UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

On advantages of cooperation in cellular systems : throughput and heavy traffic performance

Permalink

https://escholarship.org/uc/item/4gg609zt

Author

Bhardwaj, Sumit

Publication Date

2008

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

On Advantages of Cooperation in Cellular Systems: Throughput and Heavy Traffic Performance

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy

in

Electrical Engineering
(Communication Theory and Systems)

by

Sumit Bhardwaj

Committee in charge:

Professor Anthony S. Acampora, Chair
Professor Ruth J. Williams, Co-Chair
Professor Rene L. Cruz
Professor Massimo Franceschetti
Professor Joseph Pasquale

2008

© Copyright
Sumit Bhardwaj, 2008
All rights reserved

The dissertation of Sumit Bhardwaj is approved, and is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2008

To one who might have sung;

who only listened.

- Adapted from an early work of F. Scott Fitzgerald.

TABLE OF CONTENTS

LIST OF FIGURES

List of Algorithms

LIST OF TABLES

ACKNOWLEDGMENTS

Chief among those to whom I owe thanks for help in making this dissertation a reality are my advisors: Professors A. S. Acampora and R. J. Williams. Thanks for all the patience, encouragement, and valuable advice throughout my studies.

I thank my doctoral committee members, Professors Rene Cruz, Massimo Franceschetti, and Joseph Pasquale for insightful discussions, time, and effort.

I thank M'Lissa Michelson, John Minan, Karol Previte, Robert Rome, Michelle Vavra, and the rest of the staff (and the ex-staff) of the ECE department for support, advice, and for always being extremely helpful.

A Ph.D. is a long journey and I consider myself fortunate to know people who tried their best to ensure that I stayed the course. Thusly, I must thank Ramesh Annavajjala, Aditya K. Jagannatham, Anuj Mishra, Yoav Nebat, Surendra Prasad, Akshay Sharma, Michael Tan, Ron Tamari, and Nicole and Edward Truitt. I can only hope that I retain the privilege of calling you friend - in the sense of John Milton's epistle to Charles Diodati in 1637 - over my lifetime. Thanks are also due to some who must remain anonymous at their own request. Aaron Kemp and his family has been a surrogate family for my stay in San Diego and I must thank them for their kindness and generosity of spirit.

The summer of 2007, I was an intern at Motorola at Arlington Heights and I had a very rewarding experience: intellectually and otherwise. Thanks are due to all those who made those days such a wonderful time: in particular, to Phil Fleming, Scott Clapp, Ron Crocker, Tony Dean, Paul Hustedde, Alan Jette, and Randall Kohl.

My family has been a constant source of joy and support, and I am sure that any attempt at conveying my gratitude for them will be woefully inadequate. Nonetheless, here is such an attempt; thanks for your unconditional love and support.

Finally, ***gratias agimus tibi propter magnam gloriam tuam.***

Chapter 2, in part, is a reprint of the material in the following papers: A. S. Acampora, S. Bhardwaj, and R. M. Tamari, "On Best-Case Throughput of Cellular Data Networks with Cooperating Base Stations", in the Proceedings of *the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep 27–29 2006; A. S. Acampora, S. Bhardwaj, and R. M. Tamari, "A Best-Case Performance Comparison of Cellular Data Networks with Cooperating and Non-cooperating Base Stations", submitted to *Wireless Communications and Mobile Computing*. The dissertation author was the primary investigator and author of these papers.

Chapter 3, in part, is a reprint of the material in the following paper: S. Bhardwaj, R. J. Williams, and A. S. Acampora, "On the Performance of a Two-User MIMO Downlink System in Heavy Traffic", *IEEE Trans. Information Theory*, vol. 53, no. 5, pp. 1851–1859, May 2007. The dissertation author was the primary investigator and author of this paper.

VITA

2003	B. Tech., Electrical Engineering Indian Institute of Technology - Delhi
2004–2007	Research Assistant Department of Electrical and Computer Engineering University of California, San Diego
2005	M.S., Electrical and Computer Engineering University of California, San Diego
2008	Ph.D., Electrical and Computer Engineering University of California, San Diego

ABSTRACT OF THE DISSERTATION

On Advantages of Cooperation in Cellular Systems: Throughput and Heavy Traffic Performance

by

Sumit Bhardwaj

Doctor of Philosophy in Electrical Engineering
(Communication Theory and Systems)

University of California, San Diego, 2008

Professor Anthony S. Acampora, Chair

Professor Ruth J. Williams, Co-Chair

The object of interest in this dissertation is a cellular wireless system with cooperation among base stations. We study such systems from a cross-layer point of view.

In the first part of the dissertation, we investigate the maximum throughput of such a system. Assuming that the relative traffic of each mobile is specified in advance and some simplifying assumptions on the underlying channel, we show that the maximum stable throughput can be expressed in terms of the capacity of the channel. We then formulate a queueing model for this system and propose a throughput-achieving service policy. We then propose a fixed-point approximation as a tool for the performance analysis of this policy. We then proceed, via simulations, to demonstrate the advantage of

cooperation over the traditional operation of such systems. Since the proposed policy is computationally expensive, we propose a practical, albeit suboptimal, scheme for large systems. We quantify, again by means of simulations, the advantage of the proposed scheme over the traditional operation.

We next study the performance of the queueing network in heavy traffic. Specifically, we prove limit theorems justifying a diffusion approximation for a heavily loaded system operating under the policy proposed earlier. We first show that for a two-user system,the renormalized queuelength process converges in distribution to a Semimartingale Reflecting Brownian Motion (SRBM) living in a two-dimensional quadrant. Using different techniques, we next show that for an arbitrary sized system, the renormalized workload process converges in distribution to an SRBM living in the N-dimensional positive orthant where N is the number of users in the system.

CHAPTER 1

Introduction

Cellular wireless networks have been in operation for the better part of the last two decades, with improvements all along the way. (An excellent introduction to the subject matter is the book by Schwartz [36]; also see the references therein.) These networks have traditionally operated in accordance with a simple time-honored approach: divide the overall region to be served into individual radio cells, each with its own base station or access point, and serve all mobile stations within the footprint of each given cell from that cell's base station. With the possible exception of any transient that may briefly occur during an intercell hand-off, at which time a mobile station may be in simultaneous communication with two base stations, communications within all currently deployed cellular and wireless internet access systems are strictly between each individual mobile station and its respective serving base station. In fact, in most of the systems, communications between the base stations in other cells and their mobile station clients represent a source of interference with regard to mobile-to-base station communications within any given cell.

A different approach would be to let each of the mobile stations within the footprint of a set of base stations be served by all base stations in that set. Corresponding to the fact that the base stations constitute the infrastructure of the cellular wireless systems, this is known in the literature as "infrastructure cooperation" [40]. (In the sequel, we may sometimes refer to infrastructure cooperation by "base station cooperation" as well.)

Specifically, we consider a cellular wireless access network in which the base sta-

tions can cooperate over noise-free, wireline links with infinite capacity. Such links do, in fact, effectively exist within portions of today's cellular infrastructure in the form of the buried fiber optical cabling used to connect base stations with the Mobile Switching Center (MSC). However, these fiber links are, today, used strictly to enable communications between the MSC and each base station, as opposed to enabling transmit/receive cooperation among base stations.

Infrastructure cooperation seeks to better utilize theses links by allowing them to carry complex analog waveforms (or the digital representation thereof) between a central processing/control station and each base station in a small associated set. This central processing and control node computes the signal that must be sent by each base station to optimize some mobile station delivery objective and collectively processes the signals received from all base stations. In effect, each mobile is served by an array of macrodiversity antennas. In the sequel, we may sometimes refer to the cooperating base stations as the "Composite Base Stations" (CBS).

In the forward direction, the signal radiated by each base station antenna is a composite constructed from the individual data streams that are to be sent concurrently to the mobile stations collectively served by the set of base stations. The central node constructs an individual composite signal for each base station antenna such that each mobile then optimally receives its own data in accordance with the chosen objective criterion. In the reverse direction, each base station antenna receives some composite superposition of the signals sent by the mobile stations, and the central processor performs a joint detection of all signals. The focus of this dissertation will be on the downlink.

Physical layer aspects of infrastructure cooperation has been studied in earlier works (see, for example, Shamai and Zaidel [38], Ng et al. [30], Choi and Andrews [8], etc.). From the literature on physical layer aspects of infrastructure cooperation, we know that the cooperative base station approach is superior. Thus, without requiring a great deal of improvements in the infrastructure (the fiber links connecting each base station

with the MSC are assumed to be already in place), the deliverable capacity of a cellular network can be greatly improved.

However, the literature on infrastructure cooperation does not consider network traffic issues such as delay performance, optimal throughput, etc. In this dissertation, we study some of the topics related to the cross-layer analysis of cellular wireless systems enabled with infrastructure cooperation. Here by cross-layer analysis we mean viewing the physical and the network layers as one entity, unlike the OSI model [4, Chapter 1.3] where the two layers are studied separately.

While the modern communications networks have packet-based traffic, one of the key assumptions, and therefore a key limitation, in the studies of base station cooperation has been the assumption that the traffic is stream-type. Therefore, towards the next step in the study of cooperative cellular wireless systems, in this dissertation, we consider a packet-oriented traffic model.[1] To this end, in the sequel, we seek to establish bounds on the performance of the downlink of the cellular wireless systems with packet-based traffic and cooperation among base stations.

To establish the performance bounds, we use results from multiuser information theory. (A good treatment of recent results on multiuser information theory can be found in Goldsmith [15].) We start with the problem of finding the maximum absolute throughput to the mobile stations when the relative needs of each are specified in advance. The next step would be finding a queueing discipline that minimizes the overall average delay, where delay is defined by a suitable metric. However, the coupled nature of the resulting queueing network precludes an answer to this problem. Nonetheless, we propose a simple queueing discipline that achieves the maximum throughput. Again, the coupled nature of the resulting queueing network does not allow analysis of the throughput-maximizing queueing discipline. Therefore, we take an empirical approach to study the performance

[1] A packet model allows us to develop and study interesting interplays between queueing service discipline (a higher layer issue in the OSI model) and physical level channel conditions (a lower layer issue in the OSI model).

of the queueing discipline. In lieu of exact analysis, we propose two different approximations to the performance analysis of the queueing discipline. We first propose a fixed-point approximation, a useful approximation for heavily as well as lightly loaded systems. We next, as a measure of performance of the queueing network operating under this discipline in the heavy traffic, develop a diffusion approximation.

A limitation of multiuser information theory based approach is its computational complexity for real-world systems. As a remedy to this impracticality, we propose a scheduling policy that is suboptimal, but easy to implement, and gives a performance improvement over the traditional operation.

Since we are interested in establishing performance bounds, we assume perfect network synchronization with regard to bit timing and carrier phase at each base station. Moreover, we assume that the channel transfer matrix, representing the attenuation and relative phase shift occurring between the antenna at each base station and the antenna at each mobile station, is known at the central node. We also ignore the effect of time dispersion in the channel. In the sequel, we make some additional assumptions about the channel to facilitate analysis.

Admittedly, these are indeed optimistic assumptions and approximations, but our goal here is to explore the potential benefits of base station cooperation, leaving both the real-world innovations needed to achieve these benefits, and the assessment of performance degradation arising by any real world deviations from these optimistic assumptions, for future.

1.1 Organization of the Thesis

The rest of the dissertation is organized as follows. Chapter 2 investigates the problem of providing a descriptor of maximum throughput of cellular wireless systems with infrastructure cooperation. We begin with a quasi-static system with simplistic as-

sumptions on the channel model. We first develop the queuing model that will be studied throughout the dissertation. Then we show that there is a simple expression for the maximum throughput of this queuing model relating the throughput to the capacity of the underlying channel. We then propose a service policy that achieves the throughput. In both these cases, we are providing results linking the network layer (queueing) to the physical layer (the channel). Since an exact analysis is not possible, we propose an approximation to help in the analysis of the average delay under this policy.

We then propose a policy that is applicable and implementable for large real-sized systems. Through simulations, we show that, under common channel models, this policy doubles the throughput of a cellular system with infrastructure cooperation over one without cooperation among base stations.

A multi-input multi-output (MIMO) downlink system can be seen as a generalization of the downlink of the cellular wireless network with infrastructure cooperation, the object of study in Chapter 2. Moreover, the service policies that are throughput-optimal for the latter are throughput-optimal for the former as well. In Chapter 3, we analyze the performance of such a policy for quasi-static MIMO downlink system in heavy traffic. We begin with the simple case of a two-user system where the operation points are enumerable. In this case, using the results on the Skorokhod problem (Section 3.3.1.1), we show that the diffusion-scaled queuelength process converges in distribution to an SRBM (Theorem 3.3.12) in the two-dimensional positive orthant.

We next analyze the performance of such a policy for an arbitrarily-sized system. As will be seen, in this case there are $2^N - 1$ operation points where N is the number of users. Therefore, the techniques used for the two-user case can not be applied. By using results from the applied probability theory, we are able to show a result analogous to the two-user case. Specifically, Theorem 3.4.6 states that the diffusion-scaled workload process converges in distribution to an SRBM in the N-dimensional positive orthant.

CHAPTER 2

Maximizing Throughput

As mentioned in Chapter 1, our objective in this Chapter is to quantify the maximum throughput of the downlink of a cellular system with cooperation among base stations and to devise schemes that achieve this throughput. We proceed in multiple steps.

We first provide a methodology to quantify the maximum throughput of the downlink of a cellular system with infrastructure cooperation. Under some simplifying assumptions, we show that the maximum throughput of the downlink of a cellular network with cooperating base stations is related to the capacity region of the underlying channel (Section 2.1.3.2).

The next question we answer is whether or not there is a policy that can achieve this maximum throughput. Under the same assumptions as earlier, we propose a policy that achieves this objective (Section 2.1.4). Using an empirical approach, we quantify the advantage, in maximum throughput, of a system with base station cooperation over a traditional cellular system.

Unfortunately, we do not know of numerical or analytic techniques to compute the maximum throughput for even moderately sized systems (say, more than five(5) users). Therefore, we propose a practical, but suboptimal, scheme which gives a higher throughput (Section 2.2). We then quantify, again by means of simulations, the advantage of the proposed scheme over the traditional operation of cellular systems.

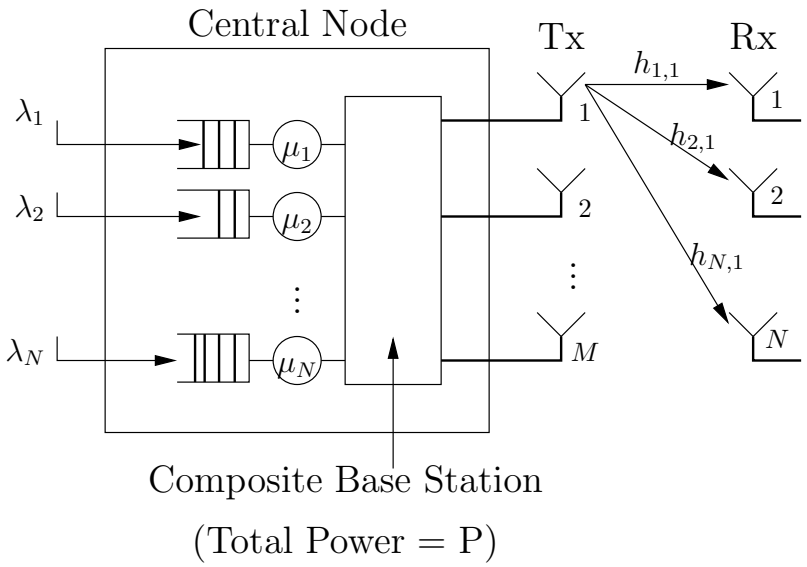


Figure 2.1 *System model: M base stations and N users; each have one antenna.*

2.1 Upper Bound on the Throughput

2.1.1 System Model

The model that we adopt to describe our system of cooperating base stations appears in Figure 2.1. Shown there are M antenna elements, each connected to a central node via a two-way, zero noise, infinite capacity link. Also shown are N non-cooperating mobile stations. The maximum power available to the central node is P. We assume that the central node can distribute this power among the cooperating base station antennas in any proportion it decides.

Arriving at the central node are N packet streams, each containing packets addressed to a single mobile station. Each packet contains a geometrically distributed number of bits with mean value $\mathbf{E}[b]$. The central node contains N queues, each holding the bits awaiting delivery to a particular mobile station. Packet arrivals for the j-th mobile station are characterized by a Poisson process with arrival rate λ_j. If we define

$$k_j \triangleq \frac{\lambda_j}{\lambda_1},\ j = 1, 2, \ldots, N, \tag{2.1}$$

then we can represent the arrival processes by the vector

$$\boldsymbol{\lambda} = \lambda_1 \boldsymbol{k} \tag{2.2}$$

where

$$\boldsymbol{k} \triangleq \{1, k_1, k_2, \ldots, k_N\}. \tag{2.3}$$

Since each packet contains a geometrically distributed number of bits, the remaining service time for a given packet in the j-th queue can be modeled as an exponentially distributed random variable with mean value

$$\tau_j \triangleq \frac{\mathbf{E}[b]}{c_j} \tag{2.4}$$

where c_j is the current rate at which the j-th queue is being emptied (expressed in bits/sec). Then, the instantaneous departure rate from the j-th queue is

$$\mu_j \triangleq \frac{1}{\tau_j} = \frac{c_j}{\mathbf{E}[b]}. \tag{2.5}$$

We represent the current set of N departure rates by a vector

$$\boldsymbol{\mu} \triangleq \{\mu_1, \mu_2, \ldots, \mu_N\}. \tag{2.6}$$

The channel between the base station antennas and the mobile station antennas is described by a channel matrix $H = [h_{j,k}]$ where $h_{j,k}$ represents the amplification or attenuation of the waveform signal $s_j(t)$ originating at base station j as observed at mobile station k, $j = 1, 2, \ldots, M$; $k = 1, 2, \ldots, N$. Added at mobile station k is white Gaussian noise, $n_k(t)$, independent from the noise added at every other mobile station. The spectral height of the noise is assumed to be $N_0/2$ and the system bandwidth is assumed to be W. Let $r_k(t)$ be the composite waveform arriving at mobile station k, $k = 1, 2, \ldots, N$. Then,

$$\boldsymbol{r}(t) = \boldsymbol{H}\boldsymbol{s}(t) + \boldsymbol{N}(t) \tag{2.7}$$

where the received signal vector $\boldsymbol{r}(t) = \{r_1(t), r_2(t), \ldots, r_N(t)\}$, the transmitted signal vector $\boldsymbol{s}(t) = \{s_1(t), s_2(t), \ldots, s_M(t)\}$, and the additive noise $\boldsymbol{n}(t)$ is given by $\boldsymbol{n}(t) = \{n_1(t), n_2(t), \ldots, n_N(t)\}$.

For this section, the elements of $\boldsymbol{H}$ are assumed to be independent complex Gaussian random variables, each with independent real and imaginary parts. Such a matrix corresponds to that produced by a flat multipath Rayleigh fading model. Therefore, at present, we are not considering path-loss and other long-term channel variations. We will introduce long-term channel variations later in this Chapter.

We define the signal-to-noise Ratio (SNR) as follows. Suppose all power P is allocated to the j-th base station for the purpose of communicating only with the k-th mobile station, and suppose further that

$$\mathbf{E}\left[[\mathrm{Re}\,(h_{j,k})]^2 + [\mathrm{Im}\,(h_{j,k})]^2\right] = 1 \tag{2.8}$$

where $\mathrm{Re}\,(\cdot)$ and $\mathrm{Im}\,(\cdot)$ denote real and imaginary parts, respectively. Then, SNR is defined to be the actual SNR observed at mobile station k, that is,

$$\text{SNR} \triangleq \frac{P}{N_0 W}. \tag{2.9}$$

2.1.2 Multiuser MIMO Downlink Capacity Region and its Properties

For a given SNR and channel matrix $\boldsymbol{H}$, the system shown in Figure 2.1 can be characterized by an N-dimensional capacity surface. A representative capacity surface, drawn for $N = 2$, appears in Figure 2.2. Each point on this surface corresponds to an allowable rate pair, that is, an allowable combination of rates (c_1, c_2) at which information can be reliably delivered from the central node to mobile stations 1 and 2, respectively.

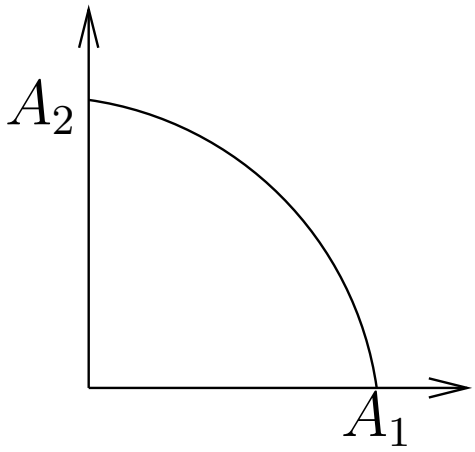


Figure 2.2 *Outline of a representative capacity surface*

For example, if all resources are allocated to the delivery of information to mobile station 1, then, for the channel matrix and SNR underlying the capacity surface of Figure 2.2, information can be delivered to mobile station 1 at rate A_1, with the rate of delivery to mobile 2 set at zero. Similarly, if all resources are allocated to the delivery of information to mobile 2, then information can be delivered to mobile station 2 at rate A_2, with the rate of delivery to mobile station 1 set to zero. In fact, for the channel matrix and SNR underlying the capacity surface of Figure 2.2, resources can be allocated to mobile stations 1 and 2 such that any point on the capacity surface can be achieved.

In fact, the downlink channel depicted in Figure 2.1 has been well-studied in the literature and is known as the multi-user MIMO broadcast channel. The problem of finding the capacity region of such a channel has also been well-studied. Based on the notion of Dirty Paper Coding (DPC) [10], an achievable rate region for this channel was proposed [6], [46], and it has been shown that the DPC region is, in fact, the capacity region of a multi-user MIMO broadcast channel [43]. Furthermore, it has been shown that the capacity region of the MIMO broadcast channel can be written in terms of the capacity region of the dual multiple access channel [20], [42]. Since the capacity region of the dual multiple access channel is easily computable, the capacity region of the primal broadcast channel can be computed by using the duality. Thus, as shown in Figure 2.3, the two-

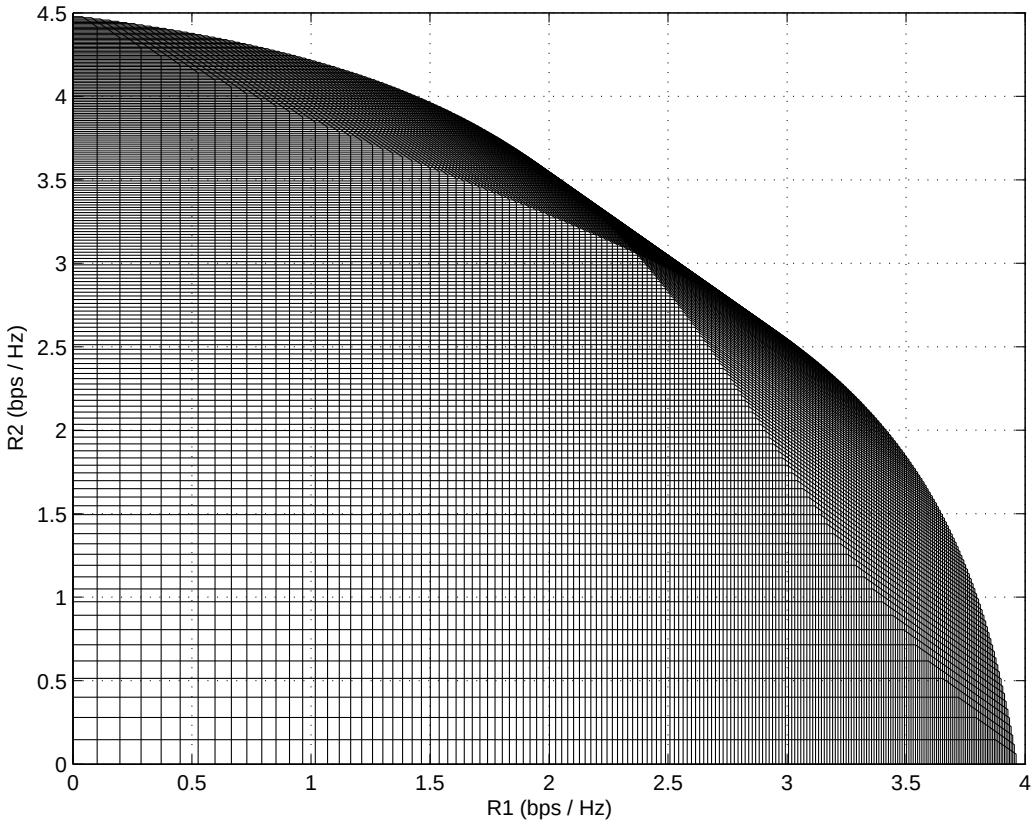


Figure 2.3 *Two-user MIMO broadcast channel Capacity region is computed using duality.*

dimensional broadcast channel capacity region is obtained as the convex hull of the dual multiple access channel regions over the set of power allocations such that the sum of allocated powers is the same as the broadcast power. Note that the number of receive antennas per user can be more than one without changing the procedure of determining the capacity region or affecting the properties of the capacity region.

The capacity region of a multi-user MIMO broadcast channel, $\mathcal{C}$, corresponding to a N-user MIMO downlink system with fixed channel $\boldsymbol{H}$ and total transmit power P has the following properties:

(i) $\mathcal{C}$ is a non-empty connected closed subset of the N-dimensional nonnegative orthant,

(ii) $\mathcal{C}$ is convex,

(iii) $\mathcal{C}$ has $(N+1)$ boundary pieces of which N are the intersections of $\mathcal{C}$ with planes passing through the origin on which exactly one of the N components is zero.

We call the $(N+1)$-th boundary of $\mathcal{C}$ the *capacity surface*. All of the components of an outward pointing unit normal at each point of the capacity surface are nonnegative. Note that there might not be a unique outward pointing normal but the abovementioned property holds for each normal. Moreover, the capacity surface has a functional form

$$f(c_1, c_2, \ldots, c_N) = k \tag{2.10}$$

where, from property (ii), $f(\cdot)$ is a convex function of N variables and k is a constant. We do not make any other assumptions about the function f.

Associated with a convex capacity region is the concept of *differentiated service capacity* (DSC). For every convex capacity region, there is a unique point where the capacity surface is intersected by a ray from the origin of slope $\boldsymbol{k} = (1, k_2, k_3, \ldots, k_N)$. We call this point the DSC for the given $\boldsymbol{k}$. In Figure 2.4, ray L_1 is the ray from the origin

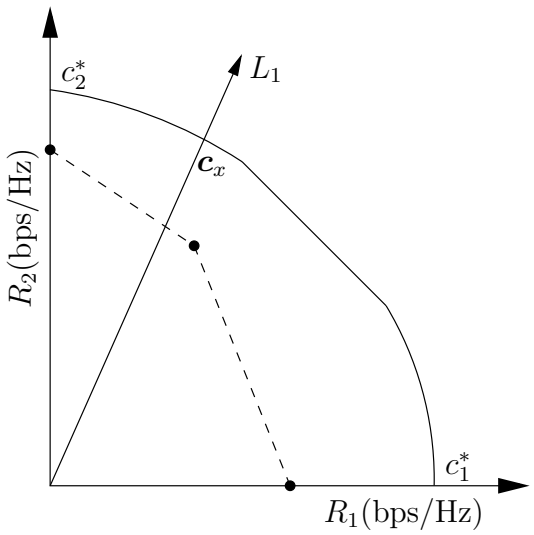


Figure 2.4 *A representative 2-user MIMO BC capacity region. The point $\boldsymbol{c}_x$ is the DSC for $\boldsymbol{k} = (1, 2)$. The capacity surface for traditional operation is indicated by the dashed lines.*

with slope $R_2 = 2R_1$. It intersects the capacity surface at the point $\boldsymbol{c}_x$. Thus, $\boldsymbol{c}_x$ is the DSC for the vector $\boldsymbol{k} = (1, 2)$. Note that symmetric capacity [34] is a special case of the differentiated service capacity with $\boldsymbol{k} = (1, 1, \ldots, 1)$.

2.1.3 Queueing Network and Maximum Throughput

In this section, we develop and study the queueing model for the system described in Sec. 2.1.1. We establish a limit on the maximum throughput that can be supported while keeping the system stable. We then propose a service policy that is throughput-optimal where a service policy is throughput-optimal if the policy results in a stable queueing system for all allowed loads.

In this section, we only consider quasi-static fading, that is, we assume that the channel matrix $\boldsymbol{H}$ is fixed for the period of interest. Note that this reduces the fading MIMO broadcast channel to a Gaussian MIMO broadcast channel.

2.1.3.1 Queueing System

As explained in Section 2.1.1, by modeling the exogenous traffic as a packet-based traffic, instead of stream-based traffic, we can consider the N flows, one for each user, at the transmit end as N queues, one for each user. The arrival process for each queue is an independent Poisson process where the average arrival rate vector is given by (2.2),

$$\boldsymbol{\lambda} = \lambda_1 \boldsymbol{k}. \tag{2.11}$$

Packets for each user are stored in their order of arrival. We assume that all the queues have infinite buffer so that no packets are lost. When service is given to a queue, it goes to the packet at the front, that is, we only consider first-in-first-out (FIFO) service disciplines. Each queue is served by a single server with average service rate vector at time t, $\boldsymbol{\mu}(t)$, given by

$$\boldsymbol{\mu}(t) = (\mu_1(t), \mu_2(t), \ldots, \mu_N(t)). \tag{2.12}$$

The service rate vector is related to the transmission rate (in bps) by (2.5), but at any given time there are infinitely many possible combinations of transmission rates, and thus, infinitely many possible ways of choosing the service rates. Thus, we restrict our attention to work-conserving policies. A policy is *work-conserving* if it serves at full rate whenever data is present in any of the queues. For our queueing system, this would correspond to choosing $\boldsymbol{c}$ on the capacity surface only. Then from (2.5) and (2.10), at all times the components of $\boldsymbol{\mu}(t)$ must satisfy a functional relationship of the form

$$g(\mu_1, \mu_2, \ldots, \mu_N) = k \tag{2.13}$$

where $g(\cdot)$, again, is a convex function because $g(\cdot)$ is an affine transform of $f(\cdot)$ and convexity is affine-transformation invariant.

Since in this queueing system, the packet arrival and service rates are different for

different users, each user can be seen as a separate class of a multi-class queueing system, thereby making this queueing system a multi-class queueing system. Furthermore, the instantaneous (and average) service rate of a queue depends on the state (service rate and the number of packets in the buffer) of every other queue in the network. Thus, we are considering a system of coupled queues.

Such queueing systems are not easy to study. Even the simple case of two coupled queues has many open questions. One case which has been studied involves a coupled two-queue system where the arrival processes to the two queues are two independent Poisson streams of equal rate and the queues are served in the order-of-arrival by a single server [25]. For such a system, an explicit expression for joint probability generating function was found in terms of an elliptic function but the probability density of the stationary distribution was not expressed in closed-form. For a coupled two-queue system where the two service rates are different and the processor resources are shared when a queue is inactive, the diffusion approximation was considered to analyze the performance in heavy traffic [24].

2.1.3.2 Maximum Stable Throughput

In this subsection we show that the maximum throughput that our queueing system can support without becoming unstable is equal to the DSC of the corresponding capacity region for vector $\boldsymbol{k}$.

We say that a queue is *stable* if the time-average delay is bounded. This leads to a natural definition for the stability of a queueing system. A *queueing system* is *stable* if each individual queue in the queueing system is stable. Consequently, we can define the *maximum stable throughput* (MST) of a queueing system to be the supremum over all arrival rate vectors for which the queueing system remains stable. (Note that one has to work with the supremum, rather than maximum, in defining MST, because it is *a priori* unknown whether the system is stable at the MST or not.) For our system, since the

average arrival rates are proportional, as given by (2.2), MST can be equivalently defined as the supremum of all α such that the system remains stable when the average arrival rate is less than or equal to $\alpha\boldsymbol{k}$.

A service policy is a function that maps the current system state to a set of service rate vectors $\{\boldsymbol{\mu}^l\}_{l=1}^{N_S}$ where N_S is the number of operating points in the policy. We say that a service policy is a *viable policy* if it satisfies the following conditions:

(i) If a queue is empty, it is not served. That is, if the number of packets in the i-th queue is zero, then the associated service rate for the i-th queue is zero.

(ii) There exists $\beta > 0$ such that for all loads $\boldsymbol{\lambda} = \alpha\boldsymbol{k} < \beta\boldsymbol{k}$, the queueing system has a steady state.

The first condition can be thought of as a "sanity-check condition" in the sense that a queue is not served if there are no packets waiting. Consequently, the zero vector, $\mathbf{0} = (0, 0, \ldots, 0)$, must be one of the operating points for a viable policy. The second condition allows us to define a viable policy without being concerned with the system load. An example of a non-viable policy for a two-user system would be a policy that always serves one of the two users, and never serves the other user.

The following lemma states the relationship between the DSC for the vector $\boldsymbol{k}$ and the MST of our queueing system.

Lemma 2.1.1. *For the queueing system of interest, if the average arrival rates are related by $\boldsymbol{\lambda} = \lambda_1\boldsymbol{k}$, the maximum stable throughput is the differentiated service capacity for the corresponding capacity region.*

Remark. This lemma only requires that the capacity region be convex, a property that all capacity regions have due to the convex hull operation.

Proof. For our queueing system, the MST is the supremum over all viable policies of the load supported by a viable policy. Consider a viable policy $\boldsymbol{\Pi} = \{\boldsymbol{\mu}^l\}_{l=1}^{N_S}$; from the

second condition, there is a load for which a steady-state can be defined. Let $\{f^l\}_{l=0}^{N_S}$ be the steady-state probability distribution, where f^l, $l = 1, 2, \ldots, N_S$, is the probability of serving at rate $\boldsymbol{\mu}^l$ in the steady-state and f^0 is the probability of serving at zero rate in the steady-state. Since $\{f^l\}_{l=0}^{N_S}$ is a probability distribution, $f^l \geq 0$, $l = 0, 1, \ldots, N_S$ and $\sum_{l=0}^{N_S} f^l = 1$. In the steady-state, the average service rate must be equal to the average arrival rate, that is, $\sum_{l=1}^{N_S} f^l \boldsymbol{\mu}^l = \boldsymbol{\lambda}$ where the vector equality is component-wise (the term corresponding to f^0 is $\mathbf{0}$). It follows from the definition of the DSC that the MST for the queueing system is the DSC for the given $\boldsymbol{k}$. □

Here we have used the fact that for a given vector $\boldsymbol{k}$, the DSC is unique for a convex capacity region. Note that strict convexity is not required for the uniqueness of DSC and the uniqueness holds even when some of the entries of the vector $\boldsymbol{k}$ are zero. An interesting problem is that of efficiently computing the DSC for a multi-user MIMO broadcast channel. To this end, Lee and Jindal [26] have proposed an algorithm to efficiently calculate the DSC (and symmetric capacity) for a multi-user MIMO broadcast channel for a fixed channel matrix, $\boldsymbol{H}$, and total transmit power P, but even their algorithm is not very effective for more than five (5) users.

2.1.4 Throughput Maximizing Service Policy

Based on Lemma 2.1.1, we consider the following service policy. If any queue is empty, set the service rate for that queue at zero. If only queue j is non-empty, $j = 1, 2, \ldots, N$, operate at the point $\boldsymbol{c} = (0, 0, \ldots, c_j, 0, 0, \ldots, 0)$ where c_j is the rate at which the j-th queue would be served if all resources were allocated to the delivery of information to the j-th mobile station. If only two queues j and l are non-empty, operate at $\boldsymbol{c} = (0, 0, \ldots, 0, c_j, 0, \ldots, c_l, 0, \ldots, 0)$ where $c_j/c_l = k_j/k_l$ and (c_j, c_l) is a point on the two-dimensional capacity surface produced when all resources are allocated to delivering information only to the users j and l. This two dimensional capacity surface corresponds to the intersection of the N-dimensional capacity surface with the plane $\{c_i = 0, i =$

$1, 2, \ldots, N, i \neq j, i \neq l\}$. Similarly, if only d out of N queues are non-empty, choose the service rate vector corresponding to the DSC for the capacity surface produced by allocating resources to serve only the d users with non-empty queues (set the $N - d$ corresponding entries of $\boldsymbol{k}$ to zero while computing the DSC). If none of the queues is empty, choose the DSC for the vector $\boldsymbol{k}$ as the operating point.

It is easy to see that for all loads below the MST, the system remains stable under this policy and therefore, this policy is throughput-optimal. The stability of the system can be seen from the fact that at all times when there are packets in a queue, the instantaneous average service rate for that queue is greater than the average arrival rate. In fact, it can be shown that a quadratic Lyapunov function [28] of the form $\sum_{i=1}^{N} Q_i^2(t)$, where $Q_i(t)$ is the number of packets in the i-th queue at time t, has a negative drift at all times.

This policy is an on-off policy where the instantaneous service rate depends only on the empty/non-empty status of each of the N queues. For a N-user system, there are $2^N - 1$ service points, each corresponding to one of the 2^N possible empty/non-empty combinations of the N queues. The case where all queues are empty is irrelevant insofar as a service point is concerned.

As an example, consider a two-user system where the underlying capacity region is given by Figure 2.4. For a two-user system, there are $(2^2 - 1) = 3$ operating points. The proposed policy will serve at $(c_1^*, 0)$ when the second queue is empty and at $(0, c_2^*)$ when the first queue is empty. If $\lambda_2 = 2\lambda_1$, this policy will serve with rate $\boldsymbol{c}_x = (c_1, c_2)$ whenever both the queues are non-empty.

2.1.5 Fixed-point Approximation

As mentioned in Section 2.1.3, the queueing system of interest is a coupled queueing system with time-varying service rates which depend on the instantaneous state of the queues. Such systems are not very amenable to analysis; in fact, we are not even aware of any work on the exact analysis of such systems. (In Chapter 3, we analyze the perfor-

mance of this policy in heavy traffic, an approximation that gives certain insights.)

Since we do not have an expression for the system average delay, we develop a fixed-point model for this policy which gives an approximation to the average delay. Fixed-point models for queueing systems are well known in the literature (see, e.g., Kelly [22] and the references therein). We first illustrate the methodology of FPA by developing the fixed-point model for a two-user (and thus, two-queue) system in Section 2.1.5.1. We then extend the method to an arbitrary number of users in Section 2.1.5.2.

2.1.5.1 Two-User System

The fixed-point approximation for the coupled two-queue system replaces the M/G/2 coupled queueing system by two M/M/1 queues where for the approximated system, the arrival processes are the same as that of the coupled system but the coupled servers are replaced by two independent servers. Note that, as in the original system, the arrival processes are assumed to be independent. The service process of each server in the approximated system is a Poisson process with constant average service rate which is a function of the service rates of the original system and the coupling between the servers in the original system.

For $i = 1, 2$, let Q_i denote the number of packets in the i-th queue. Then by the M/M/1 assumption of the approximated system, the probability that there are (q_1, q_2) packets in the approximated system is given by

$$\begin{aligned} \Pr\{Q_1 = q_1\} &= \left(1 - \frac{\lambda_1}{\mu_1^e}\right)\left(\frac{\lambda_1}{\mu_1^e}\right)^{q_1}, \\ \Pr\{Q_2 = q_2\} &= \left(1 - \frac{\lambda_2}{\mu_2^e}\right)\left(\frac{\lambda_2}{\mu_2^e}\right)^{q_2} \end{aligned} \tag{2.14}$$

where $\boldsymbol{\mu}^e \triangleq (\mu_1^e, \mu_2^e)$, the effective service rate vector for the approximated system, is

defined as

$$\begin{aligned}\mu_1^e &\triangleq \mu_1^* \Pr\{Q_1 = 0\} + \mu_1 \Pr\{Q_1 \neq 0\}, \\ \mu_2^e &\triangleq \mu_2^* \Pr\{Q_2 = 0\} + \mu_2 \Pr\{Q_2 \neq 0\}\end{aligned} \tag{2.15}$$

where from (2.5),

$$\begin{aligned}\mu_1^* &= \frac{c_1^*}{\mathbf{E}[b]}, \quad \mu_1 = \frac{c_1}{\mathbf{E}[b]}, \\ \mu_2^* &= \frac{c_2^*}{\mathbf{E}[b]}, \quad \mu_2 = \frac{c_2}{\mathbf{E}[b]}.\end{aligned} \tag{2.16}$$

Since we do not have an exact expression for the probability of either of the queues being empty, we approximate the probability of a queue being empty by using the expression for M/M/1 queues. As a result, the average service rates of the approximated system need not be the same as that of the coupled queueing system. If the average service rates are the same, then the FPA gives a lower bound on the average delay [29, Appendix C]. Furthermore, in that case, the bound becomes tight when the speed of service rate variation goes to zero or infinity corresponding, in our system, to high-load and low-load operation, respectively. Although the theorem of [29] is not applicable to our system (the average service rates of the actual and approximated systems need not be the same), our simulation results show that, for the parameters considered, the FPA produces a lower bound on the average delay.

For $i = 1, 2$, define

$$\theta_i \triangleq \Pr\{q_i = 0\}. \tag{2.17}$$

Then from (2.14),

$$\begin{aligned}\theta_1 &= \Pr\{q_1 = 0\} = 1 - \frac{\lambda_1}{\mu_1^e}, \\ \theta_2 &= \Pr\{q_2 = 0\} = 1 - \frac{\lambda_2}{\mu_2^e}.\end{aligned} \tag{2.18}$$

Substituting (2.18) in (2.15) and rearranging the terms, we obtain:

$$\begin{aligned} \mu_1^e \mu_2^e &= \mu_1^* \mu_2^e + \lambda_2(\mu_1 - \mu_1^*), \\ \mu_1^e \mu_2^e &= \mu_2^* \mu_1^e + \lambda_1(\mu_2 - \mu_2^*), \end{aligned} \tag{2.19}$$

which can be solved for $\boldsymbol{\mu}^e$ by iterative methods. Note that $\boldsymbol{\mu}^e$ is a function of the arrival and service rates of the original system. Thus, the approximated service rate will change when any of the system parameters is changed but will scale proportionally when all parameters in the original system are scaled proportionally. We thus have two M/M/1 queues with average arrival rates λ_i, $i = 1, 2$ and average service rates μ_i^e, $i = 1, 2$. Then, we can compute the average delay using the standard results from queueing theory.

2.1.5.2 FPA for N-User system

We next develop the fixed-point model for a N-user system operating under the policy proposed in Sec. 2.1.4. We represent the proposed service policy in matrix form as

$$\mathbf{\Pi} \triangleq [\boldsymbol{\mu}^l]_{l=1}^{2^N-1} \tag{2.20}$$

where the matrix $\mathbf{\Pi}$ has $2^N - 1$ columns, each corresponding to one of the the $2^N - 1$ empty/non-empty states of the N queues. Furthermore, one can write a queue-state vector

$$\boldsymbol{s}^l \triangleq (s_1^l, s_2^l, \ldots, s_N^l) \tag{2.21}$$

where $s_i^l = 0$ if the i-th queue is empty and $s_i^l = 1$ if the i-th queue is non-empty. For example, $\boldsymbol{\mu}^1 = (\mu_1^1, 0, 0, \ldots, 0)$ corresponds to only the first queue being non-empty, that is, $\boldsymbol{s}^1 = (1, 0, 0, \ldots, 0)$ while $\boldsymbol{\mu}^2 = (0, \mu_2^2, 0, 0, \ldots, 0)$ corresponds to only the second queue being non-empty, that is, $\boldsymbol{s}^2 = (0, 1, 0, \ldots, 0)$. As with the two-user case, for $i = 1, 2, \ldots, N$, define

$$\theta_i \triangleq \Pr\{Q_i = 0\} \tag{2.22}$$

where Q_i is the number of packets in the i-th queue. Then, by the M/M/1 assumption of the approximated system,

$$\theta_i = 1 - \frac{\lambda_i}{\mu_i^e}, i = 1, 2, \ldots, N, \tag{2.23}$$

where

$$\boldsymbol{\mu}^e \triangleq (\mu_1^e, \mu_2^e, \ldots, \mu_N^e) \tag{2.24}$$

is the approximated service rate vector. Also, for $i = 1, 2, \ldots, N$, define

$$\bar{\theta}_i \triangleq \Pr\{Q_i \neq 0\} = (1 - \theta_i). \tag{2.25}$$

Then, generalizing (2.15), for $i = 1, 2, \ldots, N$, we obtain:

$$\mu_i^e = \sum_{l=1:\mu_i^l \neq 0}^{2^N - 1} p_i^l \mu_i^l \tag{2.26}$$

where μ_i^l is the service rate of the i-th queue when the point of operation is the vector $\boldsymbol{\mu}^l$, and p_i^l is the probability that the packets for user i are served at the rate μ_i^l. Note that for a fixed l, p_i^l need not be the same for all i. For the service policy of interest, due to the M/M/1 assumption of each queue in the approximated system, if the queue-state vector is $\boldsymbol{s}^l$, then

$$p_i^l = \prod_{\substack{k=1 \\ k:s_k^l=0 \\ k \neq i}}^{N} \theta_k \prod_{\substack{k=1 \\ k:s_k^l=1 \\ k \neq i}}^{N} (1 - \theta_k) \tag{2.27}$$

where the first product is over the set of queues that are empty while the second product is over the set of queues that are non-empty. Note that, $\sum_{l=1}^{2^N-1} p_i^l > 1$ but $\sum_{l:\mu_i^l \neq 0} p_i^l = 1$ which corresponds to the fact that in computing the approximated service rate, the fraction of time when a queue is not served should not be considered.

The nonlinear equations (2.23)–(2.27) can be numerically solved for $\boldsymbol{\mu}^e$ using

non-linear optimization techniques such as Gauss-Newton method. Again, the approximated service rate is a function of each arrival and service rate in the original system and scales proportionally when all rates in the original system are scaled proportionally.

2.1.6 Simulation Results

We now present some numerical results demonstrating the advantage of base station cooperation over the traditional operation. Before presenting our results, we first explain what we mean by the traditional operation of a cellular network. For meaningful comparisons, we assume that the total power with the traditional operation is the same as that for the cooperative system. Moreover, we assume that in the traditional operation the total power is equally divided among all base stations.

Under traditional operation, each mobile is assigned to the base station with the strongest signal strength. When more than one mobiles are assigned to the same base station, the base station is time-shared in a way that the traffic to the mobiles satisfies the constraint on the arrival rates given by (2.2). Then, for a given configuration of active users, the rates at which data can be transferred to different users are fixed.

We next illustrate the computation of MST for traditional operation. To keep the exposition simple, we consider a system with two base stations and two users. For this system, representative achievable rate region for traditional and cooperative operation are illustrated in Figure 2.4, where the dashed line is the achievable rate region for traditional operation. A queueing model similar to that for cooperative operation can be developed for traditional operation, with the only difference being that the service rates are fixed for the traditional operation once a particular mobile assignment to base stations is chosen. As shown in Figure 2.4, the MST is given by the intersection of (1) the ray through origin with slope $\boldsymbol{k}$, and (2) the convex hull of the operation points. This can be computed by exploiting the geometric properties of the achievable rate region.

We next define the system parameters for the simulations for the general case of N

users. In the following, we normalize the system bandwidth to $W = 1$Hz. Furthermore, we define the system load as

$$\lambda_s \triangleq \lambda_1 \sum_{i=1}^{N} k_i = \sum_{i=1}^{N} \lambda_i \tag{2.28}$$

and the average system delay as

$$\bar{D} \triangleq \frac{\sum_{i=1}^{N} k_i D_i}{\sum_{i=1}^{N} k_i} \tag{2.29}$$

where D_i is the average delay for the i-th user, $i = 1, 2, \ldots, N$. This corresponds to weighing the user average delays proportional to the amount of data to be transmitted to the users. In all our simulations, we have set the mean packet size to 100 bits.

2.1.6.1 Quasi-static Systems

We now present the simulation results for traditional and cooperative operation when the channel realization is frozen, that is, the elements of the $\boldsymbol{H}$ matrix are randomly chosen but fixed. Figure 2.5 shows the gain in throughput achieved by base station cooperation for systems with two base stations and two mobiles, plotted as a function of SNR. It can be seen that for a fixed traffic vector, the throughput gain increases with SNR. Though, not true in general, for this particular channel realization, the throughput gain is higher when the relative traffic vector is symmetric. For reasonable SNRs (between 10 and 20dB), the gain is approximately 40%.

We next plot the average system delay against the system load. In Figure 2.6, we have plotted the average system delay against the system load for a two base station, two user cellular network for two different channel matrices $\boldsymbol{H}$. Thought the plots in Figure 2.6 are for specific channel realizations, they are, qualitatively, representative of the family of channel realizations. As shown in the plots, the average system delay is drawn for each of several relative traffic vectors. It can be seen that cooperation leads

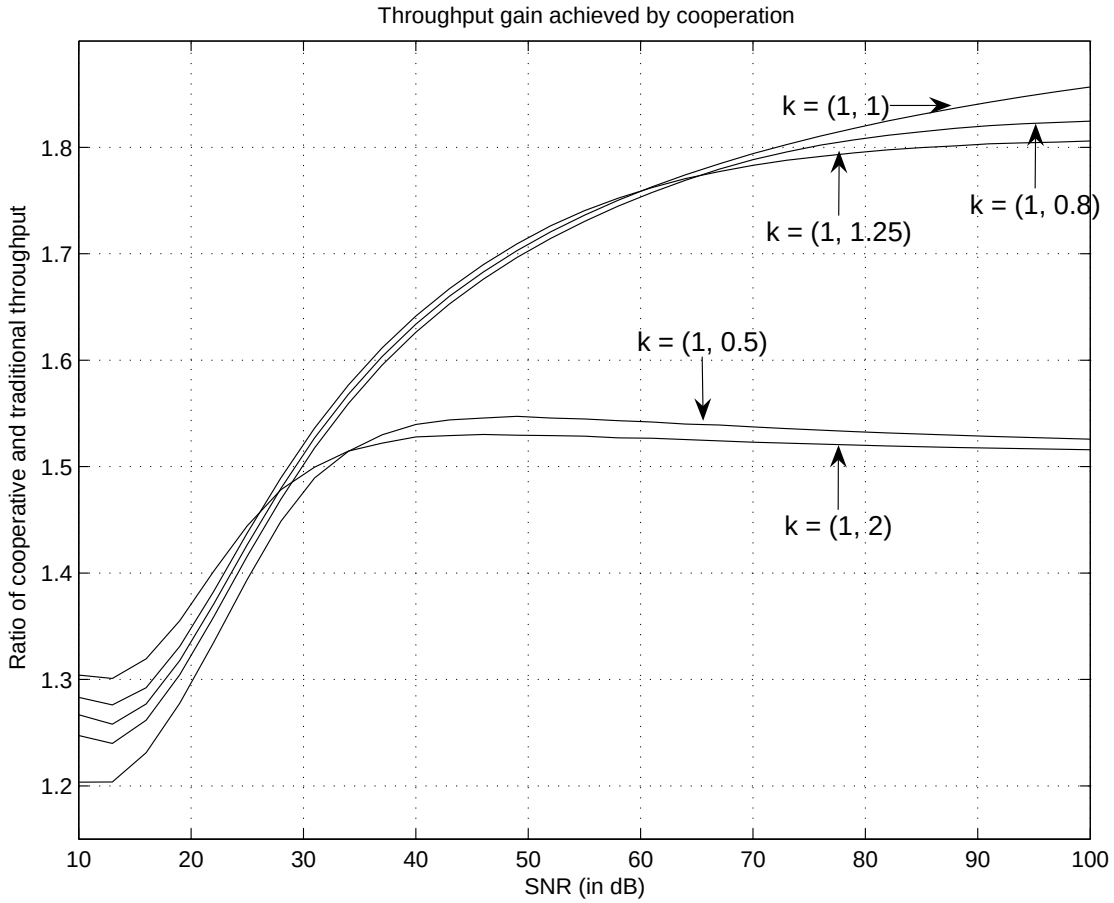


Figure 2.5 *Throughput gain achieved by cooperation for a two user system with two base stations: Channel is fixed.*

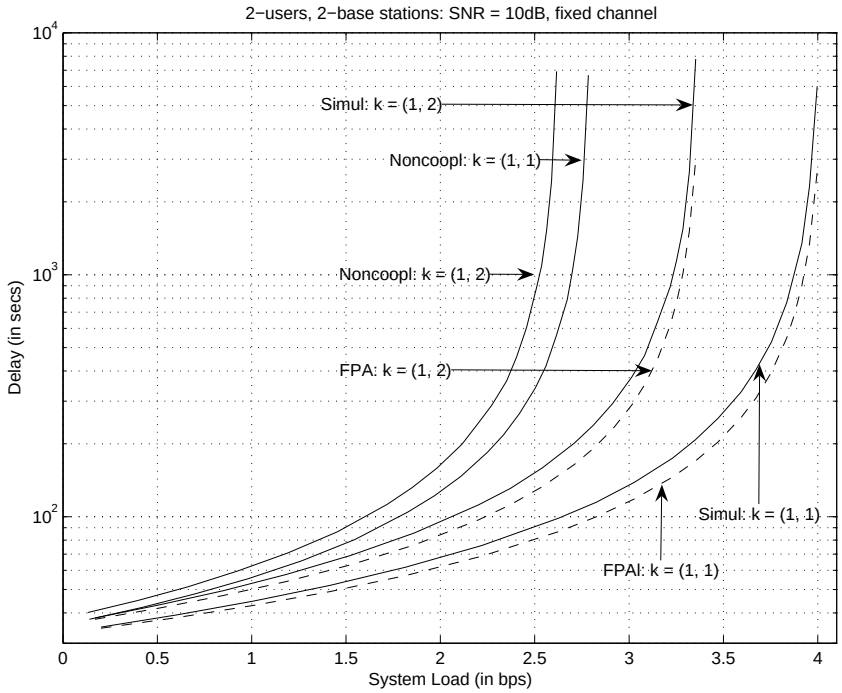
to a higher throughput, as expected, but that the throughput gain is not the same for different channel realizations. In the plots, the maximum throughput can be observed from the graph of load against delay by looking at the limit of loads as the system delay asymptotically approaches infinity (corresponding to the system approaching instability). We also see that the fixed-point approximation provides a very good approximation of the system delay.

In Figures 2.7 and 2.8, we plot the average system delay against the system load for both the three base station/three user case and the four base stations/four user case. Again, this is shown for two different channel realizations and several representative relative traffic vectors. Again, we conclude that cooperation leads to lower average delay and substantially higher system throughput, and that the fixed-point approximation gives a very good approximation to system delay. Thus, for large systems which may be difficult to simulate, we conclude that the fixed-point approximation may be safely applied, especially for the important high load regime since the maximum throughput predicted by FPA and the actual system are the same.

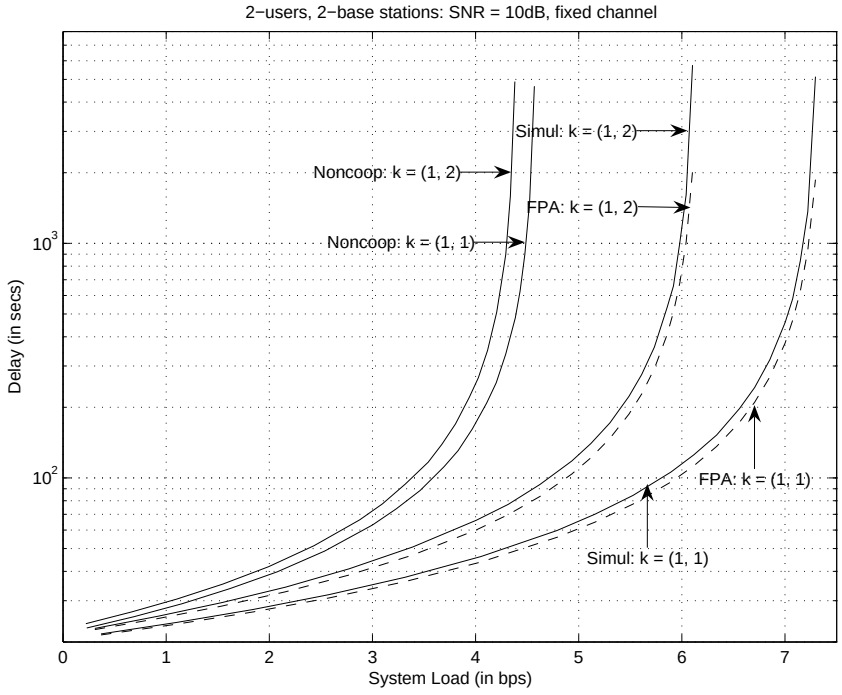
We note that the gain in maximum throughput shown in Figure 2.7 is between 20% and 70% depending on traffic vector and channel. For Figure 2.8, the gain is approximately a factor for 3 for all cases. We attribute this "gain stabilization" to the greater diversity offered by the fours base station system of Figure 2.8, and expect that a higher, more predictable, gain would be offered as the number of base station increases.

2.1.6.2 Outage Results

We next present results showing the outage probability for different system configurations. For a fixed channel realization and a relative traffic vector $\boldsymbol{k}$, we say that an *outage* has occurred if the MST of the system for that channel realization is less than the load. In our simulations, for each system configuration, we considered $10,000$ or more channel realizations. For each channel realization, we first compute the MST for cooper-

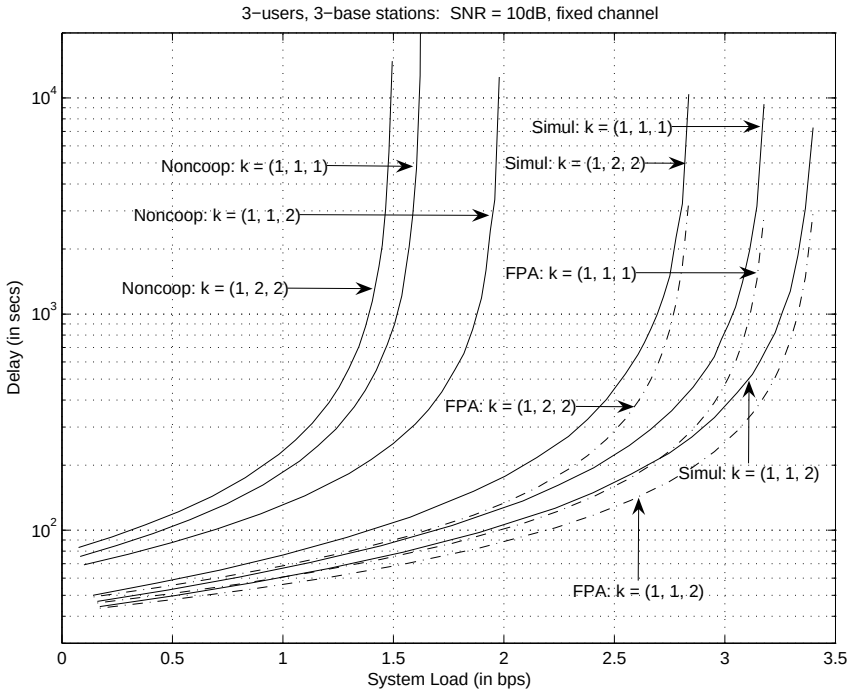


(a) *Set 1*

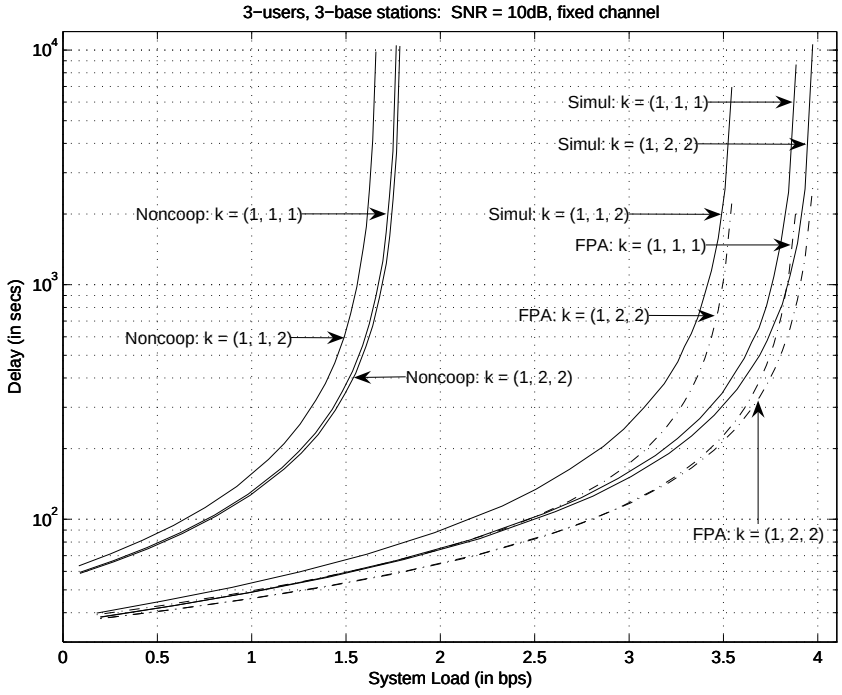


(b) *Set 2*

Figure 2.6 *Simulation results for a 2-user system: Channel is fixed in each case.*

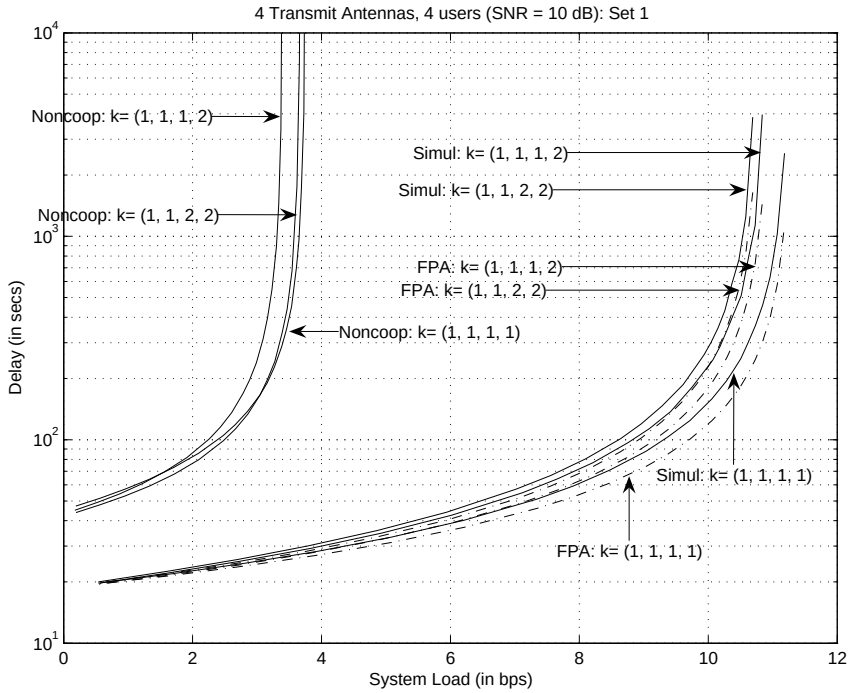


(a) *Set 1*

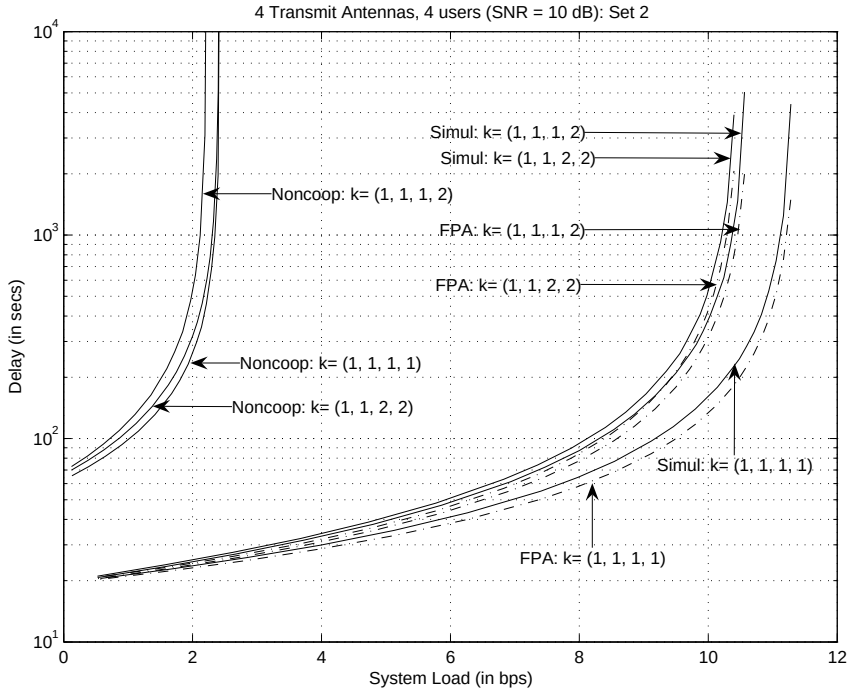


(b) *Set 2*

Figure 2.7 *Simulation results for a 3-user system: Channel is fixed in each case.*



(a) *Set 1*



(b) *Set 2*

Figure 2.8 *Simulation results for a 4-user system: Channel is fixed in each case.*

ative and traditional operation. Then the MST can be used to compute the maximum load that can be supported at a given outage probability.

In Figure 2.9, we plot the probability of outage against the offered load for different system configurations. We note that for both cooperative and traditional operation, the increase in throughput as the probability of outage increases is not very significant. For example, consider the system with three base stations and three users, wherein the throughput for cooperative operation increases by about 33% for a tenfold increase in outage probability (from 1% to 10%), and the throughput for traditional operation increases by about 50%. Similar observations hold for other system configurations. Moreover, the gain in throughput is more pronounced at low outage. To show this more clearly, for different system configurations, we plot in Figure 2.10 the throughput gain achieved by cooperation at the base station against the outage probability. As can be seen, the throughput gain decreases as the outage probability increases except for the 2×2 system where it is almost constant.

Another quantity of interest for quasi-static systems is the maximum number of users that can be supported for a given load and outage probability when the number of base stations is fixed. To this end, in Figure 2.11 we plot the maximum load for a specific user that can be supported by a two base station system with a fixed power (and thus, by our definition, fixed SNR) for different outage probabilities. (Also shown by dashed lines is the system load.) As expected, at low outage the maximum load that can be supported per user is low but it does not decrease by much when the number of users is increased; on the other hand, at high outage the maximum load per user decreases rapidly with an increase in the number of users.

2.2 A Throughput-Increasing Scheme for Large Systems

As mentioned in Section 2.1.2, computing the DPC region is computationally expensive and almost infeasible for even moderately-sized systems - in fact, the runtime to

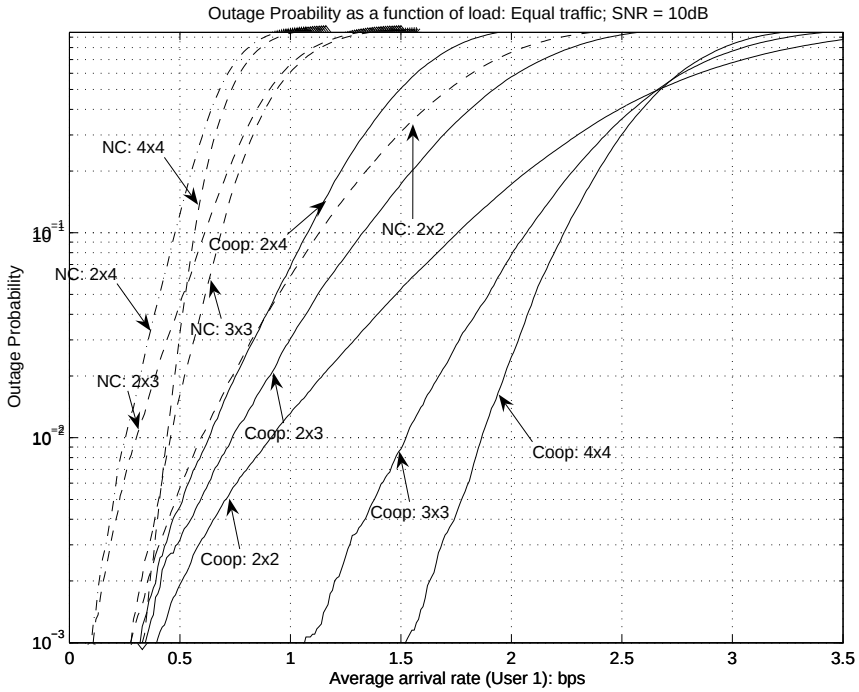


Figure 2.9 *Probability of outage for a given load is plotted for different system configurations with* $\boldsymbol{k} = \mathbf{1}$*. Average arrival rate is the throughput at a given outage probability.*

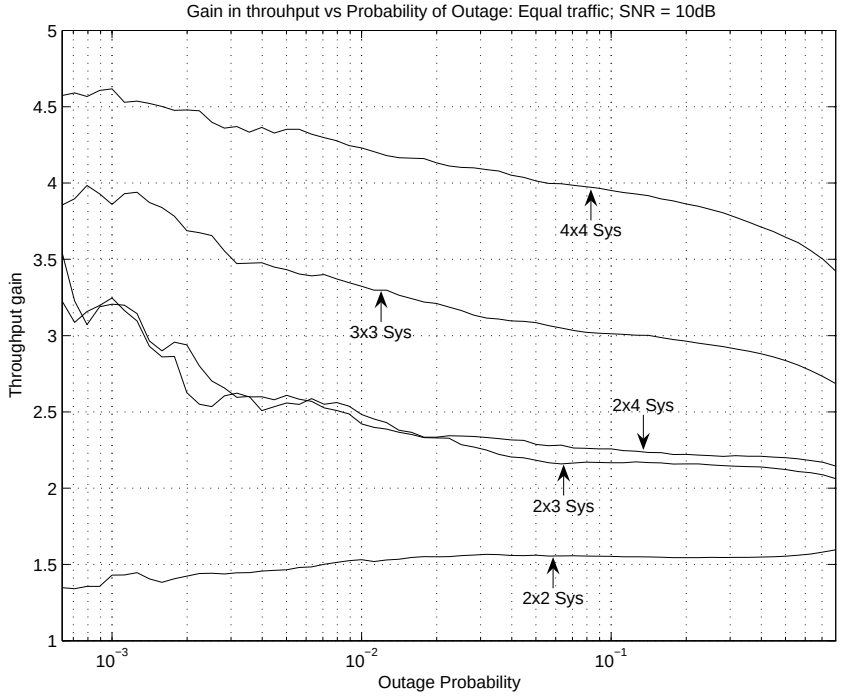


Figure 2.10 *Throughput gain is plotted against the probability of outage for different system configurations.*

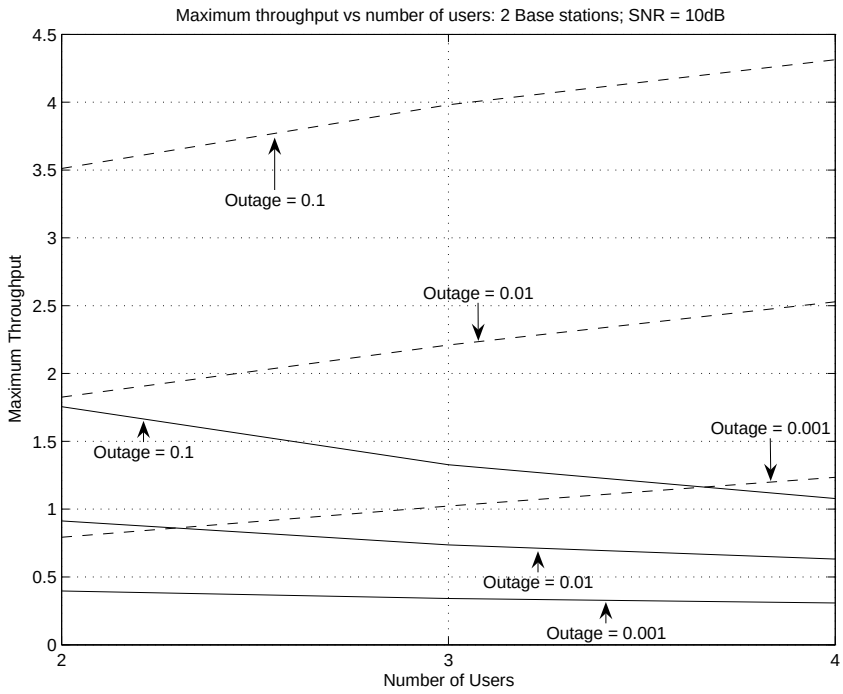


Figure 2.11 *System load and user load is plotted against the number of users where the number of base stations is fixed. Dashed lines indicate the system load.*

compute the capacity region for a system with five (5) users was in days - while real-world systems have hundreds of users. Moreover, the number of operation points increases as $2^N - 1$ where N is the number of users in the system. In light of these reasons, applying the policy proposed in Section 2.1.2 is not feasible for real-world systems. Therefore, in this Section, we propose a practicable modification of that policy which, thought suboptimal, has a higher throughput than the traditional operation. To arrive at the modified policy, we build upon several observations on the properties of the MIMO broadcast channel capacity region and the queueing network of the Section 2.1.

The first observation is that the gain in throughput is due to the fact that the capacity of MIMO channel is higher than that of the single-input single-ousput (SISO) channel. But for a MIMO channel with N_R transmit and N_T receive antennas, the gain in capacity is $\min(N_R, N_T)$. In the downlink of cellular system with cooperation, base stations act as the transmit end and, since, the base station are far fewer than the mobiles, the gain

in throughput that our policy can achieve is severely limited by the number of base station antennas in a cooperating base station set. Therefore, even if there was a method to compute the DPC region, the advantage of cooperation would be somewhat limited. (It should be noted that the base stations can have multiple antennas and therefore, it is not correct to say that the possible gain in throughput, were such a practical method available, is insignificant.)

Our next observation is based upon the conclusion of Lee and Jindal [26]. They observed that the symmetric capacity of a MIMO broadcast channel is higher when the channel is symmetric. Since the gain in throughput is primarily due to higher symmetric capacity, if possible, it is preferable for the corresponding MIMO broadcast channel to be symmetric. Unfortunately, the conclusion of [26] does not have a logical equivalent for the differentiated service capacity.

Our next observation is based upon a property of the DPC region. In a MIMO broadcast channel, the sum of service rates is higher when more users are being served. For example, consider a system with 4 transmit antennas and 4 users. In this system, the capacity when all 4 users are being served will be higher than when any 3 (or fewer) users are being served. A precise mathematical statement of this property can be seen in (3.56) and (3.57). As a consequence of this property, any policy should strive to serve the users that have data to transmit rather than serving an arbitrary subset of users who may or may not have data to transmit.

We are now ready to describe the modified policy. The key idea is to group the users in small subgroups and use the service policy from Section 2.1.4 for each subgroup. The important step in this procedure is dividing the users in subgroups. As an example of import of grouping the users, consider a policy that randomly groups users into small subgroups of, say, size four (4). This policy will be fairly easy to implement. But random grouping of users often results in asymmetric channels - sometimes highly asymmetric. As a consequence of the second observation, this will result in a much lower than expected

symmetric capacity. In fact, in our simulations, sometimes the performance of a random grouping policy was not much of an improvement over traditional operation. Moreover, the variance in performance of random grouping was high enough to make the policy unreliable for operation in practical systems.

Thus, our goal is to find a method to group the users so that the channel for each subgroup is highly symmetric. Unfortunately, good metrics for the symmetry of channel are not available. As a substitute, we use the strength of the channel from the base stations to the users as a metric to form the subgroups. Formally, consider a received signal model of the form (2.7) with no assumptions on the entries of $\boldsymbol{H}$. (While presenting the simulation results, we will specify the model used for the entries of $\boldsymbol{H}$.) Then the channel to the j-th user is given by the j-th column of $\boldsymbol{H}$, that is,

$$\boldsymbol{h}_j = [h_{1,j}, h_{2,j}, \ldots h_{M,j}] \tag{2.30}$$

where M is the number of base stations in the system. To form the groups, we sort the users on the basis of their channel strength where channel strength is defined by the norm of $\boldsymbol{h}_j$:

$$\|\boldsymbol{h}_j\| \triangleq \sum_{i=1}^{M} h_{i,j}^2. \tag{2.31}$$

In Algorithm 2.1, we present the service policy for arbitrary, but constant, sized cellular systems with possible cooperation among base stations. We are assuming that all users need data at the same rate and that the channel is quasi-static. Let G be the size of subgroup. (In our simulations, we will work with $G = 3$ and $G = 4$.) We first sort the users on the basis of their channel strength as defined by (2.31). Here, we are assuming that there is no entry to or exit from the system. (As shown in Algorithm 2.2, a straightforward modification to allow for the possibility of entry/exit from the system will be to sort the users after each iteration of the algorithm.) We then select the first G users to form the subgroup to be served. If there are less than G users to serve, we

G - the size of subgroup.
Sort the users by their channel strength (2.31).
repeat
  Select the first G users. {If there are less than G users, select all users.}
  Transmit their data as per the policy proposed in Section 2.1.4.
  Remove these users from the list of users to be served.
until All users are served.

Algorithm 2.1 *Service policy for constant sized cooperating cellular systems*

G - the size of subgroup.
repeat
  Sort the users by their channel strength (2.31).
  Select the first G users. {If there are less than G users, select all users.}
  Transmit their data as per the policy proposed in Section 2.1.4.
  Remove these users from the list of users to be served.
until All users are served.

Algorithm 2.2 *Service policy for variable sized cooperating cellular systems*

group them all together. Since we have assumed that all users need data at the same rate, all the selected users have same amount of data to transmit. Furthermore, when all users have data to transmit, the policy given in Section 2.1.4 reduces to serving all users at a constant non-zero rate. Moreover, when all users have same amount of data to transmit (that is, $\boldsymbol{k} = (1, 1, \ldots, 1)$, the service rates are equal (to the symmetric capacity of the corresponding channel matrix) and therefore, all of the queues empty at the same time. After the selected users have been served, remove them from the list of users to be served. We repeat this procedure till all users have been served.

2.2.1 Simulation Results

We next present simulation results showing the efficacy of the policy proposed in Algorithm 2.1. Before doing so, we describe what we mean by traditional operation and the system model.

Under traditional operation, each mobile is assigned to a specific cell and thereafter all communications to/from that mobile is via the assigned base station. In our

simulations, we assign the mobile to the base station from which it can get the best signal. That is, if the channel from all base stations to a specific mobile is given by (2.30), we assign that mobile to the cell given by

$$\arg\max_i |h_{i,j}| \,. \tag{2.32}$$

We assume that at any given time, the only communication in the set of cells comprising the composite base station, is between a mobile and its corresponding base station and therefore, all transmissions are inter-cell interference free.

In our simulations, we consider a cellular system with four (4) base stations (each having one antenna) and 200 users. The placement of the base stations is shown in Fig. 2.12. As shown in the Figure, the base stations are at the center of the squares comprising the grid. The mobiles are placed randomly on the grid such that they are uniformly distributed in the two-dimensional space. The channel coefficient $h_{i,j}$ between the i-th base station and j-th mobile has three components. The first component corresponds to the path-loss. We assume that the path-loss exponent is 4. The second component corresponds to the shadow fading which is assumed to be a lognormal random variable of variance 6dB. As in Section 2.1, the short-term variations in the channel are assumed to be captured by a complex Gaussian component of mean 0 and variance $1/2$ in each component ($\mathcal{CN}(0,1)$). We define the SNR as it was defined in Section 2.1.1.

We first compare the effect of grouping size. As we increase the grouping size, G, there are more antennas on the receive side and therefore, it is expected that a higher grouping size should give a higher gain in throughput. On the other hand, with an increase in G, computing the capacity region gets computationally expensive and the coordination between the transmit and receive end gets more complex.

In the first set of results, we compare the gain in throughput with $G = 3$ and $G = 4$ where SNR is 10 dB (see Section 2.1.1 for the definition of the SNR). As shown in Table 2.1, the gain in throughput for $G = 4$ is about 8% more than the gain in throughput

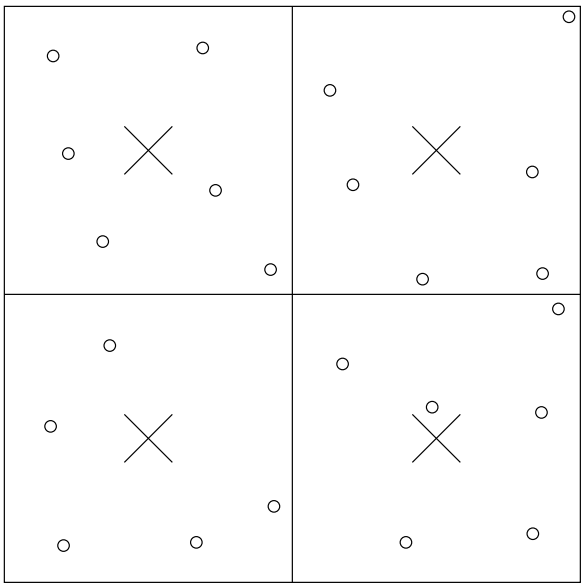


Figure 2.12 *Location of Base Stations and Mobiles in a cellular system with* 4 *cells. Base stations are denoted by "x" while the mobiles are denoted by "o".*

Table 2.1 *Gain in Throughput for different Grouping Sizes:* 4 *base stations and* 200 *mobiles. SNR in both cases is* 10 *dB.*

	Average gain	Standard Deviation	Standard Deviation/Average (in %)
$G = 3$	1.991	0.0334	1.68
$G = 4$	2.151	0.0378	1.76

for $G = 3$, a not very significant gain. Moreover, from the last entry in both rows, it is evident that in both cases, the standard deviation, as a fraction of the average gain, is low enough to give us confidence that the results are representative not outliers. Since there is no significant loss of gain in throughput by changing G from 4 to 3 and the above-mentioned computational issues are simpler for $G = 3$, in the next simulation we work with $G = 3$.

We next compare the effect of SNR on the gain in throughput. In Table 2.2, we show the average gain in throughput and the standard deviation of the gain for different SNRs with $G = 3$. Here again, the system has 4 base stations and 200 mobiles. As expected, it can be observed that the gain increases with SNR and that the variance of the gain decreases with SNR. Here again, in all cases the variance is low enough to give us

Table 2.2 *Gain in Throughput for different SNRs:* 4 *base stations and* 200 *mobiles. Group size,* G*, is* 3*.*

SNR	Average gain	Standard Deviation	Standard Deviation/Average (in %)
5 dB	1.840	0.0428	2.33
10 dB	1.991	0.0334	1.68
20 dB	2.308	0.0259	1.12

confidence in our results.

2.3 Acknowledgment

This chapter, in part, is a reprint of the material in the following papers: A. S. Acampora, S. Bhardwaj, and R. M. Tamari, “On Best-Case Throughput of Cellular Data Networks with Cooperating Base Stations”, in the Proceedings of *the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep 27–29 2006; A. S. Acampora, S. Bhardwaj, and R. M. Tamari, “A Best-Case Performance Comparison of Cellular Data Networks with Cooperating and Non-cooperating Base Stations”, submitted to *Wireless Communications and Mobile Computing*. The dissertation author was the primary investigator and author of these papers.

CHAPTER 3

Heavy Traffic Performance

3.1 Introduction

As described in Chapter 2, the cellular wireless network with infrastructure cooperation has a corresponding queueing system formulation where, even in the simple case of Poisson arrivals, independently for each user, it is not known how to minimize the average delay for a given load. Furthermore, closed-form expressions for average delay are unavailable for many simple policies; usually, this means that any meaningful comparison has to be done via simulations. However, when the ratio of the average arrival rates (also known as the relative traffic rate) is specified in advance, the maximum possible throughput can be computed and a simple policy can be shown to be throughput-optimal[1] under Markovian assumptions (see Chapter 2). But an exact expression for the performance of this policy is not available. In this Chapter, as a measure of performance, we prove limit theorems justifying a diffusion approximation for a heavily loaded system operating under this policy.

We are not aware of analysis of other policies that have been shown to be throughput-optimal for a general convex (rather than a convex polyhedral) capacity region. However, scheduling policies for certain heavily loaded wireless systems with convex polyhedral capacity regions have been studied in [37, 39] (also see references therein) under

[1]For a Markovian system, throughput-optimal means the long run average departure rate exists and equals the long run average arrival rate whenever the nominal load lies inside the capacity region, cf. [14, p. 26].

restrictive assumptions. In [39], Stolyar considered a generalized switch. He showed that under MaxWeight scheduling and certain restrictive conditions, including a resource pooling condition, in heavy traffic there is state space collapse (SSC), the workload process converges to a one-dimensional Reflecting Brownian Motion (RBM), and MaxWeight asymptotically minimizes the workload. Shakkotai et al. [37] study a throughput-optimal scheduling rule, which they call an exponential scheduling rule, and show that under resource pooling condition it is asymptotically pathwise optimal in the sense that there is SSC, the workload process is asymptotically minimized and converges to a one-dimensional RBM. In the following, we point out some of the differences between our assumptions and those in [37, 39]. The Maxweight policy [39] is designed for the case when the capacity region is a convex polyhedron while the policy we consider is designed for more general convex capacity regions. We elaborate upon this in Section 3.3.4 where we define the heavy traffic conditions. Moreover, a complete resource pooling (CRP) condition is assumed in [39] which requires that there is a unique outward pointing normal to the system stability region at the point corresponding to the mean arrival rate vector for a critical load; by comparison, we do not assume a CRP condition. The arrival process in [39] is assumed to be an ergodic Markov process while we assume that the arrival process is a renewal process. In [37], the capacity region is a convex polyhedron and a CRP condition similar to [39] is assumed; however, service is given to only one queue at a time while here we can serve more than one queue at the same time.

3.2 Organization of the Chapter

We first consider the case where there are only two users in the footprint of the cooperating base stations. For such a system, the policy under consideration has only four (4) operating points which makes it amenable to an exhaustive enumeration. We present such an approach in Section 3.3.

Unfortunately, for a system with N users, our policy has $2^N - 1$ operation points

and the approach taken in Section 3.3 is not amenable to scaling. In Section 3.4, we use other results from applied probability to analyze the performance of an arbitrarily-sized system. Though the main results in the two sections are similar, the reader will notice that the result in Section 3.3 is for the queuelength process, which is a counting process associated with the workload process (defined in the sequel), while the result in Section 3.4 is for the workload process.

We would like to mention that to maintain the completeness of the individual sections, so that an interested reader can restrict her attention to a particular section, we have some redundancy in the sequel. For example, we could combine the notations and preliminaries for the two sections, but this will, unfortunately, require a user interested in the result for the two-user case to read through unnecessary notation required for Section 3.4.

3.3 Two-User System: Queuelength

3.3.1 Notation and Preliminaries

We will use the following notation throughout this section. Let $\mathbb{Z}$ denote the set of all integers, $\mathbb{Z}_+$ the set of all non-negative integers, $\mathbb{R}$ denote the set of real numbers, and $\mathbb{R}_+$ denote the non-negative half-line, which is also denoted by $[0,\infty)$. For $d \geq 1$, $\mathbb{R}^d$ will denote d-dimensional Euclidean space and the positive orthant in this space will be denoted by $\mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x_i \geq 0 \text{ for } i = 1, 2, \ldots, d\}$. All vectors and matrices are assumed to have real valued entries. Let $0 = (0, 0, \ldots, 0) \in \mathbb{R}^d_+$. The usual Euclidean norm on $\mathbb{R}^d$ will be denoted by $\|\cdot\|$ so that $\|x\| = \left(\sum_{i=1}^d x_i^2\right)^{1/2}$ for $x \in \mathbb{R}^d$. We denote the inner product on $\mathbb{R}^d$ by $\langle \cdot, \cdot \rangle$, i.e., $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$, for $x, y \in \mathbb{R}^d$. Let $\mathcal{B}(\mathbb{R}^d)$ denote the σ-algebra of Borel subsets of $\mathbb{R}^d$. The symbol $1_{\mathcal{A}}$ denotes the indicator function of a set $\mathcal{A}$, i.e., $1_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$ and $1_{\mathcal{A}}(x) = 0$ if $x \notin \mathcal{A}$.

All stochastic processes used in this section will be assumed to have paths that are right continuous with finite left limits (r.c.l.l.). We denote by $\mathbb{D}^d$ the space of r.c.l.l. func-

tions from $[0,\infty)$ into $\mathbb{R}^d$ and we endow this space with the usual Skorokhod J_1-topology (see Ethier and Kurtz [12, Chapter 3, Section 5]). We denote by $\mathbb{C}^d$ the space of continuous functions from $[0,\infty)$ into $\mathbb{R}^d$, also endowed with the Skorokhod J_1-topology under which convergence of elements in $\mathbb{C}^d$ is equivalent to uniform convergence on compact time intervals. The σ-algebra induced on $\mathbb{D}^d$ (or $\mathbb{C}^d$) by the Skorokhod J_1-topology will be denoted by $\mathcal{M}^d$. The abbreviation *u.o.c.* will stand for *uniformly on compacts* and will be used to indicate that a sequence of functions in $\mathbb{D}^d$ (or $\mathbb{C}^d$) is converging uniformly on compact time intervals to a limit in $\mathbb{D}^d$ (or $\mathbb{C}^d$). A d-dimensional process is a measurable function from a probability space into $\mathbb{D}^d$. Consider $Q^1, Q^2, \ldots, Q$, each of which is a d-dimensional process (possibly defined on different probability spaces). The sequence $\{Q^n\}_{n=1}^{\infty}$ is said to be *tight* if the probability measures induced by the sequence $\{Q^n\}_{n=1}^{\infty}$ on $(\mathbb{D}^d, \mathcal{M}^d)$ form a tight sequence, i.e., they form a weakly relatively compact sequence in the space of probability measures on $(\mathbb{D}^d, \mathcal{M}^d)$. The notation "$Q^n \Rightarrow Q$" will mean that "$Q^n$ converges in distribution to Q as $n \to \infty$". The sequence of processes $\{Q^n\}_{n=1}^{\infty}$ is called *C-tight* if it is tight, and if each weak limit point (obtained as a weak limit along a subsequence) is in $\mathbb{C}^d$ almost surely.

3.3.1.1 Skorokhod Problem

Skorokhod problems are used in the study of approximations to certain queueing networks. Let $\mathbb{D}_+^d$ (resp. $\mathbb{C}_+^d$) denote those functions $x \in \mathbb{D}^d$ ($x \in \mathbb{C}^d$) satisfying $x(0) \geq 0$.

Definition 3.3.1 (Skorokhod Problem (SP))**.** *Fix $x \in \mathbb{D}_+^d$ and a $d \times d$ matrix R. We say that (z, y) solves the Skorokhod problem for x with respect to R, if $z, y \in \mathbb{D}_+^d$ with*

(i) $z(t) = x(t) + Ry(t)$ *for all* $t \in \mathbb{R}_+$,

(ii) $z(t) \in \mathbb{R}_+^d$ *for all* $t \in \mathbb{R}_+$,

(iii) for $i = 1, 2, \ldots, d$,

(a) $y_i(0) = 0$,

(b) y_i *is non-decreasing,*

(c) $\int_{(0,\infty)} z_i(s)dy_i(s) = 0$.

The path x *is called the driving path.*

Harrison and Reiman [16] specified some conditions on the matrix R under which there is a unique solution of the Skorokhod problem for each $x \in \mathbb{C}_+^d$. In fact these conditions also yield a unique solution for each $x \in \mathbb{D}_+^d$.

Definition 3.3.2 (Harrison-Reiman (HR) Condition)**.** *A* $d \times d$ *matrix* R *satisfies the HR condition if* $R = I - \tilde{P}$*, where* I *is the* $d \times d$ *identity matrix,* $\tilde{P}$ *has zeros along the diagonal, all of the entries of* $\tilde{P}$ *are nonnegative and* $\tilde{P}$ *has spectral radius strictly less than one.*

When $R = I - \tilde{P}$ where $\tilde{P}$ has zeros on the diagonal and the entries of $\tilde{P}$ are nonnegative, the HR condition is equivalent to the requirement that R is a non-singular M-matrix. Such matrices are discussed for example in Berman and Plemmons [3, Chapter 6].

Proposition 3.3.1. *Let* d *be a positive integer and* R *be a* $d \times d$ *matrix satisfying the HR condition. Then for each* $x \in \mathbb{D}_+^d$*, there are* $y, z \in \mathbb{D}_+^d$ *such that* (z, y) *is the solution of the Skorokhod problem for* x *with respect to* R*. Furthermore, the mapping* $\Phi : \mathbb{D}_+^d \to \mathbb{D}_+^{2d}$ *given by* $\Phi(x) = (z, y)$ *is continuous where* (z, y) *is the solution of the Skorokhod problem for* x*.*

Proof. The proof is given for $x \in \mathbb{C}_+^d$ in [16] and alluded to for $x \in \mathbb{D}_+^d$. A complete proof can be found in [44] for example. □

Fix a positive integer d, $\theta \in \mathbb{R}^d$, Γ a $d \times d$ symmetric strictly positive definite matrix and a $d \times d$ matrix R satisfying the HR condition. We can use the solvability of the Skorokhod problem to construct a Semimartingale Reflecting Brownian Motion (SRBM) associated with the data $(\mathbb{R}_+^d, \theta, \Gamma, R)$ as follows.

Given a Brownian motion X starting from the origin with drift vector θ and covariance matrix Γ, consider the pair of processes (Q, Y) that solve the Skorokhod problem for X with respect to R. Then, Q is an SRBM associated with the data $(\mathbb{R}^d_+, \theta, \Gamma, R)$ starting from the origin. Here $Q = X + RY$ where $\{X(t) - \theta t,\ t \geq 0\}$ is a continuous martingale (with respect to the filtration generated by X) and $\{RY(t) + \theta t,\ t \geq 0\}$ is a continuous locally bounded variation process adapted to the filtration generated by X. Hence, Q is a semimartingale.

3.3.2 System Model

In this subsection we specify the communication system under consideration. We consider a cellular wireless network where base stations cooperate over noise-free infinite capacity links. We do not make any distinction between a single-cell cellular system having multiple base-station antennas and the traditional cellular system with cooperating single-antenna base stations. Here, by cooperation we mean that the base stations can perform joint beamforming and/or power control but there is a constraint on the total power that the base stations can share. We do not make any assumptions about the number of receive antennas per user.

In this section, we restrict our attention to the case where there are just two mobile stations (also called users) in the footprint of the cooperating base stations. Then the downlink channel can be modeled as a two-user MIMO broadcast channel. We assume that the channel is fixed for all transmissions over the period of interest (some authors refer to this as a quasi-static channel). Moreover, we assume that the transmit end (the cooperating base stations) has perfect channel state information (CSI).

Weingarten et al. [43] have shown that for such a system, Dirty Paper Coding (DPC), introduced by Costa [10], achieves the capacity. Furthermore, the capacity region can be computed by using the duality of the MIMO multiple access channel and the MIMO broadcast channel [20]. Figure 3.1 illustrates the capacity region for an example

of a two-user MIMO broadcast channel with two transmit and two receive antennas. Here the broadcast channel capacity region is obtained by taking the convex hull of the union over the set of capacity regions of the dual MIMO multiple access channels such that the total multiple access channel power is the same as the power in the broadcast channel.

Let c_1^* (c_2^*) be the maximum rate at which data can be transmitted (in bits per second (bps)) to user 1 (2) when the rate of transmission to user 2 (1) is set at zero. If $(c_1, c_2) > 0$ is a point in the capacity region then the rate at which data can be transmitted to user 1 (2), c_1 (c_2), is strictly less than c_1^* (c_2^*). This corresponds to the fact that when the wireless resources are dedicated to a single user, the rate at which that user can be served is higher than the rate for that user when the resources are shared by the users but this higher rate comes at a cost to the sum of the rates. Indeed, when both users are being serviced, the sum of the rates is strictly greater than that for service dedicated to a single user, that is, $c_1 + c_2 > c_1^*, c_2^*$.

For a two-user system the capacity region is a two-dimensional closed convex set in $\mathbb{R}_+^2$ where the convexity follows because of the convex hull operation. The capacity region contains the origin and it has three boundary pieces of which two are along the coordinate axes while the third boundary piece is in the interior of $\mathbb{R}_+^2$. We call this third boundary the *capacity surface*. The following lemma states a key property of the capacity surface of the two-user MIMO broadcast channel.

Lemma 3.3.2. *For any point (x, y) on the capacity surface of a two-user MIMO broadcast channel, the following holds,*

$$\frac{x}{c_1^*} + \frac{y}{c_2^*} > 1. \tag{3.1}$$

Proof. As stated earlier, the capacity region is a convex set in $\mathbb{R}_+^2$, it contains the origin and it has the line segments $(0, 0)$ to $(c_1^*, 0)$ and $(0, 0)$ to $(0, c_2^*)$ along the two coordinate axes as two boundaries. Since the line segment $\{(x, y) \in \mathbb{R}_+^2 : \frac{x}{c_1^*} + \frac{y}{c_2^*} = 1\}$ lies in the capacity region (by convexity), the capacity surface must lie "along or above" this line

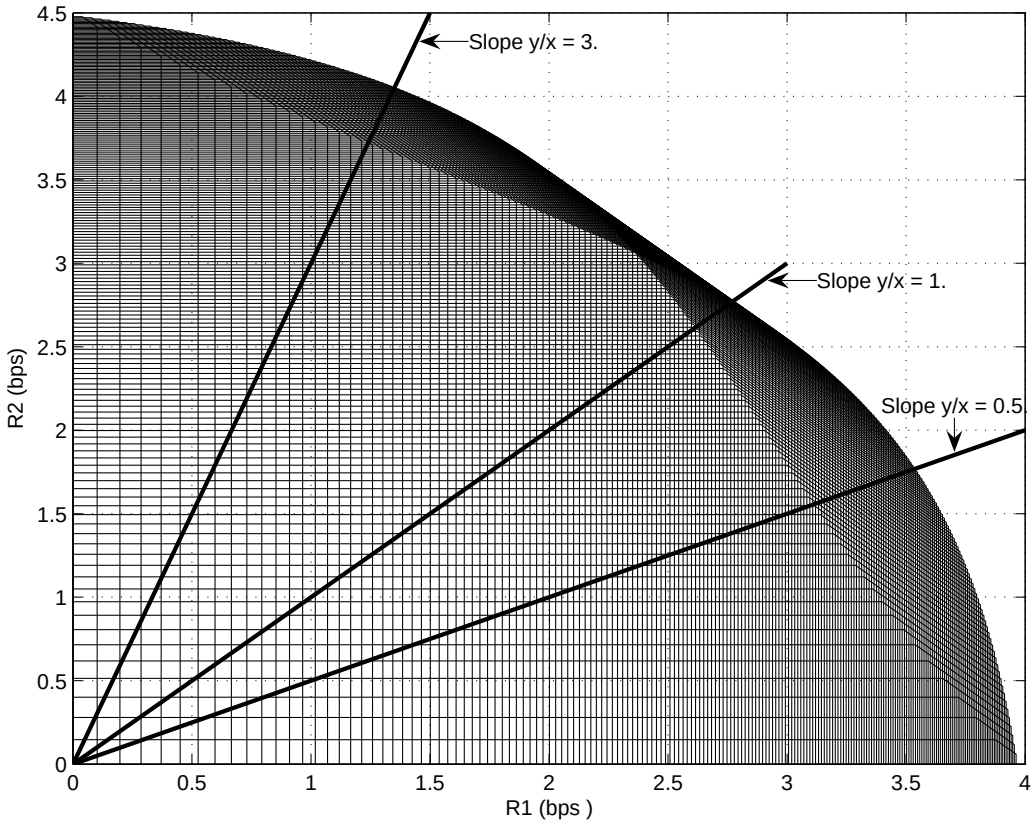


Figure 3.1 *An example of a capacity region of a 2-user MIMO broadcast channel for a fixed channel where R_1 and R_2 are the rates of user 1 and 2, respectively.*

segment and so for any point on the capacity surface we have

$$\frac{x}{c_1^*} + \frac{y}{c_2^*} \geq 1. \tag{3.2}$$

From (3.2) and the convexity of the capacity region, if there is a point on the capacity surface where (3.1) holds, it holds for every point on the capacity surface. We next show that there is at least one point on the capacity surface where (3.1) holds.

The sum-rate capacity of the MIMO broadcast channel is defined as the maximum of the sum of a pair of rates that can be transmitted. (See Viswanath et al. [42] for details.) If the sum-rate capacity of the MIMO broadcast channel is strictly greater than the single-user capacities, c_1^* and c_2^*, then (3.1) holds at the point(s) achieving sum-rate capacity. This follows by noting that if only equality held in (3.2), at a point where sum-rate capacity is achieved, the maximum sum rate would be achieved with one of x or y equal to zero (i.e., at an end-point of the line segment $\{(x, y) \in \mathbb{R}_+^2 : x/c_1^* + y/c_2^* = 1\}$) but then the sum-rate equals c_1^* or c_2^*, a contradiction. From [42, Theorem 3], the sum-rate capacity of MIMO broadcast channel is the Sato upper bound [35] which is greater than the single-user capacities. Thus, there is a point on the capacity surface where (3.1) holds, and the lemma follows. □

At the transmit end, packets arrive for each user and are buffered before transmission. The ratio of anticipated average bit arrival rates, called relative traffic rate and denoted by k_2, is specified in advance, that is, it is expected that, on average, user 2 will have k_2 times as much data as user 1. The actual traffic rate will deviate from the average due to stochastic fluctuations. Naturally, when there is no data for one of the users to transmit (the corresponding queue for that user is empty), the data for the other user should be transmitted at the maximum possible rate. That is, the data should be transmitted to user 1 (2) at the rate of c_1^* (c_2^*) when only the first (second) user has data to transmit. In Chapter 2, we have shown that under Markovian assumptions on the system, the policy that transmits at the rate (c_1, c_2) at all other times, where (c_1, c_2) is the point on

the capacity surface such that $c_2/c_1 = k_2$, is throughput-optimal. Figure 2.4 illustrates a few such operation points for sample values of $k_2 = 3, 1, 0.5$.

3.3.3 Queueing Analogue

In this subsection we develop a queueing analogue for the system described in Section 3.3.2. To this end, we describe the physical structure, the packet arrivals and sizes. Then we formalize the service discipline and specify the dynamic equations satisfied by the queuelength process.

3.3.3.1 Physical Structure

A queueing system describing our setup has two queues in parallel where each queue buffers packets intended for a given user. We assume that each of the queues has infinite buffer capacity. The queues are served by a single server corresponding to the cooperating base station.

3.3.3.2 Stochastic Primitives

We assume that the system starts empty and that there is a two-dimensional packet arrival process $E = \{(E_1(t), E_2(t)),\ t \geq 0\}$ where $E_i(t)$ is the number of packets that have arrived to the i-th queue in $(0, t]$. (Here E is used to indicate that the arrivals are exogenous.) For $i = 1, 2$, $E_i(\cdot)$ is assumed to be a (non-delayed) renewal process defined from a sequence of strictly positive i.i.d. random variables $\{u_i(k),\ k = 1, 2, \dots\}$, where for $k = 1, 2, \dots, u_i(k)$ denotes the time between the arrival of the $(k-1)$st and the k-th packet to the i-th queue. Each $u_i(k)$, $k = 1, 2, \dots$ is assumed to have finite mean $1/\lambda_i \in (0, \infty)$ and finite squared coefficient of variation (variance divided by the mean squared) $\alpha_i^2 \in [0, \infty)$. The packet lengths (in bits) for the successive arrivals to queue i are given by a sequence of strictly positive i.i.d. random variables $\{v_i(k), k = 1, 2, \dots\}$ with

average packet length $1/\mu_i \in (0,\infty)$ and squared coefficient of variation $\beta_i^2 \in [0,\infty)$, $i = 1,2$. We assume that all interarrival and service time processes are mutually independent. Note that the average bit arrival rate for user i is $b_i = \lambda_i/\mu_i$, $i = 1,2$ and we have let $k_2 = b_2/b_1$. For $i = 1,2$, we associate a renewal counting process $S_i(\cdot)$ with $\{v_i(k)\}_{k=1}^{\infty}$ such that $S_i(t) = \sup\{n \geq 0 : \sum_{k=1}^{n} v_i(k) \leq t\}$ for $t \geq 0$. We refer to the processes $E(\cdot)$ and $S(\cdot)$ as *stochastic primitives* for the system model.

3.3.3.3 Service Discipline

When service is given to a queue, it goes to the packet at the head of the line, where it is assumed that packets are queued in the order of their arrival to the queue. The service rate is a simple function of the number of packets in each of the queues. A pair (σ_1, σ_2) indicates the rates (in bps) of serving the two queues, i.e., σ_1 is the rate for queue 1 and σ_2 is the rate for queue 2. Here, given the queuelength $q = (q_1, q_2)$, the rates are given by $(\sigma_1, \sigma_2) = \Lambda(q)$ for the function[2] $\Lambda : \mathbb{R}_+^2 \to \mathbb{R}_+^2$ defined by

$$\Lambda(q) \triangleq \begin{cases} (c_1, c_2) & \text{if } q_1 > 0, q_2 > 0, \\ (c_1^*, 0) & \text{if } q_1 > 0, q_2 = 0, \\ (0, c_2^*) & \text{if } q_1 = 0, q_2 > 0, \\ (0,0) & \text{if } q_1 = 0, q_2 = 0. \end{cases} \tag{3.3}$$

Here c_1 and c_2 are chosen such that (c_1, c_2) lies on the capacity surface and $c_2/c_1 = k_2$. Also, c_1, c_2, c_1^* and c_2^* satisfy the following conditions: $0 < c_1 < c_1^*$, $0 < c_2 < c_2^*$, and $c_1^*, c_2^* < c_1 + c_2$.

Our model is a single server, two-class queueing system where the two classes correspond to the two users. The following scaling property of $\Lambda(\cdot)$ is a mathematical

[2]We only need $\Lambda(\cdot)$ defined on $\mathbb{Z}_+^2$ for the moment, but we extend the domain of $\Lambda(\cdot)$ to $\mathbb{R}_+^2$ so that later when we rescale the queuelength process $\Lambda(\cdot)$ is well-defined for the rescaled process.

statement of the property of the scheduling policy that the amount of service given to the queues in any state does not change when all queuelengths are increased/decreased proportionally.

Lemma 3.3.3. *For any* $q \in \mathbb{R}^2_+$ *and* $x > 0$, $\Lambda(xq) = \Lambda(q)$.

Proof. The proof follows easily from the definition of $\Lambda(\cdot)$. □

3.3.3.4 Queuelength Process

For $i = 1, 2$, the length of the i-th queue at time t is

$$Q_i(t) = E_i(t) - D_i(t), \tag{3.4}$$

where $D_i(t)$ is the number of packet departures from the i-th queue in $(0, t]$. Here, $D_i(t)$ is given by

$$D_i(t) = S_i(T_i(t)), \tag{3.5}$$

where $T_i(t)$, the cumulative amount of service given to queue i up to time t, is given by

$$\begin{aligned} T_i(t) &\triangleq \int_0^t \Lambda_i(Q(s))\, ds \\ &= c_i \int_0^t 1_{\{Q_j(s)>0 \text{ for all } j\}}\, ds + c_i^* \int_0^t 1_{\{Q_i(s)>0;\ Q_j=0 \text{ for all } j\neq i\}}\, ds. \end{aligned} \tag{3.6}$$

3.3.4 Heavy Traffic Assumptions

3.3.4.1 Assumptions

We consider the operation of our queueing system in the asymptotic regime where it is heavily loaded. (Kelly and Laws [23] have argued that in this regime "important features of good control policies are displayed in sharpest relief".) For this purpose one may regard a given system as a member of a sequence of systems approaching the heavy

traffic limit. To obtain a reasonable approximation, the queuelength process is rescaled using diffusion scaling. This corresponds to viewing the system over long intervals of time of order r^2 (where r will tend to infinity in the asymptotic limit) and regarding a single packet as only having a small contribution to the overall congestion level, where this is quantified to be of order $1/r$. Formally, we consider a sequence of systems indexed by r, where r tends to infinity through a sequence of values in $(0, \infty)$. These systems all have the same basic structure as that described in the last subsection; however, the arrival rates may vary with r and for determining c we assume that an estimate of the ratio $k_2 \in (0, \infty)$ of the bit arrival rates is known and is used to determine the capacity c for the whole sequence. We assume that the interarrival times in the system indexed by r are given for each $i = 1, 2$, $k = 1, 2, \ldots$, by

$$u_i^r(k) = \frac{1}{\lambda_i^r} \check{u}_i(k) \tag{3.7}$$

where the $\check{u}_i(k)$ do not depend on r, have mean one and squared coefficient of variation α_i^2. The packet lengths $\{v_i(k)\}_{k=1}^{\infty}$, $i = 1, 2$, do not change with r. [The above structure is convenient for allowing the sequence of systems to approach heavy traffic by simply changing arrival rates and keeping the underlying sources of variability $\check{u}_i(k)$ and $v_i(k)$ fixed as r varies. This type of set-up has been used previously by others in treating heavy-traffic limits (see, e.g., Peterson [31] and Bell and Williams [2]). For a first pass, the reader may like to simply choose $\lambda_i^r = \lambda_i$ for all r.] All processes and parameters that depend on r will from now have a superscript of r. Define $\lambda_i \triangleq \mu_i c_i$, $i = 1, 2$.

Assumption 3.3.4 (Heavy Traffic Assumption)**.** *For $i = 1, 2$, there is $\theta_i \in \mathbb{R}$ such that*

$$r(\lambda_i^r - \lambda_i) \to \theta_i \text{ as } r \to \infty. \tag{3.8}$$

Remark. This assumption does not restrict the direction in which the heavy traffic limit is approached, unlike that in Gans and Van Ryzin [13]. Here θ_i could be positive, negative

or zero for each i. Thus, each queue may have an arrival rate that is greater than, equal to or less than the rate yielding exact balance.

Here we may regard λ as the nominal average packet arrival rate used to set the service rates, $(c_1, c_2) = (b_1, b_2)$, for the throughput-optimal policy. The r-th system has a perturbed average packet arrival rate λ^r for which the average bit arrival rate $b^r : b_i^r = \lambda_i^r/\mu_i, i = 1, 2$, is close to (c_1, c_2).

3.3.4.2 Connection to Complete Resource Pooling (CRP)

To make a connection with the work of Stolyar [39] (and others), consider the two-user queueing system where the server is able to time-share amongst finitely many operation points chosen from the closure of the capacity surface and the origin. (To allow for viable operation when one or both queues are empty, we assume that the points $(0, 0)$, $(c_1^*, 0)$, and $(0, c_2^*)$ are included amongst the finitely many operation points.) A representative capacity surface for a two-user MIMO broadcast channel is shown in Fig. 3.2. For this system, the system stability region is the closed convex hull of the set of operation points. For example, if the operation points are $(c_1^*, 0)$, $(0, c_2^*)$, $c^1 = (c_1^1, c_2^1)$, $c^2 = (c_1^2, c_2^2)$, $c^3 = (c_1^3, c_2^3)$, and $(0, 0)$ as indicated in Fig. 3.2, then the upper surface of the system stability region, $\tilde{\mathcal{C}}$, is shown by the dashed curve.

Recall that the ray from the origin of slope k_2 intersects the boundary of $\mathcal{C}$, the capacity region, at the point $c = (c_1, c_2)$. Suppose that $\mathcal{C}$ is strictly convex at c, i.e., the capacity surface is not flat at c. The following lemma shows that then the point c must be one of the operation points, otherwise the system will be unstable in heavy traffic. Furthermore, when c is amongst the operation points, the CRP condition does not hold.

Lemma 3.3.5. *Suppose that the point $c = (c_1, c_2)$, where the ray from the origin of slope k_2 intersects the capacity surface, is an extreme point of $\mathcal{C}$. Then c must be one of the operation points of any policy that is stable whenever the arrival rate is $(1 - 1/r)\lambda$ for all $r \in (1, \infty)$. Furthermore, there is then more than one normal to $\tilde{\mathcal{C}}$ at c, and the complete*

resource pooling condition does not hold.

Proof. Consider a policy that time shares amongst finitely many operation points not including c. The average bit arrival rate vector b^r associated with the average arrival rate of $(1-1/r)\lambda$ for $r \in (1,\infty)$, approaches the point c along the ray from the origin of slope k_2. Since c is an extreme point of $\mathcal{C}$ and c is not an operation point, c is outside $\tilde{\mathcal{C}}$. Thus, there is an $\hat{r}$ such that for $r > \hat{r}$, b^r is in the capacity region $\mathcal{C}$ but not in $\tilde{\mathcal{C}}$ (as illustrated in Fig. 3.2). Thus, the time sharing policy is not stable for all b^r such that $r > \hat{r}$.

Now, if c is one of the finitely many operation points of a time-sharing policy, since c cannot be written as a convex combination of the other operating points, there is not a unique normal to the boundary of $\tilde{\mathcal{C}}$ at c. This is illustrated in Fig. 3.2 where c^1 is one of the extreme points but there is no unique normal to $\tilde{\mathcal{C}}$ at c^1. □

The analysis performed in [39] depends critically on the (CRP) assumption that there is a unique normal to $\tilde{\mathcal{C}}$ at the point where the ray in the direction of the average bit arrival rate vector intersects $\tilde{\mathcal{C}}$. Except in the special situation where c is a convex combination of two other operation points, this assumption will not be satisfied at c and hence the analysis based on the assumption that the CRP condition holds does not apply.

3.3.5 Scaling and Standard Limit Theorems

3.3.5.1 Scaling

We first consider a fluid scaled version of the system where fluid scaling corresponds to viewing the system over long intervals of time of order r^2 and simultaneously reducing the contribution of a single packet to the congestion level by a factor of $1/r^2$. The behavior of solutions of a limiting fluid model will play an important role in establishing a limit for the diffusion scaled system where diffusion scaling corresponds to looking over time intervals of order r^2 but only diminishing packet contributions to the congestion measures by a factor of $1/r$. We define the following fluid and diffusion scaled processes.

Fluid Scaling Fluid (or functional law of large numbers) scaling is indicated by placing a bar over a process. For $i = 1, 2$, $t \geq 0$, and $r > 0$, define

$$\bar{T}_i^r(t) \triangleq r^{-2} T_i^r(r^2 t), \tag{3.9}$$

$$\bar{Q}_i^r(t) \triangleq r^{-2} Q_i^r(r^2 t), \tag{3.10}$$

$$\bar{E}_i^r(t) \triangleq r^{-2} E_i^r(r^2 t), \tag{3.11}$$

$$\bar{S}_i^r(t) \triangleq r^{-2} S_i^r(r^2 t). \tag{3.12}$$

There are in fact two kinds of fluid scaling. In addition to that indicated above, one could simply accelerate time by r and scale the process by $\frac{1}{r}$ (in place of r^2 and $\frac{1}{r^2}$, respectively). Here we shall only need the first form of fluid scaling described above.

Diffusion Scaling Diffusion (or functional central limit theorem) scaling is indicated by placing a hat over a process. For $i = 1, 2$, and $r > 0$, define

$$\hat{Q}_i^r(t) \triangleq \frac{Q_i^r(r^2 t)}{r}, \; t \geq 0, \tag{3.13}$$

as the diffusion scaled version of $Q_i^r(\cdot)$. To apply diffusion scaling to the primitive stochastic processes E^r, S, we must center them before scaling. Accordingly, for $i = 1, 2$, $t \geq 0$ and $r > 0$, we define

$$\hat{E}_i^r(t) \triangleq \frac{E_i^r(r^2 t) - \lambda_i^r r^2 t}{r} \tag{3.14}$$

and

$$\hat{S}_i^r(t) \triangleq \frac{S_i(r^2 t) - \mu_i r^2 t}{r}. \tag{3.15}$$

3.3.5.2 Functional Limit Theorems for Stochastic Primitives

We will use the following functional central limit theorem (FCLT) for the stochastic primitives in the sequel.

Proposition 3.3.6 (FCLT)**.** *The diffusion scaled processes* $(\hat{E}^r(\cdot), \hat{S}^r(\cdot))$ *jointly converge in distribution to* $(B_E(\cdot), B_S(\cdot))$ *as* $r \to \infty$*, i.e.,*

$$(\hat{E}^r(\cdot), \hat{S}^r(\cdot)) \Rightarrow (B_E(\cdot), B_S(\cdot)) \text{ as } r \to \infty, \tag{3.16}$$

where $B_E(\cdot)$ *and* $B_S(\cdot)$ *are independent two-dimensional driftless Brownian motions starting from the origin with diagonal covariance matrices* $\Gamma_E \triangleq \mathrm{diag}(\lambda_1\alpha_1^2, \lambda_2\alpha_2^2)$ *and* $\Gamma_S \triangleq \mathrm{diag}(\mu_1\beta_1^2, \mu_2\beta_2^2)$*, respectively.*

Remark. As there is a single source of variability (not depending on r) for each of E_i^r, S_i, $i = 1, 2$, only the finiteness of the second moments of $\check{u}_i(k)$ and $v_i(k)$ is required for the FCLT. Furthermore, since a Brownian motion is a continuous process, the weak-convergence of $(\hat{E}^r(\cdot), \hat{S}^r(\cdot))$ to a Brownian motion implies C-tightness of the sequence $\{(\hat{E}^r(\cdot), \hat{S}^r(\cdot))\}$.

Proof. By results of Iglehart and Whitt [18], functional central limit theorems for the renewal counting processes $\hat{E}^r(\cdot)$ and $\hat{S}^r(\cdot)$ can be inferred from those for the partial sums of $\{u_i^r(k)\}_{k=1}^{\infty}$ and $\{v_i(k)\}_{k=1}^{\infty}$, respectively. Functional central limit theorems for the latter follow from Theorem 3.1 of Prokhorov [32]. □

As a corollary, we have the following functional law of large numbers (FLLN) for the stochastic primitives. For this section, from now on, for each $t \geq 0$, let $\lambda(t) \triangleq \lambda t$ and $\mu(t) \triangleq \mu t$.

Corollary 3.3.7 (FLLN)**.** *The fluid-scaled processes* $(\bar{E}^r(\cdot), \bar{S}^r(\cdot))$ *jointly converge in*

distribution to $(\lambda(\cdot), \mu(\cdot))$ *as* $r \to \infty$*, i.e.,*

$$\left(\bar{E}^r(\cdot), \bar{S}^r(\cdot)\right) \Rightarrow (\lambda(\cdot), \mu(\cdot)) \text{ as } r \to \infty. \tag{3.17}$$

Remark. The weak-convergence of $(\bar{E}^r(\cdot), \bar{S}^r(\cdot))$ to a continuous process implies C-tightness of the sequence $\{(\bar{E}^r(\cdot), \bar{S}^r(\cdot))\}$.

Proof. Proposition 3.3.6 implies that

$$\left(\frac{1}{r}\hat{E}^r(\cdot), \frac{1}{r}\hat{S}^r(\cdot)\right) \Rightarrow (0,0) \text{ as } r \to \infty. \tag{3.18}$$

The desired result follows from this and the fact that $\lambda_i^r \to \lambda_i$ as $r \to \infty$ by (3.8) for $i = 1, 2$. □

3.3.6 Fluid Model

Applying fluid scaling to the dynamic equation (3.4) satisfied by the queuelength process for the system indexed by r, we obtain for $r > 0$, $i = 1, 2$, $t \geq 0$,

$$\bar{Q}_i^r(t) = \bar{E}_i^r(t) - \bar{S}_i^r(\bar{T}_i^r(t)). \tag{3.19}$$

We next consider the behavior of $\bar{T}^r(\cdot)$, the fluid-scaled version of $T^r(\cdot)$:

$$\bar{T}^r(t) = \frac{1}{r^2}\int_0^{r^2 t} \Lambda\left(Q^r(s)\right) ds, \ t \geq 0. \tag{3.20}$$

By the change of variables $\tilde{s} = \frac{s}{r^2}$, for $t \geq 0$, (3.20) becomes

$$\bar{T}^r(t) = \int_0^t \Lambda\left(\frac{r^2 Q^r(r^2\tilde{s})}{r^2}\right) d\tilde{s} = \int_0^t \Lambda\left(\bar{Q}^r(\tilde{s})\right) d\tilde{s}. \tag{3.21}$$

where the second equality follows from the definition of $\bar{Q}^r(\cdot)$ and the scaling property of

$\Lambda(\cdot)$ (see Lemma 3.3.3). The following lemma follows from (3.21) and the fact that $\Lambda_i(\cdot)$ is bounded by c_i^* which is less than $c_1 + c_2$, for $i = 1, 2$.

Lemma 3.3.8. *For each $r > 0$, almost surely $\bar{T}^r(\cdot)$ is uniformly Lipschitz continuous with Lipschitz constant less than $c_1 + c_2$.*

Remark. This lemma is used to prove the C-tightness of the fluid-scaled stochastic processes.

For a continuous function $x : [0, \infty) \to \mathbb{R}$, we say that $t \in (0, \infty)$ is a *regular point* for x if x is differentiable at t. If x is absolutely continuous, almost every $t \in (0, \infty)$ is a regular point and x can be recovered from its almost everywhere (a.e.) defined derivative $\dot{x}$:

$$x(t) = x(0) + \int_0^t \dot{x}(s)ds, \quad t \geq 0. \tag{3.22}$$

A (uniformly) Lipschitz continuous function $x : [0, \infty) \to \mathbb{R}$ is absolutely continuous.

Lemma 3.3.9. *The sequence of processes $\left\{(\bar{E}^r(\cdot), \bar{S}^r(\cdot), \bar{T}^r(\cdot), \bar{Q}^r(\cdot))\right\}$ converges in distribution to $\left(\bar{E}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), \bar{Q}(\cdot)\right)$ as $r \to \infty$ where*

$$\bar{E}(\cdot) = \lambda(\cdot), \quad \bar{S}(\cdot) = \mu(\cdot), \quad \bar{Q}(\cdot) = 0, \quad \bar{T}(\cdot) = c(\cdot), \tag{3.23}$$

and $c(t) \triangleq (c_1 t, c_2 t)$, $t \geq 0$.

Proof. From the uniform Lipschitz continuity of $\{\bar{T}^r(\cdot)\}$ established in Lemma 3.3.8, it follows that $\{\bar{T}^r(\cdot)\}$ is C-tight. Since, $\{\bar{E}^r(\cdot)\}$ and $\{\bar{S}^r(\cdot)\}$ are also C-tight (see the remarks following Corollary 3.3.7), using (3.19) together with the random time change theorem of Billingsley [5, p. 151], we conclude that the sequence $\left\{(\bar{E}^r(\cdot), \bar{S}^r(\cdot), \bar{T}^r(\cdot), \bar{Q}^r(\cdot))\right\}$ is C-tight as well. Suppose $\left(\bar{E}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), \bar{Q}(\cdot)\right)$ is a weak limit point of this sequence. By invoking the Skorokhod representation theorem (see, e.g., [12, Theorem 3.1.8, p. 102]), we may assume without loss of generality that for a subsequence $\{r_k\}$ of $\{r\}$, $\left\{\left(\bar{E}^{r_k}(\cdot), \bar{S}^{r_k}(\cdot), \bar{T}^{r_k}(\cdot), \bar{Q}^{r_k}(\cdot)\right)\right\}_{k=1}^{\infty}$ and $\bar{T}(\cdot)$ are defined on a common probability

space such that

$$\bar{Q}_i^{r_k}(t) = \bar{E}_i^{r_k}(t) - \bar{S}_i^{r_k}(\bar{T}_i^{r_k}(t)) \text{ for } t \geq 0, i = 1, 2 \tag{3.24}$$

and almost surely as $k \to \infty$,

$$\left(\bar{E}^{r_k}(\cdot), \bar{S}^{r_k}(\cdot), \bar{T}^{r_k}(\cdot), \bar{Q}^{r_k}(\cdot)\right) \to \left(\lambda(\cdot), \mu(\cdot), \bar{T}(\cdot), \bar{Q}(\cdot)\right) \text{ u.o.c.} \tag{3.25}$$

where almost surely $\bar{Q}_i(t) = \lambda_i t - \mu_i \bar{T}_i(t)$, $t \geq 0$, $i = 1, 2$. The limit $\bar{T}(\cdot)$ inherits the Lipschitz property of $\left\{\bar{T}^r(\cdot)\right\}$ almost surely. Fix ω such that $\bar{T}(\cdot, \omega)$ is uniformly Lipschitz continuous. In the following, we suppress explicit indication of the dependence on ω, but ω is fixed throughout. Let $t > 0$ be a regular point for $\bar{T}_i$, $i = 1, 2$, then $\bar{Q}$ is differentiable at t and

$$\frac{d\bar{Q}_i(t)}{dt} = \lambda_i - \mu_i \frac{d\bar{T}_i(t)}{dt}, \ i = 1, 2. \tag{3.26}$$

We consider the following cases for $\bar{Q}_i(t)$:

Case I: $\bar{Q}_i(t) = 0$ **for** $i = 1, 2$. Fix i. Since $\bar{Q}_i(\cdot) \geq 0$, $\bar{Q}_i(t) = 0$ and $t > 0$ is a regular point for $\bar{T}$ and $\bar{Q}$, it follows from a simple analysis argument that $d\bar{Q}_i(t)/dt = 0$. Then,

$$0 = \lambda_i - \mu_i \frac{d\bar{T}_i(t)}{dt}, \tag{3.27}$$

which implies that

$$\frac{d\bar{T}_i(t)}{dt} = \frac{\lambda_i}{\mu_i} = c_i. \tag{3.28}$$

Case II: $\bar{Q}_i(t) > 0$ **for** $i = 1, 2$. Let $0 \leq u < v < \infty$ be such that $t \in (u, v)$ and for $i = 1, 2$, $\bar{Q}_i(s) > 0$ for all $s \in [u, v]$. Then, by the uniform convergence of $\bar{Q}^r(\cdot)$ to $\bar{Q}(\cdot)$ on $[u, v]$, we have for all sufficiently large r, for $i = 1, 2$, $\bar{Q}_i^r(s) > 0$ for all $s \in [u, v]$. So

for all $s > t$ in $[u, v]$ we have

$$\begin{aligned}\bar{T}_i(s) - \bar{T}_i(t) &= \lim_{r\to\infty} \left[\bar{T}_i^r(s) - \bar{T}_i^r(t)\right] = \lim_{r\to\infty} \left[\int_t^s \Lambda_i\left(\bar{Q}_i^r(z)\right) dz\right] \\ &= \lim_{r\to\infty} \left[\int_t^s c_i\, dz\right] = c_i(s-t),\end{aligned} \tag{3.29}$$

where we have used the fact that $\Lambda_i(q) = c_i$, $i = 1, 2$ when $q > 0$. Dividing by $(s-t)$ and taking the limit as $s \to t$, we obtain $d\bar{T}_i(t)/dt = c_i$ for $i = 1, 2$. Note that this implies that $d\bar{Q}_i(t)/dt = 0$ for $i = 1, 2$, by (3.26) and since $\lambda_i = \mu_i c_i$.

Case III: There is $i \in \{1, 2\}$ such that $\bar{Q}_i(t) > 0$ and $\bar{Q}_j(t) = 0$ for $j \neq i$. Since for $j \neq i$, $\bar{Q}_j(\cdot) \geq 0$, $\bar{Q}_j(t) = 0$ and $t > 0$ is a regular point, it follows that $d\bar{Q}_j(t)/dt = 0$ which implies that $d\bar{T}_j(t)/dt = c_j$. Let $0 \leq u < v < \infty$ be such that $t \in (u, v)$ and $\bar{Q}_i(s) > 0$ for all $s \in [u, v]$. Then, for all sufficiently large r, $\bar{Q}_i^r(s) > 0$ for all $s \in [u, v]$, which implies by the definition of $\Lambda_i(\bar{Q}^r(\cdot))$ that

$$c_i(s-t) \leq \bar{T}_i^r(s) - \bar{T}_i^r(t) \leq c_i^*(s-t) \text{ for all } s > t \text{ in } [u, v]. \tag{3.30}$$

Letting $r \to \infty$ yields

$$c_i(s-t) \leq \bar{T}_i(s) - \bar{T}_i(t) \leq c_i^*(s-t), \text{ for all } s > t \text{ in } [u, v]. \tag{3.31}$$

Dividing by $(s-t)$ and letting $s \to t$, we conclude that $c_i \leq d\bar{T}_i(t)/dt \leq c_i^*$. Thus from (3.26), since $\lambda_i = \mu_i c_i$,

$$d\bar{Q}_i(t)/dt \leq 0. \tag{3.32}$$

Combining cases (I)–(III) we see that at each regular point $t > 0$ for $\bar{T}(\cdot)$,

$$\frac{d}{dt}\left(\bar{Q}_1^2(t) + \bar{Q}_2^2(t)\right) = 2\left[\bar{Q}_1(t)\frac{d\bar{Q}_1(t)}{dt} + \bar{Q}_2(t)\frac{d\bar{Q}_2(t)}{dt}\right] \leq 0. \tag{3.33}$$

Since $\bar{Q}_1^2(0) + \bar{Q}_2^2(0) = 0$ and $\bar{Q}_1^2(\cdot) + \bar{Q}_2^2(\cdot) \geq 0$, it follows that $\bar{Q}_1^2(t) + \bar{Q}_2^2(t) = 0$ for

all $t \geq 0$. Hence, $\bar{Q}_1(t) = \bar{Q}_2(t) = 0$ for all $t \geq 0$ and case (I) implies that $\dot{\bar{T}}_i(t) = c_i$ at each regular point $t > 0$ for $i = 1, 2$. Such regular points, t, occur almost everywhere and $\bar{T}_i$ can be recovered from its a.e. defined derivative to give $\bar{T}_i(t) = c_i t$ for all $t \geq 0$, $i = 1, 2$.

Finally, since $\left(\bar{E}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), \bar{Q}(\cdot)\right)$ was an arbitrary weak limit point and is unique (as shown above), it follows that $\left\{\left(\bar{E}_i^r(t), \bar{S}_i^r(t), \bar{T}_i^r(t), \bar{Q}_i^r(t)\right)\right\}$ converges in distribution to $(\bar{E}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), \bar{Q}(\cdot))$ as described by (3.23). □

3.3.7 Diffusion Approximation

3.3.7.1 Pre-limit process

From (3.4), (3.5), (3.9), (3.14), and (3.15), the diffusion scaled queuelength process can be written for $i = 1, 2$, $t \geq 0$, as

$$\begin{aligned} \hat{Q}_i^r(t) &= (\hat{E}_i^r(t) + \lambda_i^r rt) - (\hat{S}_i^r(\bar{T}_i^r(t)) + \mu_i r\bar{T}_i^r(t)) \\ &= \hat{E}_i^r(t) - \hat{S}_i^r(\bar{T}_i^r(t)) + r(\lambda_i^r t - \mu_i \bar{T}_i^r(t)). \end{aligned} \tag{3.34}$$

Expanding the last term in (3.34), we have

$$\begin{aligned} r(\lambda_i^r t - \mu_i \bar{T}_i^r(t)) &= \frac{r^2\lambda_i^r t - \mu_i r^2 \bar{T}_i^r(t)}{r} \\ &= \frac{(\lambda_i^r - \lambda_i)\, r^2 t + \lambda_i \int_0^{r^2 t} ds - \mu_i \int_0^{r^2 t} \Lambda_i(Q^r(s))ds}{r}. \end{aligned} \tag{3.35}$$

Considering four different types of states for the queuelength vector Q^r and sub-

stituting the corresponding values for $\Lambda_i(Q^r(\cdot))$ from (3.3), we can rewrite (3.35) as

$$\begin{aligned} r(\lambda_i^r t - \mu_i \bar{T}_i^r(t)) = \; & (\lambda_i^r - \lambda_i)\, rt \\ & + \frac{1}{r}\Bigg[(\lambda_i - \mu_i c_i)\int_0^{r^2 t} 1_{\{Q^r(s)>0\}} ds \\ & + (\lambda_i - \mu_i c_i^*)\int_0^{r^2 t} 1_{\left\{Q_i^r(s)>0; Q_j^r(s)=0, j\neq i\right\}} ds \\ & + \lambda_i \int_0^{r^2 t} 1_{\left\{Q_i^r(s)=0; Q_j^r(s)>0, j\neq i\right\}} ds \\ & + \lambda_i \int_0^{r^2 t} 1_{\left\{Q_j^r(s)=0 \text{ for all } j\right\}} ds\Bigg]. \end{aligned} \tag{3.36}$$

Define for $t \geq 0$,

$$\begin{aligned} \hat{U}_i^r(t) &\triangleq \frac{1}{r}\int_0^{r^2 t} 1_{\left\{Q_i^r(s)=0; Q_j^r(s)>0, j\neq i\right\}} ds \\ &= r\int_0^t 1_{\left\{\hat{Q}_i^r(s)=0; \hat{Q}_j^r(s)>0, j\neq i\right\}} ds, \quad i = 1, 2, \end{aligned} \tag{3.37}$$

$$\hat{Z}^r(t) \triangleq \frac{1}{r}\int_0^{r^2 t} 1_{\left\{Q_j^r(s)=0 \text{ for all } j\right\}} ds = r\int_0^t 1_{\left\{\hat{Q}_j^r(s)=0 \text{ for all } j\right\}} ds. \tag{3.38}$$

Then, using the fact that $\lambda_i = \mu_i c_i$ and combining (3.34)–(3.38), we obtain for $i = 1, 2$, $t \geq 0$,

$$\hat{Q}_i^r(t) = \hat{X}_i^r(t) + \lambda_i \hat{U}_i^r(t) + (\lambda_i - \mu_i c_i^*)\hat{U}_j^r(t) + \lambda_i \hat{Z}^r(t), \tag{3.39}$$

where $j = i + 1 \pmod 2$ and

$$\hat{X}_i^r(t) = \hat{E}_i^r(t) - \hat{S}_i^r(\bar{T}_i^r(t)) + (\lambda_i^r - \lambda_i) rt. \tag{3.40}$$

This can be expressed in vector form for $t \geq 0$ as

$$\hat{Q}^r(t) = \hat{X}^r(t) + \begin{bmatrix} \lambda_1 & \lambda_1 - \mu_1 c_1^* \\ \lambda_2 - \mu_2 c_2^* & \lambda_2 \end{bmatrix} \hat{U}^r(t) + \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \hat{Z}^r(t). \tag{3.41}$$

Define the *reflection matrix* R as

$$R \triangleq \begin{bmatrix} 1 & \frac{\lambda_1 - \mu_1 c_1^*}{\lambda_2} \\ \frac{\lambda_2 - \mu_2 c_2^*}{\lambda_1} & 1 \end{bmatrix} \tag{3.42}$$

and for $i \in \{1, 2\}$, $j \neq i$ and $t \geq 0$, define

$$\hat{Y}_i^r(t) \triangleq \lambda_i \left(\hat{U}_i^r(t) + \frac{c_i^* c_j}{c_1^* c_2 + c_1 c_2^* - c_1^* c_2^*} \hat{Z}^r(t) \right). \tag{3.43}$$

Then, (3.41) can be written as

$$\hat{Q}^r(t) = \hat{X}^r(t) + R\hat{Y}^r(t),\ t \geq 0. \tag{3.44}$$

Note that $c_1^* c_2 + c_1 c_2^* - c_1^* c_2^* > 0$ (from Lemma 3.3.2) and $\hat{Y}_i^r$, $i = 1, 2$, can increase only when the corresponding $\hat{Q}_i^r = 0$.

We next state and prove the C-tightness of the sequence of processes $\{\hat{X}^r(\cdot)\}$ which will be used in proving the C-tightness of the sequence of diffusion-scaled queue-length processes $\{\hat{Q}^r(\cdot)\}$.

Lemma 3.3.10. *The sequence $\{\hat{X}^r(\cdot)\}$ converges in distribution to a Brownian motion with diagonal covariance matrix $\Gamma \triangleq \mathrm{diag}(\lambda_1(\alpha_1^2 + \beta_1^2), \lambda_2(\alpha_2^2 + \beta_2^2))$ and drift vector $\theta \triangleq (\theta_1, \theta_2)$, that starts from the origin.*

Proof. Let $\hat{\theta}^r(t) \triangleq r(\lambda^r - \lambda)t$, $t \geq 0$. By combining Proposition 3.3.6, Lemma 3.3.9 and Assumption 3.3.4, we have that the sequence of processes $\left\{ \left(\hat{E}^r(\cdot), \hat{S}^r(\cdot), \bar{T}^r(\cdot), \hat{\theta}^r(\cdot) \right) \right\}$ converges in distribution to $(B_E(\cdot), B_S(\cdot), c(\cdot), \theta(\cdot))$ where $B_E(\cdot)$ and $B_S(\cdot)$ are indepen-

dent two-dimensional driftless Brownian motions starting from the origin with covariance matrices Γ_E and Γ_S respectively, $c(t) = ct$, $\theta(t) = \theta t$ for all $t \geq 0$.

Then from (3.40), using the random time change theorem, $\{\hat{X}^r(\cdot)\}$ converges in distribution to a two-dimensional Brownian motion with diagonal covariance matrix $\mathrm{diag}(\lambda_1\alpha_1^2 + \mu_1 c_1 \beta_1^2, \lambda_2\alpha_2^2 + \mu_2 c_2 \beta_2^2) = \mathrm{diag}(\lambda_1(\alpha_1^2 + \beta_1^2), \lambda_2(\alpha_2^2 + \beta_2^2))$ (since $\lambda_i = \mu_i c_i$ for $i = 1, 2$), drift vector (θ_1, θ_2) and starting point $(0, 0)$. □

3.3.7.2 Limit Theorem

We next discuss the properties of the reflection matrix R and use these properties to state and prove the limit theorem, which is the main result of this section.

Define

$$\tilde{P} \triangleq I - R = \begin{bmatrix} 0 & \frac{\mu_1 c_1^* - \lambda_1}{\lambda_2} \\ \frac{\mu_2 c_2^* - \lambda_2}{\lambda_1} & 0 \end{bmatrix} \tag{3.45}$$

where I is the 2×2 identity matrix. For $i = 1, 2$, $\mu_i c_i^* - \lambda_i > 0$, since $\mu_i c_i = \lambda_i$ and $c_i < c_i^*$. Thus all of the entries of $\tilde{P}$ are nonnegative. We next show that the matrix R satisfies the HR condition described in Section 3.3.1.1.

Lemma 3.3.11. *The reflection matrix R satisfies the HR condition.*

Proof. Since $\tilde{P}$ has zeros on the diagonal and all of its entries are nonnegative, it suffices to show that $\tilde{P}$ has spectral radius strictly less than 1. The eigenvalues of $\tilde{P}$ are the solutions of the equation

$$x^2 - \frac{(\mu_1 c_1^* - \lambda_1)(\mu_2 c_2^* - \lambda_2)}{\lambda_1 \lambda_2} = 0. \tag{3.46}$$

Using $\lambda_i = c_i \mu_i$, $i = 1, 2$, and the fact that $c_1^* > c_1$, $c_2^* > c_2$, we have

$$x = \pm \sqrt{\left(\frac{c_1^*}{c_1} - 1\right)\left(\frac{c_2^*}{c_2} - 1\right)}. \tag{3.47}$$

Thus the spectral radius of $\tilde{P}$ is strictly less than 1 iff $(c_1^* - c_1)(c_2^* - c_2) < c_1 c_2$. By assumption, $c_1 + c_2 > c_1^*, c_2^*$. Thus $0 < (c_1^* - c_1) < c_2$ and $0 < (c_2^* - c_2) < c_1$. So $(c_1^* - c_1)(c_2^* - c_2) < c_1 c_2$ and the spectral radius of $\tilde{P}$ is strictly less than one. Thus R satisfies the HR condition. □

We next state and prove the main result of this section.

Theorem 3.3.12 (Main Theorem)**.** *The diffusion-scaled queuelength process $\hat{Q}^r(\cdot)$ converges in distribution to an SRBM, i.e., $\hat{Q}^r \Rightarrow \hat{Q}$ as $r \to \infty$, where $\hat{Q}$ is an SRBM associated with the data $(\mathbb{R}^2_+, \theta, \Gamma, R)$ that starts from the origin.*

Proof. Recall the results on the Skorokhod problem stated in Section 3.3.1.1. For each $r > 0$, $\hat{X}^r(\cdot)$ has paths in $\mathbb{D}^2_+$ and $\hat{Q}^r, \hat{X}^r, \hat{Y}^r$ satisfy (3.44). By definition, $\hat{Q}^r(\cdot)$ has paths in $\mathbb{R}^2_+$. Furthermore, a.s., $\hat{Y}^r(0) = 0$, $\hat{Y}^r(\cdot)$ is nonnegative, non-decreasing, continuous and for $i = 1, 2$, $\hat{Y}_i^r(\cdot)$ increases only when $\hat{Q}_i^r(\cdot) = 0$, i.e., $\int_{(0,\infty)} \hat{Q}_i^r(s) d\hat{Y}_i^r(s) = 0$. Thus, a.s., $(\hat{Q}^r(\cdot), \hat{Y}^r(\cdot))$ is a solution of the Skorokhod problem for $\hat{X}^r(\cdot)$ with respect to R. Since R satisfies the HR condition, by Proposition 3.3.1, $(\hat{Q}^r(\cdot), \hat{Y}^r(\cdot)) = \Phi(\hat{X}^r(\cdot))$ a.s. where the mapping $\Phi : \mathbb{D}^2_+ \to \mathbb{D}^4_+$ is continuous. By Lemma 3.3.10, the sequence $\{\hat{X}^r(\cdot)\}$ converges in distribution as $r \to \infty$ to a Brownian motion with drift θ and covariance matrix Γ that starts from the origin. Then by the continuous mapping theorem, $\left\{\left(\hat{Q}^r(\cdot), \hat{X}^r(\cdot), \hat{Y}^r(\cdot)\right)\right\}$ converges in distribution as $r \to \infty$ to $\left(\hat{Q}(\cdot), \hat{X}(\cdot), \hat{Y}(\cdot)\right)$ where $(\hat{Q}(\cdot), \hat{Y}(\cdot)) = \Phi(\hat{X})$ is a.s. the unique solution of the Skorokhod problem for $\hat{X}(\cdot)$ with respect to R. Here $\hat{Q}$ is a representation of the SRBM associated with the data $(\mathbb{R}^2_+, \theta, \Gamma, R)$ that starts from the origin. □

3.3.7.3 Properties of the Limit Process

The SRBM structure of $\hat{Q}$ enables us to use results from the theory of SRBMs to state some properties of the limit of the diffusion-scaled queuelength processes.

Time Spent at the Origin An important quantity for a queueing system is the time that the system is idle. It can be shown that almost surely $\hat{Q}$ spends zero Lebesgue time at the origin. Stated formally,

Proposition 3.3.13. *Almost surely, the Lebesgue measure of the time spent by $\hat{Q}$ at $(0,0)$ is zero.*

Proof. Varadhan and Williams [41] have shown that when $\theta = 0$ and the covariance matrix is the identity matrix, the associated SRBM spends zero Lebesgue time at the origin almost surely. By a scaling of the coordinates, we may conclude that the SRBM with drift $\theta = 0$ and a diagonal covariance matrix, spends zero Lebesgue time at the origin almost surely. Note that with the scaling, we end up applying a similarity transformation to the R matrix which does not alter the fact that the HR condition is satisfied. Then, by a Girsanov transformation (see [9, §9.4]) to change the drift of the driving Brownian motion, it follows that the Lebesgue measure of the time spent by $\hat{Q}$ at the origin is zero almost surely. □

Stationary Distribution Harrison and Williams [17] have shown that there is a stationary distribution for the SRBM if and only if $R^{-1}\theta < 0$ where the inequality is understood to hold component by component. As an illustration, a situation in which this condition is satisfied is depicted in Figure 3.3 with $\theta = (-1, 0)$ and $R = \begin{bmatrix} 1 & -\gamma_1 \\ -\gamma_2 & 1 \end{bmatrix}$ where $\gamma_1 = \frac{\mu_1 c_1^* - \lambda_1}{\lambda_2}$ and $\gamma_2 = \frac{\mu_2 c_2^* - \lambda_2}{\lambda_1}$. For two-dimensional SRBMs, Avram et al. [1] studied a variational problem (VP) arising from the study of SRBMs. The optimal value of the VP describes the tail behavior of the stationary distribution and the corresponding optimal paths characterize how certain rare events are most likely to occur. Dai and Harrison [11] have identified a numerical procedure for computing quantities associated with the stationary distribution for a class of SRBMs. This can be used to numerically approximate the mean of the stationary distribution of the SRBM that is a diffusion approximation of our system.

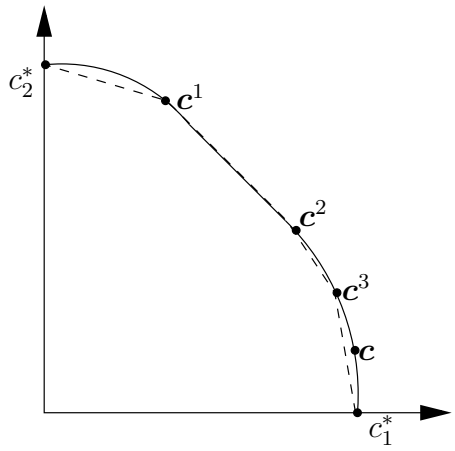


Figure 3.2 *The solid curve indicates the capacity surface while the surface of the system stability region is shown by the dashed line.*

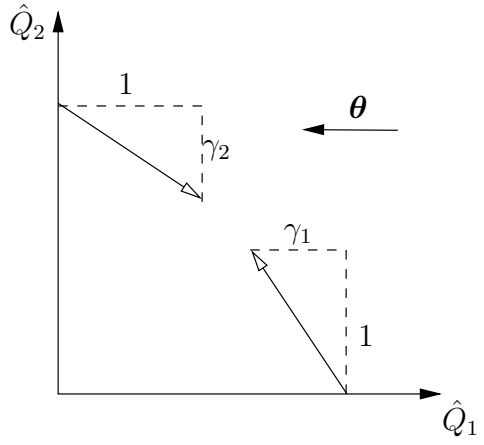


Figure 3.3 *Directions of reflection and drift for an example of an SRBM with* $\gamma_1 = \frac{\mu_1 c_1^* - \lambda_1}{\lambda_2}$, $\gamma_2 = \frac{\mu_2 c_2^* - \lambda_2}{\lambda_1}$, *and* $\theta = (-1, 0)$.

3.4 Systems with Arbitrary Number of Users: Workload

3.4.1 Notation and Preliminaries

We will use the following notation throughout this section. We will use $\mathcal{N}$ to denote the set $\{1, 2, \ldots, N\}$ where N is a finite positive integer, $\mathcal{K}$ to denote an arbitrary subset of $\mathcal{N}$, and $\mathcal{K}^c$ to denote the complement of $\mathcal{K}$ in $\mathcal{N}$. We will use $\mathcal{P}(\mathcal{A})$ to indicate the power set of an arbitrary set $\mathcal{A}$. We will use $|\mathcal{A}|$ to denote the cardinality of the set $\mathcal{A}$. The symbol $1_{\mathcal{A}}$ denotes the indicator function of a set $\mathcal{A}$, i.e., $1_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$ and $1_{\mathcal{A}}(x) = 0$ if $x \notin \mathcal{A}$.

Let $\mathbb{Z}$ denote the set of all integers, $\mathbb{Z}_+$ the set of all non-negative integers, $\mathbb{R}$ denote the set of real numbers, and $\mathbb{R}_+$ denote the set of non-negative real numbers, which is also denoted by $[0, \infty)$. The symbol $\mathbb{R}^N$ will denote N-dimensional Euclidean space and the positive orthant in this space will be denoted by $\mathbb{R}^N_+ = \{x \in \mathbb{R}^N : x_i \geq 0 \text{ for all } i \in \mathcal{N}\}$. All vectors and matrices in this section are assumed to have real valued entries. Let $0 = (0, 0, \ldots, 0) \in \mathbb{R}^N_+$. We denote the inner product on $\mathbb{R}^N$ by $\langle \cdot, \cdot \rangle$, i.e., $\langle x, y \rangle = \sum_{i=1}^N x_i y_i$, for $x, y \in \mathbb{R}^N$. The usual Euclidean norm on $\mathbb{R}^N$ will be denoted by $\|\cdot\|$ so that $\|x\| = \sqrt{\langle x, x \rangle} = \left(\sum_{i=1}^N x_i^2\right)^{1/2}$ for $x \in \mathbb{R}^N$. Let $\mathcal{B}(\mathbb{R}^N)$ denote the σ-algebra of Borel subsets of $\mathbb{R}^N$. For any non-empty set $\mathcal{K} \subseteq \mathcal{N}$ and any $x \in \mathbb{R}^N$, $x_{\mathcal{K}}$ will denote the vector whose components are those of x with indices in $\mathcal{K}$. Let $e_{\mathcal{N}} \in \mathbb{R}^N$ denote the vector whose entries are all 1. For $x, y \in \mathbb{R}^N$, we shall use $x \wedge y$ to denote the vector whose i-th component is the minimum of x_i and y_i for each $i \in \mathcal{N}$. All vector inequalities are understood to hold componentwise. For $a \in \mathbb{R}^N$, we shall use $\mathrm{diag}(a)$ to denote the $N \times N$ diagonal matrix whose diagonal entries are given by the entries in a. We will let $(\cdot)'$ denote transpose. For any set $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, we define the face $F_{\mathcal{K}}$ by

$$F_{\mathcal{K}} \triangleq \{x \in \mathbb{R}^N_+ : x_i = 0 \text{ for all } i \in \mathcal{K}\}. \tag{3.48}$$

For example $F_{\mathcal{N}} = \{0\}$, the set consisting of the origin in $\mathbb{R}^N$. When $\mathcal{K} = \{i\}$ for $i \in \mathcal{N}$,

we write F_i in place of $F_{\{i\}}$ sometimes. We define the index set of any point $x \in \mathbb{R}^N_+$ by

$$\mathcal{K}(x) \triangleq \{i \in \mathcal{N} : x_i = 0\} \tag{3.49}$$

with the convention that $\mathcal{K}(w) = \emptyset$ if $w > 0$. A domain in $\mathbb{R}^N$ is an open connected subset of $\mathbb{R}^N$. For each continuously differentiable real-valued function f defined on some non-empty domain $S \subseteq \mathbb{R}^N$, $\nabla f(x)$ is the gradient of f at $x \in S$:

$$(\nabla f(x))_i = \frac{\partial f}{\partial x_i}(x),\ i = 1, 2, \ldots, N. \tag{3.50}$$

For any set $S \subseteq \mathbb{R}^N$, we write $\overline{S}$ for the closure of S, S^o for the interior of S, and $\partial S = \overline{S} \setminus S^o$.

All stochastic processes used in this section will be assumed to have paths that are right continuous with finite left limits (r.c.l.l.). We denote by $\mathbb{D}^N$ the space of r.c.l.l. functions from $[0, \infty)$ into $\mathbb{R}^N$ and we endow this space with the usual Skorokhod J_1-topology (see Ethier and Kurtz [12, Chapter 3, Section 5]) which makes it a Polish space. We denote by $\mathbb{C}^N$ the space of continuous functions from $[0, \infty)$ into $\mathbb{R}^N$, also endowed with the Skorokhod J_1-topology under which convergence of elements in $\mathbb{C}^N$ is equivalent to uniform convergence on compact time intervals. We endow $\mathbb{D}^N$ (or $\mathbb{C}^N$) with the Borel σ-algebra induced by the Skorokhod J_1-topology and denote this σ-algebra by $\mathcal{M}^N$. The abbreviation *u.o.c.* will stand for *uniformly on compacts* and will be used to indicate that a sequence of functions in $\mathbb{D}^N$ (or $\mathbb{C}^N$) is converging uniformly on compact time intervals to a limit in $\mathbb{D}^N$ (or $\mathbb{C}^N$). An N-dimensional process is a measurable function from a probability space into $(\mathbb{D}^N, \mathcal{M}^N)$. Consider $W^1, W^2, \ldots, W$, each of which is an N-dimensional process (possibly defined on different probability spaces). The sequence $\{W^n\}_{n=1}^{\infty}$ is said to be *tight* if the probability measures induced by the sequence $\{W^n\}_{n=1}^{\infty}$ on $(\mathbb{D}^N, \mathcal{M}^N)$ form a tight sequence, i.e., they form a weakly relatively compact sequence in the space of probability measures on $(\mathbb{D}^N, \mathcal{M}^N)$. The notation "$W^n \Rightarrow W$" will

mean that "W^n converges in distribution to W as $n \to \infty$". The sequence of processes $\{W^n\}_{n=1}^{\infty}$ is called *C-tight* if it is tight and if each weak limit point (obtained as a weak limit along a subsequence) is in $\mathbb{C}^N$ almost surely.

A triple $(\Omega, \mathcal{F}, \{\mathcal{F}_t, t \geq 0\})$ will be called a filtered space if Ω is a set, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and $\{\mathcal{F}_t, t \geq 0\}$ is an increasing family of sub-σ-algebras of $\mathcal{F}$, i.e., a filtration. From now on, we will write a filtration $\{\mathcal{F}_t, t \geq 0\}$ as simply $\{\mathcal{F}_t\}$. If P is a probability measure on $(\Omega, \mathcal{F})$, then $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ is called a filtered probability space. An N-dimensional process $X = \{X(t), t \geq 0\}$ defined on $(\Omega, \mathcal{F}, P)$ is called $\{\mathcal{F}_t\}$-adapted if for each $t \geq 0$, $X(t) : \Omega \to \mathbb{R}^N$ is measurable when Ω is endowed with the σ-algebra $\mathcal{F}_t$ and $\mathbb{R}^N$ has the usual Borel σ-algebra $\mathcal{B}(\mathbb{R}^N)$, and X is said to be a continuous process if its sample paths are continuous P-a.s.

3.4.2 Communication System Model

In this subsection we specify the communication system under consideration. We consider a cellular wireless network where base stations cooperate over noise-free infinite capacity links. We do not make any distinction between a single-cell cellular system having multiple base-station antennas and the traditional cellular system with cooperating single-antenna base stations. Here by cooperation we mean that the base stations can perform joint beamforming and/or power control but there is a constraint on the total power that the base stations can share. We do not make any assumptions about the number of receive antennas per user.

The downlink channel for such a system with N users can be modeled as an N-user MIMO Broadcast Channel (BC). We assume that the channel is fixed for all transmissions over the period of interest (some authors refer to this as a quasi-static channel). Moreover, we assume that the transmit end (with the cooperating base stations) has perfect channel state information (CSI).

Weingarten et al. [43] have shown that for such a system, Dirty Paper Coding

(DPC), introduced by Costa [10], achieves the capacity. Furthermore, the capacity region can be computed by using the duality of the MIMO Multiple Access Channel (MAC) and the MIMO BC [20] where the BC capacity region is obtained by taking the convex hull of the union over the set of capacity regions of the dual MIMO MACs such that the total MAC power is the same as the power in the BC.

For an N-user system, the capacity region is an N-dimensional closed convex set in $\mathbb{R}_+^N$ containing the origin where the convexity follows because of the convex hull operation. For an example of such a capacity region in the two-user case, see Figure 3.1.

At the transmit end, packets arrive for each user and are buffered before transmission. We assume that there is given a nominal average packet arrival rate (e.g. an estimate of the true average arrival rate). The ratio of the nominal average bit arrival rate for user i relative to that for user 1 is called the relative traffic rate and is denoted by κ_i (this is assumed to be strictly positive). This nominal relative traffic rate is specified in advance with the assumption that $\kappa_1 = 1$; thus, it is expected that, on average, the i-th user will have κ_i times as much data as user 1. The actual traffic rate may deviate from this nominal average rate due to estimation error and stochastic fluctuations. Naturally, when there is no data for one (or many) of the users to transmit (the corresponding queue for that(those) user(s) is empty), the data for the other users should be transmitted at the maximum possible rate for those users. We formally state these conditions in Section 3.4.3.

3.4.3 Queueing Analogue

In this subsection, we develop a queueing analogue for the system described in Section 3.4.2. To this end, we describe the physical structure, and the stochastic primitives specifying the packet arrivals and sizes. We formulate dynamic equations satisfied by the workload process in terms of the stochastic primitives and the policy or service discipline to be used with this system.

3.4.3.1 Physical Structure

A queueing model describing our communication system has N queues in parallel where each queue buffers packets intended for a given user. We assume that each of the queues has infinite buffer capacity. The queues are served by a single server corresponding to a base station with multiple cooperating antennas.

3.4.3.2 Stochastic Primitives

We assume that the system starts empty and that there is an N-dimensional packet arrival process $E = \{(E_1(t), E_2(t), \ldots, E_N(t)), t \geq 0\}$ where $E_i(t)$ is the number of packets that have arrived to the i-th queue in $(0, t]$. (Here E is used to indicate that the arrivals are *exogenous*.) For $i \in \mathcal{N}$, $E_i(\cdot)$ is assumed to be a (non-delayed) renewal process defined from a sequence of strictly positive independent and identically distributed (i.i.d.) random variables $\{u_i(k), k = 1, 2, \ldots\}$, where for $k = 1, 2, \ldots$, the random variable $u_i(k)$ denotes the time between the arrival of the $(k-1)$-st and the k-th packet to the i-th queue (where the 0-th arrival occurs at time 0). Each $u_i(k)$, $k = 1, 2, \ldots$, is assumed to have finite mean $1/\lambda_i \in (0, \infty)$ and finite squared coefficient of variation (variance divided by the mean squared) $\alpha_i^2 \in (0, \infty)$. Then

$$E_i(t) = \max\left\{n \geq 0 : \sum_{j=1}^{n} u_i(j) \leq t\right\},\ i \in \mathcal{N},\ t \geq 0, \tag{3.51}$$

where a sum up to $n = 0$ is defined to be zero. The packet lengths (in bits) for the successive arrivals to the i-th queue are given by a sequence of strictly positive i.i.d. random variables $\{v_i(k), k = 1, 2, \ldots\}$ with average packet length $m_i = 1/\mu_i \in (0, \infty)$ and squared coefficient of variation $\beta_i^2 \in (0, \infty)$. We assume that all interarrival and service time processes are mutually independent. For $i \in \mathcal{N}$ and $n \in \mathbb{Z}_+$, we define

$$V_i(n) \triangleq \sum_{j=1}^{n} v_i(j). \tag{3.52}$$

We refer to the processes $E(\cdot)$ and $V(\cdot)$ as *stochastic primitives* for our system model. For convenience, to avoid the need to consider exceptional null sets, we assume without loss of generality that $E_i(t) < \infty$ for all $t \geq 0$ and $E_i(t) \to \infty$ as $t \to \infty$ for each $i \in \mathcal{N}$, surely.

3.4.3.3 Workload Process

For $i \in \mathcal{N}$, the workload $W_i(t)$ of the i-th queue at time $t \geq 0$ is given by

$$\begin{aligned} W_i(t) &\triangleq \sum_{j=1}^{E_i(t)} v_i(j) - T_i(t) \\ &= V_i(E_i(t)) - T_i(t), \end{aligned} \tag{3.53}$$

where $T_i(t)$ is the cumulative amount of service (measured in bits) given to the i-th queue up to time t. We next describe the service discipline which, in turn, specifies the functional form of $T_i(\cdot)$.

3.4.3.4 Service Discipline

When service is given to a queue, it goes to the packet at the head of the line, where it is assumed that packets are queued in the order of their arrival with the packet that arrived the longest time ago being at the head of the line. A vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ indicates the rates (in bits per second) of serving the N queues, i.e., σ_1 is the rate for queue 1, σ_2 is the rate for queue 2, and so on. The service rate for each queue is a very simple function of the vector of workloads. Given a workload of $w = (w_1, w_2, \ldots, w_N)$, the set of indices for the empty queues is the index set $\mathcal{K}(w)$, as defined by (3.49). The

rates $\sigma = \Lambda(w)$ are given by the function[3] $\Lambda : \mathbb{R}_+^N \to \mathbb{R}_+^N$ defined by

$$\Lambda(w) \triangleq c^{\mathcal{K}(w)} \tag{3.54}$$

where $c^{\mathcal{K}}$ is a fixed vector for each $\mathcal{K} \subseteq \mathcal{N}$ with $c_i^{\mathcal{K}} = 0$ if $i \in \mathcal{K}$ (corresponding to the fact that an empty queue should not be served) and $c_i^{\mathcal{K}} > 0$ if $i \notin \mathcal{K}$. The vector of service rates $c^{\mathcal{K}}$ is chosen such that it lies on the boundary of the capacity region and the service rate for each of the users with positive workload is related by the relative traffic rate as described below. Recall, from Section 3.4.2, $(\kappa_i, i \in \mathcal{N})$ is the given vector of nominal relative traffic rates. For all $\mathcal{K} \subsetneqq \mathcal{N}$, the non-zero entries of the service rate vector $c^{\mathcal{K}}$ are chosen such that

$$\frac{c_i^{\mathcal{K}}}{\kappa_i} = \frac{c_j^{\mathcal{K}}}{\kappa_j} \tag{3.55}$$

whenever $i, j \in \mathcal{K}^c$, and $\sum_i c_i^{\mathcal{K}}$ is as large as possible while still keeping $c^{\mathcal{K}}$ in the capacity region. (We make the non-degeneracy assumption that the capacity region is such that we can choose $c_i^{\mathcal{K}} > 0$ for all $i \in \mathcal{K}^c$.) When all of the queues are non-empty ($\mathcal{K} = \emptyset$), the service rate vector, $c^{\emptyset}$, lies on the boundary of the capacity region and for all $i \in \mathcal{N}$, $c_i^{\emptyset} = \kappa_i c_1^{\emptyset}$, i.e., $c^{\emptyset}$ is in the direction of the vector κ and is the furthest point along that direction which lies in the capacity region (see Figure 3.4 for an example of the capacity region and the service rates for a two-user system).

The following condition, which corresponds to the fact that cooperation results in an increase in the sum of the service rates, is assumed to be satisfied by the $c^{\mathcal{K}}$'s:

$$\sum_{i \in \mathcal{N} \setminus \mathcal{L}} c_i^{\mathcal{L}} > \sum_{j \in \mathcal{N} \setminus \mathcal{K}} c_j^{\mathcal{K}} \text{ for all } \mathcal{L} \subsetneqq \mathcal{K} \subseteq \mathcal{N}. \tag{3.56}$$

Moreover, the service rate for a fixed queue is assumed to be reduced as more queues are

[3]We only need $\Lambda(\cdot)$ defined on $\mathbb{Z}_+^N$ for the moment, but we extend the domain of $\Lambda(\cdot)$ to $\mathbb{R}_+^N$ so that later when we rescale the workload process, $\Lambda(\cdot)$ is well-defined for the rescaled process.

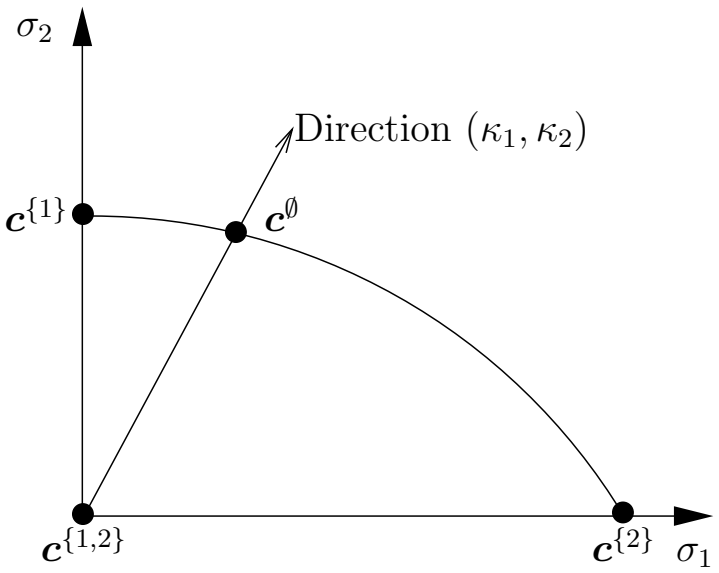


Figure 3.4 *An example of the capacity region for a two-user system. Service rate* $c^{\{1,2\}} = (0,0)$*,* $c^{\{2\}}$ *is along the direction* $(\kappa_1, 0)$ *and* $c^{\{1\}}$ *is along the direction* $(0, \kappa_2)$*.*

served concurrently. Therefore

$$\text{for all } \mathcal{L} \subsetneqq \mathcal{K} \subseteq \mathcal{N},\ c_i^{\mathcal{L}} < c_i^{\mathcal{K}} \text{ for all } i \in \mathcal{N} \setminus \mathcal{K}. \tag{3.57}$$

For example, $c_i^{\emptyset} < c_i^{\{j\}}$ for all $i \neq j$, $i, j \in \mathcal{N}$.

Our model is a single server, N-class queueing system where the N classes correspond to the N queues (users). The following scaling property of $\Lambda(\cdot)$ is a mathematical statement of the property of the scheduling policy that the amount of service given to the queues in any state does not change when all workloads are multiplied by the same factor.

Lemma 3.4.1. *For any* $w \in \mathbb{R}_+^N$ *and* $a > 0$, $\Lambda(aw) = \Lambda(w)$.

Proof. The proof follows easily from the fact that $\Lambda(\cdot)$ depends only on which queues are empty and these are unchanged by the positive scalar factor a. □

For $t \geq 0$, $i \in \mathcal{N}$, we can now give an explicit expression for $T_i(t)$ as

$$\begin{aligned} T_i(t) &\triangleq \int_0^t \Lambda_i(W(s))\, ds \\ &= \sum_{\mathcal{K} \subseteq \mathcal{N}} c_i^{\mathcal{K}} \int_0^t 1_{\{\mathcal{K}(W(s))=\mathcal{K}\}}\, ds. \end{aligned} \tag{3.58}$$

In fact, $c^{\mathcal{N}} = 0$ and so the sum could be reduced to that over $\mathcal{K} \subsetneqq \mathcal{N}$, including $\mathcal{K} = \emptyset$.

3.4.4 Heavy Traffic Assumptions

We wish to consider the behavior of the queueing system when it is heavily loaded. (Kelly and Laws [23] have argued that in this regime "important features of good control policies are displayed in sharpest relief".) For this purpose one may regard a given system as a member of a sequence of systems approaching the heavy traffic limit. To obtain a reasonable approximation, the workload process is rescaled using diffusion scaling. This corresponds to viewing the system over long intervals of time of order r^2 (where r will tend to infinity in the asymptotic limit) and regarding a single packet as only having a small contribution to the overall congestion level, where this is quantified to be of order $1/r$. Formally, we consider a sequence of systems indexed by r, where r tends to infinity through a sequence of values in $(0, \infty)$. These systems all have the same basic structure as that described in the last section; however, the arrival rates may vary with r. We assume that the interarrival times for the system indexed by r are given for each $i \in \mathcal{N}$, $k = 1, 2, \ldots$, by

$$u_i^r(k) = \frac{1}{\lambda_i^r} \breve{u}_i(k) \tag{3.59}$$

where the $\breve{u}_i(k)$ do not depend on r, have mean one and squared coefficient of variation α_i^2. The packet lengths $\{v_i(k)\}_{k=1}^{\infty}$, $i \in \mathcal{N}$, do not change with r. [The above structure is convenient for allowing the sequence of systems to approach heavy traffic by simply changing arrival rates and keeping the underlying sources of variability $\breve{u}_i(k)$ and $v_i(k)$ fixed as r varies. This type of set-up has been used previously by others in treating heavy-

traffic limits (see, e.g., Peterson [31] and Bell and Williams [2]). For a first pass, the reader may like to simply choose $\lambda_i^r = \lambda_i$ for all r.] All processes and parameters that depend on r will from now have a superscript of r appended. The nominal relative traffic rate and the service rates $\{c^{\mathcal{K}}, \mathcal{K} \subseteq \mathcal{N}\}$ are assumed fixed throughout and do not vary with r. We define $\lambda_i \triangleq \mu_i c_i^{\emptyset}$ for $i = 1, 2, \ldots, N$.

Assumption 3.4.2 (Heavy Traffic Assumption)**.** *There exists $\theta \in \mathbb{R}_+^N$ such that for each $i \in \mathcal{N}$,*

$$r(\lambda_i^r - \lambda_i)m_i \to \theta_i \text{ as } r \to \infty. \tag{3.60}$$

We may regard $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ as a nominal average packet arrival rate used to set the service rates,

$$(c_1^{\emptyset}, c_2^{\emptyset}, \ldots, c_N^{\emptyset}),$$

for the scheduling policy. The r-th system has a perturbed average packet arrival rate λ^r for which the average bit arrival rate b^r ($b_i^r = \lambda_i^r m_i, i \in \mathcal{N}$) is close to $(c_1^{\emptyset}, c_2^{\emptyset}, \ldots, c_N^{\emptyset})$.

3.4.5 Scaling, Standard Limit Theorems, and Parameters

3.4.5.1 Scaling

Fluid (or functional law of large numbers) scaling is indicated by placing a bar over a process. For $r > 0$, $i \in \mathcal{N}$, and $t \geq 0$, we define

$$\bar{E}_i^r(t) \triangleq r^{-2} E_i^r(r^2 t), \tag{3.61}$$

$$\bar{V}_i^r(t) \triangleq r^{-2} V_i^r(r^2 t), \tag{3.62}$$

$$\bar{T}_i^r(t) \triangleq r^{-2} T_i^r(r^2 t), \tag{3.63}$$

$$\bar{W}_i^r(t) \triangleq r^{-2} W_i^r(r^2 t). \tag{3.64}$$

Diffusion (or functional central limit theorem) scaling is indicated by placing a hat over a process. For $r > 0$, $i \in \mathcal{N}$, and $t \geq 0$, we define

$$\hat{W}_i^r(t) \triangleq \frac{W_i^r(r^2 t)}{r}. \tag{3.65}$$

To apply diffusion-scaling to the primitive stochastic processes $E^r(\cdot)$ and $V(\cdot)$ (note that $V(\cdot)$ does not depend on r), we must center them before scaling. Accordingly, for $r > 0$, $i \in \mathcal{N}$, and $t \geq 0$, we define

$$\hat{E}_i^r(t) \triangleq \frac{1}{r}\left(E_i^r(r^2 t) - \lambda_i^r r^2 t\right) \tag{3.66}$$

and

$$\hat{V}_i^r(t) \triangleq \frac{1}{r}\left(V_i(r^2 t) - m_i r^2 t\right). \tag{3.67}$$

3.4.5.2 Functional Limit Theorems for Stochastic Primitives

We will use the following functional central limit theorem (FCLT) for the stochastic primitives in the sequel.

Proposition 3.4.3 (FCLT)**.** *The diffusion-scaled processes* $(\hat{E}^r(\cdot), \hat{V}^r(\cdot))$ *jointly converge in distribution to* $(B_E(\cdot), B_V(\cdot))$ *as* $r \to \infty$*, i.e.,*

$$(\hat{E}^r(\cdot), \hat{V}^r(\cdot)) \Rightarrow (B_E(\cdot), B_V(\cdot)) \text{ as } r \to \infty, \tag{3.68}$$

where $B_E(\cdot)$ *and* $B_V(\cdot)$ *are independent* N*-dimensional driftless Brownian motions starting from the origin with diagonal covariance matrices*

$$\Gamma_E \triangleq \mathrm{diag}(\lambda_1 \alpha_1^2, \lambda_2 \alpha_2^2, \ldots, \lambda_N \alpha_N^2) \tag{3.69}$$

and

$$\Gamma_V \triangleq \mathrm{diag}(m_1^2\beta_1^2, m_2^2\beta_2^2, \ldots, m_N^2\beta_N^2), \tag{3.70}$$

respectively.

Remark. As there is a single source of variability (not depending on r) for each of E_i^r, V_i, $i \in \mathcal{N}$, only the finiteness of the second moments of $\breve{u}_i(k)$ and $v_i(k)$ is required for the FCLT. Furthermore, since a Brownian motion is a continuous process, the weak-convergence of $(\hat{E}^r(\cdot), \hat{V}^r(\cdot))$ to a Brownian motion implies C-tightness of the sequence $\{(\hat{E}^r(\cdot), \hat{V}^r(\cdot))\}$.

Proof. By results of Iglehart and Whitt [18], a functional central limit theorem for the renewal counting process $E^r(\cdot)$ can be inferred from that for the partial sums of $\{u_i^r(k)\}_{k=1}^{\infty}$. Functional central limit theorems for the partial sums of $\{u_i^r(k)\}_{k=1}^{\infty}$ and $\{v_i(k)\}_{k=1}^{\infty}$ follow from Theorem 3.1 of Prokhorov [32]. The joint convergence follows from the independence of $E^r(\cdot)$ and $V(\cdot)$. □

As a corollary, we have the following functional law of large numbers (FLLN) for the stochastic primitives. For each $t \geq 0$, let $\lambda(t) \triangleq \lambda t$ and $m(t) \triangleq mt$.

Corollary 3.4.4 (FLLN)**.** *The fluid-scaled processes* $(\bar{E}^r(\cdot), \bar{V}^r(\cdot))$ *jointly converge in distribution to* $(\lambda(\cdot), m(\cdot))$ *as* $r \to \infty$*, i.e.,*

$$\left(\bar{E}^r(\cdot), \bar{V}^r(\cdot)\right) \Rightarrow (\lambda(\cdot), m(\cdot)) \text{ as } r \to \infty. \tag{3.71}$$

Remark. Here again, the weak-convergence of $(\bar{E}^r(\cdot), \bar{V}^r(\cdot))$ to a continuous process implies C-tightness of the sequence $\{(\bar{E}^r(\cdot), \bar{V}^r(\cdot))\}$.

Proof. Proposition 3.4.3 implies that

$$\left(\frac{1}{r}\hat{E}^r(\cdot), \frac{1}{r}\hat{V}^r(\cdot)\right) \Rightarrow (0,0) \text{ as } r \to \infty. \tag{3.72}$$

The desired result follows from this and the fact that $\lambda^r \to \lambda$ as $r \to \infty$ (see (3.60)). $\square$

3.4.5.3 Covariance and Reflection Matrices

We next define two matrices that are part of the data for the heavy traffic limit of the workload process. We first define the *covariance matrix* Γ as the $N \times N$ diagonal matrix whose i-th diagonal entry is

$$\Gamma_{ii} \triangleq \lambda_i m_i^2 (\alpha_i^2 + \beta_i^2),\ i \in \mathcal{N}. \tag{3.73}$$

We define the *reflection matrix* R as the $N \times N$ matrix whose entries are

$$R_{ij} = \begin{cases} 1 & \text{if } i = j \\ \frac{c_i^{\emptyset} - c_i^{\{j\}}}{c_j^{\emptyset}} & \text{if } i \neq j. \end{cases} \tag{3.74}$$

For example, when $N = 3$, the reflection matrix R is

$$R = \begin{bmatrix} 1 & \frac{c_1^{\emptyset} - c_1^{\{2\}}}{c_2^{\emptyset}} & \frac{c_1^{\emptyset} - c_1^{\{3\}}}{c_3^{\emptyset}} \\ \frac{c_2^{\emptyset} - c_2^{\{1\}}}{c_1^{\emptyset}} & 1 & \frac{c_2^{\emptyset} - c_2^{\{3\}}}{c_3^{\emptyset}} \\ \frac{c_3^{\emptyset} - c_3^{\{1\}}}{c_1^{\emptyset}} & \frac{c_3^{\emptyset} - c_3^{\{2\}}}{c_2^{\emptyset}} & 1 \end{bmatrix}. \tag{3.75}$$

The matrix R defined by (3.74) has a special structure in that it satisfies the Harrison-Reiman (HR) condition [16]. We use this structure in proving the convergence of the diffusion-scaled workload process.

Definition 3.4.1 (Harrison-Reiman (HR) Condition)**.** *An $N \times N$ matrix R satisfies the HR condition if $R = I - Q$, where I is the $N \times N$ identity matrix, and the $N \times N$ matrix Q has zeros along the diagonal, all of the entries of Q are nonnegative, and Q has spectral radius strictly less than one.*

Remark. When $R = I - Q$ where Q has zeros on the diagonal and the entries of Q are

nonnegative, the HR condition is equivalent to the requirement that R is a non-singular M-matrix. Such matrices are discussed for example in Berman and Plemmons [3, Chapter 6].

Lemma 3.4.5. *The reflection matrix R satisfies the HR condition.*

Proof. It is easy to see that an $N \times N$ matrix R satisfies the HR condition if $R = I - P'$ where I is the $N \times N$ identity matrix, P is an $N \times N$ matrix whose diagonal entries are zero, and whose off-diagonal entries are nonnegative and such that each row-sum is strictly less than 1. To show that R has this form, note that the diagonal entries of R are all equal to 1 and from the condition (3.57), the off-diagonal entries are all negative. Therefore it suffices to show that the sum of each column of R is strictly greater than 0. But the sum of the j-th column of R is

$$1 + \sum_{i \in \mathcal{N} \setminus \{j\}} \frac{c_i^{\emptyset} - c_i^{\{j\}}}{c_j^{\emptyset}} = \frac{1}{c_j^{\emptyset}} \left(\sum_{i \in \mathcal{N}} c_i^{\emptyset} - \sum_{i \in \mathcal{N} \setminus \{j\}} c_i^{\{j\}} \right) \tag{3.76}$$

which is strictly greater than 0 by (3.56) with $\mathcal{L} = \emptyset$ and $\mathcal{K} = \{j\}$. □

3.4.6 Diffusion Approximation - Main Theorem

3.4.6.1 Definition of an SRBM

Before defining an SRBM, we define an $\{\mathcal{F}_t\}$-adapted Brownian motion. Given a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$, a vector $\theta \in \mathbb{R}^N$, an $N \times N$ symmetric, strictly positive-definite matrix Γ, and a probability distribution ν on $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$, an $\{\mathcal{F}_t\}$-Brownian motion with drift vector θ, covariance matrix Γ, and initial distribution ν, is an N-dimensional $\{\mathcal{F}_t\}$-adapted process, X, defined on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ such that the following hold under P:

(i) X is an N-dimensional Brownian motion whose sample paths are almost surely continuous and that has initial distribution ν,

(ii) $\{X_i(t) - X_i(0) - \theta_i t, \mathcal{F}_t, t \geq 0\}$ is a martingale for $i \in \mathcal{N}$, and

(iii) $\{(X_i(t) - X_i(0) - \theta_i t)(X_j(t) - X_j(0) - \theta_j t) - \Gamma_{ij} t, \mathcal{F}_t, t \geq 0\}$ is a martingale for $i, j \in \mathcal{N}$.

If $\nu = \delta_x$, the unit mass at $x \in \mathbb{R}^N$, we say that X starts from x.

Now, fix $\theta \in \mathbb{R}^N$, Γ an $N \times N$ symmetric strictly positive-definite covariance matrix, R an $N \times N$ matrix satisfying the HR condition, and ν a probability measure on $(\mathbb{R}^N_+, \mathcal{B}(\mathbb{R}^N_+))$. Recall the definition of F_i, $i \in \mathcal{N}$ from Section 3.4.1.

Definition 3.4.2 (Semimartingale Reflecting Brownian Motion (SRBM))**.** *A Semimartingale Reflecting Brownian Motion (abbreviated as SRBM) with the data $(\mathbb{R}^N_+, \theta, \Gamma, R, \nu)$ is an $\{\mathcal{F}_t\}$-adapted, N-dimensional process, W, defined on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ such that*

(i) P-a.s., $W(t) = X(t) + RY(t)$ for all $t \geq 0$,

(ii) P-a.s., W has continuous paths and $W(t) \in \mathbb{R}^N_+$ for all $t \geq 0$,

(iii) under P, X is an N-dimensional $\{\mathcal{F}_t\}$-Brownian motion with drift vector θ, covariance matrix Γ, and initial distribution ν,

(iv) Y is an $\{\mathcal{F}_t\}$-adapted, N-dimensional process such that P-a.s. for each $i \in \mathcal{N}$,

(a) $Y_i(0) = 0$,

(b) Y_i is continuous and non-decreasing,

(c) Y_i can only increase when W is on the face F_i, i.e., for all $t \geq 0$,

$$Y_i(t) = \int_0^t 1_{F_i}(W(s)) dY_i(s). \tag{3.77}$$

When $\nu = \delta_x$ for $x \in \mathbb{R}^N_+$, we may say that W is an SRBM with the data $(\mathbb{R}^N_+, \theta, \Gamma, R)$ that starts from x.

Remark. It is known from the work of Harrison and Reiman [16] that when R satisfies the HR condition, there is strong existence and uniqueness (and hence weak existence and uniqueness) for an SRBM given the data $(\mathbb{R}^N_+, \theta, \Gamma, R)$ and the initial distribution ν.

Remark. An N-dimensional process W defined on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ is a continuous semimartingale if W is a continuous adapted process and P-a.s., $W(t) = W(0) + M(t) + A(t)$ for all $t \geq 0$ where M is a continuous N-dimensional $\{\mathcal{F}_t\}$-adapted local martingale with $M(0) = 0$, and A is a continuous $\{\mathcal{F}_t\}$-adapted process whose paths are P-a.s of finite variation on each bounded time interval with $A(0) = 0$. In our definition of SRBM, $M(t) = X(t) - X(0) - \theta t$ and $A(t) = \theta t + RY(t)$ for all $t \geq 0$.

3.4.6.2 Main Theorem

We are now ready to state the main theorem of this section and give an outline of the proof. Recall the parameters θ, Γ, and R defined in (3.60), (3.73), and (3.74).

Theorem 3.4.6. *The diffusion-scaled workload process $\hat{W}^r(\cdot)$ converges in distribution as $r \to \infty$ to an SRBM with data $(\mathbb{R}^N_+, \theta, \Gamma, R)$ that starts from the origin.*

To prove this theorem, we first show that the sequence of processes $\{\hat{W}^r(\cdot)\}$ is C-tight (Section 3.4.7.3), i.e., any subsequence has a further subsequence that converges weakly to an almost surely continuous limit process. We then show that any weak limit point of such a subsequence is an SRBM with "extensive" data (Section 3.4.7.6), a notion that we make precise later (see Definition 3.4.3). For an SRBM with extensive data, there is a direction of reflection associated with each of the $2^N - 1$ boundary faces and there might be pushing in these directions at those boundary faces. In fact, we show that the pushing at boundary faces of dimension $N - 2$ or less is negligible (Section 3.4.8) and consequently, the SRBM with extensive data reduces to one of the simpler form as described in Theorem 3.4.6. Finally, we show that such an SRBM is unique in law and when combined with the C-tightness, we conclude that the sequence of diffusion-scaled workload processes converges in distribution to an SRBM with data $(\mathbb{R}^N_+, \theta, \Gamma, R)$ that

starts from the origin.

3.4.7 Proof of the Main Theorem

3.4.7.1 Pre-limit Workload Process

Throughout this subsection θ, Γ, and R are given by (3.60), (3.73), and (3.74) respectively. From (3.52), (3.53), (3.58), and (3.65), the diffusion-scaled workload process can be written so that for $r > 0$, $i \in \mathcal{N}$, and $t \geq 0$,

$$\hat{W}_i^r(t) = \frac{1}{r} V_i(E_i^r(r^2 t)) - \frac{T_i^r(r^2 t)}{r} \tag{3.78}$$

where

$$T_i^r(t) \triangleq \int_0^t \Lambda_i(W^r(s))ds. \tag{3.79}$$

We can rewrite (3.78) as

$$\begin{aligned}
\hat{W}_i^r(t) &= \frac{1}{r}\left[V_i(E_i^r(r^2 t)) - m_i E_i^r(r^2 t)\right] + \frac{1}{r}\left[m_i E_i^r(r^2 t) - m_i \lambda_i^r r^2 t\right] \\
&\quad + \lambda_i^r m_i r t - \frac{1}{r} T_i^r(r^2 t) \\
&= \hat{V}_i^r(\bar{E}_i^r(t)) + m_i \hat{E}_i^r(t) + (\lambda_i^r - \lambda_i)\, m_i r t \\
&\quad + \frac{1}{r}\lambda_i m_i \int_0^{r^2 t} ds - \frac{1}{r}\int_0^{r^2 t} \Lambda_i(W^r(s))ds \\
&= \hat{X}_i^r(t) + \sum_{\mathcal{K}\subseteq\mathcal{N}} \left(\lambda_i m_i - c_i^{\mathcal{K}}\right) \hat{U}^{r,\mathcal{K}}(t), \\
&= \hat{X}_i^r(t) + \sum_{\emptyset\neq\mathcal{K}\subseteq\mathcal{N}} (c_i^{\emptyset} - c_i^{\mathcal{K}}) \hat{U}^{r,\mathcal{K}}(t),
\end{aligned} \tag{3.80}$$

where

$$\hat{X}_i^r(t) \triangleq \hat{V}_i^r(\bar{E}_i^r(t)) + m_i \hat{E}_i^r(t) + (\lambda_i^r - \lambda_i)\, m_i r t, \tag{3.81}$$

$$\hat{U}^{r,\mathcal{K}}(t) \triangleq \frac{1}{r}\int_0^{r^2 t} 1_{\{\mathcal{K}(W^r(s))=\mathcal{K}\}}\, ds = r\int_0^t 1_{\left\{\mathcal{K}(\hat{W}^r(s))=\mathcal{K}\right\}}\, ds \tag{3.82}$$

and we have used the facts that for any $w \in \mathbb{R}_+^N$,

$$\sum_{\mathcal{K} \subseteq \mathcal{N}} 1_{\{\mathcal{K}(w)=\mathcal{K}\}} = 1, \tag{3.83}$$

and $c_i^{\emptyset} = \lambda_i m_i$, $i \in \mathcal{N}$. The second equation in (3.80) follows from the following simplification:

$$\begin{aligned} \frac{1}{r}\left[V_i(E_i^r(r^2t)) - m_i E_i^r(r^2t)\right] &= \frac{1}{r}\left[V_i\left(r^2\frac{E_i^r(r^2t)}{r^2}\right) - m_i r^2 \frac{E_i^r(r^2t)}{r^2}\right] \\ &= \frac{1}{r}\left[V_i(r^2\bar{E}_i^r(t)) - m_i r^2 \bar{E}_i^r(t)\right] = \hat{V}_i^r(\bar{E}_i^r(t)). \end{aligned} \tag{3.84}$$

To verify the last equality in (3.82), set $s = r^2u$ ($ds = r^2du$). Then when $s = r^2t$, $u = t$ and when $s = 0$, $u = 0$. Therefore,

$$\begin{aligned} \frac{1}{r}\int_0^{r^2t} 1_{\{\mathcal{K}(W^r(s))=\mathcal{K}\}}\, ds &= \frac{1}{r}\int_0^t 1_{\{\mathcal{K}(W^r(r^2u))=\mathcal{K}\}}\, r^2 du \\ &= r\int_0^t 1_{\left\{\mathcal{K}(r\hat{W}^r(u))=\mathcal{K}\right\}}\, du \\ &= r\int_0^t 1_{\left\{\mathcal{K}(\hat{W}^r(u))=\mathcal{K}\right\}}\, du \end{aligned} \tag{3.85}$$

where we have used Lemma 3.4.1 to arrive at the last equality. For notational convenience, we will sometimes write $\hat{U}^r(\cdot)$ in place of $\{\hat{U}^{r,\mathcal{K}}(\cdot), \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$ in the sequel.

3.4.7.2 Convergence to Brownian Motion

Our next result shows that the sequence of processes $\{\hat{X}^r(\cdot)\}$ converges in distribution to a Brownian motion. This result will be used in proving that the sequence of processes $\{(\hat{W}^r(\cdot), \hat{X}^r(\cdot), \hat{U}^r(\cdot))\}$ is C-tight (see Section 3.4.7.3) and that any weak limit point of this sequence defines an SRBM with extensive data (see Section 3.4.7.6).

Lemma 3.4.7. *The sequence of processes $\{\hat{X}^r(\cdot)\}$ converges in distribution to an N-dimensional Brownian motion that starts from the origin and has drift θ and covariance*

matrix Γ.

Proof. For all $t \geq 0$, $r > 0$, define

$$\theta(t) \triangleq \theta t, \tag{3.86}$$

$$\lambda(t) \triangleq \lambda t, \tag{3.87}$$

and

$$\hat{\theta}_i^r(t) \triangleq r(\lambda_i^r - \lambda_i) m_i t \quad \text{for all } i \in \mathcal{N}. \tag{3.88}$$

By Assumption 3.4.2, $\hat{\theta}^r(\cdot) \to \theta(\cdot)$ u.o.c. as $r \to \infty$. Combining this result with the standard functional central limit theorem (Proposition 3.4.3), we conclude that the sequence of processes $\{(\hat{E}^r(\cdot), \hat{V}^r(\cdot), \bar{E}^r(\cdot), \hat{\theta}^r(\cdot))\}$ converges in distribution to $(B_E(\cdot), B_V(\cdot), \lambda(\cdot), \theta(\cdot))$ where $B_E(\cdot)$ and $B_V(\cdot)$ are independent N-dimensional driftless Brownian motions starting from the origin with covariance matrices Γ_E and Γ_V given by (3.69) and (3.70) respectively. Then from (3.81), using the random time change lemma of Billingsley [5, p. 151], we conclude that $\{\hat{X}^r(\cdot)\}$ converges in distribution to $B_V(\lambda(\cdot)) + \mathrm{diag}(m) B_E(\cdot) + \theta(\cdot)$, which is an N-dimensional Brownian motion that starts from the origin, has drift θ, and a diagonal covariance matrix whose i-th diagonal entry is

$$\lambda_i m_i^2 \beta_i^2 + m_i^2 \lambda_i \alpha_i^2 = \lambda_i m_i^2 (\alpha_i^2 + \beta_i^2) = \Gamma_{ii}, \ i \in \mathcal{N}. \tag{3.89}$$

□

3.4.7.3 C-tightness

Theorem 3.4.8. *The sequence of processes* $\{(\hat{W}^r(\cdot), \hat{X}^r(\cdot), \hat{U}^r(\cdot))\}$ *is C-tight.*

To prove the C-tightness, we use a result from Kang and Williams [21]. In par-

ticular, we show that the Assumptions (A1)–(A5) and the Assumption 4.1 of [21] are satisfied by the geometric data and the sequence of processes, $\{(\hat{W}^r(\cdot), \hat{X}^r(\cdot), \hat{U}^r(\cdot))\}$, from which the C-tightness follows by Theorem 4.2 of [21]. This verification is carried out below.

3.4.7.4 Domain

For each $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, define $n^{\mathcal{K}}$ as the N-dimensional vector whose i-th element is $1/\sqrt{|\mathcal{K}|}$ if $i \in \mathcal{K}$ and 0 otherwise, that is, for $i \in \mathcal{N}$,

$$n_i^{\mathcal{K}} = \frac{1}{\sqrt{|\mathcal{K}|}} 1_{\{i \in \mathcal{K}\}}. \tag{3.90}$$

Then for each $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, $\left\|n^{\mathcal{K}}\right\| = 1$. For each $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, define $G^{\mathcal{K}}$ as

$$G^{\mathcal{K}} \triangleq \{x \in \mathbb{R}^N : \left\langle n^{\mathcal{K}}, x \right\rangle > 0\}. \tag{3.91}$$

Then for each $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, $G^{\mathcal{K}}$ is an open half-space of $\mathbb{R}^N$ and, therefore, a non-empty domain in $\mathbb{R}^N$. Define the domain G as

$$G \triangleq \bigcap_{\emptyset \neq \mathcal{K} \subseteq \mathcal{N}} G^{\mathcal{K}}. \tag{3.92}$$

In fact, $G = \{x \in \mathbb{R}^N : x_i > 0 \text{ for all } i \in \mathcal{N}\}$. Hence, $\overline{G} = \mathbb{R}^N_+$. (While the collection $\{G^{\{i\}}, i = 1, 2, \ldots, N\}$ is sufficient to define G, we include the other domains as well since they will have directions of reflection associated with them.)

Lemma 3.4.9. *The domain G with the representation* (3.92) *satisfies Assumptions (A1)–(A3) of [21, Section 3].*

Remark. Note that the inward unit normal vector for $G^{\mathcal{K}}$ is $n^{\mathcal{K}}$.

Proof. Since G is a finite intersection of half-spaces, $\overline{G}$ is a convex polyhedron. We also

note that for all $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, $\partial G \cap \partial G^{\mathcal{K}} \neq \emptyset$ since the origin is in $\partial G \cap \partial G^{\mathcal{K}}$. Consequently, by Lemma A.3 of [21], we only need to show that G satisfies Assumption (A1) of [21]. Recall that each $G^{\mathcal{K}}$ is a half-space. Therefore for each $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, $G^{\mathcal{K}}$ is a non-empty domain, $G^{\mathcal{K}} \neq \mathbb{R}^N$, and the boundary $\partial G^{\mathcal{K}}$ of $G^{\mathcal{K}}$ is C^1. Therefore the non-empty domain G satisfies Assumption (A1) and hence, Assumptions (A1)–(A3) of [21] hold. □

3.4.7.5 Reflection Vectors

For each $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, define the reflection vector $\gamma^{\mathcal{K}}$ such that

$$\gamma_i^{\mathcal{K}} \triangleq c_i^{\emptyset} - c_i^{\mathcal{K}} \text{ for each } i \in \mathcal{N}. \tag{3.93}$$

By this definition, if $i \in \mathcal{K}$, $c_i^{\mathcal{K}} = 0$ and therefore, $\gamma_i^{\mathcal{K}} = c_i^{\emptyset} > 0$. On the other hand, if $i \in \mathcal{K}^c$, $\gamma_i^{\mathcal{K}} = c_i^{\emptyset} - c_i^{\mathcal{K}} < 0$ by (3.57). With this definition of $\{\gamma^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$, (3.80) can be rewritten in vector form as

$$\hat{W}^r(t) = \hat{X}^r(t) + \sum_{\emptyset \neq \mathcal{K} \subseteq \mathcal{N}} \gamma^{\mathcal{K}} \hat{U}^{r,\mathcal{K}}(t). \tag{3.94}$$

Moreover, it is easy to see that the matrix whose columns are given by $\gamma^{\{1\}}, \cdots, \gamma^{\{N\}}$, is

$$R \operatorname{diag}(c_1^{\emptyset}, c_2^{\emptyset}, \ldots, c_N^{\emptyset}) \tag{3.95}$$

where R is the $N \times N$ reflection matrix defined in (3.74). To facilitate the use of [21], we define the normalized reflection vectors $\{\tilde{\gamma}^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$ by

$$\tilde{\gamma}^{\mathcal{K}} \triangleq \frac{\gamma^{\mathcal{K}}}{\|\gamma^{\mathcal{K}}\|}, \tag{3.96}$$

so that $\left\|\tilde{\gamma}^{\mathcal{K}}\right\| = 1$ for all $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$.

Lemma 3.4.10. *The reflection vectors $\{\tilde{\gamma}^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$ satisfy Assumptions (A4)–(A5) of [21, Section 3].*

Proof. Since the reflection vectors are constant, it is clear that the uniform Lipschitz continuity property of Assumption (A4) of [21] is satisfied. Also, we have normalized the vectors to be of unit length.

To verify (A5), we need to show that there is a constant $a \in (0, 1)$ such that for each $x \in \partial G$, there are nonnegative constants $(b_{\mathcal{L}}(x) : \emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x))$ and $(d_{\mathcal{L}}(x) : \emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x))$ such that

$$\sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x)} b_{\mathcal{L}}(x) = 1, \tag{3.97}$$

$$\min_{\emptyset \neq \mathcal{M} \subseteq \mathcal{K}(x)} \left\langle \sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x)} b_{\mathcal{L}}(x) n^{\mathcal{L}}, \tilde{\gamma}^{\mathcal{M}} \right\rangle \geq a, \tag{3.98}$$

$$\sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x)} d_{\mathcal{L}}(x) = 1, \tag{3.99}$$

$$\min_{\emptyset \neq \mathcal{M} \subseteq \mathcal{K}(x)} \left\langle \sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x)} d_{\mathcal{L}}(x) \tilde{\gamma}^{\mathcal{L}}, n^{\mathcal{M}} \right\rangle \geq a. \tag{3.100}$$

To this end, for any $x \in \partial G$ and $\emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x)$, set

$$b_{\mathcal{L}}(x) \triangleq 1_{\{\mathcal{L} = \mathcal{K}(x)\}} \tag{3.101}$$

and

$$d_{\mathcal{L}}(x) \triangleq 1_{\{\mathcal{L} = \mathcal{K}(x)\}}. \tag{3.102}$$

Then

$$\sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x)} b_{\mathcal{L}}(x) n^{\mathcal{L}} = n^{\mathcal{K}(x)} \tag{3.103}$$

and

$$\sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{K}(x)} d_{\mathcal{L}}(x) \tilde{\gamma}^{\mathcal{L}} = \tilde{\gamma}^{\mathcal{K}(x)}. \tag{3.104}$$

Therefore to verify that Assumption (A5) of [21] is satisfied, we only need to verify that for each $x \in \partial G$ and $\emptyset \neq \mathcal{M} \subseteq \mathcal{K}(x)$, $\left\langle n^{\mathcal{K}(x)}, \tilde{\gamma}^{\mathcal{M}} \right\rangle$ and $\left\langle \tilde{\gamma}^{\mathcal{K}(x)}, n^{\mathcal{M}} \right\rangle$ are bounded below by a strictly positive constant not depending on x or $\mathcal{M}$. We first verify that $\left\langle \tilde{\gamma}^{\mathcal{K}(x)}, n^{\mathcal{M}} \right\rangle$ has such a lower bound. From (3.93) and (3.96), for all $i \in \mathcal{K}(x)$,

$$\tilde{\gamma}_i^{\mathcal{K}(x)} = \frac{c_i^{\emptyset}}{\|\gamma^{\mathcal{K}(x)}\|} > 0. \tag{3.105}$$

Thus, using (3.90), for each $\emptyset \neq \mathcal{M} \subseteq \mathcal{K}(x)$,

$$\begin{aligned} \left\langle \tilde{\gamma}^{\mathcal{K}(x)}, n^{\mathcal{M}} \right\rangle &= \frac{1}{\sqrt{|\mathcal{M}|}} \sum_{i \in \mathcal{M}} \tilde{\gamma}_i^{\mathcal{K}(x)} \\ &\geq \frac{\min_{i \in \mathcal{M}} c_i^{\emptyset}}{\sqrt{|\mathcal{M}|}\, \|\gamma^{\mathcal{K}(x)}\|} \\ &\geq \frac{\min_{i \in \mathcal{N}} c_i^{\emptyset}}{\sqrt{N} \max_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \|\gamma^{\mathcal{L}}\|} \\ &> 0 \end{aligned} \tag{3.106}$$

where the second inequality follows because we are taking minimum over a larger set in the third line and for all $x \in \partial G$, $|\mathcal{K}(x)| \leq N$. Next, we show that for each $\emptyset \neq \mathcal{M} \subseteq \mathcal{K}(x)$, $\left\langle n^{\mathcal{K}(x)}, \tilde{\gamma}^{\mathcal{M}} \right\rangle$ has a uniform strictly positive lower bound. To this end, we have for

$\emptyset \neq \mathcal{M} \subseteq \mathcal{K}(x)$,

$$\begin{aligned}
\left\langle n^{\mathcal{K}(x)}, \tilde{\gamma}^{\mathcal{M}} \right\rangle &= \frac{1}{\sqrt{|\mathcal{K}(x)|}} \sum_{i \in \mathcal{K}(x)} \gamma_i^{\mathcal{M}} / \left\|\gamma^{\mathcal{M}}\right\| \\
&= \frac{1}{\sqrt{|\mathcal{K}(x)|}} \sum_{i \in \mathcal{K}(x)} (c_i^{\emptyset} - c_i^{\mathcal{M}}) / \left\|\gamma^{\mathcal{M}}\right\| \\
&= \frac{1}{\sqrt{|\mathcal{K}(x)|}} \left[\sum_{i \in \mathcal{N}} (c_i^{\emptyset} - c_i^{\mathcal{M}}) - \sum_{i \in (\mathcal{K}(x))^c} (c_i^{\emptyset} - c_i^{\mathcal{M}}) \right] / \left\|\gamma^{\mathcal{M}}\right\| \\
&\geq \frac{1}{\sqrt{|\mathcal{K}(x)|}} \sum_{i \in \mathcal{N}} (c_i^{\emptyset} - c_i^{\mathcal{M}}) / \left\|\gamma^{\mathcal{M}}\right\| \\
&\geq \frac{1}{\sqrt{N}} \min_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \sum_{i \in \mathcal{N}} (c_i^{\emptyset} - c_i^{\mathcal{L}}) / \max_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \left\|\gamma^{\mathcal{L}}\right\| \\
&> 0
\end{aligned} \tag{3.107}$$

where the first inequality follows from (3.57) with $\mathcal{L} = \emptyset$ and $\mathcal{M}$ in place of $\mathcal{K}$ and the last inequality follows from (3.56) and the fact that $c_i^{\mathcal{L}} = 0$ if $i \in \mathcal{L}$. □

Proof of Theorem 3.4.8. For each $r > 0$, let

$$\hat{Z}^r \triangleq (\hat{W}^r, \hat{X}^r, \hat{U}^r). \tag{3.108}$$

To prove the C-tightness of $\{\hat{Z}^r\}$, we first verify that Assumption 4.1 of [21, Section 4] is satisfied.

For any $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$ and $r \geq 0$, let $\gamma^{r,\mathcal{K}}(y, x) \triangleq \tilde{\gamma}^{\mathcal{K}}$ for all $x, y \in \mathbb{R}^N$, $\alpha^r \triangleq 0 \in \mathbb{D}^N$, $\beta^r = \{\beta^{r,\mathcal{K}} : \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$ where $\beta^{r,\mathcal{K}} \triangleq 0 \in \mathbb{D}$, $\delta^r = 1/r$, and $\hat{Y}^r = \{\hat{Y}^{r,\mathcal{K}} : \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$ where $\hat{Y}^{r,\mathcal{K}} = \left\|\gamma^{\mathcal{K}}\right\| \hat{U}^{r,\mathcal{K}}$. With these definitions, the conditions (i)–(vi) of Assumption 4.1 of [21] are satisfied with $\{(\hat{W}^r, \hat{X}^r, \hat{Y}^r)\}$ in place of $\{(W^n, X^n, Y^n)\}$. Here

$$\hat{Y}^{r,\mathcal{K}}(t) = \int_0^t 1_{\left\{\text{dist}(\hat{W}^r(s), \partial G^{\mathcal{K}} \cap \partial G) \leq \delta^r\right\}} d\hat{Y}^{r,\mathcal{K}}(s) \tag{3.109}$$

because $\hat{U}^{r,\mathcal{K}}$ can increase only when $\hat{W}^r$ is on $\partial G^{\mathcal{K}} \cap \partial G$ (see (3.82)), and $\{\hat{X}^r\}$ is

C-tight by Lemma 3.4.7. It then follows from Theorem 4.2 of [21, Section 4], that $\{(\hat{W}^r, \hat{X}^r, \hat{Y}^r)\}$, and hence $\{\hat{Z}^r\}$, is C-tight and the theorem is proved. □

3.4.7.6 SRBM with Extensive Data

We next show that any weak limit point of the sequence of processes $\{(\hat{W}^r(\cdot), \hat{X}^r(\cdot), \hat{U}^r(\cdot))\}$ is an SRBM with extensive data. Before presenting the theorem and its proof, we need to define an SRBM with extensive data. The following definition is adapted from the definition in [21, Section 2]. Recall the definition of G from (3.92), θ and Γ from (3.60) and (3.73), and $\{\gamma^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$ from (3.93). Let ν be a probability measure on $(\overline{G}, \mathcal{B}(\overline{G}))$, where $\mathcal{B}(\overline{G})$ denotes the σ-algebra of Borel subsets of the closure, $\overline{G}$, of G.

Definition 3.4.3 (SRBM with Extensive Data)**.** *An SRBM with the extensive data* $(\overline{G}, \theta, \Gamma, \{\gamma^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}, \nu)$ *is an* $\{\mathcal{F}_t\}$*-adapted,* N*-dimensional process* W *defined on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ *such that*

(i) P*-a.s., for all* $t \geq 0$*,*

$$W(t) = X(t) + \sum_{\emptyset \neq \mathcal{K} \subseteq \mathcal{N}} \int_0^t \gamma^{\mathcal{K}}(W(s)) dU^{\mathcal{K}}(s), \tag{3.110}$$

(ii) P*-a.s.,* W *has continuous paths and* $W(t) \in \overline{G}$ *for all* $t \geq 0$*,*

(iii) under P*,* X *is an* N*-dimensional* $\{\mathcal{F}_t\}$*-Brownian motion with drift vector* θ*, covariance matrix* Γ*, and initial distribution* ν*,*

(iv) for each $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$*,* $U^{\mathcal{K}}$ *is an* $\{\mathcal{F}_t\}$*-adapted, one-dimensional process such that* P*-a.s.,*

(a) $U^{\mathcal{K}}(0) = 0$*,*

(b) $U^{\mathcal{K}}$ *is continuous and non-decreasing,*

(c) for all $t \geq 0$*,*

$$U^{\mathcal{K}}(t) = \int_0^t 1_{\{W(s) \in \partial G^{\mathcal{K}} \cap \partial G\}} dU^{\mathcal{K}}(s). \tag{3.111}$$

When $\nu = \delta_x$*, for* $x \in \overline{G}$*, we may say that* W *is an SRBM associated with the data* $(\overline{G}, \theta, \Gamma, \{\gamma^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\})$ *that starts from* x*.*

Remark. We have introduced the terminology "extensive" data in this work to differentiate between the above SRBM which has reflection on the lower-dimensional faces and the simpler SRBM introduced in Definition 3.4.2.

Remark. Recall the definition of a continuous semimartingale from the remarks following Definition 3.4.2. In the above definition of an SRBM, the decomposition of the semimartingale $M(t) = X(t) - X(0) - \theta t$ and

$$A(t) = \theta t + \sum_{\emptyset \neq \mathcal{K} \subseteq \mathcal{N}} \int_0^t \gamma^{\mathcal{K}}(W(s)) dU^{\mathcal{K}}(s). \tag{3.112}$$

With this definition in hand, we can now state and prove the main result of this subsection.

Theorem 3.4.11. *Any weak limit point* $(W(\cdot), X(\cdot), U(\cdot))$ *of the sequence of processes* $\{(\hat{W}^r(\cdot), \hat{X}^r(\cdot), \hat{U}^r(\cdot))\}$ *defines an SRBM,* W*, with the extensive data* $(\overline{G}, \theta, \Gamma, \{\gamma^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\})$ *that starts from the origin.*

We need the following lemma for our proof of Theorem 3.4.11. So as to not disrupt the flow of this section, we defer the proof of this lemma to Appendix 3.A.

Lemma 3.4.12. *Suppose that* $Z = (W, X, U)$ *is a weak limit point of the sequence* $\{(\hat{W}^r, \hat{X}^r, \hat{U}^r)\}$*. Let* $\mathcal{F}_t = \sigma\{Z(s) : 0 \leq s \leq t\}, t \geq 0$*. Then* $\{X(t) - X(0) - \theta t, \mathcal{F}_t, t \geq 0\}$ *is a martingale.*

Proof. See Appendix 3.A. □

Proof of Theorem 3.4.11. The result follows from Theorem 4.3 of [21] provided Assumption 4.1 and Assumptions $(\text{vi})'$ and (vii) of Theorem 4.3 in [21] hold for $\{(\hat{W}^r, \hat{X}^r, \hat{Y}^r)\}$ where $\hat{Y}^r = \{\hat{Y}^{r,\mathcal{K}} : \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$ and $\hat{Y}^{r,\mathcal{K}} = \left\|\gamma^{\mathcal{K}}\right\| \hat{U}^{r,\mathcal{K}}$. Our proof of Theorem 3.4.8 shows that Assumption 4.1 of [21] holds. Assumption $(\text{vi})'$ of Theorem 4.3 in [21] follows immediately from Lemma 3.4.7. Assumption (vii) of Theorem 4.3 in [21] follows from Lemma 3.4.12 and the simple relationship between $\hat{U}^{r,\mathcal{K}}$ and $\hat{Y}^{r,\mathcal{K}}$. □

3.4.8 Pushing on the Lower-dimensional Faces

In this subsection, we show a result, which when combined with Theorem 3.4.11 implies that for any weak limit point, $(W(\cdot), X(\cdot)), U(\cdot))$, of the sequence of processes $\{(\hat{W}^r(\cdot), \hat{X}^r(\cdot), \hat{U}^r(\cdot))\}$, the amount of pushing done by U at any of the faces of ∂G of dimension $N-2$ or less is negligible. Formally, we prove the following.

Theorem 3.4.13. *Let $(W(\cdot), X(\cdot), U(\cdot))$ define an SRBM, $W(\cdot)$, with extensive data $(\overline{G}, \theta, \Gamma, \{\gamma^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\})$ that starts from the origin. Then for each $\mathcal{K} \subseteq \mathcal{N}$, $|\mathcal{K}| \geq 2$, for each $\emptyset \neq \mathcal{L} \subseteq \mathcal{K}$,*

$$\int_0^\infty 1_{F_{\mathcal{K}}}(W(s)) dU^{\mathcal{L}}(s) = 0 \text{ almost surely.} \tag{3.113}$$

Consequently, almost surely,

$$W(t) = X(t) + \sum_{i \in \mathcal{N}} \gamma^{\{i\}} U^{\{i\}}(t), \ t \geq 0. \tag{3.114}$$

Our proof of Theorem 3.4.13 is a generalization of the proof of the main theorem in Reiman and Williams [33]. However, there are some differences; since in [33], there were only N directions of reflection – one for each $(N-1)$-dimensional boundary face, whereas here there are $2^N - 1$, one for each boundary face. We prove the theorem in three steps. We assume that $N \geq 2$, otherwise the result is vacuous and hence trivially true. We first prove that for the case of zero drift ($\theta = 0$) the amount of pushing done

when W is at the origin is negligible (see Lemma 3.4.14). We then use a backwards induction argument on $|\mathcal{K}|$ to show that for the case of zero drift the amount of pushing done on $F_{\mathcal{K}}$ is negligible provided $|\mathcal{K}| \geq 2$ (see Lemma 3.4.15). Finally, using a Girsanov transformation, the result is extended to all constant drifts θ (see Lemma 3.4.16). We then complete the proof.

Lemma 3.4.14. *Suppose (W, X, U) is as in the hypothesis of Theorem 3.4.13 and $\theta = 0$. Then for $N \geq 2$ and $\mathcal{K} = \mathcal{N}$,* (3.113) *holds for all $\emptyset \neq \mathcal{L} \subseteq \mathcal{N}$.*

Proof. From the semimartingale representation (3.110) of W and Itô's formula, for any function f that is twice continuously differentiable in some domain containing $\overline{G}$, we have almost surely for all $t \geq 0$:

$$\begin{aligned} f(W(t)) - f(W(0)) = & \int_0^t \langle \nabla f(W(s)), dX(s) \rangle \\ & + \sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \int_0^t \left\langle \gamma^{\mathcal{L}}, \nabla f(W(s)) \right\rangle dU^{\mathcal{L}}(s) \\ & + \int_0^t Lf(W(s)) ds \end{aligned} \tag{3.115}$$

where

$$Lf = \frac{1}{2} \sum_{i=1}^{N} \Gamma_{ii} \frac{\partial^2 f}{\partial x_i^2}. \tag{3.116}$$

We shall substitute functions into (3.115) that allow us to estimate the left hand side of (3.113). Each such function will be L-harmonic in some domain containing $\overline{G}$ and for each $\emptyset \neq \mathcal{L} \subseteq \mathcal{N}$, its directional derivative in the direction of $\gamma^{\mathcal{L}}$ will be bounded below on $\overline{G}$ and be very large and positive near the origin. These functions are chosen such that they are uniformly bounded on compact subsets of $\overline{G}$.

Define

$$\tilde{\beta} \triangleq \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}_+^N. \tag{3.117}$$

Then from (3.56) with $\mathcal{L} = \emptyset$, $\mathcal{K}$ replaced by $\mathcal{L}$, the fact that $c_i^{\mathcal{L}} = 0$ if $i \in \mathcal{L}$ and (3.93), we have for all $\emptyset \neq \mathcal{L} \subseteq \mathcal{N}$,

$$\left\langle \gamma^{\mathcal{L}}, \tilde{\beta} \right\rangle > 0. \tag{3.118}$$

Therefore, there exists a vector $\beta \in \mathbb{R}_+^N$ having all components strictly positive such that for all $\emptyset \neq \mathcal{L} \subseteq \mathcal{N}$,

$$\left\langle \gamma^{\mathcal{L}}, \beta \right\rangle \triangleq \delta^{\mathcal{L}} \in [1, \infty). \tag{3.119}$$

Define

$$\alpha \triangleq \Gamma\beta. \tag{3.120}$$

For each $x \in \overline{G} = \mathbb{R}_+^N$ and $s \in (0, 1)$, define a squared distance function:

$$\begin{aligned} d^2(x, s) &\triangleq (x + s\alpha)'\Gamma^{-1}(x + s\alpha) \\ &= x'\Gamma^{-1}x + 2s\alpha'\Gamma^{-1}x + s^2\alpha'\Gamma^{-1}\alpha \\ &= x'\Gamma^{-1}x + 2s\beta'x + s^2\alpha'\Gamma^{-1}\alpha \\ &\geq s^2\hat{\alpha} \end{aligned} \tag{3.121}$$

where

$$\hat{\alpha} \triangleq \alpha'\Gamma^{-1}\alpha = \beta'\Gamma\beta > 0. \tag{3.122}$$

We have used the facts that Γ (and hence Γ^{-1}) is symmetric and strictly positive definite,

and $\beta > 0$. Then for each fixed $\varepsilon \in (0,1)$,

$$\phi_\varepsilon(x) \triangleq \begin{cases} \frac{1}{2-N}\int_\varepsilon^1 s^{N-2}(d^2(x,s))^{\frac{2-N}{2}}ds, & N \geq 3, \\ \frac{1}{2}\int_\varepsilon^1 \ln(d^2(x,s))ds, & N = 2, \end{cases} \tag{3.123}$$

is twice continuously differentiable in some domain containing $\overline{G}$, and on each compact subset of $\overline{G}$, it is bounded, uniformly in ε. Moreover, since the integrand in (3.123), for a fixed s, is L-harmonic as a function of $x \in \mathbb{R}^N \setminus \{-s\alpha\}$, it is readily verified that for each $\varepsilon \in (0,1)$,

$$L\phi_\varepsilon = 0 \tag{3.124}$$

in some domain containing $\overline{G}$.

For the verification of the directional derivative properties of ϕ_ε, for each $\emptyset \neq \mathcal{L} \subseteq \mathcal{N}$, let

$$u^{\mathcal{L}} \triangleq \Gamma^{-1}\gamma^{\mathcal{L}}. \tag{3.125}$$

Then

$$\left\langle u^{\mathcal{L}}, \alpha \right\rangle = \left\langle \Gamma^{-1}\gamma^{\mathcal{L}}, \alpha \right\rangle = (\gamma^{\mathcal{L}})'\Gamma^{-1}\alpha = (\gamma^{\mathcal{L}})'\beta = \delta^{\mathcal{L}} \geq 1. \tag{3.126}$$

Combining (3.126) with

$$\nabla\phi_\varepsilon(x) = \int_\varepsilon^1 s^{N-2}\Gamma^{-1}(x+s\alpha)(d^2(x,s))^{-N/2}ds, \tag{3.127}$$

we get

$$\left\langle \gamma^{\mathcal{L}}, \nabla\phi_\varepsilon(x) \right\rangle = \int_\varepsilon^1 s^{N-2}(\left\langle u^{\mathcal{L}}, x \right\rangle + s\delta^{\mathcal{L}})(d^2(x,s))^{-N/2}ds. \tag{3.128}$$

Let

$$\xi^{\mathcal{L}} \triangleq \frac{\delta^{\mathcal{L}}}{\|u^{\mathcal{L}}\|}. \tag{3.129}$$

Then for $\varepsilon \in (0,1)$ and $x \in \overline{G}$ satisfying $\|x\| < \varepsilon\xi^{\mathcal{L}}$, we have $\left|\left\langle u^{\mathcal{L}}, x \right\rangle\right| < \varepsilon\delta^{\mathcal{L}}$ and for

$s > \varepsilon$,

$$\begin{aligned} d^2(x,s) &\leq \left\|\Gamma^{-1}\right\| \|x + s\alpha\|^2 \\ &\leq \left\|\Gamma^{-1}\right\| (\|x\| + \|s\alpha\|)^2 \\ &\leq \left\|\Gamma^{-1}\right\| (\xi^{\mathcal{L}} + \|\alpha\|)^2 s^2 \end{aligned} \tag{3.130}$$

where $\|\Gamma^{-1}\|$ denotes the norm of Γ^{-1} as an operator from $\mathbb{R}^N$ to $\mathbb{R}^N$ with the Euclidean norm. Setting

$$\zeta^{\mathcal{L}} \triangleq \delta^{\mathcal{L}} (\left\|\Gamma^{-1}\right\| (\xi^{\mathcal{L}} + \|\alpha\|)^2)^{-N/2} \tag{3.131}$$

and substituting the above in (3.128) yields:

$$\begin{aligned} \left\langle \gamma^{\mathcal{L}}, \nabla\phi_\varepsilon(x) \right\rangle &\geq \zeta^{\mathcal{L}} \int_\varepsilon^1 s^{N-2}(s-\varepsilon)s^{-N} ds \\ &\geq -\zeta^{\mathcal{L}}[\ln\varepsilon + 1] \end{aligned} \tag{3.132}$$

for all $x \in \overline{G}$ satisfying $\|x\| < \varepsilon\xi^{\mathcal{L}}$. Note that for small ε, the term in the last line above is large and positive.

Now for any $x \in \overline{G}$, $\emptyset \neq \mathcal{L} \subseteq \mathcal{N}$,

$$\left\langle \gamma^{\mathcal{L}}, \nabla\phi_\varepsilon(x) \right\rangle = -\delta^{\mathcal{L}} \int_\varepsilon^1 s^{N-2}(\rho^{\mathcal{L}}(x) - s)(d^2(x,s))^{-N/2} ds \tag{3.133}$$

where

$$\rho^{\mathcal{L}}(x) \triangleq -\frac{\left\langle u^{\mathcal{L}}, x \right\rangle}{\delta^{\mathcal{L}}}. \tag{3.134}$$

If $\rho^{\mathcal{L}}(x) \leq \varepsilon$, then the right hand side of (3.133) is non-negative. Thus, to obtain a lower bound for $\left\langle \gamma^{\mathcal{L}}, \nabla\phi_\varepsilon(x) \right\rangle$ on $\overline{G}$, it suffices to consider $x \in \overline{G}$ such that $\rho^{\mathcal{L}}(x) > \varepsilon$. For

such x,

$$\begin{aligned}
&\int_{\varepsilon}^{1} s^{N-2}(\rho^{\mathcal{L}}(x)-s)(d^2(x,s))^{-N/2}ds \\
&\quad\leq \int_{\varepsilon}^{\rho^{\mathcal{L}}(x)} s^{N-2}(\rho^{\mathcal{L}}(x)-s)(d^2(x,s))^{-N/2}ds \\
&\quad\leq (\rho^{\mathcal{L}}(x)-\varepsilon)\max_{s\in[\varepsilon,\rho^{\mathcal{L}}(x)]}\frac{\rho^{\mathcal{L}}(x)-s}{d^2(x,s)}\max_{s\in[\varepsilon,\rho^{\mathcal{L}}(x)]}\frac{s^{N-2}}{(d^2(x,s))^{(N-2)/2}}.
\end{aligned} \tag{3.135}$$

Since $d^2(x,s)$ is quadratic in s with positive coefficients, the first maximum above is achieved at $s=\varepsilon$, and by (3.121), the second maximum is crudely dominated by $\hat{\alpha}^{(2-N)/2}$. Thus, the last term of (3.135) is bounded from above by

$$\frac{(\rho^{\mathcal{L}}(x)-\varepsilon)^2}{d^2(x,\varepsilon)}\hat{\alpha}^{(2-N)/2}. \tag{3.136}$$

Since Γ^{-1} is strictly positive definite, there is a $\eta>0$ such that $x'\Gamma^{-1}x\geq\eta\left\|x\right\|^2$ and so (see (3.121)),

$$\begin{aligned}
d^2(x,\varepsilon)&\geq\eta\left\|x\right\|^2+\varepsilon^2\hat{\alpha}\\
&\geq(\eta\wedge\hat{\alpha})(\left\|x\right\|^2+\varepsilon^2).
\end{aligned} \tag{3.137}$$

On the other hand, by the definition of $\rho^{\mathcal{L}}(x)$,

$$\begin{aligned}
(\rho^{\mathcal{L}}(x)-\varepsilon)^2&\leq 2((\rho^{\mathcal{L}}(x))^2+\varepsilon^2)\\
&\leq 2(\left\|u^{\mathcal{L}}\right\|^2\left\|x\right\|^2(\delta^{\mathcal{L}})^{-2}+\varepsilon^2)\\
&\leq 2\max(\left\|u^{\mathcal{L}}\right\|^2(\delta^{\mathcal{L}})^{-2},1)(\left\|x\right\|^2+\varepsilon^2).
\end{aligned} \tag{3.138}$$

It follows from (3.137) and (3.138) that (3.135) is bounded from above by a constant not depending on x or ε. Hence, there is a $\tilde{\zeta}^{\mathcal{L}}\geq 0$ such that for all $x\in\overline{G}$ and $\varepsilon\in(0,1)$,

$$\left\langle\gamma^{\mathcal{L}},\nabla\phi_{\varepsilon}(x)\right\rangle\geq-\tilde{\zeta}^{\mathcal{L}}. \tag{3.139}$$

We are now ready to prove that when $\mathcal{K} = \mathcal{N}$, (3.113) holds almost surely for each $\emptyset \neq \mathcal{L} \subseteq \mathcal{N}$. For each positive integer m, define

$$T_m \triangleq \inf\{t \geq 0 : \|W(t)\| \geq m \text{ or } U^{\mathcal{L}}(t) \geq m \text{ for some } \emptyset \neq \mathcal{L} \subseteq \mathcal{N}\} \wedge m. \quad (3.140)$$

Replacing f by ϕ_ε and t by T_m in (3.115), we see from (3.124) that almost surely:

$$\begin{aligned} \phi_\varepsilon(W(T_m)) - \phi_\varepsilon(W(0)) = & \int_0^{T_m} \langle \nabla \phi_\varepsilon(W(s)), dX(s) \rangle \\ & + \sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \int_0^{T_m} \left\langle \gamma^{\mathcal{L}}, \nabla \phi_\varepsilon(W(s)) \right\rangle dU^{\mathcal{L}}(s). \end{aligned} \quad (3.141)$$

Since ϕ_ε and its first derivatives are bounded on each compact subset of $\overline{G}$, by the definition of the stopping time T_m and since $\theta = 0$, the stochastic integral with respect to dX in (3.141) has zero expectation. Thus, taking expectations in (3.141) yields:

$$\begin{aligned} \mathbf{E}\left[\phi_\varepsilon(W(T_m)) - \phi_\varepsilon(W(0))\right] = & \sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \mathbf{E}\left[\int_0^{T_m} \left\langle \gamma^{\mathcal{L}}, \nabla \phi_\varepsilon(W(s)) \right\rangle dU^{\mathcal{L}}(s) \right] \\ \geq & -(\ln \varepsilon + 1) \sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \zeta^{\mathcal{L}} \mathbf{E}\left[\int_0^{T_m} 1_{\{\|W(s)\| < \varepsilon \xi^{\mathcal{L}}\}} dU^{\mathcal{L}}(s) \right] \\ & - \sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \tilde{\zeta}^{\mathcal{L}} \mathbf{E}\left[U^{\mathcal{L}}(T_m)\right], \end{aligned} \quad (3.142)$$

where the lower bounds (3.132) and (3.139) have been used to obtain the last inequality. Now, the left-hand side of (3.142) is bounded as $\varepsilon \downarrow 0$, since for $\varepsilon \in (0, 1)$, ϕ_ε is uniformly bounded on compact subsets of $\overline{G}$. Also, the last sum in (3.142) is positive and independent of ε. Thus, dividing (3.142) by $-(\ln \varepsilon + 1)$ and letting $\varepsilon \downarrow 0$ yields:

$$\lim_{\varepsilon \downarrow 0} \sum_{\emptyset \neq \mathcal{L} \subseteq \mathcal{N}} \zeta^{\mathcal{L}} \mathbf{E}\left[\int_0^{T_m} 1_{\{\|W(s)\| < \varepsilon \xi^{\mathcal{L}}\}} dU^{\mathcal{L}}(s) \right] \leq 0. \quad (3.143)$$

Since each term in the above sum is non-negative and $\zeta^{\mathcal{L}} > 0$, it follows by Fatou's lemma

that for each $\emptyset \neq \mathcal{L} \subseteq \mathcal{N}$,

$$\int_0^{T_m} 1_{F_{\mathcal{N}}}(W(s))dU^{\mathcal{L}}(s) = 0 \text{ almost surely.} \tag{3.144}$$

Letting $m \to \infty$ yields the desired result. □

Lemma 3.4.15. *Suppose (W, X, U) is as in the hypothesis of Theorem 3.4.13 and $\theta = 0$. Then* (3.113) *holds for all $\emptyset \neq \mathcal{L} \subseteq \mathcal{K} \subseteq \mathcal{N}$ where $|\mathcal{K}| \geq 2$.*

Proof. Our proof is by backwards induction on $|\mathcal{K}|$. Without loss of generality, we assume $N \geq 2$ (otherwise there is no $\mathcal{K} \subseteq \mathcal{N}$ with $|\mathcal{K}| \geq 2$). By Lemma 3.4.14, the result holds for $|\mathcal{K}| = N$ in which case the only possible $\mathcal{K}$ is $\mathcal{K} = \mathcal{N}$. Fix $2 \leq k < N$ and suppose that (3.113) holds for all $\mathcal{K} \subseteq \mathcal{N}$ and $\emptyset \neq \mathcal{L} \subseteq \mathcal{K}$, such that $k < |\mathcal{K}| \leq N$. Fix some $\mathcal{K} \subseteq \mathcal{N}$ such that $|\mathcal{K}| = k$. We need to show that for all $\emptyset \neq \mathcal{L} \subseteq \mathcal{K}$,

$$\int_0^{\infty} 1_{F_{\mathcal{K}}}(W(s))dU^{\mathcal{L}}(s) = 0 \text{ almost surely.} \tag{3.145}$$

To this end, fix $\emptyset \neq \mathcal{L} \subseteq \mathcal{K}$. Then

$$\begin{aligned} \int_0^{\infty} 1_{F_{\mathcal{K}}}(W(s))dU^{\mathcal{L}}(s) &= \int_0^{\infty} 1_{\{\mathcal{K}(W(s))=\mathcal{K}\}}dU^{\mathcal{L}}(s) \\ &\quad + \int_0^{\infty} 1_{\left\{W(s)\in\cup_{\mathcal{K}\subsetneq\mathcal{M}} F_{\mathcal{M}}\right\}}dU^{\mathcal{L}}(s) \\ &\overset{\text{a.s.}}{=} \int_0^{\infty} 1_{\{W_{\mathcal{K}}(s)=0, W_{\mathcal{K}^c}(s)>0\}}dU^{\mathcal{L}}(s) \end{aligned} \tag{3.146}$$

where by the induction assumption the second integral on the right-hand side of the first equation is almost surely 0. Thus, by monotone convergence, it suffices to prove that for each $\eta \in \mathbb{R}_+^{N-k}$, satisfying $\eta > 0$, we have

$$\int_0^{\infty} 1_{\{W_{\mathcal{K}}(s)=0, W_{\mathcal{K}^c}(s)>\eta\}}dU^{\mathcal{L}}(s) = 0 \text{ almost surely.} \tag{3.147}$$

For this, fix an $\eta \in \mathbb{R}_+^{N-k}$ with $\eta > 0$, and define a sequence of stopping times $\{T_m\}_{m=1}^{\infty}$

as follows. (Here, for notational convenience, we regard the entries in η as being indexed by $i \in \mathcal{K}^c$.)

$$\begin{aligned} T_0 &\triangleq 0, \\ T_1 &\triangleq \inf\{s \geq 0 : W_i(s) < \eta_i/2 \text{ for some } i \in \mathcal{K}^c\}, \\ T_2 &\triangleq \inf\{s \geq T_1 : W_{\mathcal{K}^c}(s) > \eta\}, \end{aligned} \tag{3.148}$$

and for $m \geq 1$,

$$\begin{aligned} T_{2m+1} &\triangleq \inf\{t \geq T_{2m} : W_i(s) < \eta_i/2 \text{ for some } i \in \mathcal{K}^c\}, \\ T_{2m+2} &\triangleq \inf\{t \geq T_{2m+1} : W_{\mathcal{K}^c} > \eta\}. \end{aligned} \tag{3.149}$$

By the continuity of the paths of W, $T_m \to \infty$ as $m \to \infty$, and we have almost surely:

$$\int_0^\infty 1_{\{W_{\mathcal{K}}(s)=0, W_{\mathcal{K}^c}(s)>\eta\}} dU^{\mathcal{L}}(s) \leq \sum_{m=0}^{\infty} \int_{T_{2m}}^{T_{2m+1}} 1_{\{W_{\mathcal{K}}(s)=0\}} dU^{\mathcal{L}}(s). \tag{3.150}$$

Consider $m \geq 0$. Then on $\{T_{2m} < \infty\}$, for $\emptyset \neq \mathcal{M} \subseteq \mathcal{N}$, $\mathcal{M} \nsubseteq \mathcal{K}$, $U^{\mathcal{M}}$ can increase only when $W_{\mathcal{M}} = 0$ and so, almost surely, for all such $\mathcal{M}$,

$$U^{\mathcal{M}}(t + T_{2m}) - U^{\mathcal{M}}(T_{2m}) = 0 \text{ for all } t \in [0, T_{2m+1} - T_{2m}]. \tag{3.151}$$

Thus, on $\{T_{2m} < \infty\}$, we have almost surely for all $t \in [0, T_{2m+1} - T_{2m}]$

$$\begin{aligned} W_{\mathcal{K}}(t + T_{2m}) - W_{\mathcal{K}}(T_{2m}) = X_{\mathcal{K}}(t + T_{2m}) - X_{\mathcal{K}}(T_{2m}) \\ + \sum_{\emptyset \neq \mathcal{M} \subseteq \mathcal{K}} \gamma_{\mathcal{K}}^{\mathcal{M}} (U^{\mathcal{M}}(t + T_{2m}) - U^{\mathcal{M}}(T_{2m})). \end{aligned} \tag{3.152}$$

Then Itô's formula, (3.115), holds on $\{T_{2m} < \infty\}$ for $f \in C^2(\mathbb{R}^k_+)$ with $(X, \{U^{\mathcal{L}} : \emptyset \neq \mathcal{L} \subseteq \mathcal{N}\}, W)$ and $\{\gamma^{\mathcal{L}} : \emptyset \neq \mathcal{L} \subseteq \mathcal{N}\}$ replaced by $(X_{\mathcal{K}}, \{U^{\mathcal{M}} : \emptyset \neq \mathcal{M} \subseteq$

$\mathcal{K}\}, W_{\mathcal{K}})((\cdot + T_{2m}) \wedge T_{2m+1})$ and $\{\gamma_{\mathcal{K}}^{\mathcal{M}} : \emptyset \neq \mathcal{M} \subseteq \mathcal{N}\}$ and with

$$Lf = \frac{1}{2} \sum_{i \in \mathcal{K}} \Gamma_{ii} \frac{\partial^2 f}{\partial x_i^2}. \tag{3.153}$$

The same proof as in Lemma 3.4.14, but with the dimension reduced from N to $k = |\mathcal{K}|$, shows that

$$\sum_{\emptyset \neq \mathcal{M} \subseteq \mathcal{K}} 1_{\{T_{2m} < \infty\}} \int_{T_{2m}}^{T_{2m+1}} 1_{\{W_{\mathcal{K}}(s)=0\}} dU^{\mathcal{M}}(s) = 0 \text{ almost surely}, \tag{3.154}$$

and hence for all $\emptyset \neq \mathcal{M} \subseteq \mathcal{K}$,

$$\int_{T_{2m}}^{T_{2m+1}} 1_{\{W_{\mathcal{K}}(s)=0\}} dU^{\mathcal{M}}(s) = 0 \text{ almost surely on } \{T_{2m} < \infty\}. \tag{3.155}$$

For this, one uses the martingale property of the Brownian motion X and the fact that there is a $\beta^k \in \mathbb{R}_+^k$ and $\delta^{\mathcal{M},k} \in [1, \infty)$ such that $\left\langle \gamma_{\mathcal{K}}^{\mathcal{M}}, \beta^k \right\rangle = \delta^{\mathcal{M},k}$ for any $\emptyset \neq \mathcal{M} \subseteq \mathcal{K}$ (this follows from the fact that (3.107) holds with $\mathcal{K}(x) = \mathcal{K}$ where $n_i^{\mathcal{K}(x)} = 0$ if $i \notin \mathcal{K}(x)$ and $n_i^{\mathcal{K}(x)} = 1/\sqrt{|\mathcal{K}(x)|}$ if $i \in \mathcal{K}(x)$). Substituting (3.155) in (3.150) then yields the desired result. □

Lemma 3.4.16. *Suppose (W, X, U) is as in the hypothesis of Theorem 3.4.13 and $\theta \in \mathbb{R}^N$. Then* (3.113) *holds for all $\emptyset \neq \mathcal{L} \subseteq \mathcal{K} \subseteq \mathcal{N}$ where $|\mathcal{K}| \geq 2$.*

Proof. Let $\mathcal{K} \subseteq \mathcal{N}$ satisfy $|\mathcal{K}| \geq 2$, $\mathcal{L} \subseteq \mathcal{K}$ and $\theta \in \mathbb{R}^N$. Without loss of generality (by considering a canonical representation on path space for example), we may assume that $(\Omega, \mathcal{F})$ is a standard measurable space and for $t \geq 0$, $\mathcal{F}_t \triangleq \sigma\{(W(s), X(s), U(s)) : 0 \leq s \leq t\}$. Let the associated probability measure be P^θ. Then X is a (θ, Γ)-Brownian motion on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P^\theta)$. By the Girsanov transformation (see Ikeda and Watanabe [19, p. 176]), there is a probability measure P^0 on $(\Omega, \mathcal{F})$ such that under P^0, X is a $(0, \Gamma)$-Brownian motion starting from 0 and for each positive integer m, P^θ and P^0 are mutually absolutely continuous on $\mathcal{F}_m$. It follows that W with the probability measure P^0 on

$(\Omega, \mathcal{F}, \{\mathcal{F}_t\})$ is an SRBM with extensive data $(\mathbb{R}^N_+, 0, \Gamma, \{\gamma^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\})$ that starts from the origin. Then by Lemma 3.4.15, for each $\emptyset \neq \mathcal{L} \subseteq \mathcal{K} \subseteq \mathcal{N}$, (3.113) holds almost surely under P^0. But since P^θ and P^0 are mutually absolutely continuous on $\mathcal{F}_m$, it follows that (3.113) holds almost surely under P^θ with m in place of the upper limit ∞ there. Letting $m \to \infty$ yields the desired result. □

Proof of Theorem 3.4.13. Combining Lemmas 3.4.14, 3.4.15, and 3.4.16, we have proved the first part of the theorem.

To prove the second part of the theorem, we use (3.113). From the definition of an SRBM with extensive data (Definition 3.4.3) and the remark following it, W has the form:

$$W(t) = X(t) + \sum_{\emptyset \neq \mathcal{K} \subseteq \mathcal{N}} \gamma^{\mathcal{K}} U^{\mathcal{K}}(t),\ t \geq 0, \tag{3.156}$$

where for $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$,

$$U^{\mathcal{K}}(t) = \int_0^t 1_{F_{\mathcal{K}}}(W(s)) dU^{\mathcal{K}}(s),\ t \geq 0. \tag{3.157}$$

From (3.113), for $\mathcal{K} \subseteq \mathcal{N}$ with $|\mathcal{K}| \geq 2$,

$$U^{\mathcal{K}}(t) = \int_0^t 1_{F_{\mathcal{K}}}(W(s)) dU^{\mathcal{K}}(s) = 0 \text{ almost surely}. \tag{3.158}$$

Thus the only non-trivial terms in the sum in (3.156) are those indexed by $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$ where $|\mathcal{K}| = 1$. Equation (3.114) immediately follows. □

3.4.9 Proof of Theorem 3.4.6

Proof. By Theorems 3.4.8 and 3.4.11, it suffices to prove that whenever (W, X, U) defines an SRBM with extensive data $(\overline{G}, \theta, \Gamma, \{\gamma^{\mathcal{K}}, \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\})$ that starts from the origin, then W is an SRBM with data $(\mathbb{R}^N_+, \theta, \Gamma, R)$ that starts from the origin and the law of the latter is unique.

By Theorem 3.4.13, $W(\cdot)$ has the representation given by (3.114). For $i \in \mathcal{N}$, define

$$Y^i \triangleq c_i^{\emptyset} U^{\{i\}}. \tag{3.159}$$

Note that from (3.111), a.s.,

$$Y^i(t) = \int_0^t 1_{F_i}(W(s))dY^i(s) \text{ for all } t \geq 0. \tag{3.160}$$

Therefore Y satisfies condition (iv) of Definition 3.4.2. From (3.114), (3.159), and the representation for $[\gamma^{\{1\}}, \gamma^{\{2\}}, \ldots, \gamma^{\{N\}}]$ given by (3.95), we have that for $t \geq 0$,

$$W(t) = X(t) + RY(t) \tag{3.161}$$

where by Lemma 3.4.5, R satisfies the HR condition, and W and X satisfy the other conditions of Definition 3.4.2 with $\nu = \delta_0$. Therefore, (W, X, Y) defines an SRBM with data $(\mathbb{R}_+^N, \theta, \Gamma, R)$ that starts from the origin. Since R satisfies the HR condition, by Harrison and Reiman [16], the law of W is unique. It follows that

$$\hat{W}^r \Rightarrow W \text{ as } r \to \infty \tag{3.162}$$

where W is an SRBM with data $(\mathbb{R}_+^N, \theta, \Gamma, R)$ that starts from the origin. □

3.A Proof of Lemma 3.4.12

To prove Lemma 3.4.12, we use Proposition 4.4 of [21]. Specifically, we prove the following lemma, a restatement of condition (II) of Proposition 4.4 in [21], from which Lemma 3.4.12 follows. Our proof of Lemma 3.A.1 is similar to the proof of Lemma 8.4 in Williams [45].

Lemma 3.A.1. *For each $r > 0$, $\hat{X}^r = \check{X}^r + \varepsilon^r$, where ε^r is a process that converges to* 0

in probability as $r \to \infty$, and

(i) $\{\check{X}^r(t) - \check{X}(0) : r > 0\}$ *is uniformly integrable for each* $t \geq 0$*,*

(ii) there is a sequence of constants $\{\theta^r\}$ *in* $\mathbb{R}^N$ *such that* $\lim_{r\to\infty} \theta^r = \theta$*,*

(iii) for each r*,* $\{\check{X}^r(t) - \check{X}^r(0) - \theta^r t, t \geq 0\}$ *is a martingale with respect to the filtration generated by* $(\hat{W}^r, \check{X}^r, \hat{U}^r)$*.*

We need to develop some preliminaries before proving Lemma 3.A.1.

For $r > 0$, $i \in \mathcal{N}$, and $n \in \mathbb{N}$, define

$$A_i^r(n) \triangleq \sum_{k=1}^{n} u_i^r(k), \tag{3.163}$$

where an empty sum is defined to be zero. Then for $r > 0$, the exogenous arrival process is defined for $i \in \mathcal{N}$, and $t \geq 0$, by

$$E_i^r(t) \triangleq \max\{n \geq 0 : A_i^r(n) \leq t\}. \tag{3.164}$$

Recall the definition of $V(\cdot)$ from (3.52).

For each $p \in \mathbb{N}^N$, let

$$\mathcal{G}_p^r \triangleq \sigma\{A^r(\cdot \wedge (p + e_{\mathcal{N}})), V^r(\cdot \wedge p)\} \tag{3.165}$$

where

$$A^r(\cdot \wedge (p + e_{\mathcal{N}})) \triangleq (A_i^r(\cdot \wedge (p_i + 1)) : i \in \mathcal{N}) \tag{3.166}$$

and

$$V^r(\cdot \wedge p) \triangleq (V_i^r(\cdot \wedge p_i) : i \in \mathcal{N}). \tag{3.167}$$

Then $\{\mathcal{G}_p^r : p \in \mathbb{N}^N\}$ is multi-parameter filtration (see [12, Section 2.8]).

Definition 3.A.1. *A multi-parameter stopping time relative to* $\{\mathcal{G}_p^r : p \in \mathbb{N}^N\}$ *is a random variable* τ *taking values in* $\mathbb{N}^N$ *such that*

$$\{\tau = p\} \in \mathcal{G}_p^r \tag{3.168}$$

for all $p \in \mathbb{N}^N$.

Lemma 3.A.2. *For each* $t \geq 0$,

$$\tau^r(t) \triangleq E^r(t) \tag{3.169}$$

is a stopping time relative to $\{\mathcal{G}_p^r : p \in \mathbb{N}^N\}$.

Remark. The reader will note that in defining $\mathcal{G}_p^r$, $e_{\mathcal{N}}$ is added to the argument of $A^r(\cdot)$. This has to be done because we need to know the first $p_i + 1$ interarrival times for the i-th user before we can determine whether $E_i^r(t) = p_i$ or not.

Proof. For $i \in \mathcal{N}$ and $p \in \mathbb{N}^N$,

$$\{E_i^r(t) = p_i\} = \{A_i^r(p_i) \leq t < A_i^r(p_i + 1)\} \in \mathcal{G}_p^r. \tag{3.170}$$

Therefore $\tau^r(t) = E^r(t)$ is a stopping time relative to $\{\mathcal{G}_p^r : p \in \mathbb{N}^N\}$. □

We next show that the diffusion-scaled workload process is adapted to the multi-parameter filtration stopped at the stopping time $\tau^r(r^2t)$. The proof of the following lemma is based on the proof of Lemma 8.3 in [45] that proves the stopping-time property of certain renewal processes for the system of interest. The following Lemma, on the other hand, proves the adaptedness of the workload, a result that in [45], unlike here, follows from the structure of the system.

Lemma 3.A.3. *The process* $\hat{W}^r(\cdot)$ *is adapted to the filtration* $\{\mathcal{G}^r_{\tau^r(r^2t)}, t \geq 0\}$, *where* $\tau^r(r^2t) = E^r(r^2t)$.

Remark. As a consequence of the adaptedness of $\hat{W}^r$, the processes $\{\hat{U}^{r,\mathcal{K}}(\cdot), \emptyset \neq \mathcal{K} \subseteq \mathcal{N}\}$ are adapted to the filtration $\{\mathcal{G}^r_{\tau^r(r^2 t)}, t \geq 0\}$ as well.

Proof. From the definition of $\hat{W}^r(\cdot)$, it suffices to show that W^r is adapted to $\{\mathcal{G}^r_{\tau^r(t)}, t \geq 0\}$. Our proof is for fixed r and so the superscript r will be suppressed in the following proof.

Since $W(0) = 0$ (and for all $\emptyset \neq \mathcal{K} \subseteq \mathcal{N}$, $U^{\mathcal{K}}(0) = 0$), it follows that $W(0)$ and $U(0)$ are $\mathcal{G}_0$-measurable. Furthermore, the process $\{(A(p + e_{\mathcal{N}}), V(p)) : p \in \mathbb{N}^N\}$ is adapted to the multi-parameter filtration $\{\mathcal{G}_p : p \in \mathbb{N}^N\}$. Then by [12, Proposition 2.8.5] and the stopping time property of $\tau(t)$ (Lemma 3.A.2), we have that for each $t \geq 0$:

$$(A(E(t) + e_{\mathcal{N}}), V(E(t))) \in \mathcal{G}_{\tau(t)}. \tag{3.171}$$

Therefore, from (3.53), we only need to show that $T(t)$ (as defined by (3.79)) is adapted to the filtration $\{\mathcal{G}_{\tau(t)}, t \geq 0\}$.

Next, we define a strictly increasing sequence of real-valued random times $\{\eta_l\}_{l=0}^{\infty}$ for the (discrete event) queueing system such that $\eta_l \triangleq 0$ and for $l = 1, 2, \ldots$, η_l is the time of the l-th change in the status of the arrival-departure process pair, i.e., η_l is the l-th time of occurrence of an arrival to, or a departure from, some user. We have $\eta_l < \infty$ for each l, and $\eta_l \to \infty$ as $l \to \infty$. (This follows by the assumption concerning the exclusion of exceptional null sets made at the end of Section 3.4.3.2.)

For each $t \geq 0$, $p \in \mathbb{N}^N$,

$$\{E(t) = p\} = \bigcup_{j=0}^{\infty} \bigcap_{l=j}^{\infty} \{E(t \wedge \eta_l) = p\}. \tag{3.172}$$

For each $l \geq 0$, $p \in \mathbb{N}^N$, define

$$B^l_p \triangleq \{E(t \wedge \eta_l) = p\}. \tag{3.173}$$

Fix $t \geq 0$. It will be shown by induction that for each $l \geq 0$, the following two properties hold for all $p \in \mathbb{N}^N$:

(i) $B_p^l \in \mathcal{G}_p$,

(ii) for

$$\mathcal{I}^l \triangleq (t \wedge \eta_l, E(\cdot \wedge t \wedge \eta_l), T(\cdot \wedge t \wedge \eta_l), W(\cdot \wedge t \wedge \eta_l)), \tag{3.174}$$

we have $1_{B_p^l}\mathcal{I}^l \in \mathcal{G}_p$.

We now proceed with the induction proof. For $l = 0$, one has $\eta_0 = 0$ and $E(0) = 0$. Moreover, for all $p \in \mathbb{N}^N$, $W(0) = 0 \in \mathcal{G}_p$ and $T(0) = 0 \in \mathcal{G}_p$. Then, (i) and (ii) are easily verified to hold for $l = 0$.

For the induction step, assume that for some $l \geq 0$, (i) and (ii) hold for all $p \in \mathbb{N}^N$. Now,

$$B_p^{l+1} = \underset{m}{\cup}(B_p^{l+1} \cap B_m^l) \tag{3.175}$$

where the union is over all $m \in \mathbb{N}^N$ such that $m \leq p$. By the induction assumption, for fixed $p \in \mathbb{N}^N$ and any $m \in \mathbb{N}^N$ such that $m \leq p$, we have

$$B_m^l \in \mathcal{G}_m, \quad 1_{B_m^l}\mathcal{I}^l \in \mathcal{G}_m. \tag{3.176}$$

Hence, from (3.174), $B_m^l \cap \{\eta_l \geq t\} \in \mathcal{G}_m$ and $B_m^l \cap \{\eta_l < t\} \in \mathcal{G}_m$.

Now, on $B_m^l \cap \{\eta_l \geq t\}$, $\eta_{l+1} \wedge t = \eta_l \wedge t$, $E(t \wedge \eta_{l+1}) = E(t \wedge \eta_l) = m$, and $\mathcal{I}^{l+1} = \mathcal{I}^l$. Thus, if $m = p$ we have

$$B_p^{l+1} \cap B_m^l \cap \{\eta_l \geq t\} = B_m^l \cap \{\eta_l \geq t\} \in \mathcal{G}_m, \tag{3.177}$$

or if $m \neq p$, then the left member of (3.177) is the empty set which is still in $\mathcal{G}_m$. Thus,

combining the above with the induction assumption (3.176), we obtain

$$1_{B_p^{l+1} \cap B_m^l \cap \{\eta_l \geq t\}} \mathcal{I}^{l+1} = 1_{\{m=p\}} 1_{B_m^l \cap \{\eta_l \geq t\}} \mathcal{I}^l \in \mathcal{G}_m. \tag{3.178}$$

On the other hand, on $B_m^l \cap \{\eta_l < t\}$, $E(\eta_l) = E(t \wedge \eta_l) = m$ and the first time after η_l that a new external arrival occurs is $\eta = \min_{i \in \mathcal{N}} A_i(m_i + 1)$. Furthermore, on the set $B_m^l \cap \{\eta_l < t\}$, we have

$$\mathcal{I}^l = (\eta_l, E(\cdot \wedge \eta_l), T(\cdot \wedge \eta_l), W(\cdot \wedge \eta_l)). \tag{3.179}$$

Recall that the rate of service given to each of the users over the period $[\eta_l, \eta_{l+1})$ is given by $\sigma^l \triangleq \Lambda(W(\eta_l))$ where, from (3.54), $\Lambda(\cdot)$ is a measurable function on $\mathbb{R}_+^N$. It follows that if we define

$$\zeta \triangleq \eta_l + \inf\{s \geq 0 : W_i(\eta_l) - \sigma_i^l s = 0 \text{ for some } i \text{ such that } \sigma_i^l > 0, i \in \mathcal{N}\}, \tag{3.180}$$

then on $B_m^l \cap \{\eta_l < t\}$, $\eta_{l+1} = \eta \wedge \zeta$ where η_{l+1} is a measurable function of $(A(\cdot \wedge (m + e_{\mathcal{N}})), \eta_l, W(\eta_l))$, and hence by the induction assumption (3.176), (3.179), and the definition of $\mathcal{G}_m$, we have

$$1_{B_m^l \cap \{\eta_l < t\}} \eta_{l+1} \in \mathcal{G}_m. \tag{3.181}$$

Moreover, on $B_m^l \cap \{\eta_l < t\}$, we can express $E(\eta_{l+1})$, $T(\eta_{l+1})$, and $W(\eta_{1+1})$ as measurable functions of η_l, η_{l+1}, $E(\eta_l)$, $A(m + e_{\mathcal{N}})$, and $W(\eta_l)$ as follows. For $i \in \mathcal{N}$,

$$\begin{aligned} E_i(\eta_{l+1}) &= E_i(\eta_l) + 1_{\{A_i(m_i+1)=\eta_{l+1}\}}, \\ T_i(\eta_{l+1}) &= T_i(\eta_l) + \sigma_i^l(\eta_{l+1} - \eta_l), \\ W_i(\eta_{l+1}) &= V_i(E_i(\eta_{l+1})) - T_i(\eta_{l+1}). \end{aligned} \tag{3.182}$$

Since on $[\eta_l, \eta_{l+1})$, E is constant and T and W are linearly increasing/decreasing at a

fixed rate, given by σ^l, on $[\eta_l, \eta_{l+1})$, on combining the above with the induction assumption (3.176), (3.179), and (3.181), we have that

$$1_{B_m^l \cap \{\eta_l < t\}}(\eta_{l+1}, E(\cdot \wedge \eta_{l+1}), T(\cdot \wedge \eta_{l+1}), W(\cdot \wedge \eta_{l+1})) \in \mathcal{G}_m. \tag{3.183}$$

In particular,

$$1_{B_m^l \cap \{\eta_l < t\}}(E(t \wedge \eta_{l+1})) \in \mathcal{G}_m \tag{3.184}$$

and hence

$$B_p^{l+1} \cap B_m^l \cap \{\eta_l < t\} \in \mathcal{G}_m. \tag{3.185}$$

On combining this with (3.177), we see that $B_p^{l+1} \cap B_m^l \in \mathcal{G}_m \subset \mathcal{G}_p$ and hence by (3.175),

$$B_p^{l+1} \in \mathcal{G}_p. \tag{3.186}$$

Thus, (i) holds with $l+1$ in place of l. Similarly,

$$B_p^{l+1} \cap \{\eta_{l+1} \leq t\} = \underset{m}{\cup}(B_p^{l+1} \cap B_m^l \cap \{\eta_l < t\} \cap \{\eta_{l+1} \leq t\}) \in \mathcal{G}_p, \tag{3.187}$$

where the union is over all $m \in \mathbb{N}^N$ such that $m \leq p$.

It remains to verify (ii) with $l+1$ in place of l. But it follows immediately from (3.183) and (3.185) that

$$1_{B_p^{l+1} \cap B_m^l \cap \{\eta_l < t\}}(t \wedge \eta_{l+1}, E(\cdot \wedge t \wedge \eta_{l+1}), T(\cdot \wedge t \wedge \eta_{l+1}), W(\cdot \wedge t \wedge \eta_{l+1})) \in \mathcal{G}_p. \tag{3.188}$$

Combining this with (3.175) and (3.178) yields that

$$1_{B_p^{l+1}} \mathcal{I}^{l+1} \in \mathcal{G}_p. \tag{3.189}$$

□

Proof of Lemma 3.A.1. An outline of our proof is as follows. The idea of the proof of part (iii) is that apart from small error terms associated with residual interarrival times, by suitably centering and scaling the primitive processes (A^r, V^r), we can reexpress $\hat{X}^r$, as given by (3.81), in terms of a martingale evaluated at a stopping time. Indeed, we use the i.i.d. and independence assumptions on the primitive sequences $\{u_i^r(k), k = 1, 2, \dots, \}$, $\{v_i(k), k = 1, 2, \dots\}$, $i \in \mathcal{N}$, to establish the martingale property. In order to conclude that the stopped process is also a martingale, we establish L^2-bounds on the martingale and on the mean of the stopping time $\tau^r(t) = E^r(t)$. The martingale property in part (iii) of the lemma follows from this stopped martingale property and the fact that U^r and W^r are adapted to $\mathcal{G}^r_{\tau^r(t)}$. The asymptotic negligibility of error terms associated with the martingale property of the renewal process $E^r(t)$ is used to show that the residual process converges in probability to 0. Part (ii) of the Lemma follows from the heavy traffic assumption (Assumption 3.4.2). Finally, the uniform integrability property in part (i) follows from L^2 bounds used in obtaining the stopped martingale property mentioned above. Now we provide the details of the proof.

For the moment, let r be fixed. Now,

$$\{\mathcal{G}_p^r\} \triangleq \{\mathcal{G}_p^r : p \in \mathbb{N}^N\} \tag{3.190}$$

defined by (3.165) is a (multi-parameter) filtration and for each $t \geq 0$, by Lemma 3.A.2,

$$\tau^r(t) = E^r(t) \tag{3.191}$$

is a (multi-parameter) stopping time relative to this filtration. If $(\Omega^r, \mathcal{F}^r)$ is the measurable space on which all of the processes indexed by r are defined, then for each $t \geq 0$ we can define a σ-algebra associated with the multi-parameter stopping time $\tau^r(t)$ as follows:

$$\mathcal{G}^r_{\tau^r(t)} \triangleq \{B \in \mathcal{F}^r : B \cap \{\tau^r(t) \leq p\} \in \mathcal{G}_p^r \text{ for all } p \in \mathbb{N}^N\}. \tag{3.192}$$

Then $\{\mathcal{G}^r_{\tau^r(t)}, t \geq 0\}$ is a filtration in the usual single-parameter sense. From Lemma 3.A.3, we have that the process W^r (and hence U^r) is adapted to this filtration.

We now introduce the fundamental multi-parameter martingales $\mathcal{M}^r$ and $\mathcal{O}^r$, and martingales associated with squares of their components. For each $p \in \mathbb{N}^N$ and $i \in \mathcal{N}$, let

$$\mathcal{M}^r_i(p_i) \triangleq \lambda^r_i A^r_i(p_i+1) - (p_i+1), \tag{3.193}$$

$$\mathcal{N}^r_i(p_i) \triangleq (\mathcal{M}^r_i(p_i))^2 - (p_i+1)\alpha^2_i, \tag{3.194}$$

$$\mathcal{O}^r_i(p_i) \triangleq V^r_i(p_i) - p_i m_i, \tag{3.195}$$

$$\mathcal{P}^r_i(p_i) \triangleq (\mathcal{O}^r_i(p_i))^2 - p_i m^2_i \beta^2_i. \tag{3.196}$$

Let $\mathcal{M}^r(p) \triangleq (\mathcal{M}^r_i(p_i) : i \in \mathcal{N})$, $\mathcal{N}^r(p) \triangleq (\mathcal{N}^r_i(p_i) : i \in \mathcal{N})$, $\mathcal{O}^r(p) \triangleq (\mathcal{O}^r_i(p_i) : i \in \mathcal{N})$, $\mathcal{P}^r(p) \triangleq (\mathcal{P}^r_i(p_i) : i \in \mathcal{N})$. Because of the independence and i.i.d. assumptions of Section 3.4.3, we have that the $4N$-dimensional process:

$$\{\mathcal{Q}^r(p) \triangleq (\mathcal{M}^r(p), \mathcal{N}^r(p), \mathcal{O}^r(p), \mathcal{P}^r(p)) : p \in \mathbb{N}^N\}, \tag{3.197}$$

is a multi-parameter martingale relative to $\{\mathcal{G}^r_p\}$.

For each $p \in \mathbb{N}^N$, let

$$\mathcal{R}^r(p) \triangleq (\mathcal{M}^r(p), \mathcal{O}^r(p)). \tag{3.198}$$

We aim to show that $\{\mathcal{R}^r(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\}$ is a martingale. However, we cannot immediately deduce this from the martingale property of $\mathcal{Q}^r$, since $\tau^r(t)$ is a possibly unbounded stopping time. So we first truncate time, apply the multi-parameter stopping theorem and then pass to the limit in the truncation using uniform integrability to deduce the desired result. The bounds obtained for the uniform integrability will also prove useful in verifying part (i) of the lemma. For $n \in \mathbb{N}$, let n^N denote the N-dimensional vector

whose components all have value n. Then, we can verify (in a similar manner to that for $\mathcal{Q}^r$) that

$$\{\mathcal{Q}^{r,n}(p) \triangleq \mathcal{Q}^r(p \wedge n^N) : p \in \mathbb{N}^N\} \tag{3.199}$$

is a multi-parameter martingale relative to $\{\mathcal{G}^r_p\}$. Then by the multi-parameter optional stopping theorem (see [12, Theorem 2.8.7]) we have that

$$\{\mathcal{Q}^{r,n}(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\} \tag{3.200}$$

is a martingale for each $n \in \mathbb{N}$. Now, for $p \in \mathbb{N}^N$ and $n \in \mathbb{N}$, let

$$\mathcal{R}^{r,n}(p) \triangleq (\mathcal{M}^r(p \wedge n^N), \mathcal{O}^r(p \wedge n^N)). \tag{3.201}$$

For each $n \in \mathbb{N}$, it follows from the martingale property of $\{\mathcal{Q}^{r,n}(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\}$ that

$$\{\mathcal{R}^{r,n}(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\} \tag{3.202}$$

is a martingale. We aim to prove that the same is true with $\mathcal{R}^r$ in place of $\mathcal{R}^{r,n}$. For $t \geq 0$ fixed, $\mathcal{R}^{r,n}(\tau^r(t)) \to \mathcal{R}^r(\tau^r(t))$ pointwise as $n \to \infty$, and so it suffices to show that $\{\mathcal{R}^{r,n}(\tau^r(t))\}_{n=1}^{\infty}$ is L^2-bounded for each $t \geq 0$, since this implies that it is uniformly integrable. By the martingale properties of the $\mathcal{N}^r$ and $\mathcal{P}^r$ elements of $\mathcal{Q}^{r,n}(\tau^r(\cdot))$ we have for all $i \in \mathcal{N}$, $n \geq 1$:

$$\mathbf{E}\left[(\mathcal{M}^r_i(E^r_i(t) \wedge n))^2 - ((E^r_i(t) + 1) \wedge n)\alpha_i^2\right] = 0, \tag{3.203}$$

$$\mathbf{E}\left[(\mathcal{O}^r_i(E^r_i(t) \wedge n))^2 - (E^r_i(t) \wedge n)m_i^2\beta_i^2\right] = 0. \tag{3.204}$$

From Lorden's inequality for renewal processes (see Lindvall [27, pp. 77–78]; Carlsson and Nerman [7]), we obtain the following upper bound for $i \in \mathcal{N}$,

$$\mathbf{E}\left[E^r_i(t) + 1\right] \leq \lambda^r_i t + \alpha_i^2 + 2 \triangleq h^r_i(t), \tag{3.205}$$

where $h_i^r(t)$ is finite. It then follows from (3.203)–(3.205) that for all $n \geq 1$, $i \in \mathcal{N}$,

$$\mathbf{E}\left[(\mathcal{M}_i^r(E_i^r(t) \wedge n))^2\right] \leq \alpha_i^2 h_i^r(t), \tag{3.206}$$

$$\mathbf{E}\left[(\mathcal{O}_i^r(E_i^r(t) \wedge n))^2\right] \leq m_i^2 \beta_i^2 h_i^r(t). \tag{3.207}$$

This establishes the desired L^2-boundedness and hence

$$\{\mathcal{R}^r(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\} \tag{3.208}$$

is a martingale for each r.

We now apply the above martingale properties to establish part (iii) of the Lemma. For $i \in \mathcal{N}$, define

$$\check{X}_i^r(t) \triangleq r^{-1}\left(\mathcal{O}_i^r(E_i^r(r^2t)) - m_i\mathcal{M}_i^r(E_i^r(r^2t)) + (\lambda_i^r - \lambda_i)m_i r^2 t\right), \tag{3.209}$$

$$\varepsilon_i^r(t) \triangleq r^{-1}m_i\left(\lambda_i^r A_i^r(E_i^r(r^2t) + 1) - (\lambda_i^r r^2 t + 1)\right), \tag{3.210}$$

$$\theta_i^r \triangleq r(\lambda_i^r - \lambda_i)m_i. \tag{3.211}$$

Then from (3.61), (3.66), (3.67), and (3.81), for $i \in \mathcal{N}$, $t \geq 0$,

$$\hat{X}_i^r(t) = \check{X}_i^r(t) + \varepsilon_i^r(t). \tag{3.212}$$

Since

$$\mathcal{R}^r(\tau^r(r^2t)) = (\mathcal{M}^r(E^r(r^2t)), \mathcal{O}^r(E^r(r^2t))), \tag{3.213}$$

it follows, from the martingale property of (3.208), that

$$\{\check{X}^r(t) - \check{X}^r(0) - \theta^r t, \mathcal{G}^r_{\tau^r(r^2t)}, t \geq 0\} \tag{3.214}$$

is a martingale. Note that by Lemma 3.A.3, $\hat{U}^r$ and $\hat{W}^r$ are adapted to the filtration $\{\mathcal{G}^r_{\tau^r(r^2t)}, t \geq 0\}$. Hence, $\{\check{X}^r(t) - \check{X}^r(0) - \theta^r t, t \geq 0\}$ is a martingale relative to the

filtration generated by $(\hat{W}^r, \check{X}^r, \hat{U}^r)$.

We next show that ε^r converges in probability to the zero process as $r \to \infty$. By the definition of E_i^r from A_i^r for $i \in \mathcal{N}$, for each $T \geq 0$,

$$\|\varepsilon^r(\cdot)\|_T \leq 2 \max_{i \in \mathcal{N}} |m_i \lambda_i^r| \left\| r^{-1} u_i^r (E_i^r(r^2 \cdot) + 1) \right\|_T + \max_{i \in \mathcal{N}} |m_i| / r \tag{3.215}$$

where, as a consequence of the functional central limit theorem (Proposition 3.4.3), the right-hand side above goes to zero in probability as $r \to \infty$ (see the proof of Lemma 6 in Iglehart and Whitt [18]).

Part (ii) of the lemma follows from the heavy traffic assumption (Assumption 3.4.2).

It remains to show part (i) of the lemma. For this it suffices to show that $\check{X}^r(t)$ as r varies is uniformly integrable for each fixed $t \geq 0$. Now by Fatou's lemma, (3.206)–(3.207) hold with the n's removed. Fix $t \geq 0$. By (3.205), we have

$$\sup_r \max_{i \in \mathcal{N}} \frac{h_i^r(r^2 t)}{r^2} < \infty. \tag{3.216}$$

Replacing t by $r^2 t$ in (3.206)–(3.207), and combining with the above, we see that

$$\{r^{-1}(\mathcal{M}^r(E^r(r^2 t)), \mathcal{O}^r(E^r(r^2 t)))\} \tag{3.217}$$

as a collection indexed by r is L^2-bounded, and hence uniformly integrable. The uniform integrability of $\{\check{X}^r(t)\}$ follows. □

3.5 Acknowledgment

Chapter 3, in part, is a reprint of the material in the following paper: S. Bhardwaj, R. J. Williams, and A. S. Acampora, "On the Performance of a Two-User MIMO Downlink System in Heavy Traffic", *IEEE Trans. Information Theory*, vol. 53, no. 5, pp.

1851–1859, May 2007. The dissertation author was the primary investigator and author of this paper.

CHAPTER 4

Conclusions

The problem of increasing the throughput of cellular wireless systems is as old as those systems and the commercial potential of any solution has made the problem well-studied. In this dissertation, we studied this problem from a different perspective. Specifically, we undertook a study of certain topics related to a cross-layer analysis of the downlink of cellular wireless systems with cooperation among base stations where our approach, unlike most of the literature, was based on queueing theory.

In Chapter 2, we investigated the maximum throughput of such a system. Starting with the assumption that the relative traffic of each of the mobiles is specified in advance - an assumption justifiable in light of the fact that higher layers often provide such a specification - we set out to find a descriptor of the maximum amount of data that can be sent to each of the mobiles. Because of the assumed knowledge of the ratio of deliverable traffic, there is a single-parameter descriptor. We first showed that if the capacity region of the underlying channel was convex and constant, the maximum stable throughput of the system can be described in terms of the differentiated service capacity of the channel.

A queueing model for the cellular wireless system with infrastructure cooperation was formulated in Section 2.1.3. The queueing network was a multi-class coupled queueing system with variable instantaneous service rate. With the same assumptions as before - quasi-static system with number of users constan - we proposed a policy that was throughput optimal, that is, the long run average departure rate exists and equals the long run average arrival rate whenever the nominal load is less than the maximum stable

throughput. Unfortunately, because of the coupled nature of the queueing network, it was not possible to perform an exact analysis of the operation of the queueing system under this policy. Therefore, we proposed a fixed-point approximation in Section 2.1.5 where the original coupled queueing system is replaced by a queueing network comprising of independent $M/M/1$ queues with the variable service rate of the queues in the original network replaced by constant service rates.

Through simulations, we were able to demonstrate the efficacy of cooperation in cellular wireless systems. In our first set of results, we presented the gain in throughput achieved by base station cooperation for a fixed channel for a cellular system with two base stations and two users. It was noticed that the throughput gain increased with SNR. We next studied the average system delay for different system loads for different system configurations. For a system with three base stations and three users, we noted that the throughput gain was between 20% and 70% depending on the relative traffic vector and the channel while it was approximately three for all of the cases for a system with four base stations and four users. It was observed that the fixed-point approximation gave a very good approximation of the system delay in the high and the low-load region. Thus, for large systems which are difficult to simulate, the fixed-point approximation can be used, especially for the important high-load region since the maximum throughput predicted by fixed-point approximation is the same as the maximum stable throughput of the system. We next studied the outage probability for different system configurations. It was noted that as the probability of outage increased, the increased in throughput was not very significant. In fact, for most of the systems the throughput gain decreased as the outage probability increased.

We next turned our attention to finding methods to increase throughput in large systems. Because of the computational complexity, a straightforward application of the service policy applicable for small systems is not possible for large systems. Therefore, we proposed a suboptimal policy that could be applied for large systems. Specifically, we proposed grouping the mobiles in the footprint of a group of cooperating base stations in

small subgroups and then applying the service policy proposed earlier on the subgroups. Clearly, this is practicable subject to an effective and inexpensive grouping mechanism. Building upon some observations from the study of small-sized systems, we were able to devise one such grouping scheme.

We next presented simulation results showing the effectiveness of our proposed scheme. For a cellular system with four base stations and two hundred (200) mobiles, with groups sizes of three (3) and four (4), we were able to get twofold gain in throughput. Moreover, the variance of the results was low enough to give confidence in our results. We next compared the effect of SNR on the gain in throughput for the same configuration. As expected the gain in throughput - not just the absolute throughput - increased with SNR. Moreover, higher SNR lead to a reduced variance in the gain in throughput. Based on these observations, we believe our scheme can be used in practical systems.

In Chapter 3, we studied the performance of the queueing policy proposed in Chapter 2. Since an exact analysis of the performance of the policy was not possible, we proved limit theorems justifying a diffusion approximation for a heavily loaded system operating under this policy. We started with the simple case of a two-user system where there were only four (4) operation points. We first proved a fluid limit result for our queueing system. This result played a role in establishing the heavy traffic limit theorem through determining the fluid scale service allocations. Then we proved the main theorem for the two-user system which said that in the heavy traffic limit, the renormalized queuelength process converged in distribution to an SRBM living in a two-dimensional quadrant. Then we discussed the properties of the limiting process.

We next analyzed the performance of the policy proposed in Section 2.1.4 for an arbitrary sized system. We proved that the renormalized workload process converged in distribution to an SRBM living in an N-dimensional positive quadrant where N was the number of users (queues) in the system. To prove this theorem, we first showed that the sequence of diffusion-scaled processes was C-tight. We then showed that any weak

limit point of such a subsequence is an SRBM with extensive data. For an SRBM with extensive data, there might be pushing at the intersection of two or more faces. We showed that such pushing is negligible and the SRBM with extensive data reduces to one of the simpler form as described in our main result. Finally, we showed that such an SRBM is unique in law and when combined with the C-tightness, we concluded that the sequence converged in distribution to an SRBM of desired form.

REFERENCES

[1] F. Avram, J. G. Dai, and J. J. Hasenbein, "Explicit solutions for variational problems in the quadrant," *Queueing Systems Theory Appl.*, vol. 37, no. 1–3, pp. 259–289, 2001.

[2] S. L. Bell and R. J. Williams, "Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy," *Ann. Appl. Probab.*, vol. 11, no. 3, pp. 608–649, 2001.

[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*. New York: Academic Press, 1979.

[4] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1992.

[5] P. Billingsley, *Convergence of probability measures*, 2nd ed. New York: Wiley, 1999.

[6] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[7] H. Carlsson and O. Nerman, "An alternative proof of Lorden's renewal inequality," *Advances in Appl. Probability*, vol. 18, no. 4, pp. 1015–1016, 1986.

[8] W. Choi and J. G. Andrews, "The capacity gain from intercell scheduling in multiantenna systems," *IEEE Transactions on Wireless Communications*, vol. 7, no. 2, pp. 714–725, Feb. 2008.

[9] K. L. Chung and R. J. Williams, *Introduction to Stochastic Integration*, ser. Probability and its Applications. Boston, MA: Birkhäuser, 1990.

[10] M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. IT-29, no. 3, pp. 439–441, 1983.

[11] J. G. Dai and J. M. Harrison, "Reflected Brownian motion in an orthant: Numerical methods for steady-state analysis," *Ann. Appl. Probab.*, vol. 2, no. 1, pp. 65–86, 1992.

[12] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, ser. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons Inc., 1986.

[13] N. Gans and G. van Ryzin, "Optimal dynamic scheduling of a general class of parallel-processing queueing systems," *Advances in Appl. Probability*, vol. 30, no. 4, pp. 1130–1156, Dec. 1998.

[14] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[15] A. Goldsmith, *Wireless Communications*. Cambridge, MA, USA: Cambridge University Press, 2005.

[16] J. M. Harrison and M. I. Reiman, "Reflected Brownian motions on an orthant," *Ann. Probab.*, vol. 9, no. 2, pp. 302–308, 1981.

[17] J. M. Harrison and R. J. Williams, "Brownian models of open queueing networks with homogeneous customre populations," *Stochastics*, vol. 22, pp. 77–115, 1987.

[18] D. L. Iglehart and W. Whitt, "The equivalence of functional central limit theorems for counting processes and associated partial sums," *Ann. Math. Statist.*, vol. 42, no. 4, pp. 1372–1378, 1971.

[19] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., ser. North-Holland Mathematical Library. Amsterdam, Holland: Birkhäuser, 1989.

[20] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels," *IEEE Transactions on Information Theory*, vol. 50, no. 05, pp. 768–783, May 2004.

[21] W. Kang and R. J. Williams, "An invariance principle for semimartingale reflecting Brownian motions in domains with piecewise smooth boundaries," *Ann. Appl. Probab.*, vol. 17, no. 2, pp. 741–779, 2007.

[22] F. P. Kelly, "Fixed point models of loss networks," *Journal of Australian Mathematical Society Series B*, vol. 31, pp. 204–218, 1989.

[23] F. P. Kelly and C. N. Laws, "Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling," *Queueing Systems Theory Appl.*, vol. 13, no. 1–3, pp. 47–86, 1993.

[24] C. Knessl, "On the diffusion approximation to two parallel queues with processor sharing," *IEEE Transactions on Automatic Control*, vol. 36, no. 12, pp. 1356–1357, Dec. 1991.

[25] A. G. Konheim, I. Meilijson, and A. Melkman, "Processor-sharing of two parallel lines," *J. Appl. Probability*, vol. 18, no. 4, pp. 952–956, 1981.

[26] J. Lee and N. Jindal, "Symmetric capacity of mimo downlink channels," in *Proc. of the IEEE ISIT*, July 2006.

[27] T. Lindvall, *Lectures on the coupling method*, ser. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons Inc., 1992.

[28] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, ser. Communications and Control Engineering Series. London: Springer-Verlag, 1993.

[29] M. J. Neely, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, Nov. 2003.

[30] B. L. Ng, J. S. Evans, S. V. Hanly, and D. Aktas, "Transmit beamforming with cooperating base stations," in *Proc. of IEEE Internation Symposium on Information Theory*, Adelaide, Australia, Sept.4–9 2005, pp. 1431–1435.

[31] W. P. Peterson, "A heavy traffic limit theorem for networks of queues with multiple customer types," *Math. Oper. Res.*, vol. 16, no. 1, pp. 90–118, 1991.

[32] Y. V. Prokhorov, "Convergence of random processes and limit theorems in probability theory," *Theory Probab. Appl.*, vol. 1, no. 2, pp. 157–214, 1956.

[33] M. I. Reiman and R. J. Williams, "A boundary property of semimartingale reflecting Brownian motions," *Probab. Theory Related Fields*, vol. 77, no. 1, pp. 87–97, 1988.

[34] M. Rupf and J. L. Massey, "Optimum sequence multisets for synchronous code-division multiple-access channels," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1261–1266, July 1994.

[35] H. Sato, "An outer bound to the capacity region of broadcast channels," *IEEE Transactions on Information Theory*, vol. IT-24, no. 3, pp. 374–377, May 1978.

[36] M. Schwartz, *Mobile Wireless Communications*. Cambridge, MA, USA: Cambridge University Press, 2005.

[37] S. Shakkotai, R. Srikant, and A. L. Stolyar, "Pathwise optimality of the exponential scheduling rule for wireless channels," *Advances in Appl. Probability*, vol. 36, no. 4, pp. 1021–1045, 2004.

[38] S. Shamai (Shitz) and B. M. Zaidel, "Enhancing the cellular downlink capacity via co-processing at the transmitting end," in *Proc. of Spring IEEE Vehicular Technology Conf.*, vol. 3, Rhodes, Greece, May6–9 2001, pp. 1745–1749.

[39] A. L. Stolyar, "Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic," *Ann. Appl. Probab.*, vol. 14, no. 1, pp. 1–53, 2004.

[40] D. N. C. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, MA, USA: Cambridge University Press, 2005.

[41] S. R. S. Varadhan and R. J. Williams, "Brownian motion in a wedge with oblique reflection," *Comm. Pure Appl. Math.*, vol. 38, no. 4, pp. 405–443, 1985.

[42] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 49(10), pp. 2658–2668, Oct. 2003.

[43] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 09, pp. 3936–3964, Sept. 2006.

[44] A. R. K. Whitley, "Skorokhod problems and semimartingale reflecting stable processes in an orthant," Ph.D. dissertation, University of California, San Diego, 2003.

[45] R. J. Williams, "Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse," *Queueing Systems Theory Appl.*, vol. 30, no. 1–2, pp. 27–88, 1998.

[46] W. Yu and J. M. Cioff, "Trellis precoding for the broadcast channel," in *Proc. of IEEE Globecom 2001*, vol. 2, 25–29 Nov. 2001, pp. 1344–1348.