Performance Evaluation of Inter-cell Interference Mitigation Techniques for OFDMA Cellular Networks

by

Weiwei Wu

Submitted in total fulfilment of
the requirements for the degree of

Doctor of Philosophy

Department of Electrical and Electronic Engineering
The University of Melbourne
Australia

September, 2010

Produced in LaTeX 2ε

The University of Melbourne

Australia

Abstract

Performance Evaluation of Inter-cell Interference Mitigation Techniques for OFDMA Cellular Networks

by Weiwei Wu

For emerging cellular wireless systems, the mitigation of inter-cell interference is the key to achieve a high capacity and good user experience. This thesis is devoted to the performance analysis of interference mitigation techniques for the downlink in an orthogonal frequency division multiple access (OFDMA) network, with a focus on the Long Term Evolution (LTE) standard. We investigate two types of coordination techniques for interference mitigation, namely reuse partitioning and resource prioritization in the frequency domain.

First, we assume best-effort elastic traffic for broad-band data networks and introduce a new metric, called the flow capacity, to indicate the maximum traffic intensity that can be supported by a base station sector while satisfying a minimum level of provided service. We develop a queueing theoretic methodology to analyse the flow capacity for standard reuse and reuse partitioning schemes with different scheduling algorithms. Using this analysis framework, we show how an improved cell-edge throughput can translate into an improvement in the flow capacity. We develop model variants for infinite (Poisson arrivals) and finite user populations; the infinite user population model is more tractable and yields simple, insightful expressions for the flow capacity, while the finite user population model has greater practical relevance. Furthermore, we develop a methodology to account for the effect of interference from neighbouring base stations with an arbitrary level of loading.

Next, we propose possible distributed realizations of interference coordination schemes in a reuse-1 environment, which are based on setting allocation priority in the frequency domain. The proposed schemes are more suited to narrow-band

services and can be implemented in a fractional loading scenario.

This is to certify that

(i) the thesis comprises only my original work,

(ii) due acknowledgement has been made in the text to all other material used,

(iii) the thesis is less than 100,000 words in length, exclusive of table, maps, bibliographies, appendices and footnotes.

Signature_____

Date_____

Acknowledgments

I would like to take this opportunity to express my sincere thanks to all the people who have given me kind help and support throughout my Ph.D study.

Firstly, I would like to specially thank my dear wife, parents and parent in-laws for their constant encouragement and unstinting support during the whole process of my Ph.D study. I am also grateful to my baby son for being such a wonderful gift just before the completion of my study.

I am immensely grateful to my supervisors Prof. Moshe Zukerman, Dr. Taka Sakurai, Prof. Jonathan Manton, and Prof. Maxim Gitlits. I thank Moshe for opening the door to research for me, assisting my obtaining a scholarship, and providing his generous encouragement at the beginning stage of my study. I am extremely grateful to Taka for his valuable guidance, patience, participation and encouragement throughout my study. Taka has devoted much time to my research and has taught me so much about doing research. Without his support, I could not have successfully completed my thesis. I thank Jonathan for being my department supervisor after Moshe left the University and overcoming various administration hurdles in the department. I am also grateful to Maxim for his participation and valuable comments and suggestions in the research of Chapter 5. I would like to extend my thanks to Prof. Bill Moran for providing supports and comments on some research topics in the early work.

I appreciate the support of the Australian Research Council (ARC) Special Research Centre for Ultra-Broadband Information Networks (CUBIN) for the excellent research environment. Acknowledgment is also extended to our CUBIN members for their friendships and valuable discussions.

Contents

List of Figures

List of Tables

List of Abbreviations

1xEv-do	Evolution-Data Optimized
3GPP	3rd Generation Partnership Project
AMC	Adaptive Modulation and Coding
BS	Base Station
CQI	Channel Quality Indicator
DPS	Discriminatory Processor Sharing
EDGE	Enhanced Data Rates for Global Evolution
eNB	Evolved NodeB
EPS	Egalitarian Processor Sharing
FDD	Frequency Division Duplex
FFR	Fractional Frequency Reuse
FH	Frequency Hopping
FPM	Fixed Population Mean
GPRS	General Packet Radio Service
GPS	Generalized Processor Sharing
GSM	Global System for Mobile Communications
HSPA	High Speed Packet Access
ICI	Inter-Cell Interference
LTE	Long Term Evolution
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PF	Proportional Fair

PSD	Power Spectral Density
RB	Resource Block
RBP	Resource Block Pair
RR	Round-robin
RS	Reference Signal
SC-FDMA	Single-Carrier Frequency Division Multiple Access
SFC	Space Frequency Coding
SFR	Soft Frequency Reuse
SIMO	Single-Input Multiple-Output
SINR	Signal to Interference and Noise Ratio
SISO	Single-Input Single-Output
TDMA	Time Division Multiple Access
TTI	Transmission Time Interval
UE	User Equipment
V-BLAST	Vertical Bell Labs Layered Space Time
WCDMA	Wideband Code Division Multiple Access
WiMAX	Worldwide Interoperability for Microwave Access

Chapter 1

Introduction

1.1 Overview and research motivation

The mobile and wireless communications industry has enjoyed tremendous growth over the past two decades, and today, there are over four billion mobile subscribers worldwide [59]. The focus in mobile and wireless systems has gradually shifted from high volume voice services to high volume and high speed data services. This has led to the development of third generation (3G) wireless technologies, which are dominated by code division multiple access (CDMA) based technologies such as wideband code division multiple access (WCDMA) and high speed packet access (HSPA) [39, 53]. The 3G technologies can support the needs of existing wireless broadband services. However, as more applications are devised and put into service, there are increasing demands for better support for even higher data rates and higher capacities. This has stimulated the need to develop orthogonal frequency division multiple access (OFDMA) based next generation networks, including 3GPP Long Term Evolution (LTE) [121] and Mobile WiMAX [55, 96]. OFDMA is well suited to support higher data rates due to its advantages over CDMA, which include bandwidth scalability, compatibility with multiple-input multiple-output (MIMO) antenna techniques, and immunity against multi-path fading [147].

In CDMA-based networks, the data transmission of one user suffers not only from inter-cell interference caused by all concurrent users in neighbouring cells or sectors (these two terms are used interchangeably), but also intra-cell interference originated by every other user transmitting in the same cell. In contrast, in OFDMA-based networks, the frequency bandwidth is divided into a set of sub-carriers, and the orthogonality of the sub-carriers effectively eliminates intra-cell interference, leaving

inter-cell interference as the most prominent source of interference in a multi-cell environment. However, inter-cell interference can be severe due to the following: (i) the scarcity of spectrum combined with the desire for high peak rates leads to an objective of universal frequency reuse (i.e. reuse-1), or as close to that as practical; (ii) a need for a high capacity per unit area causes the sector sizes to be small, which leads to a dense network. Therefore, with such deployments, the management of inter-cell interference is a challenge for wireless operators. This thesis is devoted to the performance analysis of techniques which help to mitigate inter-cell interference in OFDMA systems.

In a dense reuse-1 network, a user at the sector edge is exposed to strong inter-cell interference, which leads to a low signal to interference and noise ratio (SINR), and, consequently, limited throughput. Therefore, there is a need for interference mitigation techniques at the base station and/or the mobile terminal to improve the user performance, particularly for the edge users.

There are a number of different types of interference mitigation techniques. Interference coordination is one of the most promising approaches, and has received considerable attention in the literature. The basic principle is to apply some restrictions to the resource allocations in a coordinated way among neighbouring base stations. The coordination can be achieved via resource management or resource scheduling. OFDMA is much more amenable to resource coordination than CDMA, since OFDMA resources are partitioned in both frequency and time.

Coordination schemes involving resource management include conventional reuse schemes with a frequency reuse factor greater than 1 (such as reuse-3) and reuse partitioning [61, 128]. In reuse partitioning schemes, the available frequency bandwidth in each sector is partitioned into a reuse-1 band that is allocated to sector centre users and a reuse-3 (or larger reuse) band allocated to sector edge users. Two variations are possible, namely fractional frequency reuse (FFR) and soft frequency reuse (SFR), depending on the different use of the edge bands in neighbouring sectors. Reuse partitioning schemes can achieve a frequency reuse factor close to 1 if the resources in the network are planned well.

Many studies on reuse partitioning [33, 57, 106, 111, 115, 119, 125, 146] have ap-

plied simulation under the assumption of persistent traffic (full-buffer) to investigate the base station performance, including the sector throughput and the sector edge throughput. The general consensus is that reuse partitioning leads to an improvement in the sector edge throughput compared to the conventional reuse-1 scheme, but a reduction in the sector throughput. However, the key question of whether the improvement in the edge performance justifies the reduction in the sector throughput is left unanswered.

In coordination schemes based on resource scheduling, users are dynamically assigned those resources on which they suffer less interference. Many studies investigate this allocation problem using various optimization techniques [48, 75, 77, 118, 124, 126, 127, 149]. However, they assume the existence of a central controller which has full system state knowledge, which is not feasible in real systems. Furthermore, if the number of users in the system is large, the amount of computation required for the solutions can be prohibitive. A much simpler alternative is to do the scheduling based on prioritization of the resource allocations in the frequency domain. The key question is whether such an approach can deliver worthwhile performance gains.

1.2 Research objective

The aim of this thesis is to determine whether or not reuse partitioning or resource prioritization in the frequency domain provide benefits over conventional resource allocation and scheduling approaches. To provide a concrete example for our numerical studies, we focus on the downlink transmission direction of a LTE network. Note, however, that our analysis framework and proposed schemes are equally applicable to Mobile WiMAX or any other OFDMA-based cellular technology.

First, we assume best-effort elastic traffic for broad-band data networks and introduce a new metric, called the flow capacity, to indicate the maximum traffic intensity that can be supported by a base station sector while satisfying a minimum level of provided service. To indicate the level of service, we apply the flow throughput to be the user-level performance metric. The flow throughput gives an estimate of the average throughput experienced by a user when downloading a file

(see Bonald and Proutiére [26] and Borst [28]). We develop a queueing theoretic methodology to analyse the flow capacity for standard reuse and reuse partitioning schemes with different scheduling algorithms. We apply this framework to find the capacity benefits of reuse partitioning schemes.

Next, we propose possible distributed realizations of interference coordination schemes in a reuse-1 environment, which are based on setting allocation priority in the frequency domain. The proposed schemes are more suited to narrow-band services and can be implemented in a fractional loading scenario. We perform system-level simulation to investigate the performance improvement due to the resource prioritization schemes.

1.3 Thesis outline

The chapters in this thesis are organized as follows.

In Chapter 2, we give some background and context necessary for understanding the materials in subsequent chapters. We briefly review some basic concepts of 3GPP LTE, and survey the literature related to interference mitigation techniques. Then we discuss several important performance metrics defined for high data rate wireless networks. Since a large part of our work is about modelling the scheduler at a base station at a flow level, processor sharing queueing disciplines are introduced.

In Chapter 3 and Chapter 4, we use a flow-level queueing theoretic approach to analyse and compare the downlink performance of reuse-1, reuse-3 and static reuse partitioning schemes, namely FFR and SFR, with different scheduling algorithms. We develop model variants for infinite (Chapter 3) and finite (Chapter 4) user populations to characterise the user performance for elastic traffic, and define a new flow capacity metric as a basis for comparison of different reuse schemes. In our models, we employ a hybrid simulation/analysis approach, where a rate distribution obtained via simulation is used as input to the queueing model. Furthermore, we develop a methodology to account for the effect of interference from neighbouring base stations with an arbitrary level of loading.

In Chapter 5, we propose several inter-cell interference coordination schemes for

the LTE downlink to enable efficient utilization of the entire available bandwidth. The schemes are based on setting resource allocation priority in the frequency domain; variants are possible, depending on whether the allocation priority is assigned through off-line network planning (static) or is made adaptive to traffic load variations in neighbouring sectors. We perform system-level simulation to investigate the base station performance, namely the average sector throughput and the sector edge throughput.

In Chapter 6, we summarize the main results in this thesis, and describe potential directions for future research.

1.4 Contributions

The main contributions of this thesis are as follows.

(i) We use a flow-level queueing theoretic approach to characterise the user performance for elastic traffic, and introduce a new performance metric, which we call the flow capacity. By comparing the flow capacities of different interference mitigation schemes, we show how an improved cell-edge throughput can translate into an improvement in the flow capacity.

(ii) In addition to modelling the baseline round-robin (RR) scheduler, we show how the variations of the proportional fair (PF) scheduling gain with the number of active users can be incorporated into the analytical model. Furthermore, for the PF case, we implement a computationally efficient algorithm to solve for the flow throughputs by exploiting a fast convergence property of the multi-user diversity gain.

(iii) We perform the capacity analysis using a hybrid simulation/analysis approach, which dramatically reduces the computational effort compared to a pure simulation approach. Our approach requires simulation only of the single user rate distribution, which can be obtained from a simple static system-level simulation.

(iv) We develop model variants for both infinite user population (Poisson arrivals) and finite user population models. The infinite user population model is more tractable and yields simple, insightful expressions for the flow capacity for the round-robin scheduler. The finite user population model is more computationally complex but has greater practical appeal, since real sector populations are finite. In the finite user population model, we go further than the works by Bonald and Proutiére [26], Borst [28], and Liu and Virtamo [83] by devising computationally efficient methods to solve for the flow throughputs. Furthermore, we present an approximation method to calculate an estimate of the flow capacity with reduced computational complexity, which is particularly useful for the PF case.

(v) We extend the flow-level methodology of [26, 28, 83] to the reuse partitioning setting, which requires definition of separate processor sharing queueing systems for the sector centre and sector edge users. In the process, we significantly extend the methodology itself in several directions that are not specific to reuse partitioning, but have general applicability so that our framework can be also used to find the capacity benefits of general performance enhancement techniques. We present an example that explores the capacity benefits of MIMO schemes.

(vi) We perform the analysis in a multi-cell scenario with fractional loading, show how fractional loading in the interfering sectors can approximately be taken into account, and present an iterative approach to find the flow capacity for a homogeneous load network.

(vii) We propose possible distributed realizations of interference coordination schemes in a reuse-1 environment based on resource prioritization in the frequency domain, where the allocation priority can be assigned statically through network configuration or be made adaptive to traffic load variations. For the static schemes, we apply the idea of traditional reuse or a heuristic and greedy algorithm to assign the priorities. For the adaptive schemes, we define the interference weights to indicate different interference impacts of the neighbouring

sectors which can be pre-computed reflecting the average interference impact or be user-specific depending on the user channel conditions. Inter-base-station signalling is introduced to obtain the resource allocation information in the neighbouring sectors. The tradeoff between system performance and signalling overhead in the adaptive schemes is investigated.

All the above contributions are fully or partially presented in the following publications and submissions.

- W. Wu, M. Gitlits, and T. Sakurai. Dynamic resource allocation with inter-cell interference coordination for 3GPP LTE. In *Asia Pacific Microwave Conference, APMC 2008*, pages 1-4, December 2008.

- W. Wu and T. Sakurai. Capacity of reuse partitioning schemes for OFDMA wireless data networks. In *IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC2009*, pages 2240-2244, September 2009.

- W. Wu and T. Sakurai. Flow-level capacity of fractionally loaded OFDMA networks with proportional fair scheduling. In *IEEE Vehicular Technology Conference, VTC-2010 Fall*, September 2010.

- W. Wu and T. Sakurai. Flow capacity of different reuse schemes in OFDMA wireless data networks. Submitted to *IEEE Transactions on Wireless Communications*.

During the course of my PhD study, some other research has been made which is not included in this thesis. The following is the publication related to this research.

- W. Wu, B. Moran, J. H. Manton, and M. Zukerman. Topology design of undersea cables considering survivability under major disasters. In *International Conference on Advanced Information Networking and Applications Workshops, WAINA 2009*, pages 1154-1159, May 2009.

Chapter 2

Background

2.1 Introduction

The aim of this chapter is to present the background and context necessary for understanding the materials in subsequent chapters. We first review some basic concepts of 3GPP LTE, followed by an overview of various interference mitigation techniques. In addition, several important performance metrics defined for high data rate wireless networks are presented. Since a large part of our work is about modelling the scheduler at a base station (BS) at a flow level, processor sharing queueing disciplines are introduced.

In Section 2.2, an overview of LTE and OFDMA, and the associated inter-cell interference problem are presented. Then, a brief survey of the interference averaging, cancellation and coordination techniques is given in Section 2.3. The primary metrics for base station performance and user-level performance for evaluating resource allocation schemes are covered in Section 2.4. Lastly, the processor sharing queueing disciplines to model round-robin (RR) and proportional fair (PF) scheduling are introduced in Section 2.5.

2.2 Basic concepts in 3GPP LTE

In the past decade, the mobile and wireless communications industry has experienced an explosive growth and brought significant changes to the design of wireless network systems. While the earlier mobile communications standards focused primarily on voice traffic, the emphasis now is on the provision of high data rate service. Although 3G technologies such as WCDMA [53] and HSPA [39] provide significantly higher

speed data communication services than 2G technologies such as general packet radio service (GPRS) [51] and enhanced data rates for global evolution (EDGE) [73], there are increasing user demands for still higher data rates and higher quality mobile communication services.

Long Term Evolution (LTE) is the next step forward in the road-map of the 3GPP mobile communications standardisation body, and will be the basis on which next generation systems will be built [64]. The main objectives of LTE are to provide high data rates (greater than 100 Mbps for the peak rate in the downlink and greater than 50 Mbps in the uplink), low user plane and control plane latency, low cost (for the operators and end users) and packet-optimized radio access technology, while supporting flexible spectrum allocations (see 3GPP specification [1]).

The conflict of limited spectrum and rapidly growing user demands requires that the modulation and multiple access scheme in LTE must be much more spectrally efficient and flexible than those applied in current mobile systems. Orthogonal frequency division multiple access (OFDMA) is one such key technology. LTE has selected OFDMA for the downlink radio access, and single-carrier frequency division multiple access (SC-FDMA), which is a modified form of OFDMA, for the uplink (see 3GPP specification [4]).

OFDMA is a multi-user version of the OFDM modulation scheme, which has the great advantage of immunity against severe frequency selective fading. In OFDMA, the data is transmitted over a large number of narrow-band sub-carriers. The sub-carrier frequencies are chosen so that the sub-carriers are orthogonal to each other, which can substantially decrease the inter-carrier interference, or cross-talk between sub-carriers. Furthermore, an extra guard interval, known as the cyclic prefix (CP), is introduced in OFDM modulation to overcome the time dispersion of the channel, which helps to eliminate the inter-symbol interference.

In LTE, the smallest unit of resource that can be allocated is called a resource block (RB). According to the 3GPP standard [4], one RB spans a 0.5 ms slot in the time domain and consists of 180 kHz (12 adjacent OFDM sub-carriers) in the frequency domain, as depicted in Figure 2.1. In a frequency division duplex (FDD) LTE system, the scheduling is done on a sub-frame (1 ms) basis, and the resource

allocation is usually carried out in terms of a resource block pair (RBP), which consists of two consecutive RBs (one sub-frame) in the time domain. Within one RBP, some specific resource elements, called reference signals (RS, see Figure 2.1), are used for channel estimation, timing synchronization, and other purposes. A sub-band is a group of several adjacent RBs in the frequency domain; the size of a sub-band varies depending on the total available frequency bandwidth in the system (see 3GPP specification [5]).

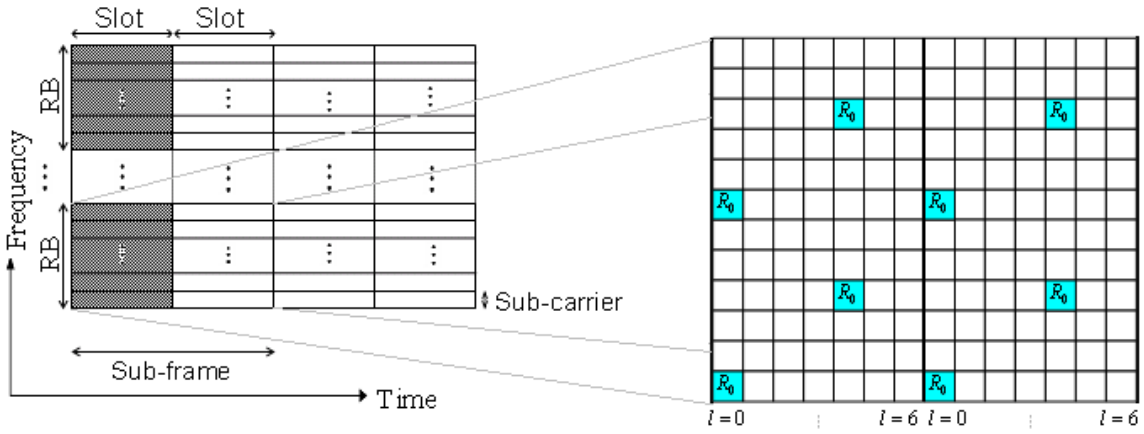

Figure 2.1: The structure of resource block and sub-frame in LTE for normal cyclic prefix.

A fixed tri-sector layout is usually considered, with one base station (referred to as e-NodeB or eNB) at the centre of each site, controlling the three sectors, as depicted in Figure 2.2. When the sector size is small, the dense network leads to an interference-limited environment.

In an attempt to maximise the capacity, a frequency reuse of 1 among sectors is typically suggested, where all the frequency resources are available everywhere in each sector. In a reuse-1 system, the interference will severely limit the user performance, particularly at the sector edge. Since intra-cell interference can be eliminated if the user equipments (UE) within one sector are allocated mutually exclusive RBPs, inter-cell interference (ICI) is the most prominent source of interference in a multi-cell environment. In the downlink, ICI comes from the neighbouring base stations using the same resource, as shown in Figure 2.2, while in the uplink, ICI is gener-

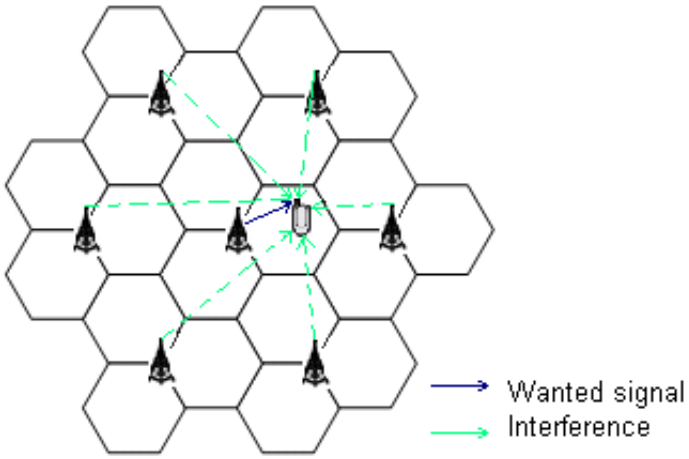

Figure 2.2: A tri-sector layout and downlink inter-cell interference.

ated by the UEs allocated the same resource in other sectors. As a consequence, an important research topic in LTE networks, and OFDMA networks in general, is how to reduce the ICI. In this chapter, we focus on downlink ICI mitigation techniques, most of which can also be applied directly or with minor changes to the uplink.

2.3 Interference mitigation techniques

A variety of interference mitigation techniques have been proposed for OFDMA networks, including interference cancellation at the UE, interference averaging, interference coordination, and other advanced techniques such as multi-antenna transmissions. The different types of techniques can often be applied in combination for more effective interference mitigation.

2.3.1 Interference averaging and cancellation

Interference averaging is the class of techniques which attempts to randomize the interfering signals and hence distribute the interference among all users evenly, such that the edge user will not always suffer strong ICI during all the transmission period.

The early 3GPP proposals for LTE [107–110] consider a distributed RB definition

in the frequency domain as a means of achieving frequency diversity. In this scheme, the sub-carriers within a RB are distributed over the operating bandwidth instead of being contiguous as shown in Figure 2.1 (also see Pokhariyal, Kolding and Mogensen [101]). Sector-specific scrambling is another proposal in 3GPP to randomize the interference from surrounding sectors (see 3GPP contributions [105, 113]).

Frequency hopping (FH) is another well-known technique to average the interference, which has been successfully applied in Global System for Mobile Communications (GSM) networks [95]. As shown in Figure 2.3, the transmit and receive carrier frequencies are dynamically assigned for each frame in GSM via hopping between available frequencies according to a specific sequence, which can be cyclic or pseudo random. FH can also be adopted in OFDMA networks (see Kim *et al.* [67] and Stolyar and Viswanathan [129]). If planned correctly, FH can reduce the possibility of using the same RBs in adjacent sectors all the time.

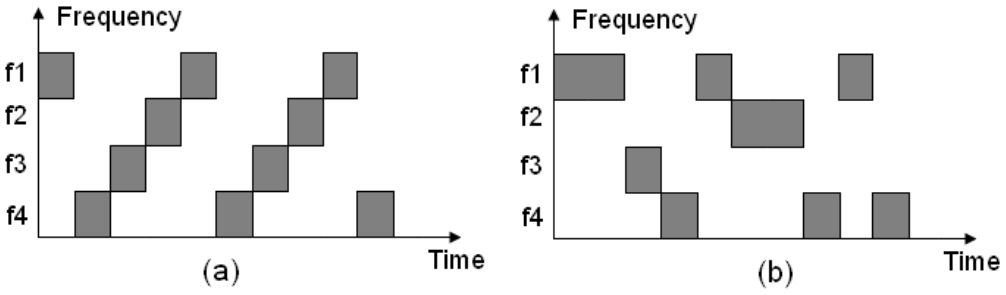

Figure 2.3: An example of FH in GSM with four hopping frequencies: (a) cyclic; (b) pseudo random.

Interference cancellation techniques use extra signal processing at the receiver to suppress the interference. The basic principle is illustrated in Figure 2.4. The receiver first carries out the channel estimation of the interference signals using approaches such as minimum mean squared estimation (MMSE) or maximum likelihood sequence estimation (MLSE) on the reference signals (see Li, Seshadri and Ariyavisitakul [78] and Beek *et al.* [134]), and subtracts the estimates from the received signal to obtain an interference-cancelled signal. There are generally three categories: successive interference cancellation, parallel interference cancellation and iterative interference cancellation (see Andrews [11] and WINNER project [56]).

Since the data detection is performed on the interference-cancelled signal instead of the original received signal, the signal quality can be improved.

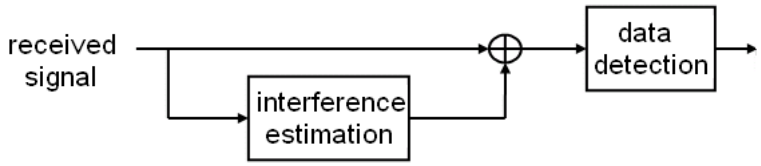

Figure 2.4: The basic principle of interference cancellation.

Impact on system design and discussion

The main impacts on the system design that need to be considered are the possible changes in system architecture, additional measurements for the channel quality, and signalling overhead for the communications between UE and base stations or the communications between base stations (see WINNER project [56, 57]).

The support of interference averaging techniques has very limited impact on the system architecture. The distributed definition of RB, sector-specific scrambling, and frequency hopping sequences may be set by network planning procedures. No additional measurements and signalling are needed. However, these types of techniques will cause interference to some users who may be initially in good signal quality conditions. Moreover, interference averaging is mainly beneficial for narrowband services with small packet sizes, and even then, the overall performance benefit can be quite small [56].

Interference cancellation in the downlink will impose additional complexity on the device. The accuracy of the channel estimation is the main concern; estimation errors may degrade the system performance greatly. Therefore, the network should be synchronized well in the time domain; the parameters of the interfering signals, such as modulation and code scheme (MCS), should be known at the receiver, which may require some control signalling. Furthermore, these techniques introduce additional processing latency, and are therefore only applied to deal with the dominating interference.

2.3.2 Interference coordination

Interference coordination (also referred to as interference avoidance) is one of the most promising approaches to solve the problem of ICI in OFDMA systems, and has received considerable attention in 3GPP for LTE. The basic principle is to apply some restrictions to the resource allocations in a coordinated way among neighbouring base stations. The coordination techniques can be classified into two categories: resource management and resource scheduling (see Hernández, Guío and Valdovinos [52], Boudreau *et al.* [29], Hu, Luo and Chen [54], and WINNER project [57]). Furthermore, depending on the time scale (days, or minutes/seconds/milliseconds), the coordination between base stations can be classified as static or adaptive.

Standard reuse and reuse partitioning schemes

LTE is designed to operate with an aggressive frequency reuse plan, with reuse-1 as an objective. Figure 2.5(a) illustrates the frequency and power allocations in the reuse-1 scheme, where all frequency resources are available everywhere in each sector. The lower pictures in Figure 2.5 depict the mapping of power (P) to frequency (F), showing the frequency partitions; the upper pictures show the allocation of the frequency partitions to the sectors, including centre and edge allocations in reuse partitioning schemes.

A conventional reuse scheme with a reuse factor greater than 1, such as reuse-3 or reuse-7, is the simplest interference coordination technique based on resource management. Figure 2.5(b) illustrates the power and frequency configuration for a reuse-3 scheme. In reuse-3, each sector only gets one third of the bandwidth of the reuse-1 case, and the allocations are configured to be orthogonal among immediately neighbouring sectors. This type of static resource management can avoid allocating the same frequency resource in the adjacent sectors, leading to substantially lower interference.

Reuse partitioning (see Katzela and Naghshineh [61] and Sternad *et al.* [128]) is another type of resource management technique to mitigate ICI and improve sector-edge performance. The essential idea is to partition the available frequency band in

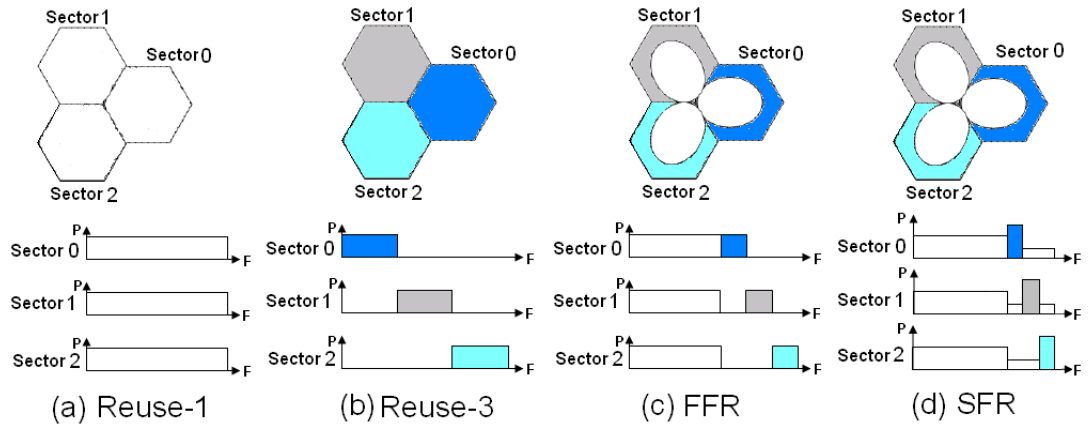

Figure 2.5: Standard reuse schemes: (a) reuse-1, (b) reuse-3 and reuse partitioning schemes: (c) FFR, (d) SFR.

each sector into a reuse-1 band that is allocated to sector centre users and a reuse-3 (or larger reuse) band allocated to sector edge users. Variants are possible, depending on whether the use of the edge bands of neighbouring sectors is strictly prohibited in the reference sector (fractional frequency reuse (FFR), shown in Figure 2.5(c), see Elayoubi, Haddada and Fourestié [42], and Simonsson [125]), or made available with reduced transmission power for sector centre users (soft frequency reuse (SFR), shown in Figure 2.5(d), see [57], Rahman, Yanikomeroglu and Wong [119], and Xiang *et al.* [146]). A metric needs to be introduced to differentiate the centre/edge users, which can be the location, geometry factor (see Won *et al.* [139] and Zhang *et al.* [148]) or an SINR threshold (see [142] or Chapter 3). If a user is deemed to be an edge user, only edge band resources can be allocated. Users in the centre area can access both the centre band and any unoccupied edge band resources.

In FFR and SFR schemes, the edge band size is an important design parameter, which can be changed through operations and maintenance intervention (static) or adaptive to traffic load variations in neighbouring sectors. Furthermore, the power levels for the centre and edge band can also be adjusted to pursue an improved overall system performance. Stolyar and Viswanathan [129, 130] propose one such algorithm, called *Multi-cell Gradient* (MGR), to adjust the transmit power of different sub-bands with periodical exchange of information between neighbouring sectors.

The information exchanged is the gradient of the system utility function with respect to the current sub-band transmit powers in the sector.

These schemes may bring in some changes to the system design and implementation, and each scheme has its own advantages and disadvantages. In the static reuse schemes and reuse partitioning schemes, the spectrum allocations in each sector remain constant over time, and can be done off-line via network planning. Therefore, no inter-base-station communication is required. However, these schemes have the disadvantage that the available bandwidth in each sector is usually much less than that of reuse-1 scheme. In particular, it is extremely low in the standard reuse schemes, for example, only one third for reuse-3 compared to reuse-1 and even lower for schemes with a larger reuse factor. For the adaptive FFR/SFR schemes, some additional information exchange will be required via inter-base-station signalling in order to agree on the edge band size, possibly on a relatively slow time scale.

Dynamic scheduling

Due to the dynamic arrivals of data flows, as well as the time-varying characteristics of the interference, a scheduling algorithm can be applied to dynamically allocate to the users those resources on which they suffer less interference. The ICI coordination schemes based on dynamic resource scheduling take a reuse-1 scheme into account. The essential problem is how to improve the user performance, especially for those in the sector edge area.

Numerous studies on dynamic resource scheduling in OFDMA systems in the single cell (see Li and Liu [76], Shen, Andrews and Evans [124], Song and Li [126, 127], Wang *et al.* [135], Wong *et al.* [140], and Zhang and Letaief [149]) and multi-cell context (see Gesbert *et al.* [48], Li and Liu [75,77], and Rahman and Yanikomeroglu [118]) are available in the literature. Most of these studies solve the problem using various optimization techniques to maximise the total throughput (or utility) or minimise the total transmit power.

In the single-cell multi-user scenario, the optimization problem can be formulated

as

$$\begin{aligned} \max \quad & \sum_{k=1}^{K}\sum_{n=1}^{N} \delta_{k,n} r_{k,n}, \\ \text{such that} \quad & \sum_{k=1}^{K}\sum_{n=1}^{N} \delta_{k,n} P_{k,n} \leq P_x, \\ & \sum_{n=1}^{N} \delta_{k,n} r_{k,n} \geq R_k \qquad \forall k, \\ & \text{if } \delta_{k,n} = 1, \text{ then } \delta_{k',n} = 0 \qquad \forall k' \neq k, \\ & P_{k,n} \geq 0, r_{k,n} \geq 0 \qquad \forall k, n, \end{aligned} \tag{2.1}$$

or

$$\begin{aligned} \min \quad & \sum_{k=1}^{K}\sum_{n=1}^{N} \delta_{k,n} P_{k,n}, \\ \text{such that} \quad & \sum_{n=1}^{N} \delta_{k,n} r_{k,n} \geq R_k \qquad \forall k, \\ & \text{if } \delta_{k,n} = 1, \text{ then } \delta_{k',n} = 0 \qquad \forall k' \neq k, \\ & P_{k,n} \geq 0, r_{k,n} \geq 0 \qquad \forall k, n, \end{aligned} \tag{2.2}$$

where it is assumed that there are K active users and N available resource units in the cell; P_x is the total transmit power; $R_k, k = 1, \ldots, K$, is the data rate requirement for user k; if the nth resource unit is assigned to the kth user, $r_{k,n}$ is the achieved throughput, $P_{k,n}$ is the transmit power, and $\delta_{k,n}$ is an indicator variable. The computational complexity in such optimization problems is very high (NP-hard), thus sub-optimal methods are usually applied to significantly reduce the amount of computation while still achieving performance close to the global optimum (see [76, 124, 135]).

The authors in [48,75,77,118] extend the optimal resource allocation problems to the multi-cell scenario. Even though several approaches are presented to reduce the complexity, it is still a very complicated problem, particularly in a traffic-varying environment. A central controller is usually assumed for the coordination among a group of neighbouring sectors, which collects all the information from the users and

base stations through signalling. By solving the optimization problem, the resource allocation can be made optimal (or sub-optimal) in the group of sectors. These models give an insight into the upper bound for the scheduling gain, however, the actual implementation of these near-optimal mechanisms are typically not feasible or economical in real systems due to the need for a global optimizer, perfect channel knowledge and huge computational complexity.

Other advanced techniques, such as graph theory, can also be applied for ICI coordination during scheduling. Necker in [92, 93] constructs an *interference graph* (see Figure 2.6); if there is an edge between two users in this graph, it indicates the possibility of high interference and the scheduler in the system should avoid allocating the same frequency resource to them. Such a scheme can maintain a minimum required SINR throughout the coverage area. However, the procedure is done based on a global interference analysis and the knowledge of full system state information (of both base stations and users). Secondly, if the number of users in the area is large, the amount of computation required for constructing the interference graph may be prohibitive.

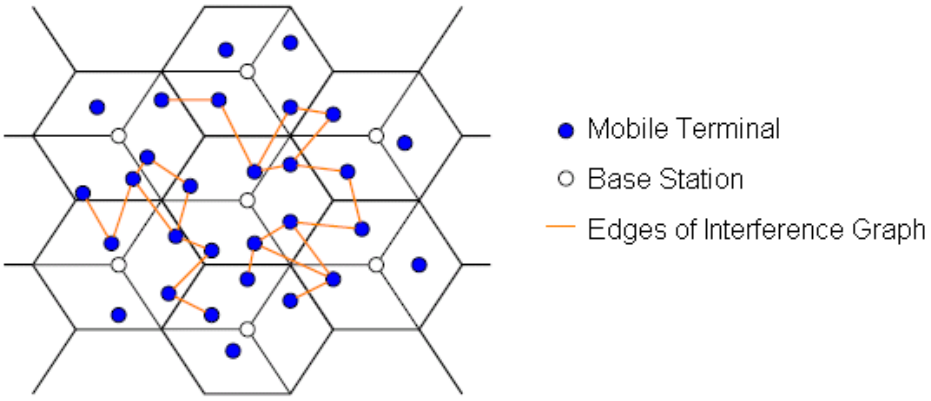

Figure 2.6: An example of interference graph [92, 93].

Some schemes therefore consider the coordination of transmissions among base stations in the time domain by switching off several base stations for a certain amount of time according to traffic in order to avoid the adverse effect of inter-cell interference (see Ahmed, Yanikomeroglu and Mahmoud [9], Bonald, Borst and Proutiére [24],

Liu and Virtamo [84]). Such time scheduling may perform well in fixed wireless systems (see Leung and Srivastava [74]), however, it is hard to implement in fully dynamic scenarios and it may seriously reduce the resource utilization.

With respect to the impact on the system design, we can see that in dynamic scheduling schemes, additional inter-base-station signalling is first required to obtain the system state information in the neighbouring sectors. Secondly, very tight time synchronization needs to be performed between the base stations. Thirdly, additional measurements should be done at the UEs and the state reporting may also require additional signalling between the UEs and base station. Lastly, the processing latency and computational complexity are other important considerations.

The support of coordination and scheduling in LTE specifications

In the LTE specifications, ICI coordination is performed through the X2 interface (see 3GPP specification [6]), which is a logical interface through which the base stations (eNBs) are interconnected with each other. A *load indication* procedure (see Figure 2.7) is used to transfer interference coordination information, namely *load information* [7], between the neighbouring eNBs. The load information indicates the interference level experienced by the originating sector on all resource blocks. Thus the receiving eNB may take such information into account for its scheduling policy.

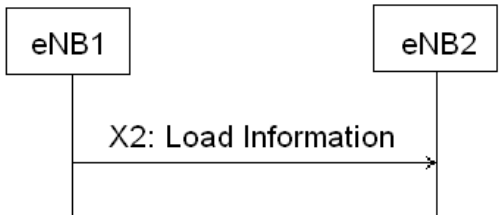

Figure 2.7: Load indication procedure in LTE.

The channel quality indicator (CQI) measurement reporting mechanism in LTE [6] also helps to support the dynamic scheduling on the downlink. As shown in Figure 2.8, the UE measures the reference signals to determine the received channel quality, which is essentially an SINR measurement; then the UE reports the CQI values of the downlink resources (on the basis of per-sub-band or wide-band) to the

eNB, which can be performed either periodically or aperiodically. Then the eNB can use the UE's reported CQI to schedule and allocate resources to the UE on the sub-bands with good channel conditions, and choose an appropriate modulation and coding scheme. To maximise the benefit from this mechanism, each UE should ideally provide accurate and timely CQI feedback. However, there are several practical issues with CQI reporting in real systems that detract from these objectives, such as reporting granularity, reporting delays and wireless errors (see Kwan and Leung [71], 3GPP contributions [116, 117], and Wunder *et al.* [145]).

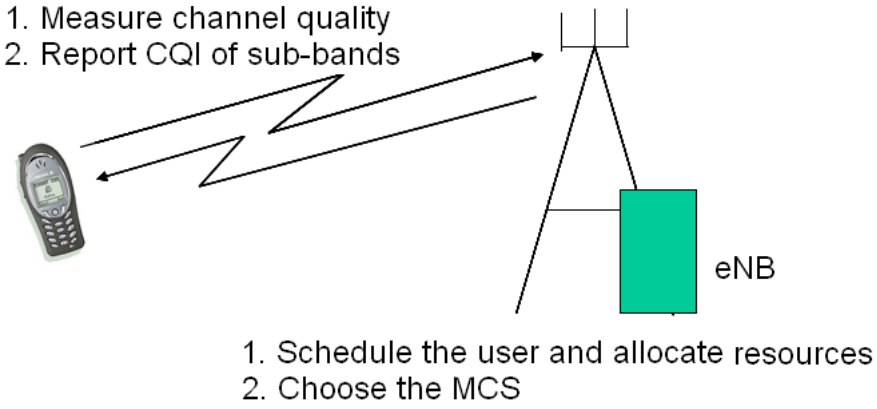

Figure 2.8: CQI reporting mechanism in LTE.

2.3.3 Other ICI mitigation techniques

There are several other techniques, which can be combined with the use of ICI averaging, cancellation and coordination for more effective ICI mitigation. For example, beamforming enables transmission in narrow beams in desired directions, resulting in very high gain and incurring less interference to other directions in the system (see Hu *et al.* [54], WINNER project [58], and Liu and Li [81]).

Adaptive antenna systems, which consist of multiple antenna elements with capability to optimize the transmission pattern automatically in response to the signal environment, are other important technologies used in wireless communication systems for ICI mitigation (see Bellofiore *et al.* [15], Karakayali, Foschini and Valenzuela [60], Liberti and Rappaport [79], and Murch and Letaief [90]). One way is to

use the smart antennas to take advantage of the spatial diversity effect at the transmitter (multiple-input single-output (MISO)), the receiver (single-input multiple-output (SIMO)), or both (multiple-input multiple-output (MIMO)) (see Andrews, Choi, and Heath [12] and Gesbert *et al.* [49]). In LTE and future broadband wireless systems, multiple antennas are available at both the base stations and the UEs. The UE can compare or combine the multiple receiving signals using Maximum Ratio Combining (MRC) or Interference Rejection Combining (IRC) approach (see Wei *et al.* [136]), and thus reduce the errors caused by the interference or other adverse effects of multi-path propagation.

Last but not least, dynamic power control (see Foschini and Miljanic [46], and Pischella and Belfiore [99]) and multi-hop transmission or relaying (see Nabar *et al.* [91], Ng and Yu [94], and Pabst *et al.* [97]) can also be used to deal with interference.

2.4 Performance metrics

The system-level performance studies of LTE systems that have appeared in the literature are typically simulation-based. To evaluate a proposed resource allocation or scheduling scheme, it is necessary to choose good performance metrics (see Tranter *et al.* [133]). There is no single universal performance metric for all systems. A good performance is defined quite differently for the user viewpoint compared to the base station viewpoint.

The metrics oriented towards measuring the base station performance include average sector throughput and sector edge throughput (see 3GPP contributions [115], Simonsson [125], and [141]). Sector throughput is the sum of the rates for all concurrent users in a sector, while the sector edge throughput is usually defined as the maximum data rate achieved by the worst 5 percent of users in the sector, and is obtained from the cumulative distribution function of the user data rate. These two metrics are considered the primary performance metrics for schedulers in wireless networks, where a persistent or full-buffer traffic model (assuming that there is never a shortage of data to transmit) and a relatively high system load are normally

assumed in the simulations. Persistent traffic is an artificial construct designed to fully exercise the lower protocol layers, with the aim of determining a bound on performance. However, these metrics do not shed much light on user-level performance.

User-level performance reflects the user experience provided by the system, such as the blocking probability, response time or session duration, or user throughput (see Fodor [43] and Wu, Williamson and Luo [144]). To estimate the user-level performance, a traffic model that accounts for the statistical nature of real traffic is needed. For non-persistent elastic traffic, Bonald and Proutiére in [26] show that the flow throughput (the ratio of the mean flow size to the mean flow duration, which is an estimate of the average throughput experienced by a user when downloading a file) is an appropriate user-level performance metric, which can be estimated analytically using a queueing theoretic approach. On the basis of flow throughput, derived metrics such as the cell capacity [21, 26] and the flow capacity [142, 143] (see Chapter 3) give important performance indicators of the provided service to the users.

2.5 Processor sharing models

To estimate the user-level performance analytically, we assume the following traffic model: the users are uniformly distributed in each cell, and the traffic of each user is elastic best-effort traffic. Each user only generates one elastic data flow at a time, where the flow size is independent and identically distributed (i.i.d.). We note that in our analysis, the user-level performance is not sensitive to flow size distributions, but depends only on the mean flow size.

The processor sharing (PS) discipline is an appropriate abstraction to model a base station on the downlink. The PS model has traditionally been used to represent resource time-sharing by jobs in a computer system (see Kleinrock [68,70], and Mitra and Weiss [88]) and bandwidth sharing on the Internet (see Ben Fredj *et al.* [47]). Recently, it has also been applied to the user-level performance analysis of wireless networks (see [26, 144], Bonald and Hegde [25], Borst [28], Cho *et al.* [32], Lei *et al.* [72] and Shankaranarayanan *et al.* [122, 123]).

Theoretically, the PS model assumes an ideal fluid resource allocation (infinitely divisible resources). In real systems such as LTE, of course, the resources have a finite minimum size in both the frequency and time domains. However, as shown in [28, 72], time-slotted systems like LTE can be represented by the PS model in continuous time if the duration of the time slot is much shorter than that of the data flow transmissions. In LTE, the scheduling is done on a 1 ms basis (see Section 2.2), which is relatively short compared to the time to transmit a typical data flow, which is in the order of seconds.

In wireless systems, the instantaneous transmission rate of a flow, denoted by r, depends in a complex way on the channel conditions and varies over time due to inter-cell interference, fast fading and user mobility. In our model, similar to Borst [28], we assume stationary users and assume that the fading is relatively fast compared to the flow transmission. The effects of interference and fast fading can be taken into account by letting $r \triangleq RY$, where $R = E[r]$ is the uncontended time-average rate achieved in the absence of other active flows when the base station allocates all its time and frequency resources to the user to send this flow. The random variable Y represents the fast fading effects and is independently and identically distributed with unit mean.

Figure 2.9 depicts the queueing model that we use for the base station. The elastic data flows are generated by the users and enter the queue, which can be characterised according to their uncontended time-average transmission rates and classified into K classes. The service discipline at the server is processor sharing, so that multiple flows share the service capacity of the sector. A flow which has finished its transmission will leave the queue. The type of processor sharing depends on the scheduling algorithms used at the base station.

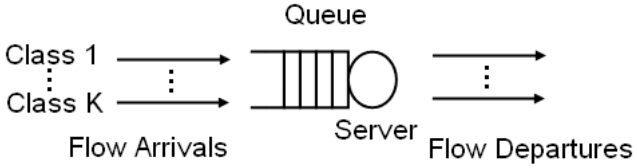

Figure 2.9: A single-server queue model.

There are several basic concepts in the queueing system, which will be used in the analysis in Chapters 3 and 4. The rate at which the flows enter the queue is called the arrival rate, denoted by λ. In absence of other flows, the average transmission time of a flow is called the mean service time, denoted by $1/\mu$. The offered traffic is defined by

$$\rho = \frac{\lambda}{\mu}. \tag{2.3}$$

The average time from the moment a flow arrives until its service is complete is called the mean duration of a flow (also called mean response time or mean sojourn time), denoted by $E[T]$. The mean number of active flows in the queue (including those in service) is called the mean queue size and denoted by $E[x]$. The utilization, denoted by U, is the proportion of time that the server is busy (when there is at least one active flow in the system). Let p_0 be the probability that the system is empty, we have

$$U = 1 - p_0. \tag{2.4}$$

An important theorem for the relationship of the arrival rate, mean duration and mean queue size is known as Little's Formula (see Kleinrock [69]), which is formulated as

$$E[x] = \lambda E[T]. \tag{2.5}$$

2.5.1 RR scheduling and EPS

Round-robin (RR) is the simplest and best-known scheduling algorithm, which assigns frequency/time resources to each flow in equal portions and in circular order, handling all flows in a fair manner. If RR scheduling is applied at the base station, the server in Figure 2.9 can be modelled as an egalitarian processor sharing (EPS) queue [26, 144]. If $x(x > 0)$ flows are active at the base station at time t, then each flow obtains during $[t, t + \Delta t]$ an amount of service equal to $\Delta t/x$. Thus a flow with time-average rate R is served with the rate R/x.

2.5.2 PF scheduling and GPS

In a wireless system with many users whose channels are fading independently, there are likely to be some users experiencing good channels and some users experiencing poor channels at any given time. Hence, an opportunistic scheduler can take advantage of the multi-user diversity by allocating the resources to the users with the best channels to improve the performance (see the book by Goldsmith [50], Svedman *et al.* [132], and Pokhariyal *et al.* [102, 103]). However, a fairness problem may be raised. The users with poor average SINRs, such as the users at the sector edge area, scarcely have the best channel and thus are rarely allocated the resources to transmit, which leads to unfairness between the users.

To extract multi-user diversity gain while maintaining a level of long-term fairness among active users, proportional fair (PF) scheduling can be applied (see Beh, Armour and Doufexi [14], Bender *et al.* [16], Kela *et al.* [62], Massoulié and Roberts [86], and Wengerter, Ohlhorst, and Elbwart [137]). The PF scheduler is widely implemented in many broadband wireless systems, such as HSPA [39], 1xEV-DO [18], and LTE [14, 62].

The basic principle is that each frequency resource unit j is scheduled to user k at time t if

$$k = \underset{i}{\operatorname{argmax}} \left(\frac{R_{i,j}(t)}{\tilde{R}_i(t)} \right), \tag{2.6}$$

where $R_{i,j}(t)$ denotes the predicted instantaneous supportable data rate if resource unit j is allocated to user i. $\tilde{R}_i(t)$ represents the estimation of the past average throughput of user i, which is calculated by

$$\tilde{R}_i(t) = \begin{cases} (1 - \frac{1}{t_c})\tilde{R}_i(t-1) + \frac{1}{t_c}R_i(t) & \text{if user } i \text{ is served,} \\ (1 - \frac{1}{t_c})\tilde{R}_i(t-1) & \text{otherwise,} \end{cases} \tag{2.7}$$

where t_c is the window size of the average throughput, and $R_i(t)$ is the actual transmission rate.

Based on the assumption that the time scale of the data flow transmission is much longer than that of the rate fluctuations, the base station can be modelled as a generalized processor sharing (GPS) queue (see Cohen [36]), which can capture

the multi-user diversity gain of PF scheduling (see Bonald, Borst and Proutiére [23] and Borst [28]). The server still assigns each flow a fraction $1/x$ of the service capacity when there are $x(x > 0)$ flows in the queue. But a flow with uncontended time-average rate R is served with time-average rate $RG(x)/x$, where $G(x)$ is the scheduling gain when there are x flows at the base station.

For the RR case, we have $G(x) \equiv 1, x = 1, 2, \ldots$, so that the GPS queue is specialized to the EPS queue. For the PF case, the function $G(x)$ is interpreted as a gain accounting for the improvements from PF scheduling over RR scheduling, which depends on the number of active users scheduled at one time and the available frequency bandwidth. In particular, $G(x)$, with $G(1) = 1$, is increasing in x and tends to some finite limit as $x \to \infty$, while the ratio $G(x)/x$ is decreasing in x. In general, the gain $G(x)$ is a function of $Y_1, \ldots, Y_x$ and is difficult to characterise analytically. An alternative approach that we employ in our work is to use system-level simulations to obtain the $G(x)$ values (see Chapter 3).

2.5.3 Open network and closed network

Depending on the user populations in the sector, the system can be modelled with either an open queueing network (see Bonald and Proutiére [26]) or a closed queueing network (see Berger and Kogan [17] and Liu and Virtamo in [83]). If we assume ideally that there is an infinite user population in the sector, the sector system can be modelled by an open queueing network like Figure 2.9, where the data flows arrive as a Poisson process. This assumption gives a reasonable approximation of the real system, and it results in a compact analytical expression for user-level performance that provides insight for parameter sensitivity (see Chapter 3).

A finite user population is closer to reality and therefore has greater practical relevance. For each user, only one elastic flow is generated at one time, and after the completion of the flow, the user is assumed to go into a thinking state, after which a new data flow is generated, and so on (as depicted in Figure 2.10). The flow sizes and the thinking times are independent and identically distributed. For this case, the sector system can be modelled as a two-node closed queueing network with K classes of user flows (see Figure 2.11). There is a fixed population of users in the

system. Node 1 is a processor sharing queue (EPS or GPS) that models the base station, while node 0 is an infinite server (IS) queue that models the users in the thinking state.

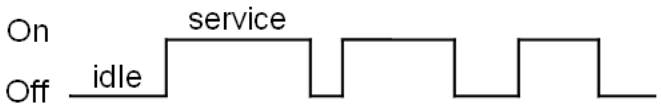

Figure 2.10: On-off elastic traffic model.

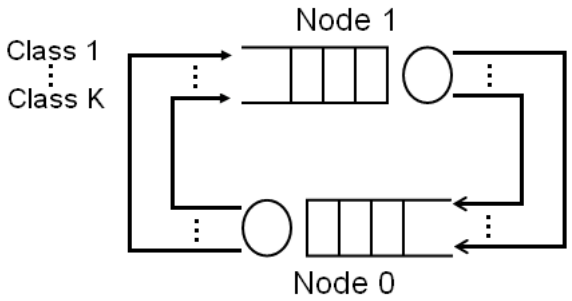

Figure 2.11: A two-node closed network model.

Most studies assume that node 1 is an EPS queue, and the theory of product form networks (see Baskett *et al.* [13]) is applied in the analysis of the user-level performance. The major difficulty is that the computational complexity of a direct calculation (especially for the normalisation constant, see Chapter 4) grows exponentially with the number of classes and the number of users. Many excellent algorithms for computing the user-level performance have been developed, such as the convolution algorithm (see Buzen [30]), the mean value analysis (MVA) algorithm (see Reiser and Lavenberg [120]), the recursion by chain algorithm (RECAL) (see Conway and Georganas [37]), the mean value analysis by chain (MVAC) algorithm (see Conway [38]), the distribution analysis by chain (DAC) algorithm (see De Souza E Silva and Lavenberg [40]), and the numerical inversion algorithm (see Choudhury , Leung and Whitt [34]). In our problem, the queueing network has few nodes and many classes, and so we apply the RECAL and MVAC algorithms, which are well-suited to this setting (see Chapter 4).

If node 1 is a GPS queue (for the PF case), the exact solution becomes more complicated and may be intractable for a large number of classes (see Chapter 4). Fortunately it is sometimes possible to obtain asymptotic results in the normal operating region (see Berger and Kogan [17], Kelly [63], McKenna, Mitra and Ramakrishnan [87], and Pittel [100]).

Other papers (see Buzen and Goldberg [31], and Protopapas [104]) use an open queueing network to approximate a closed queueing network. A more accurate open queueing network approximation, called the fixed-population-mean (FPM) method, is introduced by Whitt [138]. The aim of the approximation is to achieve the same mean number of active users as that in node 1 of the closed queueing network. The application of the FPM method to our problem will be discussed in detail in Chapter 4.

Chapter 3

Flow capacity evaluation with an infinite user population model

3.1 Introduction

In this and the next chapter, we use a flow-level queueing theoretic approach to analyse and compare the downlink performance of reuse-1, reuse-3 and static reuse partitioning schemes, namely FFR and SFR. We develop model variants for infinite and finite user populations to characterise the user-level performance for elastic traffic, and define a new *flow capacity* metric as a basis for comparison of different reuse schemes. In our models, we employ a hybrid simulation/analysis approach, where a rate distribution obtained via simulation is used as input to the queueing model. We do this because a rate distribution generated by analytical means would lack realism, yet a purely simulation-based approach to estimate flow capacities would be computationally prohibitive.

In this chapter, we assume that there is an infinite user population in the sector and that the data flows arrive as a Poisson process. This model gives a reasonable approximation of the real system, and is more tractable than a finite user population model. We define the *flow capacity* in this chapter as the maximum traffic intensity that can be supported by a base station sector while satisfying a minimum requirement on the flow throughput. In addition to modelling the baseline round-robin (RR) scheduler, we show how the gains of proportional fair (PF) scheduling can be incorporated into the analytical model. Furthermore we develop a methodology to account for the effect of interference from neighbouring base stations with an arbitrary level of loading.

Many studies (see Choi, Kim and Bahk [33], WINNER Project [57], Rahman,

Yanikomeroglu and Wong [119], Simonsson [125], Xiang, Luo and Hartmann [146] and 3GPP contributions such as [106, 111, 112, 115]) have investigated the performance of static reuse partitioning schemes in OFDMA networks using simulation. These studies assume persistent (infinite backlog) traffic and the primary performance metrics sought are the average sector throughput and the sector edge user throughput. It is observed that reuse partitioning leads to an improvement in the sector edge throughput compared to reuse-1, but a reduction in the sector throughput, and that the tradeoff point can be adjusted by moving the boundary of the frequency partition. However, the key question of whether the improvement in the edge performance justifies the reduction in the sector throughput is left unanswered.

Recently, Elayoubi and his collaborators [41, 42] present a novel analytical performance model for static reuse partitioning, which they use to compare reuse-1, reuse-3, FFR and SFR. The base station is modelled as an Erlang loss system; calls arrive according to a Poisson distribution and each call requires one circuit (one OFDMA sub-channel). The service time in the Erlang loss system is an average service time that accounts for the rate variation due to adaptive modulation and coding (AMC), its sensitivity to the location within the sector, and the interference from arbitrary loaded neighbouring sectors. Due to its loss system basis, the model in [41,42] is restricted to fixed bandwidth services, and cannot capture the behaviour of elastic services whose data rate adjusts to match the available capacity, and which are more representative of internet services today.

The seminal works by Bonald and Proutiére [26] and Borst [28] introduce flow-level models utilizing multi-class processor sharing (PS) queues to analyse user-level performance. These models employ an elastic traffic model and yield the per-class flow throughput. Bonald and his collaborators [25, 26] also introduce the notion of *cell capacity*, which they defind as the limiting traffic intensity for which the base station remains non-saturated. As such, the cell capacity is a fundamental stability limit. Borst [28] shows that the effect of PF scheduling can be approximately captured by a generalized processor sharing (GPS) queue, but does not show how to solve this system for the flow throughputs. Lei *et al.* in [72] characterise analytically the scheduling gain of PF scheduling in OFDMA systems for the idealized case of

Rayleigh fading and a linear SINR-to-Rate mapping, but only evaluate the user-level performance through simulations. The analysis of [26, 28, 72] consider primarily the single-cell scenario. A subsequent work by Bonald *et al.* [22] develops bounds and approximations for the more realistic multi-cell scenario with fractional loading in the interfering sectors.

In this chapter, we extend the flow-level methodology of [26, 28] to the setting of reuse partitioning, which requires definition of separate PS queueing systems for the sector centre and sector edge users. In the process, we significantly extend the methodology itself in several directions that are not specific to reuse partitioning, but have general applicability. The notion of flow capacity generalizes the cell capacity to the non-saturated regime, which coincides with the cell capacity when the minimum flow throughput requirement tends to zero. For the RR scheduler, we derive simple, insightful expressions for the flow capacity, which reveals that enhancement of the sector edge rate can significantly improve the capacity. For the PF case, we implement a computationally efficient algorithm to solve for the flow throughputs by exploiting a fast convergence property of the multi-user diversity gain. Furthermore, we consider the multi-cell scenario with fractional loading, and show how fractional loading in the interfering sectors can approximately be taken into account. We also develop an approximation which uses an iterative method to find the flow capacity when the reference sector has the same loading as the interfering sectors.

The rest of this chapter is organized as follows. In Section 3.2, we describe the basic assumptions and parameters used in our analysis. We introduce an approximation for fractional loading in the interfering sectors, describe the traffic model and the inputs of the analysis required for specific schemes, and present a hybrid simulation/analysis approach. In the reuse partitioning schemes, we introduce an important simplification to make the analysis tractable.

In Section 3.3, we describe an analytical model for the flow capacities for schemes without reuse partitioning. We first formulate the problem of finding the flow capacity for an arbitrary load in the interfering sectors. Then we present two different approaches to find the flow capacity in a homogeneous load network.

In Section 3.4, we provide a method to find the flow capacity for reuse partitioning

schemes. We find that the parameter used to differentiate the centre/edge user has an important effect on the capacity. Our method enables the determination of the optimal partition between the centre band and the edge band.

In Section 3.5, we explore some key aspects of the methodology, then show the simulation validation of the hybrid simulation/analysis approach. Finally, we present numerical experiments to illustrate the capacities of different schemes.

To show that our framework for evaluating capacity has more general applicability, in Section 3.6 we present an example that explores the capacity benefits of MIMO schemes.

3.2 Basic assumptions and parameters

In this section, the basic assumptions and parameters for the analysis are illustrated.

3.2.1 Fractional loading in the interfering sectors

To analyse the user-level performance in the multi-cell scenario, technically we need to model each sector by a queueing system (as shown in Figure 3.1(a)), which will make the problem analytically intractable and extremely time-consuming to simulate since the service rate in one sector is affected by the states in the surrounding queueing systems.

Therefore, we apply a simplification to approximately model fractional loading in the interfering sectors. At any arbitrary instant, we model the active/inactive state of a base station sector using a Bernoulli random variable with success probability $U_I, 0 \leq U_I \leq 1$. We call U_I the load, and assume that all interfering sectors have the same load. Then we model the reference sector by a queueing system and find the flow capacity when the reference sector is affected by the load U_I in the interfering sectors (see Figure 3.1(b)).

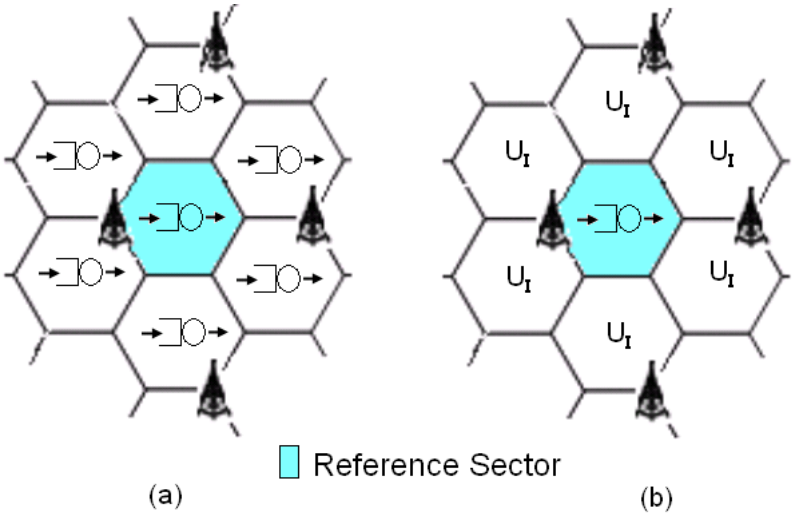

Figure 3.1: An approximation for fractional loading in a multi-cell scenario.

3.2.2 Traffic model

In our analysis, we assume that there is an infinite user population in the reference sector and that the users are uniformly distributed. The traffic of each user is elastic best-effort, and the traffic flows are generated as a Poisson process of rate λ flows/sec. We assume a single-category traffic scenario, where the flow sizes are independent and identically distributed with a mean σ bits. Note that our model could also be extended to the multi-category traffic case when there are different classes with different flow size distributions and means.

3.2.3 Reuse schemes and reuse partitioning schemes

We consider the downlink of a generic OFDMA wireless system with N resource units. A 1x2 SIMO channel is assumed where there is one transmitting antenna at the base station and two receiving antennas at the mobile terminal. The frequency and power configurations in reuse-1, reuse-3, FFR and SFR are illustrated in Figure 2.5. In reuse-1, the number of available resource units in each sector is N, while in reuse-3 it is $N/3$. In both FFR and SFR, the edge band size $n \in \{1, \ldots, \lfloor \frac{N}{3} \rfloor\}$ is an important degree-of-freedom, which gives rise to different sub-schemes, denoted by FFR-n and SFR-n. (The function $\lfloor x \rfloor$ rounds the argument to the nearest integer

less than or equal to x.)

In the FFR-n (and SFR-n) scheme, we define an SINR threshold s_n to differentiate the centre/edge users, and deem a user whose SINR on the centre band is less than s_n to be an edge user. We let s_n equal the p_n-th quantile of the SINR distribution on the centre band (see Figure 3.2), where p_n is selected to maximise the cell capacity (see Section 3.4.1). Furthermore, for tractability, we impose the restriction that centre users can only access centre band resource units; in a real system, centre users should also be permitted to access the unallocated edge band resources. Our simplified assumption should have negligible impact on the capacity when the load in the reference sector is high because there is likely to be at least one active edge user in the system. When the load is low, however, this assumption will lead to an under-estimate of the true flow capacity.

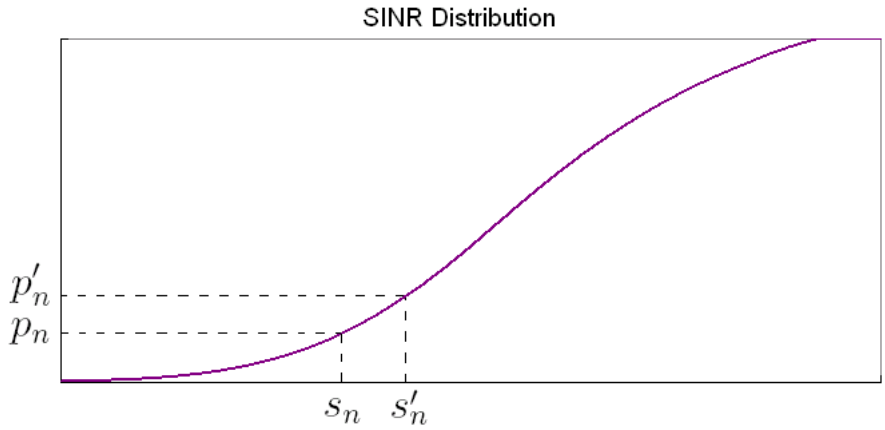

Figure 3.2: An SINR threshold for FFR/SFR to differentiate the centre/edge users.

3.2.4 Hybrid simulation/analysis approach

In our model, we employ a hybrid simulation/analysis approach to obtain the user-level performance (we will verify this approach in Section 3.5.2), the basic steps of which are illustrated in Figure 3.3. There are essentially four steps in this approach: step 1 is to obtain an SINR distribution; step 2 is to use an appropriate SINR-to-Rate mapping function to obtain a rate distribution; step 3 is to discretize the

continuous distribution; step 4 is to use a queueing theoretic approach to analyse the user-level performance, namely the per-class flow throughputs and the flow capacity. We perform the first three steps by simulation (see Section 3.5.1) while the last step is carried out using the analytical model.

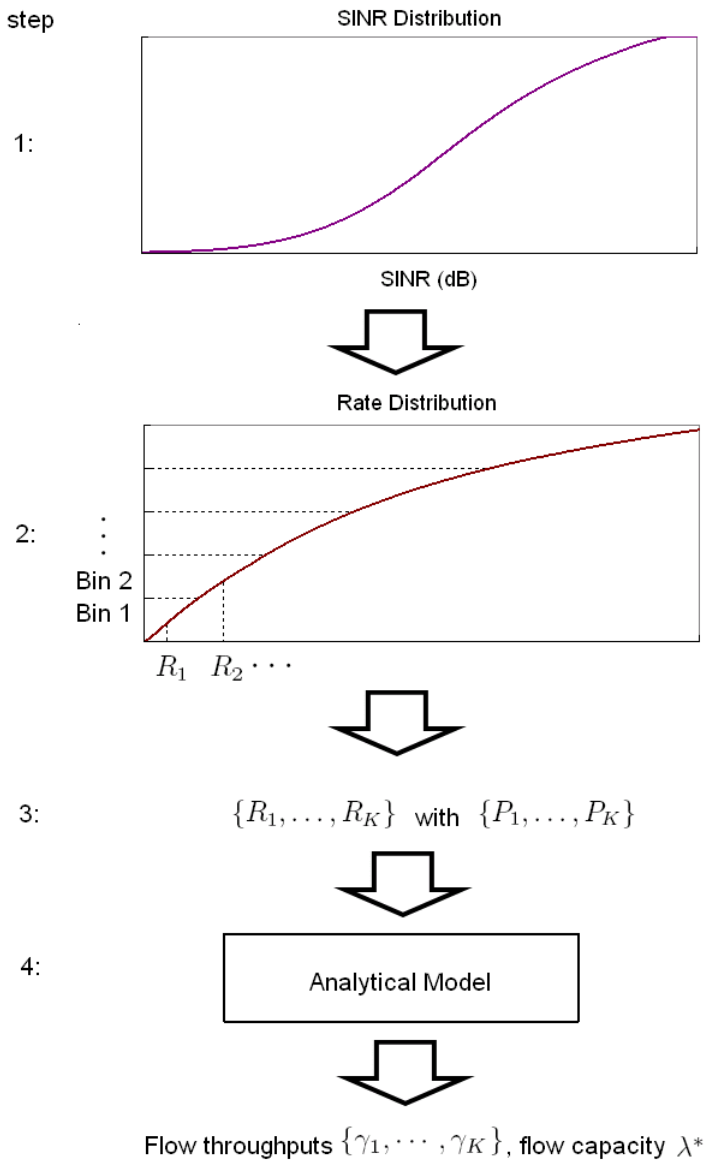

Figure 3.3: The hybrid simulation/analysis approach to obtain the flow capacity.

3.2.5 Inputs for the analytical model

In Section 2.5, we explained that we model the base station as a GPS queue (see Cohen [36]) to capture the multi-user diversity gain of PF scheduling (see Bonald, Borst and Proutiére [23] and Borst [28]), which can be specialized to the EPS queue to model RR scheduling. If x flows are active at time t, then during $[t, t + \Delta t]$ each flow obtains an amount of service $\Delta t/x$. A flow with uncontended time-average rate R is served with rate $RG(x)/x$, where $G(x)$ is the scheduling gain when there are x flows at the base station.

An input to the analytical model is a set of time-average rates (see Figure 3.3), which we obtain by discretizing a rate distribution estimated via simulation. In the reuse-1/reuse-3 schemes, we assume that there are K classes of flows characterised by discrete time-average rates $\{R_1, \ldots, R_K\}$, with corresponding probabilities $\{P_1, \ldots, P_K\}$, where P_k is the probability that an arbitrary flow belongs to class-k and $\sum_{k=1}^{K} P_k = 1$. The rates R_k (and the performance metrics introduced in Section 3.3.1 such as γ_k, ρ_k, and λ^*) depend on the load U_I in the interfering sectors, so technically we should write $R_k(U_I), k = 1, \ldots, K$. However, to avoid cluttering the notation, we will only explicitly show the functional dependence on U_I when it is helpful for understanding.

In FFR-n/SFR-n schemes, the base station can be modelled as two separate GPS queues by introducing the resource access restrictions in Section 3.2.3. We define two discrete sets of time-average rates, one for the centre band, $\{R_{c1}, \ldots, R_{cK_c}\}$ with probabilities $\{P_{c1}, \ldots, P_{cK_c}\}$ where $\sum_{k=1}^{K_c} P_{ck} = 1$, and the other for the edge band, $\{R_{e1}, \ldots, R_{eK_e}\}$ with probabilities $\{P_{e1}, \ldots, P_{eK_e}\}$ where $\sum_{k=1}^{K_e} P_{ek} = 1$.

3.2.6 Mathematical notation

The important mathematical notation that we use in our analysis is summarized in Table 3.1.

Table 3.1: Mathematical notation

Symbol	Description	Equation where symbol first appears
N	Total number of resource units in the system	–
U_I	Load in all interfering sectors	–
U_r	Load of the reference sector	(3.17)
σ	Mean flow size	(3.1)
λ	Arrival rate in the reference sector	(3.4)
λ_{sat}	Cell capacity in the reference sector	(3.12)
δ	Minimum flow throughput requirement	(3.7)
β	Proportion of users satisfying the throughput requirement	(3.8)
K	Number of classes of flows	(3.1)
R_k	Time-average rate of class-k flows	(3.3)
P_k	Probability that an arbitrary flow belongs to class-k	(3.4)
$\overline{R}$	Harmonic mean of a rate distribution	(3.6)
ρ_k	Offered traffic of class-k	(3.4)
ρ	Total offered traffic to the reference sector	(3.5)
T_k	Flow duration of one class-k flow	(3.1)
x_k	Number of active class-k users	(3.9)
γ_k	Flow throughput of class-k	(3.1)
g	Limit of PF scheduling gain	(3.25)
s_n	SINR threshold to differentiate the centre/edge users in FFR/SFR	–
p_n	p_n-th quantile of the SINR distribution	(3.39)

3.3 System model for standard reuse schemes

In this section, we describe a model for the capacity analysis of schemes without reuse partitioning, for example, reuse-1 and reuse-3 schemes. We first present a method to analyse the flow capacity for an arbitrary load U_I in all the interfering sectors, then provide an approach to find the flow capacity in a homogeneous load network when the reference sector has the same load as the interfering sectors.

3.3.1 Problem formulation for an arbitrary load U_I in interfering sectors

For a flow arrival rate λ in the reference sector, we can calculate the flow throughput of each class using

$$\gamma_k := \frac{\sigma}{E[T_k]}, \quad k = 1, \ldots, K, \tag{3.1}$$

where T_k is the flow duration for class-k. Without loss of generality, we assume that the flow throughputs are arranged in decreasing order (breaking any ties randomly), such that

$$\gamma_1 \geq \cdots \geq \gamma_K. \tag{3.2}$$

For standard reuse schemes, this will be the same index order as the list of rates in decreasing order,

$$R_1 \geq \cdots \geq R_K. \tag{3.3}$$

For the class-k flows, the arrival rate λ_k, the mean service time $1/\mu_k$ and the offered traffic ρ_k are given by

$$\lambda_k = P_k \lambda, \quad \frac{1}{\mu_k} = \frac{\sigma}{R_k}, \quad \rho_k = \frac{\lambda_k}{\mu_k} = \frac{\lambda_k \sigma}{R_k}. \tag{3.4}$$

The total offered traffic to the reference sector is

$$\rho = \sum_{k=1}^{K} \rho_k = \frac{\lambda \sigma}{\overline{R}}, \tag{3.5}$$

where $\overline{R}$ is the harmonic mean of the rate distribution, and is given by

$$\overline{R} = \left(\sum_{k=1}^{K} \frac{P_k}{R_k} \right)^{-1}. \tag{3.6}$$

Therefore, the problem of finding the flow capacity in the reference sector is formulated as

$$\lambda^* = \max \quad \lambda,$$

$$\text{such that} \qquad \gamma_L \geq \delta, \tag{3.7}$$

where δ is a pre-defined minimum flow throughput requirement. The value chosen for δ depends on the minimum user experience that the operator wants to provide. In addition, we define

$$L = \min \left\{ n : \sum_{k=1}^{n} P_k \geq \beta \right\}, \tag{3.8}$$

where β is a design parameter implying user satisfaction, defined as the fraction of the users that need to satisfy the throughput requirement. For example, if $\beta = 0.9$ then 90 percent of the users need to satisfy the requirement. The reason for introducing β is that requiring every user to meet the throughput requirement is usually too strict, and will result in a very low flow capacity. Since there are always some users experiencing poor channel conditions, it is sufficient that most users can satisfy the requirement.

The flow throughputs $\gamma_k, k = 1, \ldots, K$, are obtained by modelling the base station with a GPS queue. In the following, we analyse the flow throughputs and flow capacities for RR and PF scheduling. The state of the system at a given time is defined by the vector

$$\mathbf{x} = (x_1, \ldots, x_K), \tag{3.9}$$

where $x_k, k = 1, \ldots, K$, is the number of active class-k users, and we define

$$x = \sum_{k=1}^{K} x_k. \tag{3.10}$$

Round-robin scheduling

For RR scheduling, the flow with time-average rate R is served at rate R/x when there are x flows in the sector. Thus we have $G(x) \equiv 1$ and the GPS model reduces to EPS (see Section 2.5). For the queue to be stable, we must have

$$\rho < 1, \tag{3.11}$$

which, together with (3.5), leads to the cell capacity (see Bonald and his collaborators [25, 26]):

$$\lambda_{sat} = \frac{\overline{R}}{\sigma}. \tag{3.12}$$

The stationary distribution of the system is

$$\pi(x_1, \ldots, x_K) = H_{RR}^{-1} x! \Pi_{k=1}^{K} \frac{\rho_k^{x_k}}{x_k!}, \tag{3.13}$$

where H_{RR} is the normalisation constant. Since

$$\Sigma_{x_1=0}^{\infty} \cdots \Sigma_{x_K=0}^{\infty} \pi(x_1, \ldots, x_K) = 1, \tag{3.14}$$

we have

$$\begin{aligned} H_{RR} &= \Sigma_{x_1=0}^{\infty} \cdots \Sigma_{x_K=0}^{\infty} \left[x! \Pi_{k=1}^{K} \frac{\rho_k^{x_k}}{x_k!} \right] \\ &= \Sigma_{x=0}^{\infty} (\rho_1 + \cdots + \rho_K)^x \\ &= \frac{1}{1-\rho}. \end{aligned} \tag{3.15}$$

The probability that there are no active flows in the system is

$$\pi(0, \ldots, 0) = H_{RR}^{-1} = 1 - \rho. \tag{3.16}$$

Therefore from (2.4), we can obtain the utilization in the reference sector (the proportion of time that the system is busy), which is given by

$$U_r = 1 - \pi(0, \ldots, 0) = \rho. \tag{3.17}$$

We interpret U_r as the load of the reference sector. The mean number of active class-k users in the queue is

$$\begin{aligned} E[x_k] &= \Sigma_{x_1=0}^{\infty} \cdots \Sigma_{x_K=0}^{\infty} x_k \pi(x_1, \ldots, x_K) \\ &= H_{RR}^{-1} \rho_k \frac{\partial H_{RR}}{\partial \rho_k} \end{aligned}$$

$$= H_{RR}^{-1}\rho_k\frac{\partial H_{RR}}{\partial \rho}$$
$$= \frac{\rho_k}{1-\rho}. \tag{3.18}$$

From (3.1) and applying Little's Formula (see Section 2.5 or Kleinrock [69]), we can obtain the per-class flow throughput

$$\gamma_k = \frac{\sigma}{E[T_k]} = \frac{\lambda_k\sigma}{E[x_k]} = R_k(1-\rho) = R_k(1-\frac{\lambda\sigma}{\overline{R}}). \tag{3.19}$$

By rearranging (3.19), we can solve (3.7) and obtain the flow capacity for RR scheduling:

$$\lambda^* = (1-\frac{\delta}{R_L})\frac{\overline{R}}{\sigma}. \tag{3.20}$$

Equation (3.20) reveals the parameters that can affect the capacity: the flow throughput requirement, δ; the harmonic mean of the rate distribution, $\overline{R}$; and the worst-case rate, R_L, that meets the user satisfaction. The first parameter is determined by the system designer, the second depends on the equipment capabilities and environment, and the third depends on both sets of factors.

It is instructive to visualize the graph of λ^* versus δ, as shown in Figure 3.4. The graph is a straight line with y-intercept $\overline{R}/\sigma$ (which is λ_{sat}) and x-intercept R_L. If we reduce β, R_L increases, and the capacity region increases. If, hypothetically, we apply some enhancement technique to improve the low rates such as R_K (assume that R_L is improved too), both R_L and $\overline{R}$ increase and the capacity is significantly improved. If, on the other hand, we improve the high rates such as R_1, then only $\overline{R}$ is increased, and only slightly since a harmonic mean is less sensitive to the larger sample values, and so the capacity is not improved very much. Therefore, the best way to increase the capacity is to improve the low rates, which we shall also refer to as the cell edge rates.

Proportional fair scheduling

PF scheduling in the time and frequency domains can provide significant spectral efficiency gains over RR scheduling since the resources can be allocated to users when

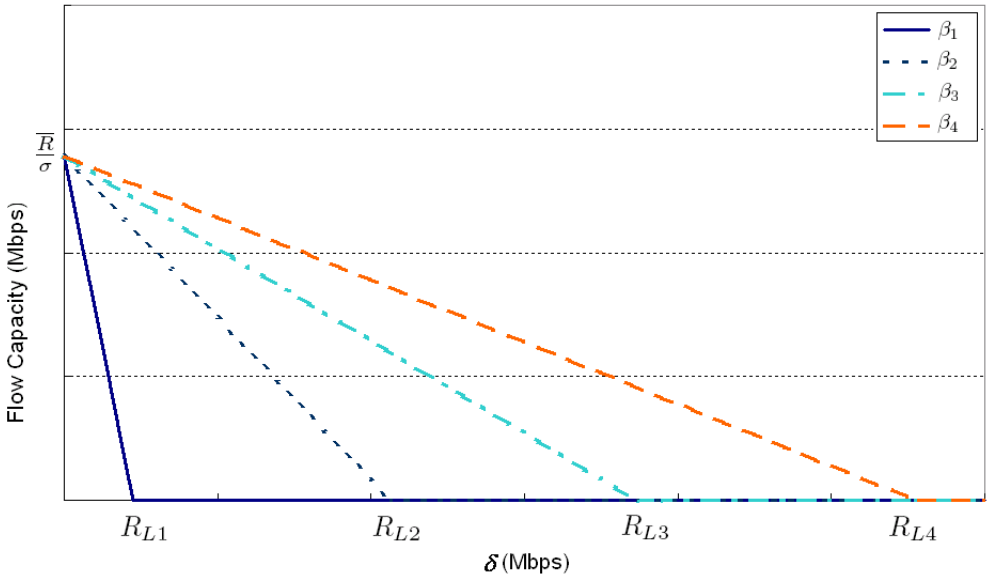

Figure 3.4: An example of the flow capacity for RR scheduling, $\beta_1 > \beta_2 > \beta_3 > \beta_4$.

they experience up-fades. The function $G(x)$ is interpreted as a gain accounting for the improvements from PF scheduling, which depends on the number of active users scheduled at one time, and the available frequency bandwidth. In general, $G(x)$ is difficult to characterise analytically, so we use simulation to obtain values (see Section 3.5.1).

For convenience, we denote

$$\phi(x) = \begin{cases} 1 & x = 0 \\ \frac{x!}{\Pi_{i=1}^{x} G(i)} & x = 1, 2, \ldots \end{cases}. \tag{3.21}$$

The stationary distribution of the system is

$$\pi(x_1, \ldots, x_K) = H_{PF}^{-1} \phi(x) \Pi_{k=1}^{K} \frac{\rho_k^{x_k}}{x_k!}, \tag{3.22}$$

where H_{PF} is the normalisation constant, given by

$$H_{PF} = \Sigma_{x_1=0}^{\infty} \cdots \Sigma_{x_K=0}^{\infty} \left[\phi(x) \Pi_{k=1}^{K} \frac{\rho_k^{x_k}}{x_k!} \right]. \tag{3.23}$$

The utilization in the reference sector is

$$U_r = 1 - \pi(0, \ldots, 0) = 1 - H_{PF}^{-1}. \tag{3.24}$$

The difficulty with PF scheduling is that for general $G(x)$, there is no easy way to solve for the flow throughputs. Fortunately, we find that for commonly used channel models, $G(x)$ converges quickly to some limit $g (g > 1)$ for $x > x_0$, as shown by our system-level simulation results in Section 3.5.1. This fast convergence property is also found by some other authors (see Beh *et al.* [14], Classson *et al.* [35], and Wengerter *et al.* [137]). Consequently, the stability condition for PF scheduling (see Borst [28]) is

$$\rho < g, \tag{3.25}$$

which leads to the definition of cell capacity (see [25, 26]):

$$\lambda_{sat} = \frac{g\overline{R}}{\sigma}. \tag{3.26}$$

We can write $\prod_{i=1}^{x} G(i) = Cg^x$ for $x > x_0$, where

$$C = g^{-x_0} \prod_{i=1}^{x_0} G(i). \tag{3.27}$$

For $x > 0$ we have

$$\phi(x) = \begin{cases} \frac{x!}{Cg^x} & x > x_0 \\ \frac{x!}{Cg^x} - x!D(x) & x \leq x_0 \end{cases}, \tag{3.28}$$

where

$$D(x) = \begin{cases} \frac{1}{Cg^x} - \frac{1}{\prod_{i=1}^{x} G(i)} & x > 0 \\ \frac{1}{C} - 1 & x = 0 \end{cases}. \tag{3.29}$$

We can then derive

$$\begin{aligned} H_{PF} &= \sum_{x=0}^{\infty} \frac{\rho^x}{Cg^x} - \sum_{x=0}^{x_0} D(x)\rho^x \\ &= \frac{1}{C\left(1 - \frac{\rho}{g}\right)} - \Sigma_{x=0}^{x_0} D(x)\rho^x. \end{aligned} \tag{3.30}$$

Therefore, the mean number of active class-k users is

$$\begin{aligned} E[x_k] &= \Sigma_{x_1=0}^{\infty} \cdots \Sigma_{x_K=0}^{\infty} x_k \pi(x_1, \ldots, x_K) \\ &= \frac{\rho_k}{H_{PF}} \frac{\partial H_{PF}}{\partial \rho_k} \\ &= \frac{\rho_k}{H_{PF}} \frac{\partial H_{PF}}{\partial \rho} \\ &= \rho_k \frac{\frac{1}{Cg(1-\frac{\rho}{g})^2} - \Sigma_{x=1}^{x_0} D(x) x \rho^{x-1}}{\frac{1}{C\left(1-\frac{\rho}{g}\right)} - \Sigma_{x=0}^{x_0} D(x) \rho^x}. \end{aligned} \tag{3.31}$$

From (3.1) and Little's Formula, the flow throughput of class-k users is

$$\gamma_k = R_k \frac{\frac{1}{C\left(1-\frac{\rho}{g}\right)} - \Sigma_{x=0}^{x_0} D(x) \rho^x}{\frac{1}{Cg\left(1-\frac{\rho}{g}\right)^2} - \Sigma_{x=1}^{x_0} D(x) x \rho^{x-1}}. \tag{3.32}$$

Since x_0 is typically not large, the sums in (3.32) can be easily evaluated.

It is not possible to obtain a simple expression for the flow capacity for PF scheduling like (3.20) for RR scheduling. (However, we can find a linear approximation, discussed in Section 3.5.1.) Therefore, we propose an iterative method below.

The task is to find the flow capacity λ^*, given the minimum flow throughput requirement δ and the user satisfaction proportion β. We let ε_1 be the error tolerance and apply a bisection search between 0 and λ_{sat}, as formalised in Algorithm 3.1. Since there is no feasible solution when $\delta > R_L$, we perform a feasibility check in Algorithm 3.1 before starting the search for the capacity. If we set $\varepsilon_1 = 10^{-3}$, we can find λ^* by at most 15 iterations of bisection search in our numerical examples.

Algorithm 3.1

(i) *Feasibility check.* If $\delta \leq R_L$, let $\lambda_l = 0$ and $\lambda_u = \lambda_{sat}$ and go to step (ii); otherwise $\lambda^* = 0$ and stop.

(ii) *Bisection search.* Let $\lambda = (\lambda_l + \lambda_u)/2$ and solve for the flow throughput γ_L using (3.32). If it satisfies the constraint of (3.7), set $\lambda_l = \lambda$; otherwise $\lambda_u = \lambda$. Continue step (ii) until $\lambda_u - \lambda_l < \varepsilon_1$.

3.3.2 Finding the flow capacity for $U_r = U_I$

We have presented how to obtain λ^* for an arbitrary load $U_I \in [0, 1]$ in the interfering sectors. However, the utilization or load U_r in the reference sector associated with this λ^* will generally not be the same as U_I. Our aim is to find the λ^* for a homogeneous load network such that $U_r = U_I$. For a given δ and β, we define ε_2 for the error tolerance and apply the iterative procedure in Algorithm 3.2.

Algorithm 3.2

(i) Start with an arbitrary initial load in the interfering sectors, for example, let $U_I = 0.5$.

(ii) Run the system-level simulation to generate the time-average rates $\{R_1(U_I), \cdots, R_K(U_I)\}$ and find the worst-case average rate, $R_L(U_I)$, meeting the proportion of satisfied users β.

(iii) Find λ^* (using (3.20) for the RR case, or Algorithm 3.1 for the PF case) and use (3.17) or (3.24) to obtain the associated utilization U_r.

(iv) If $|U_r - U_I| > \varepsilon_2$, set $U_I = U_r$ and go to step (ii); otherwise, stop.

Algorithm 3.2 finds λ^* for a homogeneous load, given δ and β. If instead of finding a single point (usually there needs to be at least 5 iterations/simulations to achieve 10^{-3} calculation accuracy for one point), we are interested in generating a curve of λ^* versus δ for homogeneous loads (the *homogeneous load curve*), then a quicker procedure can be used that avoids iterations involving additional simulations. We perform a series of simulations beforehand for a set of loads in the interfering sectors. For each load U_I, we use (3.17) or (3.24) to find the λ^* that satisfies $U_r = U_I$. Then we solve the flow throughputs (3.19) or (3.32) and find the value of δ using the constraint of (3.7) and the pre-defined β. Thus we obtain a set of (δ, λ^*) pairs which gives an approximate homogeneous load curve. We will illustrate this method graphically in Section 3.5.1.

3.4 System model for FFR and SFR schemes

3.4.1 Problem formulation for an arbitrary load U_I in interfering sectors

For reuse partitioning schemes, the reference sector with FFR-n (or SFR-n) is modelled using two separate GPS queueing systems, namely a centre system and an edge system, on the basis of resource access assumptions in Section 3.2.3. If the arrival rate in the reference sector is λ, the arrival rate in the centre system is $\lambda_c = \lambda(1-p_n)$, and in the edge system is $\lambda_e = \lambda p_n$. The total offered traffic to the centre system is

$$\rho_c = \sum_{k=1}^{K_c} \rho_{ck}, \quad \text{where } \rho_{ck} = \frac{P_{ck}\lambda_c\sigma}{R_{ck}}, k = 1, \ldots, K_c, \tag{3.33}$$

while the total offered traffic to the edge system is

$$\rho_e = \sum_{k=1}^{K_e} \rho_{ek}, \quad \text{where } \rho_{ek} = \frac{P_{ek}\lambda_e\sigma}{R_{ek}}, k = 1, \ldots, K_e. \tag{3.34}$$

For the given λ, we let $\gamma_{ck}, k = 1, \ldots, K_c$, and $\gamma_{ek}, k = 1, \ldots, K_e$, be the flow throughputs of class-k centre users and class-k edge users respectively, which can be obtained by solving two separate GPS queues. We sort the flow throughputs in a combined list in decreasing order such that

$$\gamma_1 \geq \cdots \geq \gamma_K, \quad \text{where } K = K_c + K_e. \tag{3.35}$$

For a class-k user, $\gamma_k = \gamma_{ck_c}$ (or γ_{ek_e}), the time-average rate R_k is R_{ck_c} (or R_{ek_e}) and the corresponding probability P_k is $(1-p_n)P_{ck_c}$ (or $p_n P_{ek_e}$), if the user belongs to the centre system (or edge system). Therefore, the problem formulation (3.7) can be applied to solve for the flow capacity in FFR-n (or SFR-n) with RR scheduling or PF scheduling. By solving the flow capacity for different FFR-n (or SFR-n) schemes, we can explore a judicious choice of edge band size n to maximise the capacity.

The system stability conditions for RR scheduling are

$$\rho_c < 1 \quad \text{and} \quad \rho_e < 1, \tag{3.36}$$

which lead to the cell capacity

$$\lambda_{sat} = \min\left\{\frac{\overline{R_c}}{\sigma(1-p_n)}, \frac{\overline{R_e}}{\sigma p_n}\right\}, \tag{3.37}$$

where

$$\overline{R_c} = \left[\sum_{k=1}^{K_c} \frac{P_{ck}}{R_{ck}}\right]^{-1} \quad \text{and} \quad \overline{R_e} = \left[\sum_{k=1}^{K_e} \frac{P_{ek}}{R_{ek}}\right]^{-1} \tag{3.38}$$

are the harmonic means of the rate distributions for the centre and edge systems. The next question that we address is how to choose the value for p_n. It is clear from (3.37) that if $\overline{R_c}$ and $\overline{R_e}$ are independent of p_n, then the cell capacity is maximised if p_n is chosen so that

$$p_n = \frac{\overline{R_e}}{\overline{R_c} + \overline{R_e}}. \tag{3.39}$$

In actuality, $\overline{R_c}$ and $\overline{R_e}$ depend on p_n because we choose the SINR threshold s_n as the p_n-th percentile of the SINR distribution (see Section 3.2.3). A different value of p_n gives a different SINR threshold s_n (as shown in Figure 3.2), which results in different rate distributions for centre/edge users and thus different values of $\overline{R_c}$ and $\overline{R_e}$. Nevertheless, numerical exploration shows that the setting (3.39) typically does maximise the cell capacity.

For example, Figure 3.5 shows the relationship between the cell capacity and p_n for FFR-11 with RR scheduling in a fully loaded environment ($U_I = 1$), where $\sigma = 1$ Mbits and $\overline{R_c}$ and $\overline{R_e}$ are obtained via the system-level simulation described in Section 3.5. It is clear that in this example, choosing p_n according to (3.39) leads to the maximum cell capacity.

Consequently, we apply the choice (3.39) in our numerical results, with the assumption that as well as maximising the cell capacity, it will yield a high flow capacity. The inter-dependence between p_n and $\overline{R_c}$ and $\overline{R_e}$ implies a fixed point equation, and so we find p_n and the associated rate distributions using an iterative

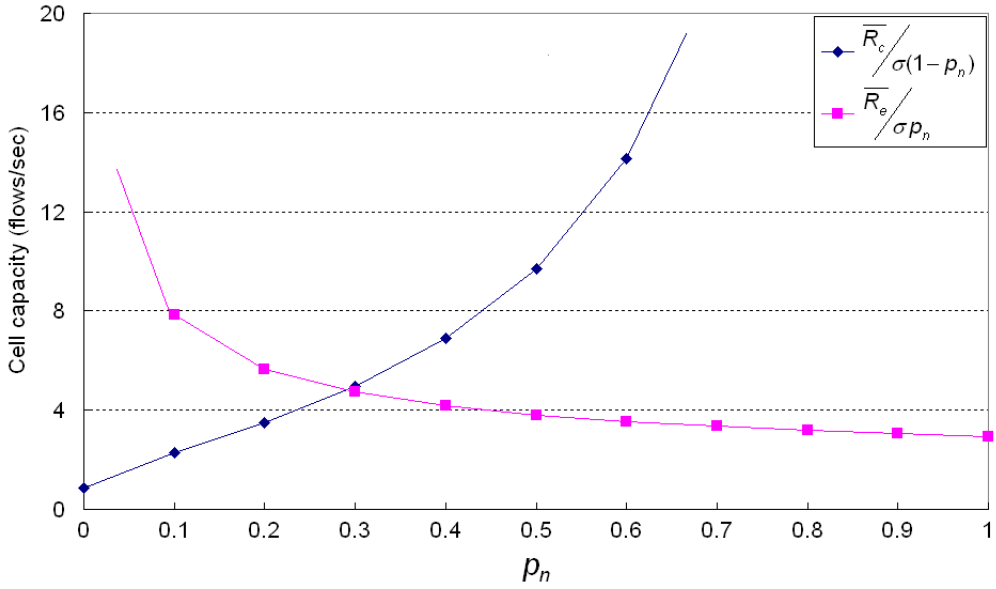

Figure 3.5: An example of the cell capacity for different p_n: FFR-11 with RR, $U_I = 1$, and $\sigma = 1$ Mbits.

algorithm, as defined in Algorithm 3.3.

The final step is to find the flow capacity; unlike the case for a standard reuse scheme, the identity of the L-th class can vary with λ since there are two systems with possibly interleaved sets of time-average rates, so it is expedient to apply an iterative search to find the flow capacity, as formalised in Algorithm 3.4.

For PF scheduling, the stability conditions are

$$\rho_c < g_c \quad \text{and} \quad \rho_e < g_e, \tag{3.40}$$

which lead to the cell capacity

$$\lambda_{sat} = \min\left\{\frac{g_c\overline{R_c}}{\sigma(1-p_n)}, \frac{g_e\overline{R_e}}{\sigma p_n}\right\}, \tag{3.41}$$

where g_c and g_e denote the limits for PF scheduling gain in the centre and edge systems. To try to maximise the cell capacity, we parallel (3.39) and use

$$p_n = \frac{g_e\overline{R_e}}{g_c\overline{R_c} + g_e\overline{R_e}}. \tag{3.42}$$

The flow capacity is found by first solving the fixed point equation between p_n and the rate distributions according to Algorithm 3.3, and then solving for the capacity itself according to the steps in Algorithm 3.4.

An approach to find rate distributions for the optimal p_n

Let ε_3 be the error tolerance. We perform a simulation a priori for the SINR distribution on the centre band, and store the relationships of p_n and s_n in a look-up table. Then we apply an iterative method, denoted by Algorithm 3.3, to determine the optimal p_n and its associated time-average rates $\{R_{c1}, \ldots, R_{cK_c}\}$ and $\{R_{e1}, \ldots, R_{eK_e}\}$ in the reference sector for an arbitrary load in the interfering sectors. In our numerical experiments, we find that we only need to run 4–5 iterations to achieve $\varepsilon_3 = 10^{-3}$ accuracy.

Algorithm 3.3

(i) Choose an initial value for p_n, for example, let it be the ratio of the edge band to the total available band ($n/(N-2n)$ for FFR or n/N for SFR).

(ii) According to the look-up table, find the SINR threshold s_n for the current p_n.

(iii) Run the system-level simulation: randomly generate a user location; if his SINR on the centre band is less than s_n, he is deemed to be an edge user and a new SINR on the edge band needs to be re-calculated; then use the SINR-to-Rate mapping function to obtain one sample for the rate distribution for the centre or edge band. Then generate enough samples to obtain two rate distributions and discretize them to $\{R_{c1}, \ldots, R_{cK_c}\}$ and $\{R_{e1}, \ldots, R_{eK_e}\}$.

(iv) Find the new value of p'_n using (3.39) for the RR case or (3.42) for the PF case. If $|p_n - p'_n| > \varepsilon_3$, then let $p_n = p'_n$ and go to step (ii); otherwise stop.

An algorithm to find the flow capacity

Given δ and β, a bisection search similar to Algorithm 3.1, denoted by Algorithm 3.4, is applied to find λ^* between 0 and λ_{sat}. Unlike Algorithm 3.1, the identity of the L-th class in the reuse partitioning case can vary with λ since there are two systems with possibly interleaved sets of time-average rates, thus it is not possible to establish the feasibility of the solution at the outset; we remove the feasibility check step and

allow Algorithm 3.4 to return $\lambda^* = 0$ if the problem is infeasible. Letting ε_4 be the error tolerance and $\lambda_l = 0$ and $\lambda_u = \lambda_{sat}$, we define the algorithm as shown below. Similar to Algorithm 3.1, we need at most 15 iterations of the bisection search to achieve 10^{-3} calculation accuracy in our numerical examples.

Algorithm 3.4

(i) *Bisection search.* Let $\lambda = (\lambda_l + \lambda_u)/2$, and set $\lambda_c = (1 - p_n)\lambda$ in the centre system and $\lambda_e = p_n\lambda$ in the edge system. Solve for the flow throughputs in each system (using (3.19) for RR and (3.32) for PF), and sort them in a single list in decreasing order so that γ_L can be identified. If γ_L satisfies the constraint of (3.7), set $\lambda_l = \lambda$; otherwise $\lambda_u = \lambda$. Continue this process until $\lambda_u - \lambda_l < \varepsilon_4$.

3.4.2 Finding the flow capacity for $U_r = U_I$

We aim to find the flow capacity λ^* for the homogeneous load case when $U_r = U_I$, where U_r is the overall utilization in the reference sector. Let U_{rc} and U_{re} denote the utilizations in the centre and edge systems. For RR scheduling, U_{rc} and U_{re} are determined from (3.17), and it can be shown that (3.39) ensures that $U_{rc} = U_{re}$; therefore

$$U_r = U_{rc} = U_{re}. \tag{3.43}$$

For PF scheduling, U_{rc} and U_{re} are computed from (3.24). The centre and edge systems will generally have different PF scheduling gains, denoted by $G_c(x)$ and $G_e(x)$ respectively, due to different frequency bandwidths (see Figure 3.10). Hence $G_c(x) \neq G_e(x)$ and $g_c \neq g_e$, and so from (3.24) and (3.30), $U_{rc} \neq U_{re}$. To err on the side of conservatism, we approximate the overall utilization of the reference sector by

$$U_r = \max\{U_{rc}, U_{re}\}. \tag{3.44}$$

Note, however, that U_{rc} and U_{re} are usually close to each other due to the following: (i) in (3.30), the first term dominates the second; (ii) (3.42) implies $\rho_c/g_c = \rho_e/g_e$; (iii) the parameter C in (3.30) depends on the ratios $G(i)/g$, $i = 1, \ldots, x_0$, which tend to be insensitive to the bandwidth.

For given δ and β, we define ε_5 for the error tolerance and apply Algorithm 3.5 to obtain the flow capacity achieving $U_r = U_I$.

Algorithm 3.5

(i) Start with an arbitrary initial load in the interfering sectors (e.g. $U_I = 0.5$).

(ii) Apply Algorithm 3.3, and obtain the time-average rates $\{R_{c1}(U_I), \ldots, R_{cK_c}(U_I)\}$ and $\{R_{e1}(U_I), \ldots, R_{eK_e}(U_I)\}$ for the centre and edge systems and the optimal partitioning p_n in the reference sector.

(iii) Find λ^* using Algorithm 3.4, where the overall utilization $U_r = U_{rc} = U_{re}$ for the RR case and the approximate overall utilization $U_r = \max\{U_{rc}, U_{re}\}$ for the PF case.

(iv) If $|U_r - U_I| > \varepsilon_5$, then set $U_I = U_r$ and go to step (ii); otherwise, stop.

Also, we can apply the alternative approach described in Section 3.3.2 to obtain the homogenous load curve, which can save a great deal of simulation effort.

3.5 Numerical experiments and discussion

In this section, we present numerical examples to illustrate the capacities of reuse-1, reuse-3, FFR-n and SFR-n. A total of 48 resource units (e.g. resource block or RB in LTE) are assumed, thus n can be selected from the set $\{1, \ldots, 16\}$. Note that FFR-16 is equivalent to reuse-3.

We employ a hybrid simulation/analysis approach to obtain the flow capacity for each scheme, where a uncontended time-average rate distribution obtained via static system-level simulation is used as the input to the queueing model. The simulation parameters and assumptions are consistent with the LTE downlink. We take into account the effects of antenna gain and antenna directivity $G(\theta)$, path loss L_d, shadow fading L_s, inter-cell interference, and thermal noise N_0 when calculating the UE's received SINR. While UEs are equipped with omni-directional receive antennas, the gain pattern (in dB) for 120° directional transmit antennas at base station (BS)

is considered to be

$$G(\theta)(\text{dB}) = G_x - \min\left\{12 \cdot (\frac{\theta}{\theta_{3dB}})^2, G_{FB}\right\}, -180 \leq \theta \leq 180, \tag{3.45}$$

where G_x is the base station antenna gain plus cable loss, θ is defined as the angle between the direction to UE and the boresight of the antenna, θ_{3dB} is the degree for the 3 dB beamwidth, and G_{FB} is the maximum attenuation or the front-to-back ratio (see 3GPP standard [2]).

The following path loss model, namely COST-231 Hata model for medium-sized cities and suburbs, has been used as defined in the book by Stuber [131],

$$\begin{aligned} L_d(\text{dB}) &= 46.3 + 33.9 \log_{10} f_c - 13.82 \log_{10} h_b \\ &\quad + (44.9 - 6.55 \log_{10} h_m) \log_{10} d, \end{aligned} \tag{3.46}$$

where f_c is the carrier frequency, h_b is the base station height, h_m is the UE height, and d is the distance between the base station and the UE.

As for the shadow fading, we use independent log-normal random variables with a standard deviation of 8 dB, with a correlation of 0.5 between different sites and 1.0 between sectors of the same sites. The thermal noise is calculated based on a power spectral density (PSD) of -174 dBm/Hz. Each resource block is assigned fixed equal power and the same modulation and coding scheme for sub-carriers is assumed within a resource block.

The parameters and assumptions are summarized in Table 3.2. The mean flow size σ is set to be 5 Mbits.

The input to the analytical model is a discrete set of rates used to approximate a more or less continuous rate distribution estimated from the simulation. For each distribution, we generate 100,000 simulation samples. There are different ways in which the distribution can be sampled to create the discrete approximation. Inspired by the important role played by the harmonic mean in the RR analysis (3.20), we divide the distribution into equi-probability "bins" (see step 2 in Figure 3.3) and take R_j to be the harmonic mean of the simulation samples in the jth bin.

Table 3.2: Simulation parameters

Cell layout	Hexagonal grid, 19 tri-sector sites
Inter-site distance	1 km (333 m cell radius)
Minimum distance to BS	35 m
Spectrum allocation	48RBs in 10 MHz at f_c 2.5 GHz for data transmission (remaining RBs used for other purposes)
Thermal noise density	-174 dBm/Hz
Propagation model	COST-231 Hata model (3.46)
Shadow fading	Log-normal shadowing with standard deviation 8 dB and correlation 0.5 between different sites and 1.0 between sectors of the same sites
BS transit power	43 dBm (20 W)
BS antenna gain plus cable loss, G_x	14 dBi
BS antenna pattern	3GPP 2-dimensional antenna pattern (3.45) with $\theta_{3dB} = 65$ and $G_{FB} = 20$dB
BS height, h_b	30 m
UE antenna gain	0 dBi
UE noise figure	7 dB
UE height, h_m	1.5 m

3.5.1 Definition of key aspects of the methodology

In this section, we use numerical exploration to decide on key parameter values for our methodology for computing the flow capacity, namely (i) definition of the mapping function between SINR and rate, (ii) quantification of the PF scheduling gain, and (iii) definition of the number of classes K. Finally, we present a linear approximation technique for the flow capacity for PF scheduling.

Sensitivity to SINR-to-Rate mappings

As an intermediate step, the simulation generates the time-average SINR distribution. To generate the rate distribution, the simulation employs a simple SINR-to-Rate mapping function. We test three different options for the mapping function (see Figure 3.6): (A) linear mapping: $S(\text{bits/s/Hz}) = 1 \times \text{SINR}$; (B) Shannon mapping: $S(\text{bits/s/Hz}) = \log_2(1+\text{SINR})$; (C) modified Shannon mapping introduced by

Mogensen *et al.* [89], which can account for various implementation issues such as the cyclic prefix, pilot channel and signalling overheads, and receiver performance, and is represented by

$$S(\text{bits/s/Hz}) = \min\left\{\text{BW}_{eff} \cdot \log_2\left(1 + \text{SINR} \,/\, \text{SNR}_{eff}\right), S_{\max}\right\}. \tag{3.47}$$

We use the results for SIMO from [89], where $\text{BW}_{eff} = 0.62$, $\text{SNR}_{eff} = 1.8$ dB and $S_{\max} = 4$ bits/s/Hz, which accounts for the 64QAM limit of modulation and coding schemes.

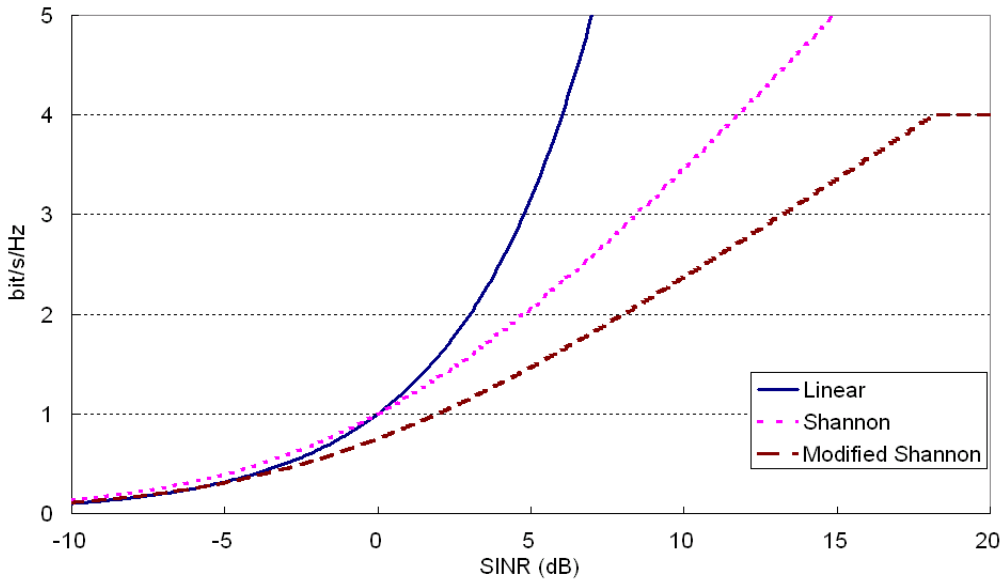

Figure 3.6: Three SINR-to-Rate mapping methods: (A) linear; (B) Shannon; (C) Modified Shannon [89].

Figure 3.7 plots the SINR distributions of reuse-1 and reuse-3 for a fully loaded environment ($U_I = 1$), as simulated for an LTE network with the parameters in Table 3.2. We can see that reuse-3 can improve the SINR significantly since reuse-3 suffers less interference than reuse-1. Figure 3.8 shows the rate distributions using the three mapping methods. For every mapping, reuse-3 increases the cell edge rates compared to reuse-1 (shown in the zoomed-in area in Figure 3.8) but sacrifices the peak rate. Note that the highest rate of reuse-3 in option (C) is restricted by the maximum spectral efficiency $S_{\max}$, hence the discontinuity.

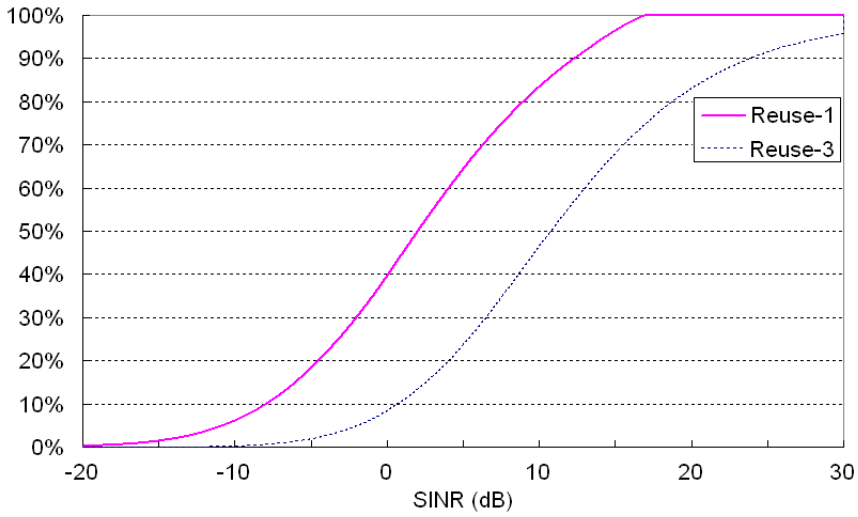

Figure 3.7: SINR distributions of reuse-1 and reuse-3 (when $U_I = 1$) for a typical LTE deployment with the parameters in Table 3.2.

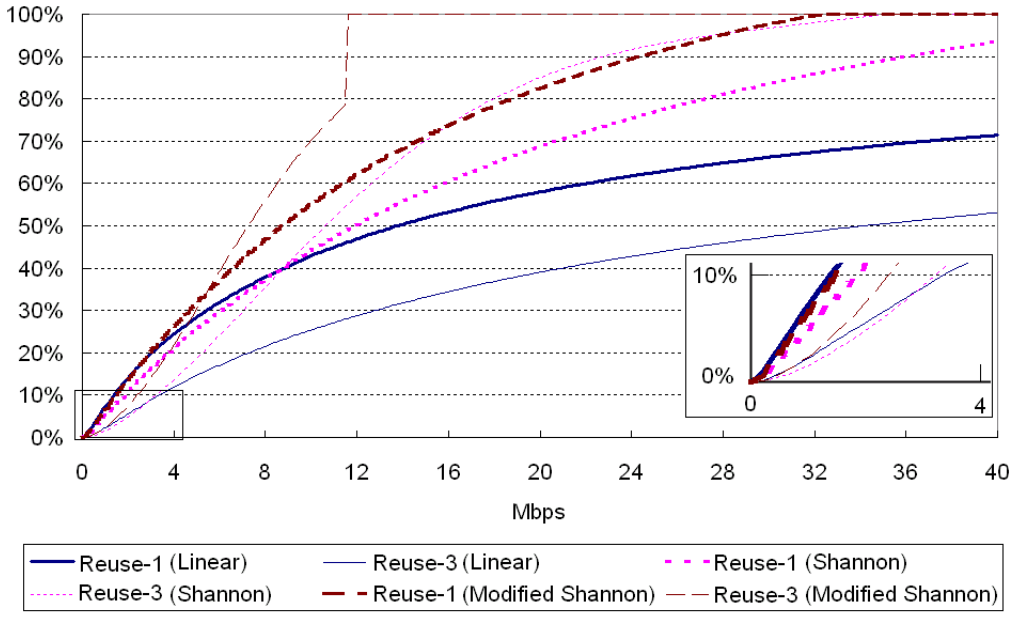

Figure 3.8: Rate distributions of reuse-1 and reuse-3 with three SINR-to-Rate mapping methods: (A) linear; (B) Shannon; (C) Modified Shannon [89]. The inset in the bottom right is a zoom-in of the low rates.

In Figure 3.9, we compare the flow capacities for reuse-1 and reuse-3 in a simplified scenario with RR scheduling, $U_I = 1$, $K = 10$ and $\beta = 1$ (10 classes in a fully loaded environment, and all the users need to satisfy the flow throughput requirement). Note that for this and all subsequent graphs of the flow capacity, we plot the product of the flow capacity λ^* and the mean flow size σ on the y-axis; this normalisation makes the plots invariant to the individual values of λ^* and σ. From Figure 3.9, we can see that for the plotted domain of δ, reuse-3 for all three mappings has higher capacity than reuse-1, reflecting the impact of a better cell edge rate. For large δ, however, the disparity in peak rates means that reuse-1 outperforms reuse-3, as we will show in later examples. In the rest of this section, we only apply the most realistic mapping, namely modified Shannon mapping (3.47).

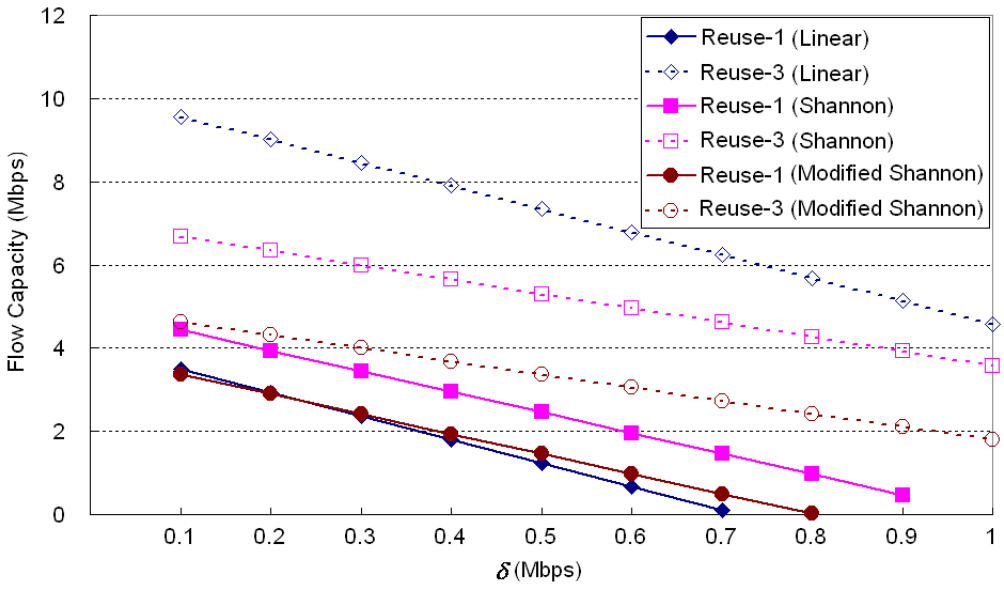

Figure 3.9: Flow capacities with three SINR-to-Rate mappings for RR scheduling with $\beta = 1$, $U_I = 1$ and low δ.

Fast convergence of PF scheduling gain

We use another dynamic system-level simulation based on the parameters listed in Table 3.2 to estimate the gain $G(x)$ of PF scheduling over RR scheduling under different bandwidths. The channel model is the extended pedestrian A 3km/h (EPA) model [3]. The predicted instantaneous data rate ($R_{i,j}(t)$ in (2.6), see Section 2.5) is

reported by CQI feedback, which is based on sub-bands (4RBs per sub-band) with delays (4 Transmission Time Intervals (TTI)) and errors (normal distribution with 1 dB variation) [5, 103, 136]. The window size t_c in (2.7) is set to 100.

Figure 3.10 depicts the gain function $G(x)$ for different bandwidths as a function of the number of active users in the sector. The gain increases with bandwidth due to increased frequency selectivity. As foreshadowed in the analysis, the gain functions are nearly flat beyond a certain number of users, $x_0 \approx 7$, which is not large and ensures that the flow throughput (3.32) has low computational complexity. We will use the $G(x)$ values in Figure 3.10 to calculate flow capacities for different schemes in Section 4.5.2.

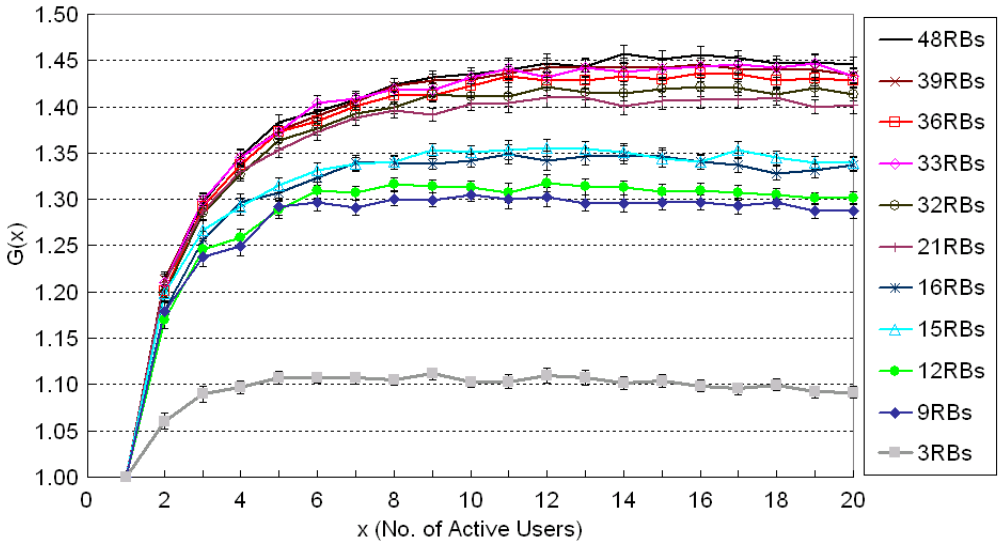

Figure 3.10: Gain functions of PF scheduling over RR scheduling as a function of the number of active users for different frequency bandwidths.

Sensitivity to the number of classes K for PF scheduling

In the discretization of the rate distribution, more classes will lead to a more accurate representation of the rate distribution but at greater computational cost in the analysis, particularly for FFR or SFR schemes. Figure 3.11 shows the capacities of reuse-1 with PF scheduling when $U_I = 1$ and $\beta = 0.9$, using different numbers of classes in the analysis. We see that 100 classes can achieve almost the same

accuracy as 1000 classes. Further, 100 classes can be comfortably managed in terms of complexity. Therefore, we apply $K = K_c = K_e = 100$ in the comparisons of different schemes in Section 4.5.2.

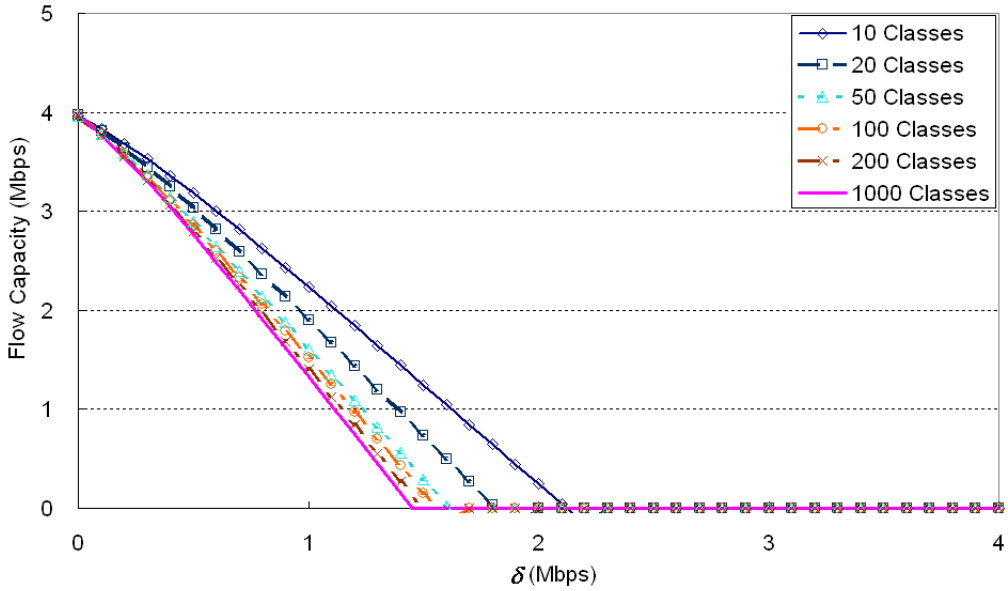

Figure 3.11: Sensitivity to the number of classes (reuse-1 with PF when $\beta = 0.9$ and $U_I = 1$).

An approximation for PF scheduling

We have shown that it is difficult to obtain a simple expression like (3.20) for the flow capacity for PF scheduling because the calculation of flow throughput (3.32) is more involved than for RR scheduling. However, we note, as a heuristic, that a simple linear approximation can be applied for the capacity for PF scheduling. Figure 3.12 displays an example for reuse-1 in a fully loaded environment. The linear approximation is defined as follows: if $\delta = 0$, then $\lambda^* = g\overline{R}/\sigma$; if $\lambda^* = 0$, then $\delta = R_L$ (R_L is dependent on β). So we have

$$\lambda^* \approx (1 - \frac{\delta}{R_L})\frac{g\overline{R}}{\sigma}. \tag{3.48}$$

Compared to the flow capacity (3.20) for the RR case, we see that in (3.48), we assume a constant multi-user diversity gain for the PF case irrespective of the number

of active users. We can see that this approximation works well for different definitions of user satisfaction β.

Figure 3.12 also shows how the parameter β (which is the proportion of users that satisfy the flow throughput requirement) impacts the flow capacity. If β is reduced, R_L increases and the capacity region increases.

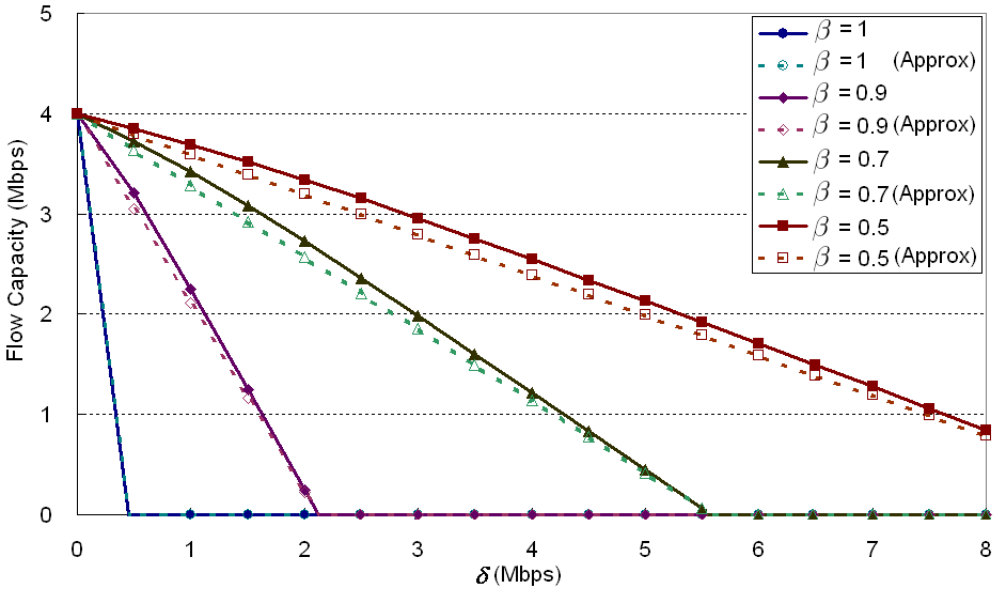

Figure 3.12: Linear approximations of the capacity (reuse-1 with PF for different values of β when $U_I = 1$).

3.5.2 Simulation validation of the hybrid simulation/analysis approach

Our hybrid simulation/analysis approach estimates the flow capacity with greatly reduced simulation effort compared to a simulation-only approach. Here we validate the accuracy of the hybrid approach with the simulation-only approach. We do not perform the comparison on the basis of flow capacity, since using a simulation-only approach to obtain the flow capacity could require many iterations on the arrival rate λ, and would prove extremely computationally expensive. Instead, we compare the per-class flow throughputs obtained by our hybrid approach with those obtained from the pure simulation approach. This is sufficient to verify the flow capacity.

The objective is to validate the following case. In the analysis, as described in Section 3.3.2, we use K classes of time-average rates as the input to calculate the per-class flow throughputs according to (3.19) or (3.32). The inputs $R_1, \ldots, R_K$ are obtained by discretizing a continuous rate distribution based on a system-level simulation using the parameters in Table 3.2, dividing it into K equi-probability bins and taking the harmonic mean of each bin. In the simulation used to validate the analysis, we determine the per-class flow throughputs via simulation. The simulation is based on the simulation used to generate the rate distribution, but is extended to also simulate the flow arrivals and transmissions. The data flows are generated by a Poisson process. For each flow, the user generating it is randomly located in the reference sector; its transmission rate depends on the location, antenna gains, shadow fading and inter-cell interference. Each flow has a fixed file size $\sigma = 5$ Mbits since the distribution of file size has no effect on the flow throughput (see Section 2.5). At the end of the simulation, we aggregate the flow throughputs into K bins, and compare the flow throughput of bin-k with the analytical flow throughput of class-k, $k = 1, \ldots, K$.

The next problem that we address is how to obtain an aggregated flow throughput for bin-k in the simulation for the comparison, since in each bin the transmission rate may vary. We present the method on a theoretic basis and take the case of reuse-1 with RR scheduling as an example. For bin-k, R_k is the harmonic mean of the rates within this bin; we assume that the rates within bin-k are constrained into a finite number J of rates, which are denoted by $\{R_{k1}, \ldots, R_{kJ}\}$ with probabilities $\{P_{k1}, \ldots, P_{kJ}\}$ ($\sum_{j=1}^{J} P_{kj} = 1$). We have

$$R_k = \left(\sum_{j=1}^{J} \frac{P_{kj}}{R_{kj}} \right)^{-1}. \tag{3.49}$$

From (3.19), we can obtain the flow throughput for the flows with the rate R_{kj}, which is

$$\gamma_{kj} = R_{kj}(1 - \frac{\lambda\sigma}{\overline{R}}), \tag{3.50}$$

where $\overline{R}$ is the harmonic mean of the rate distribution, and is given by

$$\overline{R} = \left(\sum_{k=1}^{K} \frac{P_k}{R_k} \right)^{-1} = \left(\sum_{k=1}^{K} \sum_{j=1}^{J} \frac{P_k P_{kj}}{R_{kj}} \right)^{-1}. \tag{3.51}$$

Let $E_H[\cdot]$ represent taking the harmonic mean. We have

$$E_H[\gamma_{kj}] = E_H[R_{kj}](1 - \frac{\lambda\sigma}{\overline{R}}) = R_k(1 - \frac{\lambda\sigma}{\overline{R}}) = \gamma_k. \tag{3.52}$$

Equation (3.52) shows that if we take the harmonic mean of the flow throughputs associated with the J rates, we can obtain the same value as the flow throughput obtained using the harmonic mean rate R_k of this bin. In the simulation, we use this method to calculate the aggregated flow throughput of bin-k for the comparison.

In the simulation, the RR scheduling is performed in the time domain such that at any 1 ms scheduling instant, the base station allocates all the frequency resources to one flow and at the next instant schedules the next flow on a round-robin basis. Once a data flow finishes transmission, the transmission duration is recorded. For each rate, the mean transmission duration is then calculated; from (3.1) the flow throughput can be obtained. Finally, we calculate the aggregated flow throughput by taking the harmonic mean of the flow throughputs recorded for the bin, as suggested by (3.52).

We first present the comparisons of the per-class flow throughputs between the analysis and the simulation for the case when $U_I = 1$, $K = 10$ and $J = 10$. As depicted in Figure 3.13, we see that the analysis and simulation curves of per-class flow throughputs are nearly indistinguishable (we only show the results of class-1, 2, 3, and the same accuracy holds for other classes).

Next we provide the comparisons for the notional case of $J \to \infty$ in Figure 3.14, where we simulate a continuous rate distribution in the validation. Again, we only show the results of class-1, 2, 3. It can be seen that there are only slight differences between the analysis and simulation.

For the PF case, we find that the analysis results also coincide with the simulation results when the rate fluctuations of different classes are statistically identical;

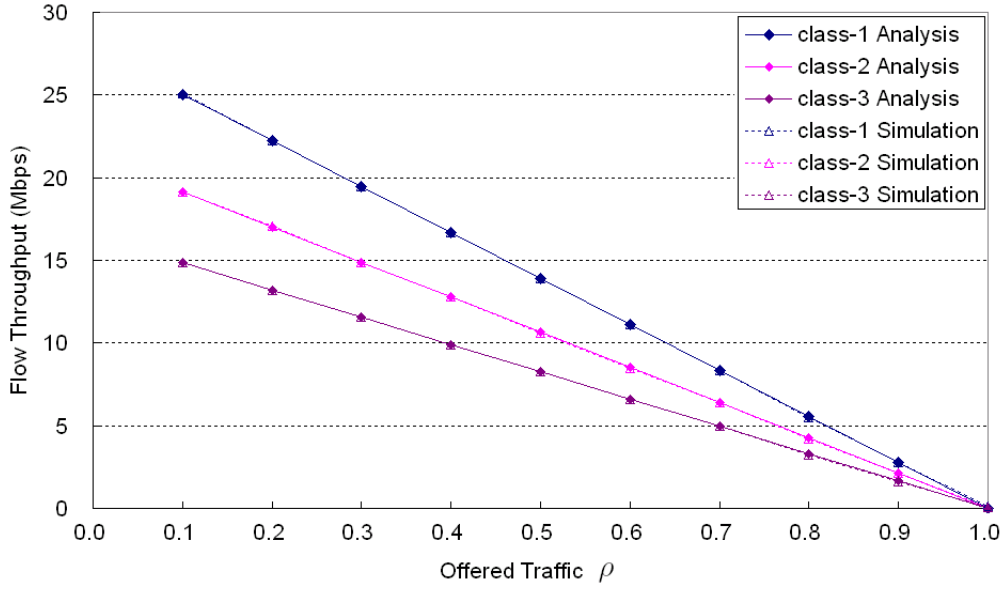

Figure 3.13: Comparison of the per-class flow throughputs between the analysis and simulation (reuse-1 with RR when $U_I = 1$ and $J = 10$).

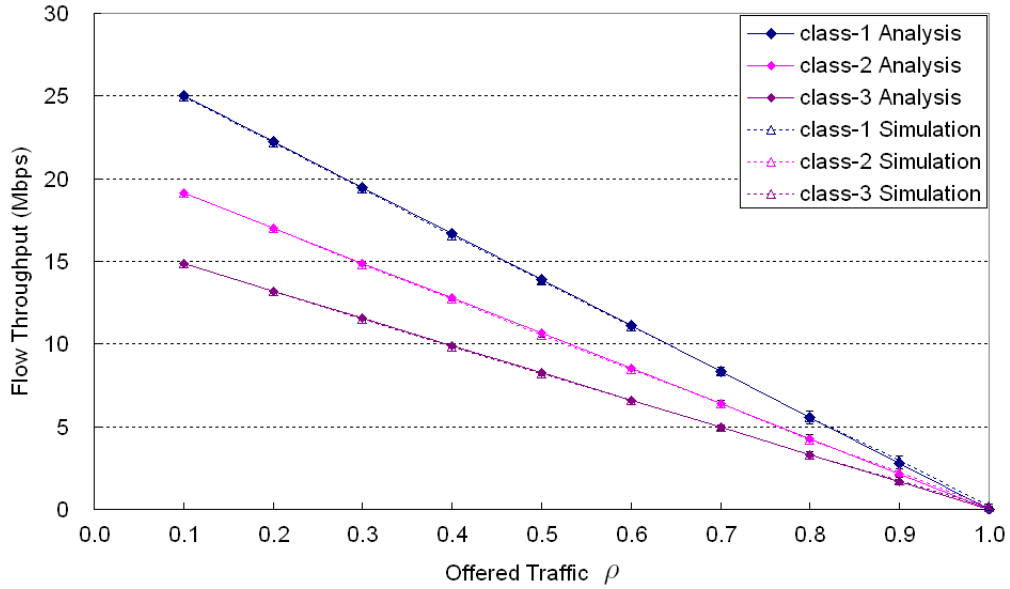

Figure 3.14: Comparison of the per-class flow throughputs between the analysis and simulation (reuse-1 with RR when $U_I = 1$ and $J \to \infty$).

similar findings are also presented by Borst [28]. However, when different classes of users experience different fast fading statistics, we observe that there are some differences between the analysis results and the simulation results, which indicates that the GPS queue model may break down for this case. A possible solution might be to apply a discriminatory processor sharing (DPS) queue to model the effects of different classes with different multi-user diversity gains (see Altman, Avrachenkov and Ayesta [10], Bonald and Proutiére [27], Mitra and Weiss [88], and Wu, Williamson and Luo [144]). In the DPS model, the division of the processor capacity is controlled by a weight vector. By changing the weights, we can effectively control the instantaneous rates of different classes, which makes DPS a more appropriate model to study the performance of the case with different statistics on the rate fluctuations. However, the DPS model is much more difficult to analyse, and we leave it for future work.

The results in this section confirm that our hybrid approach provides an accurate estimate of the user-level performance. Therefore in the next section, we use the hybrid approach to evaluate the capacities of the standard reuse and reuse partitioning schemes.

3.5.3 Comparison of flow capacities for different schemes

Plotting the homogeneous load curve

In Section 3.3.2, we described a practical approach to obtain the homogeneous load curve. Figure 3.15 shows an example of the approach for reuse-1 with PF scheduling when $\beta = 0.9$. Each approximately straight line is the flow capacity curve for a specific load U_I in the interfering sectors; the dot on the line is the point where $U_r = U_I$ (denoted as a (δ, λ^*) pair in Section 3.3.2). Thus if we analyse enough loads, we can obtain a set of (δ, λ^*) pairs and thus estimate the locus of homogeneous load, as illustrated.

We have shown in Section 3.5.1 that equation (3.48) gives a close linear approximation of the flow capacity for PF scheduling. In the following, we apply this approximation method to find the homogeneous load curve. For each specific load

U_I, we apply (3.17) for the RR case or (3.24) for the PF case in the standard reuse schemes (or (3.43) and (3.44) for FFR/SFR schemes) to find the λ^* that satisfies $U_r = U_I$. Then we apply (3.48) to find an approximate value of δ and thus obtain a (δ, λ^*) pair. Based on the set of (δ, λ^*) pairs, we plot the homogeneous load curve. Figure 3.16 illustrates such homogeneous load curves for reuse-1 and reuse-3 schemes when $\beta = 0.9$. We see that the approximations achieve very close agreement to the results from the detailed calculation. The biggest difference is about 2% at $\delta = 0.6$ Mbps for reuse-1 and about 3.1% at $\delta = 1.2$ Mbps for reuse-3. This provides a method to quickly estimate the flow capacity for the PF case if we want to avoid the complexity of the detailed calculations.

In all subsequent results, we apply the method using the detailed calculation rather than the linear approximation.

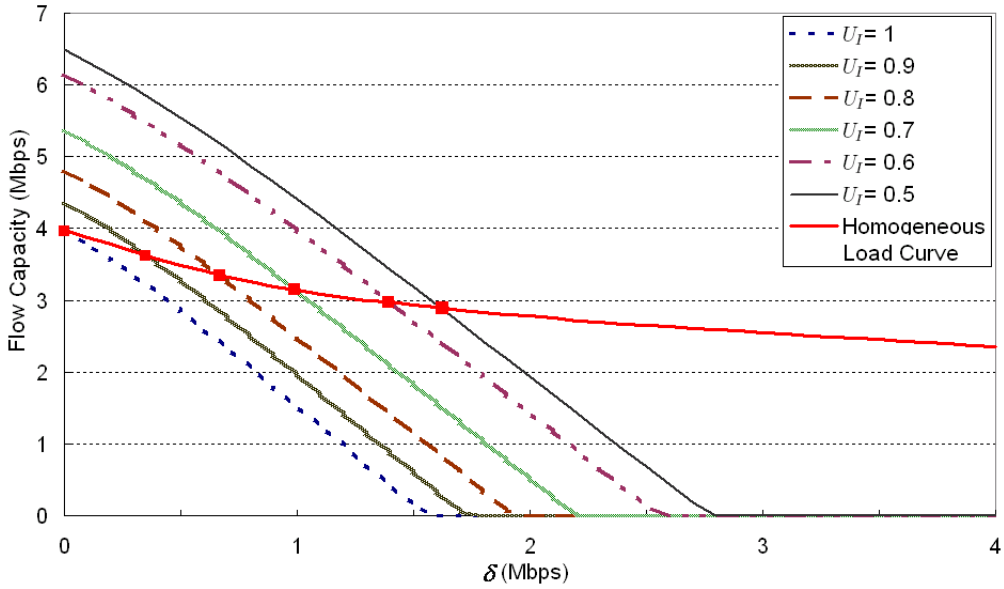

Figure 3.15: An example of plotting the approximate homogeneous load curve from the results of a set of loads (reuse-1 with PF when $\beta = 0.9$).

Choosing edge band size in FFR/SFR

We find the choice of edge band size that maximises the capacity for reuse partitioning by solving the flow capacities for different FFR-n or SFR-n schemes. Figure 3.17

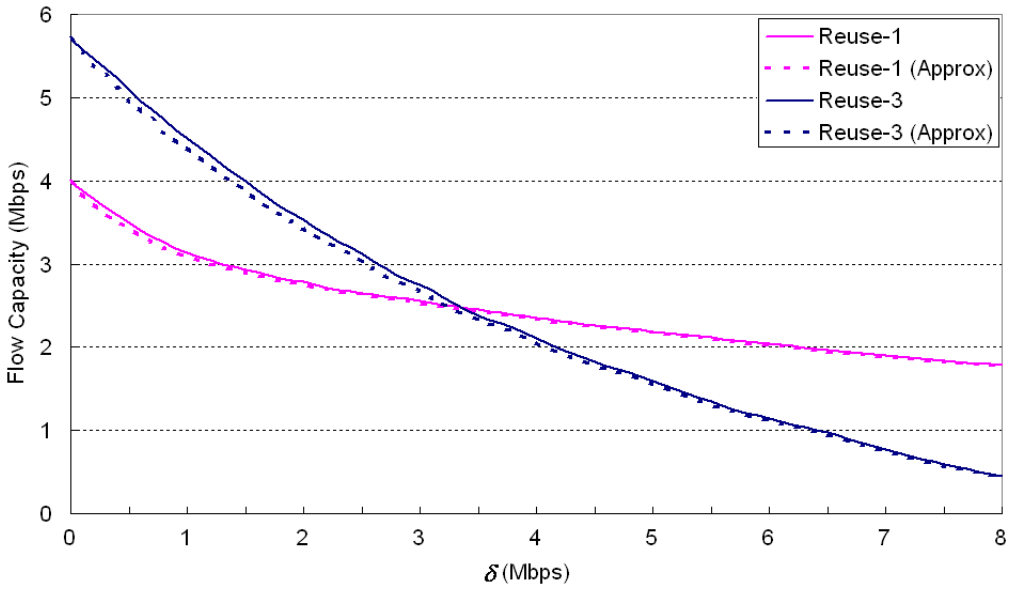

Figure 3.16: Homogeneous load curves for PF scheduling based on the linear approximation of the flow capacity using (3.48) (reuse-1 and reuse-3 when $\beta = 0.9$).

and Figure 3.18 depict the homogeneous load curves of FFR schemes with different edge band sizes when $\beta = 0.9$ for RR and PF scheduling, respectively. We find that FFR-11 is the best scheme in both RR and PF cases in terms of the cell capacity (which is the flow capacity at $\delta = 0$), though the advantage over other configurations such as FFR-12 is only slight. For $\delta > 0$, the different FFR-n schemes yield quite different flow capacities. For schemes with large n such as FFR-14 and FFR-15, the centre band resources are severely limited, and the resulting low flow throughputs of the centre band users constrain the flow capacity. For schemes with relatively small n such as FFR-8 and FFR-9, it is the edge band resources that are limited, and the edge band flow throughputs that constrain the capacity when δ exceeds 4.5 to 5 Mbps. The best all-round schemes for a wide range of δ are FFR-11 and FFR-12, which evidently provide the best balance between the edge band and centre band resources.

A similar capacity analysis for several SFR schemes with RR and PF scheduling when $\beta = 0.9$ is conducted, and the results are shown in Figure 3.19 and Figure 3.20. SFR differs from FFR in that the entire bandwidth is used in each sector, and the edge band is transmitted with higher power while the frequency bands corresponding

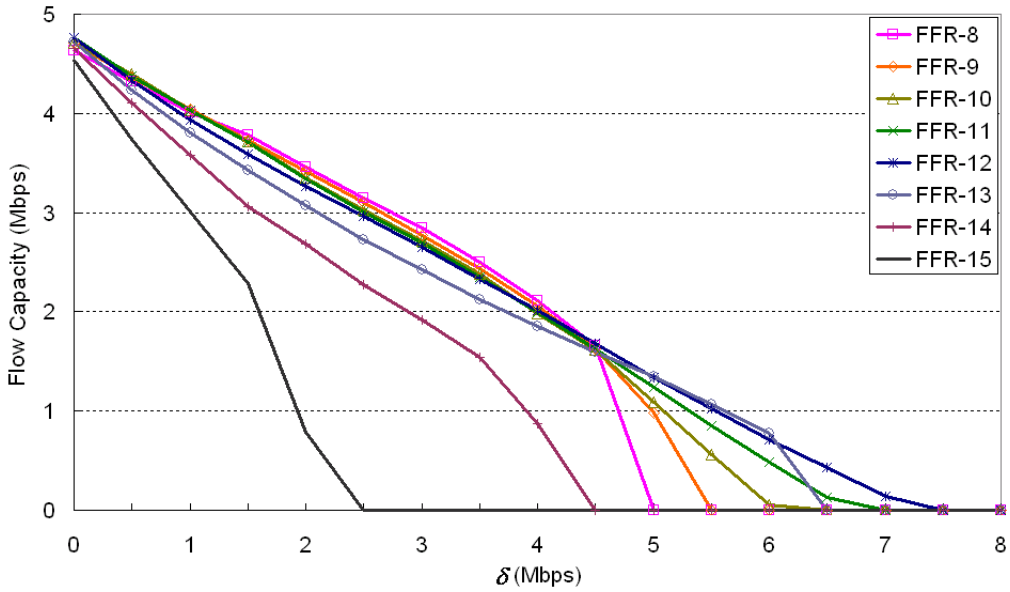

Figure 3.17: Comparison of flow capacities of different FFR schemes with RR, homogeneous load, and $\beta = 0.9$.

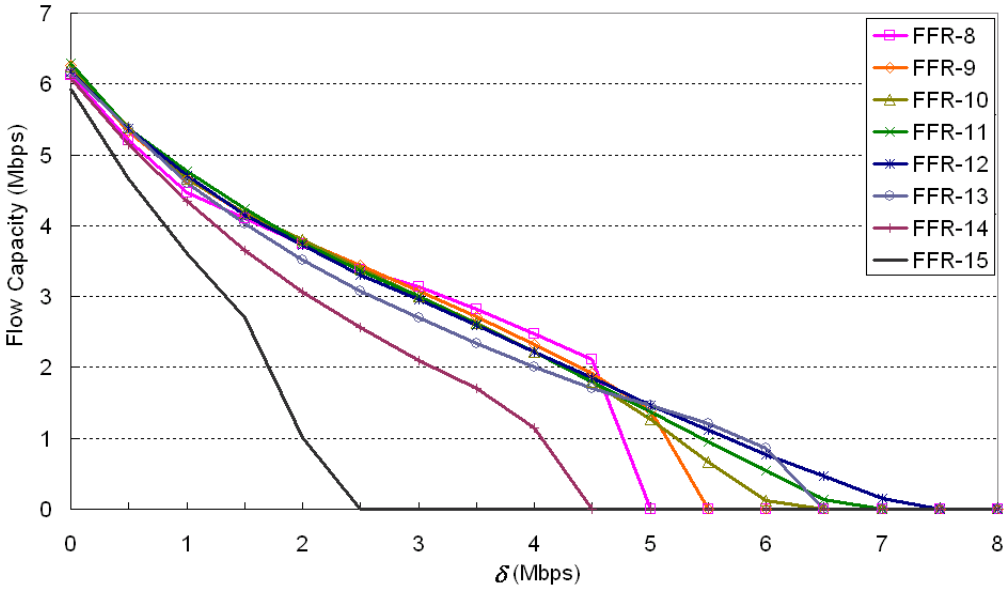

Figure 3.18: Comparison of flow capacities of different FFR schemes with PF, homogeneous load, and $\beta = 0.9$.

to the orthogonal edge bands (which can be used by sector centre users) are transmitted with lower power (see the frequency and power allocation in Figure 2.5). Therefore, for schemes with large n such as SFR-15 and SFR-16, the centre band resources are no longer severely limited. Furthermore, due to the fact that the choice of p_n plays a complicated role in determining the flow capacity, the best choice of the edge band size n depends on the throughput requirement δ. In both RR and PF cases, SFR-16 yields the maximum cell capacity (flow capacity at $\delta = 0$). For $\delta < 0.5$ Mbps or $\delta > 7$ Mbps, SFR-15 and SFR-16 have better flow capacities. For moderate $\delta \in [2.5, 5.5]$ Mbps, SFR-9 performs better; however, the edge band flow throughputs constrain the capacity when $\delta > 6$ Mbps. SFR-12 can be the best choice for some ranges of δ such as $[1, 1.5]$ and $[6, 6.5]$ Mbps.

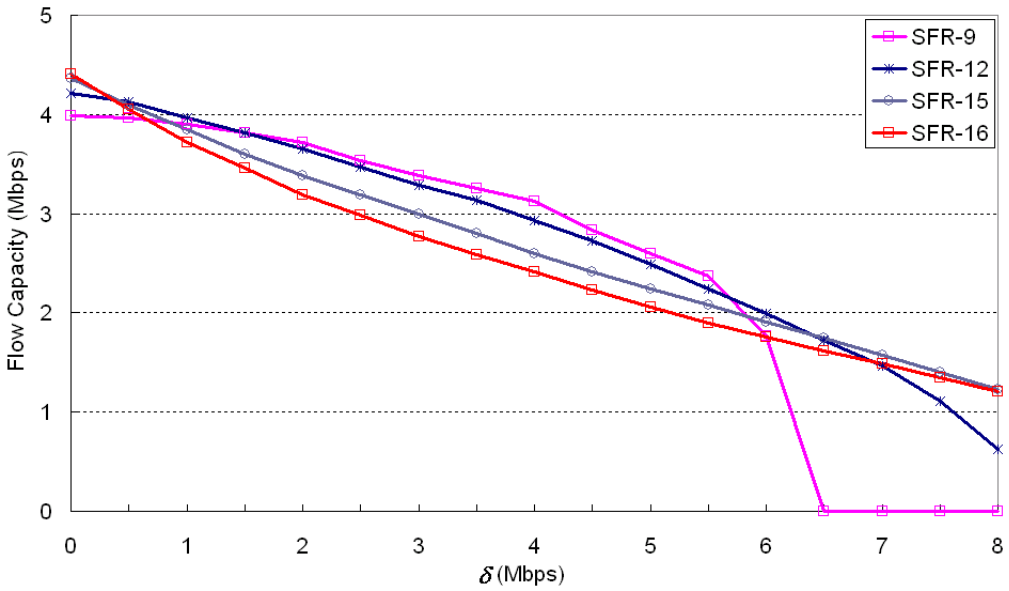

Figure 3.19: Comparison of flow capacities of different SFR schemes with RR, homogeneous load, and $\beta = 0.9$.

Capacity comparison for different schemes

We now present a capacity comparison between different reuse and reuse partitioning schemes. We apply the method in Figure 3.15 to obtain the homogeneous load curve for each scheme. Figure 3.21 depicts the flow capacities for $\beta = 0.9$ under

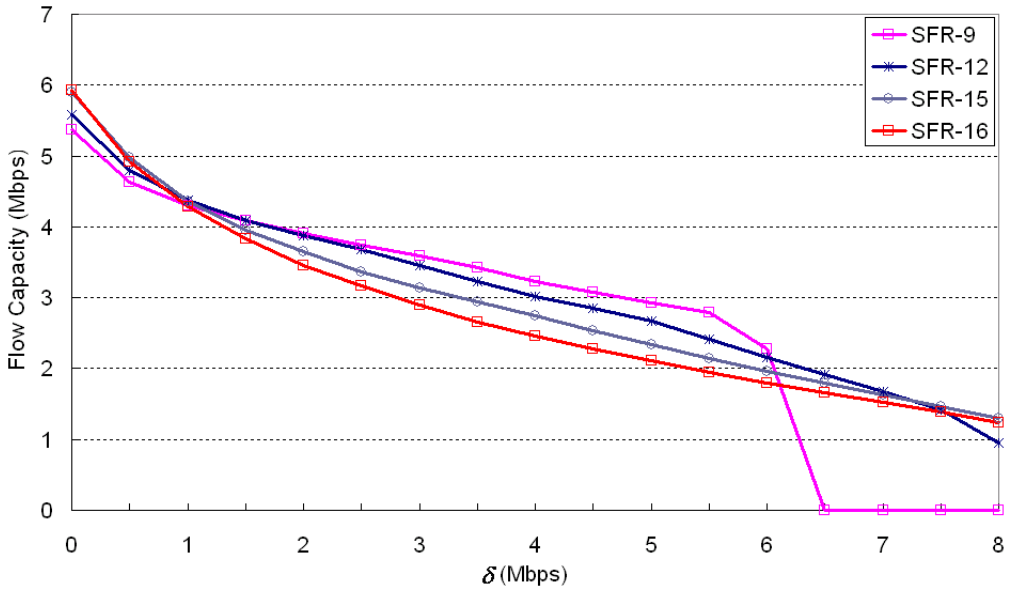

Figure 3.20: Comparison of flow capacities of different SFR schemes with PF, homogeneous load, and $\beta = 0.9$.

homogeneous loading for reuse-1, reuse-3, FFR and SFR; both RR and PF scheduling cases are plotted. Note that the FFR and SFR curves are composite curves formed by choosing the optimal edge band size n for each δ, and so are labelled as FFR-opt/SFR-opt. We see that for each scheme, PF scheduling provides a gain over RR scheduling, which decreases with increasing δ since the number of active users, and hence the multi-user diversity gain, decreases in this direction.

Let us first concentrate on the regime of low δ, such as $\delta < 1.5$ Mbps. In this regime, FFR with an optimal edge band size, namely FFR-11, is the best scheme for both scheduling methods; it can provide a gain over reuse-1 of over 50% for $\delta < 1$ Mbps and the gain decreases with increasing δ. SFR-16 and reuse-3 provide lesser gains over reuse-1.

As explained earlier, reuse-3 outperforms reuse-1 when δ is low or moderate, due to its higher cell edge rate. FFR and SFR schemes can strike a compromise between a higher cell edge rate than reuse-1 (though not as high as reuse-3) and more available resource units than reuse-3 (though not as many as reuse-1). Figure 3.21 shows that with a judicious edge band size $n = 11$ (see also Figures 3.17 and 3.18), this combination in the FFR scheme outperforms both reuse-1 and reuse-3 for low

δ. For moderate δ (such as $\delta \in [2, 6]$ Mbps), SFR with an optimal edge band size $n = 9$ (see Figures 3.19 and 3.20) yields the highest capacity. We point out that the capacity of FFR (and SFR) will be even greater if we remove the resource access restrictions for cell centre users (see Section 3.2.3).

For high δ, reuse-1 is the best scheme; reuse-1 outperforms all other schemes when $\delta > 6.5$ Mbps for the RR and PF case. This is because reuse-1 utilizes the entire bandwidth in each sector, always yields the highest peak rate, and benefits more from the reduction in interference as the network load is reduced.

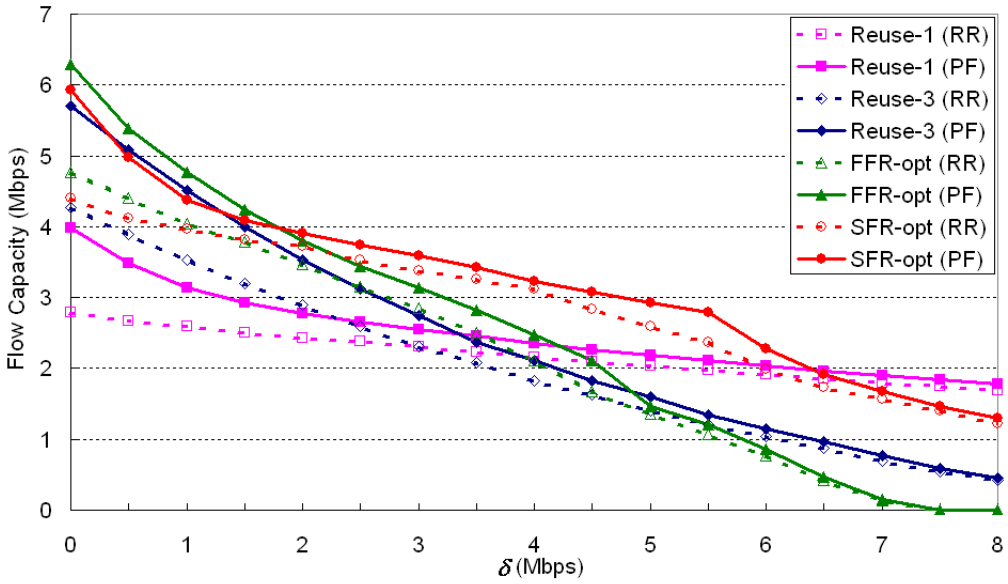

Figure 3.21: Comparison of flow capacities for reuse schemes and reuse partitioning schemes with PF and RR, homogeneous load, and $\beta = 0.9$.

3.6 Flow capacity of MIMO schemes

In this section, to show that our framework for evaluating flow capacity has applicability beyond just reuse partitioning, we apply our methodology to certain MIMO schemes.

In LTE systems, multiple antennas are available at both the base station and the UE, enabling the use of MIMO. Broadly speaking, there are two major types of MIMO: transmit diversity and spatial multiplexing. We investigate two MIMO

schemes using two antennas at both the base stations and the UEs, namely 2x2 Space Frequency Coding (SFC) and 2x2 Vertical Bell Labs Layered Space Time (V-BLAST) (see Mogensen *et al.* [89] and Wei *et al.* [136]).

SFC is a scheme that applies Alamouti Space Time Coding (see Paulraj *et al.* [98] and the book by Paulraj [19]) on groups of two neighboring sub-carriers to achieve transmit diversity. As shown in the schematic in Figure 3.22(a), two different data symbols x_{2i} and x_{2i+1} are transmitted from two antennas, respectively during the first symbol period, following which symbols x^*_{2i+1} and $-x^*_{2i}$ are launched from the two antennas. We see that effectively, only one symbol is transmitted per symbol period, but through two antennas, which can be used to improve the received instantaneous SINR for users across the sector.

V-BLAST (see Foschini [45] and Gesbert *et al.* [49]) is a scheme based on spatial multiplexing, where independent data streams are transmitted from the individual antennas (see Figure 3.22(b)). The receiver can separate the different streams under certain conditions, thus yielding a linear increase in the spectral efficiency. However, this scheme is only effective when the channels of the two streams are sufficiently decorrelated and the received SINR is high.

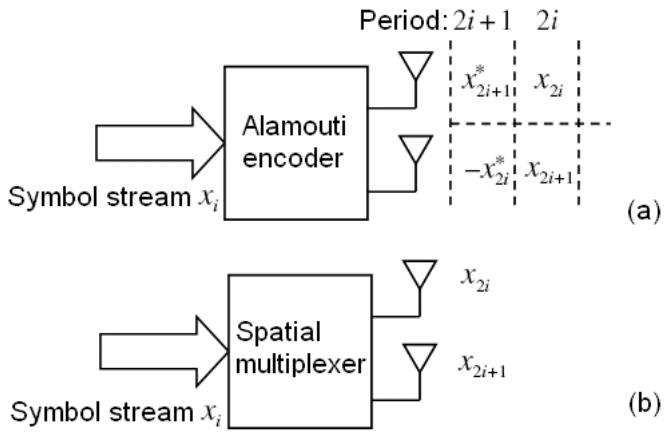

Figure 3.22: Schematic of (a) Alamouti scheme and (b) spatial multiplexing for a transmitter with two antennas.

We apply our methodology to find the flow capacities for these two schemes, where we only need to obtain their resulting rate distributions. We obtain the rate

distributions by applying SINR-to-Rate mapping functions to the SINR distribution of the reuse-1 scheme (obtained using the same parameters in Table 3.2). The mapping functions for the two MIMO schemes are obtained from [89], where the authors fit modified Shannon bounds to simulation results. The general mapping function is

$$S(\text{bits/s/Hz}) = \min\left\{k \cdot \text{BW}_{eff} \cdot \log_2\left(1 + \text{SINR} \,/\, \text{SNR}_{eff}\right), S_{\max}\right\}. \tag{3.53}$$

Equation (3.53) differs from (3.47) in the new parameter k, which represents the number of spatial streams. We apply the results for the MIMO schemes from [89]. The parameters are listed in Table 3.3, and the mapping functions for the two types of MIMO are shown in Figure 3.23, together with the mapping function for SIMO from (3.47).

Table 3.3: SINR-to-Rate mapping parameters for MIMO schemes

Scheme	BW_{eff}	SNR_{eff}	$S_{\max}$	Spatial streams k
2x2 SFC	0.62	1.4 dB	4 bits/s/Hz	1
2x2 BLAST	0.56	2 dB	8 bits/s/Hz	2

We see from Figure 3.23 that SFC has a higher spectral efficiency than V-BLAST when $\text{SINR} < \text{SINR}_{eq}$, where SINR_{eq} is the point where the two schemes achieve the same spectral efficiency. As a result, we apply a dynamic switch mechanism for the use of V-BLAST, which is denoted as hybrid SFC/V-BLAST and works as follows. When $\text{SINR} < \text{SINR}_{eq}$, we use SFC during the transmission; if $\text{SINR} \geq \text{SINR}_{eq}$, we switch to use V-BLAST.

Figure 3.24 illustrates the rate distributions using SIMO, SFC, and hybrid SFC/V-BLAST schemes. We see that due to the transmit diversity gain, SFC improves all the rates including the low rates; but the improvement is only slight because receive diversity already extracts significant diversity gain and so there is not much additional gain that can be obtained from transmit diversity. Using the spatial multiplexing or V-BLAST dynamically with SFC only increases the high rates.

Next, we perform a capacity comparison among the three schemes. Figure 3.25

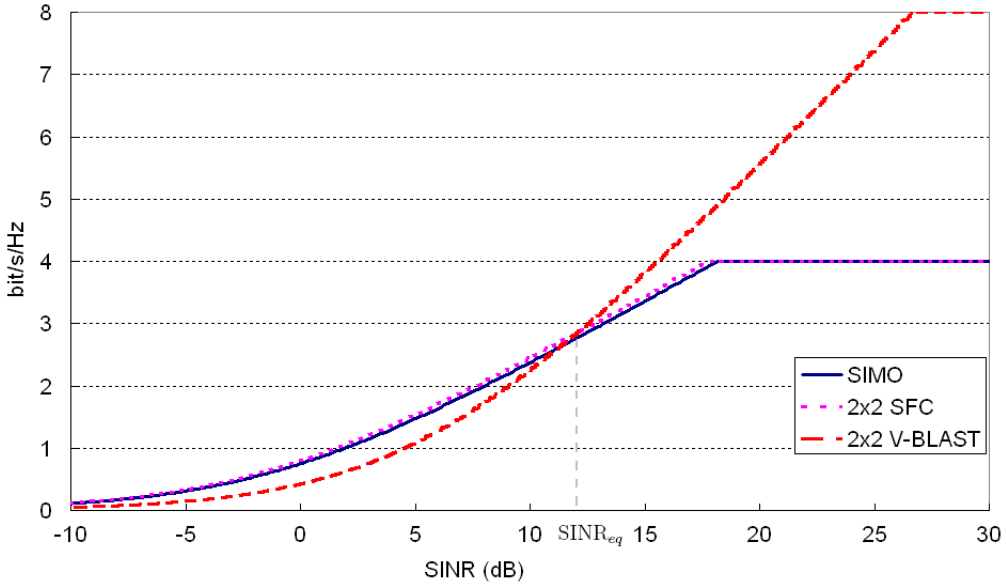

Figure 3.23: SINR-to-Rate mapping functions for the MIMO schemes.

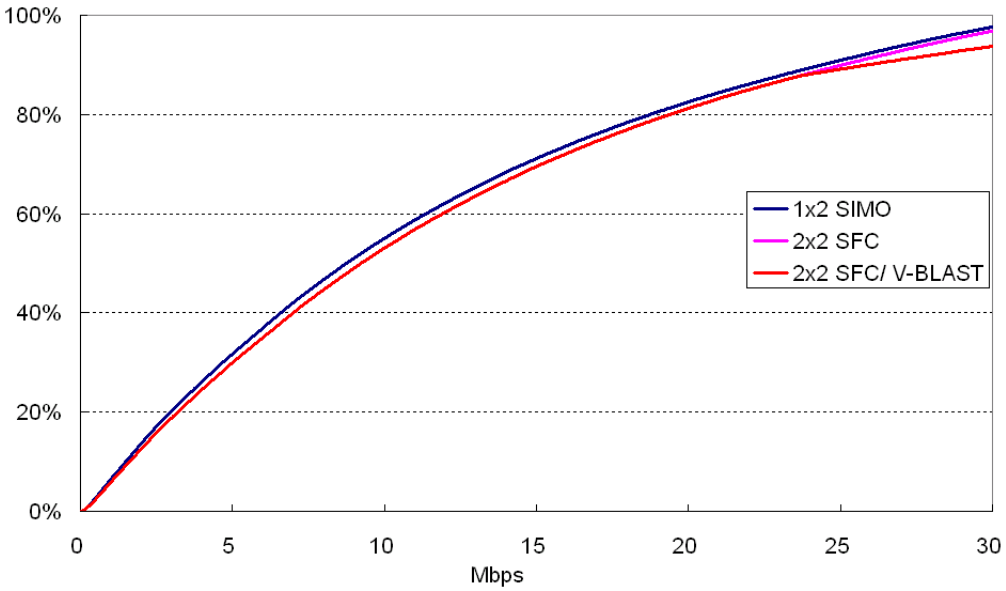

Figure 3.24: Rate distributions of SIMO, SFC and hybrid SFC/V-BLAST schemes.

depicts the flow capacities for the RR case with $\beta = 0.9$ under homogeneous loading. Since the SFC scheme improves the rates slightly relative to the SIMO scheme, there is a small improvement (about 8% gain) with respect to the flow capacity. We also see that the hybrid SFC/V-BLAST scheme only has a slight advantage over the SFC scheme. These results reinforce our findings in Section 3.3.1 that the capacity is strongly influenced by the low rates, and less sensitive to the improvement in the high rates.

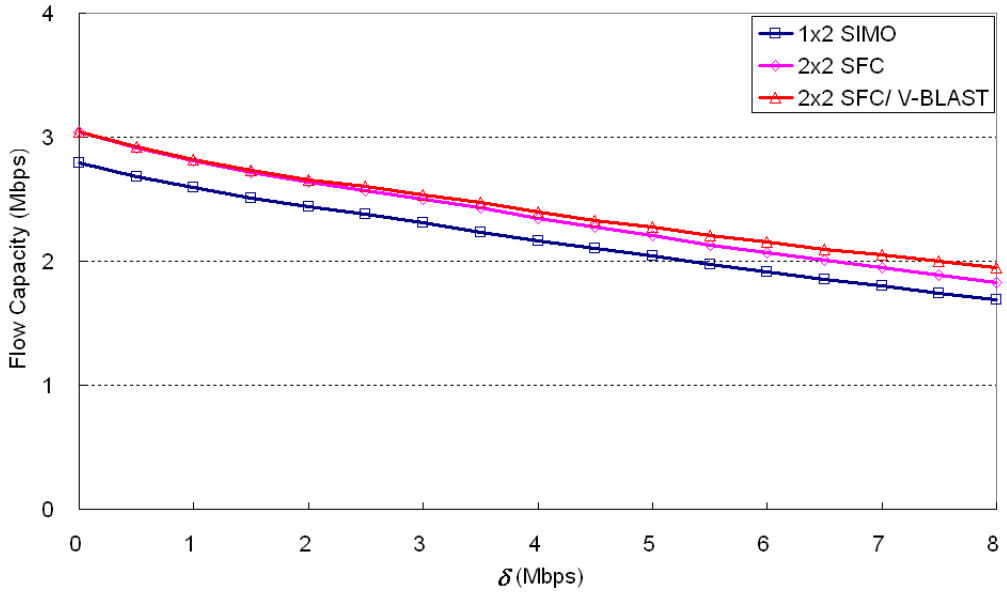

Figure 3.25: Comparison of flow capacities for SIMO, SFC and hybrid SFC/V-BLAST schemes with RR scheduling, homogeneous load, and $\beta = 0.9$.

3.7 Conclusion

In this chapter, we have introduced a new flow capacity metric to characterise the user-level performance for elastic traffic. Assuming Poisson flow arrivals from an infinite user population, we developed a methodology to find the flow capacity for reuse-1, reuse-3, FFR and SFR schemes. We applied the GPS queue to model the base station, which can approximately take the PF scheduling gains into account and be specialized to the EPS queue for the RR scheduler. We also accounted for the effect of interference from neighbouring base stations with an arbitrary level of

loading, and developed an iterative method to estimate the flow capacity when the reference sector has the same loading as the interfering sectors.

We derived a simple expression for the capacity of RR scheduling, which shows that the capacity is strongly influenced by the cell edge rate. For PF scheduling, a fast convergence property of the multi-user diversity gain was exploited, and a computationally efficient algorithm was implemented. The numerical results based on the assumption of elastic traffic indicated that with both RR scheduling and PF scheduling, FFR with a judicious choice of edge band size yields the highest capacity when the flow throughput requirement is low, SFR gives the highest capacity for a moderate throughput requirement, and reuse-1 is the best choice for a high throughput requirement.

We have validated that the per-class flow throughputs obtained using our hybrid simulation/analysis approach exhibit very close agreement with the per-class flow throughputs obtained using a pure simulation approach. Therefore, the hybrid simulation/analysis approach in our framework estimates the flow capacity with dramatically reduced simulation effort, since we only require the simulation of the rate distribution (or SINR distribution). Our framework for evaluating capacity has general applicability. We presented an example to show how our methodology can be applied to evaluate the flow capacity of MIMO schemes. Similarly, it can be also used to find the capacity benefits of other performance enhancements such as relays, higher order modulation, and bandwidth aggregation, provided that we can obtain the rate distributions. Moreover, this framework can be used by network operators to choose optimal base station configurations, or to dimension networks.

Chapter 4

Flow capacity evaluation with a finite user population model

4.1 Introduction

In this chapter, we assume a finite user population and apply a closed queueing network model to find the flow capacity for the downlink for standard reuse schemes, such as reuse-1 and reuse-3, and reuse partitioning schemes, namely FFR and SFR. We analyse user-level performance for these schemes with both RR and PF scheduling.

The *flow capacity* in this chapter is defined as the maximum number of elastic traffic users that can be supported by the system while satisfying a minimum requirement on the flow throughput. Compared to the infinite user population model in the previous chapter, the finite user population model has greater practical relevance. However, this model has the disadvantages that we cannot obtain a compact expression like (3.20) for the flow capacity, and the computational complexity of a direct calculation (especially for the normalisation constant) grows exponentially with the number of classes and the number of users. Therefore, a greater emphasis is required on computationally efficient algorithms to solve for the flow capacity.

A number of previous works have modelled the wireless base station scheduler with a processor sharing queue in a closed queueing network. Shankaranarayanan *et al.* in [122, 123] use a closed queueing network to analyse the user-level performance for a wireless Internet service; however, a weakness of their work is that a single service rate is assumed for all users. Liu and her collaborators [82, 83] consider different service rates corresponding to different channel conditions. They extend the basic model of Bonald and Proutiére [26] to the finite population case for reuse-1

using a closed queueing network model. However, similar to [26], they consider a single omni-directional cell with RR scheduling and assume a simple theoretical rate distribution, where the transmission rate is only a function of distance from the base station. Due to the computational complexity of the closed queueing model, they only calculate the flow throughputs for cases with very few users (at most 5 users in the system).

The parallel work by Maqbool, Coupechoux and Godlewski [85] provides an analytical model based on a closed queueing network to dimension an OFDMA network. They consider both single-category traffic and multi-category traffic models, and apply different scheduling policies. However, for each category traffic, they only apply an aggregate service rate, and do not differentiate the users according to their locations. Hence, the obtained performance, such as the user throughput, is a sector average value. Furthermore, at most three classes of traffic are applied in the analysis because of the computational cost.

In this chapter, we apply the closed queueing network model of [82, 83] for standard reuse schemes to find the flow capacity, and then extend it to cover reuse partitioning schemes. Unlike the EPS queue model in [82, 83], we apply a GPS queue model of the base station to account for the PF scheduling gain, which we specialize to the EPS queue for the RR case. Furthermore, we implement computationally efficient algorithms to solve for the flow throughputs for both RR and PF scheduling in different reuse schemes. In contrast to the simple distance-based rate distributions used in [26, 83], we use more realistic rate distributions generated via system-level simulations, which capture the effects of sectored antennas, shadow fading, path loss, and inter-cell interference. Furthermore, we also consider the multi-cell scenario with fractional loading, and develop an approach to find the flow capacity in a homogeneous load network.

The rest of this chapter has been arranged as follows. In Section 4.2, we first describe the basic assumptions and parameters used in our analysis. Then we present our analytical models to find the flow capacity for standard reuse schemes and reuse partitioning schemes in Sections 4.3 and 4.4, respectively. In Section 4.5, we explore some key aspects of the methodology, and present numerical experiments to illustrate

the capacities of different schemes.

4.2 Basic assumptions and parameters

We analyse the user-level performance in the reference sector by assuming that all interfering sectors have the same load $U_I, 0 \leq U_I \leq 1$. Similar to the infinite user population model (see Section 3.2.1), we approximate fractional loading in the interfering sectors by using independent Bernoulli random variables with success probability U_I to represent the active/inactive states of the interfering sectors.

The finite user population model differs from the infinite user population model in that we use a closed queueing network model in the reference sector. In this model, we assume that a total of M, $M > 0$, users are in the reference sector. The users are uniformly distributed, the traffic of each user is on-off elastic best-effort traffic (see Figure 2.10), and each user only generates one flow at one time and after completion of the transmission there is a thinking time before the next flow. We assume a single-category traffic scenario, where the flow sizes are independent and identically distributed (i.i.d.) with a mean of σ bits and the thinking times are i.i.d. according to an exponential distribution with a mean of $1/\nu$ seconds. The finite user population model can be extended to a multi-category traffic scenario when there are different categories with different distributions and means for the flow size and thinking time.

We assume that there are N resource units in the downlink (with a 1x2 SIMO channel) of an OFDMA wireless system. The available resource units in each sector is N in reuse-1 and $N/3$ in reuse-3.

The hybrid simulation/analysis approach (see Section 3.2.4) is used to find the flow capacity, where a set of time-average rates is the input of the analytical model. For the standard reuse schemes (reuse-1 or reuse-3), we assume that there are K classes of users with discrete time-average rates $\{R_1, \ldots, R_K\}$, with corresponding probabilities $\{P_1, \ldots, P_K\}$, where $\sum_{k=1}^{K} P_k = 1$.

For FFR and SFR, the edge band size n can vary over the set $\{1, \ldots, \lfloor \frac{N}{3} \rfloor\}$, denoted by FFR-$n$ and SFR-n. For tractability, we impose the same restriction as that

in the infinite user population model that centre users can only access centre band resource units. We also define two discrete sets of time-average rates for the input, one for the centre band, $\{R_{c1}, \ldots, R_{cK_c}\}$ with probabilities $\{P_{c1}, \ldots, P_{cK_c}\}$ where $\sum_{k=1}^{K_c} P_{ck} = 1$, and the other for the edge band, $\{R_{e1}, \ldots, R_{eK_e}\}$ with probabilities $\{P_{e1}, \ldots, P_{eK_e}\}$ where $\sum_{k=1}^{K_e} P_{ek} = 1$.

The rates depend on the load U_I in the interfering sectors. We only explicitly show the functional dependence on U_I when it is helpful for understanding. The important mathematic notation used in our analysis is summarized in Table 4.1.

Table 4.1: Mathematical notation

Symbol	Description	Equation where symbol first appears
N	Total number of resource units in the system	–
U_I	Load in all interfering sectors	–
U_r	Load of the reference sector	(4.15)
σ	Mean flow size	(4.6)
$\frac{1}{\nu}$	Mean thinking time	(4.6)
δ	Minimum flow throughput requirement	(4.4)
β	Proportion of users satisfying the throughput requirement	(4.5)
M	Total number of users in the reference sector	(4.1)
K	Number of classes of flows	(4.1)
R_k	Time-average of class-k flows	(4.6)
P_k	Probability that an arbitrary flow belongs to class-k	(4.1)
$\overline{m_k}$	Average number of class-k users	(4.1)
x_k	Number of active class-k users at node 1	(4.7)
y_k	Number of active class-k users at node 0	(4.7)
ρ_{0k}	Offered traffic of class-k with relative arrival rate at node 0	(4.6)
ρ_{1k}	Offered traffic of class-k with relative arrival rate at node 1	(4.6)
ρ_k	ρ_{1k}/ρ_{0k}	(4.6)
T_k	Flow duration of one class-k flow	(4.16)
γ_k	Flow throughput of class-k	(4.19)
g	Limit of PF scheduling gain	(4.30)
s_n	SINR threshold to differentiate the centre/edge users in FFR/SFR	–
p_n	p_n-th quantile of the SINR distribution	(4.60)

4.3 System model for standard reuse schemes

In this section, we present the methodology that uses a finite user population model to find the flow capacity in a homogeneous load network for standard reuse schemes, such as reuse-1 and reuse-3 schemes.

4.3.1 Problem formulation for an arbitrary load U_I in interfering sectors

If a total of M users are in the reference sector, the average number of users $\overline{m_k}$ within class-k is given by

$$\overline{m_k} = MP_k, \quad k = 1, \ldots, K, \tag{4.1}$$

which is assumed to be an integer for simplicity, otherwise rounding can be performed to make the probability mass function approximate P_k. The average population vector is defined as

$$\overline{\mathbf{m}} = (\overline{m_1}, \ldots, \overline{m_K}). \tag{4.2}$$

For simplicity, we calculate the flow throughputs based on this average population vector, which can greatly reduce the computational cost. Technically, the correct approach is to use all possible population combinations as shown in Section 4.3.3 to calculate the flow throughputs, in which the computational complexity grows exponentially with the number of classes and the number of users. However, we note that the performance metrics based on this average population vector are very close to the means considering all possible population combinations, which is discussed in Section 4.3.3.

We define the flow throughput of a class-k user based on the average population vector $\overline{\mathbf{m}}$ as $\gamma_k(\overline{\mathbf{m}}), k = 1, \ldots, K$. We order the classes in the manner of decreasing flow throughputs (breaking any ties randomly) such that

$$\gamma_1(\overline{\mathbf{m}}) \geq \cdots \geq \gamma_K(\overline{\mathbf{m}}). \tag{4.3}$$

The problem of finding the flow capacity of a system can be formulated as

$$\begin{aligned} M^* = \max \quad & M, \\ \text{such that} \quad & \gamma_L \geq \delta, \end{aligned} \tag{4.4}$$

where δ is the minimum flow throughput requirement. We define

$$L = \min \left\{ n : \frac{1}{M} \sum_{k=1}^{n} \overline{m_k} \geq \beta \right\}, \tag{4.5}$$

where β indicates the user satisfaction (see Section 3.3.1 for the reason why we introduce β in our analysis).

The flow throughput $\gamma_k(\overline{\mathbf{m}}), k = 1, \ldots, K$, can be obtained from a closed queueing network model with a finite population. Following [83], the reference sector is modelled as a two-node closed network with K classes of flows (see Figure 4.1).

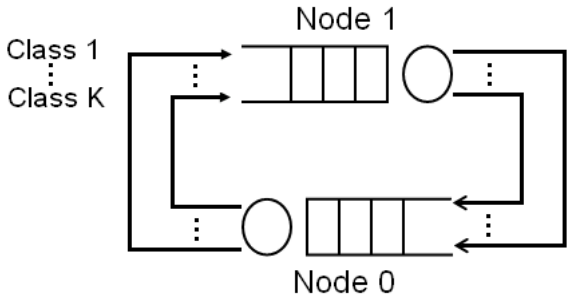

Figure 4.1: A two-node closed network model.

Node 1 is a GPS queue that models the base station, while node 0 is an infinite server (IS) queue that models the users in the thinking state. In node 0, users have an i.i.d. service time with a mean of $1/\nu$ seconds. The relative arrival rate of a class-k flow is denoted by e_{1k} at node 1 and by e_{0k} at node 0, and we have $e_{1k} = e_{0k}$ (and normalise them to 1). The service rate of a class-k flow is $\mu_{0k} = \nu$ at node 0 and $\mu_{1k} = R_k/\sigma$ at node 1, where σ is the mean flow size. We also denote

$$\rho_{1k} = \frac{e_{1k}}{\mu_{1k}} = \frac{\sigma}{R_k}, \quad \rho_{0k} = \frac{e_{0k}}{\mu_{0k}} = \frac{1}{\nu}, \quad \text{and} \quad \rho_k = \frac{\rho_{1k}}{\rho_{0k}} = \frac{\nu\sigma}{R_k}. \tag{4.6}$$

The state of node 1 at an arbitrary time is defined by the vector

$$\mathbf{x} = (x_1, \ldots, x_K), \tag{4.7}$$

where $x_k, k = 1, \ldots, K$, is the number of class-k users in node 1, and we let

$$x = \sum_{k=1}^{K} x_k. \tag{4.8}$$

The corresponding state of node 0 at this instant is

$$\mathbf{y} = (y_1, \ldots, y_K), \tag{4.9}$$

where

$$y_k = \overline{m_k} - x_k, \quad k = 1, \ldots, K. \tag{4.10}$$

By the BCMP theorem for product form networks (see Baskett *et al.* [13]), the stationary distribution of the network is given by

$$\pi_{\overline{\mathbf{m}}}(\mathbf{x}, \mathbf{y}) = H(\overline{\mathbf{m}})^{-1} \chi_{\overline{\mathbf{m}}}(\mathbf{x}) \chi_{\overline{\mathbf{m}}}(\mathbf{y}). \tag{4.11}$$

where $H(\overline{\mathbf{m}})$ is the normalisation constant, given by

$$H(\overline{\mathbf{m}}) = \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}} \chi_{\overline{\mathbf{m}}}(\mathbf{x}) \chi_{\overline{\mathbf{m}}}(\mathbf{y}). \tag{4.12}$$

Node 1 is a GPS queue, so that we have (see Cohen [36])

$$\chi_{\overline{\mathbf{m}}}(\mathbf{x}) = \frac{x!}{\Pi_{i=1}^{x} G(i)} \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right), \tag{4.13}$$

where $G(x)$ is the scheduling gain when there are x users at node 1. Node 0 is an IS queue, so we have

$$\chi_{\overline{\mathbf{m}}}(\mathbf{y}) = \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) = \prod_{k=1}^{K} \left(\frac{1}{(\overline{m_k} - x_k)!} \rho_{0k}^{\overline{m_k} - x_k} \right). \tag{4.14}$$

The probability that there are no active flows at the base station is $\pi_{\overline{\mathbf{m}}}(\mathbf{0}, \overline{\mathbf{m}})$. Therefore, from (2.4), we can obtain the utilization in the reference sector (interpreted as the load of the reference sector):

$$U_r = 1 - \pi_{\overline{\mathbf{m}}}(\mathbf{0}, \overline{\mathbf{m}}). \tag{4.15}$$

Let $E[T_{1k}]$ be the mean duration of class-k flows in node 1. From Little's formula [69], we have

$$E[T_{1k}] = \frac{E[x_k]}{\lambda_{1k}}, \tag{4.16}$$

where $E[x_k]$ is the mean number of class-k flows in node 1 and λ_{1k} is the corresponding arrival rate. From the stationary distribution (4.11), we can easily find

$$\begin{aligned} E[x_k] &= \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}} \pi_{\overline{\mathbf{m}}}(\mathbf{x}, \mathbf{y}) x_k \\ &= H(\overline{\mathbf{m}})^{-1} \rho_{1k} \frac{\partial}{\partial \rho_{1k}} H(\overline{\mathbf{m}}), \end{aligned} \tag{4.17}$$

and

$$\begin{aligned} \lambda_{1k} &= \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}} \pi_{\overline{\mathbf{m}}}(\mathbf{x}, \mathbf{y}) y_k \nu \\ &= H(\overline{\mathbf{m}})^{-1} \nu \rho_{0k} \frac{\partial}{\partial \rho_{0k}} H(\overline{\mathbf{m}}). \end{aligned} \tag{4.18}$$

Thus the flow throughput γ_k of class-k users is

$$\begin{aligned} \gamma_k(\overline{\mathbf{m}}) &:= \frac{\sigma}{E[T_{1k}]} \\ &= R_k \frac{\frac{\partial}{\partial \rho_{0k}} H(\overline{\mathbf{m}})}{\frac{\partial}{\partial \rho_{1k}} H(\overline{\mathbf{m}})}. \end{aligned} \tag{4.19}$$

For standard reuse schemes, we find that the flow throughputs are the same index order as the list of rates in decreasing order,

$$R_1 \geq \cdots \geq R_K. \tag{4.20}$$

The flow throughputs (as well as the flow capacity) are insensitive to the distributions of the file sizes and the distributions of the thinking times, but sensitive to the product of the mean service rate at the IS node ν and the mean file size σ. We show this in the following analysis.

From (4.11), (4.13) and (4.14), we can write

$$\pi_{\overline{\mathbf{m}}}(\mathbf{x},\mathbf{y}) = \hat{H}(\overline{\mathbf{m}})^{-1}\frac{x!}{\Pi_{i=1}^{x}G(i)}\prod_{k=1}^{K}\left(\frac{1}{x_k!(\overline{m_k}-x_k)!}\rho_k^{x_k}\right), \tag{4.21}$$

where $\hat{H}(\overline{\mathbf{m}})$ is a new normalisation constant

$$\begin{aligned}\hat{H}(\overline{\mathbf{m}}) &= \frac{H(\overline{\mathbf{m}})}{\prod_{k=1}^{K}\rho_{0k}^{\overline{m_k}}},\\ &= \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}}\frac{x!}{\Pi_{i=1}^{x}G(i)}\prod_{k=1}^{K}\left(\frac{1}{x_k!(\overline{m_k}-x_k)!}\rho_k^{x_k}\right).\end{aligned} \tag{4.22}$$

Thus we have

$$E[x_k] = \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}}\pi_{\overline{\mathbf{m}}}(\mathbf{x},\mathbf{y})x_k = \hat{H}(\overline{\mathbf{m}})^{-1}\rho_k\frac{\partial}{\partial\rho_k}\hat{H}(\overline{\mathbf{m}}), \tag{4.23}$$

$$\lambda_{1k} = \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}}\pi_{\overline{\mathbf{m}}}(\mathbf{x},\mathbf{y})(\overline{m_k}-x_k)\nu = \hat{H}(\overline{\mathbf{m}})^{-1}\nu\hat{H}(\overline{\mathbf{m}}-\mathbf{e_k}), \tag{4.24}$$

where $\mathbf{e_k}$ is a K-vector with 1 at position k and 0 elsewhere. The flow throughput $\gamma_k(\overline{\mathbf{m}})$ is

$$\gamma_k(\overline{\mathbf{m}}) = R_k\frac{\hat{H}(\overline{\mathbf{m}}-\mathbf{e_k})}{\frac{\partial}{\partial\rho_k}\hat{H}(\overline{\mathbf{m}})}. \tag{4.25}$$

Equations (4.22) and (4.25) show that the only dependence that the flow throughputs have on the traffic parameters is through ρ_k. Since $\rho_k, k = 1,\ldots,K$, defined in (4.6) only depends on the product of ν and σ, we can conclude that the flow throughputs depend on $\nu\sigma$.

Round-robin scheduling

For the RR case, we have $G(x) \equiv 1$ for all x, and so it follows from (4.12)-(4.14) that the normalisation constant is

$$H_{RR}(\overline{\mathbf{m}}) = \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}} \left[x! \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right) \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) \right]. \tag{4.26}$$

The principal computational difficulty associated with solving (4.4) resides in the determination of the normalisation constant (4.26). For a direct summation, the time and space requirements increase exponentially with the number of classes of flows. For example, if there are $M = 100$ users and $K = 10$ classes in the analysis, the number of states of the system will be more than 10^{10}, which is computationally prohibitive for a direct calculation.

Conway and Georganas in [37] provide an efficient recursive algorithm (illustrated in Theorem 1 of [37]) to calculate the normalisation constant, and an algorithm called the *RECAL algorithm* to calculate user-level performance such as the mean flow duration.

We can apply the algorithm for the normalisation constant in [37] to our special case. To do so, we set $H_{RR}(\overline{\mathbf{m}}) \equiv H_K(\mathbf{0})$, where $H_K(\mathbf{0})$ can be obtained recursively from

$$H_k(\mathbf{v}_k) = \sum_{\mathbf{n} \in \Psi_k} h_k(\mathbf{v}_k, \mathbf{n}) H_{k-1}(\mathbf{v}_k + \mathbf{n}), \tag{4.27}$$

where $1 \leq k \leq K, \mathbf{v}_k \in \Phi_k$ and

$$\mathbf{v}_k = (v_{1k}, v_{0k}), \text{ where } v_{1k}, v_{0k} \text{ are non-negative integers,}$$

$$\mathbf{n} = (n_1, n_0), \text{ where } n_1, n_0 \text{ are non-negative integers,}$$

$$\Phi_k = \begin{cases} \left\{ \mathbf{v}_k | v_{0k} + v_{1k} = \sum_{s=k+1}^{K} \overline{m_s} \right\} & \text{if } 0 \leq k \leq K-1, \\ \{\mathbf{0}\} & \text{if } k = K, \end{cases}$$

$$\Psi_k = \left\{ \mathbf{n} | n_1 + n_0 = \overline{m_k} \right\},$$

$$h_k(\mathbf{v}_k, \mathbf{n}) = \begin{pmatrix} n_1 + v_{1k} \\ v_{1k} \end{pmatrix} \rho_{1k}^{n_1} \cdot \frac{\rho_{0k}^{n_0}}{n_0!},$$

and the initial conditions are $H_0(\mathbf{v}_0) = 1$ for all $\mathbf{v}_0 \in \Phi_0$.

Instead of the approach described above, a more efficient approach is to apply another recursion algorithm called the *MVAC algorithm* (see Conway and his collaborators [38]) to calculate the mean number of class-k flows at node 1 $E[x_k]$ and its arrival rate λ_{1k}, and hence the flow throughputs $\gamma_k(\overline{\mathbf{m}})$. This approach avoids the need to compute the normalisation constant. The MVAC algorithm is computationally efficient for networks with few nodes and many users, hence it is well-suited to our finite user population model using RR scheduling.

From (4.27) and the MVAC algorithm, we obtain $H_{RR}(\overline{\mathbf{m}})$, $E[x_k]$ and λ_{1k}. For later use (for the PF case), we rearrange (4.17) and (4.18) to obtain the following relationships

$$\frac{\partial}{\partial \rho_{1k}} H_{RR}(\overline{\mathbf{m}}) = E[x_k] H_{RR}(\overline{\mathbf{m}}) \rho_{1k}^{-1}, \tag{4.28}$$

$$\frac{\partial}{\partial \rho_{0k}} H_{RR}(\overline{\mathbf{m}}) = \lambda_{1k} H_{RR}(\overline{\mathbf{m}}) (\nu \rho_{0k})^{-1}. \tag{4.29}$$

Proportional fair scheduling

In Section 3.5.1, we have shown that the gain function $G(x)$ of PF scheduling converges quickly to some limit $g(g > 1)$ for $x > x_0$, where the value of x_0 is not very large. For convenience, we recall two definitions here:

$$C = g^{-x_0} \prod_{i=1}^{x_0} G(i), \quad \text{and} \quad D(x) = \begin{cases} \frac{1}{Cg^x} - \frac{1}{\prod_{i=1}^{x} G(i)} & x > 0 \\ \frac{1}{C} - 1 & x = 0 \end{cases}. \tag{4.30}$$

From (4.12)-(4.14) and (4.30), we can write the normalisation constant in (4.12) for PF scheduling as

$$\begin{aligned} H_{PF}(\overline{\mathbf{m}}) &= \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}} \left[\frac{x!}{\prod_{i=1}^{x} G(i)} \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right) \cdot \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) \right] \\ &= \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}, x>x_0} \left[\frac{x!}{Cg^x} \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right) \cdot \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) \right] \\ &\quad + \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}, x \le x_0} \left[x! \left(\frac{1}{Cg^x} - D(x) \right) \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right) \cdot \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) \right] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}} \left[\frac{x!}{Cg^x} \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right) \cdot \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) \right] \\
&\quad - \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}, x \leq x_0} \left[x! D(x) \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right) \cdot \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) \right] \qquad (4.31) \\
&\equiv H_1(\overline{\mathbf{m}}) - H_2(\overline{\mathbf{m}}),
\end{aligned}
$$

where

$$
H_1(\overline{\mathbf{m}}) = \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}} \left[\frac{x!}{Cg^x} \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right) \cdot \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) \right], \qquad (4.32)
$$

$$
H_2(\overline{\mathbf{m}}) = \sum_{\mathbf{x}+\mathbf{y}=\overline{\mathbf{m}}, x \leq x_0} \left[x! D(x) \prod_{k=1}^{K} \left(\frac{1}{x_k!} \rho_{1k}^{x_k} \right) \cdot \prod_{k=1}^{K} \left(\frac{1}{y_k!} \rho_{0k}^{y_k} \right) \right]. \qquad (4.33)
$$

From (4.31), (4.17) and (4.18), the mean number and corresponding arrival rate of class-k flows in node 1 are

$$
E[x_k] = H_{PF}(\overline{\mathbf{m}})^{-1} \rho_{1k} \left(\frac{\partial}{\partial \rho_{1k}} H_1(\overline{\mathbf{m}}) - \frac{\partial}{\partial \rho_{1k}} H_2(\overline{\mathbf{m}}) \right), \qquad (4.34)
$$

$$
\lambda_{1k} = H_{PF}(\overline{\mathbf{m}})^{-1} \nu \rho_{0k} \left(\frac{\partial}{\partial \rho_{0k}} H_1(\overline{\mathbf{m}}) - \frac{\partial}{\partial \rho_{0k}} H_2(\overline{\mathbf{m}}) \right). \qquad (4.35)
$$

The first parts in (4.31), (4.34) and (4.35) can be calculated using (4.26), (4.28) and (4.29) for RR scheduling since $H_1(\overline{\mathbf{m}})$ can be put in the form of $H_{RR}(\overline{\mathbf{m}})$ by multiplying by a constant C and letting $\rho'_{1k} = \rho_{1k}/g$. The second parts of these equations can be calculated directly since the value of x_0 is not very large. Consequently, the flow throughput of a class-k user in PF scheduling can be found from

$$
\begin{aligned}
\gamma_k(\overline{\mathbf{m}}) &= \frac{\sigma \lambda_{1k}}{E[x_k]} \\
&= R_k \frac{\frac{\partial}{\partial \rho_{0k}} H_1(\overline{\mathbf{m}}) - \frac{\partial}{\partial \rho_{0k}} H_2(\overline{\mathbf{m}})}{\frac{\partial}{\partial \rho_{1k}} H_1(\overline{\mathbf{m}}) - \frac{\partial}{\partial \rho_{1k}} H_2(\overline{\mathbf{m}})}. \qquad (4.36)
\end{aligned}
$$

Rules for allocating users to rate bins

In our model, the user population of each class must be an integer. In general, we cannot ensure that (4.1) always returns integer values, and so we must perform a

rounding operation. The use of rounding can introduce ambiguity in how users are allocated to bins, so we define rules in the following to remove any such ambiguity.

To allocate the M users to the K "bins" (classes), we always define equi-probability bins ($P_1 = \cdots = P_K$) in our numerical examples and perform the following procedure. First, we allocate the bulk of the users according to the probabilities, $\lfloor MP_1 \rfloor$ for each bin. Then, starting by allocating a user to class-1, we allocate any leftover users $M - K\lfloor MP_1 \rfloor$ so that they are equi-spaced over the range of bins. An example for the case of $K = 10$ is shown in Figure 4.2. By doing this, we tend to get a nice and gradual change in the capacity results as the number of rate bins is changed. In contrast, if we place the leftover users in the lowermost rate bins, we see spurious differences in the capacity results as the number of rate bins is varied.

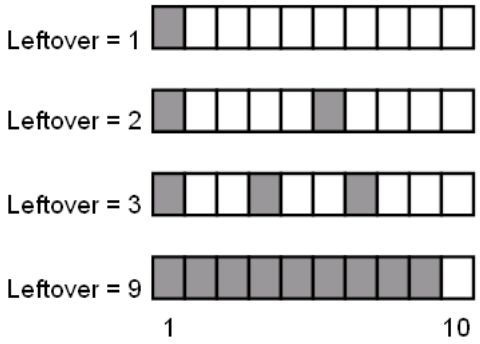

Figure 4.2: An example of the rules for allocating the leftover users to rate bins ($K = 10$).

An algorithm to find the flow capacity

Similar to Algorithm 3.1, we define a bisection search to find the flow capacity in problem (4.4) for the given minimum flow throughput requirement δ and the user satisfaction proportion β. It is necessary to define an upper bound (M_{sat}) and a lower bound (M_l) for the bisection search.

We apply a feasibility check before the bisection search in Algorithm 4.1, where we assume one user in the system and test whether the flow throughput satisfies the minimum throughput requirement. According to the above rules for allocating the

users to bins, the only user is assigned to class-1. If the requirement is not satisfied, the problem (4.4) is not feasible.

However, unlike the cell capacity in the infinite user population model, there does not exist a universal upper bound for the flow capacity since for $\delta = 0$, $M_{sat} \to \infty$. One method to find an upper bound is to choose an arbitrarily large value M_{sat} that does not satisfy the constraint of (4.4). Alternatively, we can apply a heuristic approach. Kleinrock in [70] determines an approximate saturation point for the two-node closed queueing network, which is defined by

$$M_{appr} \triangleq \left\lfloor \frac{1/\mu + 1/\nu}{1/\mu} \right\rfloor, \tag{4.37}$$

where $1/\mu$ is the average service time and $1/\nu$ is the mean thinking time. We find in our numerical experiments that it is sufficient to let

$$M_{sat} = 2M_{appr} \tag{4.38}$$

for typical $\delta > 0.1$ Mbps.

Thus we can perform a bisection search between M_l and M_{sat} to find the flow capacity. The basic algorithmic steps are defined in Algorithm 4.1.

Algorithm 4.1

(i) *Feasibility check.* Let $M_l = 1$, assign the user to class-1, and solve for the flow throughput based on the MVAC algorithm for the RR case and (4.36) for the PF case. If the constraint of the problem (4.4) is satisfied, let $M_u = M_{sat}$ and go to step (ii); otherwise the problem is infeasible (return $M^* = 0$) and stop.

(ii) *Bisection search.* If $M_u - M_l = 1$, $M^* = M_l$ and stop; otherwise let $M = \lfloor (M_l + M_u)/2 \rfloor$ and solve for the flow throughput γ_L. If the constraint of (4.4) is satisfied, $M_l = M$; otherwise $M_u = M$. Go to step (ii).

4.3.2 Finding the flow capacity for $U_r = U_I$

We aim for a homogeneous load network, thus the next question we address is how to find the flow capacity such that $U_r = U_I$.

For a given β, to obtain a flow capacity for a single δ value, we can define an algorithm similar to Algorithm 3.2 (see Section 3.3.2). The only difference is located in step (iii) in which we apply Algorithm 4.1 to solve for the flow capacity M^* and use equation (4.15) to obtain the associated utilization U_r in the reference sector. Letting ε_1 be the error tolerance, we define the steps in Algorithm 4.2.

Algorithm 4.2

(i) Start with an arbitrary initial load in the interfering sectors, for example let $U_I = 0.5$.

(ii) Run the system-level simulation to generate the time-average rates $\{R_1(U_I), \cdots, R_K(U_I)\}$.

(iii) Find M^* using Algorithm 4.1, and obtain the associated utilization U_r using (4.15).

(iv) If $|U_r - U_I| > \varepsilon_1$, set $U_I = U_r$ and go to step (ii); otherwise, stop.

Instead of a single point, we are often interested in a homogeneous load curve of M^* versus the flow throughput requirement δ. This could be determined using repeated applications of Algorithm 4.2. However, this can require a large number of simulations of rate distributions. In the following, we describe a more efficient method to find the homogeneous load curve. We perform a series of simulations beforehand for a set of loads $\{U_{I1}, \ldots, U_{IJ}\}$ in the interfering sectors, where $U_{Ij} \in [0, 1]$ for $1 \leq j \leq J$.

For each load U_{Ij}, we obtain a rate distribution and then discretize it into a set of time-average rates $\{R_1(U_{Ij}), \ldots, R_K(U_{Ij})\}$ with probabilities $\{P_1, \ldots, P_K\}$. Then, for load U_{Ij}, we use Algorithm 4.3 to find a (δ, M^*) pair such that the utilization (or load) in the reference sector U_r associated with M^* is closest to U_{Ij} and the constraint in the problem (4.4) is satisfied for the pre-defined β. We choose the initial values of M_l, M_u as in Algorithm 4.1, and perform a bisection section between M_l and M_u. Typically we find that the algorithm finishes in less than 10 iterations in our numerical examples.

In Algorithm 4.3, there are two special cases that we need to treat differently.

The first one is the case of $U_{Ij} = 0$, where we let $M^* = 0$ and define δ as

$$\delta = \max\{R_1(U_{Ij}), \ldots, R_K(U_{Ij})\}. \tag{4.39}$$

The second case is when $U_{Ij} = 1$. We define ε_2 for the error tolerance, and stop the iterations when $|U_r - U_{Ij}| < \varepsilon_2$, as shown in the step (ii) in Algorithm 4.3.

Furthermore, at the last iteration when $M_u - M_l = 1$, we let U_{rl} and U_{ru} be the utilizations associated with M_l and M_u, respectively, and choose M^* as

$$M^* = \begin{cases} M_u & \text{if } |U_{rl} - U_{Ij}| > |U_{ru} - U_{Ij}| \\ M_l & \text{otherwise} \end{cases}. \tag{4.40}$$

By applying Algorithm 4.3 for each U_{Ij}, we can finally obtain a set of (δ, M^*) pairs which gives an approximate homogeneous load curve.

Algorithm 4.3

(i) Let $M_l = 1$, and use (4.15) to find the associated load U_r. If $U_r - U_{Ij} < 0$, go to step (ii); otherwise let $M^* = 0$ and choose δ by (4.39), and stop.

(ii) Let $M_u = 2M_{appr}$, and find U_r using (4.15). If $U_r - U_{Ij} > 0$, go to step (iii). Otherwise, if $|U_r - U_{Ij}| > \varepsilon_2$, $M_u = 2M_u$ and go to step (ii); otherwise let $M^* = M_u$, solve for the flow throughput γ_L associated with M^*, define $\delta = \gamma_L$ and stop.

(iii) Let $M = \lfloor (M_l + M_u)/2 \rfloor$ and use (4.15) to find U_r. If $U_r - U_{Ij} > 0$, $M_u = M$, otherwise $M_l = M$. If $M_u - M_l > 1$, go to step (iii). Otherwise, define M^* using (4.40), solve for the associated flow throughput γ_L, and define $\delta = \gamma_L$ and stop.

4.3.3 Considering all possible population combinations

In Section 4.3.1, the performance metrics, such as the mean queue size and flow throughput, are calculated based on the average population vector $\overline{\mathbf{m}} = (\overline{m_1}, \ldots, \overline{m_K})$ with $\overline{m_k} = MP_k, k = 1, \ldots, K$, when there are M users in the reference sector. The flow throughput of class-k user is $\gamma_k(\overline{\mathbf{m}})$, and we denote the arrival rate (4.18) as $\lambda_k^c(\overline{\mathbf{m}}), k = 1, \ldots, K$, in this section.

As mentioned previously, any population combination $\mathbf{m} = (m_1, \ldots, m_K)$ belonging to the set $\mathfrak{M} = \{(m_1, \ldots, m_K) | \sum_{k=1}^{K} m_k = M\}$ is feasible. The probability for this population combination is obtained from a multinomial distribution, which is

$$P(\mathbf{m}) = \frac{M!}{m_1! \cdots m_K!} P_1^{m_1} \cdots P_K^{m_K}. \tag{4.41}$$

For a fixed population combination $\mathbf{m}$, we can use (4.18) and (4.19) to calculate the arrival rate $\lambda_k^c(\mathbf{m})$ and the flow throughput $\gamma_k(\mathbf{m})$. Therefore we can calculate the averaged arrival rate and the averaged flow throughput over all population combinations

$$E[\lambda_k^c] = \sum_{\mathbf{m} \in \mathfrak{M}} P(\mathbf{m}) \lambda_k^c(\mathbf{m}), \tag{4.42}$$

$$E[\gamma_k] = \sum_{\mathbf{m} \in \mathfrak{M}} P(\mathbf{m}) \gamma_k(\mathbf{m}). \tag{4.43}$$

Unfortunately, for a typical class number and user population ($K = 10$ and M can be greater than 100 in our numerical experiments), it is computationally prohibitive to obtain $E[\gamma_k]$ over all population combinations. This is why we use the mean population $\overline{\mathbf{m}}$ and $\gamma_k(\overline{\mathbf{m}})$ as a proxy for $E[\gamma_k]$. Intuitively, we expect that

$$\gamma_k(\overline{\mathbf{m}}) \approx E[\gamma_k]. \tag{4.44}$$

If this is true, we can use $\gamma_k(\overline{\mathbf{m}})$ estimate the flow capacity with dramatically reduced computational complexity. To test (4.44), we perform the comparison using a simplified experiment: RR scheduling is applied, and there are only $K = 4$ classes of flows with time-average rates $\{18.741, 7.626, 2.962, 0.808\}$ Mbps and probabilities $\{0.4, 0.3, 0.2, 0.1\}$.

Figure 4.3 shows the comparisons of flow throughputs $\gamma_k(\overline{\mathbf{m}})$ based on the average population and averaged flow throughput $E[\gamma_k]$ over all population combinations $\mathbf{m} \in \mathfrak{M}$ for different numbers of users in this test. We see that the values are quite close to each other, so we conjecture that this is also true in the general case.

As far as we are aware, there is no known theoretical proof for (4.44). In the

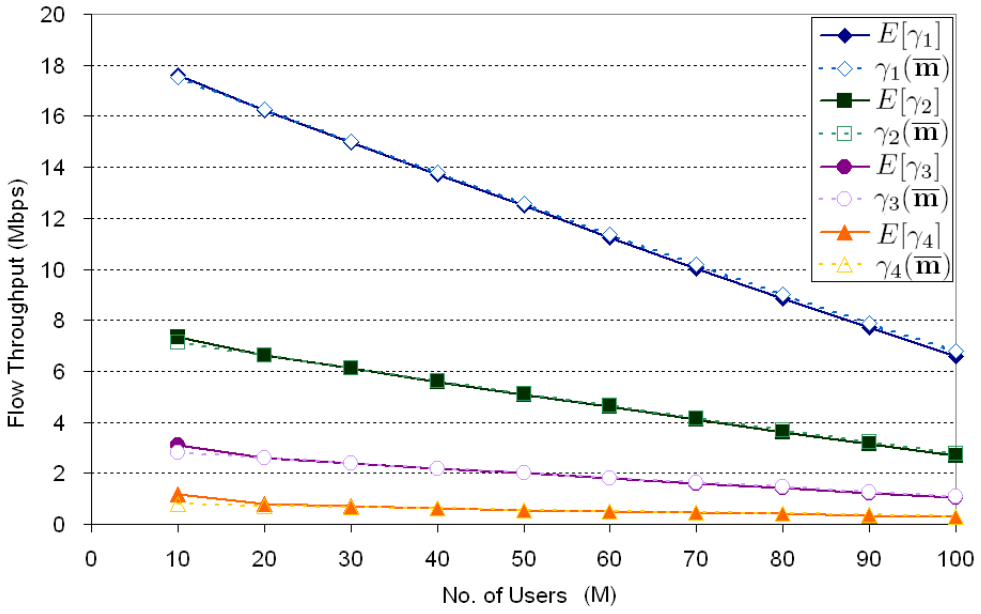

Figure 4.3: Comparison of flow throughputs with population combinations.

following, we apply some theorems of Whitt [138] to show that under some conditions this conjecture is asymptotically true. First, it is necessary to introduce the fixed-population-mean (FPM) approach of [138], which approximates the closed queueing network with an open queueing network. For simplicity, we assume that RR scheduling is applied at the base station. We impose the restriction that the population vector $\mathbf{m}$ is in the feasible set, namely $\mathbf{m} \in \mathfrak{M}$, which is not necessary in the more general setting of [138].

Recall that our two-node closed queueing network has a population vector $\mathbf{m}$, a service rate at the GPS node of $\mu_k = R_k/\sigma$ for class-k, and a mean thinking time of $1/\nu$ for class-k. To apply the FPM method to our two-node closed queueing network, we need to remove the IS node (node 0) and replace the departure processes of the K classes of the IS node by K external Poisson arrival processes to the GPS node (node 1). Then we can obtain an open queueing network (see Figure 4.4).

In the open network without the IS node which is used to approximate the closed queueing network with population vector $\mathbf{m}$, let $\lambda_k^o(\mathbf{m}), k = 1, \ldots, K$, be the external arrival rate for class-k at the GPS node. From (3.18), we can obtain the mean number of class-k users for arrival rate vector $(\lambda_1^o(\mathbf{m}), \ldots, \lambda_K^o(\mathbf{m}))$ for the RR

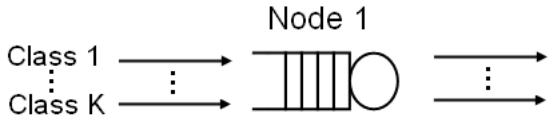

Figure 4.4: An open network approximation using the FPM approach.

case, which is

$$E\left[x_k|\left(\lambda_1^o(\mathbf{m}),\ldots,\lambda_K^o(\mathbf{m})\right)\right]=\frac{\frac{\lambda_k^o(\mathbf{m})}{\mu_k}}{1-\sum_{j=1}^{K}\frac{\lambda_j^o(\mathbf{m})}{\mu_j}}. \tag{4.45}$$

In the FPM method, Whitt [138] proposes an iterative algorithm to find the external arrival rates, where $\phi_k^{(i)}$ is the ith lower bound and $\psi_k^{(i)}$ is the ith upper bound for $\lambda_k^o(\mathbf{m}), k=1,\ldots,K$,

$$\psi_k^{(1)} = m_k\nu, \tag{4.46}$$

$$\phi_k^{(i)} = \left(m_k - E\left[x_k|\left(\psi_1^{(i)},\ldots,\psi_K^{(i)}\right)\right]\right)\nu, \quad i\geq 1, \tag{4.47}$$

$$\psi_k^{(i+1)} = \left(m_k - E\left[x_k|\left(\phi_1^{(i)},\ldots,\phi_K^{(i)}\right)\right]\right)\nu, \quad i\geq 1. \tag{4.48}$$

As $i\to\infty$, we have

$$\phi_k^{(i)}\to\lambda_k^o(\mathbf{m}) \quad \text{and} \quad \psi_k^{(i)}\to\lambda_k^o(\mathbf{m}). \tag{4.49}$$

Note that other approximation methods proposed by Buzen [31] and Protopapas [104] only use the first upper bound in the FPM method to approximate the external arrival rates.

In the following, we apply the ideas of [138] and consider a sequence of two-node closed queueing networks indexed by superscript n. In the nth network, let ν^n be the service rate of the IS node, $m_k^n, k=1,\ldots,K$, be the fixed population of class-k users, and $N_k^{cn}(0)$ be the number of class-k users at the GPS node in the initial state; m_k^n increases and ν^n decreases as n increases, but the product of them $m_k^n\nu^n$ approaches some fixed value. Let $N_k^o(0)$ be the number of class-k users at the GPS node in the initial state of the open network where the IS node is deleted according to the FPM method.

Theorem 8 of [138] shows that the stochastic processes in the GPS node in the closed queueing network converge in distribution to those in the open queueing network as $n \to \infty$ if initially $N_k^{cn}(0) \to N_k^o(0), k = 1, \ldots, K$. Furthermore, Theorem 12 of [138] shows that under these conditions, the FPM method is asymptotically correct.

Whitt shows that from these two theorems, the arrival rate to the GPS node in the closed queueing network is asymptotically equivalent to the Poisson arrival rate in the open queueing network under the conditions of Theorem 12 of [138]. In other words, $\lambda_k^c(\mathbf{m}) \to \lambda_k^o(\mathbf{m})$, where $\lambda_k^o(\mathbf{m})$ is calculated by (4.46)-(4.48). Then we use these results to show that under these conditions, $\lambda_k^c(\overline{\mathbf{m}}) \to E[\lambda_k^c]$ and $\gamma_k(\overline{\mathbf{m}}) \to E[\gamma_k]$ (see Theorem 4.1 and Corollary 4.1).

Theorem 4.1. *If $m_k^n \to \infty$, $\nu^n \to 0$, $m_k^n \nu^n \to \lambda_k^o(\mathbf{m})$, and $N_k^{cn}(0) \to N_k^o(0)$ for each k as $n \to \infty$, then we have*

$$\lambda_k^c(\overline{\mathbf{m}}) \to E[\lambda_k^c]. \tag{4.50}$$

Proof. For the average population vector $\overline{\mathbf{m}} = (\overline{m_1}, \ldots, \overline{m_K})$, the conditions of the theorem hold, namely

$$\overline{m_k} \to \infty, \nu \to 0 \text{ and } \overline{m_k}\nu \to \lambda_k^o(\overline{\mathbf{m}}),$$

where $\lambda_k^o(\overline{\mathbf{m}})$ is the arrival rate of class-k in the open queueing network approximated by the FPM method for the population vector $\overline{\mathbf{m}}$. From the two theorems of [138], we obtain

$$\lambda_k^c(\overline{\mathbf{m}}) \to \lambda_k^o(\overline{\mathbf{m}}),$$

which implies

$$\lambda_k^c(\overline{\mathbf{m}}) \to \overline{m_k}\nu.$$

For a feasible population vector $\mathbf{m} \in \mathfrak{M}$, the conditions of the theorem also hold. Also from the theorems of [138] we have

$$\lambda_k^c(\mathbf{m}) \to \lambda_k^o(\mathbf{m}) \text{ and } \lambda_k^c(\mathbf{m}) \to m_k\nu.$$

Under these conditions, $\lambda_k^c(\mathbf{m})$ is asymptotically linear in m_k, from which we have

$$E[\lambda_k^c] \to E[m_k \nu] = \overline{m_k}\nu.$$

Hence, we can obtain

$$\lambda_k^c(\overline{\mathbf{m}}) \to E[\lambda_k^c].$$

□

Corollary 4.1. *Under the conditions of Theorem 4.1,* $\gamma_k(\overline{\mathbf{m}}) \to E[\gamma_k]$.

Proof. From the two theorems of [138], the arrival process to the GPS node in the closed queueing network is asymptotically equivalent to a Poisson process under the conditions of the theorem. Therefore, from (3.19), we obtain

$$\gamma_k(\overline{\mathbf{m}}) \to R_k \left(1 - \sum_{j=1}^{K} \frac{\lambda_j^c(\overline{\mathbf{m}})}{\mu_j}\right). \tag{4.51}$$

From the proof of Theorem 4.1, we have

$$\gamma_k(\overline{\mathbf{m}}) \to R_k \left(1 - \sum_{j=1}^{K} \frac{\overline{m_j}\nu}{\mu_j}\right). \tag{4.52}$$

For a feasible population vector $\mathbf{m} \in \mathfrak{M}$, the conditions also hold, so we have

$$\gamma_k(\mathbf{m}) \to R_k \left(1 - \sum_{j=1}^{K} \frac{m_j \nu}{\mu_j}\right), \tag{4.53}$$

which implies that $\gamma_k(\mathbf{m})$ is asymptotically linear with $m_k, k = 1, \ldots, K$. Hence we obtain

$$E[\gamma_k] \to R_k \left(1 - \sum_{j=1}^{K} \frac{E[m_j]\nu}{\mu_j}\right) = R_k \left(1 - \sum_{j=1}^{K} \frac{\overline{m_j}\nu}{\mu_j}\right). \tag{4.54}$$

Therefore, we have

$$\gamma_k(\overline{\mathbf{m}}) \to E[\gamma_k]. \tag{4.55}$$

□

From Corollary 4.1, we can conclude that the results using the flow throughputs

obtained from the average population vector to find the flow capacity in problem (4.4), which is much less computationally complex, are asymptotically close to those based on the average flow throughputs over all population combinations.

4.4 System model for FFR and SFR schemes

4.4.1 Problem formulation for an arbitrary load U_I in interfering sectors

For FFR-n/SFR-n scheme (with edge band size n, $n \in \{1, \ldots, \lfloor \frac{N}{3} \rfloor\}$), we choose an SINR threshold s_n, which equals the p_n-th quantile of the SINR distribution on the centre band. We simplify the resource access restrictions (see Section 4.2), so we can model the reference sector using two separate two-node closed queueing networks, namely a centre network and an edge network. Therefore, for an arbitrary load U_I in all interfering sectors, we need two discrete sets of rate distributions as the input of the analytical model; $\{R_{c1}, \ldots, R_{cK_c}\}$ with probabilities $\{P_{c1}, \ldots, P_{cK_c}\}$ for the centre network, and $\{R_{e1}, \ldots, R_{eK_e}\}$ with probabilities $\{P_{e1}, \ldots, P_{eK_e}\}$ for the edge network.

If a total of M users are in the reference sector, the populations in the centre and edge networks are

$$M_c = M(1 - p_n) \quad \text{and} \quad M_e = Mp_n. \tag{4.56}$$

Since M_c and M_e must be integers, we perform the following rounding process

$$M_c = \text{round}\,(M(1 - p_n)) \quad \text{and} \quad M_e = M - M_c, \tag{4.57}$$

where round(x) is a function that rounds the argument to the nearest integer.

The average population vectors in the two networks are

$$\overline{\mathbf{m_c}} = (\overline{m_{c1}}, \ldots, \overline{m_{cK_c}}) \quad \text{and} \quad \overline{\mathbf{m_e}} = (\overline{m_{e1}}, \ldots, \overline{m_{eK_e}}), \tag{4.58}$$

where $\overline{m_{ck}} = P_{ck}M_c, k = 1, \ldots, K_c$, and $\overline{m_{ek}} = P_{ek}M_e, k = 1, \ldots, K_e$, are the average populations for class-k users in the centre and edge networks, respectively. Since the populations must have integer values, we apply the rules in Section 4.3.1 to assign the users to the classes in each network.

From Theorem 4.1 and Corollary 4.1 in Section 4.3.3, we can approximate the user-level performance using the average population vectors in the centre/edge network to greatly decrease the computational complexity. Let $\gamma_{ck}(\overline{\mathbf{m_c}})$ be the flow throughput of class-k centre users, and $\gamma_{ek}(\overline{\mathbf{m_e}})$ be the flow throughput of class-k edge users. We can obtain them using (4.19) by solving the two networks separately, and then sort them in one list in order of decreasing flow throughputs such that

$$\gamma_1 \geq \cdots \geq \gamma_K, \tag{4.59}$$

where $K = K_c + K_e$. For a class-k user, $\gamma_k = \gamma_{ck_c}(\overline{\mathbf{m_c}})$ (or $\gamma_{ek_e}(\overline{\mathbf{m_e}})$) if he belongs to the centre network (or the edge network). Therefore, the problem formulation (4.4) can be applied in a similar way to find the flow capacity in FFR-n (or SFR-n).

To find the flow capacity, we apply a bisection search in a similar way to Algorithm 4.1 but with the following modifications. Firstly, we need to re-define the approximate saturation point, so that it is now given by

$$M_{appr} \triangleq \min\left\{ \left\lfloor \frac{M_{c-appr}}{1-p_n} \right\rfloor, \left\lfloor \frac{M_{e-appr}}{p_n} \right\rfloor \right\}, \tag{4.60}$$

where M_{c-appr} and M_{e-appr} are the approximate saturation points for the centre/edge network, which can be obtained separately using (4.37). Secondly, during each step, we need to divide the total number of users M into two networks using (4.57), and then solve for the flow throughputs in each network and sort them in one list in a decreasing order.

The algorithmic steps are listed in Algorithm 4.4, where $M_{sat} = 2M_{appr}$ from (4.60). By solving the flow capacity for different n, we can also find the edge band size that maximises the capacity.

Another problem in finding the flow capacity for FFR/SFR schemes is how to choose an appropriate value for p_n. We have seen that p_n plays an important role

Algorithm 4.4

(i) *Feasibility check.* Let $M_l = 1$, and define M_c, M_e using (4.57). Then solve for the flow throughputs for the RR or PF case and find γ_L. If the constraint of the problem (4.4) is satisfied, let $M_u = M_{sat}$ and go to step (ii); otherwise the problem is infeasible (return $M^* = 0$) and stop.

(ii) *Bisection search.* If $M_u - M_l = 1$, $M^* = M_l$ and stop; otherwise let $M = \lfloor (M_l + M_u)/2 \rfloor$, use (4.57) to decide M_c, M_e, solve for the flow throughputs, sort them in one list with a decreasing order, and find γ_L. If the constraint of (4.4) is satisfied, $M_l = M$; otherwise $M_u = M$. Go to step (ii).

in determining the flow capacity in our infinite user population model (see Section 3.4.1). Similarly, in this finite user population model, a different value of p_n gives a different SINR threshold s_n, which results in different inputs (time-average rates) for the centre/edge network and thus different values of the flow capacity. Unfortunately, we have no definition of "cell capacity", because for a zero flow throughput requirement, the total number of users in the system grows to infinity. To avoid a great number of simulations for rate distributions and a complicated calculation for the capacity, a heuristic method is invoked where we choose the same value of p_n as that in the infinite user population model, which is (3.39) for the RR case or (3.42) for the PF case, and obtained through Algorithm 3.3. The reason is that in Section 4.3.3, we have shown that the infinite user population model is a good approximation of the finite model under certain conditions, thus we speculate that this choice of p_n will yield a high flow capacity.

4.4.2 Finding the flow capacity for $U_r = U_I$

The next step is to find the flow capacity in a homogeneous load network. We let U_r denote the overall utilization in the reference sector, and let U_{rc} and U_{re} denote the utilizations in the centre and edge networks, respectively. U_{rc} (or U_{re}) can be obtained using (4.15) in the centre (or edge) network. Due to the quantization effects (see (4.57) and the rules for allocating users to rate bins in Section 4.3.1) as well as the choice of p_n (in addition, there are different scheduling gains in the centre/edge system for the PF case), we cannot ensure $U_{rc} = U_{re}$ generally. To err on the side

of conservatism, we approximate the overall utilization of the reference sector by

$$U_r = \max\{U_{rc}, U_{re}\}. \tag{4.61}$$

To find a homogeneous load curve for FFR/SFR schemes, we perform a series of simulations beforehand for a set of loads $\{U_{I1}, \ldots, U_{IJ}\}, U_{Ij} \in [0,1]$ for $1 \leq j \leq J$, in the interfering sectors, and apply Algorithm 4.5 for each U_{Ij} to find a (δ, M^*) pair. For each U_{Ij}, we obtain the centre and edge rate distributions and discretize them to $\{R_{c1}(U_{Ij}), \ldots, R_{cK_c}(U_{Ij})\}$ with probabilities $\{P_{c1}, \ldots, P_{cK_c}\}$ for the centre network, and $\{R_{e1}(U_{Ij}), \ldots, R_{eK_e}(U_{Ij})\}$ with probabilities $\{P_{e1}, \ldots, P_{eK_e}\}$ for the edge network.

The algorithmic steps in Algorithm 4.5 are similar to Algorithm 4.3, but have the following changes. Firstly, in each step we need to solve two networks to obtain U_{rc} and U_{re}, and decide the overall utilization U_r by (4.61). Secondly, we apply (4.60) for the approximate saturation point. Thirdly, for each given M, we need to divide it using (4.57) into two networks, solve for the flow throughputs in each network, sort them in one list in decreasing order, and find the associated γ_L for δ. Finally, for the special case of $U_{Ij} = 0$, we let $M^* = 0$ and define δ as

$$\delta = \begin{cases} \max\{R_{e1}(U_{Ij}), \ldots, R_{eK_e}(U_{Ij})\} & \text{if } p_n > 0.5 \\ \max\{R_{c1}(U_{Ij}), \ldots, R_{cK_c}(U_{Ij})\} & \text{otherwise} \end{cases}. \tag{4.62}$$

Then, we can obtain a set of (δ, M^*) pairs which gives an approximate homogeneous load curve. The procedure is summarized in Algorithm 4.5 with the error tolerance ε_3.

4.5 Numerical experiments and discussion

In this section, we present numerical examples using the finite user population model to illustrate the capacities of reuse-1, reuse-3, FFR-n and SFR-n schemes (where $n \in \{1, \ldots, 16\}$ for $N = 48$ resource units, and FFR-16 is equivalent to reuse-3).

We employ a hybrid simulation/analysis approach to obtain the flow capacity

Algorithm 4.5

(i) Let $M_l = 1$, and define M_c, M_e using (4.57), solve for U_{rc}, U_{re}, and apply (4.61) to decide the approximate overall utilization U_r. If $U_r - U_{Ij} < 0$, go to step (ii); otherwise let $M^* = 0$, choose δ using (4.62), and stop.

(ii) Let $M_u = 2M_{appr}$, define M_c, M_e using (4.57), solve for the associated U_{rc}, U_{re}, and then define U_r by (4.61). If $U_r - U_{Ij} > 0$, go to step (iii). Otherwise, if $|U_r - U_{Ij}| > \varepsilon_3$, $M_u = 2M_u$ and go to step (ii); otherwise let $M^* = M_u$, solve for the associated flow throughputs in the two networks, sort them in one list to find γ_L, define $\delta = \gamma_L$ and stop.

(iii) Let $M = \lfloor (M_l + M_u)/2 \rfloor$, define M_c, M_e using (4.57), solve for U_{rc}, U_{re}, and then define U_r by (4.61). If $U_r - U_{Ij} > 0$, $M_u = M$, otherwise $M_l = M$. If $M_u - M_l > 1$, go to step (iii). Otherwise, define M^* by (4.40) solve for the associated flow throughputs in the two networks, sort them in one list to find γ_L, and define $\delta = \gamma_L$ and stop.

for each scheme. The system-level simulation is used to obtain a uncontended time-average rate distribution, where the parameters and assumptions are consistent with the LTE downlink (see Table 3.2 in Section 3.5). In the simulation, we apply the most realistic SINR-to-Rate mapping, namely modified Shannon mapping (3.47). For each distribution, we divide it into equi-probability "bins", take R_j to be the harmonic mean of the simulation samples in the jth bin, and use this discrete set of rates as the input to the analytical model. In the analysis model, we set the mean flow size σ to be 5 Mbits, and the mean thinking time $1/\nu$ to be 180 seconds.

4.5.1 Definition of key aspects of the methodology

In this section, we explore several key aspects of our methodology for computing the flow capacity in the finite user population model. We discuss (i) the sensitivity to the number of classes K, and (ii) an approximation for the flow capacity using the FPM method.

Choice of the number of classes K

Like the infinite user population model, in the finite user population model we discretize a continuous rate distribution for the input into the analysis. However, in

this case the computational costs are higher, particularly for PF scheduling, so we cannot use as many classes as we apply in the infinite user population model.

For a two-node closed queueing network with M users and K classes, if RR scheduling is applied, the computational complexity of a direct calculation is of the order of $K^{\lfloor M/K \rfloor+1}$. We apply the MVAC algorithm (see Conway *et al.* [38] or Section 4.3.1) for the calculation of flow throughputs. The algorithm has a computational cost [38]

$$4\left[\begin{pmatrix} M+2 \\ 3 \end{pmatrix} - 1\right] + 4\begin{pmatrix} K+1 \\ 3 \end{pmatrix}, \tag{4.63}$$

which reduces the computational cost dramatically comparing to a direct calculation for large values of M and K. In our numerical examples for the RR case, we can find the solution within several minutes for K up to 50.

For the PF case, the computational cost is even higher than for the RR case. We need to apply the MVAC algorithm for the calculation of the first parts of (4.31), (4.34) and (4.35), and have to perform direct summations for the second parts of these equations. Figure 4.5 shows the number of states in the direct summation of the second part of (4.31) for different class number K when $x_0 = 7$, which is a typical value in our examples. We see that even for the case of $K = 20$, we need to perform more than 200,000 additions for one performance metric.

Next, we explore the sensitivity of the number of classes K in the flow capacity analysis. Figure 4.6 illustrates the flow capacities of reuse-1 with RR scheduling in a fully loaded environment ($U_I = 1$) and $\beta = 0.9, 0.7$, using different numbers of classes ($K = 10, 20, 50$) in the analysis. We see that 10 classes can achieve almost the same accuracy as 50 classes. Furthermore, in terms of complexity we can comfortably manage the calculations for 10 classes using the MVAC algorithm. Therefore, we apply $K = K_c = K_e = 10$ for RR scheduling in the following computations. Similarly for the PF case, we apply $K = K_c = K_e = 10$ in the following analysis, which corresponds to less than 10,000 states in the direct summation of the second parts of (4.31), (4.34) and (4.35).

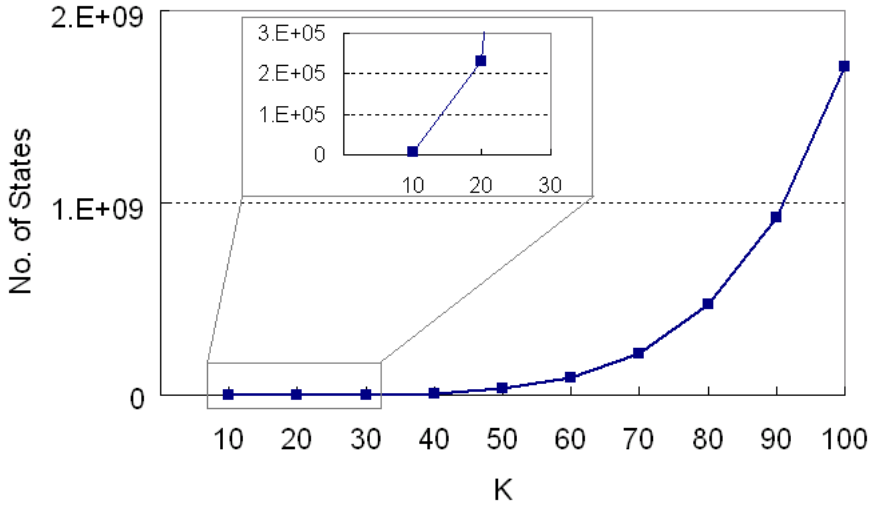

Figure 4.5: The number of states in the second part of (4.31). The inset in the top is a zoom-in for small K.

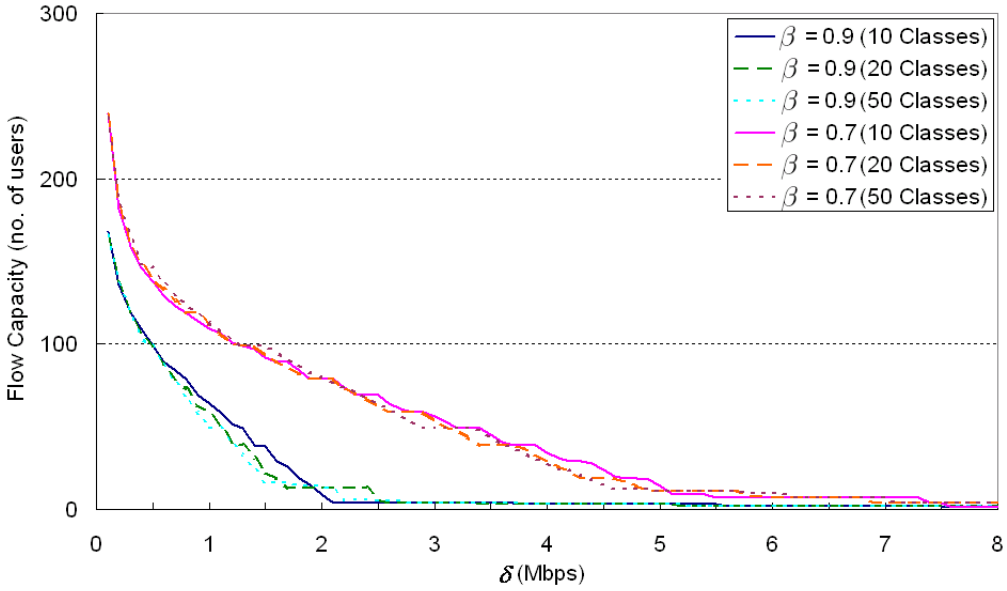

Figure 4.6: Sensitivity to the number of classes (reuse-1 with RR scheduling when $\beta = 0.9, 0.7$ and $U_I = 1$).

An approximation of the flow capacity using the FPM method

In Section 4.3.3, we showed using the FPM approximation [138] that the stochastic processes in the GPS node (node 1) of the closed queueing network can be approximated well by an open queueing network. In this section, we present an approach to find the flow capacity based on the FPM method, within which the most important step is how to find the flow throughputs for a given number of users M using the approximation.

We let ε_4 be the error tolerance, and let $\overline{\mathbf{m}} = (\overline{m_1}, \ldots, \overline{m_K})$ be the average population vector, where $\overline{m_k} = MP_k$.

For the RR case, we initialize the process by using (4.46), apply (4.45) in (4.47) and (4.48), and calculate the approximate arrival rates $(\lambda_1^o(\overline{\mathbf{m}}), \ldots, \lambda_K^o(\overline{\mathbf{m}}))$ by using (4.47) and (4.48) iteratively, where we stop the iterations when $|\psi_k^{(i)} - \phi_k^{(i)}| < \varepsilon_4$. Then we can use (3.19) from the infinite user population model (see Section 3.3.1) to solve for the flow throughputs corresponding to $(\lambda_1^o(\overline{\mathbf{m}}), \ldots, \lambda_K^o(\overline{\mathbf{m}}))$.

For the PF case, we need to modify the equations of the FPM method by letting

$$E\left[x_k \middle| \left(\lambda_1^o(\overline{\mathbf{m}}), \ldots, \lambda_K^o(\overline{\mathbf{m}})\right)\right] = \rho_k \frac{\frac{1}{Cg(1-\frac{\rho}{g})^2} - \Sigma_{x=1}^{x_0} D(x) x \rho^{x-1}}{\frac{1}{C\left(1-\frac{\rho}{g}\right)} - \Sigma_{x=0}^{x_0} D(x) \rho^x}, \tag{4.64}$$

where $\rho_k = \lambda_k^o(\overline{\mathbf{m}})/\mu_k$ and $\rho = \Sigma_{k=1}^K \rho_k$. This is simply a restatement of (3.31). Then we put (4.64) into (4.47) and (4.48), and run them iteratively with the initial arrival rate (4.46) until $|\psi_k^{(i)} - \phi_k^{(i)}| < \varepsilon_4$. Finally, we obtain the flow throughputs by using (3.32) for PF scheduling.

Thus, by substituting the above approach for the flow throughputs in Algorithm 4.1 for standard reuse schemes or Algorithm 4.4 for reuse partitioning schemes, we can find the flow capacity based on the FPM approximation method.

The next problem that we address is how to determine the regime in which we can operate the FPM approximation. We can only use the FPM approximation under some conditions because there is a stability condition in the infinite user population model, which is $\rho < 1$ for the RR case and $\rho < g$ for the PF case. We can use this condition to find the approximation operating regime. From (4.46)-(4.48), it is easy

to see that for $i \geq 1$,

$$\phi_k^{(i)} \leq \phi_k^{(i+1)} \leq \lambda_k^{(o)}(\overline{\mathbf{m}}) \leq \psi_k^{(i+1)} \leq \psi_k^{(i)}, k = 1, \ldots, K. \tag{4.65}$$

Therefore, it is sufficient that the initial arrival rates obtained from (4.46) are feasible such that the stability condition is satisfied, which is

$$\Sigma_{k=1}^{K} \frac{\psi_k^{(1)}}{\mu_k} < 1 \text{ (RR case)} \quad \text{or} \quad \Sigma_{k=1}^{K} \frac{\psi_k^{(1)}}{\mu_k} < g \text{ (PF case)}. \tag{4.66}$$

Let M_{sat}^{FPM} be an upper bound for the number of users for which (4.66) holds. From (4.46), (4.66) and $\mu_k = R_k/\sigma$, we derive

$$M_{sat}^{\text{FPM}} = \lfloor \frac{\overline{R}}{\nu\sigma} \rfloor \text{ (RR case)} \quad \text{or} \quad M_{sat}^{\text{FPM}} = \lfloor \frac{g\overline{R}}{\nu\sigma} \rfloor \text{ (PF case)}. \tag{4.67}$$

Figure 4.7 displays an example of using the FPM approximation to find the flow capacities for reuse-1 with RR scheduling in a fully loaded environment ($U_I = 1$). We observe the results for different definitions of user satisfaction β. If the capacity is greater than M_{sat}^{FPM} (which occurs when δ is small), we represent it by M_{sat}^{FPM} since we cannot apply the approximation beyond that value. We can see that this FPM approximation approach works well within the operating regime where $M^* \leq M_{sat}^{\text{FPM}}$ for different β.

We show another example for the PF case in Figure 4.8 for reuse-1 when $U_I = 1$. Again, we see that for the range $M^* \leq M_{sat}^{\text{FPM}}$, the approximations achieve very close agreement to those from the detailed algorithm.

Furthermore, we note that the results always give a lower bound for the flow capacity. The FPM approximation provides us with a method to calculate an estimate of the flow capacity with reduced computational complexity, and is particularly useful for the PF case. To estimate the performance over a wider operating regime, we could apply asymptotic methods like [63, 87], which we leave for future work.

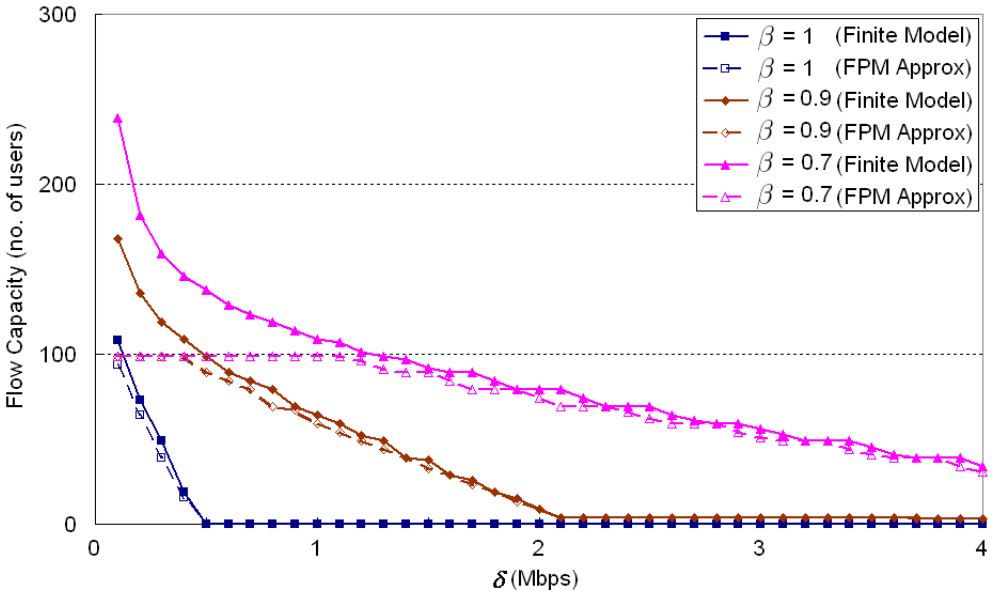

Figure 4.7: The approximations of the capacity based on FPM method (reuse-1 with RR when $U_I = 1$).

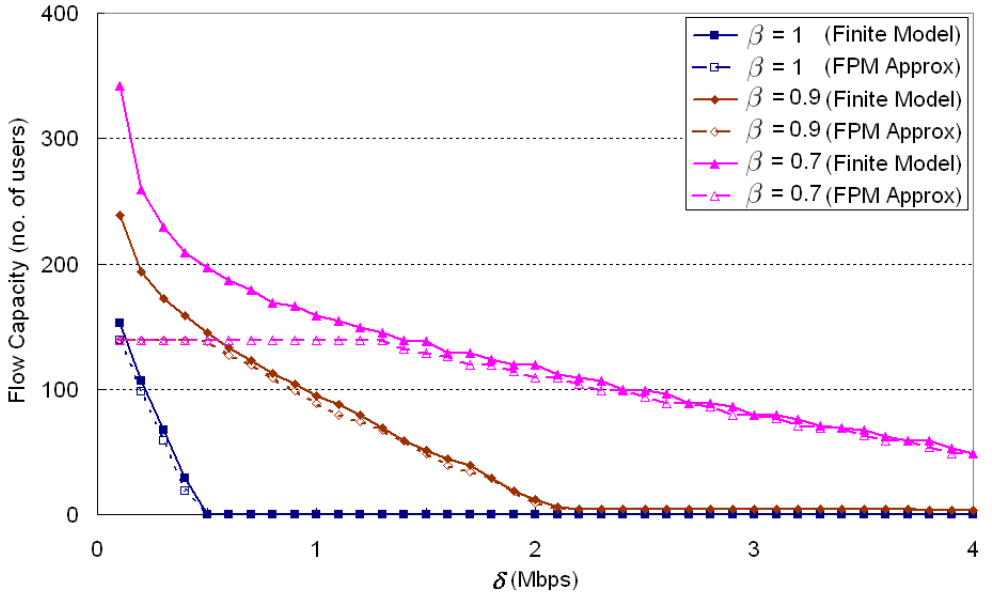

Figure 4.8: The approximations of the capacity based on FPM method (reuse-1 with PF when $U_I = 1$).

4.5.2 Comparison of flow capacities for different schemes

In this section, we compare the flow capacities of different reuse and reuse partitioning schemes on the basis of their homogeneous load curves. First, we give an example showing how the approximate homogeneous load curve for a finite user population can be generated using Algorithm 4.3. The example is for reuse-1 with PF scheduling when $\beta = 0.9$. As depicted in Figure 4.9, we obtain the capacity curves for a set of loads; for a specific load U_I, we apply Algorithm 4.3 to find a flow capacity whose associated utilization U_r is closest to U_I (shown as the dot on the line); from these dots, we estimate the locus of homogeneous load. To achieve a more accurate estimation, more loads need to be analysed. In the following results, for each scheme we perform the analysis for 21 loads ($U_I \in \{0, 0.05, \ldots, 1\}$, which can provide acceptable accuracy with manageable computational cost.

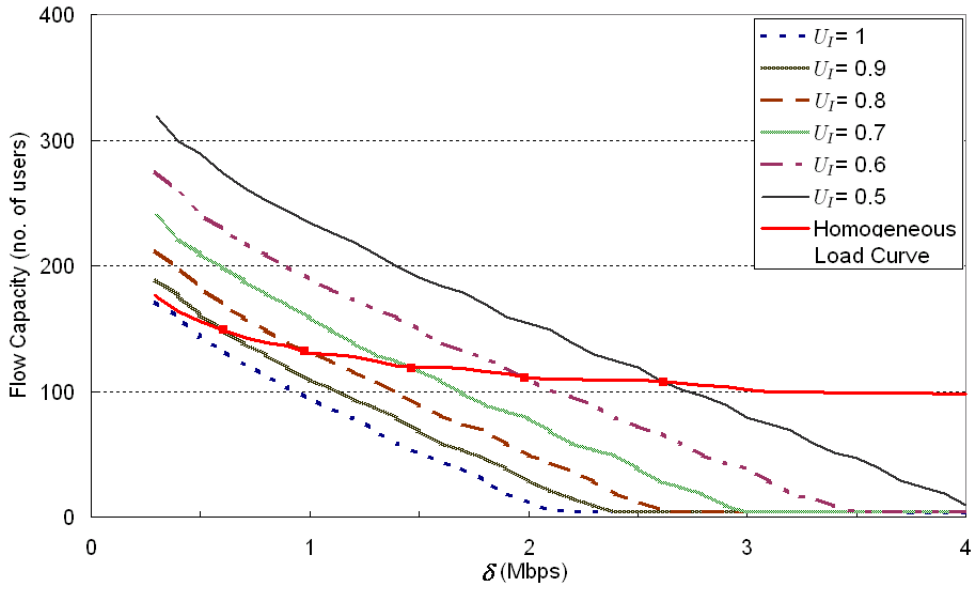

Figure 4.9: An example of plotting the approximate homogeneous load curve (reuse-1 with PF when $\beta = 0.9$).

By solving the flow capacities for different FFR-n/SFR-n schemes, we can find a judicious choice of edge band size that maximises the capacity for a given throughput requirement δ and user satisfaction β. Figure 4.10 and Figure 4.11 show the homogeneous load curves of FFR schemes with different edge band sizes when $\beta = 0.9$ for

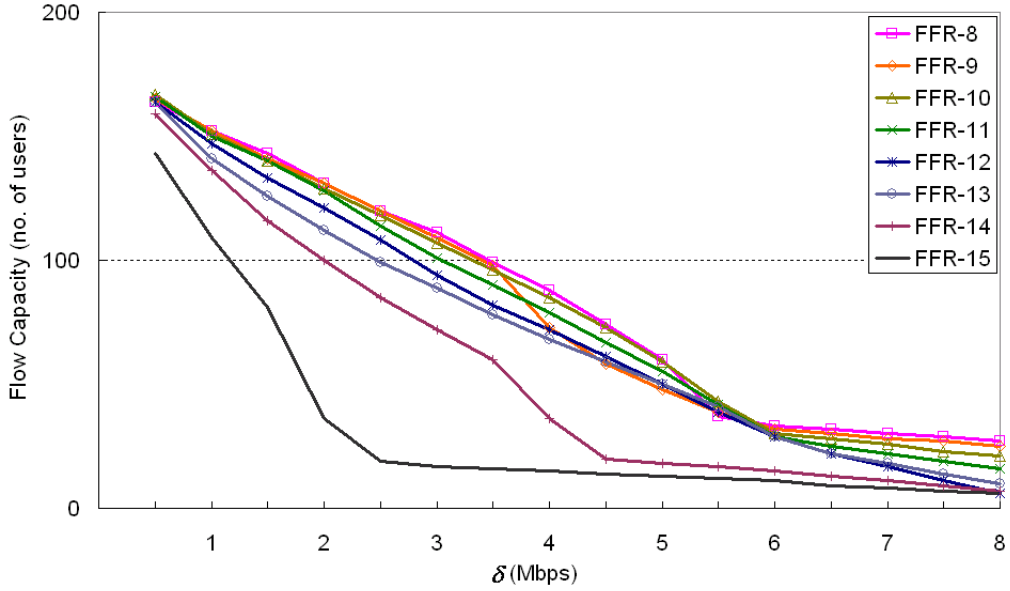

Figure 4.10: Comparison of flow capacities of different FFR schemes with RR, homogeneous load, and $\beta = 0.9$.

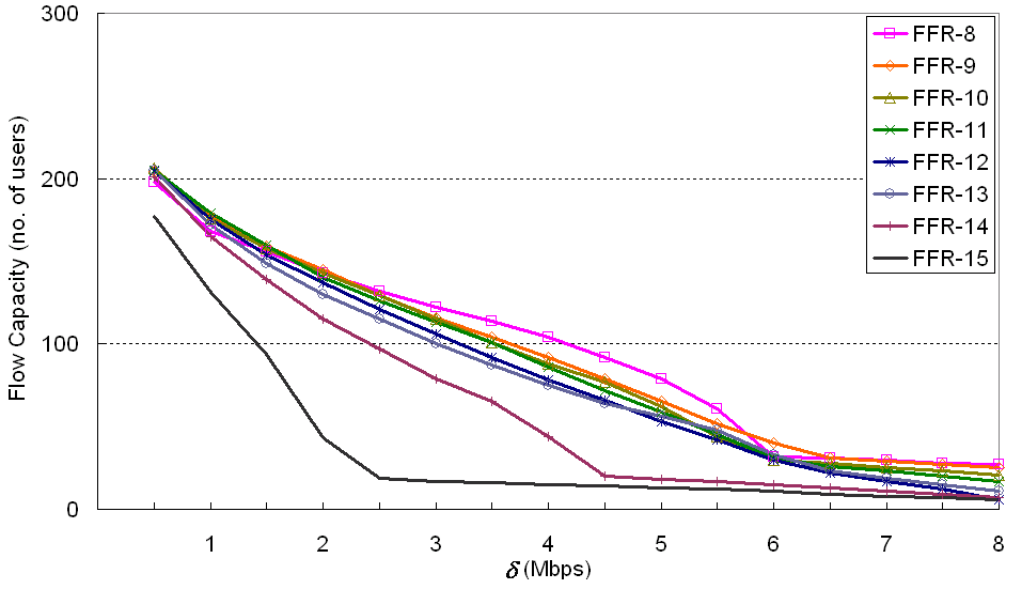

Figure 4.11: Comparison of flow capacities of different FFR schemes with PF, homogeneous load, and $\beta = 0.9$.

RR and PF scheduling, respectively. We see that the best choice of the edge band size for FFR schemes depends on the value of δ. For very low δ such as $\delta < 0.5$ Mbps, FFR-11 has the best performance and other FFR schemes have similar capacities, except for schemes with large n such as FFR-14 and FFR-15 which have low capacities due to their severely limited centre band resources. For larger δ, schemes with relatively small n such as FFR-8 and FFR-9 have the highest capacities. However, there is the possibility that the results here are influenced by quantization effects, since as discussed in Section 4.3.1, the user populations in each class must be integral. Because the edge band resources in these schemes are limited, there are few users allocated to the edge band and it is highly possible that there is no user allocated to the class with the worst rate, which results in better performance. This holds particularly for scenarios where the total number of users is small, namely cases with high δ.

Results for several SFR schemes with RR and PF scheduling when $\beta = 0.9$ are shown in Figure 4.12 and Figure 4.13. Unlike FFR, schemes with large n such as SFR-15 and SFR-16 are not severely limited on the centre band resources, so they achieve better flow capacities for very low $\delta < 0.5$ Mbps. SFR-9 and SFR-12 perform better for a moderate δ, but the limited edge band resources constrain the capacities for a high δ.

Finally, we present a capacity comparison between different reuse and reuse partitioning schemes. Figure 4.14 depicts the flow capacities for $\beta = 0.9$ under homogeneous loading. Note that for the FFR/SFR scheme, we always choose an optimal edge band size for a given δ. We see that we can improve the capacities by applying PF scheduling comparing to RR scheduling. In the regime of low δ, such as $\delta < 1.5$ Mbps, FFR with an optimal edge band size is the best scheme for both scheduling methods. Note that reuse-3 with PF scheduling achieves competitive capacities in this regime. For moderate δ (such as $\delta \in [2, 6]$ Mbps), SFR with an optimal edge band size provides the highest capacity. For high δ, reuse-1 is the best scheme since it yields the highest peak rate; reuse-1 outperforms all other schemes when $\delta > 6.5$ Mbps for both RR and PF cases.

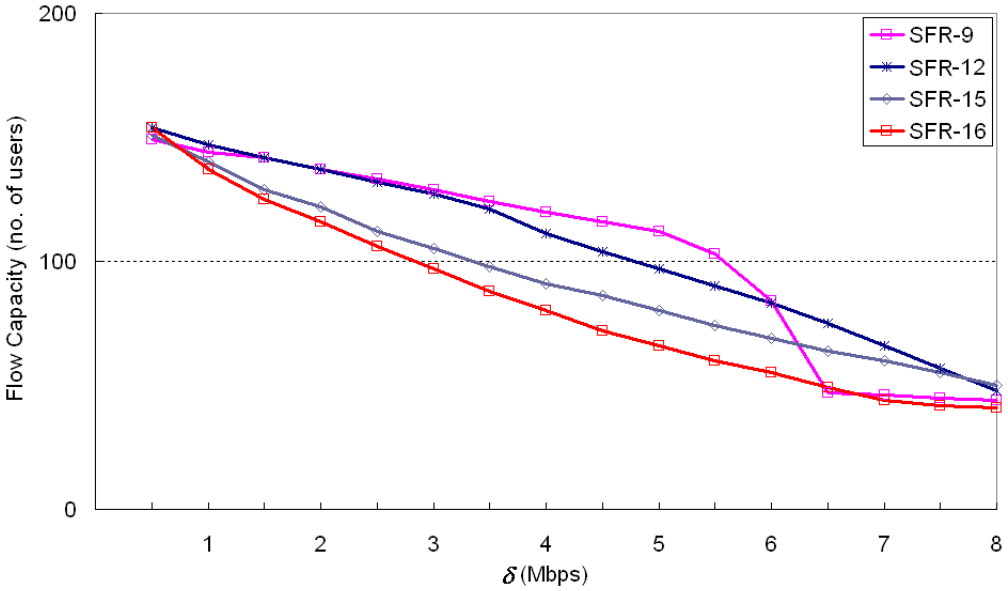

Figure 4.12: Comparison of flow capacities of different SFR schemes with RR, homogeneous load, and $\beta = 0.9$.

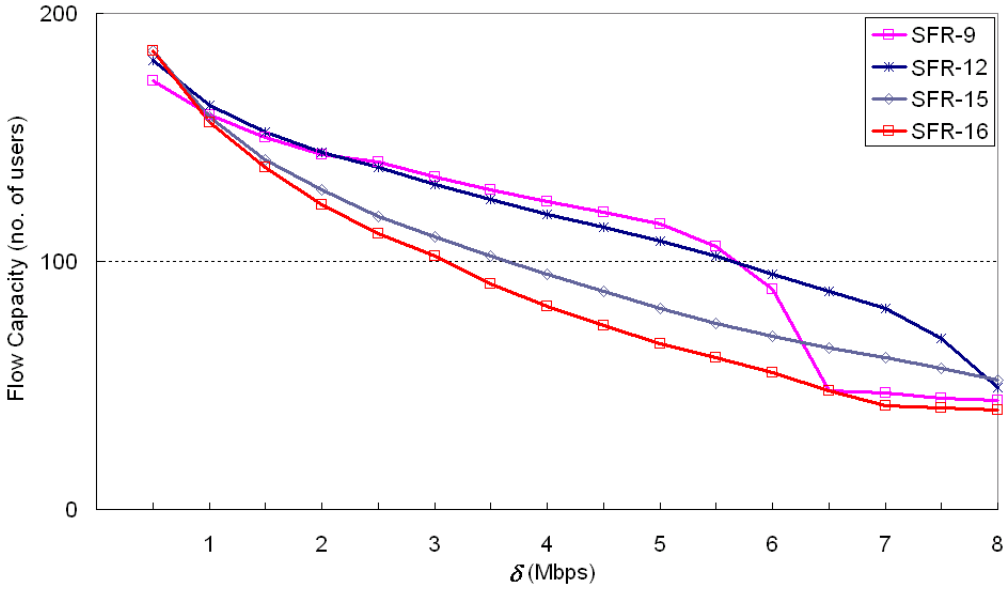

Figure 4.13: Comparison of flow capacities of different SFR schemes with PF, homogeneous load, and $\beta = 0.9$.

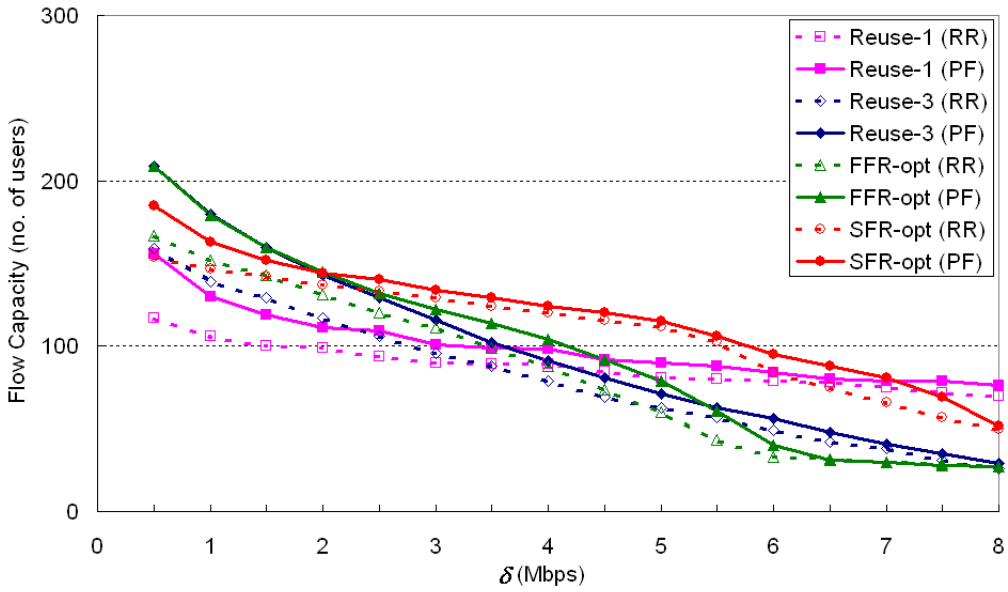

Figure 4.14: Comparison of flow capacities for reuse schemes and reuse partitioning schemes with RR and PF, homogeneous load, and $\beta = 0.9$.

4.6 Conclusion

In this chapter, we assumed a finite user population and developed a methodology to find the flow capacity for standard reuse and reuse partitioning schemes by using a closed queueing network model. We performed a capacity analysis using a hybrid simulation/analysis approach for both RR and PF scheduling, where the GPS queue was used to model the base station to take the PF scheduling gains into account. We performed the analysis in the multi-cell scenario with fractional loading, showed how fractional loading in the interfering sectors can approximately be taken into account, and found the flow capacity for a homogeneous load network.

To deal with the greater computational cost of this model, we applied an efficient algorithm to find the flow throughputs for the RR case. For the PF case, we exploited a fast convergence property of the multi-user diversity gain to convert the problem into a form where the same computationally efficient algorithm as for the RR case could be applied. The numerical results indicated that the choice of the best scheme depends on the throughput requirement δ. FFR with a judicious choice of edge band size yields the highest capacity for a low δ (reuse-3 achieves a competitive capacity

for a low δ), SFR gives the highest capacity for a moderate δ, and reuse-1 is the best choice for a high δ. Then we presented an open queueing network approximation based on the FPM method, and found that it works well in the regime of moderate to large δ.

Like the infinite user population model in the previous chapter, this framework for evaluating capacity also has applicability to general performance enhancement techniques, and can be used to dimension networks.

Chapter 5

Interference coordination based on allocation priority in the frequency domain

5.1 Introduction

In the previous two chapters, we applied a flow-level queueing theoretic approach to analyse the performance under a fractional loading scenario with elastic broadband traffic, where all the resources of a base station sector are allocated whenever there are any active flows. Since there are idle periods when no flows are active, the loading is fractional in the time domain. In this chapter, we consider a different fractional loading scenario which arises when the traffic consists of narrow-band services. When the number of concurrent narrow-band connections is such that the base station only allocates a portion of the frequency resources, it results in a fractional load in the frequency domain. Fodor, Telek, and Koutsimanis [44] and Pokhariyal *et al.* [102] also consider this type of traffic. We propose several inter-cell interference (ICI) coordination schemes for the LTE downlink under this fractional loading scenario, which involve prioritization of resource allocations between neighbouring base stations (BS). We use system-level simulations to evaluate the performance. The primary performance metrics sought are oriented towards measuring the base station performance, including average sector throughput and sector edge throughput.

In an attempt to maximise the sector capacity, LTE is designed to operate with reuse-1, thus it is important to improve the user performance in the sector edge area. A number of studies have applied interference coordination to achieve this goal. Gesbert *et al.* [48], Li and Liu [75, 77], Rahman and Yanikomeroglu [118], and Zhang and Letaief [149] solve the problem using various optimization techniques in a multi-cell environment. Necker in [92, 93] applies other advanced techniques,

such as graph theory, for interference coordination during scheduling. However, the aforementioned works assume a central controller for the coordination among a group of neighbouring sectors and the knowledge of full system state information (of both base stations and users), which is not feasible in real systems. Furthermore, if the number of users in the system is large, the amount of computation required for the solutions may be prohibitive.

In this chapter, we propose possible distributed realizations of interference coordination schemes for future cellular networks to enable efficient utilization of the entire available bandwidth. The schemes are based on setting resource allocation priority in the frequency domain; variants are possible, depending on whether priority setting is through off-line network planning (static) or adaptive to traffic load variations in neighbouring sectors. Note that when all sectors are fractionally loaded, it may be possible to avoid collisions with resource allocations in neighbouring sectors (if the same resource is allocated in more than one sector at the same time, we call it a collision). Our schemes assign higher allocation priority to the resources with less potential to cause allocation collisions. However, as the level of fractional loading increases, more resources need to be allocated such that more collisions occur. At full loading, the prioritization does not help since the collisions are unavoidable. By setting the allocation priority, the scheduling algorithms in our schemes are simple, so that fast resource allocation decisions can be made at the base stations. Furthermore, the resource allocation function is distributed across the base stations, thus a central controller is not required.

In the adaptive schemes, to enable coordination between base stations, we require inter-base-station signalling of system state information at some cycle time, defined as the status report update interval. There is a tradeoff between system performance and signalling overhead, which we investigate. The static schemes do not need inter-base-station signalling but require clever network planning.

We perform system-level simulation to investigate the base station performance, namely average sector throughput and sector edge throughput. When the sector load is not too high, our schemes yield gains in average sector throughput and sector edge throughput compared to an uncoordinated reuse-1 system. While our schemes are

applicable to any OFDMA or TDMA wireless access technology, we focus on their application to LTE in this chapter.

The rest of this chapter is organized as follows. In Section 5.2, we first describe the basic assumptions in our model. Then we propose interference coordination schemes based on adaptive and static assignment for the resource allocation priority in the frequency domain in Section 5.3 and 5.4, respectively. In Section 5.5, we describe our system-level simulation, present numerical experiments to illustrate the performance of different schemes, and explore the tradeoff between performance and signalling overhead in the adaptive schemes.

5.2 Model assumptions

5.2.1 Basic assumptions

We restrict attention to a single-input-single-output (SISO) downlink channel in a frequency division duplex (FDD) LTE network, where there is one transmitting antenna at the base station and one receiving antenna at the user equipment (UE). A fixed tri-sector layout (see Figure 5.1) is assumed, with one base station at the centre of each site, controlling the three sectors numbered 0, 1, and 2 (0, 120, and 240 degree antenna boresight orientations). As explained in Section 2.2, the smallest unit of resource that can be allocated is a resource block pair (RBP), which consists of two consecutive resource blocks, occupies 1 ms in the time domain, and can only be allocated to one user at a time in each sector.

We assume that there are N RBPs in total in the frequency domain available in each sector. In accordance with the LTE standard, we assume that the transmission power applied to each RBP is equal and the same modulation and coding scheme is assumed for all sub-carriers and all RBPs allocated to the same user. While adaptive power allocation over the sub-carriers is the optimal approach, it has been shown that adaptive power allocation only offers a small gain over fixed power allocation with adaptive modulation and coding (AMC) (see Biglieri, Proakis and Shamai [20], and Li and Liu [77]).

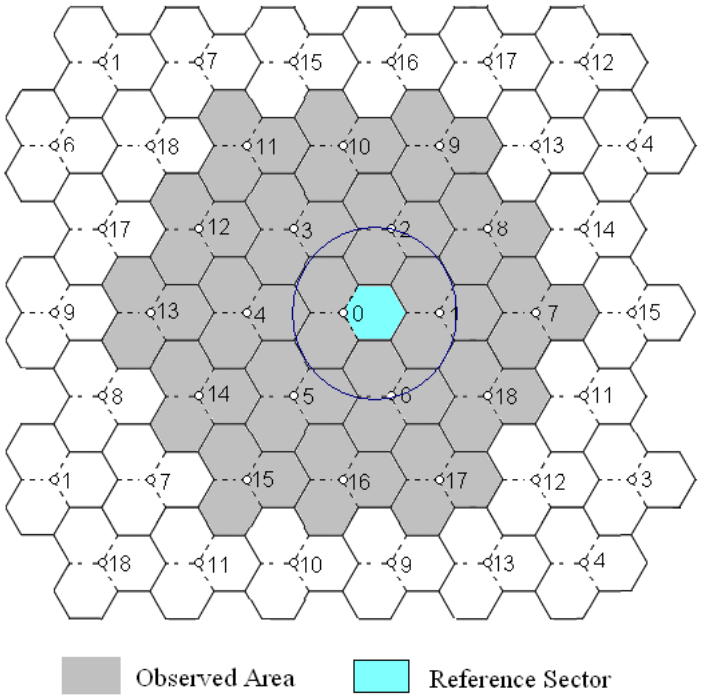

Figure 5.1: Wrap-around hexagonal cell layout.

For each RBP $n, n = 1, \ldots, N$, we set an allocation priority in the frequency domain, denoted by α^n. The larger the value of α^n, the higher the priority. The scheduling rule for resource allocation in each sector is highest priority first, that is, the scheduler always allocates available RBPs with highest priority to the users.

We consider only narrow-band services where each user only requires one RBP at each scheduling instant, so that the number of simultaneously active UEs is equal to the number of RBPs that are allocated. We also assume that all sectors have the same load.

5.2.2 Mathematical notation

The important mathematical notation that we use in this chapter is summarized in Table 5.1.

Table 5.1: Mathematical notation

Symbol	Description	Equation where symbol first appears
N	Total number of resource units in the system	–
N_0	Thermal noise	(5.1)
T	Status report update interval	–
$\delta_{i,j}^n$	State of the nth RBP in sector j of site i	(5.1)
ζ^n	Sector-average RB usage of the nth RBP	(5.3)
α^n	Sector-average allocation priority of the nth RBP	(5.4)
$w_{k,l}^{i,j}$	Sector-average interference weight from sector l of site k to sector j of site i	(5.3)
$\zeta^n(u)$	User-specific RB usage of the nth RBP	(5.6)
$\alpha^n(u)$	User-specific allocation priority of the nth RBP	(5.7)
$w_{k,l}^{i,j}(u)$	User-specific interference weight from sector l of site k to sector j of site i	(5.6)
$\phi_{i,j}$	Priority setting in sector j of site i	(5.8)
$W_{i,j}$	Interference impact suffered by sector j of site i	(5.11)
W	Overall interference impact in the system	(5.12)

5.3 Adaptive priority setting schemes

In this section, we propose interference coordination schemes based on resource allocation priority in the frequency domain, where the priority is adaptive to the resource utilizations in the surrounding sectors.

5.3.1 System model

Considering the central 19 sites of the network in Figure 5.1, we take the central site (site 0) as the reference site and sector 0 of the reference site as the reference sector (shaded in blue in Figure 5.1). In the downlink, the interference is predominantly generated by the other two sectors of the same site and the first two tiers of surrounding sites. If the nth RBP is allocated in the reference sector, the SINR at the UE receiver is

$$\mathrm{SINR}^n = \frac{\frac{P_x}{q_{0,0}}}{\sum_{j=1}^{2} \delta_{0,j}^n \frac{P_x}{q_{0,j}} + \sum_{i=1}^{18} \sum_{j=0}^{2} \delta_{i,j}^n \frac{P_x}{q_{i,j}} + N_0}, \quad n = 1, \ldots, N, \tag{5.1}$$

where P_x is the transmit power for each RBP, N_0 is the background thermal noise, and $q_{i,j}$ is the propagation loss between the sector j of site i and this UE, which includes the effects of antenna gains and antenna directivity, path loss, and shadowing fading. We also introduce an indicator variable $\delta^n_{i,j}(i = 0, \ldots, 18; j = 0, 1, 2; n = 1, \ldots, N)$ to denote the state of the nth RBP in sector j of site i. The value of $\delta^n_{i,j}$ is equal to 1 if the nth RBP is currently allocated, and 0 otherwise. The first item in the denominator of the right hand side of (5.1) describes the interference due to the other two sectors of the reference site, while the second item represents the interference caused by the first two tiers of surrounding sites.

Different neighbouring sectors contribute dissimilarly to the total interference in (5.1) because of the different effect of $q_{i,j}$. For instance, the first layer sectors, which are enclosed by the circle in Figure 5.1, would be the most dominant interferers. Our adaptive schemes assign the resource allocation priorities based on the system state in the neighbouring sectors, in order to enable resource allocation decisions to be made that minimise interference. To achieve this goal, our schemes require inter-base-station signalling between base stations. We assume that this signalling is sent periodically with a fixed update interval T.

A consequent problem that we address is the volume and frequency of the inter-base-station signalling. To limit the volume of signalling information that needs to be transported and processed, we require that only per RBP binary state information $\delta^n_{i,j}$ is sent to indicate whether the nth RBP is allocated or not. In our model, we assume that each base station sends state information of all three sectors to its first two tiers of neighbouring base stations; consequently, each base station receives state information of all three sectors from its first two tiers of neighbouring base stations. This is a reasonable simplification, since the first two tiers usually contain the dominant interferers. Note that to further reduce the volume of this inter-base-station signalling, each base station could send/receive the state information to/from only the strongest interfering sites, or we could accept less precise status reports based on groups of RBPs, rather than individual RBPs.

As for the frequency of this inter-base-station signalling, we first assume that it is performed at every scheduling interval such that $T = 1$ ms. However, it would

be possible to allow a longer status report update interval (such as hundreds of milliseconds or seconds), but at the expense of possibly degraded performance, which is discussed in Section 5.5.3.

To account for the fact that different neighbouring sectors contribute dissimilarly to the total interference, we introduce *interference weights* which reflect the interference impact of each sector of the two tiers of surrounding base stations. Based on the state information of neighbouring base stations and the interference weights, we set the priority higher for those resources with the least likelihood of suffering or causing interference.

Two schemes, called *Adaptive Scheme 1* and *Adaptive Scheme 2*, are defined, which differ in the way in which the interference weights are determined.

Adaptive Scheme 1

We denote the interference weight from sector l of site k to sector j of site i as $w_{k,l}^{i,j}(i,k=0,\ldots,18;j,l=0,1,2)$, with the convention that $w_{k,l}^{i,j}=0$ if $i=k$ and $j=l$. In this scheme, the interference weight $w_{k,l}^{i,j}$ represents the downlink interference power received from sector l of site k, averaged over all user locations in sector j of site i and described as

$$w_{k,l}^{i,j}=\begin{cases} E[P_{k,l}^{i,j}] & \text{if } i\neq k \text{ and } j\neq l \\ 0 & \text{otherwise} \end{cases}, \tag{5.2}$$

where $P_{k,l}^{i,j}$ is a random variable representing the received interference power from sector l of site k for user locations in sector j of site i.

From (5.2), it is clear that those sectors which tend to cause larger inter-cell interference, such as sector 2 of site 2 or sector 1 of site 6 with respect to the reference sector (see Figure 5.1), will have larger interference weights. The advantage of this definition of $w_{k,l}^{i,j}$ is that it does not depend on the specific user's channel state, so it can be pre-computed (for example, we determine it by a test program in our numerical experiments) and stored in a look-up table in each base station.

The priority setting procedure for this scheme is as follows (taking the reference

sector as an example). We define a metric, called the *RBP usage*, to indicate the interference impact and use it to assign the allocation priority. At any scheduling instant, we calculate the RBP usage ζ^n in the reference sector by

$$\zeta^n = \sum_{l=1}^{2} w_{0,l}^{0,0} \delta_{0,l}^n + \sum_{k=1}^{18} \sum_{l=0}^{2} w_{k,l}^{0,0} \delta_{k,l}^n, \quad n = 1, \ldots, N. \tag{5.3}$$

The first term of (5.3) describes the usage in the other two sectors of the same site, while the last term captures the usage in the first two tiers of surrounding sites. Note that the smaller the value of ζ^n, the less the usage of the nth RBP in the interfering sectors. Therefore, we can assign the priority of the nth RBP by

$$\alpha^n = -\zeta^n, \quad n = 1, \ldots, N. \tag{5.4}$$

The scheduler for the reference sector then selects an unassigned RBP with the largest value of α^n for the next user, hence avoiding (or at least minimising) the use of resources with strong interference.

The scheduling algorithm for this scheme is described by Algorithm 5.1, and needs to be performed at every scheduling instant. We first need to assign the priority to each RBP in the frequency domain using (5.4) based on the most recent $\delta_{i,j}^n$ values. Then we can allocate the resources to the active users according to the resource allocation priorities. During this procedure, we introduce one more ordering step for all active users, which determines the priority order in which the users should be allocated resources. This step allows for a scheduling discipline to be imposed, such as round-robin or a channel-dependent discipline such as time-based proportional fair (see Kela *et al.* [62], and Wengerter, Ohlhorst, and Elbwart [137]).

Adaptive Scheme 2

A disadvantage of Adaptive Scheme 1 is that the average interference weight $w_{k,l}^{i,j}$ is only an approximation to the interference that would be experienced by a given user, and may be a severe over-estimate or under-estimate. Adaptive Scheme 2

Algorithm 5.1

(i) Calculate the priority for each RBP using (5.4).

(ii) Assign RBPs to the users according to the following steps:

(a) Order the active users, according to the scheduling discipline.

(b) For the next user in the ordered list, find the RBP with the highest priority α^n. If this RBP is unallocated, assign it to this user. Otherwise, check the next highest priority α^n, and so on.

(c) Repeat step (b) for all users in the ordered list, or until all available RBPs have been allocated.

addresses this shortcoming by making the interference weights dependent on the user's channel state. Specifically, for any user u in sector j of site i, we define the user-specific interference weights by

$$w_{k,l}^{i,j}(u) = \begin{cases} \frac{P_{k,l}(u)}{P_{i,j}(u)} & \text{if } i \neq k \text{ and } j \neq l \\ 0 & \text{otherwise} \end{cases}, \tag{5.5}$$

where $P_{k,l}(u)$ is the received interference power from sector l of site k for this user. We assume that the UE takes measurements of the pilot channel powers of the serving and neighbouring sectors, and sends this information to the serving base station in CQI feedback messages (see Section 2.3.2), from which the base station can determine $P_{k,l}(u)$. Clearly, $w_{k,l}^{i,j}(u)$ in this scheme is user-specific and time-varying, and therefore, cannot be pre-computed.

For the user u in the reference sector, we can calculate the priorities using

$$\zeta^n(u) = \sum_{j=1}^{2} w_{0,j}^{0,0}(u)\delta_{0,j}^n + \sum_{i=1}^{18}\sum_{j=0}^{2} w_{i,j}^{0,0}(u)\delta_{i,j}^n, \quad n = 1, \ldots, N, \tag{5.6}$$

$$\alpha^n(u) = -\zeta^n(u), \quad n = 1, \ldots, N. \tag{5.7}$$

For each scheduling instant, we perform the algorithm defined in Algorithm 5.2 for this scheme. The resource allocation information in the surrounding sectors is still required by inter-base-station signalling for every update interval T. Now the priorities of the RBPs are user-dependent, so that they need to be calculated

separately for each user according to the channel conditions. Then the resources are allocated based on the user-specific priorities. Again, we introduce an ordering step to implement some scheduling discipline.

Algorithm 5.2

(i) Assign RBPs to the users according to the following steps:

(a) Order the active users, according to the scheduling discipline.

(b) For each user u in the ordered list, calculate the interference weights $w_{k,l}^{i,j}(u)$ using (5.5) and find the RBP with the highest priority $\alpha^n(u)$ by (5.7). If this RBP is unallocated, assign it to this user. Otherwise, check the next highest priority $\alpha^n(u)$, and so on.

(c) Repeat step (b) for all users in the ordered list, or until all available RBPs have been allocated.

5.4 Static priority setting schemes

In this section, we investigate interference coordination schemes based on static resource allocation priority in the frequency domain. Compared to the two adaptive schemes in the previous section, the priority setting is performed off-line via network configuration, thus the static schemes do not need inter-base-station signalling but still utilize the whole frequency spectrum in each sector.

5.4.1 System model

The reference sector is numbered 0, and sector $1, \ldots, M-1$ are the $M-1$ dominant interfering sectors. The N RBPs are divided in the frequency domain into $M (M \leq N)$ sub-bands, and each sub-band is assigned a unique priority. The idea is to assign different sub-band priority lists to the sectors to minimise the chance of the same sub-bands being used in sectors with strong interference coupling.

We index the sub-bands consecutively from lowest to highest frequency, form a circular list of indices, and assign different starting indices for different sectors. In each sector, the sub-bands are allocated commencing from the starting index and

moving in the direction of increasing index. Consequently, the problem in this model is to assign the starting indices to the sectors to achieve the least interference for different traffic loads.

Figure 5.2 illustrates a general example of sub-band priority lists for the M sectors, where each sector has a different starting sub-band index. This type of resource allocation scheme has the advantage of avoiding interference from the dominant interfering sectors when the load is low to moderate. For example if the sector load is less than $1/M$ in all sectors, only the resources in the highest priority sub-band will be allocated and there would be no interference between the sectors.

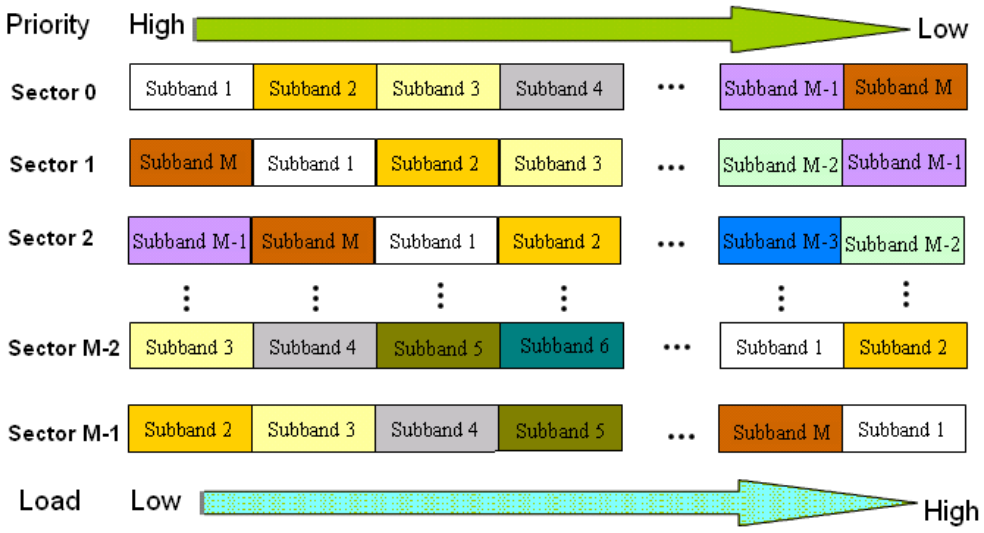

Figure 5.2: System model for static priority setting in the frequency domain.

When the load increases, say to $2/M$, only one sector (sector 1) causes interference to sub-band 1 of the reference sector and another sector (sector $M-1$) introduces interference to sub-band 2 of the reference sector (see Figure 5.2), while other sectors do not cause any interference. If the load increases further, sub-band collisions are inevitable. However, it is still possible to avoid using the same sub-band in sectors with the strongest interference if the allocation priority setting in the network is planned well.

The challenge in this approach is to decide how many sub-bands are needed and how to determine the starting sub-band index in each sector to incur the least

interference as possible for different loads. This is an optimal network configuration problem. Clever resource planning schemes are interesting as they offer additional flexibility in mitigating interference with very low complexity. In the following, we propose two schemes, called *Static Scheme 1* and *Static Scheme 2*, for the static priority setting.

Static Scheme 1

The strongest interference comes from the immediately adjacent sectors, which means that the three sectors of the same site should also avoid using the same sub-bands as far as possible. We adopt the idea of the conventional reuse-3 (see Figure 2.5) to set the priority for sub-bands, on the basis of which we define Static Scheme 1. We divide the resources into three sub-bands and define three categories of priority settings, denoted by "A", "B", and "C". The allocation priority in each sector is shown in Figure 5.3. We see that for each sector, its six immediately adjacent sectors, which are the strongest interfering sectors, always use different highest priority sub-bands. Furthermore, this scheme can avoid interference from the same site provided that the three sectors have a load of less than one-third.

Static Scheme 2

In this scheme, we consider that there is only one RBP in each sub-band. We let $\phi_{i,j}$ be the starting sub-band index in sector j of site i in the network, and $w_{k,l}^{i,j}$ be the average interference weight from sector l of site k to sector j of site i, which is obtained using (5.2). Thus, if $\phi_{i,j} = n_0, n_0 \in \{1, \ldots, N\}$ in one sector, we define the priorities α^n as follows

$$\phi_{i,j} = n_0 \quad \Rightarrow \quad \alpha^n = \begin{cases} N & \text{for } n = n_0 \\ N + n_0 - n & \text{for } n = n_0 + 1, \ldots, N \\ n_0 - n & \text{for } n = 1, \ldots, n_0 - 1 \end{cases} . \tag{5.8}$$

Given the starting indices $\phi_{i,j} = n_0$ in sector j of site i and $\phi_{k,l} = n_1$ in sector l of site k, we define a function $\Phi(\phi_{i,j}, \phi_{k,l}, c)$ which gives the number of allocation

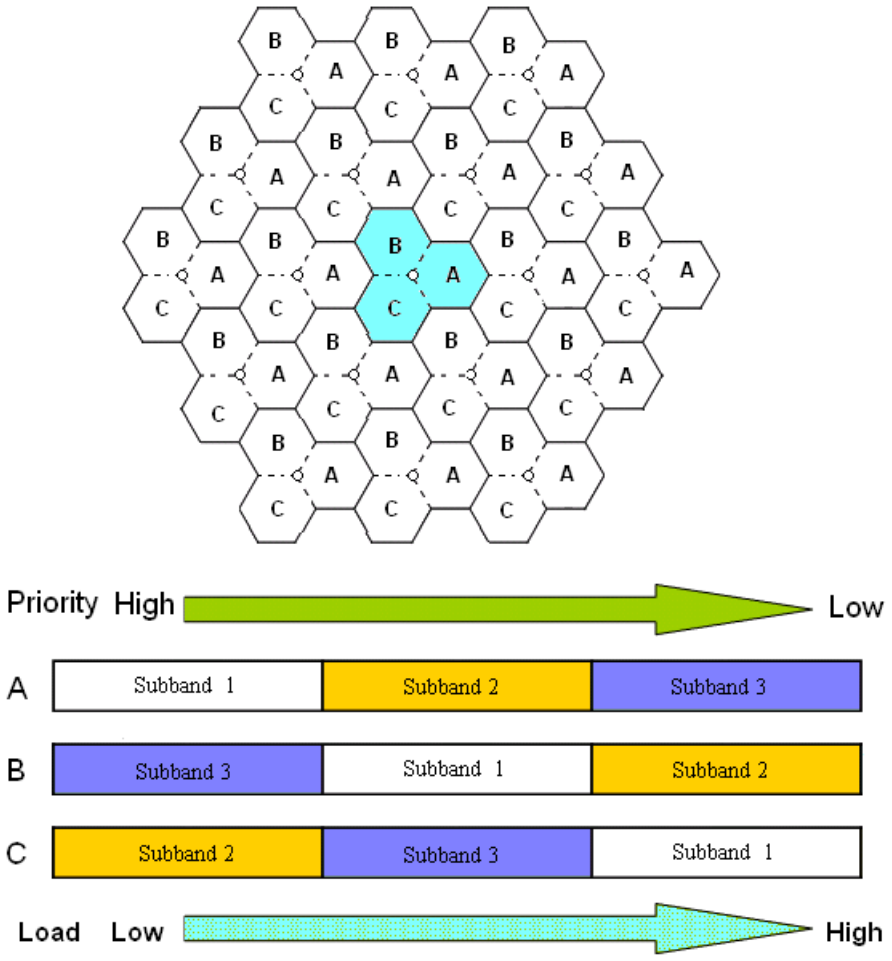

Figure 5.3: Network plan for Static Scheme 1.

collisions between the two sectors under the sector load c. The load $c, 0 \leq c \leq 1$, is defined as the fraction of RBPs that need to be allocated to meet the offered traffic. Since the scheduler always allocates the available sub-bands with highest priority to the users, at most $N_c = \lceil N \cdot c \rceil$ RBPs need to be allocated, where $\lceil x \rceil$ denotes rounding to the nearest integer greater than or equal to x. Without loss of

generality, we let $n_0 \leq n_1$. It can be shown (see Figure 5.4) that

$$\Phi(\phi_{i,j}, \phi_{k,l}, c) = \begin{cases} 0 & \text{if } n_0 + N_c - 1 < n_1, n_1 + N_c \leq N + n_0 \\ n_1 + N_c - N - n_0 & \text{if } n_0 + N_c - 1 < n_1, n_1 + N_c > N + n_0 \\ n_0 + N_c - n_1 & \text{if } n_0 + N_c - 1 \geq n_1, n_1 + N_c \leq N + n_0 \\ 2N_c - N & \text{if } n_0 + N_c - 1 \geq n_1, n_1 + N_c > N + n_0 \end{cases}. \tag{5.9}$$

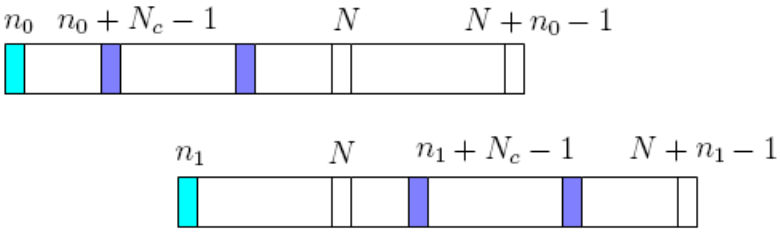

Figure 5.4: Frequency collisions calculation.

We define another function $\Psi(\phi_{i,j}, \phi_{k,l})$ to represent the total number of allocation collisions between the two sectors for all loads, which is

$$\Psi(\phi_{i,j}, \phi_{k,l}) = \int_0^1 \Phi(\phi_{i,j}, \phi_{k,l}, c) dc. \tag{5.10}$$

Then we define the overall interference impact $W_{i,j}$ suffered by sector j of site i:

$$W_{i,j} = \sum_{k=0}^{18} \sum_{l=0}^{2} w_{k,l}^{i,j} \Psi(\phi_{i,j}, \phi_{k,l}), \tag{5.11}$$

where $w_{k,l}^{i,j}$ is defined in (5.2). The total interference impact W in the observed area (see Figure 5.1) is defined as

$$W = \sum_{i=0}^{18} \sum_{j=0}^{2} W_{i,j} = \sum_{i=0}^{18} \sum_{j=0}^{2} \left(\sum_{k=0}^{18} \sum_{l=0}^{2} w_{k,l}^{i,j} \Psi(\phi_{i,j}, \phi_{k,l}) \right). \tag{5.12}$$

Our aim is to find a starting index assignment $\phi_{i,j} (i = 0, \ldots, 18; j = 0, 1, 2)$ such that the total interference impact W is minimised.

An exhaustive search for the optimal solution for such a problem is usually

computationally prohibitive. For example, if we assume a total of 10 MHz bandwidth in the system (50 sub-bands in the frequency domain), the computational cost of an exhaustive search is in the order of 50^{57}, which is impractical. Consequently, we apply a heuristic and greedy algorithm to define $\phi_{i,j}$ which is operated on a site-by-site basis, sequentially starting from the central site. The algorithm works as follows.

Let $N_{r3} = \lfloor \frac{N}{3} \rfloor$. We have seen that the three sectors of the same site should avoid using the same sub-bands whenever possible. Thus we define a heuristic rule for the priority setting in one site: if sub-band $n, n \in \{1, \ldots, N_{r3}\}$ is assigned as the starting index in one sector, the other two sectors will set sub-band $n + N_{r3}$ and sub-band $n + 2N_{r3}$ as their starting indices.

We start from the central site 0, and set $\phi_{0,0} = 1, \phi_{0,1} = 1 + N_{r3}, \phi_{0,2} = 1 + 2N_{r3}$ (it does not matter for the initial setting as long as it follows the priority setting rule for one site). Then in every step only one site is added into the network and we try to find a priority setting for this site such that the total interference impact in the current network is minimised.

Assuming the priorities of sub-bands in sites $0, \ldots, k-1$ have been assigned, for the next site k, there are 6 options to set the priority when we choose a sub-band index $n, n \in \{1, \ldots, N_{r3}\}$,

$$
\begin{array}{lllll}
\text{Option 1:} & \phi_{k,0} = n, & \phi_{k,1} = n + N_{r3}, & \phi_{k,2} = n + 2N_{r3}; \\
\text{Option 2:} & \phi_{k,0} = n, & \phi_{k,1} = n + 2N_{r3}, & \phi_{k,2} = n + N_{r3}; \\
\text{Option 3:} & \phi_{k,0} = n + N_{r3}, & \phi_{k,1} = n, & \phi_{k,2} = n + 2N_{r3}; \\
\text{Option 4:} & \phi_{k,0} = n + N_{r3}, & \phi_{k,1} = n + 2N_{r3}, & \phi_{k,2} = n; \\
\text{Option 5:} & \phi_{k,0} = n + 2N_{r3}, & \phi_{k,1} = n, & \phi_{k,2} = n + N_{r3}; \\
\text{Option 6:} & \phi_{k,0} = n + 2N_{r3}, & \phi_{k,1} = n + N_{r3}, & \phi_{k,2} = n.
\end{array}
$$

For these options and for each $n, n \in \{1, \ldots, N_{r3}\}$, we calculate the total interference impact W_k in the current network and choose an option and an n which minimises it. W_k is defined as

$$W_0 = 0; \tag{5.13}$$

$$W_k = W_{k-1} + \sum_{i=0}^{k-1}\sum_{j=0}^{2}\left(\sum_{l=0}^{2} w_{k,l}^{i,j}\Psi(\phi_{i,j},\phi_{k,l})\right) + \sum_{l=0}^{2}\left(\sum_{i=0}^{k-1}\sum_{j=0}^{2} w_{i,j}^{k,l}\Psi(\phi_{k,l},\phi_{i,j})\right). \quad (5.14)$$

The first part in (5.14) is the total interference in the existing sites (sites $0,\ldots,k-1$), and the second part is the additional interference introduced by the new site k to those existing sites, and the last part is the interference from the existing sites to the new site k. By using this approach, we reduce the computational cost for one site from N^3 to $6N_{r3}$.

In our numerical examples, we assume a 10 MHz bandwidth with 50 sub-bands in the frequency domain (see Section 5.5.1), and apply the above heuristic and greedy algorithm to assign the resource allocation priorities in each sector, which is shown in Figure 5.5. The number in each sector is the value of $\phi_{i,j}$, and the priority setting is given by (5.8).

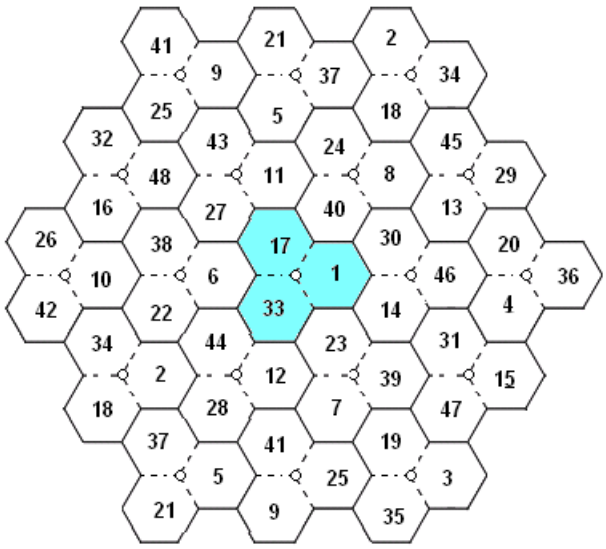

Figure 5.5: Network plan for Static Scheme 2.

5.5 Numerical results and discussion

In this section, we present numerical results that compare the performance of our schemes with that of an uncoordinated resource allocation, where RBP allocations in each sector are made randomly from the pool of available RBPs and independently to RBP allocations in other sectors. We make the comparisons on the basis of average sector throughput and average sector edge throughput.

5.5.1 Simulation description

We use a static system-level simulation, and simulate a regular hexagonal layout of 19 sites. A wrap-around model is used to avoid boundary effects, as shown in Figure 5.1. A small inter-site spacing of 1 km is used to create a scenario that is predominantly interference-limited.

A bandwidth of 10 MHz in the 2.5 GHz frequency band is available in each sector giving a total of 600 sub-carriers, each of which has a bandwidth 15 kHz [4]. A RBP consists of 12 sub-carriers (180 kHz) in the frequency domain and 1 ms in the time domain. Therefore, a total of 50 RBPs can be allocated within the 1 ms scheduling interval in each sector. With 50 RBPs, the three sub-bands in Static Scheme 1 cannot have the same number of RBPs; we choose to assign 16 RBPs to the two highest priority sub-bands in each sector, and 18 RBPs to the lowest priority sub-band.

We assume a reuse-1 scheme with fractional loading, where all RBPs are available everywhere in each sector and the number of resources allocated depends on the sector load. A SISO channel is assumed. When calculating the UE's instantaneous SINR using (5.1), we take into account the effects of antenna gain and directivity, path loss, shadow fading, inter-cell interference, and thermal noise. The achieved rate for each user is estimated from the SINR using an SINR-to-Rate mapping. For the mapping, we adopt an attenuated and truncated form of the Shannon bound (see 3GPP standard [8]), which is given by

$$S(\text{bits/s/Hz}) = \min\{\eta \log_2(1 + \text{SINR}), S_{\max}\}, \tag{5.15}$$

where we apply $\eta = 0.6$ and $S_{\max} = 4$ bits/s/Hz to the SISO case. The other simulation parameters and assumptions are the same as listed in Table 3.2.

For the two adaptive schemes, inter-base-station signalling is assumed such that each base station has perfect knowledge about current RBP allocation information in the two tiers of surrounding sites for each scheduling instant ($T = 1$ ms, and we will relax this assumption in Section 5.5.3). In all schemes, if there are several unassigned RBPs with the same priority, we break the tie randomly during the resource allocation.

We run the simulation for different pre-defined levels of sector load, where the sector load is defined as the fraction of RBPs in each sector that need to be allocated to satisfy the offered traffic. Since we assume a narrow-band service where each user only requires one RBP, the number of simultaneous UEs is equal to the number of RBPs that are allocated. We assume that the sector load and the number of users are the same for each sector. The users are randomly located in each sector, and the traffic model for each user is full buffer.

For each level of sector load, we run the simulation for tens of thousands of iterations. We define the user throughput as the number of successfully transmitted bits per user per unit time. For each iteration of the simulation, the aggregate throughput for all concurrent users in a sector yields a sample of the sector throughput. A user is deemed to be in the sector edge if its SINR is below the 10th percentile of the SINR distribution. The edge user throughput is the user throughput experienced by a sector edge user. We define the sector edge throughput in a different way, where the aggregate throughput for all concurrent sector edge users in a sector yields a sample of the sector edge throughput.

For the two adaptive schemes, we execute Algorithm 5.1 or 5.2 to allocate the resources in each sector of the 19 sites, and average over all samples of sector throughput and sector edge throughput over all sectors to obtain estimates of the average sector throughput and the average sector edge throughput. For the static schemes, we only run the resource allocations in the three sectors of the reference site according to their pre-defined priorities since we can explicitly calculate which resources are allocated in the neighbouring sectors based on the current sector load, which can

save a lot of simulation effort. Thus the average sector throughput and the average sector edge throughput are only averaged over the samples in the three sectors of the reference site.

5.5.2 Comparison of the performance of different schemes

In this section, we present numerical results based on the simulation described in the previous section. We compare the average sector throughput and average sector edge throughput obtained by applying our proposed schemes with those obtained by an uncoordinated resource allocation scheme.

Figure 5.6 presents the normalised average sector throughput as a function of sector load for Adaptive Schemes 1 and 2, Static Schemes 1 and 2, and the uncoordinated approach. (Note that Figures 5.6, 5.8 – 5.11 are plotted with confidence intervals, but they are sometimes difficult to distinguish because they are very tight.) For the results, we normalise the average sector throughput with the average sector throughput at full loading. As expected, we can avoid allocation collisions among the immediately adjacent sectors by prioritization of resource allocations when the level of sector load is not very high. A consequent result is that the users suffer less interference such that the user throughput per user (and thus the sector throughput) can be improved. When the sector load increases, more collisions occur and at full loading, the prioritization does not offer any gains over the uncoordinated scheme since the collisions are unavoidable and we cannot avoid any interference by using our schemes.

From Figure 5.6, we see that the two adaptive schemes can achieve better or the same performance as the uncoordinated scheme, depending on the sector load; the two static schemes can only perform better for low to moderate loads while the uncoordinated scheme slightly outperforms the static schemes for very high loads.

Compared to the uncoordinated scheme, the maximum gain for Adaptive Scheme 1 is approximately 25% and is achieved when the sector load is between 0.15 and 0.20. Adaptive Scheme 2 yields gains over a much larger range of sector loads than Adaptive Scheme 1; the maximum gain of about 43% is achieved when the sector load is between 0.25 and 0.35. As expected, Adaptive Scheme 2 achieves higher

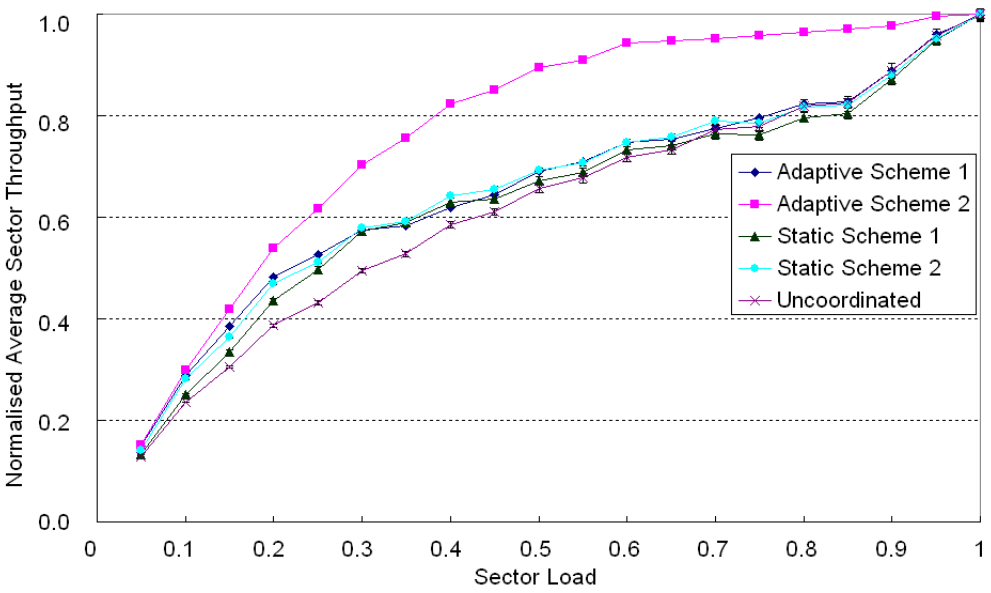

Figure 5.6: Normalised average sector throughput by applying different schemes.

gains than Adaptive Scheme 1 because the use of channel quality information gives a more accurate picture of the potential interference. Our results indicate that when the sector load is very low or very high (i.e. full load), our interference coordination schemes with adaptive priority assignments give the same performance as the uncoordinated approach. When the load is low, the probability of encountering significant interference with the uncoordinated approach is low, so the introduction of adaptive prioritization does not significantly improve average performance. When the load is high, as we have stated, the collisions between RBP allocations in neighbouring sectors are inevitable, so the prioritization does not help.

The static schemes achieve smaller gains than the adaptive schemes since the resources are allocated irrespective of the system state in the surrounding sectors. The maximum gain over the uncoordinated scheme is about 15% in Static Scheme 1, achieved when the sector load is about 0.30, and is approximately 21% in Static Scheme 2, achieved when the sector load is about 0.20. When the load is not very high, the performance of these two schemes is good as expected since the immediate adjacent sectors use different sub-bands. We also see that since the sub-bands defined in Static Scheme 2 are more fine-grained, the resource allocation can be done with

more flexibility, yielding better performance. However, the gains are less than those achieved in Adaptive Scheme 1 (about 25%) and Adaptive Scheme 2 (about 43%), which clearly indicates that the static schemes eliminate inter-base-station signalling at the expense of degraded performance.

We also observe that when the load is high, the static schemes do not help. For instance, when the load is greater than 0.75 (and less than 1), the uncoordinated scheme slightly outperforms Static Scheme 1 with respect to the average sector throughput. To explain how this might happen, we introduce an example in which we analyse the number of interfering sectors when the load is 0.8.

We consider the resource allocations in the reference sector (sector 0 of the reference site). As illustrated in Figure 5.3, the priority assignments in Static Scheme 1 are made according to category "A". Under the load of 0.8, all RBPs in sub-bands 1 and 2 are definitely allocated, and some RBPs in sub-band 3 are allocated and are randomly chosen. For any RBP in sub-band 1, among the 56 surrounding sectors in the considered network, there are 37 sectors definitely causing interference and 19 sectors with probability 0.4 of causing interference. Therefore, we can derive a probability distribution for the number of sectors which generate interference. The same analysis can be carried out for other sub-bands in Static Scheme 1. In the uncoordinated scheme, all RBPs have the same distribution (binomial distribution) for the number of interfering sectors. The cumulative distribution functions (CDF) of the number of interfering sectors for any RBP in each sub-band in the reference sector using Static Scheme 1 and the uncoordinated scheme are presented in Figure 5.7.

Next we determine the mean number of interfering sectors to give some hints on performance. For the RBPs in sub-bands 1 and 2, the mean number of interfering sectors is 44.6, which is slightly less than that for the uncoordinated scheme, which has a mean of 44.8. However, the mean number of interfering sectors for sub-band 3 is 45.2, which is greater than that for the uncoordinated scheme. In addition, the six immediately adjacent sectors definitely cause interference to the resources in sub-band 3. Therefore, from this point of view, the uncoordinated scheme could perform better than Static Scheme 1. Such an analysis for the number of interfering sectors can also be carried out for any other load. We conclude that the average sector

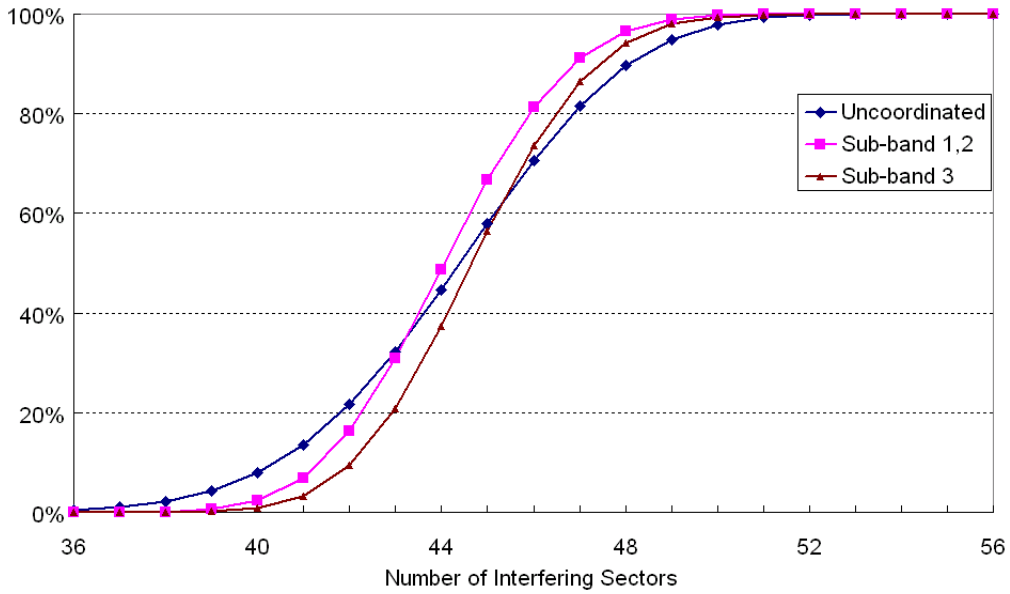

Figure 5.7: CDF of the number of interfering sectors for resources in the reference sector using Static Scheme 1 and the uncoordinated scheme when load= 0.8.

throughput in the static schemes may be not as good as that of the uncoordinated scheme when the sector load is high.

A fairness problem between the different sectors is also considered. We see that Adaptive Schemes 1 and 2 and Static Scheme 1 treat all sectors fairly because for a given load, all sectors have the same probability of suffering interference. In contrast, Static Scheme 2 does not treat all sectors fairly. Figure 5.8 shows the normalised average sector throughput in three sectors of the reference site as a function of sector load for Static Scheme 2 and the uncoordinated approach. We see that sector 0 and 2 achieve a higher gain on the average sector throughput than sector 1. To obtain a fair treatment, we need to introduce additional constraints during the resource priority planning; new heuristics, or other advanced techniques such as the use of graph theory or conventional reuse cluster, may help to obtain a better and fairer resource planning, which we leave for future work.

Next, we observe the performance improvement of our proposed schemes on the sector edge throughput. In Figure 5.9 we plot the normalised average sector edge throughput (normalised with the average sector throughput at full loading) as a function of the sector load for the adaptive and static schemes and the uncoordinated

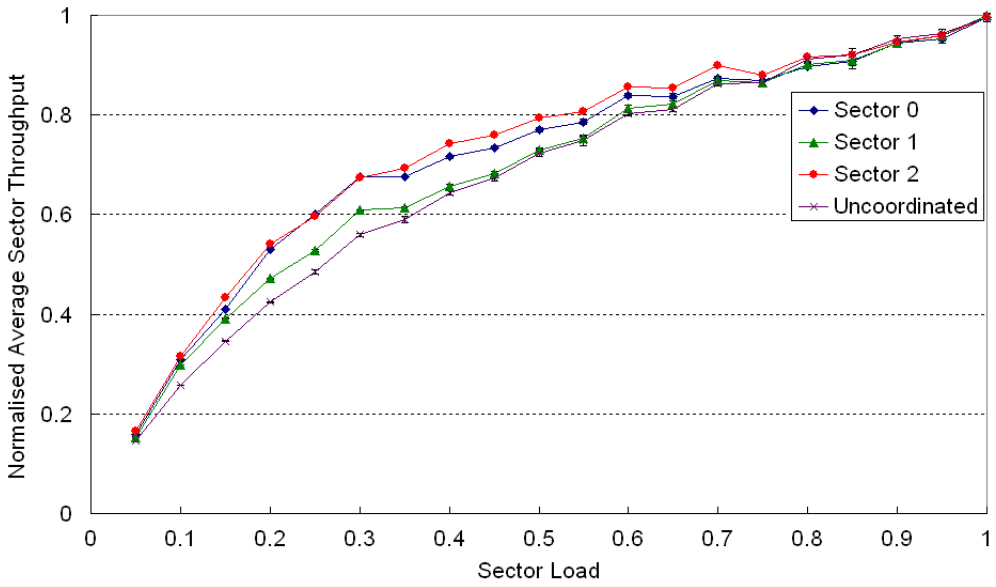

Figure 5.8: The throughput of Static Scheme 2 for different sectors of the reference site.

scheme. We see that compared to the uncoordinated approach, the adaptive schemes achieve higher gains than the static schemes because they use knowledge of the system state in the surrounding sectors to assign the priority. Static Scheme 2 performs better than Static Scheme 1 for the low loads due to the higher resolution of the priority assignment and the uncoordinated scheme beats the static schemes at high load.

For the static schemes, the gain on the average sector edge throughput for Static Scheme 1 reaches its maximum of approximately 21% when the sector load is about 0.30, while for Static Scheme 2, the gain reaches its maximal value of nearly 35% when the sector load is between 0.15 and 0.20. For the adaptive schemes, the maximum gain is about 41% for Adaptive Scheme 1 at about 0.20 sector load, and it is nearly 72% for Adaptive Scheme 2 at about 0.40 sector load. Note that Adaptive Scheme 2 offers substantial gain for a wide range of loads. For example, when the load is 0.75, Adaptive Scheme 2 can still achieve a gain of 60% on the sector edge throughput compared to the uncoordinated allocation. This is due to the fact that the resources with the potential of causing severe interference to the sector edge users are prohibited in neighbouring sectors.

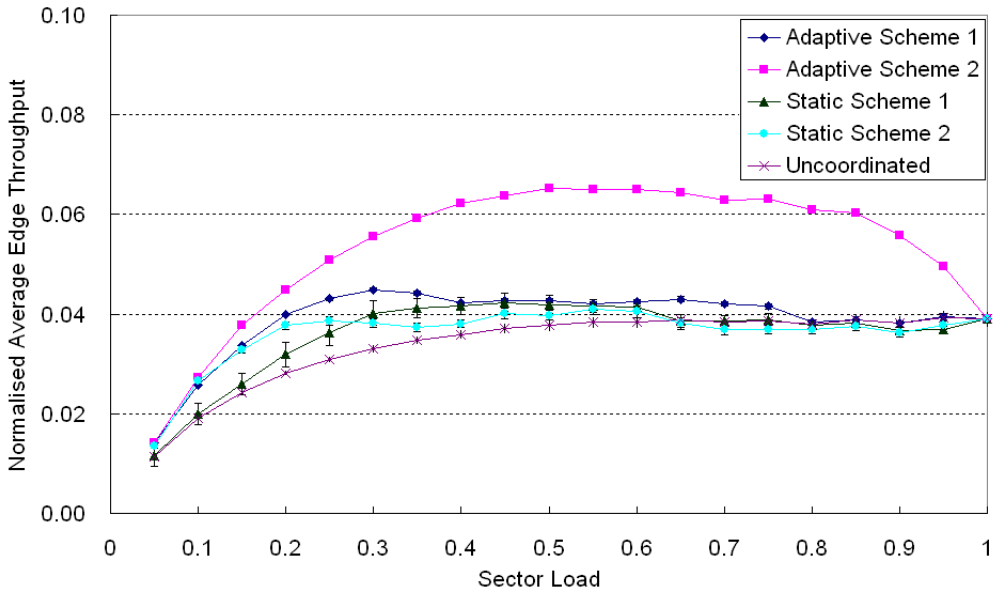

Figure 5.9: Normalised average sector edge throughput by applying different schemes.

From Figures 5.6 and 5.9, we also observe that for the static schemes, the sector load at which the maximum gain is achieved depends on the number of sub-bands. As shown in Figure 5.2, if there are M sub-bands in the system where there are $M-1$ dominant interfering sectors to the reference sector, the maximum gain should be achieved near the load $1/M$. Therefore, in the real implementation of the static schemes, we can decide the number of sub-bands according to the usual operating load of the network.

5.5.3 Sensitivity to the update interval T

In the previous section, we assumed full buffer traffic and that the update interval for the system state in the neighbouring base stations is $T = 1$ ms. The results show that the adaptive schemes achieve better performance by introducing additional inter-base-station signalling. However, as discussed in Section 5.3.1, a resulting issue is the volume and frequency of the inter-base-station signalling. In Adaptive Schemes 1 and 2, we impose the restriction that the base station only sends status reports to the neighbouring two tiers of sites to keep the signalling volume manageable. In

this section, we investigate the performance impact of reducing the frequency of the status reports.

The previous assumption of 1 ms status report update interval will incur very high signalling overhead to the network. One method to reduce the overhead is to relax this assumption and adopt an update interval in the order of tens of milliseconds or even seconds. First, we change the user traffic model to an on-off traffic model [114] (see Figure 2.10). In this model, each user generates one flow at one time, followed by a thinking time after completion of this flow, and a new flow is generated, and so on. We assume that the flow size follows a truncated log-normal distribution with a mean 2 Mbits, a standard deviation 0.722 Mbits, and a maximum 5 Mbits. The thinking time is based on an exponential distribution with mean 30 seconds.

We apply all other parameters and assumptions as described in Section 5.5.1 to the simulation, and simulate the flow arrivals and transmissions. We still assume a narrow-band service such that at any scheduling instant, only one RBP is allocated to one active user. We run the simulation for different status report update intervals, and observe the average user throughput and average edge user throughput (defined in Section 5.5.1) for our two adaptive schemes. By comparing with the performance of the uncoordinated scheme, we investigate the tradeoff between the system performance and signalling overhead.

According to our traffic assumptions, the flow transmission is in the order of seconds. Thus we propose four options for the update interval T: 1 ms, 10 ms, 100 ms, 1 s, which are smaller than the flow transmission duration. Figure 5.10 and Figure 5.11 depict the average user throughput and average edge user throughput for the two adaptive schemes and the uncoordinated scheme. For the new traffic model, Adaptive Scheme 2 still achieves higher gains than Adaptive Scheme 1; Adaptive Scheme 1 only performs better than the uncoordinated scheme when the number of users is not very large.

Comparing the performance for different update intervals, we find that, as expected, the achieved performance degrades with increasing T. However, the most degradation occurs in the step from $T = 1$ ms to $T = 10$ ms; the degradation in the

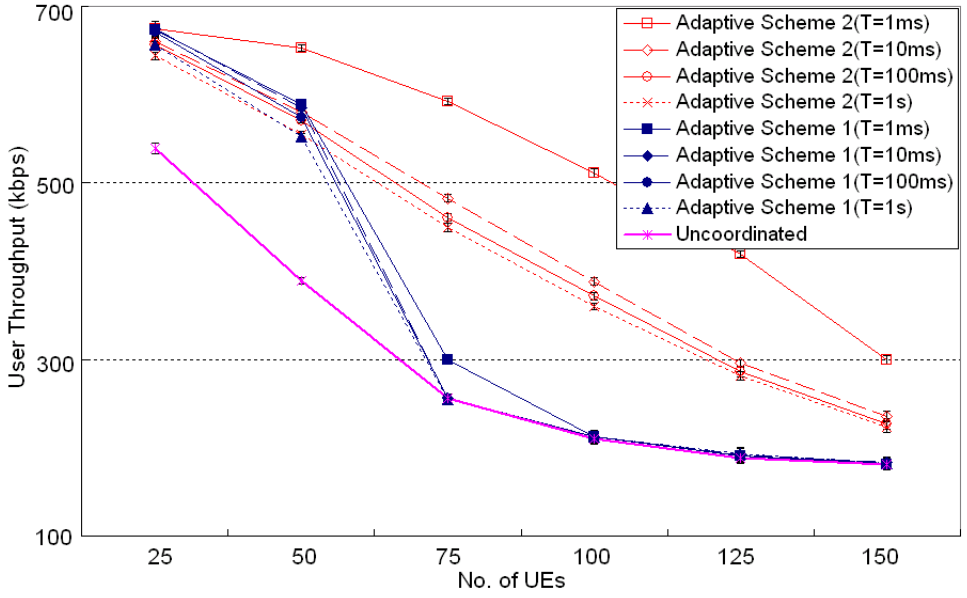

Figure 5.10: Comparison of average user throughput for two adaptive schemes and the uncoordinated scheme for different T.

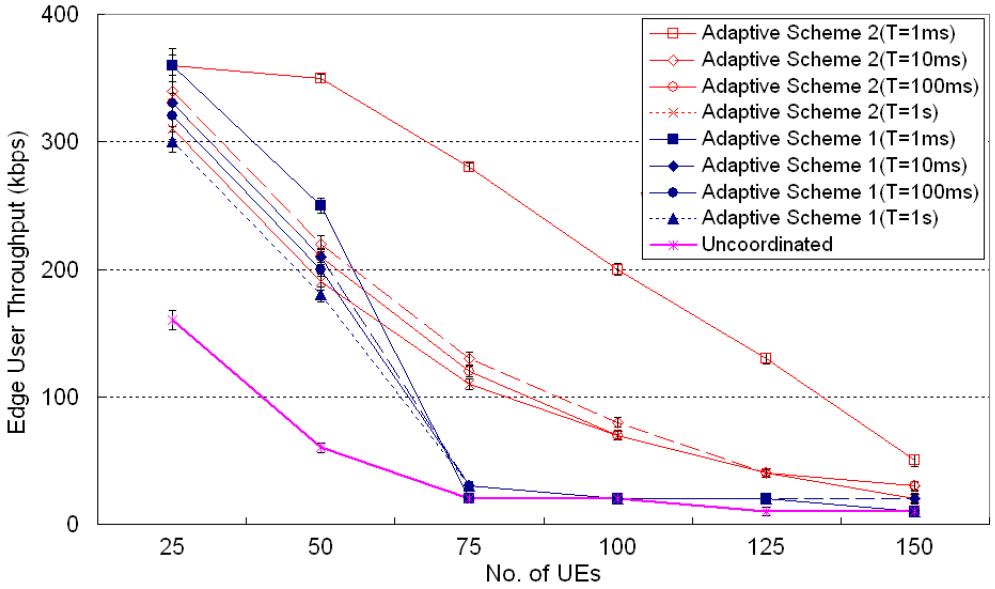

Figure 5.11: Comparison of average edge user throughput for two adaptive schemes and the uncoordinated scheme for different T.

performance from $T = 10$ ms to $T = 1$ s is only slight. For Adaptive Scheme 2, the gain is halved in most cases when we increase T from 1 ms to 1 s. However, we see that Adaptive scheme 2 can still give better performance than the uncoordinated approach even when the update interval is $T = 1$ s, while the overall signalling overhead is reduced to one thousandth of that in the case of $T = 1$ ms. For low numbers of users, Adaptive Scheme 1 with $T = 1$ s also gives better performance than no coordination.

5.6 Conclusion

In this chapter, we proposed possible distributed realizations of interference coordination schemes in a reuse-1 environment, which are based on setting allocation priority in the frequency domain. The allocation priority can be assigned statically through network configuration or made adaptive to traffic load variations in neighbouring sectors.

We proposed two schemes for the static case; Static Scheme 1 is based on the idea of the traditional reuse-3 scheme, and Static Scheme 2 uses a heuristic and greedy algorithm to assign the priorities. In the adaptive schemes, we introduced inter-base-station signalling to obtain the resource allocation information in the neighbouring sectors. Interference weights were defined to calculate the priority. Two adaptive schemes were proposed which differ in whether the interference weights are pre-computed reflecting the average interference impact (Adaptive Scheme 1) or user-specific depending on the user channel conditions (Adaptive Scheme 2). Numerical simulation results indicated that our schemes yield significant gains over conventional uncoordinated resource allocation in interference-limited networks when the sector load is not very high.

Furthermore, we investigated the tradeoff between the system performance and signalling overhead when applying the adaptive schemes. We saw from the simulation results that our adaptive schemes still achieve gains over the uncoordinated scheme when the update interval T is large but still relatively short compared to the transmission time of a flow. For Static Scheme 2, we need to seek new heuristics, or

other advanced techniques, to obtain a better and fairer resource planning, and we leave this for future work.

Chapter 6

Conclusion and future work

6.1 Introduction

In this chapter, we summarize the thesis by discussing the practical significance of the work, and we also present possible extensions for future work.

6.2 Summary of the work

For emerging cellular wireless systems such as LTE, the mitigation of inter-cell interference is the key to achieve a high capacity and good user experience. The central work of this thesis is to investigate the performance benefits of two interference mitigation techniques for the LTE downlink, namely reuse partitioning and resource prioritization in the frequency domain. The major significance of this thesis is summarized as follows.

In Chapter 3 and Chapter 4, a new metric, called the flow capacity, was introduced to characterise the user-level performance for elastic traffic. We used a generalized processor sharing queue to model a sector of a base station, and developed a queueing theoretic methodology, assuming infinite (Poisson arrivals) and finite user populations respectively, to calculate the flow capacity for standard reuse (reuse-1 and reuse-3) and reuse partitioning (FFR and SFR) with RR or PF scheduling. We showed that the capacity is strongly influenced by the sector edge rate, and is less sensitive to any improvement in the higher rates. The numerical results indicated that with both RR and PF scheduling, it is preferable to implement FFR with a judicious choice of edge band size when the level of provided service (or the flow throughput requirement) is low, to choose SFR with a judicious choice of edge band

size for a moderate level of service, and to configure a reuse-1 system for a high level of service. Furthermore, we applied our methodology to find the capacity benefits of MIMO schemes and saw that the spatial multiplexing MIMO scheme only provides a slight benefit since spatial multiplexing only increases the higher rates.

Within our analysis framework, we successfully took the PF scheduling gains into account by exploiting a fast convergence property of the multi-user diversity gain, and performed the analysis in the multi-cell scenario with fractional loading. Furthermore, to deal with the greater computational cost of the finite user population model, we applied efficient recursive algorithms (namely the MVAC and RECAL algorithms) to find the per-class flow throughputs for the case with a large number of classes and users. We also presented an open queueing network approximation based on the FPM method, and found that it provides good estimates in the regime of moderate to large flow throughput requirements.

In Chapter 5, we proposed possible implementations of interference coordination schemes in a reuse-1 environment, which are based on resource prioritization in the frequency domain. We showed that the proposed static and adaptive schemes yield significant gains with respect to average sector throughput and sector edge throughput over a conventional random allocation approach (uncoordinated allocation) in a fractionally loaded network when the sector load is not very high. Furthermore, for the adaptive schemes, we also investigated the effect of the inter-base-station signalling. We saw that the adaptive schemes still achieve gains over the uncoordinated scheme when the period of inter-base-station signalling is large but still relatively short compared to the transmission time of a flow.

6.3 Discussion and future research

In this section, we propose some possible extensions for future research.

In Section 3.5.2, with the assumption of Poisson flow arrivals from an infinite user population, we validated for RR scheduling that the per-class flow throughputs obtained using our hybrid simulation/analysis approach exhibit very close agreement with the per-class flow throughputs obtained using a pure simulation approach; for

PF scheduling, we found that the analysis results coincide with the simulation results only when the rate fluctuations of different classes are statistically identical. When different classes of users experience different fast fading statistics, we find that there are some differences between the analysis results and the simulation results, which indicates that the GPS queue is not a good model for this case. A possible solution might be to apply a more appropriate model, discriminatory processor sharing (see Altman, Avrachenkov and Ayesta [10] and Wu, Williamson and Luo [144]), to model the effects of multi-user diversity for this case.

In Chapter 3 and Chapter 4, we chose the flow throughput as the user-level performance metric to indicate the level of service, which is the ratio of the mean flow size to the mean flow duration. A more relevant throughput measure from the user's perspective is the call average throughput, which is the mean of the ratio of flow size to flow duration (see Kherani and Kumar [65, 66] and Litjens, Berg and Boucherie [80]). Therefore, an interesting extension would be to develop approximations to estimate the call average throughput for both infinite and finite user population models with multiple classes.

In Chapter 4, we presented an open queueing network approximation based on the FPM method, however, we found that it only works well in a certain operating regime. It may be better to apply asymptotic methods like [63, 87] to estimate the performance over a wider operating regime.

In Chapter 5, we saw that different sectors perform differently in Static Scheme 2, so we need to seek new heuristics or other advanced techniques to obtain a better and fairer resource planning. A possible solution might be that we apply the idea of the conventional reuse cluster and use our greedy algorithm to assign the priority within the cluster. Furthermore, we would try to apply our queueing analysis framework to determine the capacity benefits of resource prioritization.

Bibliography

[1] 3GPP TR 25.913. Requirements for evolved UTRA (E-UTRA) and evolved UTRAN (E-UTRAN) (Release 7) V7.3.0. March 2006.

[2] 3GPP TR 25.996. Spatial channel model for Multiple Input Multiple Output (MIMO) simulations (Release 9) V9.0.0. December 2009.

[3] 3GPP TS 36.101. Evolved universal terrestrial radio access (E-UTRA); user equipment (UE) radio transmission and reception (Release 9) V9.4.0. June 2010.

[4] 3GPP TS 36.211. Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation (Release 9) V9.1.0. March 2010.

[5] 3GPP TS 36.213. Evolved universal terrestrial radio access (E-UTRA); physical layer procedures (Release 9) V9.1.0. March 2010.

[6] 3GPP TS 36.300. Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description; stage 2 (Release 9) V9.3.0. March 2010.

[7] 3GPP TS 36.423. Evolved universal terrestrial radio access network (E-UTRAN); X2 application protocol (X2AP) (Release 9) V9.2.0. March 2010.

[8] 3GPP TS 36.942. Evolved universal terrestrial radio access network (E-UTRAN); radio frequency system scenarios (Release 8) V8.1.0. December 2008.

[9] M. H. Ahmed, H. Yanikomeroglu, and S. Mahmoud. Interference management using basestation coordination in broadband wireless access networks. *Wireless Communications and Mobile Computing*, 6(1):95–103, January 2006.

[10] E. Altman, K. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Systems*, 53(1):53–63, June 2006.

[11] J. G. Andrews. Interference cancellation for cellular systems: a contemporary overview. *IEEE Wireless Communications*, 12(2):19–29, April 2005.

[12] J. G. Andrews, W. Choi, and R. W. Heath. Overcoming interference in spatial multiplexing MIMO cellular networks. *IEEE Wireless Communications*, 14(6):95–104, December 2007.

[13] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2):248–260, April 1975.

[14] K. C. Beh, S. Armour, and A. Doufexi. Joint time-frequency domain proportional fair scheduler with HARQ for 3GPP LTE Systems. In *IEEE Vehicular Technology Conference, VTC-2008 Fall*, pages 1–5, September 2008.

[15] S. Bellofiore, C. A. Balanis, J. Foutz, and A. S. Spanias. Smart-antenna systems for mobile communication networks. Part 1. Overview and antenna design. *IEEE Antennas and Propagation Magazine*, 44(3):145–154, June 2002.

[16] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi. CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users. *IEEE Communications Magazine*, 38(7):70–77, July 2000.

[17] A. W. Berger and Y. Kogan. Dimensioning bandwidth for elastic traffic in high-speed data networks. *IEEE/ACM Transactions on Networking*, 8(5):643–654, October 2000.

[18] N. Bhushan, C. Lott, P. Black, R. Attar, Y.-C. Jou, M. Fan, D. Ghosh, and J. Au. CDMA2000 1xEV-DO Revision A: a physical layer and MAC layer overview. *IEEE Communications Magazine*, 44(2):37–49, February 2006.

[19] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor. *MIMO Wireless Communications*. Cambridge University Press, 2007.

[20] E. Biglieri, J. Proakis, and S. Shamai. Fading channels: information-theoretic and communications aspects. *IEEE Transactions on Information Theory*, 44(6):2619–2692, October 1998.

[21] T. Bonald. Flow-level performance analysis of some opportunistic scheduling algorithms. *European Transaction on Telecommunications*, 16(1):65–75, January 2005.

[22] T. Bonald, S. Borst, N. Hegde, and A. Proutiére. Wireless data performance in multi-cell scenarios. In *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems*, volume 32, pages 378–380, June 2004.

[23] T. Bonald, S. Borst, and A. Proutiére. How mobility impacts the flow-level performance of wireless data systems. In *Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2004*, volume 3, pages 1872–1881, March 2004.

[24] T. Bonald, S. Borst, and A. Proutiére. Inter-cell scheduling in wireless data networks. In *European Wireless*, April 2005.

[25] T. Bonald and N. Hegde. Capacity gains of some frequency reuse schemes in OFDMA networks. In *IEEE Global Telecommunications Conference, GLOBECOM'09*, pages 1–6, November 2009.

[26] T. Bonald and A. Proutiére. Wireless downlink data channels: user performance and cell dimensioning. In *the 9th annual international conference on mobile computing and networking, MobiCom 03*, pages 339–352, September 2003.

[27] T. Bonald and A. Proutiére. On stochastic bounds for monotonic processor sharing networks. *Queueing Systems*, 47(1):81–106, May 2004.

[28] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. *IEEE/ACM Transactions on Networking*, 13(3):636–647, June 2005.

[29] G. Boudreau, J. Panicker, N. Guo, R. Chang, Ne.Wang, and S. Vrzic. Interference coordination and cancellation for 4G networks. *IEEE Communications Magazine*, 47(4):74–81, April 2009.

[30] J. P. Buzen. Computational algorithms for closed queueing networks with exponential servers. *Communications of the ACM*, 16(9):527–531, September 1973.

[31] J. P. Buzen and P. S. Goldberg. Guidelines for the use of infinite source queueing models in the analysis of computer system performance. In *AFIPS74: Proceedings of May 6-10, 1974, National Computer Conference and Exposition*, pages 371–374, May 1974.

[32] M. Cho, N. Kim, and H. Yoon. User-perceived QoS in a wireless packet network with multiple channel conditions. *IEICE Transactions on Communications*, E89-B(2):342–349, February 2006.

[33] Y.-J. Choi, C. S. Kim, and S. Bahk. Flexible design of frequency reuse factor in OFDMA cellular networks. In *IEEE International Conference on Communications, ICC06*, volume 4, pages 1784–1788, June 2006.

[34] G. L. Choudhury, K. K. Leung, and W. Whitt. Calculating normalization constants of closed queueing networks by numerically inverting their generating functions. *Journal of the ACM*, 42(5):935–970, September 1995.

[35] B. Classon, P. Sartori, V. Nangia, X. Zhuang, and K. Baum. Multi-dimensional adaptation and multi-user scheduling techniques for wireless OFDM systems. In *IEEE International Conference on Communications ICC03*, volume 3, pages 2251–2255, May 2003.

[36] J. W. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12(3):245–284, October 1979.

[37] A. E. Conway and N. D. Georganas. RECAL - a new efficient algorithm for the exact analysis of multiple-chain closed queueing networks. *Journal of the ACM*, 33(4):768–791, October 1986.

[38] A. E. Conway, E. De Souza E Silva, and S. S. Lavenberg. Mean value analysis by chain of product form queueing networks. *IEEE Transactions on Computers*, 38(3):432–442, March 1989.

[39] E. Dahlman, S. Parkvall, J. Skold, and P. Beming. *3G Evolution: HSPA and LTE for Mobile Broadband*. Acadamic Press, 2007.

[40] E. de Souza e Silva and S. S. Lavenberg. Calculating joint queue-length distributions in product-form queuing networks. *Journal of the ACM*, 36(1):194–207, January 1989.

[41] S.-E. Elayoubi and B. Fourestié. On frequency allocation in 3G LTE systems. In *IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC2006*, pages 1–5, September 2006.

[42] S.-E. Elayoubi, O. B. Haddada, and B. Fourestié. Performance evaluation of frequency planning schemes in OFDMA-based networks. *IEEE Transactions on Wireless Communications*, 7(5):1623–1633, May 2008.

[43] G. Fodor. Performance analysis of a reuse partitioning technique for OFDM based evolved UTRA. In *IEEE International Workshop on Quality of Service, IWQoS 2006*, pages 112–120, June 2006.

[44] G. Fodor, M. Telek, and C. Koutsimanis. Performance analysis of scheduling and interference coordination policies for OFDMA networks. *Computer Networks*, 52(6):1389–1286, January 2008.

[45] G. J. Foschini. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Technical Journal*, 1(2):41–59, Summer 1996.

[46] G. J. Foschini and Z. Miljanic. A simple distributed autonomous power control algorithm and its convergence. *IEEE Transactions on Vehicular Technology*, 42(4):641–646, November 1993.

[47] S. B. Fredj, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts. Statistical bandwidth sharing: a study of congestion at flow level. *ACM SIGCOMM Computer Communication Review*, 31(4):111–122, October 2001.

[48] D. Gesbert, S. G. Kiani, A. Gjendemsjo, and G. E. Oien. Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks. *Proceedings of the IEEE*, 95(12):2393–2409, December 2007.

[49] D. Gesbert, M. Shafi, D. Shiu, P. J. Smith, and A. Naguib. From theory to practice: an overview of MIMO space-time coded wireless systems. *IEEE Journal on Selected Areas in Communications*, 21(3):281–302, April 2003.

[50] A. Goldsmith. *Wireless Communications*. Cambridge University Press, 2005.

[51] T. Halonen, J. Romero, and J. Melero. *GSM, GPRS, and EDGE Performance: Evolution towards 3G/UMTS*. Wiley Publishing, 2003.

[52] A. Hernández, I. Guío, and A. Valdovinos. Radio resource allocation for interference management in mobile broadband OFDMA based networks. *Wireless Communications and Mobile Computing*, July 2009.

[53] H. Holma and A. Toskala. *WCDMA for UMTS C HSPA Evolution and LTE*. Wiley Publishing, 2007.

[54] H. Hu, J. Luo, and H. Chen. Radio resource management for cooperative wireless communication systems with organized beam-hopping techniques. *IEEE Wireless Communications*, 15(2):100–109, April 2008.

[55] IEEE 802.16e-2005. Part 16: air interface for fixed and mobile broadband wireless access systems. Feburary 2006.

[56] IST-4-027756 WINNER II. D4.7.1 v.1.0 Interference averaging concepts. June 2007.

[57] IST-4-027756 WINNER II. D4.7.2 v.1.0 Interference avoidance concepts. June 2007.

[58] IST-4-027756 WINNER II. D4.7.3 v.1.0 Smart antenna based interference mitigation. June 2007.

[59] ITU-D. *ITU-D Measuring the Information Society 2010*. 2010.

[60] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela. Network coordination for spectrally efficient communications in cellular systems. *IEEE Wireless Communications*, 13(4):56–61, August 2006.

[61] I. Katzela and M. Naghshineh. Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey. *IEEE Personal Communications*, 3(3):10–31, June 1996.

[62] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M. Moisio. Dynamic packet scheduling performance in UTRA Long Term Evolution downlink. In *3rd International Symposium on Wireless Pervasive Computing, ISWPC 2008*, pages 308–313, May 2008.

[63] F. P. Kelly. On a class of approximations for closed queueing networks. *Queueing Systems*, 4(1):69–76, March 1989.

[64] F. Khan. *LTE for 4G Mobile Broadband*. Cambridge University Press, 2009.

[65] A. A. Kherani and A. Kumar. Stochastic models for throughput analysis of randomly arriving elastic flows in the internet. In *Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2002*, volume 2, pages 1014–1023, November 2002.

[66] A. A. Kherani and A. Kumar. On processor sharing as a model for TCP controlled HTTP-like transfers. In *IEEE International Conference on Communications, ICC04*, volume 4, pages 2256–2260, June 2004.

[67] Y. Kim, B. J. Jeong, J. Chung, C.-S. Hwang, J. S. Ryu, K.-H. Kim, and Y. K. Kim. Beyond 3G: vision, requirements, and enabling technologies. *IEEE Communications Magazine*, 41(3):120–124, March 2003.

[68] L. Kleinrock. Time-shared systems: a theoretical treatment. *Journal of the ACM*, 14(2):242–261, April 1967.

[69] L. Kleinrock. *Queueing Systems Volume I: Theory.* A Wiley-Interscience publication, United States of America, 1975.

[70] L. Kleinrock. *Queueing Systems Volume II: Computer Applications.* A Wiley-Interscience publication, United States of America, 1976.

[71] R. Kwan and C. Leung. A survey of scheduling and interference mitigation in LTE. *Journal of Electrical and Computer Engineering*, 2010:10 pages, 2010.

[72] L. Lei, C. Lin, J. Cai, and X. Shen. Flow-level performance of opportunistic OFDM-TDMA and OFDMA networks. *IEEE Transactions on Wireless Communications*, 7(12):5461–5472, December 2008.

[73] P. Lescuyer and T. Lucidarme. *Evolved Packet System (EPS) the LTE and SAE Evolution of 3G UMTS.* Wiley Publishing, 2008.

[74] K. K. Leung and A. Srivastava. Dynamic allocation of downlink and uplink resource for broadband services in fixed wireless networks. *IEEE Journal on Selected Areas in Communications*, 17(5):990–1006, May 1999.

[75] G. Li and H. Liu. Downlink dynamic resource allocation for multi-cell OFDMA system. In *IEEE Vehicular Technology Conference, VTC-2003 Fall*, volume 3, pages 1698–1702, October 2003.

[76] G. Li and H. Liu. Dynamic resource allocation with finite buffer constraint in broadband OFDMA networks. *IEEE Wireless Communications and Networking, WCNC 2003*, 2:1037–1042, March 2003.

[77] G. Li and H. Liu. Downlink radio resource allocation for multi-cell OFDMA system. *IEEE Transactions on Wireless Communications*, 5(12):3451–3459, December 2006.

[78] Y. Li, N. Seshadri, and S. Ariyavisitakul. Channel estimation for OFDM systems with transmitter diversity in mobile wireless channels. *IEEE Journal on Selected Areas in Communications*, 17(3):461–471, March 1999.

[79] J. C. Liberti and T. S. Rappaport. *Smart Antennas for Wireless Communications: IS-95 and Third Generation CDMA Applicationss*. Prentice-Hall, 1999.

[80] R. Litjens, H. van den Berg, and R. J. Boucherie. Throughputs in processor sharing models for integrated stream and elastic traffic. *Performance Evaluation*, 65(2):152–180, February 2008.

[81] L. Liu and P. Li. Performance analysis of inter-cell interference mitigation with beam-forming for E-UTRA system. *SciencePaper Online*, October 2006.

[82] S. Liu. Data traffic performance analysis of a cellular system with finite user population. Master's thesis, Department of Electrical and Communications Engineering, Helsinki University of Technology, August 2004.

[83] S. Liu and J. Virtamo. Performance analysis of wireless data systems with a finite population of mobile users. In *19th International Teletraffic Congress, ITC-19*, pages 1295–1304, August 2005.

[84] S. Liu and J. Virtamo. Inter-cell coordination with inhomogeneous traffic distribution. In *Conference on Next Generation Internet Design and Engineering, NGI'06*, pages 64–71, April 2006.

[85] M. Maqbool, M. Coupechoux, and P. Godlewski. Dimensioning methodology for OFDMA networks. In *Proceedings of Wireless World Research Forum (WWRF), 22nd meeiting*, May 2009.

[86] L. Massoulié and J. Roberts. Bandwidth sharing: objectives and algorithms. *IEEE/ACM Transactions on Networking*, 10(3):320–328, June 2002.

[87] J. McKenna, D. Mitra, and K. G. Ramakrishnan. A class of closed Markovian queueing networks: integral representations, asymptotic expansions and

generalizations. *The Bell System Technical Journal*, 60(5):599–641, May-June 1981.

[88] D. Mitra and A. Weiss. A closed network with a discriminatory processor-sharing server. *ACM SIGMETRICS Performance Evaluation Review*, 17(1):200–208, May 1989.

[89] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela. LTE capacity compared to the Shannon bound. In *IEEE Vehicular Technology Conference, VTC2007-Spring*, pages 1234–1238, April 2007.

[90] R. D. Murch and K. B. Letaief. Antenna systems for broadband wireless access. *IEEE Communications Magazine*, 40(4):76–83, April 2002.

[91] R. U. Nabar, H. Bolcskei, and F. W. Kneubuhler. Fading relay channels: performance limits and space-time signal design. *IEEE Journal on Selected Areas in Communications*, 22(6):1099–1109, August 2004.

[92] M. C. Necker. Towards frequency reuse 1 cellular FDM/TDM systems. In *International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems, MSWiM 2006*, pages 338–346, October 2006.

[93] M. C. Necker. Local interference coordination in cellular OFDMA networks. In *IEEE Vehicular Technology Conference, VTC-2007 Fall*, pages 1741–1743, September 2007.

[94] T. C. Y. Ng and W. Yu. Joint optimization of relay strategies and resource allocations in cooperative cellular networks. *IEEE Journal on Selected Areas in Communications*, 25(2):328–339, February 2007.

[95] T. T. Nielsen and J. Wigard. *Performance Enhancements in a Frequency Hopping GSM Network*. Kluwer Academic, 2000.

[96] L. Nuaymi. *WiMAX: Technology for Broadband Wireless Access*. Wiley Publishing, 2007.

[97] R. Pabst, B. H. Walke, D. C. Schultz, P. Herhold, H. Yanikomeroglu, S. Mukherjee, H. Viswanathan, M. Lott, W. Zirwas, M. Dohler, H. Aghvami, D. D. Falconer, and G. P. Fettweis. Relay-based deployment concepts for wireless and mobile broadband radio. *IEEE Communications Magazine*, 42(9):80–89, September 2004.

[98] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei. An overview of MIMO communications - a key to gigabit wireless. *Proceedings of the IEEE*, 92(2):198–218, Feburary 2004.

[99] M. Pischella and J. C. Belfiore. Power control in distributed cooperative OFDMA cellular networks. *IEEE Transactions on Wireless Communications*, 7(5):1900–1906, May 2008.

[100] B. Pittel. Closed exponential networks of queues with saturation: the Jackson-type stationary distribution and its asymptotic analysis. *Mathematics of Operations Research*, 4(4):357–378, November 1979.

[101] A. Pokhariyal, T. E. Kolding, and P. E. Mogensen. Performance of downlink frequency domain packet scheduling for the UTRAN Long Term Evolution. In *IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC2006*, pages 1–5, September 2006.

[102] A. Pokhariyal, G. Monghal, K. I. Pedersen, P. E. Mogensen, I. Z. Kovacs, C. Rosa, and T. E. Kolding. Frequency domain packet scheduling under fractional load for the UTRAN LTE downlink. In *IEEE Vehicular Technology Conference, VTC2007-Spring*, pages 699 –703, April 2007.

[103] A. Pokhariyal, K. I. Pedersen, G. Monghal, I. Z. Kovacs, C. Rosa, T. E. Kolding, and P. E. Mogensen. HARQ aware frequency domain packet scheduler with different degrees of fairness for the UTRAN Long Term Evolution. In *IEEE Vehicular Technology Conference, VTC-2007 Spring*, pages 2761–2765, April 2007.

[104] D. A. Protopapas. Finite queueing approximation techniques for analysis of computer systems. In *AFIPS81: Proceedings of the May 4-7, 1981, National Computer Conference*, pages 423–429, May 1981.

[105] 3GPP R1-050589. Pilot channel and scrambling code in evolved UTRA downlink. In *3GPP TSG RAN WG1 Ad Hoc on LTE*, June 2005.

[106] 3GPP R1-050896. Description and simulations of interference management technique for OFDMA based E-UTRA downlink evaluation. In *3GPP TSG-RAN WG1 Meeting No.42*, August 2005.

[107] 3GPP R1-051043. Downlink time-frequency diversity transmission. In *3GPP TSG-RAN WG1 Meeting No.42bis*, October 2005.

[108] 3GPP R1-051517. Multiplexing distributed & localized allocations. In *3GPP TSG-RAN WG1 Meeting No.43*, November 2005.

[109] 3GPP R1-060038. Distributed OFDMA transmission for shared data channel in E-UTRA downlink. In *3GPP TSG-RAN WG1 LTE Ad Hoc Meeting*, January 2006.

[110] 3GPP R1-060095. E-UTRA DL - Localized and distributed transmission. In *3GPP TSG-RAN WG1 LTE Ad Hoc Meeting*, January 2006.

[111] 3GPP R1-060291. OFDMA downlink inter-cell interference mitigation. In *3GPP TSG RAN WG1 Meeting No.44*, February 2006.

[112] 3GPP R1-060864. Overview of resource management techniques for interference mitigation in EUTRA. In *3GPP TSG-RAN WG1 Meeting No.44bis*, March 2006.

[113] 3GPP R1-062712. Scrambling code in E-UTRA downlink. In *3GPP TSG-RAN WG1 Meeting No.46*, October 2006.

[114] 3GPP R1-070674. LTE physical layer framework for performance verification. In *3GPP TSG-RAN WG1 Meeting No.48*, February 2007.

[115] 3GPP R1-071967. DL E-UTRA Performance Checkpoint. In *3GPP TSG-RAN WG1 Telephone conference*, April 2007.

[116] 3GPP R1-080887. CQI measurement methodology. In *3GPP TSG-RAN WG1 Meeting No.52*, February 2008.

[117] 3GPP R1-082014. Further results on CQI measurement methodology. In *3GPP TSG-RAN WG1 Meeting No.53*, May 2008.

[118] M. Rahman and H. Yanikomeroglu. Interference avoidance through dynamic downlink OFDMA subchannel allocation using intercell coordination. In *IEEE Vehicular Technology Conference, VTC-2008 Spring*, pages 1630–1635, May 2008.

[119] M. Rahman, H. Yanikomeroglu, and W. Wong. Interference avoidance with dynamic inter-cell coordination for downlink LTE system. In *IEEE Wireless Communications and Networking Conference, WCNC 2009*, pages 1–6, April 2009.

[120] M. Reiser and S. S. Lavenberg. Mean-value analysis of closed multichain queuing networks. *Journal of the ACM*, 27(2):313–322, May 1980.

[121] M. Rumney. *LTE and the Evolution to 4G Wireless: Design and Measurement Challenges*. Agilent Technologies, 2009.

[122] N. K. Shankaranarayanan, Z. Jiang, and P. Mishra. User-perceived performance of web-browsing and interactive data applications in TDMA packet wireless networks. *Multiaccess, Mobility and Teletraffic for Wireless Communications*, 5:73–84, 2000.

[123] N. K. Shankaranarayanan, Z. Jiang, and P. Mishra. Performance of a shared packet wireless network with interactive data users. *Mobile Networks and Applications*, 8(3):279–293, June 2003.

[124] Z. Shen, J. G. Andrews, and B. L. Evans. Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints. *IEEE Transactions on Wireless Communications*, 4(6):2726–2737, November 2005.

[125] A. Simonsson. Frequency reuse and intercell interference co-ordination in E-UTRA. In *IEEE Vehicular Technology Conference, VTC-2007 Spring*, pages 3091–3095, April 2007.

[126] G. Song and Y. Li. Cross-layer optimization for OFDM wireless network - Part I: theoretical framework. *IEEE Transactions on Wireless Communications*, 4(2):614–624, March 2005.

[127] G. Song and Y. Li. Cross-layer optimization for OFDM wireless network - Part II: algorithm development. *IEEE Transactions on Wireless Communications*, 4(2):625–634, March 2005.

[128] M. Sternad, T. Svensson, T. Ottosson, A. Ahlen, A. Svensson, and A. Brunstrom. Towards systems beyond 3G based on adaptive OFDMA transmission. *Proceedings of the IEEE*, 95(12):2432–2455, December 2007.

[129] A. L. Stolyar and H. Viswanathan. Self-organizing dynamic fractional frequency reuse in OFDMA systems. In *The 27th Conference on Computer Communications, IEEE INFOCOM 2008*, pages 691–699, April 2008.

[130] A. L. Stolyar and H. Viswanathan. Self-organizing dynamic fractional frequency reuse through distributed inter-cell coodination: the case of best-effort traffic. *Bell Labs Technical Memo*, June 2008.

[131] G. L. Stuber. *Principles of Mobile Communication, Second Edition.* Kluwer Academic Publishers, 2002.

[132] P. Svedman, S. K. Wilson, L. J. Cimini, and B. Ottersten. Opportunistic beamforming and scheduling for OFDMA systems. *IEEE Transactions on Wireless Communications*, 55(5):941–952, May 2007.

[133] W. H. Tranter, K. S. Shanmugan, T. S. Rappaport, and K. L. Kosbar. *Principles of Communication Systems Simulation with Wireless Applications.* PRENTICE HALL, 2003.

[134] J.-J. van de Beek, O. Edfors, M. Sandell, S. K. Wilson, and P. O. Borjesson. On channel estimation in ofdm systems. In *IEEE Vehicular Technology Conference, VTC-1995*, volume 2, pages 815–819, July 1995.

[135] W. Wang, K. C. Hwang, K. B. Lee, and S. Bahk. Resource allocation for heterogeneous services in multiuser OFDM systems. In *IEEE Global Telecommunications Conference, GLOBECOM'04*, volume 6, pages 3478–3481, November 2004.

[136] N. Wei, A. Pokhariyal, T. B. Sorensen, T. E. Kolding, and P. E. Mogensen. Performance of MIMO with frequency domain packet scheduling in UTRAN LTE downlink. In *IEEE Vehicular Technology Conference, VTC2007-Spring*, pages 1177–1181, April 2007.

[137] C. Wengerter, J. Ohlhorst, and A. G. E. von Elbwart. Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA. In *IEEE Vehicular Technology Conference, VTC-2005 Spring*, volume 3, pages 1903–1907, May 2005.

[138] W. Whitt. Open and Closed Models for Networks of Queues. *AT&T Bell Laboratories Technical Journal*, 63(9):1911–1979, November 1984.

[139] S. H. Won, H. J. Park, J. O. Neel, and J. H. Reed. Inter-cell interference coordination/avoidance for frequency reuse by resource scheduling in an OFDM-based cellular system. In *IEEE Vehicular Technology Conference, VTC-2007 Fall*, pages 1722–1725, September 2007.

[140] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch. Multiuser OFDM with adaptive subcarrier, bit, and power allocation. *IEEE Journal on Selected Areas in Communication*, 17(10):1747–1758, October 1999.

[141] W. Wu, M. Gitlits, and T. Sakurai. Dynamic resource allocation with inter-cell interference coordination for 3GPP LTE. In *Asia Pacific Microwave Conference, APMC 2008*, pages 1–4, December 2008.

[142] W. Wu and T. Sakurai. Capacity of reuse partitioning schemes for OFDMA wireless data networks. In *IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC2009*, pages 2240–2244, September 2009.

[143] W. Wu and T. Sakurai. Flow-level capacity of fractionally loaded OFDMA networks with proportional fair scheduling. In *IEEE Vehicular Technology Conference, VTC-2010 Fall*, pages 1–5, September 2010.

[144] Y. Wu, C. Williamson, and J. Luo. On processor sharing and its applications to cellular data network provisioning. *Performance Evaluation*, 64:9–12, October 2007.

[145] G. Wunder, C. Zhou, H.-E. Bakker, and S. Kamins. Throughput maximization under rate requirements for the OFDMA downlink channel with limited feedback. *EURASIP Journal on Wireless Communications and Networking*, 2008:1–11, January 2008.

[146] Y. Xiang, J. Luo, and C. Hartmann. Inter-cell interference mitigation through flexible resource reuse in OFDMA based communication networks. In *European Wireless 2007*, April 2007.

[147] H. Yin and S. Alamouti. OFDMA: a broadband wireless access technology. In *IEEE Sarnoff Symposium 2006*, pages 1–4, March 2006.

[148] H. Zhang, X. Xu, J. Li, X. Tao, S. Tommy, and B. Carmen. Performance of power control in inter-cell interference coordination for frequency reuse. *The Journal of China Universities of Posts and Telecommunications*, 17(1):37–43, February 2010.

[149] Y. Zhang and K. B. Letaief. Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems. *IEEE Transactions on Wireless Communications*, 3(5):1566–1575, September 2004.