta, P. Bahl, and Y. Yuan. 2008. Load-Aware Spectrum Distribution in Wireless LANs. In *Proceedings of IEEE ICNP*.

[82] Muller, C., S. Lederer, and C. Timmerer. 2012. An Evaluation of Dynamic Adaptive Streaming over HTTP in Vehicular Environments. In *Proceedings of ACM MoVid*.

[83] Navda, V., A. P. Subramanian, K. Dhanasekaran, A. Timm-Giel, and S. Das. 2007. MobiSteer: Using Steerable Beam Directional Antenna for Vehicular Network Access. In *Proceedings of ACM MobiSys*.

[84] Patra, R. K., S. Nedevschi, S. Surana, A. Sheth, L. Subramanian, and E. A. Brewer. 2007. WiLDNet: Design and implementation of high performance WiFi based long distance networks. In *Proceedings of USENIX NSDI*.

[85] Paulraj, A., R. Nabar, and D. Gore. 2003. Introduction to Space-Time Wireless Communications. In *Cambridge University Press*.

[86] Pfletschinger, S., G. Muenz, and J. Speidel. 2002. Efficient Subcarrier Allocation for Multiple Access in OFDM Systems. In *Proceedings of 7th International OFDM-Workshop*.

[87] Quek, T., Z. Lei, and S. Sun. 2009. Adaptive Interference Coordination in Multi-cell OFDMA Systems. In *Proceedings of IEEE PIMRC*.

[88] Ramachandran, K., R. Kokku, K. Sundaresan, M. Gruteser, and S. Rangarajan. 2009. R2D2: Regulating Beam Shape and Rate as Directionality Meets Diversity. In *Proceedings of ACM MobiSys*.

[89] Satyanarayanan, M., P. Bahl, R. Caceres, and N. Davies. Oct. 2009. The Case for VM-Based Cloudlets in Mobile Computing. *IEEE Pervasive Computing* 8(4):14–23.

[90] Sen, S., N. K. Madabhushi, and S. Banerjee. 2010. Scalable WiFi Media Delivery Through Adaptive Broadcasts. In *Proceedings of USENIX NSDI*.

[91] Seong, K., M. Mohseni, and John M. Cioffi. 2006. Optimal Resource Allocation for OFDMA Downlink Systems. In *Proceedings of IEEE ISTS*.

[92] Shrivastava, V., S. Rayanchu, J. Yoon, and S. Banerjee. 2008. 802.11n Under the Microscope. In *Proceedings of ACM IMC*.

[93] Sun, Y., R. P. Jover, and X. Wang. 2012. Uplink Interference Mitigation for OFDMA Femtocell Networks. *IEEE Transactions on Wireless Communications* 11(2):614–625.

[94] Sundaresan, K., K. Ramachandran, and S. Rangarajan. 2009. Optimal Beam Scheduling for Multicasting in Wireless Networks. In *Proceedings of ACM MobiCom*.

[95] Sundaresan, K., and S. Rangarajan. 2009. Efficient Resource Management in OFDMA Femto Cells. In *Proceedings of ACM MobiHoc*.

[96] Takai, M., J. Martin, R. Bagrodia, and A. Ren. 2002. Directional Virtual Carrier Sensing for Directional Antennas in Mobile Ad Hoc Networks. In *Proceedings of ACM MobiHoc*.

[97] Tan, K., J. Fang, Y. Zhang, S. Chen, L. Shi, J. Zhang, and Y. Zhang. 2010. Fine-grained Channel Access in Wireless LAN. In *Proceedings of ACM SIGCOMM*.

[98] Vickers, B., C. Albuquerque, and T. Suda. 2000. Source-Adaptive Multilayered Multicast Algorithms for Real-Time Video Distribution. *IEEE/ACM Transactions on Networking* 8(6):720–733.

[99] Viswanathan, H., and A. L. Stolyar. 2010. Interference Management in Femto/Small Cell and Macro Environments. In *Proceedings of IEEE CTW*.

[100] Wong, C., R. Cheng, K. Letaief, and R. Murch. 1999. Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation. *IEEE Journal on Selected Areas in Communications* 17(10):1747–1758.

[101] Yang, L., W. Hou, Z. Zhang, B. Zhao, and H. Zheng. 2010. Jello: Dynamic Spectrum Sharing in Digital Homes. In *Proceedings of IEEE Infocom*.

[102] Yeh, S., S. Talwar, S. Lee, and H. Kim. 2008. WiMAX Femtocells: A Perspective on Network Architecture, Capacity, and Coverage. *IEEE Communications Magazine* 46(10):58–65.

[103] Yuan, Y., P. Bahl, R. Chandra, P. Chou, I. Ferrell, T. Moscibroda, S. Narlanka, and Y. Wu. 2007. KNOWS: Kognitiv Networking Over White Spaces. In *Proceedings of IEEE DySPAN*.

[104] Yun, J., and K. G. Shin. 2010. CTRL: A Self-Organizing Femtocell Management Architecture for Co-Channel Deployment. In *Proceedings of ACM MobiCom*.

[105] Zhang, Q., L. Cheng, and R. Boutaba. May 2010. Cloud Computing: State-of-the-art and Research Challenges. *Journal of Internet Services and Applications* 1(1):7–18.

[106] Zhang, Z. May 2005. DTRA: Directional Transmission and Reception Algorithms in WLANs with Directional Antennas for QoS Support. *IEEE Network* 19(3):27–32.